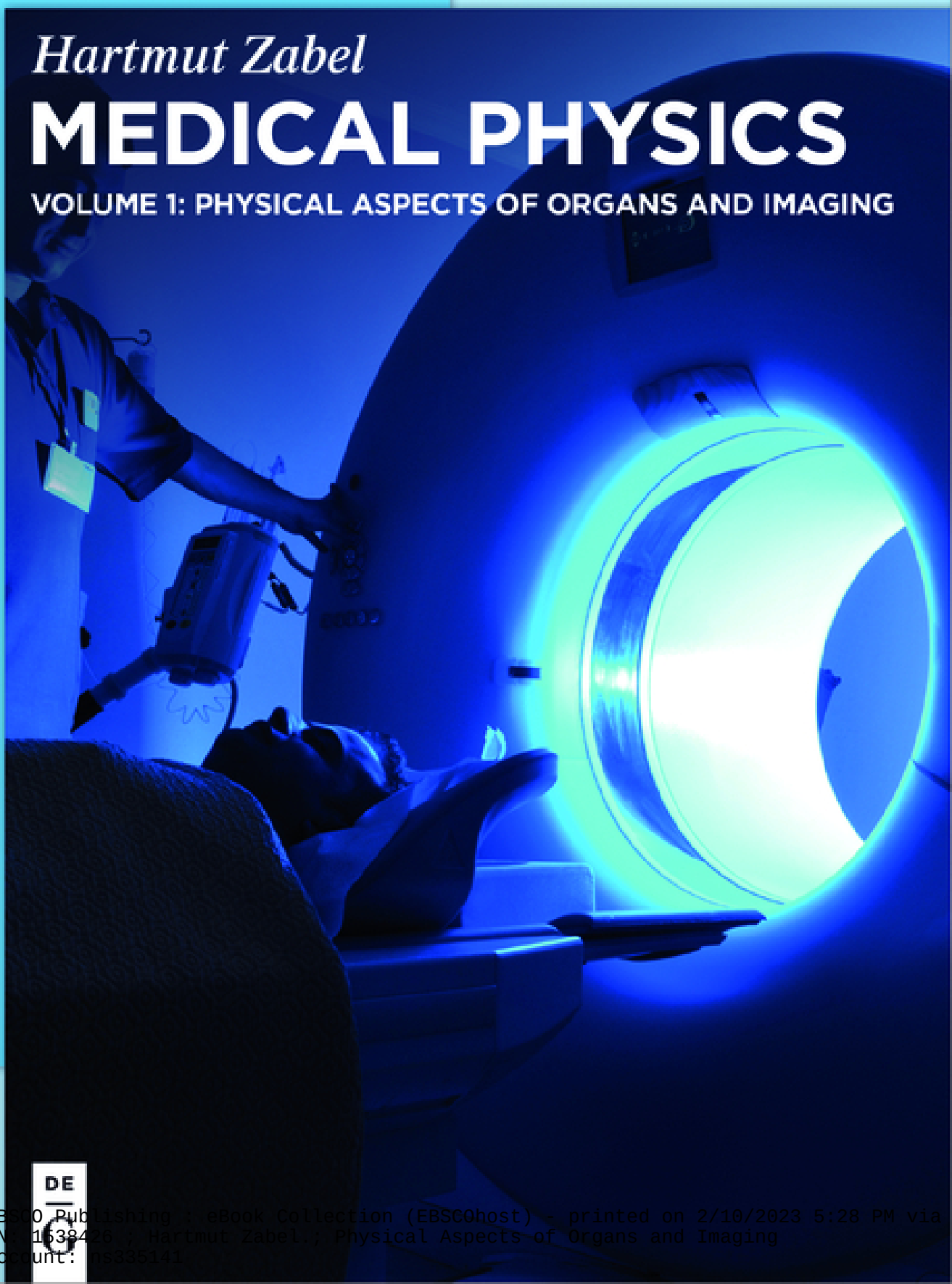
DE GRUYTER

GRADUATE

Hartmut Zabel

MEDICAL PHYSICS

VOLUME 1: PHYSICAL ASPECTS OF ORGANS AND IMAGING



Hartmut Zabel

Medical Physics

De Gruyter Graduate

Also of Interest

Medical Physics.
Volume 2: Radiology, Lasers, Nanoparticles and Prosthetics
Hartmut Zabel, 2017
ISBN 978-3-11-055310-9, e-ISBN (PDF) 978-3-11-055311-6
Volume 1 & Volume 2: also available as a set.
Set-ISBN: 978-3-11-055957-6

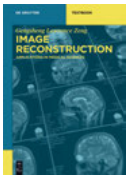
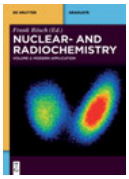
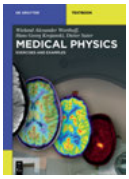


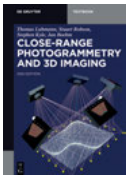
Image Reconstruction.
Applications in Medical Sciences
Gengsheng Lawrence Zeng, 2017
ISBN 978-3-11-050048-6, e-ISBN (PDF) 978-3-11-050059-2



Nuclear- and Radiochemistry.
Volume 2: Modern Applications
Frank Rösch (Ed.), 2016
ISBN 978-3-11-022185-5, e-ISBN (PDF) 978-3-11-022186-2



Medical Physics.
Exercises and Examples
Wieland Alexander Worthoff, Hans Georg Krojanski, Dieter Suter, 2013
ISBN 978-3-11-030675-0, e-ISBN (PDF) 978-3-11-030676-7



Close-Range Photogrammetry and 3D Imaging.
Thomas Luhmann, Stuart Robson, Stephen Kyle, Jan Boehm, 2013
ISBN 978-3-11-030269-1, e-ISBN (PDF) 978-3-11-030278-3

Hartmut Zabel

Medical Physics

Volume 1: Physical Aspects of Organs and Imaging

DE GRUYTER

Physics and Astronomy Classification Scheme 2010
Medical imaging: 87.57.-s
Magnetic resonance imaging: 87.61.-c
Ultrasonography: 87.63.dh
Fiber optical imaging: 42.81.Uv
Biomechanics: 87.85.G-
Electrophysiology in neuroscience: 87.19.lb
Hemodynamics, 87.19.U-

Author
Prof. Dr. Dr. h. c. Hartmut Zabel
Ruhr-University Bochum
Institute for Experimental Physics
44780 Bochum
hartmut.zabel@ruhr-uni-bochum.de

ISBN 978-3-11-037281-6
e-ISBN (PDF) 978-3-11-037283-0
e-ISBN (EPUB) 978-3-11-037285-4

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Cover image: LeventKonuk/iStock/thinkstock
Typesetting: PTP-Berlin, Protago-TEX-Production GmbH, Berlin
Printing and binding: CPI books GmbH, Leck
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

These volumes are dedicated to my grandchildren
Hanna, Henriette, Katharina, Niklas, Raphael, and Sonja.
Now it's your turn again.

Preface

According to the American Association of Physicists in Medicine, Medical Physics is an applied branch of physics concerned with the application of concepts and methods of physics to diagnostics and therapeutics of human diseases. In recent years many academic institutions have started offering Medical Physics in their Bachelor, Master, and PhD programs in response to the expanding job market and increasing demand for professionals in this discipline. The Volumes 1 and 2 on Medical Physics are introductory texts intended to guide bachelor students during their first semesters through the broad range of topics relevant in this field. There are many excellent books available that specialize on more specific topics of Medical Physics, such as imaging and radiation therapy. However, there is rarely a single text covering the entire field, including the physics of the body, imaging methods, radiology, laser methods, nanomedicine, and prosthetics. Because of the multitude of topics to be covered and the rapid development of the field chances are that such a script might be incomplete or partially outdated by the time of publication. Nevertheless, these two volumes try to provide a reasonably balanced overview as well as guidance to more specialized and more comprehensive literature. Each chapter is rounded off with a summary, references, hints for further reading and useful webpages, and with questions and answers to the respective chapters. Historical remarks are kept at a minimum to keep the text at a reasonable total length.

This textbook does not provide any medical interpretation or any medical judgement. It is entirely focused on methods and procedures from a physical point of view. Furthermore, this text is neither an introduction into elementary physics nor into medical physiology. It tries to connect the principles of physics with the functionality of the body and with physical methods used for diagnostics and therapeutics. Furthermore, medical physics is located somewhere in between medical sciences and engineering with regard to diagnostic and therapeutic equipment. As such, medical physics acts as a mediator between these disciplines. While the physical principles of medical equipment and instrumentation are introduced, for more detailed engineering aspects of hardware and software developments references are provided for further reading. Throughout the text it is assumed that the reader has a basic understanding of physics corresponding to about one to two years of college physics.

Physics and medicine have their own distinct terminology. Here we will mainly use a “physics language”, but medical and physiological terms will not be avoided. In fact it is important for students of medical physics to know at least the most important terms in order to be able to communicate with physicians in medical practice. Therefore in these volumes medical and physiological terms will be introduced and defined whenever they are used. In addition, a list of acronyms and definitions is provided in the appendix. In physiology it is still common practice to use non-MKS (SI) units, such as mmHg for pressure. In this text we strictly adhere to the SI system with one

DOI 10.1515/9783110372830-001

exception: for practical reasons we make use of the energy unit 'electron volt' (eV) in the context of electromagnetic and particle radiation.

The present Volume 1 contains 15 chapters divided into two parts. The first 12 chapters in Part A focus on physical and physiological aspects of body parts, organs, and sensors. This is a selection of body systems that have a strong physical component, such as body mechanics including bones and muscles, energy household, electrophysical aspects, circulatory system, respiration, kidneys, and the sensory systems for light and sound. Fundamental to all these topics is an understanding of the resting potential and the action potential and how the action potential travels along nerve fibers for communication and motor control. Part B covers imaging modalities without the use of ionizing radiation, which includes sonography, endoscopy, and magnetic resonance imaging (MRI). Sometimes MRI is also listed under radiography. But this is not justified as the typical frequencies involved are in the GHz regime, far away from ionizing conditions. For reading and learning about imaging modalities in Part B a basic understanding of wave propagation, interference phenomena, and resonance conditions would be beneficial.

Volume 2 Part A complements the imaging modalities by using ionizing radiation: x-ray radiography, scintigraphy with γ-rays, and positron emission tomography. The other parts of Volume 2 treat external beam radiation therapy with x-rays, protons, and neutrons as well as brachytherapy. Volume 2 concludes with chapters on diagnostics and therapeutics beyond radiology: laser applications, multifunctional nanoparticles, and prosthetics.

Each chapter is concluded by a summary of the main points. A number of contextual questions pertaining to each chapter are listed in the appendix. They are complemented by corresponding answers as a guide and for providing an incentive for further studies.

The more I read, the more I recognize the immense scientific activity that has taken place in the past in the field of medical physics and that will continue to take place in the future, and the little that I know about it. Any suggestions and hints for missing aspects or inadequate presentations are highly welcome. Nevertheless, I would like to express my hope that this introductory text may turn out to be a useful companion to students enrolled in a course on Medical Physics during their first semesters.

Bochum and Mainz, January 2017

Acknowledgments

Writing this text would not have been possible without the help of many. First, I am highly thankful to authors and publishers who make their work freely accessible in open access journals, and to the organization Wikipedia including their contributors who do an amazing job for the benefit of those who are eager to know. Furthermore, I am deeply indebted to colleagues and experts who devoted their precious time to critically reading and correcting various chapters. In particular I would like to thank, in alphabetic order, Professor Ping Ao and Dr. Xiaomei Zhu, Shanghai Jiao Tong University, for sharing their deep insight into the causes and development of cancer; Dr. Andrea Denker, Helmholtz Zentrum Berlin, for pointing out intricacies of proton therapy and eliminating misconceptions; Professor Helmut Ermert, University of Erlangen, for sharing his expertise of ultrasonic imaging; Professor Ulf Eysel, Ruhr-University Bochum, for critically reading and correcting the chapter on the mechanism of vision; Professor Werner Havers, University Clinic Essen, for improving the chapter on x-ray radiotherapy; Dr. Margot Jonas, University Clinic Knappschaft Bochum, for valuable suggestions concerning scintigraphy and positron annihilation tomography, and for sharing most instructive images on the comparison of x-ray imaging versus PET; Professor Jan Meijer, University Leipzig, for important amendments to the chapter on dosimetry; Professor Werner Meyer and Dr. Gerhard Reicherz, Ruhr-University Bochum, for their critical reading of and valuable contributions to the chapter on magnetic resonance imaging; Professor Winfried Petry and Professor Franz Wagner, Technical University Munich, for their modifications of the chapters on neutron radiotherapy and on nuclei and isotopes, and for providing proper numbers of neutron energies and dose distributions; Professor Franz Pfeiffer, Technical University Munich, for his improvements to the chapter on x-ray radiography and sharing images on phase contrast imaging; Professor Lutz Pott, Ruhr-University Bochum, for critically reading and rectifying the chapter on electrophysical aspects of the heart; Professor Katharina Theis-Bröhl, University Bremerhaven, for valuable suggestions towards improvement of the chapter on x-ray generators; Dr. Sebastian Tripple, CFEL Hamburg, for his expert input and revisions of the chapter on laser applications. Last but not least, I am grateful to my wife Rosemarie for reading and correcting all chapters with respect to spelling and solecisms. For any remaining inaccuracies or ambiguities I take the sole responsibility.

Furthermore, I would like to thank Professor Mathias Kläui for his hospitality and that provided by the Mainz Graduate School of Excellence at the Johannes Gutenberg University Mainz while this text was being written. My hearty thanks also go to Astrid Seifert and Nadja Schedensack of de Gruyter Verlag, who patiently waited for the submission of this manuscript, and Anne Hirschelmann and her staff, who did a superb job on layout and graphical enhancement of the figures.

DOI 10.1515/9783110372830-002

Contents

Part B: **Imaging modalities without ionizing radiation**

Part A: **Physical and physiological aspects of the body**

1 Brief overview of body parts and functions

1.1 Introduction

Body parts and functions are, as we know, controlled by chemical and biochemical processes. It may be less evident that our body also strictly follows physical principles. This becomes obvious when inspecting different parts of our body. Any movement requires forces, torques, and mechanical stability. The energy household of the body follows the principles of thermodynamics. The sensory system including hearing and sight is based on the principles of acoustics and optics, respectively. The respiratory system, blood circulation, and kidneys all obey the laws of diffusion, hydrostatics, and hydrodynamics. Signal propagation along nerve fibers can be described by electrical circuitry, etc. Part A of this book comprising Chapters 2–12 is dedicated to the physiology of some organs, sensors and systems, with special emphasis on their connection to basic physical principles. This can and will not replace textbooks on physiology but it will provide a solid background for the comprehension of the remaining Part B of Volume 1 and Parts A–C of Volume 2, dealing with imaging, radiotherapy, and prosthetics. Before starting with the kinematics of the body in Chapter 2, we briefly review the basic building blocks, some organs, and systems of the human body.

1.2 Overview

1.2.1 Cells

The human body consists of about 60×10^{12} *cells*. They form the building blocks of the body like bricks of a house. The cross section of a cell is shown in Fig. 1.1. All cells contain double-stranded helical nucleic acid chains known as deoxyribonucleic acid (DNA). Each DNA molecule contains a sequence of paired nucleobases, which read like letters of the genetic code. About 20 000 genes in the DNA encode about 2×10^6 proteins in the human body that do their daily job. They build ion channels and molecular motors, they form receptors, enzymes and hormones, they take care of oxygen transport, strengthen tissues and bones in the body, regulate water and ion concentrations, and are responsible for many more tasks. Proteins are big; some contain more than 100 000 atoms. Furthermore, they are folded up from a chain of amino acids into complex quaternary structures. So far only a few proteins have been described with atomic resolution. Although almost all cells contain the complete and identical genetic information, they specialize in different tasks. Muscle cells develop a surprising tensile force, liver cells are specialized in performing important tasks for the metabolism of food, and nerve cells can transmit electrical signals as fast as 100 m/s. This specialization and combined action of certain tasks is only pos-

DOI 10.1515/9783110372830-003

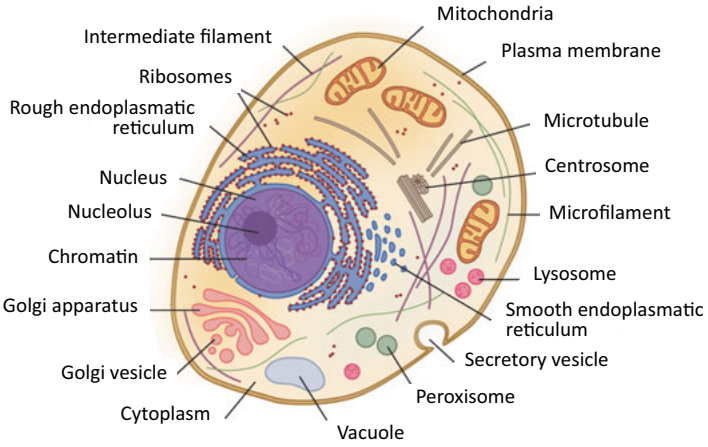


Fig. 1.1: Schematic of a human cell. The cell membrane separates the cytoplasm from the extracellular space. The cytoplasm contains among many other parts the genetic information inside the nucleus and the mitochondria with its own genetic information, acting as a powerhouse for the cell functions. The cell membrane is a double lipid layer that is perforated by a number of ion channels for maintaining an electrical potential across the cell membrane, or rather changing it by depolarization upon a stimulus (adapted from OpenStax Anatomy and Physiology, 2016, © Creative Commons).

sible through a high degree of self-organization and communication among the cells. Comparing cells with bricks is, after all, a gross oversimplification. When we build a complex system, then we indeed start with simple building blocks that fit together to form something more complex. In contrast, in organisms the complexity does not start at the cell level but already at the molecular level of DNA, ribosomes, lysosomes and many more. This incredible complexity organizes life. First, cells organize to form tissues: epithelial tissue, connective tissue, muscular tissue, and nervous tissue. Tissues, in turn, assemble to form organs with characteristic and distinct shapes and functions like the liver and the kidneys. Several organs work together in a system, such as the digestive system that involves ten different organs. The body contains eleven distinguishable systems: (1) cardiovascular/circulatory system; (2) respiratory system; (3) digestive system/excretory system; (4) endocrine system; (5) lymphatic system/immune system; (6) sensory system; (7) locomotor system; (8) nervous system; (9) renal system/urinary system; (10) integumentary system; (11) reproductive system. All eleven systems interact and are responsible for the life of the human body. These interactions take place under the promise of constancy in a variable environment. For instance, core body temperature, arterial blood pressure, and blood partial oxygen/carbon dioxide pressures are kept constant by an active negative feedback system, like a thermostat. This control mechanism is known as homeostatic control or homeostasis. Body temperature homeostasis is further discussed in Chapter 4.

1.2.2 Circulation

For maintaining all body functions, blood *circulation* is essential. No blood circulation, no life. As the name implies, blood circulation is a closed circuit, where the flow is mechanically powered by the heart acting as a pump. Any body part is penetrated by a dense mesh of blood vessels. Blood circulation is a transport system for oxygen, nutrients, and heat to their destinations: the muscles, organs, and bones. Oxygen is taken up from the surrounding air during inhalation and binds to hemoglobin by diffusing through membranes in the lung. In return, during expiration carbon dioxide as a residue of combustion is exhaled. A simplified schematic of blood circulation is shown in Fig. 1.2. It has an oxygen rich and an oxygen poor part, which goes along with a high pressure part and a low pressure part, respectively. The color coding (red: oxygen rich, blue: oxygen poor) originates from our visual perception. Oxygen rich blood has a much brighter red color than oxygen poor blood. Circulation is discussed in more depth in Chapter 8.

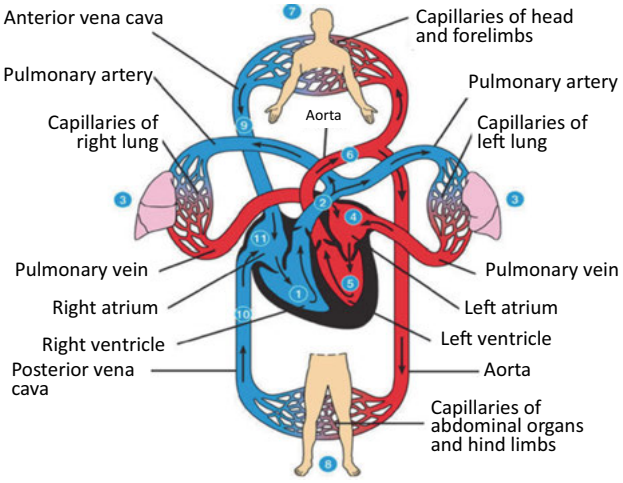


Fig. 1.2: Schematic of the human circulatory system. The right ventricle of the heart (1) takes in oxygen poor blood from the extremities (9, 10) and pumps it into the lungs (2, 3) for oxygen enrichment. After returning into the left atrium and ventricle of the heart (4, 5), the blood is ejected into the periphery (6–8) to supply oxygen to all body tissues. Blood pressure in the left circulation (red) is higher than in the right circulation (blue) at similar locations. The muscle of the left ventricle is also stronger than the muscle of the right ventricle (reproduced from www.online-sciences.com/, © Creative Commons).

1.2.3 Heart

The *heart* has an electrical and a mechanical component. The electrical component takes care that after self-stimulation of the sinus-atrial node the excitation is distributed over the entire heart muscle (myocardium) for sequential contraction. The contracting part of the myocardium increases the blood pressure and performs volume work, i.e. it ejects the cardiac volume periodically into the arteries and veins, thereby maintaining a pulsatile blood flow. A cross section of the heart is shown in Fig. 1.3. The electrical and the mechanical components are topics of Chapters 7 and 8, respectively. The heart and circulation form one unit, known as the *cardiovascular system*.

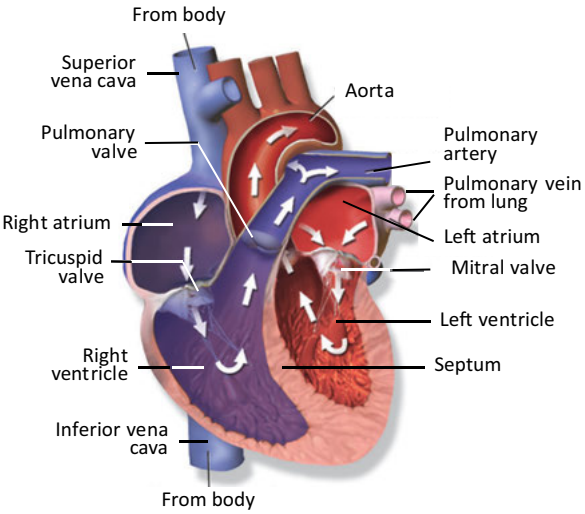


Fig. 1.3: Main parts of the heart, consisting of four chambers (left atrium, right atrium, left ventricle, right ventricle) and four valves (tricuspid, pulmonary, mitral, aortic) (adapted from *Wikiversity Journal of Medicine*, Blausen gallery 2014, © Creative Commons).

1.2.4 Kidneys

The *kidneys* are part of the urinary system that also contains ureters, bladder, and urethra. The main function of the kidneys is the filtering of blood with respect to residual products of metabolism and the elimination of toxic substances resulting from protein breakdown, such as urea and urea acid. Everything that is filtered out goes into the urinary tract. The location of the kidneys in the body is shown in Fig. 1.4.

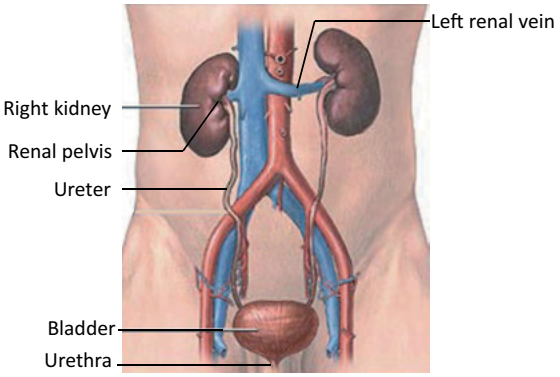


Fig. 1.4: Location of the kidneys in the body. The kidneys are connected to the blood circulatory system and receive about 20 % of the cardiac output, which sums up to about 180 liters per day. After filtering, most of the fluid is resorbed, only 1–2 liters per day is collected in the bladder and excreted through the urethra (adapted from https://digestplus.files.wordpress.com/).

Another important task of the kidneys is the maintenance of the sodium and potassium ion concentration and the regulation of the acid-base balance (pH value), the water balance, and the osmotic pressure of the blood. The third task of the kidneys is the production of hormones involved in regulating blood pressure and blood flow. Kidneys and their various physical properties of filtration and clearance are discussed in Chapter 9, including artificial filtering via dialysis.

1.2.5 Respiratory system

The *respiratory system*, schematically shown in Fig. 1.5, consists of the upper part for inhalation through the nose, oral cavity, and throat, and a lower part of air transportation that goes through the trachea and the bronchial tree down into the lungs. In addition, inspiration and expiration requires a combined action of the rib cage and the diaphragm for exerting volume work of the thorax. The main task of breathing is the exchange of gases in the lungs, i.e. the uptake of oxygen by the blood and the release of carbon dioxide back into the air. With the term breathing we understand all organ functions that allow a flow of air into and out of the lungs for gas exchange with the blood. The mechanics of breathing works by generating a pressure difference between atmospheric gas pressure and gas pressure in the thorax (intrapulmonary pressure). By expanding the thorax during inspiration, the intrapulmonary pressure is lower than outside so air flows in. Vice versa, during expiration the intrapulmonary volume shrinks and the pressure increases by the combined action of thorax contraction and diaphragmatic recess. The flow rate and the flow resistance are governed by the laws of aerodynamics, as discussed in further detail in Chapter 10.

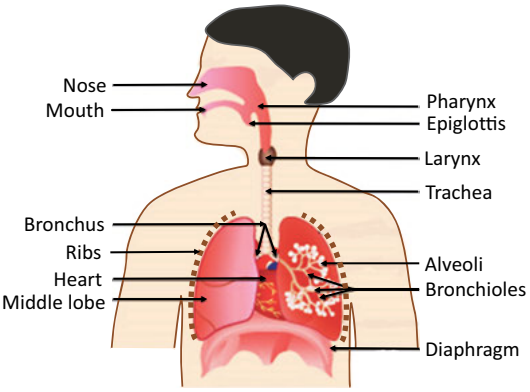


Fig. 1.5: The respiratory system consists of an upper part for inspiration and expiration through mouth and nose, and a lower part in the lungs for exchanging oxygen and carbon dioxide in the blood. Thorax and diaphragm work together to generate overpressure for expiration and under pressure for inspiration.

1.2.6 Digestive system

The *digestive system* is the system that takes up nutrients and subsequently breaks them down into chemical components for combustion with oxygen to generate energy. The digestive system combines most of the organs in the body (Fig. 1.6): oral cavity including teeth and tongue, esophagus, stomach, liver, spleen, gallbladder, pancreas, intestines, colon, and rectum. A particularly important organ is the liver. The liver is responsible for the production of vitally important proteins such as the

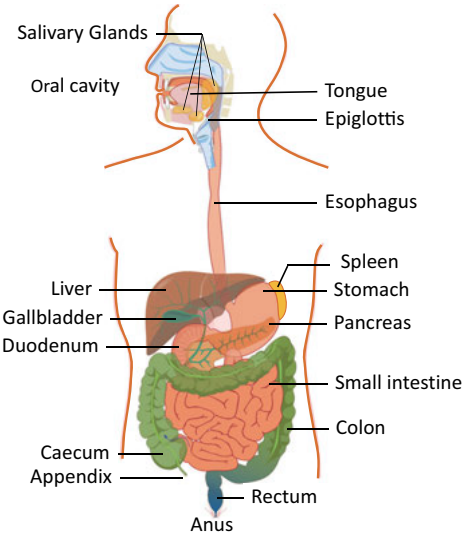


Fig. 1.6: The digestive system is responsible for the energy supply of the body. For this task it encompasses the largest number of organs: mouth, esophagus, stomach, liver, spleen, gallbladder, intestines, pancreas, colon, and rectum (adapted from OpenStax Anatomy and Physiology, 2016, © Creative Commons).

synthesis of albumin and clotting factors, storage of carbohydrates and vitamins, for the synthesis of bile, for the inactivation of medications, detoxification of the blood, and for producing biochemicals required for breaking down nutrition components. While other organs can be compensated for, the liver is unique in its multitude of vital tasks. In case of malfunction, the liver can presently only be replaced from a donor but not substituted for. The energy production of nutrients and the energy household of the body are further discussed in Chapter 4.

1.2.7 Sensory organs

The body is equipped with an array of detectors (*sensory organs*) that constantly provide information on the external environment and the status of the inner organs for processing in the central nervous system. The most important sensory organs are the eye for visual information, the ear for audible perception, combined with the vestibular organ for sensing changes in positon and acceleration. Furthermore we have a nose for smell, a tongue for taste, and in particular a large area of skin for sensing temperature, pressure, humidity, flow of air, water, etc. and any kind of injury. The fingers can determine the surface finish (smoothness and roughness) and the shape of objects. Some sensory organs are shown in Fig. 1.7. All these sensors require the brain for signal processing and decision-making. In addition we have sensors for pressure in bladder and colon, sensors for hunger and thirst, and sensors for sexual appetite.

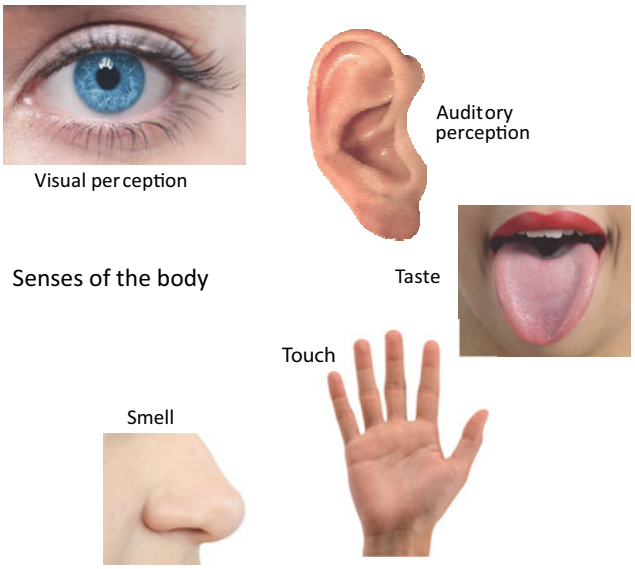


Fig. 1.7: Some senses are shown that provide information on the external environment. The human body contains many more sensors, the largest one is the skin.

However, we do not have sensors for electrical and magnetic fields as some fish and birds do, or sensors for altitude. Altitude is sensed indirectly by the lack of oxygen and depression in the middle ear. We discuss the physical principles of two sensors in more detail: visual perception and auditory perception in Chapters 11 and 12, respectively. The basic principle from stimulus to action potential also works for the other senses.

1.2.8 Nervous system

The *nervous system*, consisting of brain, spinal cord and nerve fibers, has three main functions. It collects information from the environment by receiving signals through specialized receptors. The information is transmitted along nerve fibers to the brain and spinal cord for integration and processing. After processing and decision-making, signals are sent from the brain through the spinal cord to muscles, inner organs and glands for motion. The control of inner organs, such as heart, lung, digestive system, etc. runs semi-automatically. Therefore we distinguish between a *somatic* nervous system for conscious sensation and deliberate movement, and a *vegetative* nervous system for the autonomous regulation of the inner organ functions. Both nervous systems encompass sensory (*afferent*) and motor (*efferent*) connections. Afferent connections transmit signals from the periphery to the center (brain or spinal cord). Efferent connections conduct signals from the center to the periphery. The center is called the central nervous system (CNS), the peripheral nervous system includes all somatic and vegetative nerves. An overview is shown in Fig. 1.8.

For the vegetative nervous system another distinction is necessary: the *sympathetic* and the *parasympathetic* nervous system. Both nerve fibers intertwine and control antagonistically the function of most inner organs such as the heart rate and blood flow. Sympathetic nerves stimulate performance enhancement and fast reactions in

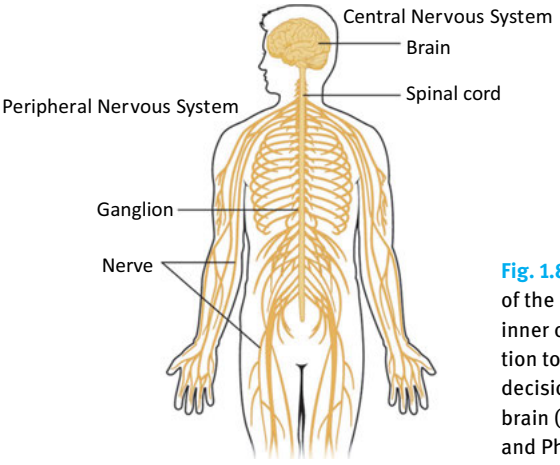


Fig. 1.8: The nervous system reaches all parts of the body. It automatically controls the inner organs as well as the deliberate reaction to external stimuli and the execution of decisions made in the cognitive center of the brain (reproduced from OpenStax Anatomy and Physiology, 2016, © Creative Commons).

dangerous circumstances, while parasympathetic nerves reduce the activity. Both systems are required for an optimal functioning of the organs. In Chapters 5 and 6 the basic principles of resting potentials, action potentials, and signal transmission in neurons are discussed.

1.2.9 Locomotor system

The locomotor system of the body consists of all movable and supporting bones, joints, skeletal muscles, tendons, cartilage, and connective tissues that allow the body to move. In contrast to insects and crustaceans, where the inner organs are protected by a hard shell, the skeletal systems provides less protection but more mobility and flexibility of the whole body. Only the brain and the spinal cord are well protected in the skull and in the vertebral canal, respectively. The skeletal muscles (Fig. 1.9) are triggered by nerve fibers and respond by contraction upon receiving electrical signals (action potentials) from the brain to the muscles. The controlled sequence of action potentials results in a deliberately initiated course of motion. Mechanical, elastomechanical, and kinematic aspects of the locomotor system are presented in the Chapters 2 and 3.

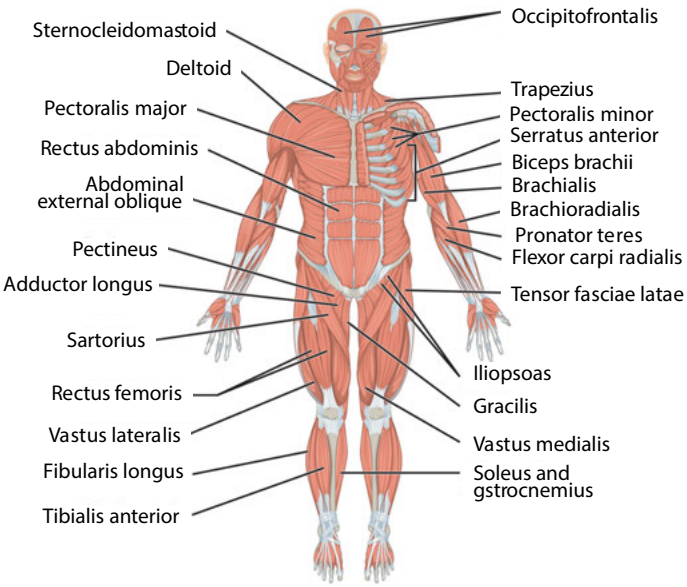


Fig. 1.9: The locomotor system consists of all movable and supporting bones, joints, skeletal muscles, tendons, cartilage, and connective tissues (reproduced from OpenStax Anatomy and Physiology, 2016, © Creative Commons).

1.2.10 Skin

As part of the integumentary system the *skin* is the largest organ of the body, spanning about 2 m^2 for adults; a cross section is shown in Fig. 1.10. It is not only a sensor for temperature, pressure, and humidity, as already mentioned. It is also an essential part of body temperature regulation, through radiation and in extreme cases through sweating for additional heat loss and shivering for heat production. Furthermore, the skin protects the inner organs from mechanical impact, short-term temperature variations, from microorganisms, chemicals, and radiation. In fact, melanin in the skin protects against UV radiation, which has a hazardous ionization effect. Furthermore, the skin is an exchange system, excreting waste products on the one hand and producing various compounds and essential vitamins on the other hand. The skin consists of two main layers, the highly vascular dermis covered by the avascular epidermis. There is a gradient in the skin from nourished deeper layers to unnourished and dying cells on the surface of the skin. The nerve fibers and sensory receptors are located in the dermis. Hair follicles and with arrector pili muscles attached are also embedded in the dermis. Beneath the dermis, the hypodermis contains fat that helps insulate the rest of the body from temperature fluctuations.

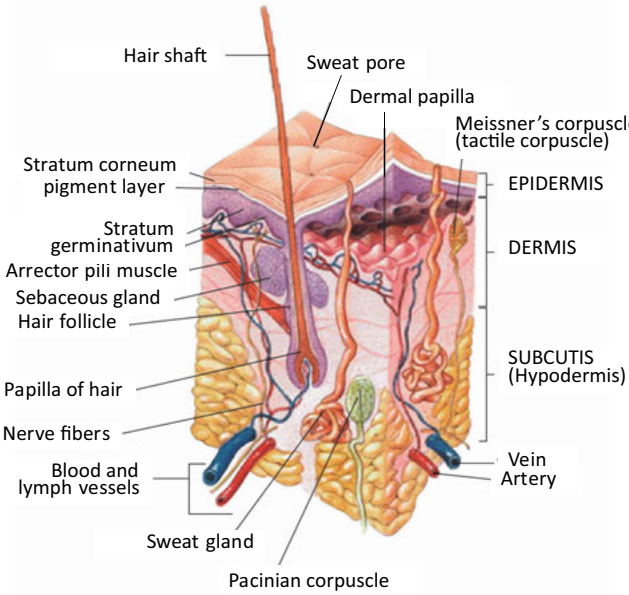


Fig. 1.10: Layers of the skin. Only the dermis and hypodermis layers are supplied with oxygen through the blood vessels. The epidermis is a "dead" layer of cells, eventually stripped off and replaced by dermis cells moving up (reproduced from Wikimedia, Human skin, © Creative Commons).

1.2.11 Reproductive system

Sexual differences are present all over the body but are not essential, aside from the reproductive system. The *reproductive organs* of men and women are fundamentally different and are grouped in inner and outer parts (see Fig. 1.11). The reproductive organs of males comprise the penis, testis, epididymis, spermatic duct, and prostate. The reproductive organs of females comprise two ovaries, two fallopian tubes, uterus, vagina, and female breast (mamma) for the nutrition of newborns. It is interesting to note that prostate and mamma are the organs which most often are affected by cancerous tissues. Cancer and various cancer treatment methods are topics of Chapters 8–12/Vol. 2.

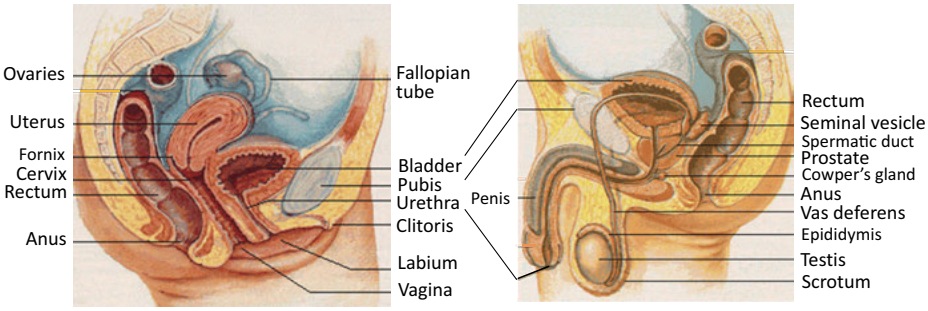


Fig. 1.11: The reproductive organs of males and females, consisting of ovaries, two fallopian tubes, uterus, and vagina on the left-hand side (female) and of penis, testis, epididymis, vas deferens, spermatic duct, and prostate on the right-hand side (male). As a consequence, the urethra tube of females is much shorter than that of males (adapted from Wikimedia Commons, © Creative Commons).

This brief and partial overview of body parts and functions will be deepened in the following Chapters 2–12 for those organs that exhibit clear physical components and aspects, such as the eye, ear, heart, respiratory system and others. The intention is not a physiological description, but a description that makes clear the physical principles that determine the functionality of these organs. For further information on the anatomy of the body and functions of its organs reference is made to a large body of textbooks on these topics. Some are listed and recommended for further reading. Whatever aspect of the body we touch upon, we will immediately recognize its incredible complexity. The complexity starts already at the level of the building blocks, the cells, and continues on any level of the body up to the brain that provides consciousness. The blueprint for body parts and functions is more or less the same for all mammals. Therefore, animal studies are an important part of medical and physiological investigations. This research has provided us with tremendous insights into all physiological aspects of our body and will continue to do so. Medical physics has focused

in the past mainly on instrumentation for diagnostic imaging modalities and therapeutic radiology. New fields are emerging and will become increasingly important in the future, such as studies of interfaces between the nervous system and electronics, biocompatible materials, intelligent prostheses, and more specific local therapies utilizing functionalized nanoparticles. Some of these aspects are discussed in Chapters 14 and 15/Vol. 2.

1.3 Summary

1. The circulatory system provides oxygen through blood flow to the organs for power consumption.
2. The heart is the mechanical pump of the circulatory systems which functions via electrical signals.
3. The kidneys filter the blood, control blood pressure and viscosity, and take care of the electrolyte and water balance.
4. The respiratory system is responsible for the uptake of oxygen into the blood and expiration of carbon dioxide from the blood.
5. The digestive system takes care of the energy supply of the body.
6. The sensory system is the interface to the outside world and connects to the central nervous system for data processing.
7. The nervous system connects the sensors to the center and from the center to organs and muscles.
8. The locomotor system allows us to move around and to do mechanical work.
9. The reproductive system allows us not only to have partnerships but also to maintain the human race.

Further reading

Faller A, Schünke M, Schünke G. The human body. An introduction to structure and function. Stuttgart, New York: Thieme Verlag; 2004.

Boron WF, Boulpaep EL. Medical physiology. 2nd edition. Saunders W.B. Elsevier; 2012.

Martini FH, Nath J, Bartholomew EF. Essentials of anatomy and physiology. 7th edition. Pearson; 2017.

Marieb EN, Hoehn KN. Human anatomy and physiology. 9th edition. Pearson; 2013.

Seeley R, Vanputte C, Russo A. Seeley's anatomy and physiology. McGraw Hill Book Co.; 2016.

Tortoba GJ, Derrickson B. Principles of anatomy and physiology. 14th edition. John Wiley & Sons; 2015.

Guyton AC, Hall JE. Textbook of medical physiology. 11th edition. Elsevier Saunders; 2006.

Pape H-C, Kurtz A, Silbernagel S (eds.). Physiologie. 7th edition. Stuttgart, New York: Thieme Verlag; 2014.

Purves D, Augustine GJ, Fitzpatrick D (eds.). Neuroscience. 2nd edition. Sunderland (MA): Sinauer Associates; 2001. Online textbook can be accessed but not browsed: www.ncbi.nlm.nih.gov/books/NBK11059/

Campbell NA, Reece JB. Biology. 9th edition. Benjamin Cummings; 2009.

Alberts BE, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P. Molecular biology of the cell. 6th edition. Garland Science; 2014.
Watson JD. The double helix: A personal account of the discovery of the structure of DNA. 3rd edition. New York, London, Toronto, Sydney, Singapour: Simon & Schuster; 1968.
Thomas L. The lives of a cell: Notes of biology watcher. Penguin Books; 1980.

Useful websites

https://openstax.org/details/books/anatomy-and-physiology
https://openstax.org/details/books/biology
3D Anatomy of the human body: https://human.biodigital.com/m/anatomy/

2 Body mechanics and muscles

2.1 Introduction

In this chapter on body mechanics we will make the standard distinction between kinematics of a point mass and kinematics of an extended body. From a mechanical point of view the human body is extended, irregular in shape, and inhomogeneous with respect to its mass distribution. Nevertheless, we can ascribe to the human body volume, weight, density, center of mass, momentum of inertia about different axes, torque, and various levers. Resting posture as well as movement of the body require the activity of muscles. How muscles are structured is discussed at the end of this chapter and how they are activated is presented in Chapter 6 after introducing the concept of action potentials in Chapter 5. Energy and power – topics which also belong to the kinematics of bodies – are treated separately in Chapter 4.

2.2 Static mechanical properties

2.2.1 Density

The density of a body is an important physical parameter. The density of the human body has three main contributions: bones, muscles, and fatty tissues. By determining the average density of a body, we can estimate the proportion of fatty tissues with respect to other parts of the body.

For a precise determination of the average *density* $\langle\rho\rangle$ we need to know the mass m and the volume V: $\langle\rho\rangle = m/V$. It is trivial to determine the mass, but it is less trivial to determine the volume of an irregular body. One method would be to submerge a person briefly and completely into a tub brimful of water and measure the water overflow. Another slightly more elegant method follows the procedure of Archimedes, which measures the density directly without determining the volume. First the person's weight $F_1 = m_1 g$ is measured under normal atmospheric conditions, and then again after immersion into water, yielding $F_2 = m_2 g$. g is the gravitational acceleration of the earth. From these two measurements the average density of the body follows according to:

$$\langle\rho_{\text{body}}\rangle = \rho_{\text{water}} \frac{F_1}{F_1 - F_2}.$$

The main components of the body are bones, muscles, and fatty tissue. Their densities, relative proportions, and weights are listed in Tab. 2.1.

DOI 10.1515/9783110372830-004

Tab. 2.1: Average values for densities, relative mass proportion, weight and volume for a person with a mass of 75 kg and a volume of 0.07 m^3. Sexual differences are not taken into account.

	Density [g/cm^3]	Relative mass proportion Percent [%]	Mass [kg]	Volume [m^3]
Bones	1.2	12	9	0.0075
Muscles	1.04	29	21.75	0.0209
Other tissues	0.92	59	44.25	0.048

The average density of the body, composed of bones, muscles and the rest, is:

$$\langle \rho_{\text{body}} \rangle = \frac{m_{\text{bone}} + m_{\text{muscle}} + m_{\text{fat}} + \cdots}{V},$$

and varies between 1.03 g/cm^3 and 1.08 g/cm^3. Thus the body almost floats in water and only little effort is required to keep it aloft.

2.2.2 Center of mass

The *center of mass* (c.m.) of the human body not only is difficult to determine, its location varies by changing posture. We first discuss the upright position at rest and introduce a coordinate system adapted to the symmetry planes of the body.

The bilaterally symmetric human body can be divided into three orthogonal planes shown in Fig. 2.1, corresponding to symmetry planes in the Cartesian coordinate system: *xz* plane or *sagittal plane*, *yz* plane or *coronal plane*, *xy* plane or *transverse plane*. The medial line is defined by the crossing of the coronal plane and the sagittal plane. Furthermore we distinguish between front (ventral or anterior) and back (dorsal or posterior), between up (cranial) and down (caudal), and between left and right, both lateral directions.

Considering the symmetry planes of the body, we expect to find the center of mass along the medial line. We can find the c.m. by first taking the weight of the body F_{body}, then placing the person on a flat board, supported on one side by a pivot point and on the other side by a balance.

Using the lever rule, we obtain:

$$l_{\text{c.m.}} F_{\text{body}} = l_{\text{balance}} F_{\text{balance}},$$

where $l_{\text{c.m.}}$ is the lever arm from the pivot point to the c.m., and l_{balance} is the lever arm from the pivot point to the balance. Reading the weight displayed on the balance and correcting for the weight of the board, we find the center of mass, which – as suspected – lies in the sagittal plane at the position of the pelvic space.

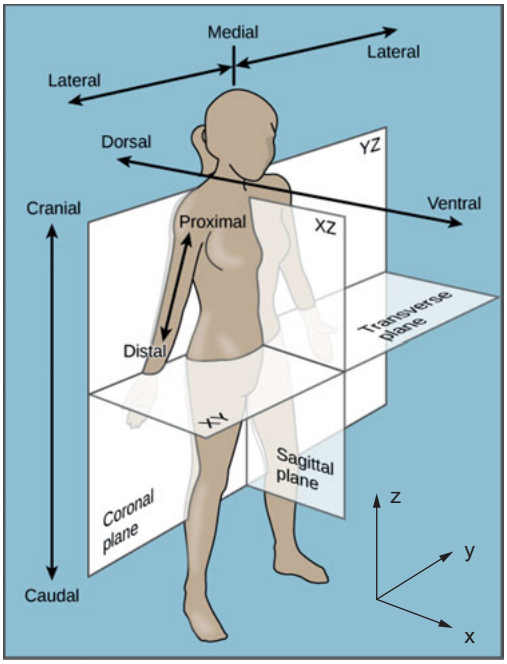


Fig. 2.1: Three main symmetry planes of the human body: Coronal plane, sagittal plane, and transverse plane (reproduced from openstax.org/details/books/anatomy-and-physiology, © Creative Commons).

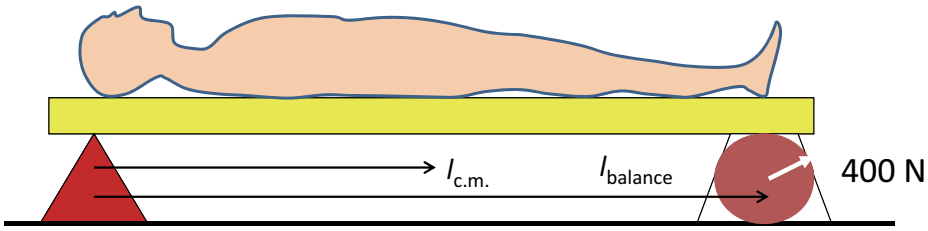


Fig. 2.2: Arrangement for determining the c.m. of a person.

It is useful to distinguish between a c.m. of the upper body half, the cranial c.m., and of the lower body half, the caudal c.m. Those points are indicated in Fig. 2.3 and horizontal lines separate the upper from the lower body. When a person bends over, the c.m. will shift. The new position of the c.m. can be found by connecting the cranial and the caudal c.m. via a straight line. The body's c.m. then lies half way in between, as shown in Fig. 2.3.

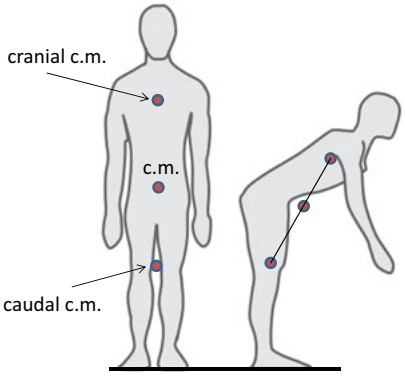


Fig. 2.3: Center of mass of the body in upright position and after bending over (adapted from [1]).

Knowing the location of the c.m. and controlling it is extremely important for sport activities, ballet dancing, ice skating, and various other body activities. For instance in high jumping, the c.m. passes below the bar while the high jumper crosses over the bar, as sketched in Fig. 2.4.

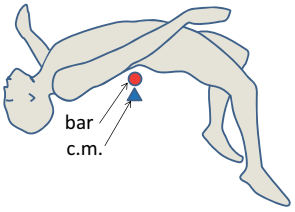


Fig. 2.4: The center of mass of a high jumper crosses below the bar (adapted from [1]).

2.3 Body mechanics

2.3.1 Mechanical models

The mechanics and statics of a body can be modeled by sticks representing bones, counterbalanced by strings representing muscles. A simple case is the upright posture. If the body weight is BW, then the upper part of the body has about 70 % of the total weight, or 0.7 BW. This weight is balanced by two legs, each one supporting 0.35 BW. If we lift one leg, the other one has to support the total weight, i.e. 0.7 BW. However, balancing on one leg requires the contraction of the adductor muscles of the hip, which in fact amplifies the force on the hip, specifically on the femur, not by a factor of 2 but up to 2.5 BW, which is a factor of 7! Leaning over without dipping, as shown in Fig. 2.5, requires a muscle action by the erector spinae muscle connecting the middle of the spine with the lumbar vertebrae to balance the weight of the upper body. With a *stick and string model* this connection is represented by the points D and C, respectively. The weight of the upper body is composed of $W_1 (\approx \frac{1}{2}\text{BW})$ for the chest and

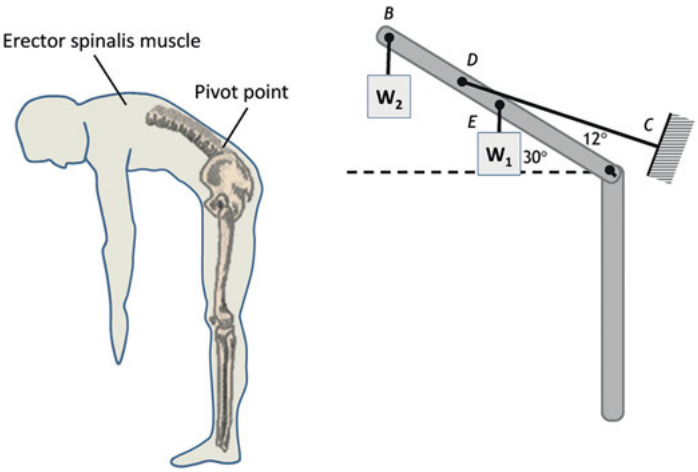


Fig. 2.5: Stick and string model for the action of muscles to counterbalance a particular posture of the body (adapted from [1]).

$W_2 (\approx \frac{1}{4}\mathrm{BW})$ for head and arms. Considering the weight distribution and the angles for spine (30°) and erector spinae muscle (12°), it can be calculated that the force on the lumber vertebra, where the erector spinae muscle is attached, corresponds to a total of 3 BW.

The mechanical equivalent of other postures can be discussed similarly, which gives a hint at the strength of various muscles. It should be remembered, however, that the body does not act in terms of a static balance. The balance is always dynamic, requiring the action of many muscles incorporated into a control and feedback system.

2.3.2 Levers

There are a number of *levers* in the body which help us doing our daily work, such as climbing stairs, lifting weights, or eating apples. Before discussing those in more detail, we first introduce the concept of the *mechanical advantage*.
We distinguish between three classes of levers, which are sketched in Fig. 2.6. They are distinguished by the position of the load $F_1 = mg$ with respect to the pivot point and the position of the counteracting lifting force F_2. Case (a) is termed a *two armed lever* or *1st class lever*. Here the load and the lift are on opposite sides of the pivot point. The lever is in equilibrium, when

$$l_1 F_1 = l_2 F_2,$$

such that the force F_2 required to lift the mass m is:

$$F_2 = \frac{l_1}{l_2} F_1 \ll F_1.$$

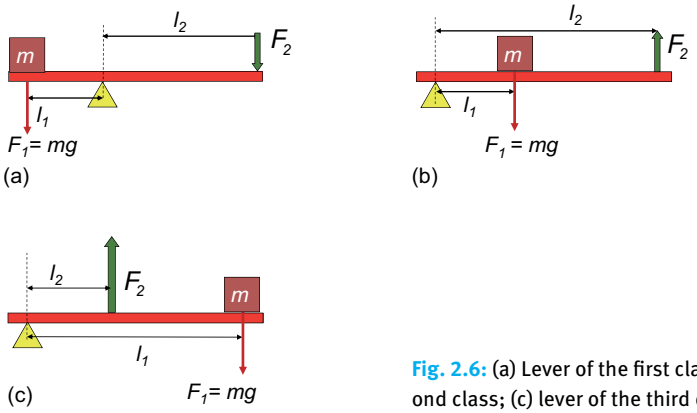


Fig. 2.6: (a) Lever of the first class; (b) lever of the second class; (c) lever of the third class.

Here the load arm l_1 is much shorter than the lift arm l_2. Obviously this is of mechanical advantage.

Second class levers (panel b) have their load and lift on the same side of the pivot point. Therefore they are termed *single armed levers*. Pushcarts are good examples for second class levers. The equilibrium condition and the mechanical advantage is the same as for the two armed lever.

Third class levers (panel c) are also single armed levers. But in contrast to 2nd class levers, load and lift are exchanged so that the load arm l_1 is longer than the lift arm l_2. In this case

$$F_2 = \frac{l_1}{l_2} F_1 \gg F_1,$$

which is obviously not of great help.

The *mechanical advantage* is defined as the ratio: F_1/F_2. For 1st and 2nd class levers the ratio is larger than 1, for 3rd class levers it is smaller than 1.

In the body we can find more often 3rd class levers than of the other two classes. The most prominent example is our arm combined with the forearm, shown schematically in Fig. 2.7. Here the load arm, i.e. the length of the forearm l_1, is much longer than the lift arm l_2, which is only the distance between the anchoring of the biceps at the ulna and the elbow joint. For lifting a weight with the hand the biceps contracts exerting a force F_2 while the triceps is relaxed. The contraction of the biceps results in a torque $T_2 = l_2 \times F_2$ counterbalancing the torque $T_1 = l_1 \times F_1$ by the weight. For pushing down a bar the triceps contracts while the biceps is relaxed. This action has an even lower mechanical advantage. Although the mechanical advantage is minor for 3rd class levers, they provide us with much higher mobility and speed than 1st or 2nd class levers would do. Thus our body is not designed for lifting heavy weights but for high agility. Two more examples are illustrated in Fig. 2.8: the jaw (left) and the foot (right). Only at the foot do we have a 2nd class lever. Indeed, the foot and leg are particularly strong. We can easily lift up our complete body by standing on our toes.

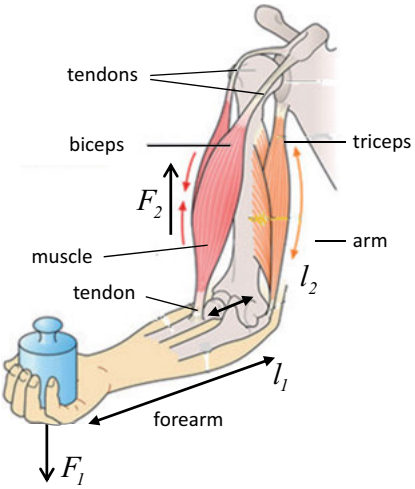


Fig. 2.7: Lever for lifting objects with upper arm and forearm is of the 3rd class displaying no mechanical advantage. The illustration also shows the tendons and muscles of the biceps and triceps (adapted from [2] with permission of Thieme Verlag).

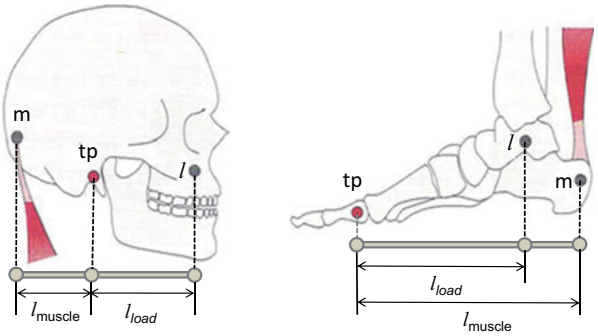


Fig. 2.8: Levers in the jaw and in the foot. Only the foot is a 2nd class lever, providing a very strong lifting force. m = muscle anker point, tp = turning point, l = load point (adapted from [3] with permission of Thieme Verlag).

2.3.3 Femur

In the context of torques and levers the *femur* warrants special attention. The femur is the longest, heaviest, and strongest bone in the body. The upper body weight rests on both femurs when standing still, and additional forces act when walking, running, or jumping. Extreme forces are exerted on the femur due to the strength of the muscles of the hip and the thigh that act on the femur to move the leg.

The pair of forces F_1 and F_2 from the upper body weight, the muscle tension, and the supporting legs, all together exert a torque (Fig. 2.9):

$$T = 2|F|l \sin \alpha.$$

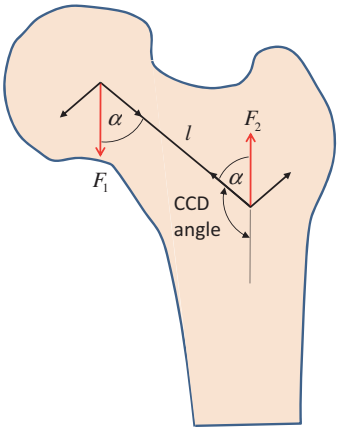


Fig. 2.9: Torque acting on the femur, which is the upper end of the long bone. CCD is the caput-collum-diaphyseal angle. The CCD angle decreases with age.

This torque is estimated to be about 20 Nm for a body standing at rest, but varies between ±50 Nm during walking.

The *Caput-Collum-Diaphyseal angle* (CCD) is the angle between the neck and the shaft of the femur in the hip: CCD = 180° – α. Because of the constant load on the femur, the CCD angle changes with age. At young age the CCD angle is about 140°, decreasing over time to about 115°. A decreasing CCD angle implies an increasing angle α and thus increasing torque acting on the femur. Considering that with higher age the strength of bones decreases, the increasing torque at the femur is a big problem and often results in cracks or even fracture. Fortunately, hip fracture can be repaired by inserting an artificial hip joint. Bone fracture is discussed in Chapter 3 and artificial hip replacement is treated in Chapter 15/Vol. 2.

2.3.4 Degrees of freedom

An extended body has six *degrees of freedom* for motion: three translational and three rotational. In the case of the human body we do not need to discuss the motion of the whole body but the motion of parts of the body that lends the body flexibility, agility, and fine motor skills that are better than a robot. For instance, the head can be bent forwards, backwards, and can be rotated. This makes a total of four degrees of freedom. A special joint, called the condyloid joint, connecting the atlas with the occipital bone (back side of the skull), providing this greater range of motion as compared to the rest of the vertebrae.

Six different types of joints can be identified in the body. Apart from the condyloid joint, the other joints are: ball and socket joint, hinge joint, pivot joint, saddle joint, plane joint. These joints are sketched in Fig. 2.10 together with examples where they can be found in the body. They have the following properties:

1. Pivot joints and
2. hinge joints have one axis of rotation, providing either left-right rotation or back and forward movements, respectively. An example for pivot joints is the radioulnar joint and for hinge joints it is the elbow joint and the knee.
3. Saddle joints comprise two main axes of rotation perpendicular to each other and allow four directions of movement. Examples are the saddle joint of the thumb, i.e., the wrist between the first metacarpal bone and the trapezium.
4. Ball-and-socket joints have three axes of rotation perpendicular to each other and therefore provide six directions of movement. Typical examples are the hip and the shoulder joints.
5. Condyloid joints are distinct from ball-and-socket joints by their elliptical shape. They have two axes of rotation perpendicular to each other, allowing four main movements. Examples are the joints between the forearm and the wrist and the joint between the atlas and the occipital condyles, as already mentioned.
6. Plane joints have no axis of rotation but a glide plane allowing translational motion. Examples are the small joints of the vertebrae.

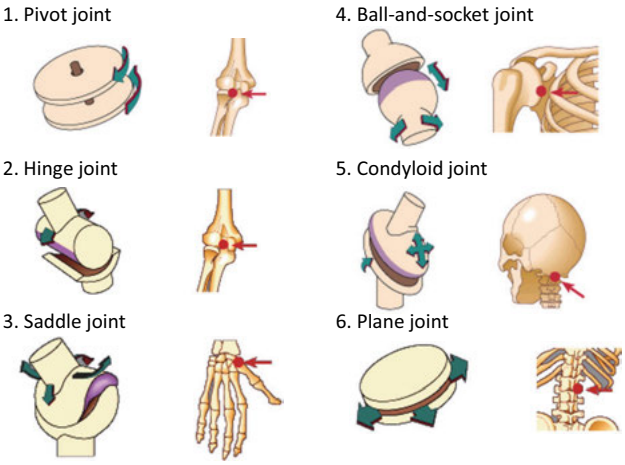


Fig. 2.10: Different types of joints in the skeleton (adapted from [3] with permission of Thieme Verlag).

Joints would not function unless they were tightened by muscles and tendons. Tendons attach to bones and muscles move bones by contraction and relaxation. For instance, the forearm in Fig. 2.7 moves up by contraction of the biceps and simultaneous relaxation of the antagonistic triceps. Lowering the forearm requires the reverse course of actions. Muscles, tendons and joints act together to deliver torque for equilibrating load and to enable various motions of the body. Figure 2.11 exemplarily shows the mobility of the vertebral column with three main movements: (a) lateral bending

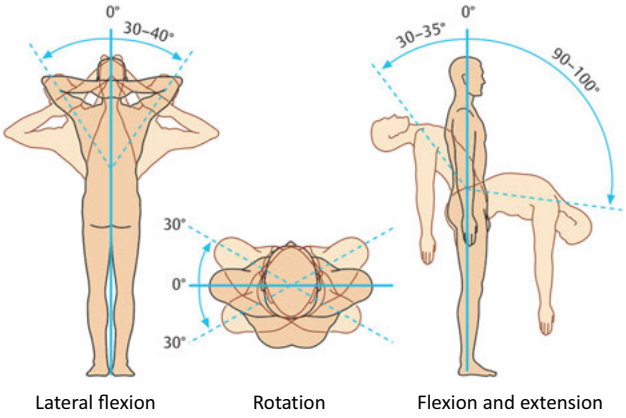


Fig. 2.11: Mobility of the vertebra column. The maximum extension for each movement from zero position is given in degrees (adapted from [2] with permission of Thieme Verlag).

or lateral flexion in the coronal plane; (b) rotation around the vertical axis; (c) forward and backward bending (flexion and extension) in the sagittal plane (see Fig. 2.1 for the definition of the planes).

2.3.5 Biomechanics of walking

Normal walking (gait) is a complex three-dimensional activity, requiring the coordination of a number of joints, bones and muscles [4, 5]. The gait pattern can be subdivided into two major phases and up to 16 partial phases. In Fig. 2.12 the two major phases are shown: stance phase and swing phase. The stance phase starts as soon as the heel of one foot (here the right foot) touches ground (heel strike, HS) and lasts until the same foot is taken off ground again (toe off, TO). In this moment the swing phase starts, supported by the contralateral leg. The stance phase takes about 60 % of the cycle time, the swing phase about 40 %. There is a certain overlap of these two phases between the left and right leg.

The stance phase can be further subdivided into four additional phases: initial heel strike (HS), contralateral toe off (CTO), mid-stance (MS), and contralateral heal off (CHO). During the latter three phases the body is supported by just one leg (single support phase). Correspondingly, during this phase the force and the torque on femur and knee reach maximum values. During the swing phase we can distinguish three additional phases: toe off, mid-swing, and heel strike, which closes the cycle. The force pattern during one gait cycle is shown in Fig. 2.13 for the foot striking the ground (red line) and the force at the joint between femur and tibia (tibiofemoral force) (blue line). The ground reaction force shows two characteristic maxima: the first maximum at CTO is due to upward acceleration of the body towards the mid-stance, the second

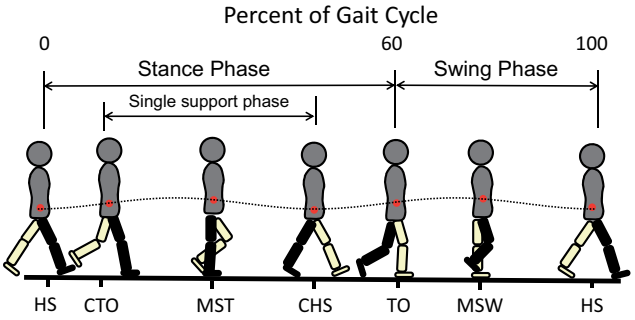


Fig. 2.12: Phases of a normal gait consist of a stance phase and a swing phase. HS = heals strike; CTO = contralateral toe-off; MST = mid-stance; CHS = contralateral heal strike; TO = toe off; MSW = mid swing. The red dots indicate the body c.m. and the black dashed line is a guide for the eye.

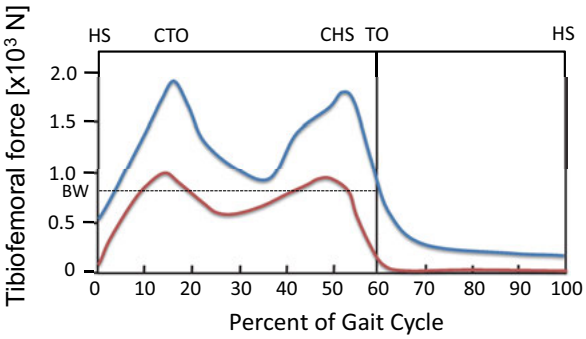


Fig. 2.13: Red curve: vertical ground reaction force of the right leg during normal gait. Blue curve: forces acting on tibiofemoral joint during a gait cycle. BW = body weight, the other acronyms are explained in Fig. 2.12.

maximum is from deceleration as the body drops down from mid-stance to CHO. The maximum ground force is about 120 % of body weight. The red dots indicate the c.m. and the dashed black line is a guide to the eye for the sinusoidal movement of the c.m. during gait with two maxima at mid-stance and mid-swing. The tibiofemoral force mimics the ground force at a higher force level because of the concomitant muscle tension reaching up to 2 × BW.

In the past, tests were conducted using a straight walk path with a centered force plate that can measure three orthogonal components of force and torque on the ground [4]. Evaluating the acting joint forces is then achieved by biomechanical simulation using three-dimensional models of the lower limb [5]. More recently forces and torques in the joints have been measured with strain gauges inserted in artificial hips and knees, telemetrically connected to the outside [6]. An example is shown

in Fig. 15.11/Vol. 2. The force on the hip can be as high as six times the body weight. This force is a combination of total static body weight, accelerated body weight when hitting the ground, and muscle tension. The gait pattern is repeated about 5000 to 8000 times per day for an average person.

2.4 Skeletal muscles

2.4.1 Structure of skeletal muscles

Without muscles there would be no movement – neither of feet on the street nor of intestinal content through the guts. Within the body there are three types of muscles: (a) muscles that are responsible for the functioning of large parts of the inner organs; (b) skeletal muscles, which control the movement of bones in the human body; (c) and cardiac muscle (myocardium), which regulates the contraction of the heart. These muscles are also distinguished by their appearance. Muscles of the inner organs are *smooth muscles*. Skeletal muscles have a characteristic cross-striped appearance and are called *striated muscles*. The cardiac muscle is also cross-striped, but differs nevertheless from skeletal muscles. In the following we will restrict the discussion to skeletal muscles.

There are about 700 different skeletal muscles that make up roughly half of a person's body weight. Skeletal muscles consist of variously shaped muscle bellies (macroscopic shape of a contracted muscle) and tendons on either end, attaching them to bones. The origin of muscles may be single, double, triple, or quadruple headed, but they always end in a single tendon for moving a particular bone. Accordingly, these muscles have additional names such as biceps, triceps, etc. Examples of biceps brachii and triceps brachii are presented in Fig. 2.7.

Movement of skeletal bones is always achieved by contraction of muscle fibers. But unlike a spring, muscles will not relax after release. They require an antagonist pair of muscles for returning to the resting position. The muscles biceps and triceps of the forearm work exactly according to this principle (see Fig. 2.7). Both contract: the muscle biceps brachii moves the forearm up, the muscle triceps brachii moves it down.

Muscles have a hierarchically organized structure of four levels, sketched in Fig. 2.14. On the first level they are composed of several bundles also known as *fasciculus*, which can be distinguished by the naked eye and give the muscle its striped appearance. Each bundle is separated from other bundles by a connective tissue sheath called *perimysium*. All bundles together are wrapped up in an epimysium.

On the second level we find about 150 muscle fibers within a bundle. The fibers are separated by a connective tissue called *endomysium*. Each muscle fiber is an elongated multinucleate cell containing hundreds of *myofibrils*, which make up the third level. The myofibrils are separated by a connective tissue called *sarcoplasm* and wrapped in

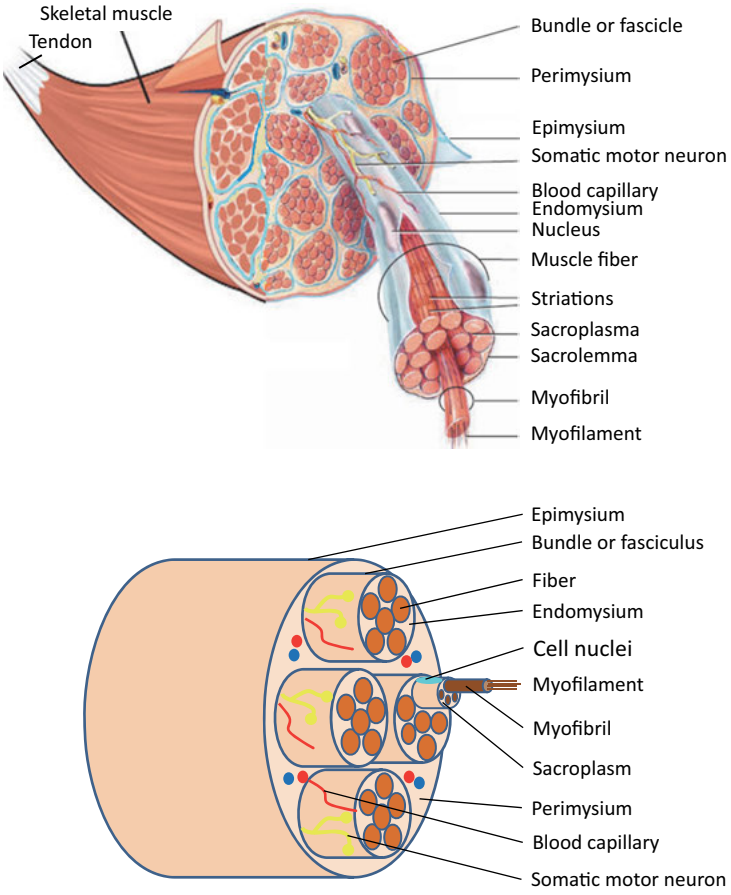


Fig. 2.14: Cross section through a skeletal muscle. Left: anatomical cross section. Right: not to scale schematic view of the hierarchical order of bundles, fibers, myofibrils, and myofilaments.

by a *sarcolemma* that also contains a *sarcoplasmic reticulum*. The sarcoplasmic reticulum is important as a reservoir for Ca^{2+} ions, required for fiber contraction. Finally on the fourth level we recognize within each myofibril hundreds of *myofilaments* organized in *sarcomeres*. Sarcomeres are themselves highly ordered and organized. This unusual fibrous cell structure stretches over several millimeters to centimeters from one tendon end to the other with a diameter of 10–100 micrometers. Blood capillaries and *somatic motor neurons* round up the content of muscles as a discrete organ. Oxygen is required for the metabolism of muscles; and somatic motor neurons connecting to muscle fibers tell the sarcomeres when to contract. In summary, the four hierarchy levels of muscles from macroscale to nanoscale are: bundles → fibers → myofibrils → myofilaments.

Now we take a closer look into a single myofibril, one of them is sketched in Fig. 2.15. Each myofibril contains a highly ordered hexagonal arrangement of myofilaments. The hexagonal order is so perfect that sharp Bragg reflections can be observed by x-ray scattering [7]. Myofibrils are periodically structured by membranes perpendicular to the long axis called *z-disks*, forming a transverse (T)-bamboo like appearance of the tubule. The distance from one z-disk to the next is referred to as a sarcomere. The sarcomeres line up parallel to the long axis of muscle cells and between z-disks. The distance between z-disks is about 2.5 µm at rest, but can stretch and contract considerably. One sarcomere is composed of thin *actin filaments* and thick movable *myosin filaments*. Together they form a *molecular motor*.

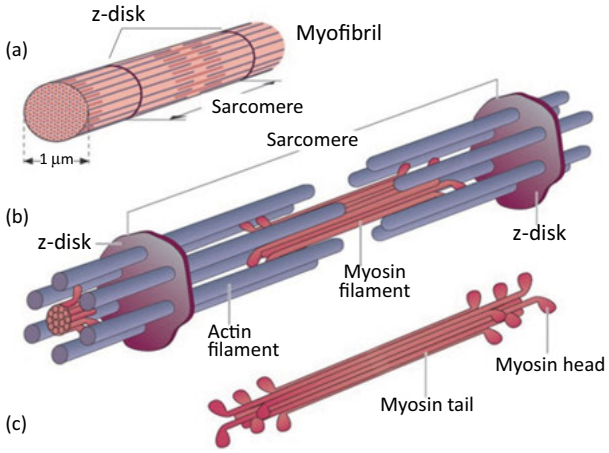


Fig. 2.15: (a) Schematic of one myofibril consisting of many sarcomeres. (b) Each sarcomere consists of fixed actin filaments and movable myosin filaments. (c) The heads of myosin bond to the actin filaments and their movement pulls together the entire sarcomere (reproduced from [2] with permission of Thieme Verlag).

The actin filaments are attached to consecutive z-disks, while the myosin filament is free to move inside the surrounding actin filaments. The myosin filaments have a head and a tail. When activated by motor neurons the *myosin heads* make a swinging and sliding motion along the surrounding actin fibers. The result is a shortening of the total length of a sarcomere while maintaining the length of the filaments. The total contraction is about 50 % of the original length. The contraction is reversed by releasing the bonds between the myosin head and the actin filament and by the action of an antagonistic muscle.

2.4.2 Muscle contraction

Each sarcomere, composed of actin and myosin, is equivalent to one *action unit* of a muscle. The myosin heads are the active molecular motors that develop force upon contraction. One myosin head generates a force of about 1–5 pico-Newton (pN). This has been estimated by optical tweezer experiments [8]. There are about 500 myosin heads in one myosin filament and about 10^6 sarcomeres per bundle. Therefore, there are about 10^9 active actin/myosin cross-bridges per bundle, developing together a maximum force of about 1 mN. The maximum force that can be developed scales with the number of bundles per cross section. For a typical bundle density it is about 10^4 N/m^2. This force is not reached at the beginning or at the end of the contraction, but at a level of about 50 % of sarcomere contraction.

While physics defines mechanical work W by the path integral $W = \int \vec{F} \cdot \mathrm{d}\vec{s}$, where the integration is taken along the entire distance over which a force is applied, the physiological definition of work is $W = F\,\Delta x$, where Δx is the muscle contraction and F is the muscle tension times the bundle cross section. The mere holding of a weight requires muscle tension which consumes energy. The physiological work is only zero if the muscle is contracted without load (force), or if the force is increased without muscle contraction. Now we take a more microscopic view on this definition.

The total force generated by a muscle is the sum of forces generated by many independently cycling *actin-myosin cross-bridges*. The number of cycling cross-bridges depends on the initial length of the muscle fiber and on the pattern of stimulation. When muscles are stimulated by motor neurons to contract, myosin heads move up the actin helix, tending to pull them toward each other at the attachment points on either end. On a microscopic scale this implies pulling together the z-disks and on a macroscopic scale pulling on the tendons. This contracting force per cross section of the muscle is referred to as *tension*.

Tension development during contraction becomes clear when examining the coordinated effort of actin-myosin bonds, sketched in Fig. 2.16. As already mentioned, myosin heads bond to actin molecules during muscle contraction by forming cross-bridges. To increase tension they perform a swinging and sliding motion. If all 500 myosin heads of one filament bond at once, they together can shorten a sarcomere by only about 1 %. For a 50 % contraction a fast stepwise bonding and release motion is required of at least 50 times in rapid progression.

For the release of bonds, energy is required and delivered by adenosine triphosphate (ATP). ATP is synthesized in mitochondria spread all over the muscle fibers. ATP synthesis requires breaking down carbohydrates, which are the main source of fuel, into carbon dioxide and water by reaction with oxygen (see Section 4.2 for more details). The process of supplying energy by consuming carbohydrates using oxygen for breakdown is known as *aerobic metabolism*. For a highly intense and short period of time, energy can also be supplied without oxygen supply, which then is called *anaerobic metabolism*. But this produces lactate in the fibers, which quickly tires the

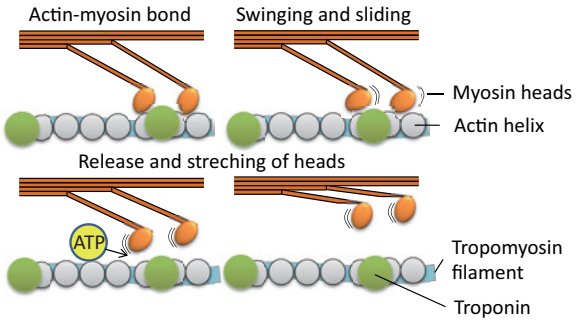


Fig. 2.16: Actin-myosin bridges get bonded and released again in order to obtain a swinging and sliding motion along the actin helix, resulting in sarcomere contraction. Troponin and ATP play roles for binding and unbinding of myosin heads to actin molecules, respectively.

muscles and must be carried away to the liver for transformation back to glucose. Persistent muscle work is of aerobic nature.

Figure 2.17 shows the development of tension during fiber contraction. In the elongated state (c) there is little overlap between myosin heads and actin fibers and therefore the tension that can be developed in this state is rather low. Highest tension is developed in case (b) when there is an optimal overlap between myosin heads and actin fibers forming the highest number of cross-bridges. Further contraction as in case (a) is contraproductive, since actin filaments start to overlap loosing potential cross-bridges. The sarcomere length under optimal tension conditions is about 2.5 µm long. Longer or shorter lengths result in lower tension.

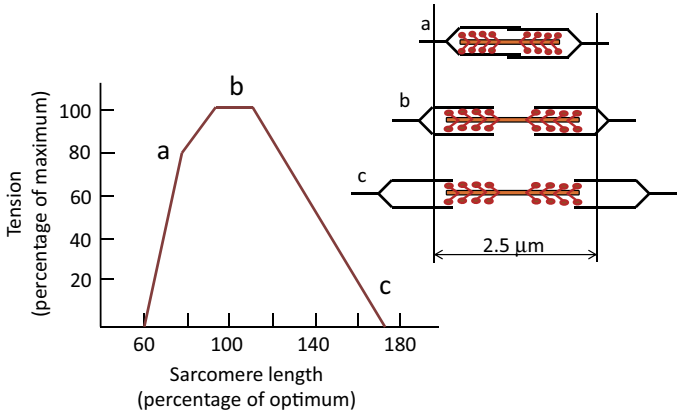


Fig. 2.17: Development of tension as function of sarcomere length.

According to Fig. 2.17 muscles can only develop their maximum strength and tension when they do not shorten or when they shorten by a small fraction. We distinguish between *isometric* and *isotonic* tension development, highlighted in Fig. 2.18. During isometric conditions the muscle is tensed without changing the length from inactive at rest to active under tension. A Newton meter displays the increasing force development without length change. During isotonic muscle contraction the muscle shortens its length without changing its tension, for instance by lifting an object of constant weight.

Isometric contraction shows us how strong a muscle can be, for instance the erector spinae muscle for keeping a position like the one shown in Fig. 2.5. Isometric contraction is therefore important for maintaining posture, for holding joints together while other muscles cause movement, and for holding objects in fixed positions. Although isometric contraction does not result in body movement or lifting, energy is consumed nevertheless.

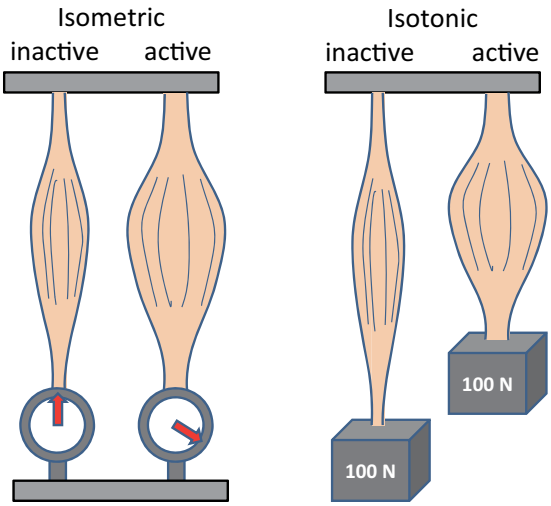


Fig. 2.18: Isometric and isotonic muscle tension. Isometric action implies an increase of tension without change of muscle length; isotonic action refers to a muscle contraction without change of tension.

Isotonic contraction is important for body movement and daily work. Daily experience tells us that we can lift a single apple faster than a sack of apples. Thus weight and speed of contraction are inversely related. This relation becomes clear on a microscopic level. With rapid shortening of the muscle, the myosin-actin filaments glide against each other very quickly. This requires that myosin-actin bonds are continuously and quickly broken in order to create new cross-bridges. Heavy weights can only be lifted when the tension increment is relatively small, which requires more time resulting in lower speed.

2.4.3 Muscle activation

Muscles are activated by somatic motor neurons (neurons are discussed in Chapter 6). Although each muscle fiber has only one neuromuscular junction, the axon branches out and makes connections to as many as 150 muscle fibers. The innervation of one motor neuron with fibers in a muscle is sketched in Figs. 6.1 and 6.13. One junction together with all fibers that it can excite forms one *motor unit*. Motor units are to be distinguished from action units; the latter are identical with one sarcomere. When a motor unit is stimulated, all 150 fibers or so contract together synchronously. However, these motor units are not clustered together; they are dispersed throughout all fibers within a bundle. Excitation of motor units is sketched in the lower panel of Fig. 2.20. The number of motor units per muscle depends on the precision of movement that is needed. For instance, muscles that control eye movement consist of 10–20 muscle fibers per motor unit. In contrast, skeletal muscles responsible for large movements such as those of arm or leg have up to 2000 muscle fibers connected to one motor unit.

Muscle fibers contract according to their stimulus. A *twitch contraction* is a short contraction of all fibers within one motor unit in response to one action potential arriving at a somatic motor neuron. Action potentials can arrive as a single pulse or in short sequences. We first discuss a single action potential. An action potential lasts for about 5 ms (see Chapter 5), the responding twitch may last between 20–200 ms. First there is some delay between the firing of an action potential and the start of contraction, which is known as the *latent period* (green line in Fig. 2.19). During the latent period the action potential spreads over the sarcolemma and Ca^{2+} ions are released from the sacroplasmatic reticulum. Next is the contraction period, which lasts for about 10–100 ms (blue line). During this time Ca^{2+} ions trigger a chemical reaction that allows bonding of myosin heads to actin filaments, which otherwise is hindered

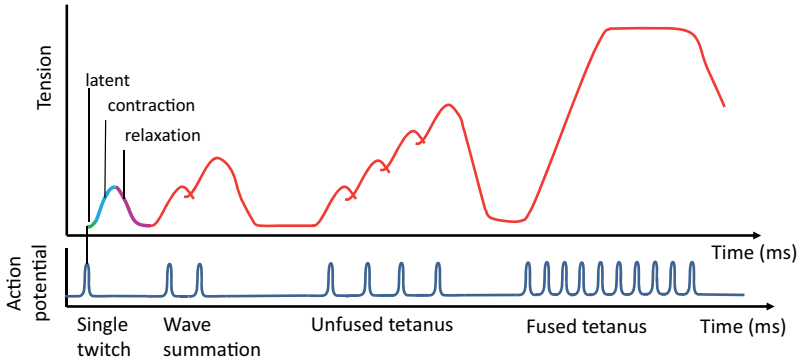


Fig. 2.19: Tension development in succession of a single twitch, a double twitch, multiple but low frequency twitches resulting in unfused tetanus, and high frequency twitches causing a fused tetanus.

by a protein called *troponin*. The third period is the relaxation period lasting for another 10–100 ms (purple line). During this time Ca^{2+} ions are actively pumped back into the sarcoplasmic reticulum, myosin heads detach from actin, and tension in the fibers is released. For breaking myosin-actin bonds ATP is required, as already discussed. Fast muscle movement such as the movement of the eye have contraction and relaxation periods as short as 10 ms, skeletal muscles react more slowly.

If two stimuli occur at very short time intervals, shorter than the *refractory period* of an action potential, the muscle will respond to the first one but not to the second. During a refractory period of about 5 ms the muscle fiber is immune to another stimulus. Refractory periods can be as short as 5 ms for skeletal muscles and as long as 300 ms for the cardiac muscle (see Chapter 7). However, when a second stimulus arrives after the refractory period of the first one and before relaxation has taken place, the second contraction will be added to the first and together the contraction will be stronger than each single one. This phenomenon is known as *wave summation*. When the rate of stimulation is 20–30 Hz, the fiber can only partially relax in between and the resultant wavelike contraction is called unfused or *incomplete tetanus*. When increasing the stimulating frequency to 80–100 Hz, the fiber does not relax at all and the result is a *fused or complete tetanus*. Individual twitches can no longer be distinguished and the tension strength reaches a maximum, which is about 5 to 10 times stronger than the peak tension after a single twitch. Only skeletal muscles have the ability of tetanic contraction for increasing tension. The cardiac muscle, in contrast, is not capable of producing tetanus because of the long refractory time during which the heart completely relaxes before a new contraction can take place (further details are discussed in Chapter 7).

Figure 2.20 summarizes skeletal muscle contraction. As long as the stimulus strength or action potential of somatic motor neurons is below threshold, no contraction will take place. At the threshold level some motor units will be activated and result in fiber contraction. As more and more motor units are added with increasing stimulus strength and frequency, tension increases up to a maximum, beyond which saturation is reached. At the same time the diameter of the muscle belly increases.

Muscle and muscle contraction is only one part of the story. The other part is *coordination*. Any movement requires not only a nerve signal from the CNS to the muscle to contract, but also a reverse signal from the muscle reporting the progress of contraction. For illustration we consider balancing on one foot. This is definitely not an equilibrium position. It requires the other foot and arms to counterbalance. Furthermore, the eyes and the vestibular organ provide important information on the position, which eventually needs correction to avoid dropping. Sensors at the muscles together with sensors of the eyes and vestibular tubes send signals via the nerve fibers to the brain, the brain has to process this information and send back signals to the muscles for maintaining upright position. Now we close the eyes and try to balance blindly on one foot. It is still possible, but much more challenging. If we could switch off the vestibular organ, balancing would not be possible at all. Likewise,

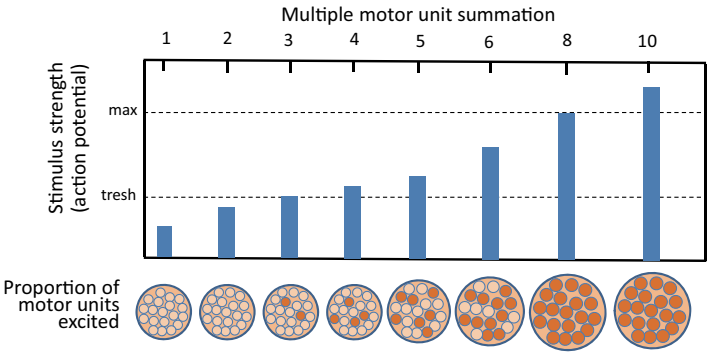


Fig. 2.20: The contraction of muscle fibers depends on the stimulation via somatic motor neurons. If their action potential is higher than the threshold potential, fibers belonging to a motor unit will contract. One circle represents one motor unit. Contracted fibers belonging to one motor unit are colored darker. With increasing stimulus strength and frequency the number of contracted motor units increases up to saturation.

brain injuries, strokes, or other degenerations like Alzheimer and Parkinson dramatically affect our coordination capability. The coordination efficiency can also severely be reduced temporarily by alcohol and drugs.

2.5 Summary

1. Determination of body density is important for gaining information on the proportion of fat.
2. The center of mass of an upright standing person lies in the region of the pelvic space, but can easily be shifted up or down or even to an outside position upon changing the posture of the body.
3. Forces on bones are often much greater than the gravitational force due to the tension of muscles.
4. Most of the levers in the body are of the third type. They do not offer mechanical advantage, but they lend increased mobility and agility to the body.
5. For movement of the body, six different types of joints are available that provide the body with a maximum degree of freedom.
6. Muscles have a hierarchical structure of four levels from bundles, fibers, myofibrils, to myofilaments.
7. Sarcomeres consisting of actin helix and myofilaments form the basic unit of molecular motors, called action unit.
8. Contraction of sarcomeres leads to development of tension.
9. Each myosin-actin pair delivers only one pico-Newton of force. The cooperation of all pairs produce the force required via contraction of sarcomeres.
10. Isometric action means an increase in tension without a change in muscle length.
11. Isotonic action refers to muscle contraction without a change in tension.
12. Contraction of muscle fibers is initiated by somatic motor neurons.

13. A motor unit consists of a neuromuscular junction and all fibers that it can excite.
14. Following an action potential, a twitch contraction occurs in all fibers belonging to one motor unit.
15. Two action potentials at time increments shorter than the refractory time will not result in additional tension.
16. Two action potentials at time increments larger than the refractory time will cause a wave summation.
17. Stimulation at high frequency may cause an unfused or a fused tetanus.

References

[1] Levangie PK, Norkin CC. Joint structure and function. 5th edition. Philadelphia: Davis; 2011.
[2] Faller A, Schuenke M. The human body. An introduction to structure and function. Stuttgart: Georg Thieme Verlag KG; 2004.
[3] Zabel H. Kurzlehrbuch Physik. Stuttgart: Georg Thieme Verlag KG: 2016.
[4] Mann RA, Hagy J. Biomechanics of walking. The American J. of Sports Medicine. 1980; 8:345–350.
[5] Shelburne KB, Torry MR, Pandy MG. Muscle, ligament, and joint-contact forces at the knee during walking. Medicine & Science in Sports & Exercise. 2005; 37:1948–1956.
[6] Damm P, Graichen F, Rohlmann A, Bender A, Bergmann G. Total hip joint prosthesis for in vivo measurement of forces and moments. Med Eng Phys. 2010; 32:95–100.
[7] Geeves MA, Holmes KC. The molecular mechanism of muscle contraction. In: Squire JM, Parry DAD, eds. Advances in protein chemistry. Vol. 71: Fibrous proteins: Muscles and molecular motors. Elsevier Academic Press; 2005.
[8] Mehta AD, Rief M, Spudich JA, Smith DA, Simmons RM. Single-molecule biomechanics with optical methods. Science. 1999; 283:1689.

Further reading

OpenStax CNX Biology, a free web-based textbook on biology.
Guyton AC, Hall JE. Textbook of medical physiology. 11th edition. Elsevier Saunders; 2006.
Tortoba GJ, Derrickson B. Principles of anatomy and physiology. 14th edition. John Wiley & Sons; 2015.

3 Elastomechanics: bones and fractures

3.1 Introduction

The human body consists of a variety of materials with different elastic properties, such as bones, skin, blood vessels, lung, bladder, or the heart. Before discussing their elastic and plastic properties in this and following chapters, we will first review some basic elastic and plastic properties of materials in general.

Any solid body can be deformed by a pair of forces $\vec{F}_1$ and $-\vec{F}_2$ acting on opposite surfaces and in opposite directions, such that the body will not move, since the resulting force cancels: $\vec{F}_1 + \vec{F}_2 = 0$. We consider the magnitude of the force component $F = |\vec{F}_1| = |\vec{F}_2|$ perpendicular to the surface A. The ratio F per surface area A yields a pressure $p = F/A$, also called stress or tension $\sigma = F/A$. If the stress is exerted in only one direction, whereas all other surfaces of the body are free, the stress is called *uniaxial*. Depending on the direction of forces, uniaxial stress can be either tensile or compressive. If all surfaces of a body experience the same pressure (stress), the acting pressure is called *hydrostatic*.

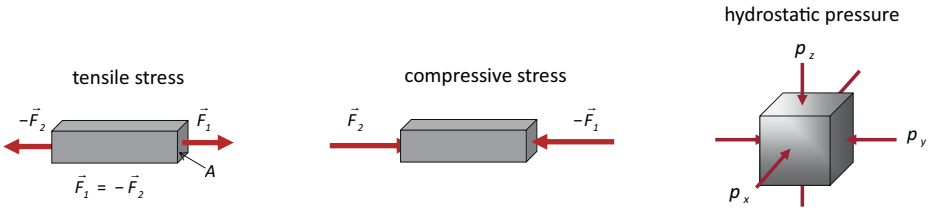


Fig. 3.1: Tensile stress and compressive stress are due to pairs of forces acting on opposite surfaces with area A in one direction; hydrostatic pressure is due to pairs of forces acting on surfaces in all three spatial directions.

Bones are exposed to all kinds of forces from inside, due to body weight and muscle tension, or from outside by lifting objects or getting impacts via striking a hard floor. The ability of bones to lend the body protection, stability, and mobility in conjunction with muscles critically depends on the material properties of the bone matrix. Bones display an amazingly rich hierarchical structure from the nanoscale to the macroscale. This explains their viscoelastic response to impact, their fracture behavior, and their ability to self-repair. We come back to these points in Section 3.5 after introducing further elastomechanical concepts.

DOI 10.1515/9783110372830-005

3.2 Elastic deformation

If an extended solid body is under uniaxial external pressure or *stress*, it will react by changing its shape. The resulting deformation is called *strain*. Strain is measured in terms of a relative change of length: $\varepsilon = \Delta l/l$. *Tensile* stress results in elongation, *compressive stress* in contraction. For small deformations the relation between stress σ and strain ε is linear, which is known as *Hooke's law* (Fig. 3.2):

$$\sigma = E\varepsilon.$$

The proportionality constant E is the *elastic modulus* also known as *Young's modulus*. E characterizes a material as being elastically soft or hard, easy to deform like rubber, or hard to deform like steel. $\sigma = F/A$ is defined as force F per unit area A on which the force acts and therefore has the unit $[\sigma] = \mathrm{N/m^2} = \mathrm{Pascal}$. As ε is dimensionless, the unit of σ and the unit of E is identical. Typical values of E range from 0.5 GPa for rubber to 200 GPa for steel. Some elastic moduli are listed in Tab. 3.1.

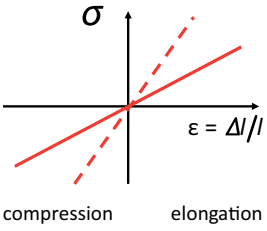


Fig. 3.2: Hooke's law for soft material (solid red line) and hard material (dashed red line).

Tab. 3.1: Elastic moduli of some selected materials.

Material	*E* [GPa]
Alumina	70
Steel	200
Wood	13
Bones	15
Rubber	0.5

In addition to the length change parallel to the applied stress, solid bodies also react by changing their thickness t in the two perpendicular directions. For instance, a tensile stress causes an elongation $\Delta l/l$ in the direction of the stress and a contraction $\Delta t/t$ in the two perpendicular directions. The ratio

$$\mu = \frac{\Delta t/t}{\Delta l/l}$$

is called the *Poisson ratio*. The total volume change upon uniaxial stress is then

$$\frac{\Delta V}{V} = \frac{\sigma}{E}(1 - 2\mu).$$

The factor of 2 is due to the two perpendicular directions to the applied stress. The volume change is zero for $\mu = 0.5$. Typically μ values range from 0.2 to 0.4. As the Poisson contraction does not play any role for the human body, we will neglect this effect in the following chapters.

Other forms of elastic deformation occur by applying tangential forces at opposite sides of a body that change the angles of the body but not its volume. The tangential force is the force component projected into the surface area A. Pairs of forces act such that the body does not gain angular momentum. We differentiate between shear deformation, torsional deformation, and bending deformation. These three types are sketched in Fig. 3.3. In simple terms, the elastic shear deformation is described by a linear equation, which has the same form as Hooke's law for uniaxial stress:

$$\tau = G \cdot \alpha,$$

where $\tau = F/A$ is the shear stress and α is the shear angle. The proportionality constant G is the *shear modulus*. Shear is particularly important for bone fracture. In the case of bending a beam shown in Fig. 3.3 (c) there is a strain gradient from top to bottom from tension to compression, divided by a neutral axis (plane) where the strain is zero. Through bending, the *stiffness* k of a material can be tested. The stiffness is defined as the ratio of force F applied and deflection δ measured: $k = F/\delta$. A high k value indicates a stiff material.

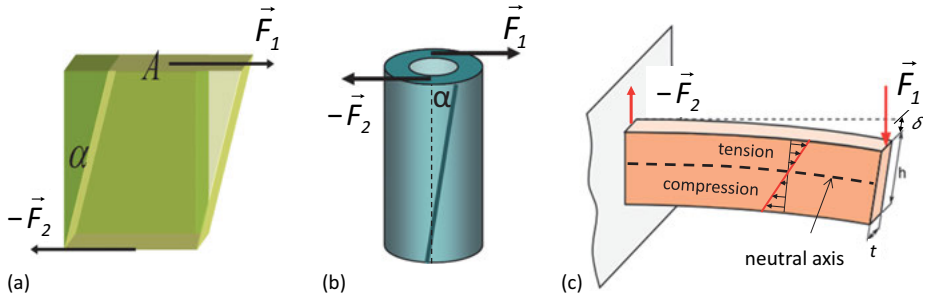


Fig. 3.3: Elastic deformations by tangential forces. (a) Shear; (b) torsion; (c) bending. Note that in all cases pairs of shear forces are applied parallel to a surface. When bending a beam, the counterforce F_2 is exerted by a fixture.

3.3 Plastic deformation

As long as the stress does not exceed a critical value, the strain is completely reversible, i.e., the deformation relaxes after the stress is released. This is the validity range of Hooke's law. Elastic strain changes the distance between atoms by a tiny amount, but does not change the arrangement of atoms. However, beyond a critical

stress value, which is known as *yield stress*, ductile materials start to deform by *plastic flow*, see Fig. 3.4. This deformation is not reversible: after removing the stress a residual strain remains. When cycling stress and strain from positive to negative values beyond the yield point, a hysteresis opens up. The area covered by the hysteresis corresponds to the total elastic energy required for deformation. In the case of metal sheets the possibility of plastic deformation is immensely important for fabrication of all kinds of goods such as pots and pans, automobiles, trains, aircrafts, etc. Plastic deformation goes along with the formation of dislocations in crystallites and grain boundaries between crystallites. Progressive deformation leads to work hardening of metals, i.e., dislocations become entangled and hinder further plastic flow. Work hardening is very important for metals that are to be used as construction materials or tools, since defect free metals are usually too soft. Note that the yield stress σ_Y is lower than the maximum stress σ_{max} after work hardening. With increasing stress beyond hardening, ductile materials will eventually break. Crack formation, crack propagation and eventually fracture is the result of material thinning by elongation or material fatigue by repeated load.

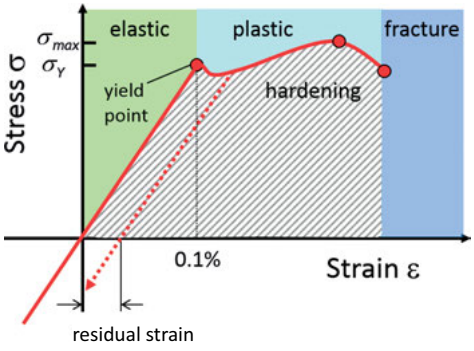


Fig. 3.4: Stress-strain relationship beyond the elastic regime into plastic deformation, ending in fracture. The yield point is the maximum stress without plastic deformation. The dashed line indicates irreversibility of strain after release of stress beyond the yield point. The shaded area is defined as the toughness of materials.

Ductility is the ability of a solid material to deform under tensile or compressive stress. An essential prerequisite of ductile deformation is a charge distribution, which does not change when atomic planes in the material start to flow under the action of stress. This is indeed the case for metals, but not for materials with ionic or covalent bonds. The latter materials turn out to be brittle. Stress exceeding the yield stress result immediately in a break instead of plastic deformation. For many materials the yield stress is reached at a strain of 0.1 %. We experience the difference between ductile and brittle materials by dropping cups of different materials on a hard floor: the metal cup may eventually deform but not break, the porcelain cup will not deform but shatter into pieces. Bones tend to be brittle, but elastic and plastic properties are much more complex because bones are composite and inhomogeneous materials, as we will see later in Sections 3.5 and 3.6.

Toughness of a material is defined as the energy density stored in the material up to the point of fracture:

$$\frac{E}{V} = \int_0^{\varepsilon_{\text{frac}}} \sigma \, \mathrm{d}\varepsilon.$$

The integral corresponds to the shaded area in Fig. 3.4. The larger the area, the tougher the material. Ductile materials are tough, since they can absorb large amounts of strain energy. Brittle materials break already at the yield point and therefore their toughness is low. On the other hand, brittle materials may have a higher *strength*, because they resist a high stress (yield point) before breaking, whereas ductile materials tend to have lower strength, though the strength increases after work hardening.

3.4 Elastic properties of beams

Bones have a hard shell and an open pore structure in the interior, which makes them lightweight without losing elastic rigidity, similar to the design of lightweight constructions of buildings, bridges, and airplanes. The bone material is composed of collagen and minerals. The right mix is important. Too many minerals make them brittle like glass (glassy bones), lack of minerals causes softening and loss of rigidity. There is no unique elastic modulus for bones as we will see later. For the present discussion we make the simplifying assumption that long bones can be described by hollow cylinders like a tube.

For bending a beam, a torque T is required:

$$T = L \times F,$$

where L is the length of the beam and F is the applied tangential force, as shown in Fig. 3.3 (c). The torque can also be expressed in terms of a moment of resistance W:

$$T = W \times \tau$$

where τ is the shear stress. The moment of resistance is defined by the integral $W = 1/L \int L^2 \, \mathrm{d}A$, where L is the distance from the c.m. and A is the cross-sectional area of a beam. Obviously W depends on the geometry of the beam and for a rectangular beam (Fig. 3.5 (a)) one finds:

$$W = \frac{th^2}{6}.$$

Here t is the thickness and h the height of the beam. Obviously edgewise beams have higher bending resistance than sideways beams, which is used in civil engineering for construction of houses, bridges, etc. If we consider instead a cylindrical beam (Fig. 3.5 (b)) of the same cross-sectional area and with radius $R = h$, the moment of resistance is even higher than for the rectangular beam:

$$W = \frac{\pi}{4} R^3.$$

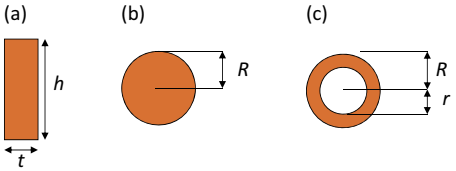


Fig. 3.5: Moment of resistance for rectangular beams (a), full cylindrical beams (b), and hollow cylindrical beams (c).

Finally we consider a hollow cylinder of outer radius R and an inner radius of r (Fig. 3.5 (c)). In the case that the shell is not too thin, we can approximate the moment of resistance as:

$$W = \frac{\pi}{4}\frac{R^4 - r^4}{R} \approx \frac{\pi}{4}R^2\Delta R.$$

This shows that a hollow cylinder has a moment of resistance similar to that of a bulk cylinder with the decisive advantage that the hollow cylinder is much lighter. Long bones can be considered as hollow cylinders and use exactly this elastic advantage. In addition, the pore structure in the trabecular part of bones makes them lightweight without losing stability, and the hollow part of long bones does not compromise elastic rigidity.

3.5 Structure of bones

In the body we have some 270 bones when born. After fusing together some 206 distinct bones are still left at adult age. Those 206 bones are of very different size and shape. We distinguish long and hollow bones for the extremities, short bones for hands and feet, irregularly shaped bones for the spine and knee, flat bones for the skull, shoulder blade (scapula), ribs, and breastbone (sternum). All bones share in common a hard shell (compacta or cortical shell) and a spongy bone interior (trabecular bone). Figure 3.6 shows two examples: a long bone (femur) and a spine bone segment, both featuring the characteristic bone structure: hard cortical shell and a spongy trabecular interior. But the femur has in addition a medullary cavity, which is

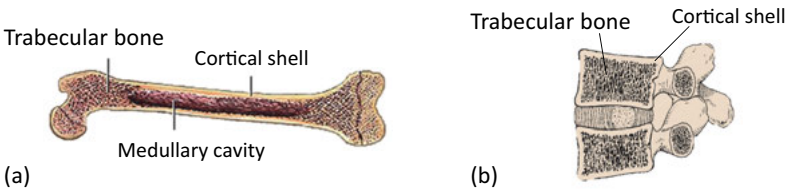


Fig. 3.6: Structure of long bone (a) and skeletal spine bone (b).

characteristic only for long bones. Bones have many tasks. Bones protect vital organs like the heart and the lung, they lend stability and mobility to the body via joints between bones and attachment points for muscles, and they act as production and storage centers. They produce blood cells in the bone marrow and store minerals and lipids, which are released when needed.

Now we take a more microscopic view on the structure of bones and learn that bones are organized in a hierarchical order similar to muscle tissue discussed in Chapter 2. Up to seven levels have been distinguished starting from the molecular level up to the macroscopic scale.

On the molecular level we find collagen and minerals. Collagen is an elongated fiber like protein that serves as connective tissue in most parts of the body, delivering strength and protection. Collagen fibers are shown schematically in Fig. 3.7. The fundamental structural unit is a right-handed triple helix, consisting of three coiled polymer chains intertwined with each other. The chains are held together by hydrogen bonds. The collagen triple helix is 300 nm long and only 1.5 nm wide.

Fig. 3.7: Right-handed triple helix structure of collagen. The individual strands are differently colored (reproduced from www.rcsb.org/pdb/home/home.do).

Thousands of collagen fibers are packed together in the lateral direction and on top of each other to form cylindrical fibrils. Fibrils have a diameter of 25–500 nm, depending on the collagen type and number of fibers. The ends of adjacent collagen fibers are displaced from one another by a distance of 67 nm, which produces a striated appearance visible in electron micrographs. In bones these gaps are filled by nanocrystalline minerals of calcium hydroxylapatite ($Ca_5(PO_4)_3OH$)) (HA), the crystal structure is shown in Fig. 3.8. HA is also known as bone mineral. Bones contain up to 50 % HA by volume and 70 % by weight. Collagen fibers and HA minerals together form organic-inorganic composite biomaterials that constitute the building blocks of fibrils in the cortical bone structure. The collagen fibers provide bending resistance and resilience to bones, whereas the minerals give the bones hardness as well as a high elastic modulus for compressional load. Together the mineralized collagen fibers show viscoelastic properties, which are discussed further below. Nanocrystals of HA are also found in dental

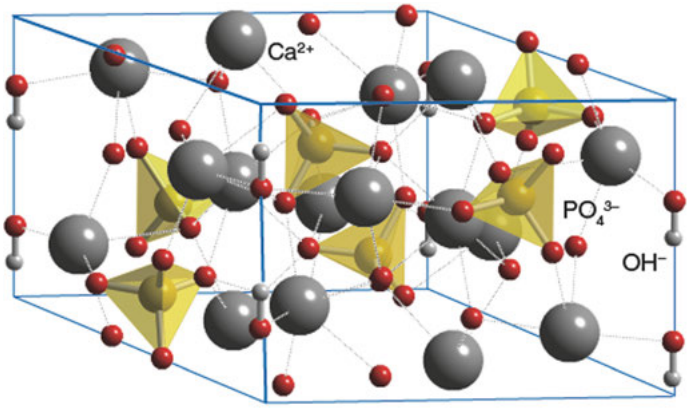


Fig. 3.8: Crystal structure of naturally occurring calcium hydroxylapatite ($Ca_5(PO_4)_3OH$)). In bones nonstoichiometric and calcium deficient forms are found, which form plate-like crystallites (reproduced from www.chemtube3d.com/images/craigimages/CraigMichael/i624fg51.png © Creative Commons).

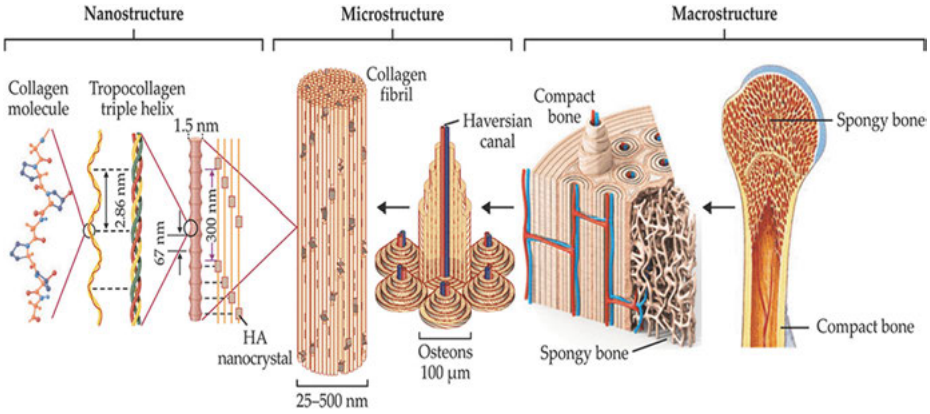


Fig. 3.9: Hierarchical structure of compact bones (reproduced from [1, 2] with permission of Macmillan Publishers Limited).

enamel and they are injected into the skin for correcting fold depressions. HA as a biocompatible material is also important for the design of nanoparticles (Chapter 14/Vol. 2) and implants (Chapter 15/Vol. 2).

Figure 3.9 gives an overview of the hierarchical architecture of bones. Collagen fibers intertwine and form a triple helix structure. The helical structure leaves gaps that are filled with HA nanocrystals. Many fibers combine to collagen fibrils. The collagen fibrils, in turn, are the building blocks for the next level in the hierarchical structure that constitutes cylindrical lamellas forming osteons in the compact part of bones. Osteons have a diameter of about 100 µm. The osteons are cemented together

and form highly regular structures in the cortical part of bones. Each cylindrical osteon has in its center a Haversian canal that contains blood vessels and nerve fibers. The longitudinal Haversian canals are interconnected by transverse Volkmann canals, which link different osteons. The blood vessels carry away and distribute newly generated blood cells from the bone marrow and minerals from the osteons.

In bones one can find four types of cells: osteoblast, osteoclast, osteocytes and osteogenic cells. Osteogenic cells are stem cells that develop osteoblast cells. Osteoblast cells are responsible for the construction of bone material. They segregate calcium phosphate and calcium carbonate, which crystallize in watery environment along the collagen fibrils and form HA nanocrystals. Upon secretion the osteoblast cells become trapped in pores of the matrix called lacunae, thereby transforming themselves into osteocyte cells that can no longer divide. The tissue hardens constituting the typical bone structure. A transverse top view of this structure is shown in Fig. 3.10. The osteocyte cells are interconnected by channels called canaliculi that transport minerals across the bone and sense any damage due to fracture.

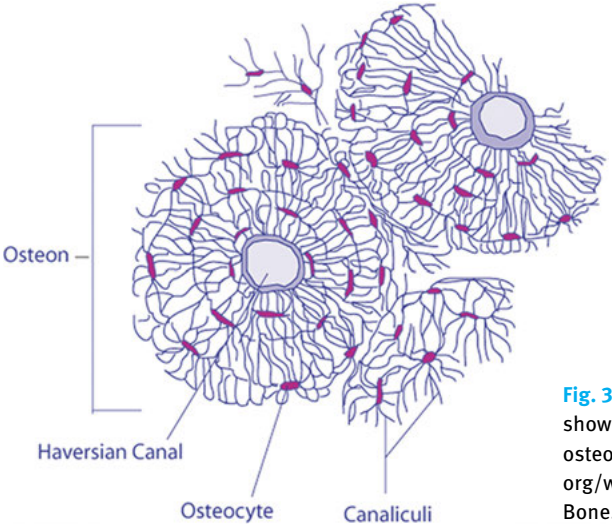


Fig. 3.10: Cross section of a long bone, showing the bone structure including osteocytes (from https://en.wikipedia.org/wiki/File:Transverse_Section_Of_Bone.png © Creative Commons).

The antagonist of osteoblast cells are osteoclast cells. The latter can eliminate old bone structure to be replaced by new ones from osteoblast cells. Osteoclast cells also deliver Ca ions to the body when needed, for instance in muscles. There is a dynamic equilibrium between formation and annihilation of bone cells. This dynamical equilibrium is also responsible for morphological changes of bone structures, such as changes of the CCD angle upon sustained load. With increasing age the osteoclast cells outbalance the osteoblast cells. Then the bone structure becomes weaker and the loss of minerals results in osteoporosis.

The trabecular or spongy part of bones has the appearance of an open pore laticework, a scanning electron microscopy picture is shown in Fig. 3.11. They contain mineralized collagen fibrils for reinforcement, just like in the cortical fibrils, however they do not enclose Haversian canals. In contrast to cortical tissue the fibrils are not parallel to each other and not closely packed. They do, however, contain three of the four bone cells: osteocytes, osteoblasts, and osteoclasts as well as lacunae and canaliculi. The main difference to the cortical part is the open network structure with much lower density.

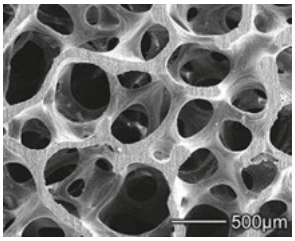


Fig. 3.11: Scanning electron microscopy image of human trabecular bone. The network of trabeculae and pores is in vivo filled with marrow (reproduced from [3] with permission of Elsevier Publisher, Inc.).

3.6 Elastic and plastic properties of bones

From an elastomechanical point of view bones are composite biomaterials, which consist of mineralized collagen fibrils. The mineral reinforced fibrils constitute the elementary building blocks for a large variety of bones structures. They may be parallel aligned and closely packed as in the cortical part of the bone, or randomly arranged forming a trabecular network. Therefore bones are heterogeneous materials with different densities and different elastic properties. The main elements: collagen, fibrils, and osteons are summarized again in Fig. 3.12. Elastic properties and fracture formation of bones have been studied on the macroscopic scale as well as on the microscopic scale. We will review the main results for both scales. Any description of the

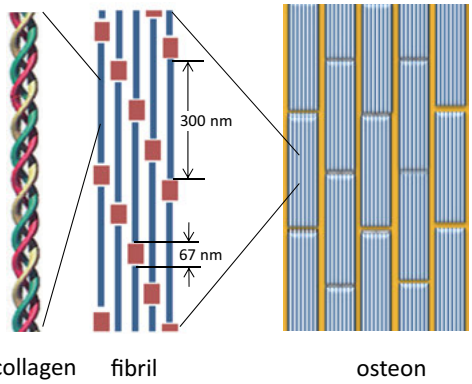


Fig. 3.12: Main building blocks of bones. On the molecular level we find triple helix collagen fibers which line up in a stacked fashion on the fibrillary level, leaving gaps that are filled by mineral nanocrystals. Fibrils, in turn, bundle together in lamellar structures, forming osteons.

bone's elastomechanical properties has to take into account differences in the elastic response of collagen and minerals as well as differences in the three-dimensional packing of fibrils, cortical versus trabecular. A description on a macroscopic level is much more complex as all strains from different components and their spatial distribution have to be considered although they are not visible on the macroscale.

3.6.1 Macroscopic level

For a general characterization of bones, macroscopic strain-stress tests are justified because of their similarity to in vivo behavior. However, they do not provide any insight into the mechanism of strain resistance nor reasons for failure. Bones have an intrinsic anisotropy. When applying a force parallel to the long axis of a long bone, tensile and compressive load can be tested; by applying forces in the perpendicular direction, bending load is probed. From such static measurements the yield stress and the Young's modulus can be derived and the results are listed in Tab. 3.2 from [4]. We notice that on the average the strain resistance is lower for bending load than for compressional load. Therefore we expect that fractures are more likely to occur due to bending than due to compressional or tensile load. Torsional load is particularly likely to cause fracture with the lowest yield stress and elastic modulus. Different types of macroscopic fractures associated with the loads discussed are highlighted in Fig. 3.13.

Tab. 3.2: Yield stress and Young's modulus of bones for compressive, tensile, bending, and torsional load. Range of values from various test measurements are taken from [4].

	Compression	Tension	Bending	Torsion
Yield stress [MPa]	167–213	107–170	103–238	65–71
Young's modulus [GPa]	14.7–34.3	11.4–29.2	9.8–15.7	3.1–3.7

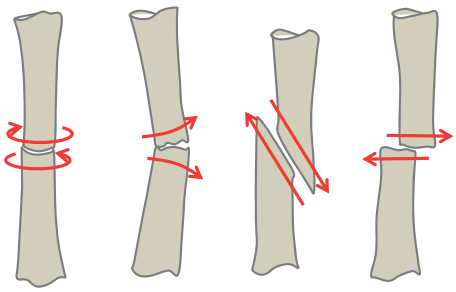


Fig. 3.13: Fracture of long bones due to different types of loads: (a) torsional load; (b) bending load; (c) compressional load; (d) shear load.

The complex structure of bones combining hard minerals and collagen fibrils with high tensile strength results in a viscoelastic response, i.e., the elastic modulus depends on the stress rate: the higher the strain rate, i.e., the faster the stress increases, the higher is the elastic modulus (slope) and the yield stress, as shown in Fig. 3.14. Viscoelastic properties can be modeled by mechanical equivalents: spring for the elastic part and a dissipative element for the damping part. These elements can be arranged in parallel (Kelvin–Voigt model), in series (Maxwell model), or in combinations thereof. In any case, these mechanical models are rather poor representatives of the complex mechanical properties of bones and their usefulness is questionable.

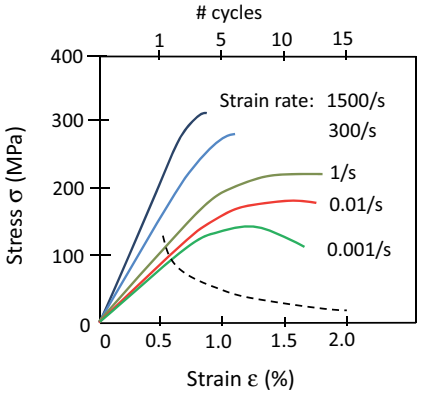


Fig. 3.14: Viscoelastic modulus of bones depends on the rate of applied stress. The lower dashed line shows the yield stress as a function of number of load cycles indicated on the top scale.

The lowest curve in Fig. 3.14 with a rate of 0.001/s can be considered as representative for a static strain-stress curve. It consists of a linear regime up to about 130 MPa and a strain of 0.9 %, followed by a post-yield regime, where the stress levels off, but the strain still increases. This behavior may indicate that a decoupling of two components in the bone takes place, for instance between fibrils and extrafibrillar matrix.

Bones also show *fatigue*, as shown by the dashed line in Fig. 3.14. Fatigue is a general material property, meaning that the strength decreases with the number of load cycles applied. Fatigue depends on many different parameters such as the magnitude of load, type of load, rate applied, temperature, humidity, etc. In contrast to most other materials, fatigue in bones occurs already after a few cycles, not after thousands or millions of cycles as normally observed. This indicates that microfractures occur at an early stage that dramatically reduces the strength of bones. If no time is provided for self-repair, these microfractures accumulate resulting finally in failure [5].

Elastic properties have also been tested for cortical and trabecular bones independently by taking samples from both parts. First there is quite a dramatic difference in structure (compare Figs. 3.6 and 3.11) and in density. The cortical bone has a much higher density (1.9 g/cm^3) than the trabecular bone (0.43 g/cm^3) [4]. This correlates well with differences in porosity P defined as the ratio of void volume to total volume,

which can be redefined as $P = 1 - BV/TV$, where BV is the bone volume and TV is the total volume. For cortical bone the porosity $P = 0.05–0.1$, whereas for trabecular bone it is $0.75–0.95$. Accordingly, the elastic properties are quite distinct. Cortical bones are strong but not tough. In contrast, trabecular bones are less strong, but much tougher. The difference can easily be recognized in Fig. 3.15, which compares the stress-strain relationships of both bone structures. In both cases the *bone mineral density* (BMD) is important. If the BMD is higher than average, bones become brittle. If the BMD is lower than average, bones lose strength.

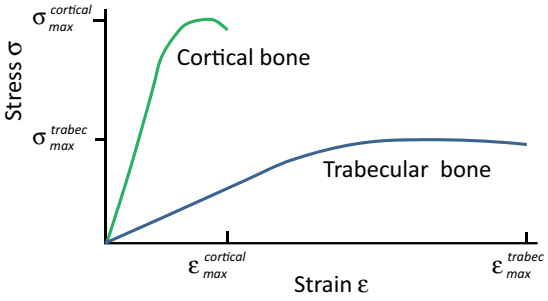


Fig. 3.15: Stress-strain relationship for cortical bone and trabecular bone. Note the different slopes and toughness for the different bone structures (adapted from [4]).

Although the network of trabecular fibers appears to be random, this is actually not the case. A good example is the femur with its highly complex stress distribution. Taking samples from different areas in the trabecular part of the femur for conducting elastomechanical tests, it turns out that the elastic properties are anisotropic. Most of the fibrils align along the lines of strongest load, and the remaining fibrils are used for crosslinking [5].

3.6.2 Microscopic level

For explaining the different biomechanical and elastomechanical characteristics, it has been suggested that toughness is the result of a molecular slip mechanism that allows weak bonds to break and stretch the composite without destroying it. Alternatively, the bone minerals could be responsible for the toughness of bones. The mineral crystallites are considered to be too small for breaking, thus contributing to the overall strength of bones. Only recently has more light been shed on the mechanical behavior of bones on the molecular level by using a variety of methods for imaging, x-ray scattering, and testing stress-strain response on the nanoscale [3, 6, 7].

Strain in response to stress on bone tissue varies largely from the microscale to the nanoscale. This has been revealed by measuring strain via x-ray scattering independently in the filaments and in the apatite nanocrystallites [6]. The measurements yield a surprising result: tissue, fibrils, and mineral particles are exposed to succes-

sively lower levels of strain in a ratio of 12 : 5 : 2, i.e., only about 42 % of the strain at the tissue level is transmitted to the fibrils, and only 12 % of the original strain arrives at mineral particles. This implies that fibrils and minerals are much less strained in comparison to the actually applied strain. The authors explain this apparent discrepancy by the hierarchy of the bone structure, displayed schematically in Fig. 3.16. Much of the strain is taken up by shearing forces of crosslinked fibrils in the interfibrillar matrix. On the next lower level the minerals in the fibrils are again crosslinked, dissipating the tensile strain into shear strain. By this shear transfer mechanism the brittle apatite minerals remain shielded from overload. Nevertheless, the strain at the mineral level is still excessively high by a factor of 2–3 with respect to the yield strain. Here size is indeed important; crack formation is prevented by lack of nucleation points in these nanocrystals. Therefore the mineral particles can withstand stress and strain beyond the yield point.

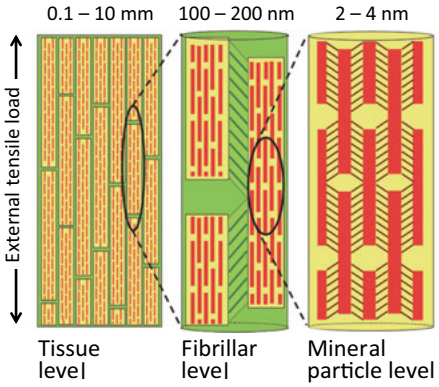


Fig. 3.16: The hierarchical structure of bones imposes a hierarchical deformation transfer from tissue level to mineral particle level. Yellow cylinders denote the mineralized collagen fibrils in longitudinal section of bone tissue such as in osteons. Red tablets denote the mineral apatite crystallites embedded within the collagenous matrix of the fibrils. Green background denotes interfibrillar matrix and slanted lines indicate crosslinks between fibrils as well as between mineral plates. The strain decreases from tissue level to mineral particle level in a ratio of approximately 12 : 5 : 2 (adapted from [6] with permission of National Academy of Sciences, USA, © 2006).

Using scanning electron microscopy (SEM) and atomic force microscopy (AFM), researchers have studied crack formation in bones [7]. They conclude that opposite sides of a scission in mineralized collagen fibrils are held together by some “glue” (see Fig. 3.17). The nature of the glue is not clear at present, but it appears to be formed by an unmineralized organic but nonfibrillar material. Most likely it is the same interfibrillar matrix that is also responsible for transmitting shear stress between the fibrils. Independent of the true molecular nature of the glue, it is evident that it contributes to the toughness of bone before it finally ruptures. The glue in-

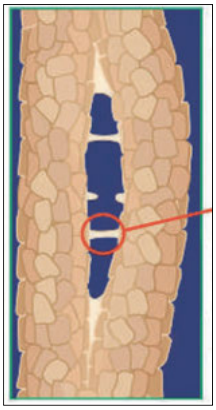


Fig. 3.17: When fibrils are pulled apart, they are bonded by "glue" filaments between the fibrils (modified and reproduced from [7] with permission of Macmillan Publishers Limited).

creases the energy required to stretch and ultimately break the tissue. Conversely, when removing the load, the bonds can reform providing a mechanism of energy dissipation during repeated load cycles. The time scale for bond reformation is not clear yet, but this mechanism could eventually explain the viscoelastic properties and the fatigue of bones on a molecular level.

Before closing this chapter it should be remembered that many other parts of the body also display elastic properties, such as lung, blood vessels, bladder, and eyeballs. Their elastic properties are discussed in the respective chapters.

3.7 Summary

1. We distinguish between elastic deformation and plastic deformation.
2. Hooke's law describes the linear elastic response between stress and strain.
3. Materials can be characterized as ductile or brittle. Ductile materials can be plastically deformed, brittle materials are not deformable.
4. Brittle materials have a high yield stress but low toughness. Ductile materials have lower yield stress, but higher toughness.
5. Long bones are hollow cylinders in their middle part; they display a moment of resistance similar to solid cylinders while being much lighter.
6. Bones are complex composite biomaterials.
7. On a macroscopic level we distinguish between cortical bone and trabecular bone. Cortical bone forms the hard shell of bones, trabecular bone is found inside.
8. Cortical bone has high density and low porosity. Trabecular bone has low density and high porosity.
9. Bone tissue orders on several hierarchical levels.
10. The most important constituents of bones on the nanoscale are minerals (apatite) and collagen, which stack together to form fibrils.
11. Fibrils order in lamellar structures, which form osteons.
12. Osteons contain channels for blood vessels and nerve fibers.

13. Two antagonistic cell types are present in osteons, osteoblast cells and osteoclast
14. Osteoblast cells generate bone material, osteoclast cells eliminate old bone structures.
15. Creation by osteoblast cells and annihilation by osteoclast cells are in a dynamical equilibrium.
16. Bones have (visco-)elastic properties up to a yield point, but no plastic deformation upon sudden impact.
17. Large compressional strain on bones is transformed into shear strain by crosslinks in the interfibrillar matrix; glue filaments between fibrils help to absorb strain energy and to suppress fracture.

References

[1] Grandfield K. Bone, implants, and their interfaces. Physics Today. 2015; April:40.
[2] Wegst UGK, Bai H, Saiz E, Tomsia AP, Ritchie RO. Bioinspired natural materials. Nat Mater. 2014;14:23.
[3] Fantner GE, Rabinovych O, Schitter G, Thurner P, Kindt JH, Finch MM, Weaver JC, Golde LS, Morse DE, Lipman EA, Rangelow IW, Hansma PK. Hierarchical interconnections in the nano-composite material bone: Fibrillar cross-links resist fracture on several length scales. Composites Science and Technology. 2006; 66:1205–1211.
[4] Caeiro JR, González P, Guede D. Biomechanics and bone (& II): Trials in different hierarchical levels of bone and alternative tools for the determination of bone. Rev Osteoporos Metab Miner. 2013; 5:99–108.
[5] Endo K, Yamada S, Todoh M, Takahata M, Iwasaki N, Tadano S. Structural strength of cancellous specimens from bovine femur under cyclic compression. Peer J. 2016; 1562:1–15.
[6] Gupta HS, Seto J, Wagermaier W, Zaslansky P, Boesecke P, Fratzl P. Cooperative deformation of mineral and collagen in bone at the nanoscale. Proc Nat Acad Science. 2006; 103:17741–17746.
[7] Fantner GE, Hassenkam T, Kindt JH, Weaver JD, Birkedal H, Pechenik L, Cutroni JA, Cidade GA, Stucky GD, Morse DE, Hansma PD. Sacrificial bonds and hidden length dissipate energy as mineralized fibrils separate during bone fracture. Nature Materials. 2005; 4:612.

Further reading

Nordin M, Frankel VH. Basic biomechanics of the musculoskeletal system. 4th edition. Wolters Kluwer & Lippincott Williams & Wilkins; 2012
Tortoba GJ, Derrickson B. Principles of anatomy and physiology. 14th edition. John Wiley & Sons; 2015.

Useful website

Bone biology: http://hubpages.com/education/Osteoblasts-Osteoclasts-Calcium-and-Bone-Remodeling

4 Energy household of the body

4.1 Thermodynamics

The human body can be considered as a thermodynamic machine that obeys the laws of thermodynamics. The first law of thermodynamics states that the stored internal energy ΔU of a body can be changed either by exchanging heat ΔQ with a thermal reservoir or by performing work ΔW:

$$\Delta U = \Delta Q - \Delta W.$$

The internal energy of a body increases by heat flow from the reservoir to the body ($+\Delta Q$ and decreases by heat flow from the body to the environment ($-\Delta Q$). In the case of the human body, $+\Delta Q$ is mainly provided by chemical energy, i.e., through metabolizing food, like in a combustion engine, while there are several ways for heat loss ($-\Delta Q$), to be discussed in Section 4.6.

Similarly, the internal energy changes by adding or subtracting mechanical work. Lifting up a person by an elevator increases the person's internal energy ($-\Delta W$), whereas work performed by the person such as chopping wood consumes internal energy ($+\Delta W$).

The *efficiency* for mechanical work is defined by:

$$\varepsilon = \frac{\Delta W}{\Delta Q},$$

i.e., the work performed per unit of thermal (chemical) energy taken up. As in any thermodynamic machine, this ratio is smaller than one.

In equilibrium, $\Delta Q = \Delta W$ and $\Delta U = 0$. However the equilibrium is always a dynamic one, governed by temporal changes of ΔQ and ΔW. Thus it is reasonable to consider the rate of changes:

$$\frac{\mathrm{d}U}{\mathrm{d}t} = \frac{\mathrm{d}Q}{\mathrm{d}t} - \frac{\mathrm{d}W}{\mathrm{d}t}.$$

Here $\mathrm{d}U/\mathrm{d}t$ is known as the *metabolic rate*, $\mathrm{d}Q/\mathrm{d}t$ is the *in*- and *out*-flow of heat, and $\mathrm{d}W/\mathrm{d}t$ is the mechanical power.

4.2 Caloric oxygen equivalent (COE)

The body gains energy by burning food. Metabolism is the combustion of edible organic compounds with oxygen into its constituents carbon dioxide CO_2 and water H_2O. The metabolic energy gain can therefore be determined by the equivalent oxygen consumption. A pretty good approximation shows that 1 liter of oxygen is required for the production of 20 kJ of energy:

$$1\,\mathrm{l}[O_2] \approx 20\,\mathrm{kJ}.$$

DOI 10.1515/9783110372830-006

This relation is known as the *caloric oxygen equivalent* (COE).

The most important food items are:

1. carbohydrates (sugar, glucose): $C_m(H_2O)_n$
2. Fat (i.e., triglyceride): $C_{55}H_{98}O_6$
3. Proteins (i.e., alanine): $CH_3–HCNH_2–COOH$
4. Alcohol (e.g., ethanol): $C_2H_5(OH)$

Let's consider as an example the COE of 1 mol of glucose $C_6H_{12}O_6$. The reaction path with oxygen is as follows:

$$C_6H_{12}O_6 + 6O_2 = 6CO_2 + 6H_2O + 2.9\,\text{MJ}.$$

First we determine the mol mass of the reaction:

$$180\,\text{g} + 192\,\text{g} = 264\,\text{g} + 108\,\text{g},$$

from which we calculate the mol volumina of the reaction:

$$22.4\,\text{l of } C_6H_{12}O_6,\ 134.4\,\text{l of } O_2,\ 134.4\,\text{l of } CO_2,\ \text{and } 134.4\,\text{l of } H_2O.$$

Using the oxygen volume, we find:

$$\frac{2.9\,\text{MJ}}{134.4\,\text{l}[O_2]} = 21.5\frac{\text{kJ}}{\text{l}[O_2]}.$$

When calculating the COE for the other food items, we obtain the values listed in Tab. 4.1. Note that alcohol and fat have the highest energy density but still roughly the same COE.

Tab. 4.1: Energy density and caloric oxygen equivalent (COE) for the main food items.

Food item	Energy density [kJ/g]	COE [kJ/liter O_2]
Carbohydrates	16.1	21.5
Proteins	17.6	18.7
Ethanol	29.8	20.4
Fat	27	19.7

4.3 Metabolic rate

Having calculated the energy content of food, we consider next the energy requirement of the body and the rate of energy consumption. The energy requirement of the body scales with the mass of the body rather than with its surface. The *basal metabolic rate* (BMR) is then defined as:

Energy consumption per kilogram body weight per hour required for maintaining all functions of the inner organs without performing any physical work.

For giving numbers we need first to define a reference frame. The reference frame is a healthy body at rest after actively digesting food (3–4 hours after eating) and in a neutral environment that does not require additional body energy for maintaining body temperature. The use of energy in this state is sufficient only for the basic functioning of vital organs, including heart, lungs, nervous system, kidneys, liver, intestine, sex organs, muscles, brain, and skin. An alternative and related measurement that has less restrictive conditions is the *resting metabolic rate* (RMR).

BMR and RMR can be determined by gas analysis of the respiratory system as the COE is proportional to the energy produced by combusting carbohydrates, fats, and proteins. The following empirical formula, known as the *Mifflin St Jeor equation* [1], provides the BMR for males and females as a function of mass m in kg, height h in cm, and age a in years:

$$P = (10m + 6.25h - 5a + s)\, 4.184 \frac{\text{kJ}}{\text{day}}.$$

Here s is +5 for males and −160 for females.

For an average person the BMR corresponds to about 7–8 MJ/day or 80 Watts. This power is broken down into the consumption by the different organs as listed in Tab. 4.2.

Tab. 4.2: Energy consumption broken down into the consumption by different organs of the body.

Energy requirement of the body	
Liver	27 %
Brain	19 %
Skeletal muscles	18 %
Kidneys	10 %
Heart	7 %
Other organs	19 %

Although the brain has only 2 % of the total body mass, it consumes 19–20 % of the oxygen uptake and therefore 19–20 % of the total energy. This consumption is independent of whether we sleep or whether we are awake. The difference in oxygen consumption is minute, but nevertheless detectable by magnetic resonance imaging (see Chapter 15). The reason for similar oxygen consumption independent of being awake or asleep is the mere shift of the brain activity to different cerebral regions. Also during intense brain activity for solving, for instance, a mathematical problem, the oxygen consumption in this particular “math” region is enhanced, whereas it is lowered in other regions such that the average remains constant.

Oxygen uptake and energy consumption increase rapidly with body physical activity. Some representative examples are listed in Tab. 4.3. We notice that the body's energy consumption is much higher than the actual physical energy produced, for instance by bicycling.

Tab. 4.3: Body energy consumption for some representative activities.

Activity	Oxygen consumption [ml/min kg]	Equivalent power consumption [Watts]
Rest	3.5	80
Slow walking	10	230
Bicycling at 16 km/h	20	460
Jogging	30–40	500–600
Squash	30	700
Bicycle racing 40 km/h	70	1600

The maximum oxygen that a person can take up depends strongly on personal fitness and age. A body with very low fitness level can take up only about 20 ml/min kg, which is sufficient for slow bicycling but not for more demanding physical activities. The age dependence of oxygen uptake for differently trained people is plotted in Fig. 4.1.

Highest performance with highest power can only be delivered for a short period of time. If the muscle power requires more oxygen supply than can be delivered by the circulating blood, then the body is in an anaerobic phase. For longer periods of high power physical exercise, like bicycle tours or jogging, oxygen consumption must balance the oxygen uptake. This is the aerobic phase of exercise.

During anaerobic exercise the extra energy is taken from splitting *creatine phosphate* (CP) into creatine and phosphate and from the conversion of glucose to lactate. This delivers high power quickly, but only for a short period of time. The temporal

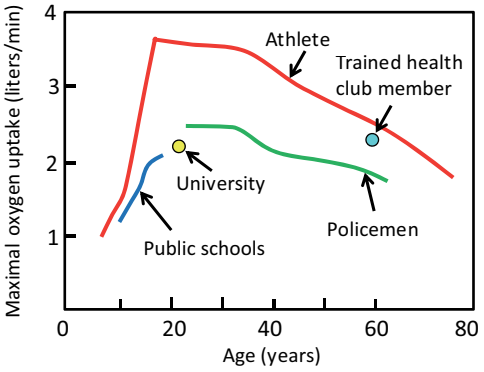


Fig. 4.1: Age dependence of maximal oxygen uptake for differently trained people as a function of age.

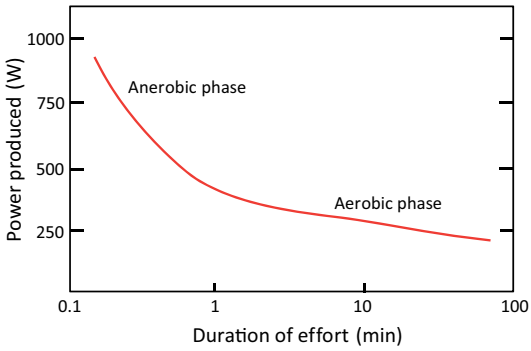


Fig. 4.2: Anaerobic and aerobic power production as a function of duration of effort.

dependence of power production is shown in Fig. 4.2, where it is also seen that the anaerobic phase may be as short as one minute before these extra energy reservoirs are depleted.

Finally, we want to determine the efficiency of human physical work. We take as an example bicycling at 16 km/h, which consumes according to Tab. 4.3 chemical energy of 460 Watts. The mechanical power when measured with an ergometer will be in the order of 100 Watts. Thus the efficiency is in the order of 20–21 %. Bicycling is the most efficient mechanical work that can be performed by the body. A comparison of efficiencies during different tasks is listed in Tab. 4.4.

Tab. 4.4: Efficiency of various human activities.

Task	Efficiency [%]
Bicycling	≈ 20
Swimming at surface	< 2
Swimming under water	≈ 4
Shoveling snow	≈ 3

4.4 Metabolic heat production of the body

We can estimate the total stored thermal energy of the body, knowing that the equilibrium body temperature is 37 °C or 310 K, and assuming that the body consists to 80 % of water or water equivalent substances. For a body mass of 70 kg we then find the thermal energy:

$$\Delta Q_{\text{tot}} = c_{\text{water}} m \, \Delta T = C_{\text{water}} \Delta T = \frac{245\,\text{kJ}}{\text{K}} \, \Delta T,$$

where c_{water} (3.5 kJ/kg K) is the specific heat of water and C_{water} is the heat capacity.

For a body temperature of 310 K this yields a total stored thermal energy of 76 MJ. Metabolic heat production ensures that the stored energy remains constant over time, i.e., any heat loss must be compensated by an equivalent *metabolic heat production* rate (MHR):

$$\left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{tot}} = \left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{MHR}}.$$

The MHR is the sum of BMR and heat produced by physical activity:

$$\left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{MHR}} = \left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{BMR}} + \left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{work}}.$$

We have already confirmed that we need at rest about 7 MJ/d or roughly 300 kJ/h. The BMR is therefore:

$$P_{\mathrm{BMR}} = \left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{BMR}} = 300\,\frac{\mathrm{kJ}}{h},$$

and the metabolic heat production rate can be written as:

$$\left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{MHR}} = P_{\mathrm{BMR}} + \left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{work}} = \delta P_{\mathrm{BMR}}.$$

The *metabolic activity factor* δ can vary between 1 (rest) and 20 (heavy work).

4.5 Heat losses of the body

Having calculated the metabolic rate, we are now prepared to derive the rate of temperature change. The rate is

$$\frac{\Delta T}{\Delta t} = \frac{1}{C}\left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{metab}} = \delta \cdot 300\left[\frac{\mathrm{kJ}}{\mathrm{h}}\right] \cdot \frac{1}{245}\left[\frac{\mathrm{K}}{\mathrm{kJ}}\right] = 1.2 \cdot \delta\left[\frac{\mathrm{K}}{\mathrm{h}}\right].$$

From the last equation we conclude that the temperature increase of the body per hour is 1.2 K at rest and much more during activity. Without heat loss from the body to the environment, the body would overheat. Heat overproduction is essential for any temperature control.

In the body we distinguish four different types of heat loss. Three are conventional, but one of them occurs only in living matter, flora and fauna. The conventional ones are:

1. Heat conduction
2. Radiation
3. Convection or wind chill

The unconventional one is:

4. Evaporation or sweating

In the following we discuss all four types of heat loss separately.

4.5.1 Heat conduction

Heat conduction requires a temperature gradient and a medium that transports the heat from the heat source to the heat sink. The medium can be either gas, liquid, or solid. But vacuum is an insulator. Heat conduction is a dissipative process, similar to electric conduction in metals. In electrical conductors, charges are transported, in thermal conductors phonons (energy) is transported. In both cases Ohm's law applies for the resistance, and Kirchhoff's laws apply for the parallel and serial flow through conductors.

In analogy to Ohm's law for electrical conductance, we write for heat flow or heat rate:

$$\left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{cond}} = \frac{\Delta T}{R_{\mathrm{cond}}}.$$

Here $\mathrm{d}Q/\mathrm{d}t$ is the heat loss per unit time, ΔT is the temperature difference and

$$R_{\mathrm{cond}} = \rho_{\mathrm{cond}} \frac{L}{A}$$

is the thermal resistance for heat flow through a material that has a length L and a cross section A. The material specific parameter is the heat resistivity ρ_{cond}, or the inverse is the thermal conductivity: $\alpha_{\mathrm{cond}} = 1/\rho_{\mathrm{cond}}$. In Tab. 4.5 some representative values for the thermal conductivity are listed. Using the expression for thermal conductivity, we have the expression for heat flow:

$$\left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{cond}} = \alpha_{\mathrm{cond}} \frac{A}{L} \Delta T.$$

This equation confirms our intuition: heat loss increases with increasing thermal conductance of the materials α_Q, with increasing area A, and with increasing temperature difference ΔT.

Kirchhoff's first law states that the total thermal resistance is the sum of all individual ones if they are lined up in series:

$$R_{\mathrm{cond,tot}} = R_{\mathrm{cond},1} + R_{\mathrm{cond},2} + \cdots .$$

Kirchhoff's second law states that the reciprocal thermal resistances add up if the individual resistances are lined up in parallel:

$$\frac{1}{R_{\mathrm{cond,tot}}} = \frac{1}{R_{\mathrm{cond},1}} + \frac{1}{R_{\mathrm{cond},2}} + \cdots .$$

This has consequences for our clothing strategy. In the summer we may wear only one T-shirt. As winter approaches, we dress up with additional shirts and coats. These items of clothing are arranged in series and increase the thermal resistance, i.e., decrease the heat loss. Air between a series of clothing items provides, in addition, good thermal isolation. If, however, due to rain or sweating the clothing becomes partially

wet, then the high heat conductivity of water acts as a short where heat will leak out. This becomes particularly dangerous when dropping into cold water with clothes on. Not only the swimming ability is hindered, but also undercooling is a severe danger due to the high heat conductivity of water. Divers protect themselves by special neoprene overalls that feature low heat conductance. A similar problem occurs when touching a very cold metal pole below freezing temperature with bare hands. The very high heat conductivity of metals drains the heat from the fingers rapidly and at the same time the humidity from the hand freezes at the interface to the metal, resulting in the fingers sticking to the metal pole. This will be hazardous either way: keeping on or pull off. The only strategy is to warm up the hand with a heat blower as quickly as possible. Table 4.5 lists the thermal conductance of some familiar materials.

Tab. 4.5: Thermal conductivity of some typical materials.

Material	Thermal conductivity [Watt K^{-1} m^{-1}]
Air	0.01–0.1
Clothes (cotton)	0.1
Water	0.6
Tissue	0.1–0.2
Metal	200–400

4.5.2 Heat radiation

Heat loss via heat radiation is omnipresent. We may consider our body as a black body radiator that absorbs and emits electromagnetic radiation but does not reflect nor transmit. According to the Stefan–Boltzmann law the heat emitted per second by a black body with total surface area A at a temperature T into the solid angle 4π is given by:

$$\left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\mathrm{rad}} = \varepsilon\sigma A T^4.$$

Here $\sigma(= 5.66 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}})$ is the Stefan–Boltzmann constant, T is the absolute temperature, and ε is the emissivity (a dimensionless number) that characterizes the black body surface with respect to its roughness versus smoothness. For the skin, ε is usually taken as one. The heat flux increases with the body temperature to the power of 4!

Usually one black body of area A at a temperature T_1 is not alone. There is another one nearby with a temperature T_2. Body 1 emits radiation according to its temperature T_1 and absorbs radiation from body 2 at temperature T_2. Vice versa, body 2 emits radiation according to its temperature T_2 and absorbs radiation from body 1. The effective heat flux of body 1 to 2 is then:

$$\left(\frac{dQ}{dt}\right)_{rad} = \varepsilon\sigma A(T_1^4 - T_2^4) \approx \varepsilon\sigma A T^3 \Delta T.$$

The approximation is taken for the case that the temperature difference ΔT between bodies 1 and 2 is only a fraction of their temperatures on the absolute temperature scale. This equation is known as Newton's law of heat loss, which states that the radiation of a black body 1 in an environment of 2 is proportional to the third power of the temperature T times the temperature difference ΔT. It can be shown that a naked person with a body temperature of 37 °C = 310 K in an environment of 20 °C = 293 K would be undercooled in a short time just by radiative heat loss. Thus, clothing is an essential protection, at least in the northern hemisphere. In the case of babies the surface/volume ratio is unfavorable, as a baby has a large surface area for heat radiation but only a small volume for heat storage. The result can be a rapid undercooling if not protected by clothes, by a warm environment, or by incubators for new born babies.

4.5.3 Convection or wind chill

Thermal conduction assumes that the mediating material, such as air or water, is at rest. However, if there is a laminar or turbulent flow of gas or liquid touching a hot surface, than more heat can be carried away per unit time. In an open system air will always flow from hotter to colder areas, which is the usual chimney effect. In hot countries the direction of air flow can be controlled to cool off buildings. Applied to the skin, air flow that carries away body heat is called wind chill. While wind chill is felt comfortable in a summer heat wave, it can be become rather dangerous in a winter storm and may lead to undercooling.

The energy loss rate due to convection depends in first approximation linearly on the temperature difference between uncovered skin and the environment:

$$\left(\frac{dQ}{dt}\right)_{conv} = K_{conv} A(T_{skin} - T_{environ}).$$

Here A is the total uncovered area of the skin, and K_{conv} is a velocity dependent constant with the unit [J/m^2 s K], which can be approximated by:

$$K_{conv} = (10.5 - v + 10\sqrt{v}),$$

where v is the air flow velocity. Heat loss by convection can be substantial.

4.5.4 Sweating and shivering

If all other mechanisms for body heat control are exhausted, the body can still counteract overheating by sweating and undercooling by shivering.

Sweating requires the ability to open pores in the skin, where water or sweat can escape. The heat required for evaporation of the sweat cools off the skin. The heat of evaporation for water is particularly high. For evaporation of one liter of water at the condensation point, an energy of 2.26 MJ is required.

Heat loss through evaporation of water on the skin depends on the relative humidity of the surrounding air. If the partial pressure of water vapor in air is greater than the vapor pressure produced on the skin at a temperature of 37 °C (6.3 kPa), then water (sweat) cannot evaporate (sauna effect). Vice versa, in still and dry air with $T_{\text{environ}} > T_{\text{body}}$, evaporation is the only possibility to lose body heat.

The energy loss rate due to evaporation is given by:

$$\left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\text{evap}} = K_{\text{evap}} A \Delta p,$$

where K_{evap} is a heat coefficient, A is the surface area of the sweating body, and Δp is the partial pressure difference for water vapor in air and at the skin surface.

Shivering is the opposite reaction of the body to fight against undercooling. Shivering implies muscle activity, which requires energy and releases heat.

Summarizing, the combined heat loss of the body is given by:

$$\begin{aligned}\left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\text{tot}} &= \left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\text{cond}} + \left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\text{rad}} + \left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\text{conv}} + \left(\frac{\mathrm{d}Q}{\mathrm{d}t}\right)_{\text{evap}} \\ &= \underbrace{\alpha_Q \frac{A}{L} \Delta T}_{\text{Conduction}} + \underbrace{\varepsilon \sigma A T^3 \Delta T}_{\text{Radiation}} + \underbrace{K_{\text{conv}} A \Delta T}_{\text{Convection}} + \underbrace{K_{\text{evap}} A \Delta p}_{\text{Evaporation}}.\end{aligned}$$

All terms have already been explained.

4.6 Temperature regulation

The body temperature is kept constant by a sophisticated control system. For a healthy person the central body temperature is (37 ± 0.5) °C. This temperature has to be kept constant in the central part of the body where the organs are situated and in the brain. The peripheral parts of the body belonging to the locomotor system may have higher or lower temperatures depending on the environment. Isotherms for an air temperature of 20 °C and 30 °C are shown in Fig. 4.3.

Any temperature control system including the body consists of three essential parts: (1) temperature sensor, (2) heat source, and (3) heat loss. The balance of heat source and heat loss keeps the temperature constant. Temperature sensors register the actual temperature and compare it with the set-point temperature. Deviations between actual and set-point temperature require a negative feedback system: if the actual temperature is higher than the set-point, heat production is lowered, and vice versa.

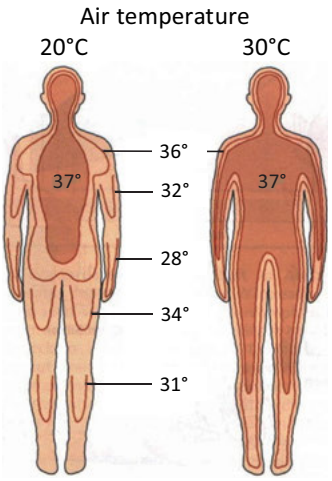


Fig. 4.3: Isotherms of the body temperature in different environments. All temperatures are in centigrade.

In the case of a resting adult human body under normal conditions heat production of about 90 W is required to compensate for heat losses due to heat conduction, radiation, convection, and sweating. Heat production is provided by metabolism. Heat sensors are distributed in the skin all over the body, but with particular high density on the back. Too high temperatures are compensated for by widening of blood vessels in the skin, increasing blood circulation, and opening pores for sweating. Too low temperatures are counteracted by constricting blood vessels in the skin and diverting blood from the skin to the inner parts of the body, and finally by shivering. In all cases the thermostat that compares the set-point temperature with the actual temperature and reaction upon deviations sits within the hypothalamus, which, in turn, is part of the cerebellum. The regulation of the core body temperature is known as *body temperature homeostasis* with negative feedback.

4.7 Summary

1. Caloric oxygen equivalent (COE) measures the energy content of food via oxygen consumption. One liter of oxygen is required for the production of 20 kJ of energy.
2. The basic metabolic rate (BMR) is the energy required to maintain the basic body functions at rest. The BMR depends on weight and age and is about 7–8 MJ per day.
3. The ability to take up the oxygen required for additional activity depends on the fitness of the body.
4. The metabolic heat production rate (MHR) is required for maintaining body temperature. It is composed of the BMR and additional physical activity.
5. Constant body temperature is achieved if the MHR and heat loss are in equilibrium.
6. Heat loss is provided by heat conduction, radiation, convection or wind chill, and evaporation or sweating.

7. For keeping the body temperature constant, an elaborate temperature control system is activated, including temperature sensors everywhere in the skin and in the inner parts of the body. The information is transferred to the hypothalamus, acting as a thermostat.

References

[1] Mifflin MD, St Jeor ST, Hill LA, Scott BJ, Daugherty SA, Koh YO. A new predictive equation for resting energy expenditure in healthy individuals. The American Journal of Clinical Nutrition. 1990; 51:241–247.

Further reading

McArdle WD, Katch FI, Katch VL. Essentials of exercise physiology. 5th edition. Wolters Kluwer; 2015.
Brooks GA, Fahey TD, Baldwin KM. Exercise physiology: Human bioenergetics and its applications. McGraw-Hill Higher Education; 2004.
Herman IP. Physics of the human body. Berlin, Heidelberg: Springer; 2008.

5 Resting potential and action potential

5.1 Introduction

The human body is composed of about 60 trillion (60×10^{12}) cells. Most of them are rather similar in size and shape, but there are exceptions such as nerve and muscle cells, which are quite distinct, elongated and macroscopically visible. Although each cell contains the complete genomic information, they fulfill a diverse range of tasks in organs, nerves, muscles, brain, and skin.

Here is not the place to discuss the complete complexity of a biological cell. Instead we refer to text books in Physiology and Biology, which cover this topic in great detail [1, 2, 3]. We consider an extremely simplified model of the cell that is sufficient to explain the *resting potential* and the *action potential*. The action potential is the basic electrochemical process by which receptors, nerves, and muscles operate. But before an action potential can be fired, cells are at a resting potential, which is the first point of consideration.

The main task of the cell membrane is the distinction between inside and outside. Figure 5.1 is a not to scale sketch of a cell, but indicating the typical extension of a nerve cell's soma. Everything inside is called *cytoplasm* or *cytosol*, everything outside is the *extracellular* space. The cell membrane consists of a double lipid layer with hydrophobic tails inside of the membrane and the hydrophilic heads in contact with the watery solution inside and outside of the cell. The thickness of the membrane, given by the chain length of the double lipid molecules, is about 4 nm. The watery solution is an electrolyte that contains as main components the salts NaCl and KCl. The cytoplasm exhibits an excess of potassium cations (K^+), while the extracellular space has a sodium cation (Na^+) surplus. The pH value of about 7.4 is the same on both sides of the membrane, but the cation concentrations are drastically different. This difference is the main ingredient for the electrical potential of cells at rest, called *resting potential* or *resting membrane potential*.

There is also a difference in the anion concentration inside and outside the cell. Outside, the positive charge of the cations is mainly balanced by the negative charge of the anion Cl^-. However, in the cytoplasm the Cl^- concentration is rather low and the anions are mainly of organic type. The cation and anion concentrations of cells at rest are listed in Tab. 5.1. The anion disproportion does not play a role for the resting potential nor for the action potential of cells as anions cannot diffuse through the cell membrane and take no part in ion transport.

The cell membrane is perforated by pores, called *channels* that serve for ion exchange between cytoplasm and extracellular space (Fig. 5.2). Some of the channels are always open, and some are open or closed in response to a stimulus. The stimulus can be a voltage change, a ligand, or a mechanical strain. Accordingly, ion channels are classified as voltage gated, ligand gated, or mechanical gated. We will consider here

DOI 10.1515/9783110372830-007

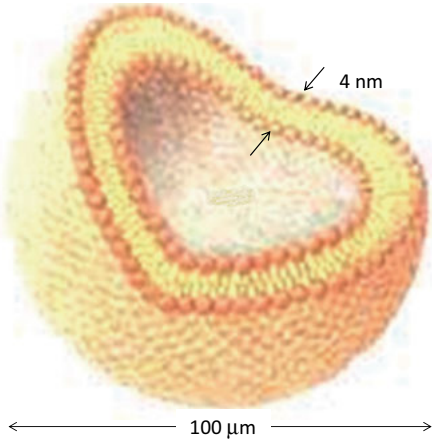


Fig. 5.1: Not to scale sketch of a cell, emphasizing the double lipid membrane, which separates the cytoplasm from the extracellular space.

Tab. 5.1: Ion concentrations inside and outside a cell. The numbers are average values. They may vary depending on the cell type.

Ions	C_{inside} [mmol/l H_2O]	$C_{outside}$ [mmol/l H_2O]
K^+	140	4
Na^+	12	145
Cl^-	15	117
Organic anions	140	

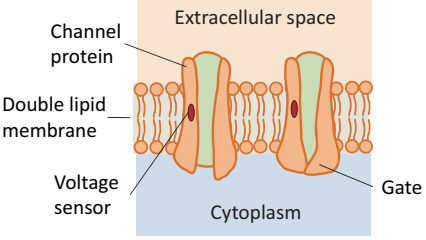


Fig. 5.2: Membrane with open and closed ion channels.

only voltage gated ion channels. Furthermore, ion channels are cation selective, they are specialized for sodium (Na^+), potassium (K^+), calcium (Ca^{2+}), and other cations. There are no channels for anions. The density of K^+ and Na^+ channels is still disputed in the literature [4], partly because their density varies greatly depending on location, such as soma or dendrites.

5.2 Resting potential

Now we focus our attention on Na^+ and K^+ ions. Figure 5.3 shows a model cell including ion channels for diffusion of Na^+ and K^+ ions in and out of the cell. The channels that allow K^+ ion diffusion are always open, the Na^+ channels are closed unless stimulated to open.

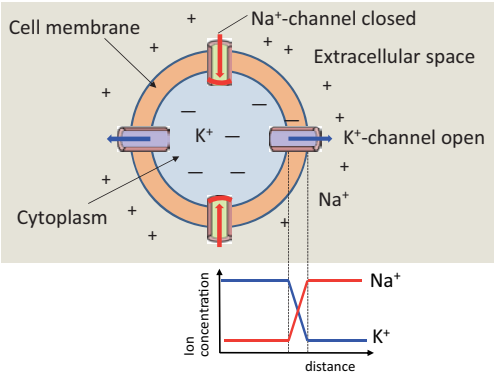


Fig. 5.3: Schematic of a biological cell including specific ion channels for the diffusion of the cations K^+ and Na^+.

Because of the huge concentration difference of K^+ ions across the cell membrane, this difference is apt for a diffusional exchange following *Fick's first law*:

$$I_n = PA\Delta n.$$

Here I_n is the particle current (unit: [mol s^{-1}]), i.e., the number of particles per unit time crossing a barrier (membrane) of total area A, $\Delta n = n_{\text{inside}} - n_{\text{outside}}$ is the difference in particle densities inside and outside, $P = D/\Delta x$ is the *permeability* of the membrane (unit: [m/s]), where D is the diffusion constant (unit: [m^2/s]) and Δx is the membrane wall thickness. The K^+/Na^+ ion density profiles are schematically shown in the lower panel of Fig. 5.3.

A complete K^+ cation exchange, as indicated by the green line in Fig. 5.4, is hindered by Coulomb attraction. Each K^+ cation that diffuses through the ion channel from the inside out leaves a negative unbalanced charge behind. In equilibrium there must be a balance between thermal energy that promotes ion exchange, and electrostatic potential that imposes an increasing barrier against out diffusion. Thus the concentration ratio in equilibrium is given by:

$$\frac{n_{\text{inside}}}{n_{\text{outside}}} = \exp\left(\frac{Q\Delta U}{RT}\right) = \exp\left(\frac{ZF\Delta U}{RT}\right).$$

Here Q is the total charge transported by diffusion, ΔU is the Coulomb potential difference, R is the gas constant for one mole of gas (8.314 J K^{-1} mol^{-1}), and T is the absolute temperature. The charge flow can be expressed in terms of the Faraday constant $F = 96\,487$ C. The Faraday constant is the charge of one mole of singly charged

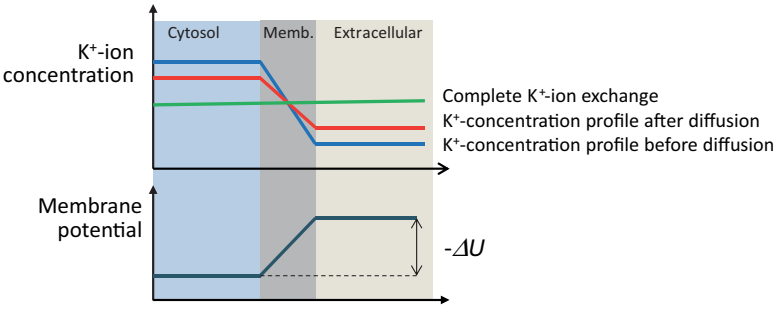


Fig. 5.4: Potassium concentration profile across the cell membrane. The large concentration difference cannot be balanced completely because of Coulomb forces hindering the outflow of K^+ ions. The remaining concentration difference is responsible for the resting potential.

ions, and Z is the ionicity, which is 1 for K^+. This equation is known as the Nernst equation, according to the physical-chemist Walther Nernst (1864–1941), who taught at the Humboldt University in Berlin. The K^+ ion concentration profile before and after diffusion is shown by the blue and red lines in Fig. 5.4, respectively.

According to the concentration profile after K^+ ion exchange, the expected Coulomb potential difference is:

$$\Delta U = \frac{-RT}{ZF} \ln \frac{n_{\text{inside}}}{n_{\text{outside}}}.$$

Putting in numbers, at body temperature of 310 K the prefactor yields a potential difference of 0.0267 J/C = −26.7 mV. Converting the natural logarithm to the decadal logarithm, a factor of 2.3 has to be taken into account. Thus the prefactor becomes −61 mV:

$$\Delta U = -61\,\text{mV} \log \frac{n_{\text{inside}}}{n_{\text{outside}}}.$$

Now we consider the ratio of the cation concentrations according to Tab. 5.1 and obtain a resting potential difference of −90 mV. We note that the resting potential is a consequence of the potassium surplus in the cytoplasm.

In a more advanced consideration one may take into account the finite permeability of the cell membrane for Cl^- anions and a backflow of Na^+ cations. All this leads to the *Goldman–Hodgin–Katz equation* [5], which reads:

$$\Delta U = \frac{-RT}{ZF} \times \ln \frac{P_{K^+}[K^+]_i + P_{Na^+}[Na^+]_i + P_{Cl^-}[Cl^-]_i}{P_{K^+}[K^+]_o + P_{Na^+}[Na^+]_o + P_{Cl^-}[Cl^-]_o}.$$

Here P are the respective permeabilities and the subscripts i, o stand for inside and outside, respectively. This equation yields a resting potential difference of about −75 mV, which is a typical value for experimentally determined resting potentials, using the so called patch-clamp technique [6, 7]. Hence, at rest there is a potential difference between the cytoplasm and the extracellular space maintained by a delicate gradient

of the potassium cation concentration across the cell membrane. The cell may be compared with a battery, where cations and anions are also spatially separated to create a voltage difference. The resting potential is not a fixed equilibrium but a floating one that can also be disturbed by changes to the pH value in the extracellular space. Extensive global changes due to either too high or too low salt ion concentrations can be life threatening. For more details see Chapter 6.

5.3 Action potential

As the name indicates, resting potential is constant in time (Fig. 5.5), unless triggered by an external stimulus that tells the cell to depolarize, often referred to as "firing an action potential". The external stimulus can be anything from a sting by a mosquito to a decision in the brain to lift up our body or to listen to music. How does the activation proceed? This is one of the most fascinating tricks that evolution invented. It is similar to the control of the flotation depth in submarines. Triggering implies opening Na^+ channels. There is always a bit of potential fluctuation and leakage of Na^+ from outside in. But this does not do any harm as long as the fluctuations are smaller than the threshold potential. However, once the stimulus is so strong that it surpassed a threshold level, then all Na^+ channels open at once. The Na^+ ions flush in and depolarize the cytoplasm in a matter of less than a millisecond. For this purpose the Na^+ channels are much faster than the K^+ channels. It has been estimated that 10^6 to 10^8 ions per second can pass a single sodium channel during depolarization.

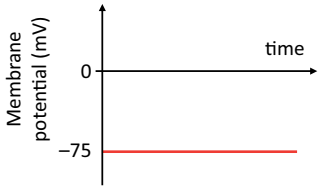


Fig. 5.5: Ideal resting membrane potential neglecting thermal fluctuations.

Each sense has a different "trick" to activate an action potential. This can be a chemical stimulus for tasting and smelling, a mechanical tilt of tiny hair cells for hearing, a photo-isomerization for vision, or a self-activated stimulus for the heartbeat. These different stimuli are discussed later in the respective chapters. The result is always the same. Once a cell is depolarized in response to a stimulus, it has to fulfill two tasks: first communicate the state of depolarization to the neighboring cells, for instance through nerve conductivity; second, repolarize as fast as possible and get back to the resting potential to be ready for the next stimulus. The *refractory period* from depolarization to repolarization takes about 3 ms, whereas the complete cycle from resting potential, activation to repolarization requires about 5 ms. During the refractory period it is not possible to fire another action potential.

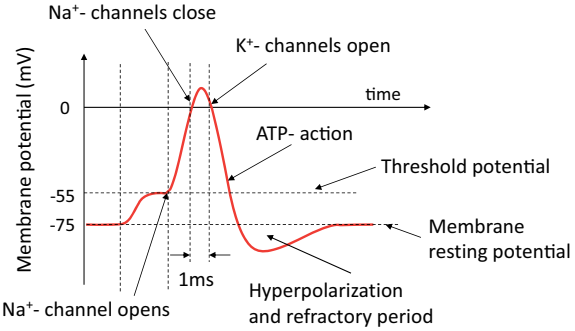


Fig. 5.6: Time development of the action potential.

The sequence of events during one cycle of depolarization is shown in Fig. 5.6. After passing the threshold level, the depolarization of the cell is limited by automatic closing of the Na^+ channels as soon as the cell potential crosses the zero level. Then the K^+ channels, which were temporarily closed to speed up the depolarization, will open again. And finally the most important action is the launching of an ion pump that exchanges $3\,Na^+$ ions from the inside (cytoplasm) outwards against $2\,K^+$ ions from the extracellular space to the inside. This is an active ion transport against the ion concentration gradients that requires energy. The energy is delivered by a molecule called *adenosine triphosphate* or short ATP. The action of ATP pumps is discussed in the next section. As we notice, depolarization is faster than repolarization of the cell. This is due to the fact that depolarization merely requires opening Na^+ channels and diffusion does the rest, whereas for repolarization an active process against the concentration gradient is initiated via the ATP pump, demanding more time.

The response of a voltage gated receptor to stimuli is illustrated in Fig. 5.7. If the stimulus is on a subthreshold level, the membrane potential changes for a short period

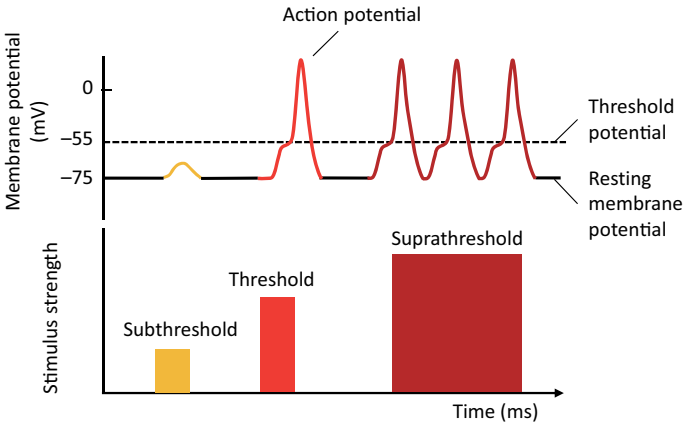


Fig. 5.7: Response of a voltage gated receptor cell for stimuli below and above threshold.

of time, but no action potential is triggered. If the stimulus reaches threshold values and is short in time, a single action potential occurs. In the case that the stimulus is strong and lasts for a longer period of time, a sequence of action potentials will be activated, the frequency is proportional to the amplitude of the stimulus and the total number represents the duration of the stimulus. The frequency is limited to about 200 Hz due to the refractory period, during which no new action potential can be initiated. This analog to digital conversion of stimuli is the fundamental working principle for all sensory receptor cells and will be discussed further in Chapter 6.

5.4 Channel conductivity

Using very simplified model assumptions we can nevertheless estimate a number of physical parameters of the cell's electric properties. We assume that the cell has a spherical shape and that the cell membrane can be considered as a capacitor with a capacitance $C = \varepsilon_0 A/d$, where A is the surface of the sphere, d is the thickness of the membrane, and ε_0 is the dielectric constant of vacuum. Taking the time derivative of the capacitor equation $Q = CU$, we find for the Na^+ charge current during depolarization:

$$I_{Na^+} = C\frac{\Delta U}{\Delta t}.$$

Note that the charge current has to be distinguished from the diffusional current I_n. They may be identical if each ion carries one elementary charge. The capacitance of a cell can easily be determined to be about 100 pF for a typical cell of 100 µm diameter, membrane thickness of 4 nm, and surface area of 3×10^4 µm^2. Furthermore, the depolarization from the resting potential −75 mV to zero takes about 1 ms. This indicates that the current during depolarization is in the order of 100 µA.

We may also ask how many Na^+ ions are required for depolarization. This must be the same number of charges, although carried by different ions, that are responsible for generating the resting potential. Using again $Q = CU$, we find for a resting potential of $\Delta U = -75$ mV that an influx carrying a total charge of 7.5 pC, equivalent a total number of 5×10^7 Na^+ ions, is required to cross the membrane for compensating the K^+ charge imbalance. This number has to be set into perspective with the total number of K^+ ions in the cytoplasm. We have from Tab. 5.1 the concentration of K^+ ions in the cytoplasm: $n_{K^+} = 140$ mmol/l. The total number of K^+ ions is then $N_{K^+} = V_{cell} \times n_{K^+} \times N_{Av}$, where V_{cell} is the cell volume and N_{Av} is the Avogadro number. This yields about 4×10^{13} K^+ ions in the cytoplasm. Only a fraction $5 \times 10^7/4 \times 10^{13} \approx 10^{-6}$ or 1 ppm K^+ ions leak out through the open K^+ channels generating the resting potential. This is indeed a rather small fraction of K^+ ions that are involved in generating a resting potential. Vice versa, only a small number of Na^+ ions are required for depolarization. The absolute numbers quoted here depend on the cell size. Some cells are large, others are very small. The number of K^+ and Na^+ ions required for the resting poten-

tial and the depolarization varies accordingly, but the concentrations and fractions remain the same.

With these numbers at hand, we can now determine the membrane conductivity and eventually the single *channel conductivity*. According to Ohm's law $I_c = \Delta U/R = g \cdot \Delta U$, where $g = 1/R$ is the conductance. We find for all Na^+ channels in the membrane a conductance of 1.3 mS, which may or may not be similar to the K^+ channel conductance. The total current is distributed over N open Na^+ ion channels: $I_c = N \times i$. The conductance of each single channel depends on the density of channels per unit area, which is quite uncertain and varies strongly with location. Conversely, one can determine the conductance of a single channel by investigating the I-V characteristics using the patch-clamp technique and then determine the density of Na^+ ion channels. Such techniques indicate a current of 2 pA and a conductance of about 10 nS per channel [7]. This would indicate that the number of ion channels per cell is in the order of 10^6 or about 3000 channels per μm^2, which appears to be not unrealistic. Certainly these rough and ready estimates can be refined by considering contributions from other channels, nonlinearity effects, and proximity effects by neighboring channels, etc. Early models of channel conductivity were derived by Hodgkin and Huxley [8] and were later refined by a number of authors. With equivalent RC circuits for the cell the time constants for depolarization and repolarization can also be modeled. For a more detailed discussion of these properties we refer to [9] and the book by Alberts and coauthors listed under Further reading.

5.5 ATP pump

ATP consists of a carbon compound as a backbone and a phosphorous part as depicted in Fig. 5.8. Three phosphorous groups P_α, P_β, P_γ are bound by oxygen ions, and they also have side oxygen ions attached, which are negatively charged. These negative charges have a high repulsive potential. Removing one phosphate group from the end converts ATP to adenosine diphosphate (ADP). The reaction ATP → ADP + P releases energy of 30.6 kJ/mol. The body stores ATP as an energy source during the metabolic process. About one third of the total body energy consumption is used for synthesizing fresh ATP and converting ADP + P back to ATP. Indeed, ATP is the fuel that keeps the ion pumps and therefore the entire body running. In fact, ATP is the ubiquitous carrier of metabolic energy in all living cells.

The action of the Na^+/K^+ ion exchange pump powered by ATP is sketched in Fig. 5.9 (A–F). First we recognize that the ion pump is asymmetric and either open to one or to the other side of the cell membrane. This is clearly different from ion channels, which are always open or may have a gate that opens for diffusive transport via a proper gate potential. In the starting position of panel A the ion pump is open to the cytoplasm allowing Na^+ ions to diffuse in and find three proper specific lock sites. As soon as all three positive ions have settled down in their pockets, their combined

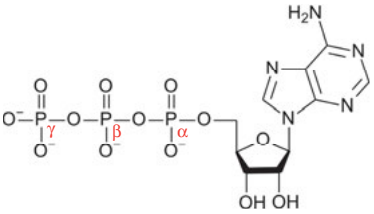


Fig. 5.8: Chemical structure of adenosine triphosphate (ATP) (adapted from https://commons.wikimedia.org/wiki/File:ATP_structure.svg, © Creative Commons).

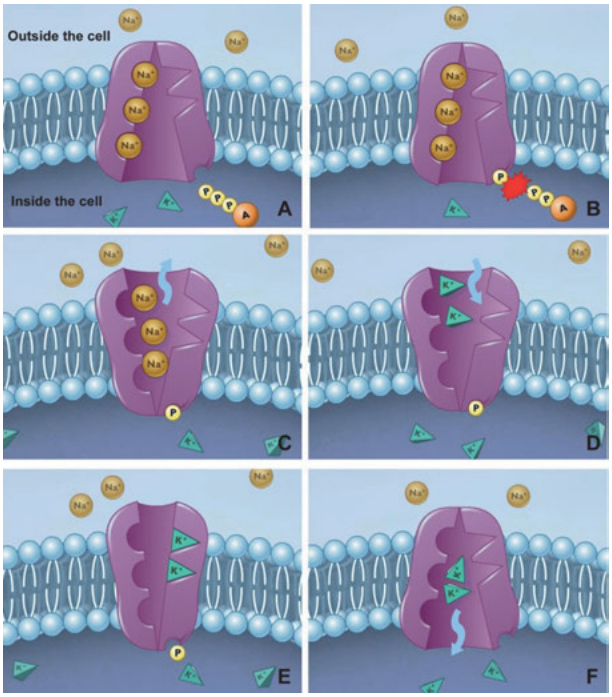


Fig. 5.9: The sodium-potassium exchange pump mechanism. The Na^+/K^+ pump moves two potassium ions from outside the cell to inside and three sodium ions from inside the cell to outside by the breakdown of ATP molecules. Further information on all six steps from A to E is given in the text (reproduced from https://openi.nlm.nih.gov/, open access).

positive charge triggers the phosphate end group P_γ to dissociate and attach to a side group of the ion channel (B). The energy release, in turn, results in a conformational change of the ion pump, which now opens to the outside, thereby releasing the Na^+ ions to the extracellular space, shown in panel (C). This allows two K^+ ions to diffuse into the ion pump (D). Once they have found their sites, these two positive charges repel the singly charge phosphate group, which will leave its bonding site (E). Once the phosphate group departs, the shape of the ion channel reverts back to the original conformation, releasing the K^+ ions (F) and starting a new cycle.

The rocking-like ion exchange pump powered by ATP is an extremely clever mechanism for active charge transport in wet electrochemical environments and for resetting the action potential to rest. This pump can be found in the membrane of virtually all human or mammal cells, but is particularly abundant in muscle and nerve cells. Similar mechanisms apply to the H^+/K^+ exchange in the intestines, to the resorption of Na^+ ions in the urine of kidneys, and to the Na^+/Ca^{2+} ion exchange in muscles.

5.6 Summary

1. Cell membranes define areas inside the cell (cytoplasm) and outside (extracellular space).
2. The cell membrane consists of a double lipid layer.
3. Without ion channels the membrane is impermeable to ion exchange.
4. There are two main types of channels distinguished by passive versus active ion transport.
5. Passive ion channels are either open or closed.
6. Passive ion channels are gated to open or to close either by voltage, ligands, or mechanical stress.
7. Passive channels are specific for the type of ion that is allowed to permeate through.
8. The cytoplasm has a much higher K^+ concentration than the extracellular space.
9. Vice versa, the extracellular space has a much higher Na^+ ion concentration than the cytoplasm.
10. The outflow of K^+ ions through open K^+ channels is responsible for generating a resting potential while Na^+ channels remain closed.
11. Only about 1 ppm of K^+ ions crossing the K^+ channels are required for setting the resting potential.
12. The ion exchange in equilibrium follows from the ratio of electric energy to thermal energy.
13. The resting potential is about –75 mV and is described by the Nernst equation.
14. Stimuli from receptors cause the Na^+ channels to open which depolarizes the cell and causes an action potential.
15. Depolarization requires about 1 ms.
16. Repolarization requires active ion transport against the ion concentration gradients in the cell.
17. Active transport is achieved by the ATP pump, resetting the action potential to a resting potential.
18. The total action potential last for about 5 ms.
19. During the refractory period cells cannot be activated

References

[1] Campbell NA, Reece JB. Biology. 9th edition. Benjamin Cummings; 2009.
[2] OpenStax CNX Biology, a free web-based textbook on biology, which can be accessed but not browsed.
[3] Tortoba GJ, Derrickson B. Principles of anatomy and physiology. 12th edition. John Wiley & Sons; 2009.

[4] Kole MPH, Ilschner SU, Kampa BM, Williams SR, Ruben PC, Stuart GJ. Action potential generation requires a high sodium channel density in the axon initial segment. Nature Neuroscience. 2008; 11:178–186.
[5] Goldman DE. Potential, impedance, and rectification in membranes. Journal of General Physiology. 1943; 27:37–60.
[6] Neher E, Sackmann B. The patch clamp technique. Sci Am. 1992; 3:44–51.
[7] Hamill OP, Marty A, Neher E, Sackmann B, Sigworth FJ. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. Pflügers Arch. 1981; 391:85–100.
[8] Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol. 1952; 117:500–544.
[9] Bezanilla F. Electrophysiology and the molecular level for excitability. The nerve impulse. Available at: http://nerve.bsd.uchicago.edu/

Further reading

Guyton AC, Hall JE. Textbook of medical physiology. 11th edition. Elsevier Saunders; 2006.
Alberts BE, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P. Molecular biology of the cell. 6th edition. Garland Science; 2014.

6 Signal transmission in neurons

6.1 Introduction

The nervous system has three functions: (1) detection of stimuli from the environment, translation into action potentials, and transmission of signals along nerve fibers to the central nervous system (CNS); (2) integration and processing of received data in the CNS; (3) emission of signals along nerve fibers from the center to the periphery for reactive movement. Signal reception causes graded receptor potentials; signal transmission requires action potentials, discussed in Chapter 5. In this chapter we will describe how these two potentials act together to transmit signals over long distances, i.e., information transport along neuron fibers. Since neurons are neither metal wires for charge transport nor glass fibers for photon transport, another "trick" is required to make signal transmission along nerve fibers responsive, fast, and reliable. In this short chapter we will only treat the transmission aspect of neurons; for all other aspects of the nervous system we refer to standard text books on human physiology and neurology listed at the end.

6.2 Overview on signal transmission

Neurons are cells, i.e., one neuron is one single cell. Nerves, in contrast, may consist of a whole body of neurons, which are interconnected. The neuron cell has four characteristic parts (Fig. 6.1): (1) cell body, also called *soma*; (2) multiples of *dendrites* going into the soma; (3) a single *axon* going out of the soma; (4) and an *axon terminal* connecting to a target cell. Dendrites and axons together are called nerve fibers. Dendrites are usually short; axons can be very long, stretching from brain to toe. Arrows in Fig. 6.1 indicate information flow. Dendrites collect signals within the areal reach of the cell, integrate those signals, and funnel them to the *axon hillock* (Fig. 6.1). The axon hillock is a specialized part of a neuron cell that connects to the axon. Receptor potentials that pass the axon hillock will fire an action potential that travels down the axon towards the axon terminal. To speed up signal transmission, the axon is coated by insulating sheathes, called *Schwann cells* after the physiologist Theodor Schwann (1810–1882). These cells are arranged like beads on a string and are separated by constrictions, called *Nodes of Ranvier* named after the anatomist Louis-Antoine Ranvier (1835–1922). The purpose of Schwann cells and nodes of Ranvier is a jump type (saltatory) conduction of the action potential from one node to the next, where the signal strength is boosted up and refreshed at the same time. Finally the action potential will reach its destination, for instance a muscle cell, where it may trigger a muscle contraction. Referring to the numbers in Fig. 6.1, the four essential steps of signal transduction are as follows:

DOI 10.1515/9783110372830-008

1. Signal reception: incoming signals are received at the dendrites and change the membrane potential.
2. Signal integration: changes in membrane potential which pass the axon hillock initiate action potentials;
3. Signal conductance: action potentials travel along axon fibers coated with Schwann cells and are transported to axon terminals;
4. Signal transmission: neurotransmitters are released at synapses to activate target cells.

We distinguish between *afferent* and *efferent neurons*. Afferent neurons – also called *sensory neurons* – transmit signals from the periphery to the center, i.e., to the spinal cord and from there to the brain. Efferent connections – also called *motor neurons* –

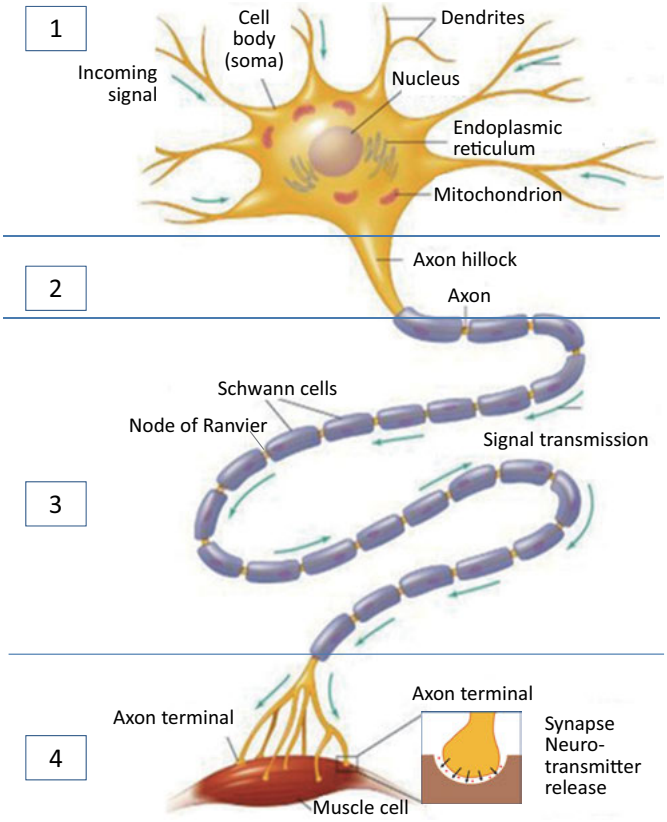


Fig. 6.1: Neuron consisting of dendrites, soma, axon, and axon terminal. The numbers refer to: (1) signal reception, (2) integration, (3) conductance, and (4) transmission through an axon from the central nervous system to the target cell. At the axon terminal the connection is established by neurotransmitters.

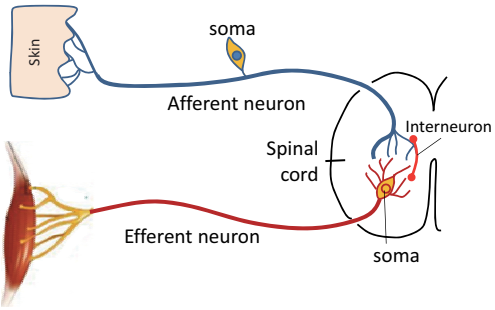


Fig. 6.2: Afferent neurons connect receptors with the CNS, efferent neurons conduct signals to action centers like a muscle. Afferent neurons are unipolar neurons, efferent neurons are multipolar neurons.

conduct signals from the center to the periphery to initiate actions (see Chapter 1 and Fig. 6.2). The neuron shown in Fig. 6.1 is an efferent neuron. The dendrite endings of this neuron are somewhere in the brain or in the spinal cord. There is no direct connection between a receptor and a muscle cell. All receptor signals first go to the center and from there back to the action center. There is a third type of neuron, called *interneuron* or *association neuron*, which is located within the CNS between sensory and motor neurons, making – in a few cases – a direct connection between them. Note that the afferent and efferent neurons have different shapes. Afferent neurons are unipolar neurons, whereas efferent neurons are multipolar neurons.

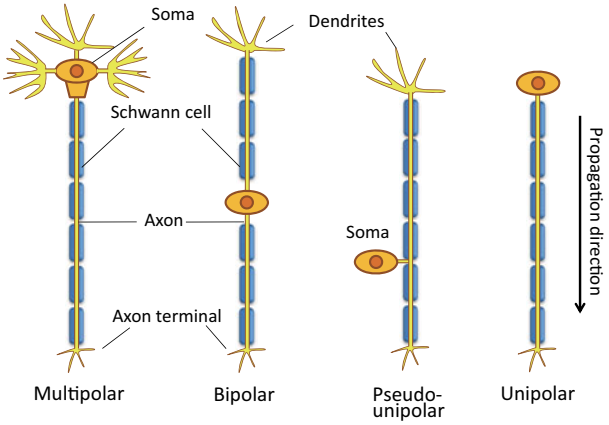


Fig. 6.3: Four types of neurons: multipolar, bipolar, pseudo-unipolar, and unipolar.

Furthermore we distinguish neurons according to the number of excitable dendrites connected to an axon (Fig. 6.3): multipolar, bipolar, pseudo-unipolar and unipolar. Multipolar neurons have multiple dendrites connected to the soma and one axon coming out. Most neurons of the brain and the spinal cord are of this type. Bipolar neurons have one branched dendrite going into the soma and one axon coming out. They can be found in the retina of the eye (Chapter 11), in the inner ear (Chapter 12), and in the olfactory (smell) center of the brain. Pseudo-unipolar neurons have one

branched dendrite going directly into an axon connecting to the terminal, whereas the cell body (soma) lies outside. Pseudo-unipolar neurons are in most cases afferent neurons, which connect a sensory receptor to the CNS. In unipolar neurons the axon connects the soma directly with the terminal. Figure 6.2 shows the typical situation where a pseudo-unipolar afferent neuron connects from the periphery to the CNS and a multipolar efferent neuron conducts the action potential back to the target tissue.

6.3 Sensory receptor potential

In the body we have many types of receptors which react to various stimuli: mechanoreceptors sensing touch, pressure, vibration, and sound (Chapter 12); thermoreceptors detecting changes in temperature; nociceptors responding to pain from physical or chemical damage of the skin; photoreceptors sensing light that hits the retina (Chapter 11); chemoreceptors for smell and taste; osmoreceptors for detecting osmotic pressure changes of body fluids. Although this looks pretty complete, we lack receptors for electrical or magnetic fields, unlike some animals.

Three main types of sensory receptor cells are sketched in Fig. 6.4. The receptors may be directly embedded in dendrites such as for temperature sensitivity (top panel), or the dendrites may be encapsulated for pressure sensing (middle panel), or the primary sensor may be a chemical sensor for taste or smell (bottom panel). All three types of receptors connect to afferent unipolar neurons that transport the information to the CNS. The first two receptors are called *primary sensory receptors*. The receptor potential from the taste and smell receptors are conveyed first into a re-

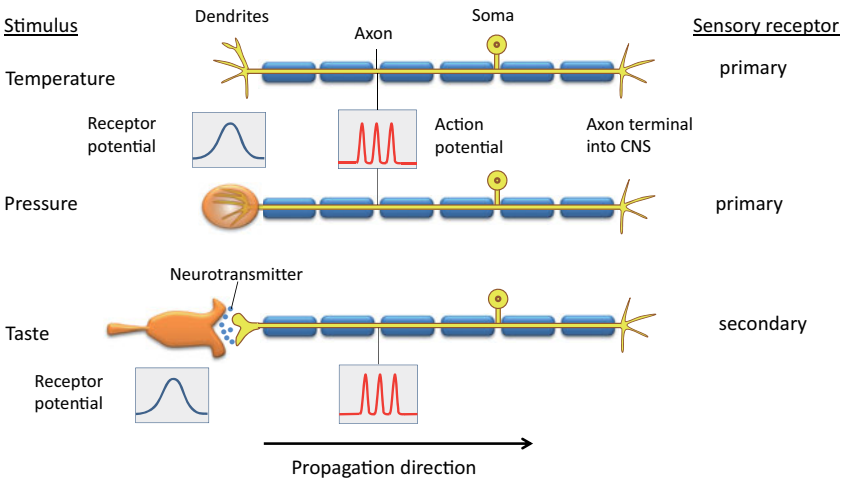


Fig. 6.4: Three types of sensory receptor neurons specialized for sensing temperature, pressure, and chemicals for taste and smell.

lease of neurotransmitters that are sensed by specific dendrites before being converted into action potentials. Sensors of this type are called *secondary sensory receptors*. The sensory part of primary and secondary receptors generates a change in the membrane potential in response to the stimulus, called the *receptor potential*.

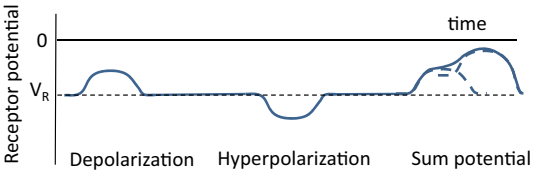


Fig. 6.5: Receptor potentials are graded potentials. The potential change may be positive (depolarization), or negative (hyperpolarization), or be the sum of two not temporally separated responses. V_R is the resting membrane potential.

The receptor potential responds to the stimulus continuously, without threshold and without refractory period. Therefore these potential changes are called *graded potentials*. The graded potential change may be positive (*depolarization*) or negative (*hyperpolarization*), and can be added to a sum potential (Fig. 6.5). Depending on the receptor type they may respond proportionally to the stimulus (*p-receptors*), or they may emphasize changes of the stimulus, called differential receptors (*d-receptors*). Pain receptors are of the p-type, temperature and smell receptors are of the d-type. Sense of smell is at the beginning strongest and dies off after a short time, so they are d-type. Most sensors are mixed pd-type. The difference between receptor potentials and action potentials is essential and important for our subsequent discussion. Therefore their main characteristics are compared in Tab. 6.1.

Tab. 6.1: Comparison between receptor potentials and action potentials.

	Receptor potential	**Action potential**
Threshold	no	yes
Refractory time	none	2–3 ms
Summation	yes	no
Polarization direction	Positive and negative	Only positive
Potential change	graded	all or none
Propagation characteristics	Passive and damped	Active and regenerated at each node
Initiation	Receptors for specific stimuli	All membranes containing fast voltage gated Na^+ channels

6.4 Analog-digital conversion

In Fig. 6.6 the sequence from stimulus to receptor potential and action potential is schematically plotted. The receptor potential in the dendrites reacts to the stimulus, where a pd-type reaction is shown. At the first node of Ranvier the potential signal is a mixed signal of graded receptor potential and action potential. A few nodes further down the graded potential is filtered out and only the action potential continues to propagate along the axon. The analog to digital conversion (ADC) encodes the strength of the stimulus and the duration. The stronger the stimulus, the higher is the frequency of action potentials. The frequency may decrease over time, but action potentials continue to be issued as long as the stimulus is "on" and produces a sufficient voltage at voltage gated Na^+ channels above threshold.

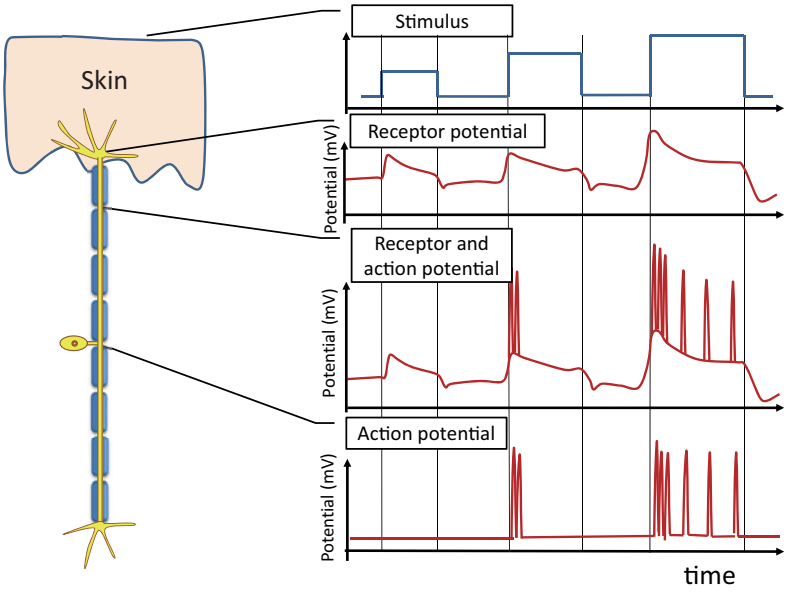


Fig. 6.6: Sequence of potentials from receptor potential to action potential.

ADC conversion is an absolute necessity if information needs to be transported fast over long distances. If the charge and discharge of receptor cells were continued over the total distance of an axon from the periphery to the brain, it would never arrive there. The depolarization decays exponentially with distance l from the receptor cell according to $\exp(-l/\lambda)$, where the extinction length λ is in the order of 2–5 mm. Therefore it is mandatory that receptor potentials are converted into action potentials as soon as they reach an axon. Only action potentials are refreshed at any node and can travel any distance without loss of signal strength.

The generation of action potentials in an axon close to a receptor cell is demonstrated in Fig. 6.7. Schwann cells coating the axon have a high electrical resistance and block all ion channels such that within its extension depolarization is suppressed. Transmembrane charges that have accumulated at the end of a receptor cell generate an electric field, which can jump to the next node of Ranvier. Voltage gated ion channels in the gap sense this strong electric field and depolarize. The depolarization is accomplished as soon as the threshold potential is surpassed. If the electric field from the receptor continues to be present, another action potential will be initiated as soon as the refractory time of the first one has ended, and so on. Therefore the completely different electrical properties of graded receptor potentials versus action potentials are responsible for the ADC conversion. Schwann cells help in speeding up this conversion, but even without Schwann cells the conversion is possible. After a few nodes of Ranvier the receptor potential is filtered out, as we have already seen in Fig. 6.6, and only the digital action potential propagates by jumping. At each node the action potential delivers a refreshed and complete depolarization that continues forever or at least to the end of an axon.

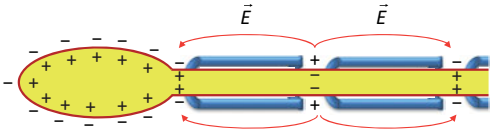


Fig. 6.7: Conversion of receptor potential into action potential. Red arrows indicate electrical field lines.

6.5 Saltatory polarization current

Now we consider the propagation of action potentials along axons after having described the conversion of receptor potentials into action potentials. At rest axons have positive charges (Na^+ excess) on the outside, negative charges (K^+ deficit) on the inside. The potential difference of an axon across its membrane has the usual value of about −75 mV. This potential is called the *transmembrane resting potential*. Simultaneously, the electric dipole polarization $\vec{p}$ points from the inside outwards. When activated, Na^+ channels open, reversing the transmembrane potential and the membrane polarization turns from outside in. In Fig. 6.8 each ion channel is symbolized by one pair of charges. The depolarization successively propagates down the axon (in Fig. 6.8 from left to right), which corresponds to a polarization current I_p oriented perpendicular to the polarization $\vec{p}$. As there is no potential gradient to tell the polarization current in which direction to flow, there must be another mechanism that determines the directional preference. This is achieved by the refractory time during which the polarization current can move forward to areas which have not yet been depolarized, but not backwards. Repolarization during the refractory time is the key to the directional preference of the polarization current.

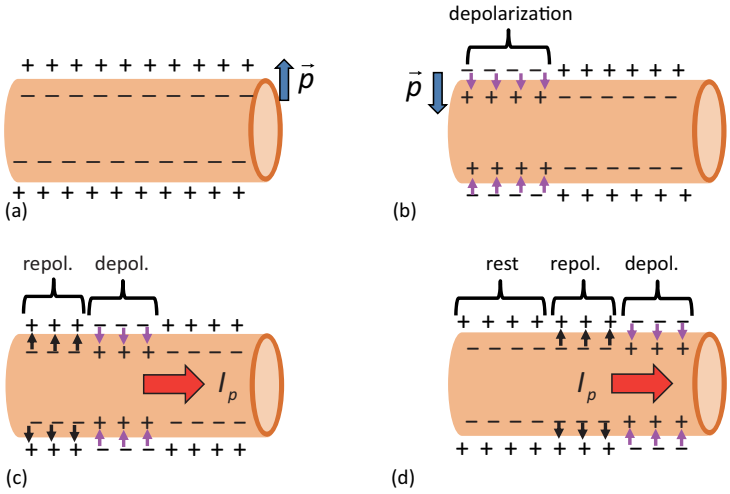


Fig. 6.8: Propagation of depolarization along axon fibers. Areas that have been depolarized have a refractory period for repolarization, which prevents the action potential from moving backwards.

Nerve fibers, which are not myelinated, have a signal propagation speed of about 2 m/s. This is not particularly fast, but sufficient for short distances such as in the brain. Higher speeds are required if long distances need to be traversed like the distance from the brain to the lower limbs. This is achieved by myelination of the nerve fiber. Myelin is a dielectric material forming sheaths that wrap around axon fibers. A myelin sheath can contain up to 300 bilipid membrane layers, each one adding to the resistance of the total transmembrane resistance. The cells that produce myelin are called Schwann cells, as already mentioned. Schwann cells inhibit any depolarization in their area. Then the transmembrane potential is forced to jump over a Schwann cell from one node of Ranvier to the next, as shown schematically in Fig. 6.9. In these nodes the axonal membrane is not insulated, permitting a rapid depolarization via a high density of Na^+ channels. By means of this depolarization the action potential becomes refreshed and maintains the original signal height. Nodes of Ranvier are po-

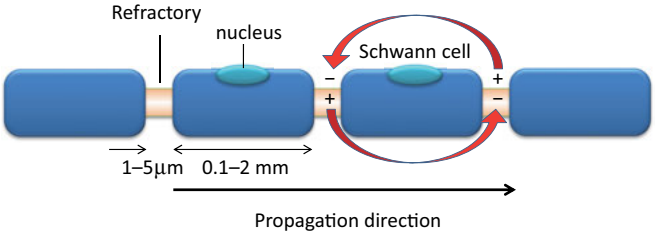


Fig. 6.9: Schwann cells along an axon interrupted by nodes of Ranvier, arranged like beads on a string, allow jump-like fast signal propagation.

sitioned every 0.2 to 2 mm apart and their length is about 1–5 µm. With myelination the jump-like or saltatory propagation of the polarization current reaches speeds of up to 130 m/s.

6.6 Communication across axons

Neurons form networks with other neurons. A single axon may have as many as 15 000 connections with which it communicates. Signal transmission across different axons is termed *synaptic transmission*. Synaptic transmission can be either of electrical or chemical nature.

In electrical junctions, also known as *gap junctions*, cells almost touch each other and are connected via common ion channels that allow electrical and ion transport. The ion channels form pairs, one channel from either membrane. In order to function, the ion channel pairs have to be in perfect registry as illustrated in Fig. 6.10. Gap junctions can be found in all tissues of the body guaranteeing very fast signal response. The channels of gap junctions are not continuously open, but opening and closing is regulated by Ca^{2+} ions. As such the permeability of gap junctions can change rapidly (within seconds) and reversibly, allowing an efficient communication between cells.

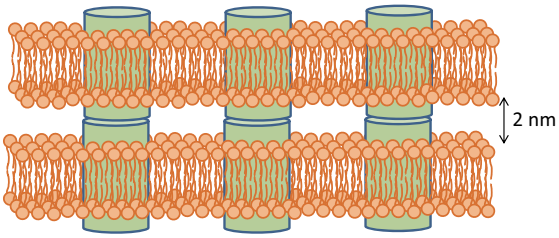


Fig. 6.10: Gap junctions between two membranes. Ion channels from both membranes pair up to a joined channel.

More common are *chemical synapses* which link two cells together by diffusion of chemical neurotransmitters across a large gap. In contrast to gap junctions, which are separated by only 2 nm, chemical synapses are separated by about 30 nm. Neurotransmitters are stored in presynaptic vesicles and are released into the synaptic space, also called *synaptic cleft*, in response to arriving action potentials, schematically shown in Fig. 6.11. The most frequent neurotransmitter in the CNS is glutamate (sum chemical formula: $C_5H_{10}N_2O_3$). The neurotransmitter molecules diffuse across the synaptic deft and activate the opposing cell by binding to specific receptor sites on the cell membrane. Thus the arriving digital action potential is converted into an analog signal. Low frequency action potentials release fewer chemical transmitters than high frequency action potentials. After arriving at the receptor of the postsynaptic cell, the analog signal again triggers a digital action potential (see Fig. 6.12). The frequency of the latter depends on the input not only of the presynaptic cell but also

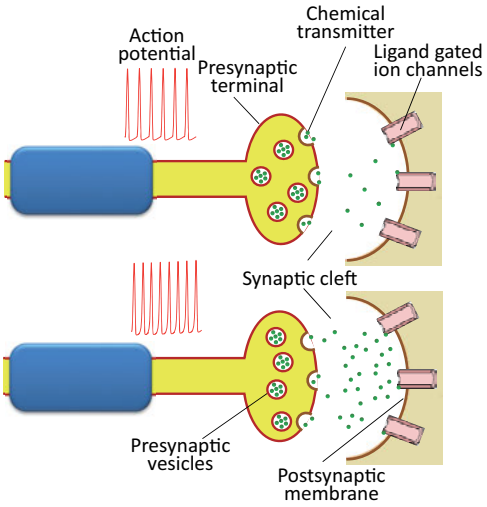


Fig. 6.11: Chemical transmitter is released in response to action potentials, converting a digital signal into an analog signal. Higher frequency of action potentials causes more neurotransmitters to be released.

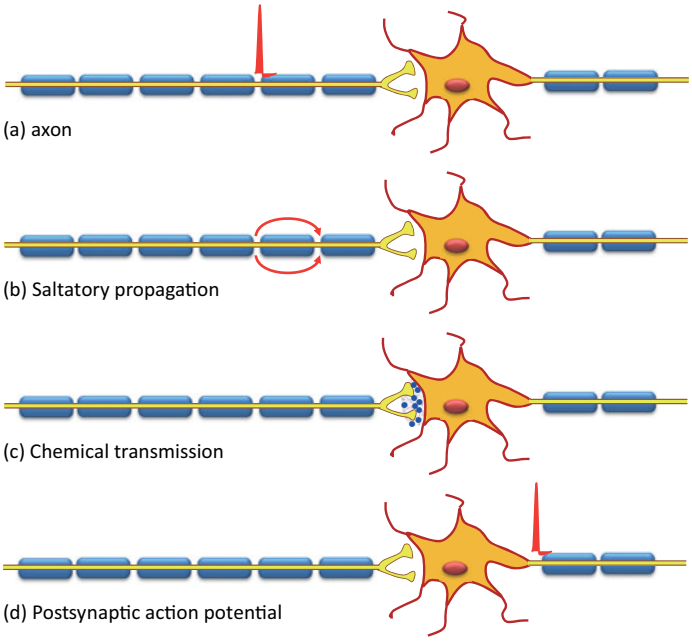


Fig. 6.12: Sequence of events at a chemical synapsis. (a) Action potential at nodes of Ranvier; (b) polarization jumps across Schwann cells; (c) release of neurotransmitters in presynaptic vesicles into the synaptic gap; (d) the chemical signal is converted back into a digital action potential in the postsynaptic cell.

on the input of all other neurons that bind to the synapsis. This input can be both excitatory (enhancing) or inhibitory (suppressing). If the sum of all input analog signals at the receptor site reaches the threshold potential of the postsynaptic cell, it will fire an action potential. Conversely, if the sum signal is too low, no action potential will result. Obviously chemical synapses offer a variety of different signal transmissions, including excitatory, inhibitory, and manipulatory by external chemicals such as anesthetics. The complete sequence of events is shown in the summary Fig. 6.12.

6.7 Neuromuscular junction – triggering muscle contraction

Efferent nerve fibers (motor neurons) terminate at muscle fibers and bring them to contraction via action potentials. Each individual muscle fiber is innervated by only one motor neuron. However, a single motor neuron is usually split up into many branches and can innervate many muscle fibers, between 10 to 1000 depending on muscle size. The combination of an individual motor neuron and all muscle fibers that it innervates is called one *motor unit*. There may be as many as thousand motor units that connect to one muscle. Figure 6.13 illustrates one motor unit that innervates with only three muscle fibers. For the distinction between muscles, bundles, fibers, myofibrils and myofilaments we refer to Section 2.4. Now we take a closer look at the processes that occur at one of the motor axon terminals when activated.

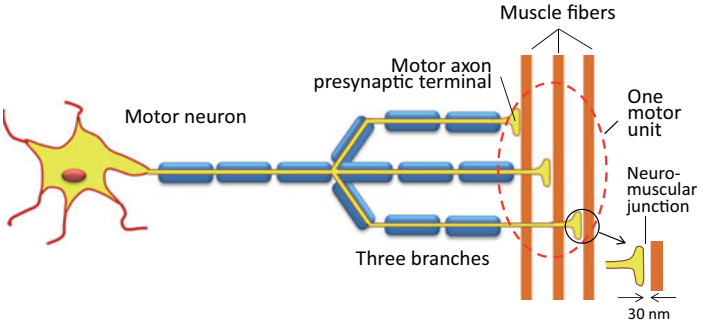


Fig. 6.13: Innervation of one motor neuron with fibers in a muscle.

The 30 nm wide gap between the motor axon terminal and a muscle fiber is the so called *neuromuscular junction* (NMJ). After the action potential is transmitted across the NMJ, an action potential is stimulated in all the innervated muscle fibers of that particular motor unit. The sum of the electric activity leading to the action potential is called *motor unit action potential* (MUAP). The NMJ is a chemical synapse, but differs in shape from synapses between neurons. The cross section of an NMJ is schematically shown in Fig. 6.14.

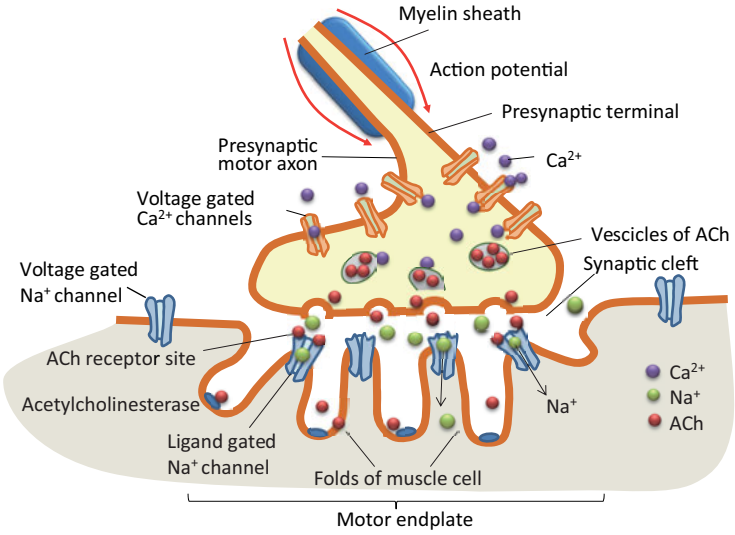


Fig. 6.14: Neuromuscular junction between an efferent presynaptic motor axon and a muscle fiber.

The presynaptic motor axon is not myelinated but contains a large number of voltage gated Ca^{2+} channels. Beyond the gap the muscle fiber is folded up to increase the surface area of the membrane exposed to the synaptic gap. These folds form the motor end plate of muscle cells hosting a huge number of receptors for the neurotransmitter acetylcholine (ACh) to bind on.

Now we consider the course of actions that take place at the NMJ once an action potential arrives at the end of a motor neuron. First, voltage gated calcium channels open up in the membrane of the presynaptic terminal. As Ca^{2+} ions enter the cytoplasm of the neuron, they bind to sensor proteins on vesicles containing the neurotransmitter ACh. This triggers the vesicles to fuse with the active zone of the membrane and to release the ACh that was previously synthesized in the terminal button and stored in the vesicles. Then ACh diffuses through the synaptic cleft and binds to ligand gated Na^+ channels in the folds of the motor end plate. The binding of ACh to Na^+ channels opens them. As Na^+ ions flow into the postsynaptic muscle cell, it depolarizes up to the threshold potential, triggering an action potential. This depolarization is called *end plate potential* (EPP). As the action potential spreads along the muscle cell, it causes the muscle cell to contract (see Chapter 2 for details). Then the ligand gated Na^+ channels close and repolarize, letting the muscle cell relax. ACh, which was attached to the Na^+ receptor site, is released and splits into acetate and choline by the enzyme acetylcholinesterase (ACHase) sitting in the muscle cell membrane. While this is taking place, the acetate and choline are actively transported back up into the presynaptic terminal to be resynthesized in ACh and encapsulated in vesicles. Then the sequence of events can start over again.

6.8 Spinal reflexes

Some afferent connections do not go to the brain but are automatically controlled by the spinal cord, from where they go back to the periphery. These connections are made possible by interneurons, also called association neurons, suitable for the control of the connectivity as well as any possible injuries in the spinal cord. The best known automatic reflex is the patellar tendon or knee-jerk reflex shown in Fig. 6.15.

Striking the patellar tendon just below the patella with a hammer (stimulus) will stretch the femur muscle and receptors within the muscle spindle. The signal from the receptors travels along the afferent sensory neuron to the spinal cord completely independent without interference from higher centers. In the spinal cord the afferent and efferent fibers are directly connected via an association neuron (interneuron). From the spinal cord efferent motor neurons conduct the signal back to the quadriceps and hamstring muscles, triggering a contraction of the quadriceps muscle and a relaxation of the hamstring muscle causing the leg to kick. There are a couple of other automatic reflexes active in the body, such as the heat reflex and the eye blink reflex. In the case of the heat reflex, the receptor afferent neuron connects to the spinal cord and automatically retracts the finger from the heat source by contracting the biceps brachii.

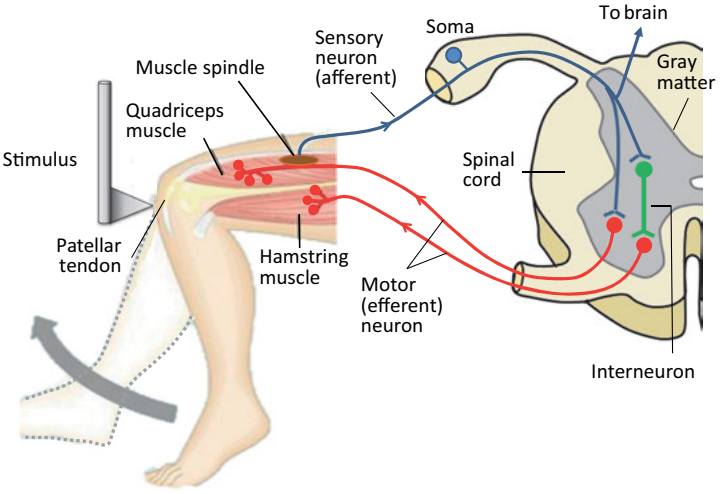


Fig. 6.15: Patellar-kicking reflex. By stretching the patellar tendon with a swift knock, the muscle spindle reflex is activated, sending an action potential along the afferent sensory neuron to the spinal cord. The direct synaptic connection between sensor neuron and motor neurons via the interneuron bypassing the brain causes a contraction of the quadriceps muscle and a relaxation of the antagonistic hamstring muscle, resulting in a swift upturn of the lower limb. The coordination of muscle contraction on the one hand and interneuron inhibitory signal to the hamstring muscle on the other hand allows unconscious balance on our feet.

6.9 Electromyography (EMG)

Electromyography (EMG) is a technique for evaluating and recording the electrical activity produced by skeletal muscles. EMG detects the electric potential generated by muscle cells when these cells contract and relax. Electrodes for EMG detection are either intramuscular needle electrodes or extramuscular surface electrodes. In the first case fine needles are inserted into the muscle and the transmembrane potential is measured. In the second case the extracellular potential is determined over a larger area.

There are two main applications of EMG: (1) observation of potential fluctuations as a function of time and in response to muscle contraction; (2) measurement of signal propagation along nerve fibers after stimulation. Both types of tests concern the proper functioning of muscles and fibers versus malfunctioning due to various diseases that affect myelination, neurotransmitters, ion channels, muscle fibers, and the peripheral nervous system in general.

A typical EMG procedure is shown in Fig. 6.16. Two electrodes are placed on the muscles and extracellular potential fluctuations are recorded. By taking the difference, the noise drops out, while small differences between position 1 and 2 remain. The power spectral density of the difference signal is characteristic for the twitch pattern of muscles in the relaxed state as compared to the contracted state. The amplitude and frequency distribution of the recorded potential fluctuations can hence be used for assessing normal versus diseased muscle contraction or nerve conductance. Indeed, EMG can be used for diagnosis of neurogenic or myogenic diseases. Furthermore, EMG is often used in sports medicine for measuring muscle contractibility.

Another use of EMG is to measure nerve conductivity, i.e., the ability to transmit action potentials to muscles. Nerve fibers that are degenerated, injured by accidents, or mechanically squeezed produce a numb feeling. In these cases a diagnosis via EMG is indicated. An electrode is attached to a motor nerve fiber and the motor nerve is stimulated to fire an action potential by exposing a short voltage pulse, schematically indicated in Fig. 6.17. At the muscle site the arrival time T_A is measured by the reaction of the muscle and respective potential change. Then the electrode is moved to position B and the arrival time T_B is measured again. From the time difference $\Delta T = T_A - T_B$ and the separation of the electrodes D_{AB} the motor neuron conduction velocity (MNCV) $v_{\mathrm{MNCV}} = \Delta T / D_{AB}$ can be determined and compared with standard values for particular nerves and muscles. Care has to be taken that the nerve tested is indeed an efferent motor nerve and not an afferent sensory nerve that has different velocity characteristics. Also branching of nerve fibers may give false results or, alternatively, have to be taken into account by using the standard Kirchhoff law for current branching when the nerve conductance is analyzed. For further information on EMG procedures we refer to [1, 2].

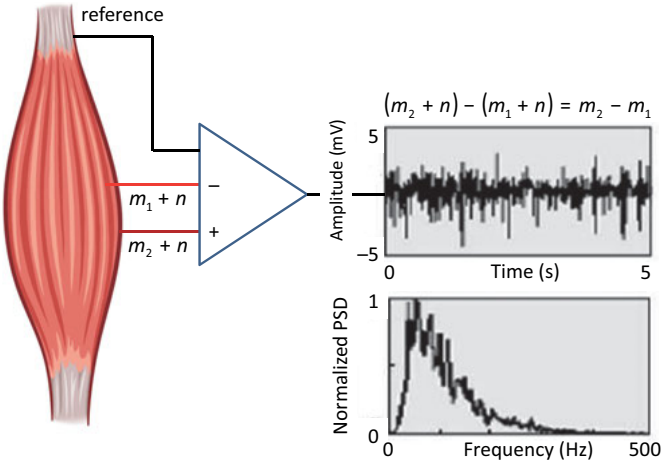


Fig. 6.16: Electrical potential fluctuations produced by muscle contraction can be tested by inserting electrodes into the muscle. n is the noise that cancels out by measuring at two positions. PSD is the power spectral density of the measured signal.

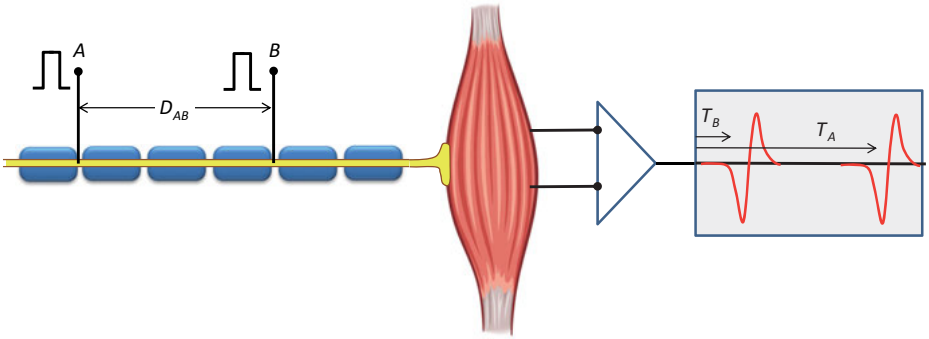


Fig. 6.17: Determination of the motor neuron conduction velocity using two electrodes separated by a certain distance.

6.10 Electroencephalography (EEG)

EEG is a method to record electrical potentials of the brain on the scalp as a function of time. Electrodes are attached with the help of a cap to different positions on the scalp (Fig. 6.18). The pattern of electrical activity that is recorded originates from nervous activities in the brain and is characteristic for the brain at rest, like sleeping, or for particular evoked reactions, like visual perception, muscle contraction, cognitive activity, etc. EEG is performed on patients to recognize normal brain behavior and to diagnose abnormal activities indicative for brain tumor, epilepsy, or stroke.

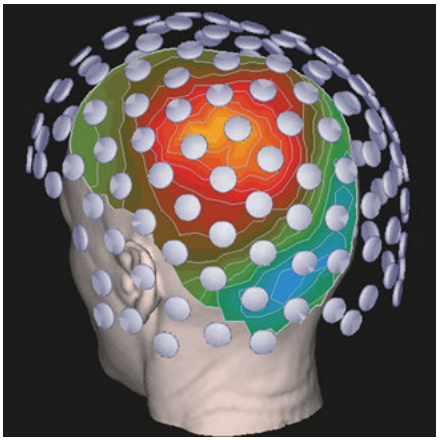


Fig. 6.18: EEG cap with electrodes attached (reproduced courtesy Los Alamos National Laboratory).

EEG competes with functional magnetic resonance imaging (fMRI) of brain activity. The latter method, discussed in Chapter 15, has the advantage of producing high spatial resolution images of brain activity in various slices through the brain. Compared to fMRI, EEG has much lower spatial resolution for mapping potential or voltage fluctuations generated in the upper part of the brain close to the scalp. But EEG has much higher time resolution of 1 ms or below. Another important difference to fMRI is the fact that EEG shows brain activity directly, whereas fMRI signals result from enhanced oxygenation of blood flowing near active neurons. Therefore EEG can also be used for studying brain activity during sleep. Furthermore, EEG is cost efficient, mobile, noninvasive, and patients do not need to be immobilized as is the case with fMRI. Therefore EEG is frequently used in clinical practice and in research although the patterns are not easy to interpret. An example is shown in Fig. 6.19.

In spite of all kinds of ion currents constantly flowing in and out of cells, on the whole the body is charge neutral. But transient electrical dipoles occur all over the body and in particular in the brain, where the frequency of action potentials is highest. From this we conclude that EEG registers the electrical stray fields of dipoles. However, considering that the dipoles are arranged randomly in space we expect that the sum over all dipoles yields a zero potential. It turns out that in the upper part of the brain an EEG signal can indeed be recorded from synchronous activity of aligned neurons. Nevertheless, the noise level is rather high. Aggravating the problem are jittered potential fluctuations at variable latencies. Therefore the best results are achieved by precisely phase locking EEG signals to preceding events. Further examples are shown on the webpage listed in [3].

An alternative method for recording brain activity is magnetoencephalography (MEG), which records magnetic Oersted fields generated by ionic currents in the brain (Fig. 6.20) instead of electrical dipole fields in EEG. These magnetic fields are extremely weak, recorded amplitudes are only in the order of 50 pT. The earth's magnetic field is 50 μT, which is 10^6 times stronger than the Oersted field of the brain.

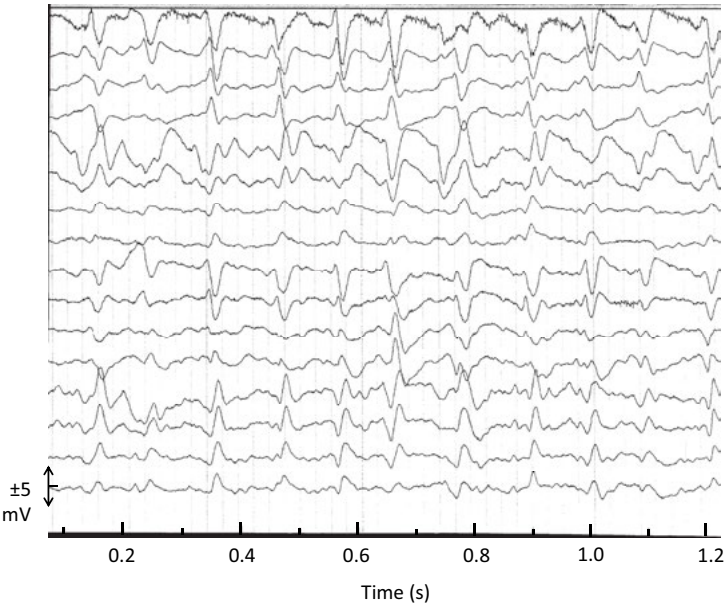


Fig. 6.19: EEG recording of brain activity.

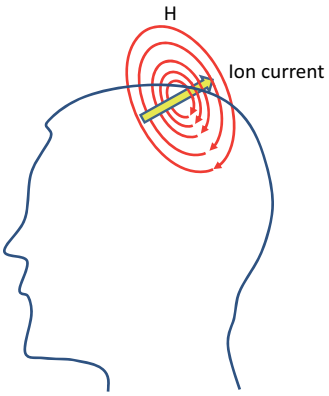


Fig. 6.20: Magnetoencephalography recording of magnetic fields generated by ion currents in the brain. The recording is performed with SQUID sensors distributed over the scalp in a magnetically well shielded hutch.

Therefore the most sophisticated SQUID sensors are required to detect these weak fields on the scalp of a person who must be extremely well shielded against any magnetic stray fields from the environment. It has been estimated that 50 000 neurons need to fire simultaneously in order to give a detectable signal. Time resolution and information gain is similar to EEG, but at a much higher price. The advantage over EEG is the better localization of brain activity because magnetic stray fields have a shorter range compared to electric dipole fields. Because of the technical difficulties and high investment costs, MEG is not often found in clinical practice but rather more so in research.

6.11 Summary

1. Neurons are single cells containing four parts: cell body, dendrites, axon, and axon terminal.
2. Nerves consist of an ensemble of interconnected neurons.
3. Long axons are myelinated by Schwann cells and interrupted by nodes of Ranvier.
4. The four essential steps of signal transduction are: (a) signal reception; (b) signal integration; (c) signal conductance; (d) signal transmission.
5. Afferent or sensory neurons conduct information from the periphery to the central nervous system.
6. Efferent or motor neurons conduct action potentials to muscles for motion.
7. Afferent and efferent neurons are not connected and have different pathways via the spinal cord to the brain and back.
8. In a few exceptions there are direct connections between afferent and efferent neurons in the spinal cord via association neurons. One example is the patellar reflex.
9. Stimuli from the environment are detected by specialized receptors.
10. The receptor signal can be proportional or differential.
11. The amplitude of the stimulus is encoded in the frequency of the action potential.
12. The action potential propagates via a salutatory polarization current along the axon fiber.
13. The propagation velocity of signals in myelinated axons can be up to 130 m/s.
14. Different axons communicate via synaptic transmission.
15. Synaptic transmissions can be electrical or chemical.
16. Chemical synapses convert digital axon potentials into analog chemical transmitters, and then back again into digital potentials in the postsynaptic cell.
17. Efferent axons are connected to muscles fibers via neuromuscular junctions.
18. Muscle contraction is triggered by the reception of neurotransmitters that depolarize the motor end plate.
19. One axon controls one motor unit of a muscle. One motor unit may contain up to 1000 neuromuscular junctions.
20. Electromyography is used for testing muscle activity and neuron conductance.
21. Electroencephalography and magnetoencephalography are used for analyzing brain activities.

References

[1] Broman H, Bilotto G, de Luca CJ. Myoelectric signal conduction velocity and spectral parameters: Influence of force and time. J Appl Physiol. 1985; 58: 1428–1437.
[2] Rodriguez-Falces J, Duchateau J, Muraoka Y, Baudry S. M-wave potentiation after voluntary contractions of different durations and intensities in the tibialis anterior. J Appl Physiol. 2015; 118: 953–964.
[3] Brain mapping: www.cerebromente.org.br/n03/tecnologia/eeg.htm

Further reading

Kandel ER, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth AJ. Principles of neural science. McGraw Hill; 2012.

Purves D, Augustine GJ, Fitzpatrick D, Katz LC, LaMantia AS, McNamara JO, Williams SM, editors. Neuroscience. 2nd edition. Sunderland (MA): Sinauer Associates; 2001. Online textbook can be accessed but not browsed: www.ncbi.nlm.nih.gov/books/NBK11059/

Tortoba GJ, Derrickson B. Principles of anatomy and physiology. 14th edition. John Wiley & Sons; 2015.

Malmivuo J, Plonsey R. Bioelectromagnetism: Principles and applications of bioelectric and biomagnetic fields. Oxford University Press; 1995. Available at: www.bem.fi/book/

7 Electrophysical aspects of the heart

7.1 Introduction

The circulatory system of all vertebrates including humans consists of three main functional parts: A pump (*heart*) keeps a liquid (*blood*) circulating through a closed system of tubes (*vessels*). In this chapter we focus on the electrophysical aspects of the heart that make the heart contract and thus act as a pump. The mechanical aspects of the heart and the circulatory system are topics of Chapter 8. For further physiological information we refer to standard textbooks on medical physiology and cardiology listed at the end of this chapter.

The heart is unquestionably a remarkable pump. With a power consumption of merely 6 Watt the 300 g light muscle pumps on average 7 l/min with each ventricle, which makes roughly 10 000 l per day corresponding to one cubic meter of blood, and twice this amount for both ventricles. Left and right ventricles of the heart operate in series and they must do this well balanced and synchronously. The heart consists of four chambers with complete separation of oxygenated and de-oxygenated blood. The right ventricle pumps blood to the lungs, while the left ventricle pumps blood to the rest of the body. Some electrophysical aspects of the heart are discussed here. Figure 7.1 shows a cross section of the heart with the designations of different parts that are discussed later in more detail.

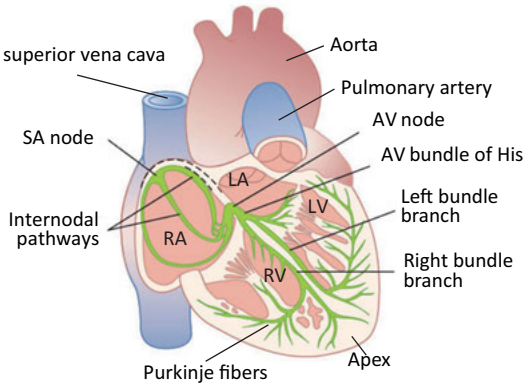


Fig. 7.1: Cross section of the heart showing the electrical conduction system. Arrows indicate the propagation direction of the cardiac action potential. Details are discussed in the text. AV = atrioventricular, SA = sinoatrial, RA = right atrium, LA = left atrium, RV = right ventricle, LV = left ventricle (adapted from www.medicinehack.com/).

DOI 10.1515/9783110372830-009

7.2 Cardiac action potential

The heart can be considered as an extended muscle that depolarizes upon stimulus just like a normal skeletal muscle cell. However, the *cardiac muscle*, also called *myocard* or *myocardium*, is distinguished from skeletal muscles by three important differences: (1) cardiac muscles have a much longer lasting action potential and a longer refractory period before the next depolarization can be activated; (2) the cardiac muscle contains pacemaker cells which are self-excitatory; (3) cardiac muscles do not have the ability of tetanic contraction for increasing tension, as skeletal muscles do (see Fig. 2.17). We recall from Chapter 5 that the action potential of nervous cells last for about 2–3 ms from depolarization to repolarization. In contrast, the *cardiac action potential* (CAP) is characterized by a fast depolarization, fast but partial repolarization, followed by an extended plateau phase before complete repolarization takes place. The whole action potential as seen in Fig. 7.2 lasts for about 200 to 400 ms. This is 100 times longer than the action potential of a nerve fiber. The *absolute refractory time* during which the heart is insensitive to new action potentials stretches from depolarization to repolarization. The *relative refractory period* allows depolarization but only with an enhanced threshold potential. The resting phase between two ventricular action potentials corresponds to the diastolic or filling phase. During the plateau phase the action potential, starting from the sinoatrial node, spreads over the entire myocardium causing the myocardium to contract. Hence the plateau phase corresponds to the ejection phase.

The rapid depolarization is achieved by opening of fast Na^+ ion channels. Complete depolarization is reached at a potential of approximately +40 mV. Early repolarization occurs when the Na^+ ion channels close and a small number of K^+ ion channels

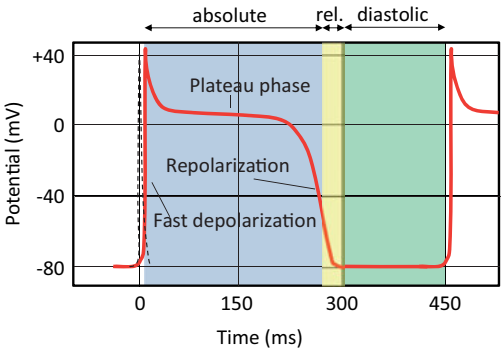


Fig. 7.2: Action potential of the heart is characterized by very fast depolarization, extended plateau region and repolarization after about 300 ms. The action potential is typical for a ventricular muscle fiber, as recorded by means of microelectrodes. The dashed black line shows the action potential of a skeletal muscle cell on the same timescale for comparison. Absolute and relative refractory periods as well as the diastolic phase are indicated by shaded areas.

open. A complete repolarization through K^+ ion channels is hindered by an influx of Ca^{2+} ions. Ca^{2+} ion control is typical for skeletal and cardiac muscle cells, but in the latter it is essential for maintaining a prolonged plateau phase. The Ca^{2+} ion concentration is rather low, but due to the double ionization, it very efficiently reduces the permeability of K^+ channels that otherwise would be open. Repolarization by K^+ ion channels is only possible once the Ca^{2+} concentration is exhausted. The time evolution of the ion channel permeability for Na^+, Ca^{2+}, and K^+ is plotted in Fig. 7.3. From this plot it becomes clear that slow Ca^{2+} channels control the long plateau phase which is most characteristic for cardiac action potential.

Cardiac excitation starts in the *sinoatrial node* (SA), which is located at the entry of the right atrium as indicated in Figs. 7.1 and 7.4. The SA node consists of specialized cardiac cells, called *pacemaker cells*, generating action potentials spontaneously and periodically. The action potential is fired at the SA node during the diastolic phase and then spreads out via conducting fibers in a highly organized fashion to other parts of the heart, causing Na^+ ion channels to open resulting in further action potentials, which stimulate the myocardium to contract.

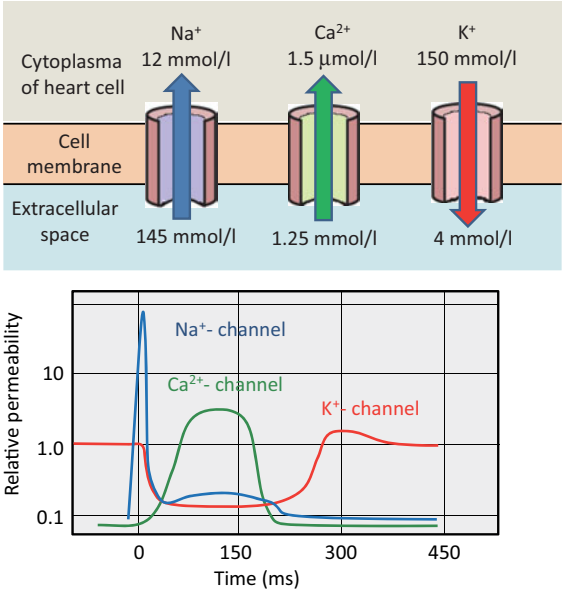


Fig. 7.3: The cell membrane of the heart contains three main ion channels for Na^+, K^+, and Ca^{2+}. The temporal opening of these channels, increasing their permeability during a cardiac action potential, is shown in the lower panel. Cardiac cells express six different types of K^+ channels, which are required for maintaining the resting potential and for shaping the plateau phase.

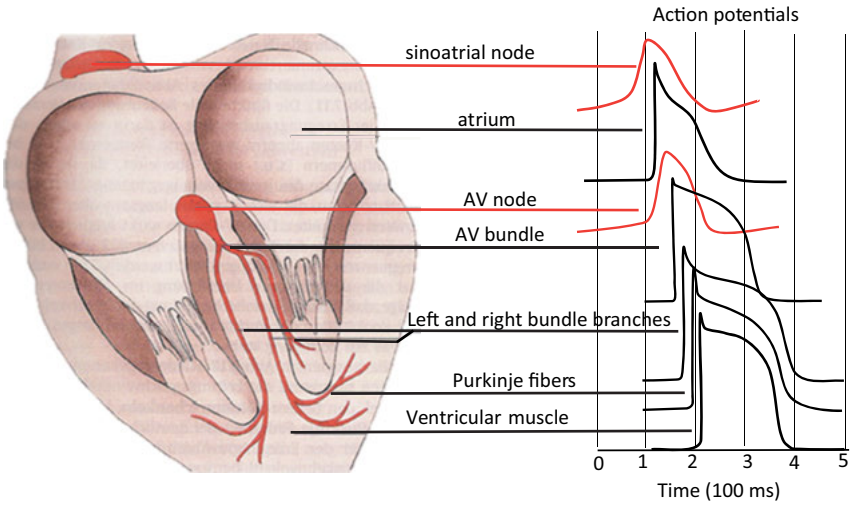


Fig. 7.4: Spreading of action potential across the heart starting from the sinoatrial node as the self-organized pacemaker (adapted from Pape, Kurtz, and Silbernagel, with permission of Thieme Verlag, Stuttgart, New York, 7th edition, 2014).

The *action potential* at the SA node as plotted in Fig. 7.4 differs from action potentials of muscle cells in several ways. It does not have a stable resting potential. Starting from a maximum diastolic potential of about –60 mV – much less than the surrounding –90 mV of the myocardial muscle – there is a slow spontaneous depolarization, called diastolic depolarization or pacemaker potential. The threshold for starting an action potential lies at about –50 mV. The depolarization is mainly caused by a Ca^{2+} current carried by voltage gated Ca^{2+} selective ion channels. A contribution of voltage gated Na^{+} channels is lacking, since these are deactivated due to the less negative diastolic membrane potential (–60 mV) as compared to myocardial cells at –90 mV. The action potential in the SA node lacks a distinct plateau phase and repolarization starts immediately after reaching the maximum. From threshold to repolarization it takes about 100 ms. The SA pacemaker under resting conditions produces a heart rate of about 70–80 beats per minute or 1.2–1.3 Hz.

Once the SA node has fired an action potential, the conducting system of the heart relays the action potential to other parts of the cardiac muscle. This is possible because the muscle cells are highly interconnected and electrical current flows between muscle cells via *gap junctions*, which provide a low resistance cell-to-cell coupling (see Section 6.6). The sequence of depolarizations at different locations in the heart is shown in Fig. 7.4. First the atrium depolarizes, than the action potential jumps over to the *atrioventricular (AV) node*. It is the only conductive connection between atrium and ventricle. From there the action potential is conducted via the AV bundle, which splits in left and right bundles, down to the *Purkinje fibers*, named after the Czech phys-

iologist J. E. Purkinje (1787–1869), and finally captures the ventricular muscle, which than contracts. It is important to note that the cardiac pulse traveling down the bundle branches arrives at the same time, ensuring that both ventricular chambers contract simultaneously. This is essential for effective pumping power. The AV bundle and the Purkinje filaments are highly conductive fibers, even more so than the cardiac muscle cells that convey the action potential quickly to the lower ventricle. The high conductivity of the AV bundle and Purkinje fibers is achieved by gap junctions between the cells. The SA node is the primary pacemaker of the heart. But the AV node also shows this ability, however at a lower frequency of 0.6 to 1 Hz or 40–60 beats per minute. The AV acts as a low pass filter: if the SA node frequency is too high, the AV node will not pass it on. Both the SA node and AV node have efferent innervations via the sympathetic and parasympathetic nervous system, which controls the Ca^{2+} ion channel permeability and thereby the beat frequency. Indeed the heart rate goes up when we become excited.

7.3 Electric polarization of the heart

Having described the spreading of the action potential from the SA node to the ventricular muscle, i.e., from the diastole phase to the systole phase, we are now prepared to analyze the electrical potential changes which can be detected on the skin during a heart cycle. The action potentials plotted in Fig. 7.4 are characterized as *transmembrane potentials*. They are measured with a pair of electrodes, one punching through the cell membrane into the cytoplasm and the other one on the outside. The extracellular potential is measured outside along a muscle or a nerve fiber. The correspondence between transmembrane potential and extracellular potential during an action potential is shown step by step in Fig. 7.5.

We start with the resting potential (panel (a)), which has the highest negative transmembrane potential difference. On the outside, in contrast, no potential difference is measured. When the cell is partially and locally depolarized (panel (b)), the transmembrane potential decreases and the extracellular potential shows a peak. At the same time an electric dipole moment occurs pointing from the negatively charged (depolarized) side to the positive side. Upon complete depolarization in the plateau phase (panel (c)) the transmembrane potential is reversed, while the extracellular potential is again zero. On the way back to the resting potential, the electric dipole moment increases again pointing in the opposite direction, while the extracellular potential has a negative amplitude (panel (d)) before returning to zero in panel (e). The extracellular potential observed here pretty much matches the potential changes found during a cardiac cycle, as we will see further below.

Each dipole represents the excitation (depolarization) of one muscle cell at a particular time and space during an action potential. When the CAP spreads over the atrium to the ventricle, billions of muscle cells are affected, all featuring their local

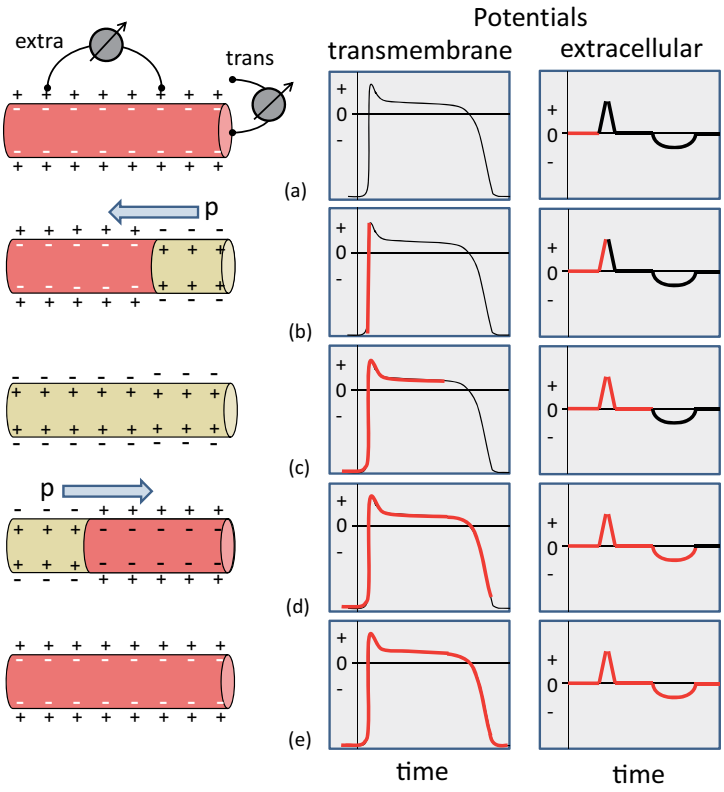


Fig. 7.5: Time evolution of the extracellular and transmembrane potential of a cardiac muscle cell during an action potential.

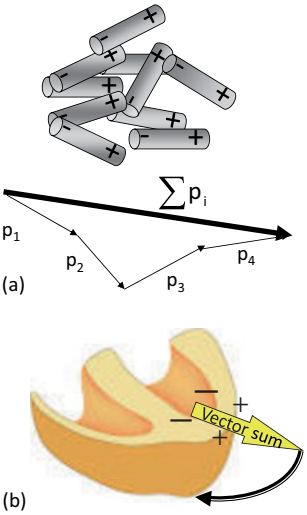


Fig. 7.6: Resultant vector of dipole moments generated during an action potential in the muscle cells.

and temporal extracellular potential difference forming electric dipoles. The vectors of all individual dipole moments superimpose yielding a resultant vector sum, as illustrated in Fig. 7.6 (a).

The effective sum of all dipoles characterizes the temporal and spatial spreading of the cardiac excitation. Panel (b) of Fig. 7.6 shows one snapshot. When we follow this vector moving in space, we will notice a rather complex threefold vector looping: a small loop for the atrial depolarization (P-loop), and a much bigger loop for the apical and ventricular depolarization (QRS-loop), and at the end another small loop during ventricular repolarization (T-loop). These three loops are shown with different projections in Fig. 7.7.

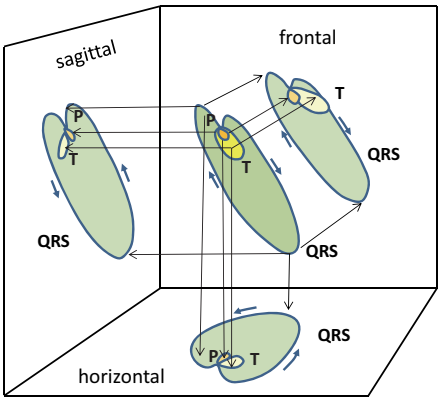


Fig. 7.7: Loops of the resultant dipole moment during a cardiac cycle and their projections on the sagittal, horizontal, and frontal plane. Three loops can be distinguished, labeled P, QRS, and T-loop (adapted from [1]).

The electric dipole moment is the result of ion mobility and charge fluctuations in the heart taking place during the *cardiac cycle*. The dipole moment, in turn, generates electrical fields and potential differences that can be measured externally on the skin as sketched in Fig. 7.8. Isopotential lines between different charges exhibit measurable potential differences providing information on the internal dynamics of the heart. Furthermore, when the potential difference is measured with an electrometer for different orientations on the skin, for instance parallel to the lines indicated by the blue triangle

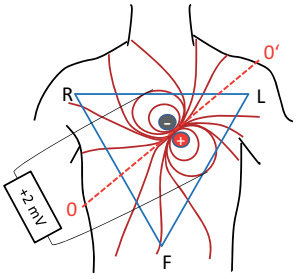


Fig. 7.8: Electric dipole moment due to charge fluctuations during the cardiac cycle. The blue triangle is the Einthoven triangle.

in Fig. 7.8, the orientation of the dipole can be reconstructed. This is the essential point of *electrocardiograms*, invented by Willem Einthoven at Leiden University, for which he received the Nobel Prize in medicine in 1925.

7.4 Electrocardiography (ECG)

Before continuing the analysis of the circulating dipole moment, we first describe the non-Cartesian triangular coordinate system introduced by Einthoven, which is better adapted to the physiology of the body than the Cartesian coordinate system. With reference to Fig. 7.9, the projections I, II, and III, also called leads, relate to potential differences measured when electrodes are attached to the extremities: between both arms (I), between right arm and left foot (II), and between left arm and left foot (III). In lead I the negative electrode connects to the right arm and the positive electrode to the left arm. Therefore when isopotential lines from negative charges point to the right arm and the isopotential lines from positive charges point to the left arm, the reading on the electrometer is positive. The other two leads II and III have corresponding sign conventions. The blue arrows in Fig. 7.9 reflect the sign convention for positive potential differences: in lead I from R to L, in lead II from right arm to left foot, and in lead III from left arm to left foot.

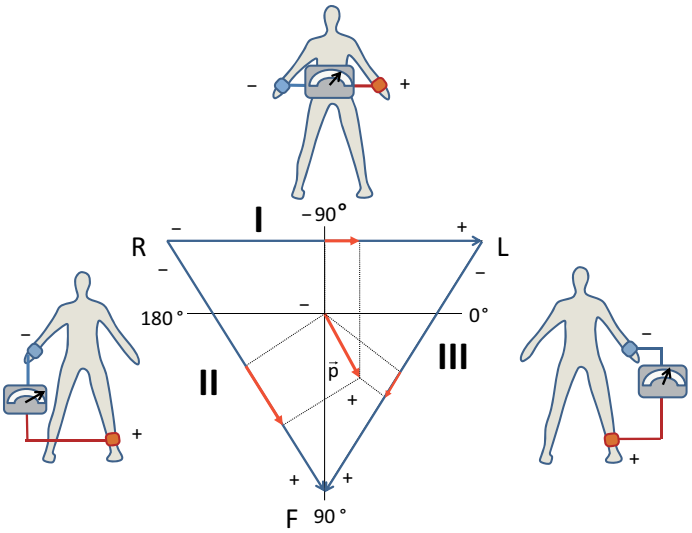


Fig. 7.9: Einthoven triangle defining the projections of electric dipole moments and for determining potential differences between extremities. I, II, III are the leads; R, L, F stand for right, left, and foot.

Formally the potential differences are defined as follows, where Φ_i are the respective potentials at the limbs:

$$\begin{aligned} \text{Lead I:} \quad & \Delta V_{\mathrm{I}} = \Phi_{\mathrm{L}} - \Phi_{\mathrm{R}} \\ \text{Lead II:} \quad & \Delta V_{\mathrm{II}} = \Phi_{\mathrm{F}} - \Phi_{\mathrm{R}} \\ \text{Lead III:} \quad & \Delta V_{\mathrm{III}} = \Phi_{\mathrm{F}} - \Phi_{\mathrm{L}} \end{aligned}$$

Since the *Einthoven triangle* is over determined, from the potential difference of two leads the third one follows according to Kirchhoff's law:

$$\Delta V_{\mathrm{II}} = \Delta V_{\mathrm{I}} + \Delta V_{\mathrm{III}}.$$

By way of example we assume that the potential at the right arm is −0.2 mV with respect to the average potential of the body (reference point), at the left arm a potential of +0.3 mV is measured, and at the left foot a potential of +1.0 mV. Then the potential difference ΔV_{I} measured in lead I between right and left arm is +0.5 mV, in lead II it is +1.2 mV, and in lead III +0.7 mV. Both potential differences add up to the third one, if not only the signs of the potential differences are respected but also the proper vector addition.

Now we come back to the dipole projection. If a dipole projection is parallel to one of the blue arrows, the potential difference is positive for that particular direction, otherwise it is negative. The length of the projected dipole onto one of the three leads is proportional to the potential difference measured along these leads. The standard nomenclature for the angular orientation of dipoles is also indicated in the graph in Fig. 7.9. If the resultant dipole is oriented horizontally, it may have either 0° or 180°. The orientation −90° is equivalent to 270°, but −90° is the preferred terminology in medical literature.

Let's consider one particular dipole shown in red in Fig. 7.9 with an angular orientation of 60°. For this orientation the dipole has positive projections on all three leads. As the projection on lead II is largest, the potential difference measured in this direction is highest. The other two projections are also positive but smaller.

We are now prepared to connect the potential differences determined by leads I, II, and III, with the actual cardiac cycle. For the same phase of the cardiac cycle each lead will show different but related potential variations from which the functioning of the heart, or in pathological cases the malfunctioning, can be inferred. This is known as *electrocardiography* (ECG) and the plots of potentials versus time are *electrocardiograms*. One example is shown in Fig. 7.9 and in more detail in Fig. 7.10 (a), which represents the beginning of the cardiac cycle when the depolarization starting at the SA node has spread over the atria. This causes a strong dipole moment pointing parallel to the heart axis. The positive amplitude measured in all three leads is known as the *P wave*. It occurs after about 0.1 seconds and the amplitude is about 0.5 mV. Shortly after the depolarization of the atria is completed, the P wave returns to zero potential, the action potential crosses the AV node. The propagation through the AV node is very sluggish resulting in a delay in the progress of activation. The delay

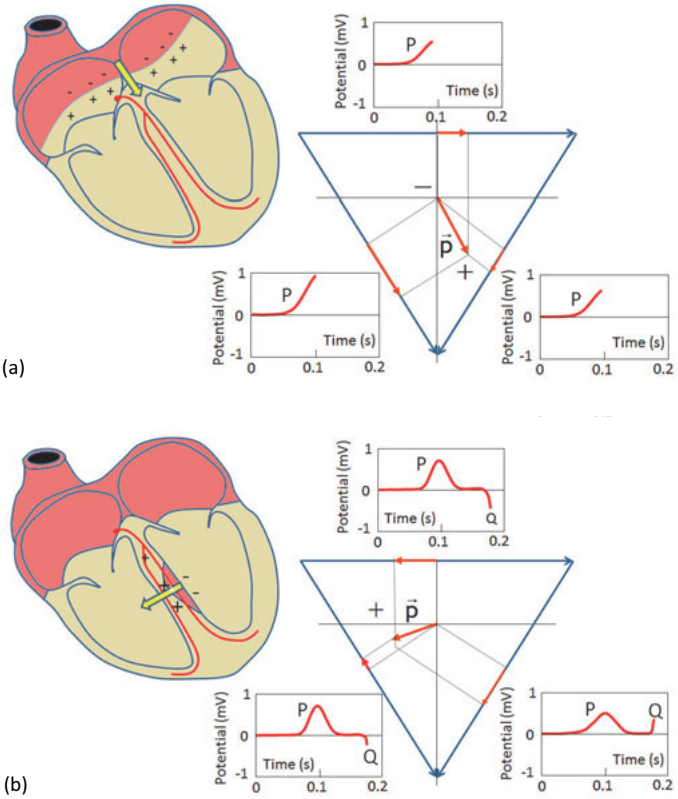


Fig. 7.10: Depolarization of the heart during two different phases: (a) depolarization of the atria generating the P wave; (b) depolarization of the left side of the ventricular septum generating the Q spike. The potential changes are shown for all three leads as a function of time. Red color in the heart contours corresponds to a depolarized state, ocher color to a resting state.

allows completion of ventricular filling. After activating the AV node, the ventricular septum becomes depolarized and a dipole occurs at an angle of about 170°. The projection of this dipole is negative in leads I and II, but positive in lead III. This point is referred to as the Q point and is shown in Fig. 7.10 (b). In lead III the Q point merges with the R wave and is therefore difficult to recognize. The time period between the P and the Q wave is called the PQ interval, which takes about 0.16 seconds. Sometimes it is also called PR interval, because the Q wave is not very pronounced. After passing the Q spike, both ventricles and the apex become depolarized, which generates the largest dipole moment and the highest potential difference in the R peak of about 4 mV. Then follows a rapid change where only the left ventricle outside the wall shows a polarization resulting in the S spike. Now all parts of both ventricles are depolarized and the potential returns to zero. Repolarization of the ventricles causes the T wave. The time sequence of the polarization changes are plotted in Fig. 7.11 together with the potential changes according to lead II.

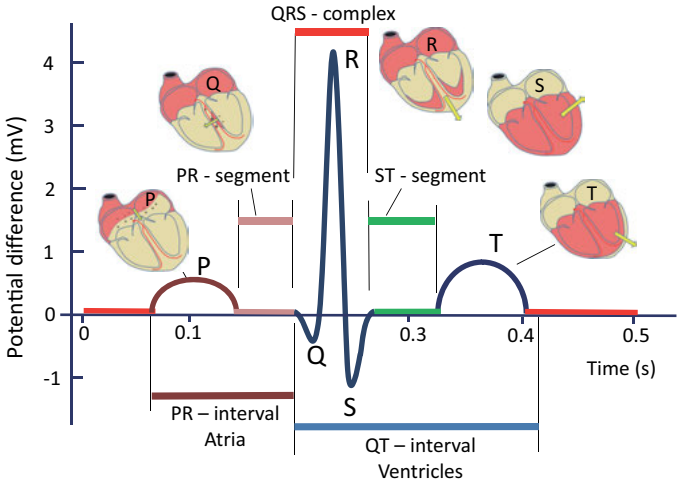


Fig. 7.11: Sequence of polarization changes during a cardiac action potential as recorded in lead II. The graph also gives the standard nomenclature for time intervals and characteristic points. Red colored segment of the heart are depolarized, ocher colored segments correspond to a resting state or to repolarization.

The *QRS complex* corresponds to the largest of the three loops shown in Fig. 7.7, the other loops are connected to the P and T wave. It is interesting to note that the P wave corresponds to a depolarization of the atria, whereas the T wave, although having the same sign, is due to the repolarization of the ventricles. The repolarization of the atria is not noticeable, because it is submerged in the large and dominating QRS complex. Summarizing: the P wave reflects the depolarization of the atria, the QRS complex the depolarization of both ventricles while the atria recover, and the T wave reflects the repolarization of the ventricles.

7.5 Leads according to Goldberger and Wilson

According to Einthoven the leads are referred to as *bipolar leads*. Alternatively, *unipolar leads* can be defined according to the scheme of Emanuel Goldberger. The *Goldberger leads* define a neutral reference point and the potential difference is measured between this neutral point and one of the extremities. The neutral point is defined to lie between two equally large resistances corresponding to the center of the Einthoven triangle. The leads indicated in Fig. 7.12 are labeled aVL, aVR, and aVF for augmented voltage left, augmented voltage right, and augmented voltage foot, respectively. Originally these leads were measured without resistances defining the central point, but the signals were too small. The subsequent use of high ohmic resistances augmented the signal, therefore the name ‘augmented voltage’. The Goldberger leads are equiva-

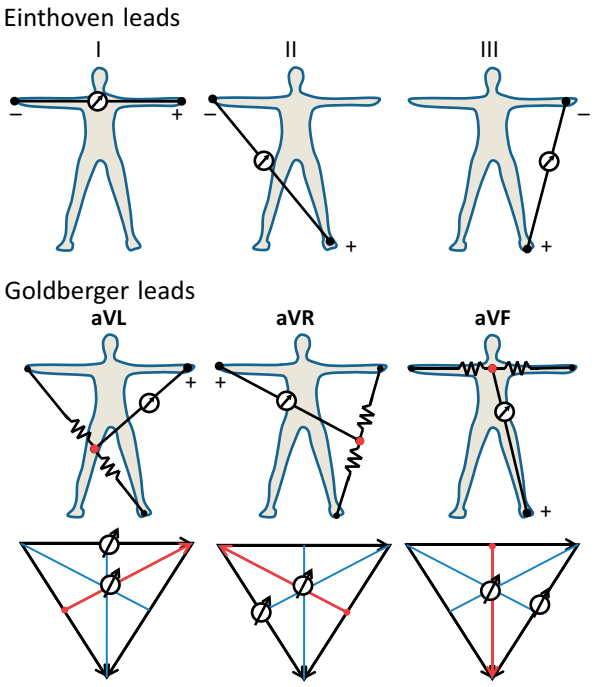


Fig. 7.12: Comparison of leads according to Einthoven and to Goldberger. In the leads according to Goldberger a neutral point is defined by two high and equal ohmic resistances between extremities. In the bottom panels the equivalence of both lead systems becomes obvious.

lent to the Einthoven leads I, II, and III. Although they yield redundant information, Goldberger leads allow more precise measurements of potential changes at the extremities than possible with Einthoven leads and they confirm characteristic patterns in cases of heart diseases.

Combining the Einthoven projections and the Goldberger projections every thirty degrees a vector projection is achieved in the frontal plane, as shown in Fig. 7.13. The circle is known as the *Cabrera circle*. The Einthoven projections are located every 60° and the Goldberger projections lie in between. Note that with this system orthogonal projections can be compared. For instance, if one of the leads shows a large amplitude in the QRS complex, the projection in the orthogonal direction should vanish. With this system the precision of determining a vector orientation is about ±10°.

As an example we discuss the axis of the heart. In the R spike the electrical dipole moment lies parallel to the *heart axis*. The standard orientation is about 60° with respect to the horizontal line (see Fig. 7.9 for the definitions of angles). Therefore lead II should have the highest positive signal. Lead aVL is oriented at right angles to lead II and therefore should display a zero crossing on passing through the perpendicular orientation. This is indeed the case when comparing all projections displayed in Fig. 7.15.

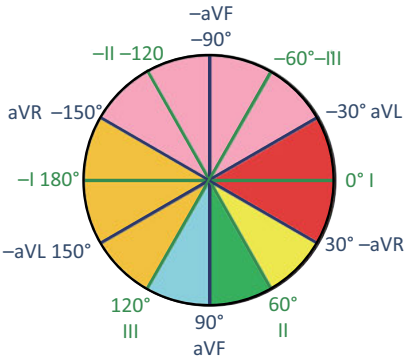


Fig. 7.13: The Cabrera circle combines Einthoven leads and Goldberger leads. The colors refer to the orientation of the heart axis. Yellow: standard indifferent, green: steep; blue: right; red: left. All other regions are extremes and usually not observed.

At the same time the projection onto aVR should be negative, which is actually observed. Heart axes which lie in the yellow angular region (30–60°) of Fig. 7.13 are called standard or indifferent. The green region (60–90°) refers to a "steep" heart axis, the blue region (90–120°) to a right-oriented heart axis, whereas the red colored region (30–30°) corresponds to a left-oriented heart. Extreme deviations from the standard orientation may indicate a heart disease or lung emboli, which pushes the heart to the right.

Note that all projections taken so far according to Einthoven and Goldberger yield a vector projection into the frontal plane. However, as we have seen in Fig. 7.7 the electrical sum dipole is a vector describing loops in three dimensions. In order to determine the third dimension in the horizontal plane, an additional system of leads is used according to Norman Wilson. These are leads which are placed directly on the left side of the chest on the left and right side of the heart. The exact location of the electrodes, labeled V1 to V6, is shown in Fig. 7.14.

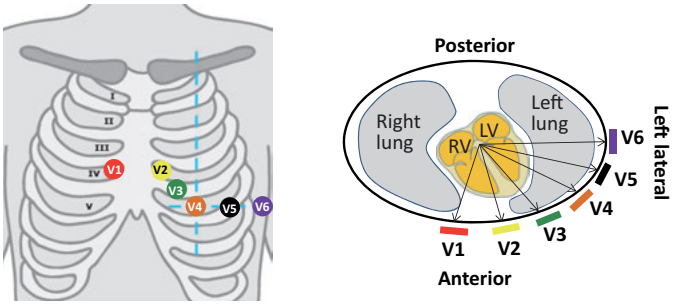


Fig. 7.14: Left panel: placement of the Wilson leads on the chest. Right panel: lead projections in the horizontal plane.

The *Wilson leads* yield information on the horizontal vector projection. The potential is positive during periods when the dipole points in the direction of the electrode; the potential is negative whenever the vector points away from the electrode.

These 12 leads (3 Einthoven, 3 Goldberger, 6 Wilson) are called *standard leads*. They are most frequently used in clinical practice. Electrocardiograms according to all 12 leads are shown in Fig. 7.15. In fact they are highly redundant. Only three independent leads are required to determine all three vector components. For ECG diagnosis in clinical practice we refer to [2].

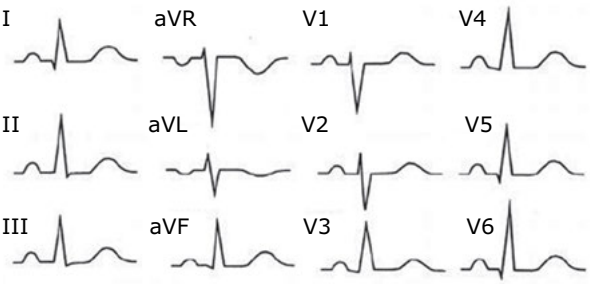


Fig. 7.15: Electrocardiograms recorded according to the leads of Einthoven (I,II,II), Goldberger (aVR, aVL, aVF), and Wilson (V1–V6).

ECG recordings can be performed at rest, during exercise, for controlling long-term heart rhythms or abnormal and eventually life-threating rhythm perturbations (*cardiac arrhythmias*), for checking heart functions during illnesses, after strokes, and after accidents. In all cases the ECG provides invaluable and indispensable information on the functioning of the heart. For instance, cardiac fibrillation is a condition of uncoordinated action potentials which leads to improper ventricular contraction. An example of ventricular fibrillation determined by lead II is shown in Fig. 7.16. Under these conditions blood pumping is completely ineffective. As no oxygen is transported, the state is lethal unless stopped by an electroshock through the heart using a defibrillator. The defibrillator provides a high dc voltage of 1000–4000 V for 3–40 ms to the chest on both sides of the heart, resetting the action potential, so that a new and fresh coordinated depolarization via the AV node or SA node can be started. The total energy supplied by the defibrillator is limited by the charge of a capacitor which is usually 100 to 300 Joules. Ventricular fibrillation is the most commonly identified arrhythmia in cardiac arrest patients.

Fig. 7.16: Ventricular fibrillation measured at lead II.

7.6 Methods, procedures, and new developments

7.6.1 Electrocardiography

In earlier times ECGs were recorded with a single channel strip chart recorder connected to an electrometer. This required recording each lead one after the other rendering comparisons between different projections difficult if not impossible. More recently the single channel strip chart recorder was replaced by 12 channels in groups of two for all 12 leads; an example is shown in Fig. 7.17. Nowadays the electrometer is connected to a digital oscilloscope after signal amplification and analog digital conversion. For long-term ECG monitoring, patients keep the electrodes attached for a day or so, and the data are recorded in a small electronic storage device for later readout. Alternatively, the data can be transmitted instantaneously to the clinic via the internet, known as real-time telemedical (home-)care.

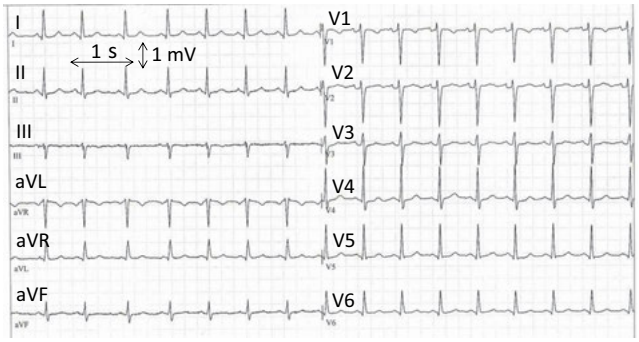


Fig. 7.17: ECG with 12 leads documented with a strip chart recorder.

Traditionally electrodes for measuring ECGs are attached to the skin using conducting Ag/AgCl electrolytic gels that reduce the ohmic resistance in the air gap between skin and electrode. A typical setup is shown in Fig. 7.18 including ergometer, recording unit, and attachment electrodes to the skin of a subject. In principle, electrical potentials can also be recorded by capacitive coupling.

Capacitive sensors are well known in various industrial environments, but for measuring ECGs with very small millivolt potential changes their sensitivity was not yet sufficient. New capacitive sensors using high dielectric constant materials (high k materials) have remedied this problem and provide the necessary sensitivity. The capacitive sensors are dry, do not need any adhesives or electrolytes that could cause allergic reactions on the skin, and they are easy to apply. Therefore, in the long run capacitive sensors will likely supersede ohmic contacts. They are now already being used on wrist watches to monitor one channel ECGs usually applying the Einthoven lead I. One electrode on the back side of the watch is in permanent contact with the

Fig. 7.18: Typical setup for ECG recording with ergometer, recording unit, and placement of electrodes on the chest.

skin. In order to record the ECG, the second front facing electrode has to be touched with a finger from the opposite hand. An Electric Potential Integrated Circuit (EPIC) sensor records the ECG signal and transmits it to the display of the wristwatch or to any other mobile device like a smartphone. With the same sensor technology and using a handheld two EPIC thumb pad a device is now becoming available to read out the cardiac signal from lead I. An example of such a handheld device is shown in Fig. 7.19 [3].

Fig. 7.19: Capacitive handheld device for recording cardiac signals. This device utilizes EPIC™ sensing technology (reproduced from www.plesseysemiconductors.com/products/impulse/, image provided courtesy of Plessey).

7.6.2 Magnetocardiography

The ion currents in the heart that lead to the myocardial contraction also produce magnetic fields that are albeit much weaker (≈ 100 pT) than electric fields. Nevertheless, with magnetic field sensors of high sensitivity these weak magnetic fields can be detected and a magnetocardiogram (MCG) can be recorded similar to the ECG. Two types of sensors are presently being tested and are in use: superconducting interference devices (SQUID) based on high-temperature superconductors that require liquid nitrogen instead of helium for cooling [4], and laser optical pumped Cs–cells [5]. The advantage of MCG is a contact-free recording of cardiac signals. Furthermore, a magnetic field mapping of cardiac activities can be performed over the entire chest with an array of sensors. Therefore MCG can be carried over from detecting local signals to imaging. In principle this should also be possible with ECG and in particular with the new EPIC sensor technique. However, the spatial resolution would be lower because of the wider range of electrical fields compared to magnetic fields as already discussed in Chapter 6 with respect to MEG versus EEG. The disadvantage of MCG is the weak signal, which requires averaging over many cycles in order to get a decent signal-to-noise ratio (SNR). This works fine for a healthy person, but in cases of cardiac arrhythmias high time resolution is required but lacking in MCG. Nevertheless, this new technique is now available in a few clinics.

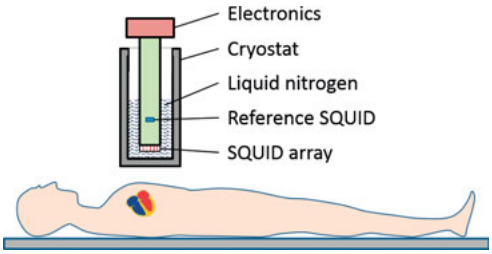


Fig. 7.20: Magnetocardiography uses SQUID sensors for recording weak magnetic fields from cardiac activity.

One promising application of MCG is the recording of fetal cardiac function [6]. ECG has been used in the past but with mixed success. This is due to a low SNR that results from interfering maternal ECG and from the small size of the fetal heart. On the other hand, fetal MCG, although very weak, has been demonstrated to yield results on a reasonable SNR level that allows analyzing fetal cardiac arrhythmia before birth for potential treatment.

7.6.3 Artificial pacemaker

An artificial pacemaker is a small electronic device that helps control the heart rate when the natural heartbeat is too slow or irregular (arrhythmias). Artificial pacemakers consist of three main parts: battery, pulse generator, one or more leads with electrodes on each lead. A flat thin metal box houses battery and electronics, which is implanted under the skin just below the collarbone through an incision in the chest. The leads are funneled through the vein leading into the right ventricle. The electrodes have a double purpose: they record the ECG from the heart and signal the pulses back to the electronics. The electronics compares the actual heart rate with standard settings. The pacemaker may also monitor temperature, breathing rate, and other factors and can adjust the heart rate to changes in activity. If the rate is too low or arrhythmic, the leads will stimulate the heart muscle by electric discharges. An Li-ion battery lasts for about 8–10 years and when it needs to be replaced, the metal case has to be opened by minor surgery. Modern pacemakers have bluetooth connections to the outside so that the heart rate can be monitored by the patient or a doctor via the internet (telemedicine). Depending on the design, pacemakers may also contain a defibrillator, which delivers stronger pulses than a standard pacemaker. Developments for recharging batteries via microwave radiation have not been successful. More promising are efforts to increase the battery's lifetime.

Recent developments have the potential to revolutionize the application of artificial pacemakers [7]. A tiny pen-shaped wireless pacemaker has been developed which contains all parts, including battery, electronics, sensors, and pulse generator. The dimensions of 6 mm wide and 42 mm long are such that it can be inserted directly into the heart without surgery. Insertion is achieved by a steerable catheter delivery sys-

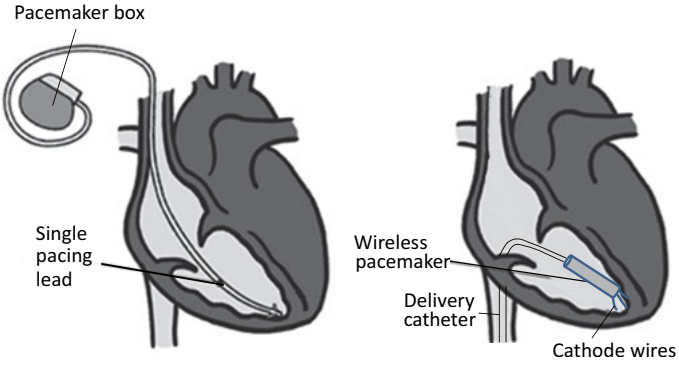


Fig. 7.21: Comparison of single lead pacemaker and wireless pacemaker. The pacemaker box contains battery, electronics, and pulse generator. A single lead in the right ventricle is shown, but up to three leads are used depending on design. The wireless pacemaker consists of one metal tube that contains all parts in a miniaturized fashion inserted into the lower part of the right ventricle. Cathode wires are anchored [7].

tem threaded from a leg vein through the inferior vena cava to the right ventricle of the heart, where it is anchored in the lower part of the right ventricle. After fixation and electrical approval, the catheter is removed. As no wires are inserted considerably fewer complications are expected such as infections at the veins where the wires are inserted and vein injuries. The battery with a lifetime of about 12 years can be replaced by removing the pacemaker with the help of the same delivery catheter. Both types of pacemaker, traditional and wireless, are shown schematically in Fig. 7.21. In either case, all leads and parts that are in contact with blood and tissue must be made biocompatible in order to avoid blood clotting. Biocompatibility is discussed in Section 14.2.8/Vol. 2.

7.7 Summary

1. The cardiac action potential is much longer than the action potential of normal cells.
2. The cardiac action potential is characterized by a very fast depolarization, fast but partial repolarization, followed by an extended plateau region before complete repolarization takes place.
3. The myocardium does not show tetanic contraction.
4. The action potential is self-excitatory and starts at the sinoatrial node.
5. The atrioventricular node acts as low pass filter.
6. The sum vector of electric dipoles describes three loops during one cardiac cycle, characterized by a P wave, PR interval, Q spike, QRS complex, S spike, ST interval, and T wave.
7. The P wave reflects the depolarization of the atria, the QRS complex the depolarization of both ventricles while the atria recover, and the T wave reflects the repolarization of the ventricles.
8. The Einthoven triangle describes projections of the temporal electric dipole onto three directions, called leads.
9. The potential differences according to the three projections reflect the depolarization and repolarization of the heart during one cardiac cycle.
10. Goldberger leads are taken between the center and the extremities.
11. The Cabrera circle combines Einthoven leads and Goldberger leads.
12. Combining Einthoven and Goldberger leads allows a determination of the heart axis within ±10°
13. Wilson leads are placed on the chest and add vector information in the third dimension.
14. Standard ECG recordings use 12 leads: 3 according to Einthoven, 3 according to Goldberger, and 6 according to Wilson.
15. New capacitive sensors allow ECG recording by simply touching contacts with fingers from both hands.
16. Magnetocardiography shows promising applications for recording fetal heart activity.
17. Artificial pacemakers record the heart rate and stimulate ventricular contraction if the heart rate drops below normal.
18. Miniaturized pacemakers are inserted directly into the right ventricle.

References

[1] Zabel M, Acar B, Klingenheben T, Franz MR, Hohnloser SH, Malik M. Analysis of 12-lead T-wave morphology for risk stratification after myocardial infarction. Circulation. 2000; 102: 1252–1257.
[2] Vecht R, Gatzoulis MA, Peters NS. ECG diagnosis in clinical practice. 2nd edition. Springer-Verlag; 2009.
[3] Plessey Semiconductors: www.plesseysemiconductors.com/
[4] Zhang Y, Wolters N, Lomparski D, Zander W, Banzet M, Schubert J, Krause HJ, van Leeuwen P. Multi-channel HTS rf SQUID gradiometer system recording fetal and adult magnetocardiograms. IEEE Trans Appl Supercond. 2005; 15: 631–634.
[5] Bison G, Castagna N, Hofer A, Knowles P, Schenker JL, Kasprzak M, Saudan H, Weis A. A room temperature 19-channel magnetic field mapping device for cardiac signals. Applied Physics Letters. 2009; 95: 173701.
[6] Sameni R, Clifford GD. A review of fetal ECG signal processing: Issues and promising directions. Open Pacing Electrophysiol Ther J. 2010; 3: 4–20.
[7] Reynolds D, et al. A leadless intracardiac transcatheter pacing system. N Engl J Med. 2016; 374: 533–541.

Further reading

Crawford MH, DiMarco JP, Paulus WJ. Cardiology. 3rd edition. Elsevier Saunders; 2010.
Seeley R, Vanputte C, Russo A. Seeley's anatomy and physiology. McGraw Hill Book Co.; 2016.
Guyton AC, Hall JE. Textbook of medical physiology. 11th edition. Elsevier Saunders; 2006.
Pape HC, Kurtz A, Silbernagel S, editors. Physiologie. 7th edition. Stuttgart, New York: Thieme Verlag; 2014.
Malmivuo J, Plonsey R. Bioelectromagnetism: Principles and applications of bioelectric and biomagnetic fields. Oxford University Press; 1995. Available at: www.bem.fi/book/
Martini FH, Nath J, Bartholomew EF. Essentials of anatomy and physiology. 10th edition. Pearson; 2015.

8 The circulatory system

8.1 Introduction and overview

Circulation implies the flow of a medium in a closed system of tubes. In a physiological sense circulation is the blood flow in the body powered by the heart as a pump. This chapter is based on information provided in Chapter 7. Therefore it is advisable to first read the previous chapter or to gain some knowledge of the cardiac action potential before proceeding with this chapter.

Circulation has the task of delivering oxygen and nutrition to all parts of the body and of disposing of metabolic byproducts, such as carbon dioxide, urea, water, and heat. There are further tasks of blood flow connected with defending against diseases via leucocytes, transporting hormones, and supplying thrombocytes to injured blood vessels. Oxygen is taken up during inspiration in the lungs and carbon dioxide is disposed of during expiration. How this is achieved is the topic of Chapter 10 on the respiratory system. Oxygen molecules (O_2) bind to the hem complex in hemoglobin proteins situated in erythrocytes. These small disk-shaped oxygen carriers float in the blood stream to even the remotest parts of the body and through vessels ranging from wide to extremely narrow.

Figure 8.1 schematically shows the circulatory system consisting of two pumps. Oxygen arriving from the lungs is pumped to the capillary system and delivered to organs and muscles, while carbon dioxide is resorbed from the tissue and pumped back to the lungs for expiration. The left ventricle is the pump for oxygenated blood, and the right ventricle is the pump for deoxygenated and carbon dioxide rich blood. In the following we will describe the circulatory system in more detail from a physical point of view. For physiological aspects literature is provided at the end of this chapter.

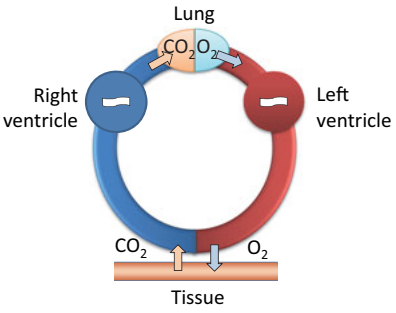


Fig. 8.1: Schematic of the circulatory system. O_2 is taken up in the lungs and pumped by the left ventricle to the organs and muscles, whereas metabolic CO_2 is pumped back to the lung for expiration.

We may also look at the circulatory system in a different way as consisting of two circulations: a lung circuit or *pulmonary circuit* and a body circuit or *systemic circuit*, both circuits displayed in Fig. 8.2 require the action of both ventricles:

DOI 10.1515/9783110372830-010

1. The pulmonary circuit pumps deoxygenated blood from the right ventricle to the lung and returns oxygenated blood back to the left atrium.
2. The systemic circuit pumps oxygenated blood through the left ventricle to the body and returns deoxygenated blood back to the right atrium.

Sometimes these circuits are also referred to as *small circuit* and *large circuit*. Both circuits run synchronously and in phase and they pump the exact same blood volume of about 70 ml per systolic ejection called *stroke volume* (SV).

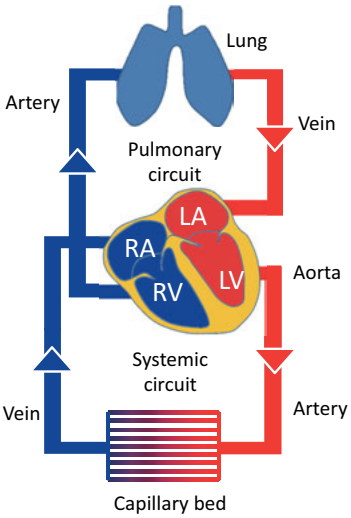


Fig. 8.2: Circulatory system consisting of two circuits, the pulmonary circuit and the systemic circuit. RA = right atrium, LA = left atrium, RV = right ventricle, LV = left ventricle.

There are three types of blood vessels in the body, as indicated in Fig. 8.3:

1. *Arteries* and *arterioles* carry blood away from the heart;
2. *Veins* and *venules* carry blood to the heart;
3. *Capillaries* allow diffusion of oxygen and nutrients from the blood to the tissue and diffusion of waste products to the blood. The *capillary bed* is the counterpart to the lung. Both serve the purpose of exchange into and out of the blood.

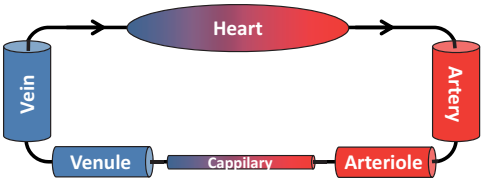


Fig. 8.3: Blood vessels in the circulatory system.

There is still another way to categorize the circulatory system according to blood pressure. Arteries carry high pressure blood whereas veins support the low pressure blood system. These differences are also expressed in the structure of the blood vessels as we will discuss later.

8.2 The heart as a pump

The heart consists of two chambers and four valves, two for each chamber (see Figs. 8.4 and 1.3). The two *atrioventricular valves* (AVV) are filling valves: *tricuspid valve* and *mitral valve*. The *tricuspid valve* with three flaps fills blood from the body into the right atrium, and the *mitral valve* with two flaps fills blood from the lung into the left atrium. The other two *ventricular valves* are ejection valves also called *semilunar valves: pulmonary valve* and *aortic valve*. The pulmonary valve in the right ventricle opens for ejecting blood into the lung, and the aortic valve in the left ventricle opens for ejecting blood into the body.

The cardiac pumping cycle can be divided into four distinct phases. They are shown schematically in Fig. 8.5 (1–4) for the left atrium and ventricle. The red color indicates oxygenated blood flowing through the left ventricle. The right ventricle operates in exactly the same way for deoxygenated blood. In Fig. 8.6 the corresponding pressures in the atrium and in the ventricle are plotted, the cardiac volume and the

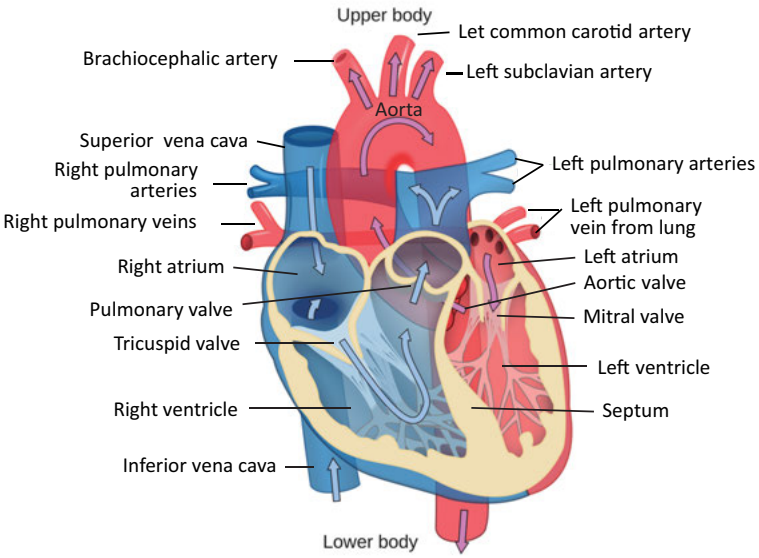


Fig. 8.4: Valves in the heart and blood flow directions (adapted from Wikimedia, © Creative Commons).

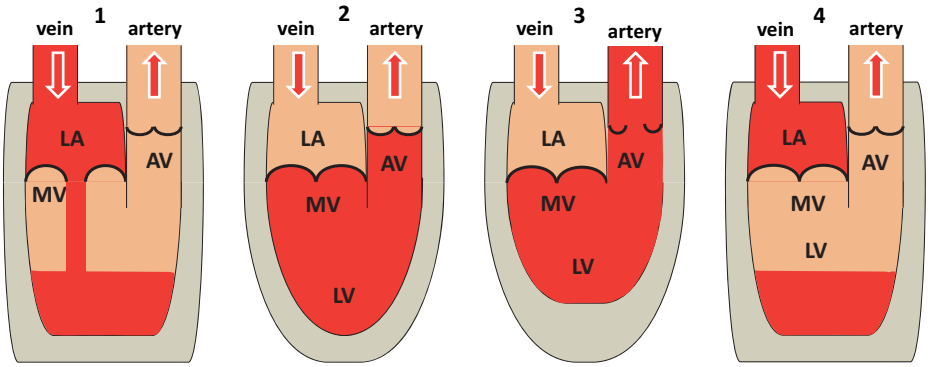


Fig. 8.5: Four phases of the left ventricle acting as a pump: (1) filling phase; (2) contraction phase; (3) ejection phase; (4) relaxation phase. LA = left atrium, LV = left ventricle, AV = aortic valve, MV = mitral valve.

relation to the ECG. Also indicated at the bottom are the opening and closing phases of the valves and the time span for diastole and systole.

The four phases of the pumping cycle are characterized as follows:

1. *Inflow phase*: The mitral valve is open, the semilunar aorta valve is closed. This phase is initiated by the P wave of the ECG (see Chapter 7), which leads to a contraction of the atrium. As the atrium contracts, pressure increases and more blood flows through the open mitral valve, leading to a rapid filling of the ventricles.
2. *Isovolumetric contraction phase*: all valves are closed. This phase begins with the QRS complex of the ECG, which represents ventricular depolarization. The depolarization triggers excitation contraction of the ventricle and a rapid increase in intraventricular pressure. When the atria-valves close, a heart sound can be recognized. As the tricuspid valve in the right ventricle closes slightly before the mitral valve in the left ventricle by about 40 ms, actually a double sound can be picked up with the help of a stethoscope.
3. *Outflow phase*: The semilunar valves (aortic and pulmonic) are open, the AV valves remain closed. This phase represents rapid ejection of blood into the arteries (aorta and pulmonary arteries). Ejection begins as soon as the intraventricular pressures exceed the pressures within the artery, causing the aortic and pulmonic valves to open. This is the case immediately after the S wave of the ECG. Highest outflow velocity is reached early in the ejection phase, and maximum systolic pressure is achieved.
4. *Isovolumetric relaxation phase*: All valves are closed and the pressure decreases. Late in the ejection phase the blood flow drops back to low values until it may reverse. At this point the semilunar valves close, defining the onset of the diastole at the end of the T wave.

These four phases are commonly separated into two sequences: systole, including phase 2 and 3, and diastole, comprising phase 1 and 4. The four phases can also be identified with the ECG, where phase 1 corresponds to the P wave, phase 2 and 3 to the QRS complex, phase 4 to the T wave and to the subsequent resting phase. In a normal heart cycle at rest with a rate of 75 cycles per minute, corresponding to a cycle duration of 800 ms, systole occupies about 300 ms and diastole about 500 ms. With increasing heart rate diastole shortens, whereas systole remains roughly constant in time.

The pressure dependence and the left ventricular volume are plotted in Fig. 8.6 in relation to the electrocardiogram. The filling phase is the early low pressure phase during diastole with a blood pressure in the order of 3 hPa. In the succeeding contraction phase the pressure steeply rises to a maximum pressure of about 140 hPa immediately before ejection. The ventricular blood volume after filling and just before ejection, called the *end-diastolic volume* (EDV) is about 120 ml. After ejection the volume decreases to the *end-systolic volume* (ESV) of about 50 ml. This is the rest volume remaining in the ventricle. The *stroke volume* (SV) is defined as the difference between EDV and ESV. Therefore the stroke volume is typically:

$$V_{\mathrm{SV}} = V_{\mathrm{EDV}} - V_{\mathrm{ESV}} = 120\,\mathrm{ml} - 50\,\mathrm{ml} = 70\,\mathrm{ml}.$$

The *ejection fraction* (EF) is defined as:

$$\mathrm{EF} = \frac{V_{\mathrm{SV}}}{V_{\mathrm{EDV}}} = \frac{70\,\mathrm{ml}}{120\,\mathrm{ml}} \approx 0.6.$$

An EF of 0.6 is typical for a healthy person. The stroke volume and the ejection fraction are independent of the heart frequency up to about 180 beats per minute. Beyond this the SV and EF decrease rapidly, because there is not sufficient diastolic time for refilling the ventricles.

Now we can also determine the *cardiac output*, which is defined as the product of heart rate and stroke volume. Assuming 75 beats per minute at rest, the heart rate is 1.25 Hz. Then we find for the cardiac output (I_{CO}), defined as volume rate, i.e., volume per time:

$$I_{\mathrm{CO}} = V_{\mathrm{SV}} \times f_{\mathrm{heart}} = 70\,\mathrm{ml} \times 1.25\,\mathrm{s}^{-1} = 88\,\mathrm{ml\,s}^{-1}.$$

The I_{CO} amounts to 5.3 l/min and 7600 l/day for the left ventricle, and 15 200 l/day for both ventricles at rest.

The heart has a built-in mechanism that allows adapting to fluctuations in the blood volume arriving at the venous side. Whenever blood volume increases, cardiac output will increase at the same heart frequency. This feedback system is known as the *Frank–Starling law* of the heart. Increased blood volume stretches the cardiac muscle, and during systole the myocardium is required to contract more strongly in order to eject the extra volume. Important is that the left and right ventricles always pump exactly the same amount at the same frequency and in phase.

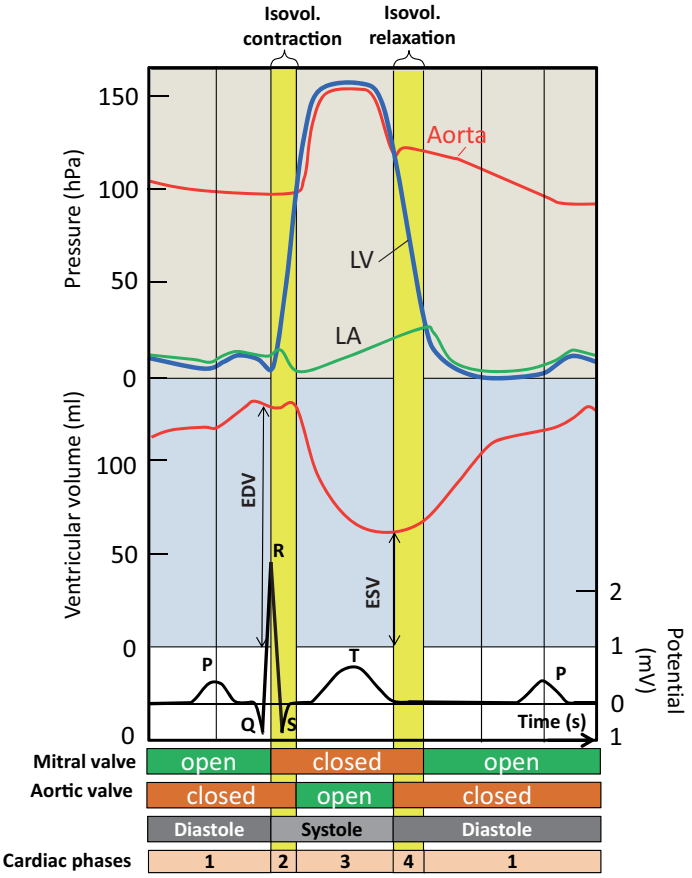


Fig. 8.6: Pressure changes during a cardiac cycle in the left ventricle, aorta, and left atrium. Also shown are the ventricular volume and the electrocardiogram. The time axis spans about 1 s. EDV = end-diastolic volume, ESV = end-systolic volume, LA = left atrium, LV = left ventricle. PQRST have their usual meaning.

8.3 Energy, power, and efficiency of the heart

Next we consider how much energy the heart requires in order to keep it running. The total energy of the heart is composed of volume or potential energy, ejection or kinetic energy, and tension or thermal energy:

$$E_{\text{tot}} = E_{\text{pot}} + E_{\text{kin}} + E_{\text{tension}}.$$

The potential energy of the heart can be determined considering a pressure–volume phase diagram using the information from the volume versus time and pressure versus time diagrams presented in Fig. 8.6. In Fig. 8.7 the ventricular pressure is plotted versus volume for the left ventricle. The cycle nature of the heart becomes directly visible

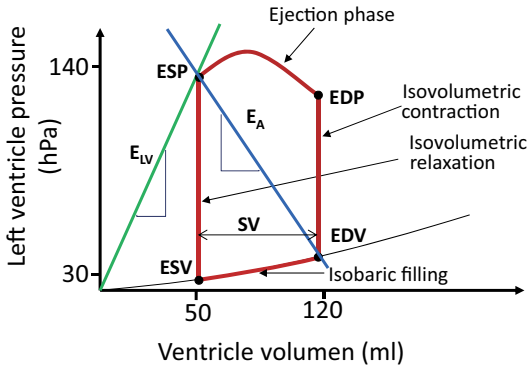


Fig. 8.7: pV-diagram of the cardiac cycle. ESV = end-systolic volume; EDV = end-diastolic volume; ESP = end-systolic pressure; EDP = end-diastolic pressure; SV = stroke volume; E_A = arterial elastance; E_{LV} = end-systolic elastance.

in this pV-diagram reminding us of the pV-diagram of a Stirling engine. The first phase from ESV to EDV is the filling phase, which is almost isobaric with only a slight pressure increase. Then follows the isovolumetric (isochoric) contraction from EDV to the *end-diastolic pressure* (EDP), where the tension of the ventricular muscle increases. During the ejection phase from EDP to the end-systolic pressure (ESP) the pressure first slightly increases, but then relaxes when approaching the relaxation phase from ESP to ESV, which again is isovolumetric.

The mechanical volume work (potential energy) performed by the left ventricle equals the enclosed area, which can be estimated as follows:

$$E_{\text{pot}} = \int_{\text{ESP}}^{\text{EDP}} p_{\text{ejection}}\,\mathrm{d}V - \int_{\text{ESV}}^{\text{EDV}} p_{\text{filling}}\,\mathrm{d}V$$
$$\approx \Delta p \int_{\text{ESV}}^{\text{EDV}} \mathrm{d}V = \Delta p \times V_{\text{SV}}.$$

According to Figs. 8.6 and 8.7 the ejection pressure is about 140 hPa and the pressure during filling is about 3 hPa. Therefore we obtain for the potential energy of the left ventricle:

$$E_{\text{pot}} = \Delta p_{\text{LV}} \times V_{\text{SV}} = 137\,\text{hPa} \times 70\,\text{ml} = 0.96\,\text{J}.$$

The kinetic energy is the ejection energy of the blood into the aorta with a flow velocity of about 1 m/s. Assuming a density of 1 g/cm^3 for the blood, we obtain:

$$E_{\text{kin}} = \frac{1}{2} m_{\text{SV}} v^2 = 0.5 \times 0.07\,\text{kg} \times 1\,\text{m/s} = 0.035\,\text{J}.$$

The total mechanical energy $E_{\text{mech}} = E_{\text{pot}} + E_{\text{kin}}$ is therefore about 1.0 J for one stroke. With a heart rate of 1.25 Hz, we find for a person at rest a mechanical power consumption $P_{\text{mech}} = E_{\text{mech}} \times f_{\text{heart}}$ in the order of 1.25 Watt.

The right ventricle operates at a much lower pressure difference Δp_{RV} of only about 20 hPa. Accordingly, the potential energy and the mechanical power are only 1/7 of those of the left ventricle. The power of both chambers together is roughly $P_{mech} = 1.4$ Watt. This difference is directly expressed in the wall thicknesses of the left and right ventricles as shown in a cross section in Fig. 8.8.

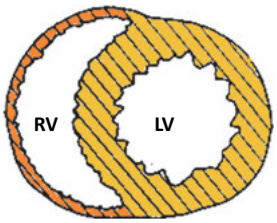


Fig. 8.8: Cross section through the heart showing the myocardial wall thicknesses of the left and right ventricle. The left ventricle (LV) works at a much higher pressure level and therefore the tension of the wall needs to be much larger than for the lower pressure right ventricle (RV).

During exercise, f_{heart} increases while V_{SV} remains constant. Therefore also the power consumption will go up, but never much beyond 2.5 Watt.

It is a bit more difficult to estimate the tension energy. During isometric tension the muscle length stays constant. But to keep the tension, energy is required; otherwise the muscle would stretch and relax (see Fig. 2.16). The aerobic energy required is supplied by ATP and oxygen and converted into heat. Therefore this thermal heat is also referred to as tension heat. It is proportional to the tension of the myocardial muscles T_{muscle} and the duration of the stress Δt:

$$E_{tension} = kT\,\Delta t,$$

where k is a conversion factor to SI units.

We know from Chapter 4 that a person at rest has a basal metabolic rate of 80 Watt, 7 % of this goes to the heart, which makes $P_{heart} = 5.6$ Watt. As $P_{heart} = P_{mech} + P_{heat}$, we find for $P_{heat} = 4.2$ Watt. This is 75 % of the total energy requirement, and only 25 % is needed for the mechanical work. The *efficiency* of the heart, defined as:

$$\eta = \frac{P_{mech}}{P_{heart}},$$

is thus 25 %. We have already compared the heart with a Stirling engine, the latter has an efficiency of up to 50 %. In comparison, the heart is still reasonably efficient and as we have already seen, the cardiac output of the heart is amazingly high.

The slope of the blue line $E_A = ESP/SV$ in Fig. 8.7 characterizes the *arterial elastance*, which is a measure of the total compliance of the heart and impedance of the vascular system. The slope of the green line $E_{LV} = ESP/ESV$ is the end-systolic elastance. The ratio E_A/E_{LV} is inversely proportional to the ejection fraction EF: $E_A/E_{LV} = (EF)^{-1}$. Both slopes and their ratio are indicators for the functionality of the heart and the arterial system and deviations from normal behavior are early signs for potential heart failure.

All numbers derived so far are average numbers not taking into account the gender or the size of a person. However, this can be considered by normalizing the stroke volume by the *body surface area* (BSA), which is often done in cardiovascular physiology.

For estimating the energetics and efficiency of the myocardium we have mainly used information on the stroke volume and left ventricular pressure difference. In the past this information could only be gained invasively with the use of catheters inserted into the heart. However, with modern imaging methods, in particular Doppler sonography (Chapter 13) and MRI (Chapter 15), myocardial efficiency can be evaluated by noninvasive methods. An overview of present day methods for determining cardiac efficiencies is given in [1].

8.4 Fluid statics of the circulatory system

Neglecting the pulsing nature of the cardiac output, we consider first the time averaged hydrostatic properties of the heart and the circulatory system, concentrating on the systemic circuit. Without cardiac activity, the pressure in the blood vessels is the same everywhere. This pressure is called the *mean systemic filling pressure* (MSFP), which is approximately 10 hPa. This pressure depends on the filling capacity of all the blood vessels including the heart and the compliance of the blood vessels. As soon as the heart is turned on, the pressure on the artery side increases from 10 hPa to 140 hPa and lowers the pressure on the venous side from 10 hPa to about 3 hPa. Thus the pressure difference is about 137 hPa, as already quoted in the previous sections and schematically shown in Fig. 8.9. High and low pressures are symbolized by pitot tubes. The high pressure drops off continuously with increasing distance from the aorta but mainly in the capillary bed to a low level on the venous side.

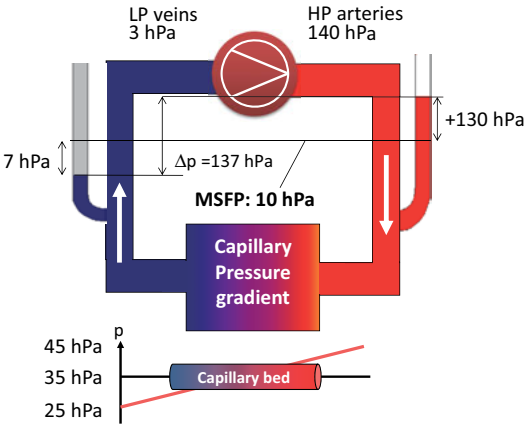


Fig. 8.9: The heart as a mechanical pump, maintaining a pressure difference. MSFP = mean systemic filling pressure is the remaining pressure when the heart stands still. LP = low pressure, HP = high pressure.

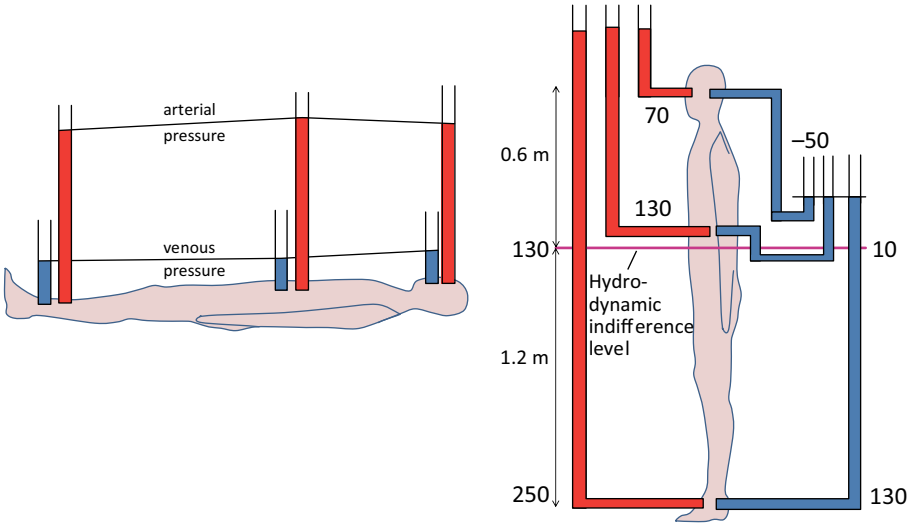


Fig. 8.10: Hydrostatic pressure of the blood system when lying horizontally and standing upright. All numbers in the right panel are in units of hPa, but the heights are in meters.

In fact, we have to distinguish between a pressure distribution of a person in horizontal and in vertical position, as shown in Fig. 8.10. In horizontal position the aortic pressure is high and constant across the body. The blood pressure drops off across the capillary bed, which represents the main flow resistance. On the venous side the pressure is also constant on a much lower level.

When standing up the circulatory system has to adapt to the superimposed pressure gradient due to gravitation: $p(h) = \rho g h$. Here ρ is the blood density, which is essentially the same as water, g is the gravitational acceleration, and h is the height of blood vessels above ground. $p(h)$ is called the gravitational pressure or hydrostatic pressure. At a certain height the pressure in the blood system is independent of the position. This height is known as the *hydrodynamic indifference level*. Its position is below the heart. Let us assume that it is at about 120 cm above ground. Then all pressures below this line follow from $p_{\text{indiff}} + \rho g h$, all pressures above are given by $p_{\text{indiff}} - \rho g h$.

When standing up, the 1.2 m high blood tube from floor to the hydrodynamic indifference level causes an additional pressure of 120 hPa. Therefore the arterial pressure of 130 hPa for horizontal posture goes up to 250 hPa at the feet and drops to 70 hPa in the head.

Similar for the venous side: the pressure of 10 hPa for the horizontal position increases to 130 at the feet and drops to −50 hPa in the head in vertical posture. Thus the arterial pressure in the head drops by roughly half in upright position, whereas the pressure in the feet doubles. Overall the pressure difference Δp between arterial and venous blood vessels remains constant at a level of 120 hPa.

The structure of blood vessels reflects this pressure difference. The veins have large inner diameters and thin walls, whereas the arteries have small inner diameters and thick walls. Cross sections of both vessel types are shown in Fig. 8.11 at the level just above the capillary bed. Because of the larger diameter most of the blood volume resides in the venous system (75 %) which can be considered as a blood reservoir.

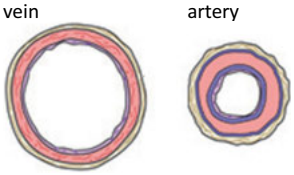


Fig. 8.11: Cross sections of blood vessels. Note the different inner diameters and wall thicknesses for veins and arteries.

When considering blood vessels and their response to pressure changes we need to regard them as elastic tubes. The diameter depends on the pressure difference between the internal pressure p_i and the external pressure p_e. The pressure difference across the vessel wall is called *transmural pressure difference* $\Delta p_\mathrm{tm} = p_\mathrm{i} - p_\mathrm{e}$. As we increase the transmural pressure, the radius and the tension T in the walls increases, similar to inflating a balloon. The wall tension is equivalent to the surface tension and depends on the wall thickness h and the radius r of the tube:

$$T_\mathrm{tm} = \frac{r}{h}\Delta p_\mathrm{tm}.$$

For low pressure differences the relationship between T and Δp_tm is linear. This is the elastic reversible region. If the tension increases beyond a critical value, rupture of the vessel wall may occur, as indicated in Fig. 8.12.

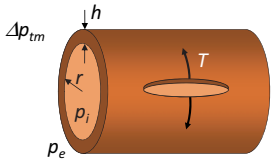


Fig. 8.12: Blood vessel as an elastic tube with radius r and wall tension T. Δp is the transmural pressure difference between inside and outside pressures p_i and p_e, respectively.

Although veins and arteries both react elastically on pressure changes, their elastic modulus is quite different. This can be best characterized by considering the elastic compliance. According to Chapter 3, volume change and pressure difference are related as:

$$\frac{\Delta V}{V} = \frac{\Delta p_\mathrm{tm}}{E},$$

where E is the elastic modulus. Rearranging, we find:

$$\frac{\Delta V}{\Delta p_\mathrm{tm}} = \frac{V}{E} = C.$$

C is called the *elastic compliance*, here specifically of the blood vessels. C is basically the inverse of the elastic modulus. The higher the compliance, the more responsive the elastic medium is. Applied to blood vessels we find that veins and arteries show very different compliances, as can be recognized from Fig. 8.13. Veins have a very high compliance, but beyond 15 hPa volume expansion stops or rupture would follow. This implies that veins have a low yield point. In contrast, the compliance of arteries is much lower and pressures can rise up to 250 hPa without rupture, i.e., the yield point is much higher than for veins.

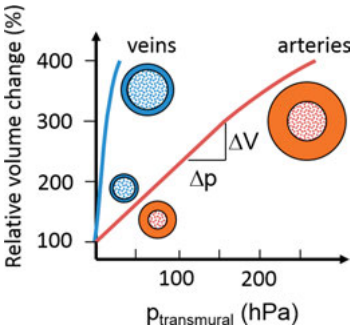


Fig. 8.13: Volume-pressure relationship for blood vessels. Blue: veins, red: arteries. The slope $\Delta V/\Delta p$ marks the compliance of the vessels. Note that the plot is rotated by 90° compared to standard plots of stress-strain relationships, for instance in Figs. 3.1 or 3.3.

8.5 Hemodynamics of the circulatory system

8.5.1 Basic equations and assumptions

Fluid dynamics of the circulatory systems is called *hemodynamics* [2]. This is because of the special characteristics that blood circulation has as a non-Newtonian type fluid in a circuit of expandable vessels. Nevertheless, we start with classical fluid dynamics (hydrodynamics), which is governed by four fundamental laws:

1. *Continuity equation*, expressing the conservation of mass of a fluid in a closed system:

$$\rho\vec{\nabla}\cdot\vec{v}+\dot{\rho}=0,$$

from which follows the continuity equation in the familiar form:

$$I_{\mathrm{V}}=\frac{\mathrm{d}V}{\mathrm{d}t}=\dot{V}=Av=\text{const.},$$

where I_{V} is the volume flow or volume rate, A is the tube cross section, and v is the flow velocity. The period above a character stands for the first time derivative.

2. *Bernoulli equation*, stating the conservation of energy for ideal fluids:

$$\Delta p=\frac{1}{2}\rho v^2+mgh,$$

where Δp is the pressure difference in the fluid maintaining flow.

3. *Hagen–Poisseuille equation*, recognizing the dissipation of energy by viscous flow, expressed by Ohm's law:
$$I_V = \frac{\Delta p}{R_{flow}},$$
where R_{flow} is the flow resistance.
4. *Kirchhoff's laws* describing the total flow resistance composed of serial and parallel resistances in a circuit.
For serial resistances
$$R'_{tot} = \sum_i R_i$$
and for parallel resistances:
$$\frac{1}{R'_{tot}} = \sum_i \frac{1}{R_i}.$$

These equations are based on a number of inherent assumptions:
(a) incompressibility of fluids;
(b) independence of viscosity on shear rate;
(c) rigid tubes with zero compliance;
(d) continuous laminar flow;
(e) closed circuit without fluid volume gain or drain.

All these assumptions are more or less violated for the flow of blood in the circulatory system. The compressibility of blood is a bit higher than that of water due to the suspension of squeezable red blood cells; the viscosity of blood depends on the shear rate; the flow is pulsatile, the blood vessels respond to changes of transmural pressure, and the blood volume varies. Nevertheless, we will first apply these hydrodynamic laws to obtain some useful insights into blood flow. Afterwards we will refine our conclusions by taking nonlinear effects into account.

The pressure of the left ventricle sends a periodic pulse wave down the blood stream through the artery. Due to this the blood flow is not continuous but pulsatile, as indicated in Fig. 8.14. If the blood flow were continuous, there would not be much of a difference between a rigid tube and an elastic rubber-like vessel. However, because of the pulsing flow in combination with compliant blood vessels, the flow resistance is no longer a constant but changes with the pulse amplitude. Altogether the flow resistance has to be replaced by a complex fluid impedance that properly describes dissipation and phase retardation between source and resistor response. The further away from the aorta, the more the pulsing blood flow is damped out and replaced by more continuous flow. Ohm's law can only be applied to the time average mean pressure, indicated by the red line in Fig. 8.14. Beyond the arterioles the pulsatile pressure oscillations have ceased and the local pressure is also the time average pressure.

The pulsatile pressure oscillation can be used for measuring systolic versus diastolic blood pressure according to the method of Riva–Rocci (Fig. 8.15): using an inflatable cuff around the upper arm or the wrist connected to a pressure manometer,

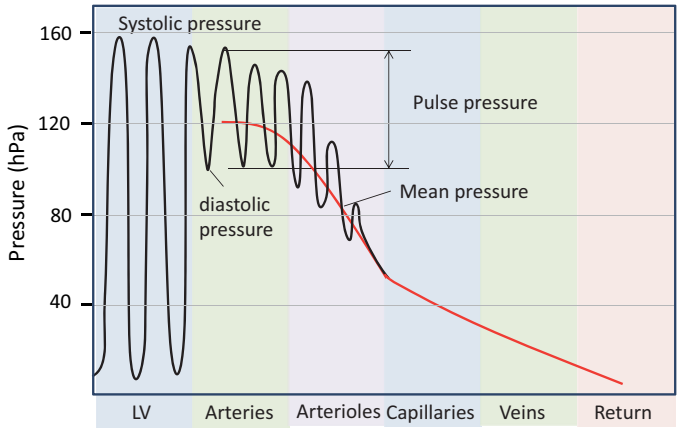


Fig. 8.14: Pressure-time relation at different distances to the left ventricle (LV).

the pressure is first increased to a level that blood flow is stopped and pulsing is no more noticeable (panel (a)). Then the pressure is released again until pulsing occurs as soon as the internal systolic pressure in the blood vessel is higher than the external pressure. This point fixes the systolic pressure and blood flow occurs during systole while the flow is still cut off during diastole (panel (b)). With a stethoscope, noise from turbulent flow behind the constriction can be discerned during systole. Further pressure release defines the point where blood flows in laminar fashion and noise is no longer noticeable, which corresponds to the diastolic pressure (panel (c)).

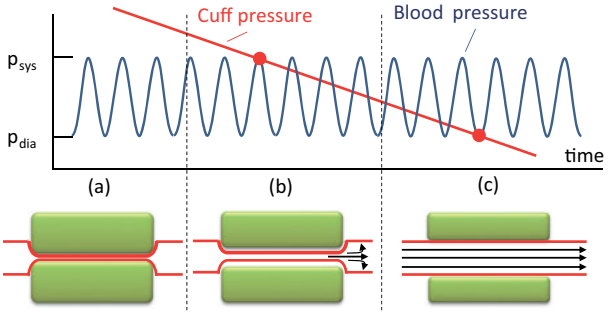


Fig. 8.15: Blood pressure measurement according to the method of Riva–Rocci. The blue line indicates the pressure amplitude oscillating between systolic and diastolic pressure. The red line is the pressure applied by a cuff around the upper arm. When releasing pressure in the cuff, first the systolic pressure level and later the diastolic pressure level is passed, indicated by red dots. The corresponding squeezing and expanding of the blood vessel caused by the cuff (in green) is shown in the lower panel.

8.5.2 Flow resistance

The flow resistance for cylindrical tubes of length ΔL and radius r or cross section $A = \pi r^2$ according to Hagen–Poisseuille is:

$$R_{\text{flow}} = \frac{8}{\pi}\eta\frac{\Delta L}{r^4} = 8\pi\eta\frac{\Delta L}{A^2},$$

where η is the viscosity of the fluid. In analogy to electric resistors, the flow resistance has a geometric dependence (length and radius of tube) and a material-specific dependence (viscosity). Since

$$I_{\text{V}} \propto \frac{1}{R_{\text{flow}}} \propto r^4 \propto A^2,$$

the volume flow increases with the fourth power of the radius or the square power of the cross section. This has severe consequences if constrictions of blood vessels occur (stenosis), which may lead to a stroke unless removed in time, for instance by insertion of a stent (see Section 15.4/Vol. 2).

Although the Hagen–Poisseuille resistance depends on the cylindrical geometry of tubes, the essential proportionalities always apply to any cross section of tubes:

$$R_{\text{flow}} = k\eta\frac{\Delta L}{r^\alpha},$$

where k is a constant and α is an exponent that depends on the geometry.

The capillary bed connects the artery with the veins and the porosity of the capillaries allows exchange of oxygen and nutrients (Fig. 8.16). These blood vessels have the smallest radius and therefore the highest flow resistance. Using the Hagen–Poisseuille equation, we can estimate the flow resistance for a short capillary tube. As an example we consider a tube of 10 millimeter length, a radius of 5 µm, and a viscosity of 4 mPa s. For such a capillary the flow resistance is about 1.6×10^{12} Pa s/l. The blood flow in the capillaries runs in parallel circuits, which reduces the overall resistance according to Kirchhoff's second law:

$$\frac{1}{R_{\text{tot}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots = \frac{N}{\langle R_i\rangle}.$$

Thus the total resistance R_{tot} decreases with the number of parallel branches N, where $\langle R_i\rangle$ is the average resistance of a single branch. On the other hand, the *total peripheral flow resistance* (TPFR), which is the sum of the resistances of all peripheral vasculature in the systemic circulation, can be estimated from the known blood pressure and the cardiac output:

$$R_{\text{tot}} = \frac{\Delta p}{I_{\text{V}}}.$$

Δp is the mean systolic pressure, which is about 120 hPa, and I_{V} is the cardiac output of about 5 l/min. Then the TPFR is:

$$R_{\text{flow}} = \frac{120\,\text{hPa}}{5\,\text{l}\,\text{min}^{-1}} = 24\frac{\text{hPa}\,\text{min}}{\text{l}} = 1.4 \times 10^5\frac{\text{Pa}}{\text{l/s}}.$$

This is much lower than the resistance of a short capillary tube. Therefore we conclude that about $N \approx 10^7$ capillaries run in parallel, which is a pretty good estimate. Cross section, local velocity, and pressure drop across the capillary bed is plotted in Fig. 8.16. We notice that the velocity decreases as the total cross section in the capillary bed increases according to the continuity equation $A \cdot v = \text{const}$. The low velocity of about $v_{\text{capillary}} = 0.005$ m/s in the capillary bed is physiologically important since it provides sufficient time for exchange of oxygen and nutrients.

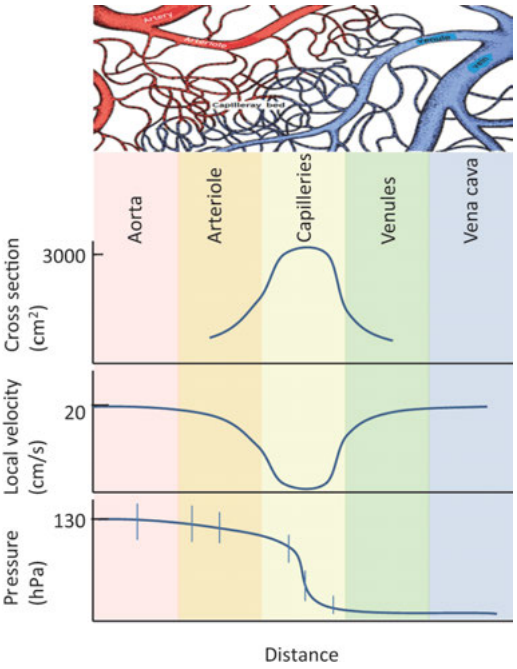


Fig. 8.16: Top panel: Branching of capillaries in the capillary bed connecting arteries and arterioles with veins and venules. Lower panels: Dependence of cross section, local velocity, and blood pressure as a function of position across the capillary bed. The blue vertical bars in the lowest panel indicate the pressure pulse between diastole and systole.

8.5.3 Turbulent flow and windkessel

Low velocity and a narrow cross section are beneficial for assuring *laminar flow*. However, in the aorta the conditions are just opposite: high ejection velocity combined with a large cross section. With the dimensionless *Reynolds number* (*Re*), defined as

$$Re = \frac{\rho v d}{\eta},$$

where $d = 2r$ is the diameter of a tube, one can estimate whether *turbulent flow* is expected. For *Re* numbers below 1000 the flow is *laminar*, but for *Re* numbers beyond 2000 turbulent flow is very likely. Turbulent flow is characterized by a disruption of flow lines and formation of curls. Turbulent flow not only is chaotic, it is also destructive and must definitely be avoided. In the case of the aorta the ejection velocity is about 1 m/s, the diameter is 20 mm, and the average viscosity of blood is $\eta = 3–4$ mPa s. Therefore we obtain a Reynolds number for blood flow after ejection into the aorta:

$$Re\ (\text{aorta}) = \frac{(10^3\ \text{kg/m}^3) \times (1\ \text{m/s}) \times (20 \times 10^{-3}\ \text{m})}{4 \times 10^{-3}\ \text{Ns/m}^2}$$
$$= 5000.$$

This number is critically high. To avoid turbulences, nature has invented a scheme that is known in engineering as *Windkessel*, meaning an air chamber or more generally an elastic reservoir. The aorta is an elastic piece of artery with extra high compliance serving the purpose of absorbing part of the kinetic energy and transforming it temporarily into potential energy. The potential energy is returned back into kinetic energy maintaining organ perfusion during diastole when cardiac ejection has ceased. The left ventricle with the aorta including the Windkessel is shown schematically in Fig. 8.17. Hence the Windkessel is an energy storage system that suppresses turbulence and regulates blood pressure fluctuations. With age the compliance of the arteries decreases, resulting in a less effective Windkessel conversion and increased systolic pulse pressure. For modeling blood flow, often an equivalent circuit diagram is used as shown in Fig. 8.18 that starts from the aortic valve (AV) providing an end systolic pulsed pressure p_{ESP}, crossing the Windkessel of mechanoelastic volume capacitance V_C, and continuing into the vasculature circulation with a peripheral resistance R [3].

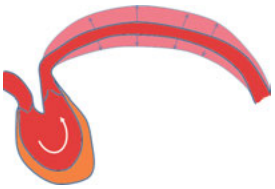


Fig. 8.17: Left ventricle and Windkessel function of the aorta.

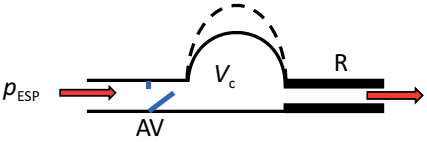


Fig. 8.18: Equivalent circuit diagram for modeling the pulsed flow from the aortic valve (AV) with end systolic pressure p_{ESP}, passing by the Windkessel with mechanoelastic volume capacitance V_C, and continuing into the vasculature circulation of peripheral resistance R.

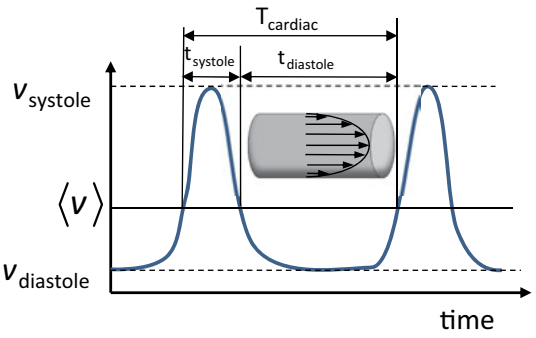


Fig. 8.19: Velocity versus time for pulsatile blood flow determined at a fixed distance Δx from the aorta. The same plot can be made for the time structure of the pressure amplitude.

8.5.4 Flow velocity and pulse wave velocity

Now we return to the pulsatile nature of blood flow. The left ventricle ejects a stroke volume at high pressures into the artery. The systole-diastole pressure wave then propagates down the aorta and up into the carotid artery. This results in two effects: at any distance Δx from the aortic valve and after a propagation time Δt we measure a local flow velocity $v(\Delta x, \Delta t)$ in the fluid and a local pressure amplitude $\Delta p(\Delta x, \Delta t)$ of the vessels, both having pulsing and propagating character, as plotted in Fig. 8.19.

Hence we distinguish between a time independent *mean flow velocity*:

$$\langle v_{\text{mean}}(\Delta x)\rangle = \frac{v_{\text{systole}}\, t_{\text{systole}} + v_{\text{diastole}}\, t_{\text{diastole}}}{T_{\text{cardiac}}},$$

where $T_{\text{cardiac}} = t_{\text{systole}} + t_{\text{diastole}}$, and a time dependent *pulse wave velocity* v_{PWV}. The mean velocity is measured at a fixed distance Δx from the aorta and averaged over time, which can be done, for instance, via Doppler shift sonography (see Section 13.7). For this purpose laminar flow is assumed and the velocity is determined at the peak of the velocity profile at the center of a cylindrical tube as indicated in Fig. 8.19. Table 8.1 lists typical values for the mean velocity at different locations. With increasing distance from the aorta, the pulsatile property of the blood flow ceases and the flow becomes more continuous. The pulsatility of the flow is quantified by a *pulsatility index* (PI), defined as:

$$\text{PI} = \frac{v_{\text{systole}} - v_{\text{diastole}}}{\langle v_{\text{mean}}\rangle}.$$

For $\text{PI} \to 0$ we find $v_{\text{PWV}} = \langle v_{\text{mean}}\rangle$.
The PWV is the propagating velocity of the velocity peak or pressure peak as it travels along the artery, resembling a soliton-like pulse propagation along a stretched rope illustrated in Fig. 8.20.

The pulse propagation can be expressed as:

$$\begin{aligned} t = 0: &\quad y(x,0) = f(x) \\ t > 0: &\quad y(x,t) = f(x - v_{\text{PWV}} t), \end{aligned}$$

Tab. 8.1: Mean velocities and cross sections of blood vessels.

Blood vessel	Mean velocity [cm/s]	Total cross section [cm^2]
Aorta	92	3–5
Pulmonary artery	65	
Capillaries	0.05	4500–6000
Vena cava	15	14

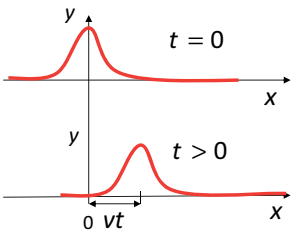


Fig. 8.20: Pulse propagation in one direction.

where $f(x)$ is a function describing the shape of the pulse and v_{PWV} is the pulse wave velocity. Because of the compliance of the blood vessels, the propagating pressure wave can directly be observed by a bulging of the arteries (Fig. 8.21), which is similar to the Windkessel effect discussed above. A pressure wave in a rigid tube would have pure longitudinal character. However, in a blood vessel due to the bulging of the walls the pressure wave has both longitudinal and transverse components.

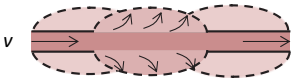


Fig. 8.21: Local bulging of blood vessels due to propagating pressure wave.

The bulging effect is used for determining the blood pressure according to the method of Riva–Rocci as presented before, but also for detecting the arrival time of a pulse wave with a pressure sensitive device, such as a tonometer or a piezoelectric transducer. The amplitude of the pressure wave is a question of vessel compliance. The softer the vessel wall, the greater is the amplitude. Since elastic deformation such as bulging costs energy, the PWV decreases with increasing compliance C. This is expressed in the Moens–Korteweg equation, which relates the PWV to the elastic modulus E_{wall} of the wall, the wall thickness h, the inner diameter d of the blood vessel, and the density of the fluid ρ [3]:

$$v_{\mathrm{PWV}} = \sqrt{\frac{E_{\mathrm{wall}}}{\rho}\frac{h}{d}}.$$

This equation resembles that for longitudinal sound velocity in solids (Chapter 12):

$$v_{\text{sound}} = \sqrt{\frac{E_{\text{solid}}}{\rho}},$$

but the interpretation is quite different. v_{PWV} is a transverse velocity whereas v_{sound} is a longitudinal velocity. Furthermore, v_{PWV} connects the physical properties of two different materials: elastic modulus and thickness of the wall with density of the fluid inside. In contrast, v_{sound} is related to elastic modulus and density of the same material.

PWV provides information on the elastic properties of the arterial system, in particular on the compliance or stiffness of the arterial vessels. The mechanical properties of the arterial walls change along the arterial system. From the large arteries to the periphery the stiffness of the walls increases. Hence the PWV is not a constant but increases with distance from the heart. Furthermore, elastic stiffness also changes with age: the vessels become less compliant, which results in increased PWV and increased blood pressure. Therefore PWV measurement is also being used for an early indication of cardiovascular disease that stiffens up the vessels.

There are a number of methods to determine the PWV, which can be categorized in direct contact methods (tonometer, piezoelectric transducer), imaging methods (sonography, magnetic resonance imaging), and optical interferometry. We will only discuss nonlocal direct contact modes, and refer to the respective chapters for discussions of imaging methods.

In contact mode, PWV is determined by measuring the distance Δx from AV over travel time Δt. Favorable points for measuring the pulse amplitudes are those where the pulse can be easily detected because of its proximity to the skin, such as at the carotid artery and the femoral artery. In one-point measurements the travel time Δt is taken as the time difference between the R point of systole, which is the time when the isovolumetric contraction starts (see Fig. 8.6), and the arrival time of the pulse at the carotid artery (t_1) or femoral artery (t_2). The arrival time is defined as the ascending foot of the pressure pulse, indicated by a red dot in Fig. 8.22. Thus one-point measurements require simultaneous recording of ECG and pressure amplitudes at specific points on the skin. The distance Δx between the AV and points C or F is usually measured with a tape measure on the skin. This is not a precise measurement but usually sufficient for the purpose.

In a two-point measurement ECG recording can be avoided. In this case the *transfer time* (TT), which is the difference in the arrival times t_1 for the carotid pulse and t_2 for the femoral pulse, is measured: $\text{TT} = t_2 - t_1$. This time difference corresponds to the length differences of the femoral point $\Delta x_2 = F$ compared to the carotid point $\Delta x_1 = C$, as indicated in Fig. 8.22.

Therefore the PWV can be calculated according to:

$$v_{\text{PWV}} = \frac{F - C}{\text{TT}}.$$

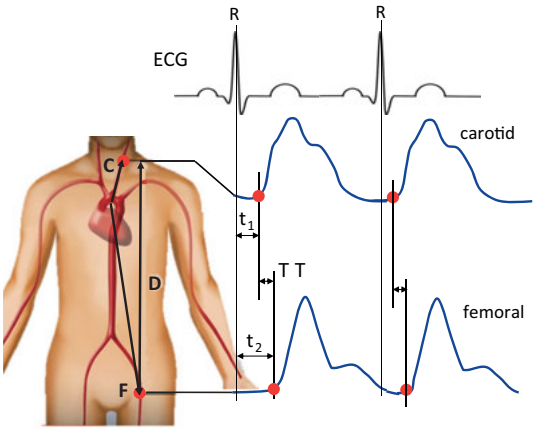


Fig. 8.22: Determining pulse wave velocity using either a one-point measurement together with an ECG recording or a two-point measurement by measuring the time difference TT and the distance D between the carotid and femoral arteries.

F and C are again measured with a tape measure on the skin that may yield some systematic errors. A much less accurate method is, however, obtained by taking the distance D between the carotid and femoral arteries and calculating the velocity according to:

$$v_{\mathrm{PWV}} = \frac{D}{\mathrm{TT}},$$

which, nevertheless, is often practiced. The oversimplification can be rectified by taking a correction factor into account. As D can be easily measured, and the ratio $\alpha = C/F$ is roughly constant independent of body size, a better approximation is:

$$v_{\mathrm{PWV}} = \frac{D}{\mathrm{TT}} \left(\frac{1-\alpha}{1+\alpha} \right).$$

Typical values for C are 100 mm and for F are 560 mm, such that $\alpha = 0.18$ and $(1-\alpha)/(1+\alpha) = 0.7$. PWV determined in such a manner is in the order of 6–8 m/s. When the PWV goes up to 10–12 m/s, a cardiovascular disease is likely. It should be noted that the PWV is much higher than the mean blood flow velocity in this pulsatile part of the artery. In addition, the pulse shape shown in Fig. 8.22 changes drastically from point to point. This is not only due to the varying stiffness of the blood vessels with distance but also due to partial reflection of pressure waves as the impedance changes along the pathway. Furthermore, pulses vary slightly from one systole to the next. Therefore, the foot of pulses or the initial slope is a better indicator for pulse positions than the peak.

8.5.5 Viscosity of blood

Blood is a complex fluid. It consists of extracellular fluid called plasma and blood cells in suspension. The plasma contains water (92 %), plasma proteins (7 %), and solutes (1 %). The blood cells are mostly red blood cells (RBC, *erythrocytes*, 95 %), and to a

much smaller fraction thrombocytes (4.8 %) and leucocytes called white blood cells (WBC, 0.2 %). 1 µl (= 1 mm^3) of blood contains about 5×10^6 RBCs. Male adults have 5–6 l blood, female adults 4–5 l blood in the body.

The lifetime of erythrocytes is 100–120 days. Two to 4 million new erythrocytes are being synthesized any second from special stem cells (hemocytoblasts) located in the red bone marrow at the heads of long bones. The newly formed RBCs use the bone marrow vasculature as a channel into the systemic circulation of the body.

The fraction of erythrocytes to the total blood volume is called *hematocrit value* (Hct) or *packed cell volume* (PCV). The hematocrit value is determined by centrifugation of a test tube filled with a blood volume $V1$ (Fig. 8.23). After centrifugation three areas are recognized: yellowish plasma on top, middle zone containing thrombocytes and leucocytes called buffy coat, and the remaining volume $V2$ containing erythrocytes. The ratio:

$$\text{Hct} = \frac{V2}{V1}$$

is the hematocrit value. Typically, Hct = 40–45 %.

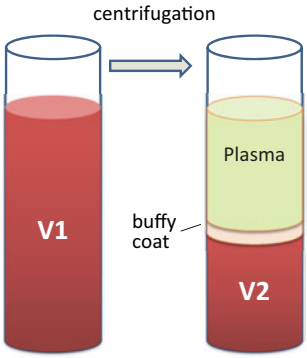


Fig. 8.23: Test tube filled with blood separates into three zones after centrifugation: plasma, buffy coat, and erythrocyte volume.

RBCs contain the red pigment hemoglobin, which binds and transports O_2 and CO_2, discussed in more detail in Section 8.6. Each RBC is a biconcave disc, with a diameter of 7–8 µm and disk thickness of 2.5 µm (Fig. 8.24).

The viscosity of fluids is an important intrinsic and material-specific parameter determining flow resistance. There are many methods for determining the viscosity of fluids, which are discussed in standard textbooks. For practical reasons we discuss here the gliding plate method (Fig. 8.25): two plates, one fixed and the other movable, are separated by a fluid of constant thickness h. The top free plate is moved with a constant speed v_1 by a tangential shear stress $\tau = F/A$, where F is the force and A is the area of the plate. Assuming that the fluid layers next to the surface of the plates stick to the plates by adhesion and do not slip, there will be a velocity gradient of the fluid from the lower plate ($v = 0$) to the upper plate ($v = v_1$). The shear rate $\dot{\gamma}$ is defined as $\dot{\gamma} = v_1/h$ (units $[\dot{\gamma}] = s^{-1}$). Then the viscosity η is defined by the shear stress divided

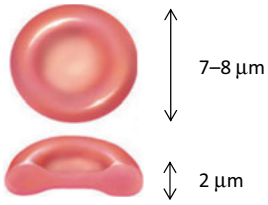


Fig. 8.24: Shape and size of red blood cells.

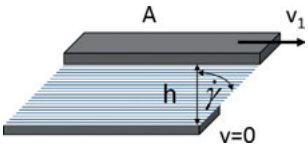


Fig. 8.25: Gliding plate method for determining the viscosity of a fluid.

by the shear rate:

$$\eta = \frac{\tau}{\dot{\gamma}} = \frac{F/A}{v_1/h}.$$

The unit for the viscosity is $[\eta] = \mathrm{N\,s/m^2} = \mathrm{Pa\,s}$. In medical textbooks the cgs unit Poise is often used: 1 Poise (P) = 0.1 Pa s.

The viscosity of fluids depends on temperature and sometimes also on the shear rate. Usually the viscosity of fluids decreases with increasing temperature due to the increasing mobility of molecules in fluids. Typical examples are glycerin or honey, which stick at low temperatures but flow at higher temperatures. Table 8.2 lists a few viscosity values for water and blood. From these values we recognize that at the body temperature of 37 °C the viscosity of blood is five times higher than for water. In fact the viscosity of blood depends on the Hct value as shown in Fig. 8.26. The viscosity quoted in Tab. 8.2 refers to an average Hct of 45 % typical for healthy people.

Tab. 8.2: Viscosity values for water and blood at different temperatures.

Viscosity [mPa s]	0 °C	20 °C	37 °C
Water	18	1	0.8
Blood			4

Fluids that have a viscosity independent of the shear rate $\dot{\gamma}$ are called *Newtonian fluids*. Fluids whose viscosity depends on $\dot{\gamma}$ are called *non-Newtonian fluids*. Blood is a non-Newtonian fluid. The viscosity of blood depends on the shear rate: the viscosity decreases with increasing shear rate (see Fig. 8.27). Because of this characteristic the flow resistance decreases with increasing flow velocity. A Newtonian liquid has a parabolic velocity profile when flowing through a cylindrical tube, i.e., in the center the fluid velocity is highest and decreases to zero at the walls. In contrast, for non-Newtonian fluids the velocity profile is flattened in the center.

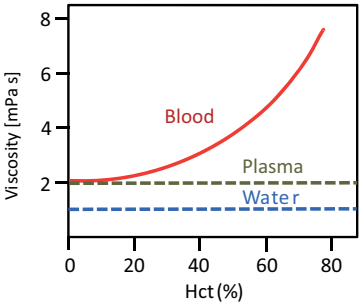


Fig. 8.26: Dependence of viscosity on the hematocrit value.

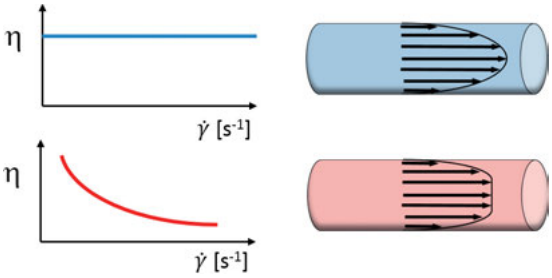


Fig. 8.27: Velocity profiles for a Newtonian fluid with constant viscosity (top panel) and a non-Newtonian fluid with shear rate dependent viscosity (bottom panel). The Newtonian fluid shows a parabolic velocity profile, the non-Newtonian fluid has a flattened velocity profile in the center.

The special properties of blood are due to the shape of the RBCs. Because of the biconcave disk shape, erythrocytes can arrange randomly in streams of wide vessels, but when squeezed through narrow vessels the plates arrange themselves in a more organized fashion parallel to the blood flow and on top of each other. One can distinguish three regions with increasing shear rate (Fig. 8.28 (a)): in region I the orientation of the RBCs is random; in region II the RBCs arrange parallel to each other for better slip; in region III the RBCs have to form chains and can only pass through narrow capillaries one after the other. In fact, they can squeeze through arterioles in the capillary bed as

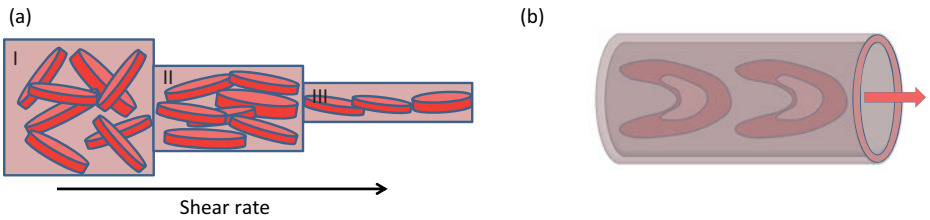
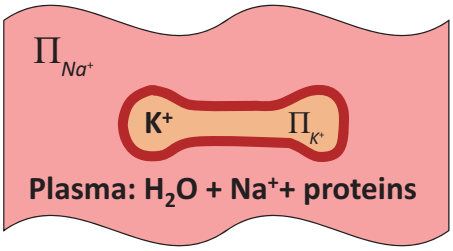


Fig. 8.28: (a) Three regions of blood cell organization can be distinguished, depending on the diameter of blood vessels. With increasing shear rate the RBCs organize themselves in plates or chains. (b) Curled-up shape of erythrocytes in a narrow capillary tube.

narrow as 4–5 µm in diameter without getting stuck by curling up and bending over to reduce their size, as indicated in Fig. 8.28 (b). The high deformability of the RBCs is due to a network of fibers on the inside of the double lipid cell membrane of the erythrocytes called *cytoskeleton*.

8.5.6 Osmotic pressure

Red blood cells contain no nuclei, no mitochondria, and no ion channels. Therefore RBCs cannot perform action potentials and they cannot reproduce themselves. They are synthesized from stem cells in the red bone marrow, as already mentioned. Because of the lack of ion channels the ion concentration in the intracellular space (mainly K^+) remains constant. However, the cell membrane is permeable to water, which is essential for controlling the *osmotic pressure* of RBCs.



$$\Pi_{K^+} \cong \Pi_{Na^+}$$

Fig. 8.29: Osmotic pressure in plasma and inside the blood cell need to be well balanced.

The osmotic pressure follows from the solubility S defined as $S = n/V$, where n is the mole number of the solute and V is the volume of the solvent. The osmotic pressure Π_i of a component i is then:

$$\Pi_i = S_i RT,$$

where R is the gas constant and T is the absolute temperature. This relation is known as Henry's law.

The osmotic pressure in RBCs is determined by the K^+-solubility and the osmotic pressure in the plasma is determined by the Na^+-solubility. The osmotic pressures are high ($\approx$ 7.5 bar). If they are about equal inside and outside ($\Pi_{K^+} \approx \Pi_{Na^+}$), the condition is called *isotonic* and the blood cells have their regular shape as shown in Fig. 8.30. However, any imbalance of the osmotic pressure results in a movement of water across the cell membrane. For $\Pi_{K^+} > \Pi_{Na^+}$ the condition is hypotonic. Additional water penetrates into RBCs, the cells expand and may eventually burst. Hypotonic conditions occur due to lack of Na^+ in the plasma. Vice versa, hypertonic conditions are present if the Na^+-concentration in the plasma is too high. Then $\Pi_{K^+} < \Pi_{Na^+}$ and water moves from RBCs

Fig. 8.30: Red blood cells under isotonic, hypotonic, and hypertonic conditions.

into the plasma to balance the pressures. However, the RBCs will then shrink and lose their ability as carriers for gas exchange. Furthermore, the changing shape under hypotonic and hypertonic conditions changes the elasticity and deformability of RBCs and therefore also the viscosity.

In the case of high loss of blood volume and lack of blood plasma, an *isotonic saline solution* can be injected as immediate and temporary first aid. The isotonic fluid consists of a 0.9 % solution which is 9 g of NaCl per 1000 ml of H_2O.

8.6 Binding of oxygen to heme

Red blood cells are carriers of oxygen (O_2) and carbon dioxide (CO_2). Their color ranges from bright red (O_2-rich blood, oxygenated blood) to dark red (O_2-poor blood, deoxygenated blood). Gas transport by RBCs requires permeability of cell membranes for the gases in question, binding during transport, and release at targeted destinations. RBCs are predestinated for these tasks as they contain the protein hemoglobin. Each erythrocyte hosts 2.5×10^8 *hemoglobin* molecules. The mean hemoglobin concentration in an RBC is about 32–36 %, the rest of the cell is filled with intracellular fluid and other proteins.

8.6.1 Structure of hemoglobin

Each hemoglobin molecule shown in Fig. 8.31 consists of four helical chains, two α-chains, and two β-chains. The polypeptide chains have slightly different lengths: 141 amino acids in α-chains and 146 amino acids in β-chains. More important is the fact that all chains form a pouch that contains one heme molecule and a channel for oxygen access. Each flat heme molecule (Fig. 8.32) has one Fe^{2+} ion at its center for binding oxygen molecules during inspiration. The description of the protein hemoglobin by Max Perutz in 1959 (Nobel Prize in Chemistry 1962) was an early milestone in biochemistry and protein crystallography. With advanced and partially automated techniques using synchrotron radiation, nowadays thousands of protein structures are determined every year.

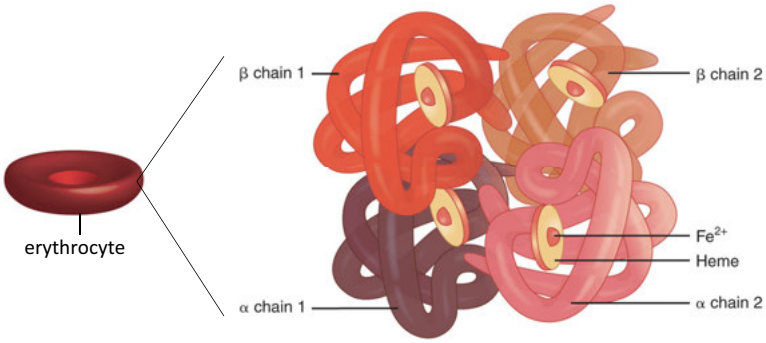


Fig. 8.31: Hemoglobin with four side chains, each hosting one heme molecule (reproduced from Wikimedia, © Creative Commons).

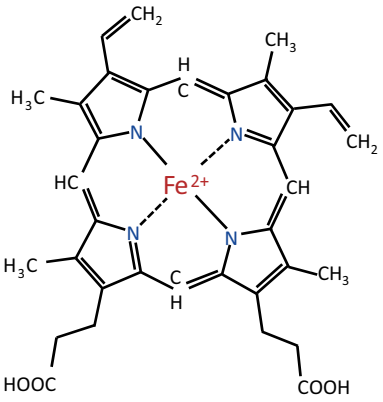


Fig. 8.32: Heme molecule with an Fe^{2+} ion at the center.

8.6.2 High spin-low spin transition

The heme molecule (Fig. 8.32), also called Fe(II)-protoporphyrin, is a large flat molecule that binds the Fe^{2+} ion to four nitrogen atoms. Fe^{2+} has two important properties: it is highly reactive and likely to bind oxygen, and it is magnetic. The atomic configuration of Fe^{2+} is $3d^6$. As such the spin quantum number according to Hund's rule is $S = 2$, the orbital quantum number is $L = 2$, and the total angular momentum quantum number is $J = 4$. The Landé factor g_J is 3/2 and therefore the expected magnetic moment is:

$$\begin{aligned} m_J &= g_J \sqrt{J(J+1)}\mu_B \\ &= \frac{3}{2}\sqrt{4(4+1)}\mu_B = \frac{3}{2}\sqrt{20}\mu_B = 6.7\mu_B. \end{aligned}$$

However, Fe^{2+} is surrounded by a local *crystal electric field* E_{CF} and therefore atomic considerations need to be modified, depending on whether the local crystal electric field E_{CF} is stronger ($E_{CF} \gg \lambda \vec{L} \cdot \vec{S}$) or weaker than the *spin-orbit coupling* ($E_{CF} < \lambda \vec{L} \cdot \vec{S}$). In any case, the local electric field E_{CF} lifts the degeneracy of the atomic orbital moments m_L into two subsets of orbitals, e_g and t_{2g}, as shown in Fig. 8.33 for octahedral local symmetry.

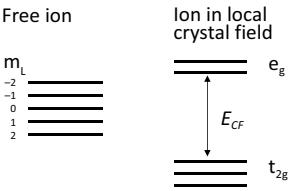


Fig. 8.33: Degenerate orbital levels of 3d electrons in transition metals split into two subsets of energy levels when exposed to a crystal electrical field. For octahedral symmetry the three t_{2g} are lowered in energy and the two e_g states are lifted up.

In the heme molecule the Fe^{2+} ion is well protected by a so called βHis92 molecule. Without this protection, O_2 would immediately bind to Fe^{2+} and form a complex $Fe^{3+}(O_2)^-$, i.e., Fe^{2+} would be reduced by electron transfer to O_2, yielding FeO_2 This bond is so strong that it would be impossible to recycle O_2 during respiration. Therefore, with the help of the globin crevice and in particular the histidine chain βHis92, electron transfer is only partial and much weaker, permitting a reversible bond. The binding of O_2 to hemoglobin is usually written: $Hb + O_2 \rightarrow HbO_2$.

When O_2 binds to the heme group, the Fe^{2+} ion is co-planar within the heme molecule and experiences a strong *octahedral crystal field*. Then only the t_{2g} states are occupied and the total spin moment cancels to zero: $S = 0$. Upon oxygen release, Fe^{2+} moves slightly out of the heme plane by 0.04 nm pushing the βHis92 molecule up. This weakens the crystal field and the splitting between e_g and t_{2g} states becomes much reduced. Then also the energetically higher e_g states can be occupied, resulting in a high spin $S = 2$ state. The schematics of the spin occupancies are shown in the bottom panel of Fig. 8.34. The reversible oxygen binding in hemoglobin simultaneously drives a *high spin – low spin* (HSLS) *transition* in Fe^{2+} as we can recognize from the upper panels in Fig. 8.34. For the physiology of the respiratory system this spin transition is not important. However, for functional imaging of brain activity via magnetic resonance imaging, discussed in Chapter 15, the HSLS transition is essential, because the proton relaxation time T2 depends on the local Fe spin state. Using scans that are sensitive to the T2 relaxation time one can map out those parts of the brain which consume more oxygen in response to an external stimulus, like listening to music or associating words, as compared to other nonactivated parts of the brain. This type of brain mapping is referred to as functional MRI (fMRI).

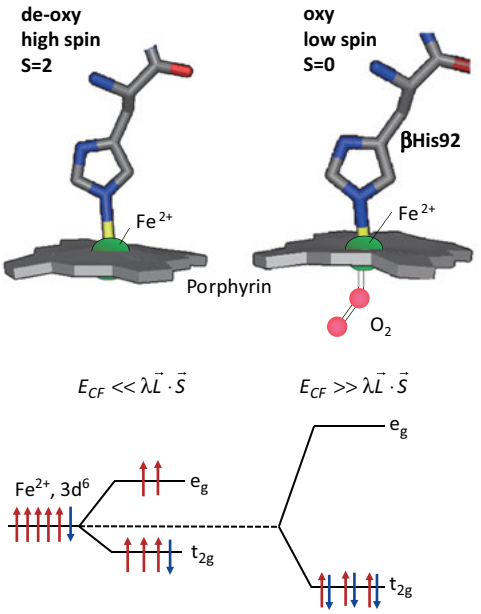


Fig. 8.34: Correlation between oxygen binding to Fe^{2+} in porphyrin and the magnetic properties, featuring a high spin–low spin transition. Top panel shows the heme molecule and the βHis92 bundle attached. Without oxygen binding, the Fe^{2+} ion is in the heme plane and in a high spin state. With oxygen binding, Fe^{2+} is pushed out of the plane and assumes a low spin state. Lower panel: each arrow represents the spin of one electron. The degenerate state to the left is lifted by local crystal fields. For low crystal fields in deoxy state the splitting between t_{2g} and e_g levels is small, resulting in a high spin state. In the oxy state the crystal field is stronger, resulting in a high splitting of the t_{2g} and e_g levels, causing a repopulation of the electron levels to a low spin state.

8.6.3 Saturation curve

The reversible binding of oxygen to hemoglobin in erythrocytes or myglobin in muscle cells can be represented in oxygen binding curves where the fractional saturation of oxygen is plotted versus the partial pressure of oxygen during inspiration. The *fractional saturation of oxygen* is defined as:

$$Y = \frac{[HbO_2]}{[HbO_2] + [Hb]},$$

where $[HbO_2]$ and $[Hb]$ are the concentrations of hemoglobin with and without bonded oxygen, respectively. The oxygen concentration in the blood is according to Henry's law (Section 8.5.6) proportional to the partial oxygen pressure p_{O_2} in the alveolar space where the oxygen exchange takes place: $[O_2] \approx p_{O_2}$. The rate of O_2 bonding to Hb depends on the concentration of O_2 in the plasma and the concentration of vacant Hb sites:

$$\dot{Y} = k_b[O_2]N(1 - Y).$$

Here N is the total number of hemoglobin molecules, $N(1 - Y)$ is the number of sites that can still be occupied by O_2, and k_b is a bonding coefficient. Conversely, the rate of dissociation depends on the oxygen concentration that is bound to Hb, where k_d is a dissociation constant:

$$\dot{Y} = k_d[O_2]NY.$$

In equilibrium bonding and dissociation rates must equilibrate at a constant pressure and temperature, and therefore we have:

$$k_b[O_2]N(1 - Y) = k_d[O_2]NY.$$

Solving for the fractional occupation Y, we obtain using $K = k_d/k_b$:

$$Y = \frac{[O_2]}{K + [O_2]}.$$

This is the standard *Langmuir equation*. It holds for the bonding of oxygen to myoglobin, but not for hemoglobin. Hemoglobin has an S-shaped saturation curve at low oxygen pressures before going into saturation (Fig. 8.35), described by:

$$Y = \frac{[O_2]^n}{K + [O_2]^n}.$$

In this form the equation is known as the *Hill equation*, n is called the *Hill coefficient*. n is a measure of the cooperativity of different binding sites. If $n > 1$, the binding is cooperative, meaning the affinity to O_2 increases with the number of O_2 already bonded. If $n < 1$ the binding is counter-cooperative. For $n = 1$ the bonding-dissociation rates follow the Langmuir rate equation.

For hemoglobin the Hill coefficient n is about 2.5. Myoglobin is another protein that contains a heme molecule for reversible oxygen binding and transport in muscles. In the case of myoglobin, the S-shape is missing and the saturation follows the more simple Langmuir curve with $n = 1$.

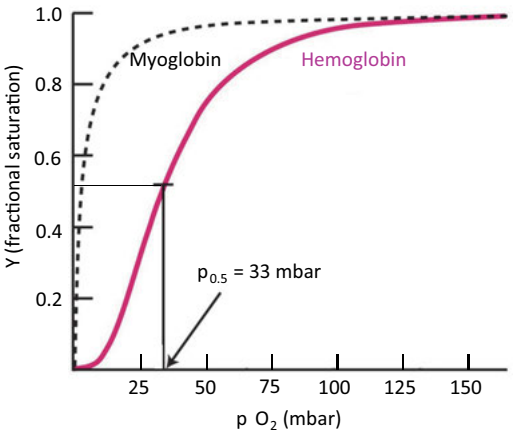


Fig. 8.35: Fractional saturation of oxygen in hemoglobin and myoglobin as a function of partial oxygen pressure.

From a physiological point of view there are several remarkable features about these saturation curves compared in Fig. 8.35. First, they show that oxygen molecules are chemically bound to hemoglobin (myoglobin), because both exhibit a steep rise at low pressures and saturation occurs at higher pressures. Physical solution of gaseous O_2 is also present but at a much lower level and following a linear dependence as a function of oxygen pressure $p(O_2)$ (see Fig. 9.4). Furthermore, the S-shape for hemoglobin indicates that at low pressures binding of O_2 is weaker and increases with increasing fractional saturation. This implies that by a positive feedback the affinity to oxygen increases with each oxygen molecule that is already bound. Indeed, the α- and β-chains rearrange in response to oxygen uptake from a tense form (T) without oxygen to a relaxed form (R) with oxygen. This cooperation is typical for hemoglobin, requiring the cooperative action of four chains and is not present in myoglobin, which consists of only a single chain. Consequentially for myoglobin the saturation curve is not S-shaped.

In saturation each heme molecule carries one O_2 molecule. This makes $4 \times 2.5 \times 10^8 = 10^9$ per erythrocyte, and 5×10^{21} per liter blood. Assuming that 2 l are fully oxygenated on average, and the remaining 3 l have an oxygen content of 75 % after expiration, then the total oxygen content of blood is 0.03 mole of oxygen.

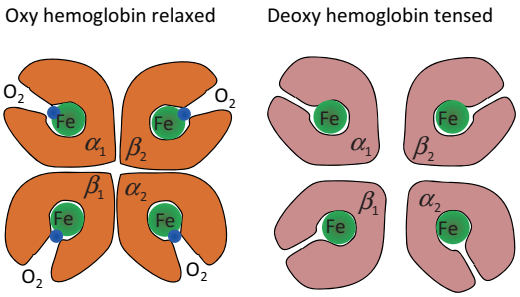


Fig. 8.36: Tense and relaxed states of hemoglobin consisting of four chains (α_1, α_2, β_1, β_2) that cooperate positively for increased oxygen affinity: the more oxygen is bound, the more relaxed the four chains become and the affinity to more oxygen binding increases.

The partial pressure at half saturation $p_{0.5}$ is a measure of the affinity for oxygen binding. In the case of hemoglobin, $p_{0.5}$ is about 33 mbar, whereas saturation is reached beyond 100 mbar. Myoglobin has a much higher oxygen affinity: $p_{0.5}$ is 2–5 mbar and saturation is already reached at 50 mbar. At sea level the partial pressure of atmospheric O_2 is 210 mbar = 210 hPa, sufficient for saturation. However, at higher altitudes, such as Mount Everest the partial O_2 pressure drops to 70 mbar, not sufficient for sustainable life without an oxygen mask.

Oxygen is not the only molecule with a high affinity to heme. Carbon monoxide can also bind to Hb with an affinity that is higher than that of O_2 by a factor of 200.

Carbon monoxide poisoning results from two factors: first, CO blocks sites in Hb for O_2 uptake and therefore hinders O_2 transport. And second, due to the presence of CO the bonding strength in remaining HbO_2 is enhanced and therefore impedes O_2 release at target sites in tissues.

8.6.4 Ferritin

Above we recognized the importance of iron for oxygen transport. We have also noticed that a huge amount of erythrocytes including hemoglobin are being synthesized every second in the bone marrow. For each hemoglobin molecule synthesized, four Fe atoms are required. Therefore the body must run an Fe recycling system and store Fe for buffering fluctuations. About 3.7 g of Fe are in the body at any time. Of these 2.5 g are bound in hemoglobin for oxygen transport, some 0.2 g are bound in myoglobin, and another 0.02 g are distributed over various proteins. The remaining 1 g is the reserve and is stored in ferritin.

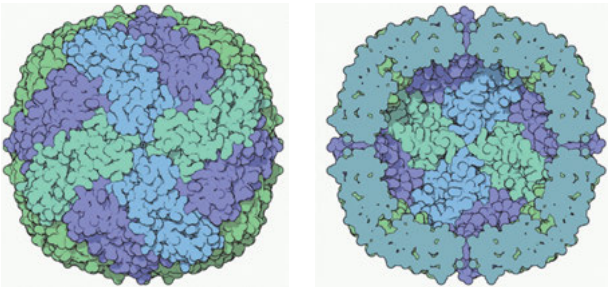


Fig. 8.37: Ferritin is formed by 24 proteins forming a hollow sphere which houses Fe in form of small crystals. Left side: closed structure; right side: cross section of the hollow ferritin sphere (from www.rcsb.org/pdb/molecules/pdb35_1.html).

Ferritin is a highly interesting biomaterial [4]. It consists of 24 identical proteins which together form a hollow sphere, shown in Fig. 8.37. Only a few pores allow access to the interior. Once Fe ions have penetrated through the pores, they are bound in ferritic molecules, forming small crystals. Each ferritin stores about 4500 Fe ions. Ferritin can be found mainly in the liver and in the spleen.

In ferritin Fe is bound in the oxidation state Fe^{3+} within small crystals with the stoichiometry $[FeO(OH)]_8[FeO(H_2PO_4)]$. If needed for synthesis of hemoglobin, Fe^{3+} must first be reduced to Fe^{2+}. Then $Fe(H_2O)_6^{2+}$ goes into solution and can leave ferritin through the pores. The entire metabolism of iron is described in textbooks on physiology. Here we only want to mention that ferritin is used in nanomedicine for transporting targeted drugs or magnetic nanoparticles protected by a biocompatible shell. This is discussed in more detail in Chapter 14/Vol. 2.

8.6.5 Absorbance

The color of RBCs changes dramatically from bright red of well-oxygenated blood to deep red with blueish hue of deoxygenated blood. This is the origin of the corresponding color coding for arteries and veins in schematic drawings of the circulatory system. The color difference can be quantified by recording light absorption spectra in the visible regime of oxygenated and deoxygenated blood. According to the *Lambert–Beer equation*, the light intensity after penetrating a cuvette of thickness t filled with hemoglobin is:

$$I(\lambda, t) = I_0(\lambda) \exp(-\varepsilon(\lambda) s t) = I_0 10^{-A}.$$

Here I_0 is the incident intensity and the absorbance A as a function of wavelength λ is:

$$A = \log_{10} \frac{I_0(\lambda, t)}{I(\lambda)} = \frac{\varepsilon(\lambda) s t}{\ln 10}.$$

$\varepsilon(\lambda)$ is the wavelength dependent extinction coefficient, and $s = f[\mathrm{Hb}] + (1 - f)[\mathrm{HbO_2}]$ is the combined concentration of oxygenated and deoxygenated hemoglobin in the sample with fraction $(1 - f)$ and f, respectively. The absorbance of hemoglobin is plotted in Fig. 8.38. The color that we observe is the one that is less absorbed and more reflected. In the case of oxygenated hemoglobin, absorption bands are at 535 nm and 575 nm whereas little absorption occurs at 560 nm and above 600 nm. For deoxygenated hemoglobin, an absorption band exists at 560 nm, but absorption is low beyond 650 nm. The absorption spectra reproduced in Fig. 8.38 are characteristic and can be quantified with respect to the oxygen concentration bound to hemoglobin [5]. Absorbance measurements are usually performed in vitro. In vivo measurements are difficult to execute as blood vessels are always covered by skin which affects spectral characteristics and contributes noise. Nevertheless, attempts have been undertaken to determine in vivo blood oxygen saturation by diffuse reflectance spectroscopy. However, the technical efforts are still so extensive that routine application in clinics for patients with myocardial problems still appears to be far ahead, although this would

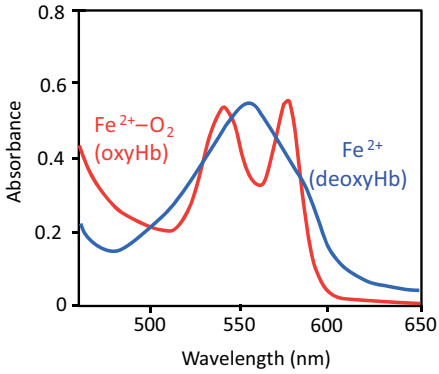


Fig. 8.38: Absorbance spectra of oxygenated and deoxygenated hemoglobin.

be the ultimate test for successful heart surgery. After all, the heart is supposed to pump oxygen enriched blood to the periphery, and this needs verification. Alternatively, oxygen perfusion in blood vessels can be studied by the use of isotope enriched oxygen ^{17}O in conjunction with MRI imaging techniques (see Chapter 15). But this is obviously a rather elaborate and expensive method. Hence a more practical solution is still to be developed.

8.7 Summary

1. The heart contains two separate but not independent pumps.
2. The right and left ventricles work synchronously and have to pump the exact same amount of blood, which is about 15 000–20 000 liter per day.
3. The right ventricle pumps deoxygenated blood from the body to the lung at low pressure. The left ventricle pumps oxygenated blood from the lung to the body at high pressure.
4. The pumping cycle consists of four phases: inflow phase or isobaric filling phase, isovolumetric contraction phase, outflow phase or isobaric ejection phase, and isovolumetric relaxation.
5. The mechanical power of the heart is about 1.25 Watt with a cardiac output of about 90 ml/s. The total power consumption of the heart is about 5.6 Watt.
6. The efficiency of the heart is about 25 %.
7. The pressure difference between the arterial and the venous side during systole is about 140 hPa.
8. The Windkessel suppresses turbulence in the aorta and stores potential energy during ejection. The potential energy is converted back into kinetic energy during the diastole phase.
9. The elastic properties of blood vessels for high pressure (arteries) circulation are distinctly different from blood vessels for low pressure (veins) circulation. Veins have a high elastic compliance but low yield; arteries have a low compliance but high yield.
10. Blood flow into the arteries is pulsatile.
11. Pulsatile blood flow can be used for measuring blood pressure according to the method of Riva–Rocci.
12. The pulse wave velocity is higher than the mean blood flow velocity in the arteries.
13. The pulse wave velocity is proportional to the elastic modulus of blood vessels.
14. The mean blood flow resistance is about 100 kPa per liter and second, which is rather low and due to the wide branching in the capillary bed.
15. Blood consists of plasma and mostly red blood cells. The volume fraction of red blood cells is called hematocrit and is about 40–45 %.
16. Blood is a non-Newtonian fluid. The viscosity of blood decreases with increasing shear rate.
17. Red blood cells have no ion channels but have pores for water exchange. The osmotic pressure inside and outside must be balanced (isotonic conditions).
18. Red blood cells are carriers of oxygen and of carbon dioxide. Oxygen binds to the Fe^{2+} in heme molecules, sitting in a pouch surrounded by amino acid chains.
19. Upon binding of oxygen, the Fe^{2+} ion undergoes a high spin – low spin transition, which is important for functional MRI.
20. Fe^{3+} is stored in ferritin. Ferritin is an important biocompatible hollow nanoparticle formed by proteins.
21. The absorption spectrum in the visible regime is characteristically different for oxygenated and deoxygenated hemoglobin explaining the color of blood.

References

[1] Knaapen P, Germans T, Knuuti J, Paulus WJ, Dijkmans PA, Allaart CP, Lammertsma AA, Visser FC. Contemporary reviews in cardiovascular medicine, myocardial energetics and efficiency: Current status of the noninvasive approach. Circulation. 2007; 116: 434–448.
[2] Milnor WR. Hemodynamics. Baltimore: Williams & Wilkins; 1982.
[3] Butlin M. Structural and functional effects on large artery stiffness: An in-vivo experimental investigation. PhD thesis. The University of New South Wales, Australia; 2007.
[4] Theil EC. Ferritin: structure, gene regulation, and cellular function in animals, plants and microorganisms. Annual Review of Biochemistry. 1987; 56: 289–315.
[5] Pittman RN. Regulation of tissue oxygenation. San Rafael (CA): Morgan & Claypool Life Sciences; 2011. Chapter 10, Measurement of oxygen.

Further reading

Feher J. Quantitative human physiology: An introduction. Academic Press imprint of Elsevier; 2012.
Rhodes RA, Bell DR. Medical physiology. 3rd edition. Wolters Kluwer, Lippincott Williams and Wilkins; 2009.
Boron WF, Boulpaep EL. Medical physiology. 2nd edition. Saunders W.B. Elsevier; 2012.
Guyton AC, Hall JE. Textbook of medical physiology. 11th edition. Elsevier Saunders; 2006.
Blundel S. Magnetism in Condensed Matter. Oxford University Press; 2001.

9 The respiratory system

9.1 Introduction

While some of our organs fulfill their impressive tasks in complete silence such as kidneys, liver, and pancreas, the respiratory system, in contrast, is a comparatively noisy organ always noticeable by chest and diaphragm movement during inspiration and expiration. Like heart beats, respiration is a most obvious vital sign. Both heart and lung have their own rhythms and both are self-stimulated.

The respiratory system consists of several parts: nose and oral cavity allow for inhalation, the trachea channels air to the lower bronchial tree, and the lungs ensure gas exchange with the blood. Respiration delivers oxygen from the environment to the circulatory system. Hemoglobin in the erythrocytes is the carrier of oxygen to the tissue wherever oxygen is needed for energy production. On its return, CO_2 as metabolic end product is taken up by the circulation for expiration to the environment (Fig. 9.1). In capillary beds at both ends specialized arterioles take care of the gas exchange.

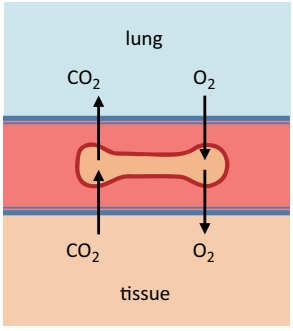


Fig. 9.1: Gas exchange with erythrocytes in the lung and in the tissue following pressure differences.

During inspiration the inhaled air is warmed up to body temperature, filtered, and moistened before entering tiny sacs in the lungs called the *alveoli*. In the alveoli the O_2-CO_2 gas exchange with blood takes place. Oxygen is then transported and distributed to various destinations in the tissue and finally to individual target cells.

From the point of view of physics, elastomechanics of the chest and aerodynamics of the respiratory tract are central for understanding the respiratory system. These topics are the focus of this chapter after introducing some basic physiological aspects.

DOI 10.1515/9783110372830-011

9.2 Respiratory organs

The *respiratory system* includes all organs that serve the purpose of transporting oxygen into and carbon dioxide out of the cardiovascular system (Fig. 9.2). Therefore the respiratory system and the cardiovascular system are strongly interlinked. The respiratory system is the delivery service to the conveyer belt of the blood stream that carries oxygen to the cells. This loop is necessary, since in our body direct diffusion of oxygen to the cells is not possible, unlike in small animals, such as insects, salamander, frogs, etc. where gas exchange through the skin takes place without specialized respiratory membranes, known as cutaneous respiration.

Within the respiratory system, the lungs are responsible not only for *gas exchange* but also for the pH level of the blood, for heat exchange, moistening the inflow of air, control of air flow through the vocal cords for articulation, and for cleaning inhaled air from dust and contaminations. However, the dust and smoke filter is easily exhausted by overload from industry, traffic, and tobacco smoking with the result that gas exchange in the alveoli is hindered. Severe diseases such as silicosis and bronchial carcinoma may result from such adverse effects.

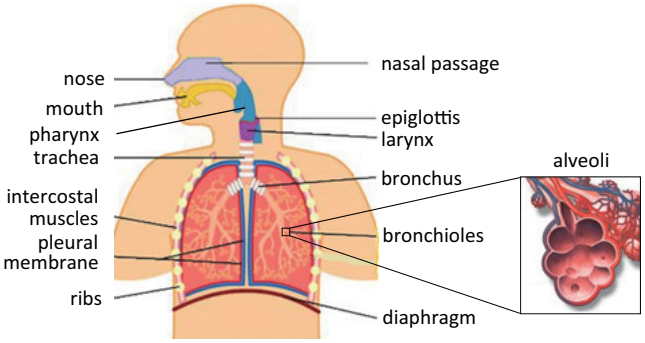


Fig. 9.2: The respiratory system indicating the main organs from the upper parts (mouth, nose, pharynx, larynx) to the lower parts (bronchus, lung, diaphragm). Bronchioles terminate in spherical air sacs called alveoli where gas exchange takes place. The alveoli are surrounded by capillaries.

The different parts of the respiratory tract sketched in Fig. 9.2 are as follows:

The *pharynx* is the common pathway above the epiglottis of both the digestive and the respiratory system. It receives air from the nasal passage and air, food, and liquids from the mouth. The pharynx is connected to the respiratory system at the larynx and to the digestive system at the esophagus.

Air and food both pass through the pharynx and are separated at the *epiglottis*. The epiglottis is like a two way switch: it either opens for air to enter the trachea, or it closes the entrance of the trachea and opens the esophagus for swallowing food.

Muscles control the folding of the epiglottis being either upright for breathing or touching the base of the tongue thereby blocking the glottis and directing food to the esophagus.

The *larynx* is part of the trachea and has the function of a gatekeeper for the lower part of the respiratory ducts and for the vocal cords. The larynx consists of an assembly of cartilage, which is connected to joints, ligaments, and membranes. The larynx containing the voice box (glottis) can regulate the position of the cartilage and the tension of the vocal cords. It is the organ that makes the sound. Larynx and glottis are again topics of Section 12.9 on the origin of sound.

The *trachea* branches off into left and right *bronchus* conducting air to the left and right lungs. From there the primary bronchi branch off further into many more secondary and tertiary bronchi. Air is then channeled through bronchioles into terminal alveoli where the gas exchange takes place. There are about 23 generations of branching from the trachea to the final bronchioles. The final branching together with terminating alveoli is indicated in Fig. 9.2. The diameter of alveoli ranges from 75 µm to 300 µm. There are about 3×10^8 alveoli in both lungs with an average surface area of about 80 m^2 and a total volume of about 5–6 l. All these numbers are mean values; actual numbers depend on body size and sex.

The lungs are the main organ of the respiratory system and with respect to volume and weight the largest organ in the body. Each lung weighs about 500–600 g. The lungs are contained in the *thorax cavity* consisting of the *thoracic wall* and the *diaphragm* at the lower end. The thoracic wall is made up of thoracic vertebrae, ribs, intercostal muscles, and the breastbone (sternum). The diaphragm is a large dome-like muscle separating the thoracic cavity from the abdominal cavity. The muscles associated with the thorax cavity are responsible for respiration as we will see later.

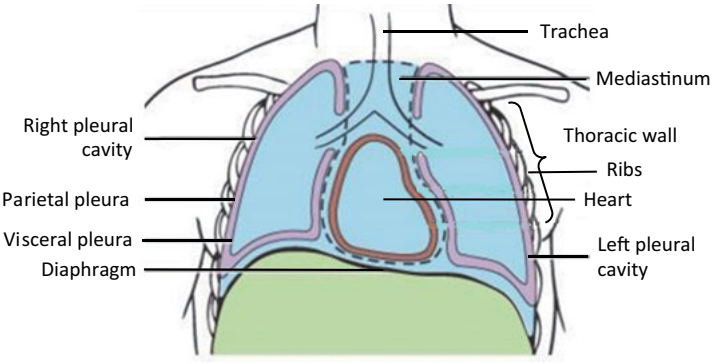


Fig. 9.3: Thoracic cavity consisting of thoracic wall on the sides and diaphragm at the bottom. Each lung is enclosed by a pleural cavity. Left and right pleural cavities are separated by the mediastinum formed by the heart, trachea, esophagus, and associated structures.

The lungs would not be ventilated without the interaction of the thorax with the *pleural cavity* surrounding the lungs, together maintaining an underpressure of about −500 Pa with respect to the environment. Each lung is enclosed by a separate pleural cavity formed by a membrane, as shown in Fig. 9.3. The membrane is folded back onto itself forming a pleural sac with an inner wall covering the surface of the lung (visceral pleura) and an outer wall covering the inner thoracic wall (parietal pleura). The pleural cavity is filled with a pleural fluid that serves as a lubricant between the gliding inner and outer walls when the lungs change shape during respiration. The pleural fluid also holds the two membranes together when the thoracic volume changes. In the case of rupture of the pleural cavity the underpressure is lost, the lung will collapse and respiration is no longer possible. The condition is known as *pneumothorax*. Since the pleural cavities of the left and right lungs are not connected, rupture of one cavity will not affect the other and limited breathing with one lung can be continued.

9.3 Gas exchange

Blood is pumped from the right atrium into the pulmonary circuit at a low pressure of about 3 kPa. The body holds about 6–7 l of blood; about 1 l is always present in the lung. This blood volume spreads out over the capillary bed in the lung. In the capillary bed three actions take place:

1. *Perfusion*: blood must flow from the pulmonary circuit to the capillary bed and back again.
2. *Ventilation*: air must reach the alveolar surface in the periodic rhythm of inspiration and expiration.
3. *Diffusion*: gas exchange across the membrane walls of the blood vessels and the walls of the alveoli occurs by diffusion, following a concentration gradient.

Now we take a closer look at gases and their respective partial pressures on both sides of the membrane, which control the gas exchange.

First we consider the *mole fractions* that make up our atmosphere. If n_i is the mole number of one of the component i in the gas, then the sum of all components yields the mole number n:

$$n = n_1 + n_2 + n_3 + \cdots = \sum_i n_i .$$

According to the general gas law, these mole components generate a total pressure:

$$\begin{aligned} p_{\text{tot}} &= \frac{RT}{V} n = \frac{RT}{V}(n_1 + n_2 + n_3 + \cdots) \\ &= p_1 + p_2 + p_3 + \cdots . \end{aligned}$$

Here R is the gas constant, T is the absolute temperature, and V is volume filled with gas. Each component therefore has the partial pressure:

$$p_i = \frac{RT}{V} n_i = \frac{n_i}{(n_1 + n_2 + n_3 + \cdots)} p_{\text{tot}} = \gamma_i p_{\text{tot}},$$

where γ_i is the mole fraction of the component i. Our atmosphere has the following mole fractions for the respective gases:

$$\gamma_i = 78\,\%\,\text{N}_2,\ 20.95\,\%\,\text{O}_2,\ 0.93\,\%\,\text{Ar},\ 0.03\,\%\,\text{CO}_2.$$

Assuming *standard ambient temperature and pressure* conditions (SATP), defined by 293.15 K (20 °C) and an absolute pressure of 100 kPa, the partial pressures of the gases are accordingly:

$$p_{\text{O}_2} = 21\,\text{kPa},\quad p_{\text{N}_2} = 78\,\text{kPa},\quad p_{\text{Ar}} = 930\,\text{Pa},\quad p_{\text{CO}_2} = 30\,\text{Pa}.$$

So far we have neglected the pressure produced by water vapor. At 20 °C and a relative humidity of 50 %, the water vapor pressure is 1.2 kPa. All mole fractions and partial pressures of the air are listed in Tab. 9.1.

Tab. 9.1: Mole fractions and partial pressures of air components at normal SATP conditions.

Gas component	Mole fraction [%]	Pressure [kPa]
N_2	78	78
O_2	20.95	21
Ar	0.93	0.93
CO_2	0.03	0.03
H_2O	1.2	1.2

Next we consider the pressure conditions for inspiration and expiration. During inspiration the air volume inhaled, called the *tidal volume*, is about 0.5 l of fresh air at a partial oxygen pressure of 21 kPa. In the lungs or, better, in the alveolar space the fresh air is mixed with 2 l of alveolar residual air. Due to this mixing the oxygen partial pressure drops to 13 kPa. At the same time the CO_2 partial pressure from expiration is relatively high at 5 kPa, much higher than in the atmosphere. Those are the conditions that we need to consider later when discussing gas exchange and diffusion across membranes and vessel walls.

The expiration volume, which is discussed in more detail in the next section, contains about 0.15 l of residual air from the trachea that was not in contact with the alveoli. In the residual volume the partial pressure of O_2 is higher than in the alveoli and the partial pressure of CO_2 is lower than in the alveoli. The ratio of CO_2 expiration pressure to O_2 inspiration pressure is called the *respiratory coefficient* (RC):

$$\text{RC} = \frac{p_{\text{CO}_2}}{p_{\text{O}_2}}.$$

RC is roughly 1 when metabolizing carbon hydrates and about 0.7 when burning fat because of the higher COE of fat compared to carbon hydrates (see Chapter 4). The mole fractions and partial pressures during inspiration and expiration together with the conditions in the alveoli are listed in Tab. 9.2.

Tab. 9.2: Mole fractions and partial pressures during respiration.

	% O_2	p_{O_2} [kPa]	% CO_2	p_{CO_2} [kPa]
Inspirational air in the trachea	21	21	0.04	0.04
Alveoli air space	14	13	5.6	5.3
Expirational air in the trachea	16	15	4.5	4.3

The pressure difference on both sides of the membrane controls the direction of diffusion of gases across the barriers. Furthermore, the partial pressure of oxygen in the alveoli governs the solution of oxygen gas in the blood according to *Henry's law*:

$$S_i = \frac{n_i}{V} = K_H p_i.$$

Here K_H is the Henry constant and p_i is the partial external pressure of gas component i. This equation shows that the partial gas pressure p_i controls the solubility $S_i = n_i/V$ of this particular gas component i in the liquid. The unit of K_H is $[K_H] = \text{mol/l Pa}$. Table 9.3 gives an overview of some solubilities of gases in water. This table also shows that the *solubility of oxygen* in water is comparatively low, about a factor of 18 less than the solubility of CO_2 in water. These considerations also apply to oxygen solubility in lakes, rivers, and sea. If the oxygen solubility in water drops from the normal level of 2.7×10^{-4} mol/l to less than half, life in water that depends on oxygen ceases.

Assuming that the solubility of oxygen in blood plasma is the same as in normal water, the oxygen solubility at a partial pressure of 13 kPa, corresponding to the alveolar partial oxygen pressure during inspiration, is only about 4 % of what is really needed to support the oxygen consumption of the body. If physical solubility were the only mechanism to supply oxygen to the cells, about 150 l of blood would be required

Tab. 9.3: Henry constants and solubility of gases in water according to their partial pressures in the atmosphere.

	K_H [10^2 mol/l Pa]	S [10^{-4} mol/l]
Ar	1.5	0.14
CO_2	23	0.07
N_2	0.7	5.4
O_2	1.3	2.7

instead of the actually available 6 l in the body. From this we infer that there must be another mechanism active to increase the oxygen concentration in blood. The other mechanism is chemical bonding of oxygen to hemoglobin according to the reaction:

$$Hb + O_2 = HbO_2.$$

The bonding process of oxygen to hemoglobin is described in Section 8.6.3. With chemical bonding of oxygen to hemoglobin in the erythrocytes – four O_2 molecules per hemoglobin – at an alveolar pressure of 13 kPa, 98 % saturation is reached. Physical solution and chemical bonding of O_2 in hemoglobin is compared in Fig. 9.4 as a function of oxygen pressure. This graph shows that chemical bonding and transport by hemoglobin is by far the dominating process of oxygen supply in the body.

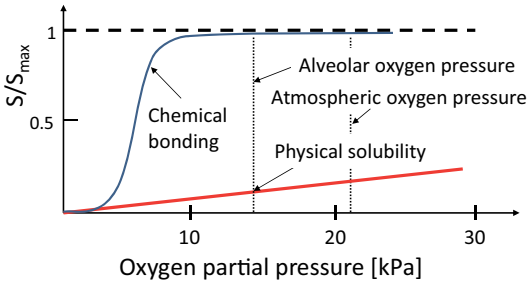


Fig. 9.4: Oxygen solution in blood and oxygen bonding in red blood cells.

Each erythrocyte carries about 10^9 O_2 molecules in saturation. This makes about 5×10^{21} oxygen molecules per liter blood, or roughly 0.01 mol O_2/l, much more than physical solution would deliver. The venous blood still contains about 75 % of the original oxygen uptake. In the case that the oxygen pressure drops in muscles due to activity, more oxygen is exchanged. Only in the coronary vessels is the $O_2 \leftrightarrow CO_2$ exchange complete.

CO_2 transport in blood is very different from O_2 transport. There are three main pathways for CO_2 from the tissue, where metabolic CO_2 is produced, to the lung. (1) According to Tab. 9.3 about 7–8 % of CO_2 is physically dissolved in blood. (2) Another 5–8 % diffuses across the membrane of erythrocytes and is reversibly bonded to the β-chains of hemoglobin but not to the heme molecule, forming carbaminohemoglobin. (3) The majority of about 85 % CO_2 reacts with water within the cytoplasm of the erythrocytes forming carbonic acid (H_2CO_3) with the help of the enzyme carbonic anhydrase. Carbonic acid, also known as respiratory acid, is highly unstable and decomposes into bicarbonate (HCO_3^-) and protons (H^+). These two reactions are as follows:

$$CO_2 + H_2O \rightleftarrows H_2CO_3$$
$$H_2CO_3 \rightleftarrows HCO_3^- + H^+.$$

Two thirds of the bicarbonate diffuses back from the cytoplasm to the blood plasma and is transported via the blood stream to the lungs. In the capillaries of the lung all

reactions run backwards: bicarbonate moves back to the erythrocytes, where it is split up into CO_2 and water. Together with CO_2 from carbaminohemoglobin and physically dissolved O_2, CO_2 finally diffuses through the air-blood barrier to the outside, following the pressure gradient of 5.3 kPa inside versus 4.3 kPa in the alveolar space. As a side note, carbon monoxide in contrast to carbon dioxide is bonded to the same site in hemoglobin as O_2 but with a much higher affinity. Therefore CO is toxic, while CO_2 is exhaled.

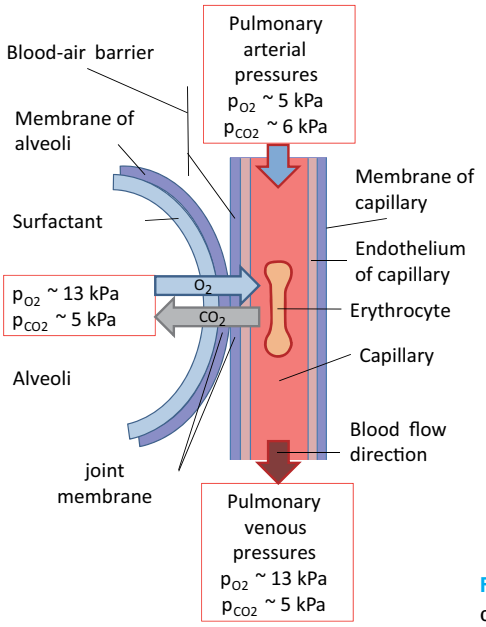


Fig. 9.5: Partial pressures on both sides of the blood-air barrier.

Now we are prepared to once more inspect the $O_2 \leftrightarrow CO_2$ gas exchange in the alveoli and pulmonary capillaries. Referring to Fig. 9.5, blood arrives on the pulmonary arterial side of the lung with O_2 and CO_2 concentrations that correspond to partial pressures of 5 kPa and 6 kPa, respectively. The partial pressure of O_2 is higher in the alveoli than in the erythrocytes, while for CO_2 the reverse applies. The pressure difference is $\Delta p_{O_2} = 8$ kPa and $\Delta p_{CO_2} = -1$ kPa for oxygen and carbon dioxide, respectively. Therefore oxygen will diffuse into the erythrocytes and carbon dioxide will diffuse in the opposite direction. On the pulmonary venous side of the capillary the erythrocytes are oxygen enriched and carbon dioxide depleted according to the partial pressures. The joint membrane acting as blood-air barrier, also known as the *alveolar-capillary barrier*, consists of four layers: a surfactant, an alveolar epithelium, a basal lamina, and endothelium, adding up to a total thickness of about 1 µm. Altogether the oxygen molecules have to pass 9 membrane walls on their way from the alveolar airspace to the inside of the erythrocytes. Nevertheless, the gas exchange rate, controlled by the

diffusivity and permeability of the blood-air barrier, is high enough for a sufficient O_2 supply to the body. The rate of oxygen uptake by the lung is 250 ml O_2/min or 360 l/day. This agrees well with our daily oxygen requirement according to the metabolic rate, estimated in Chapter 4. Diffusion depends on pressure differences. If more oxygen is required during hard work or exercise, the partial pressure difference across the blood-air barrier increases and thereby also the rate of oxygen uptake.

9.4 Tidal volume and vital capacity

The inspiratory and expiratory air volume of a test person can be measured with the help of a spirometer that records gas volumes as a function of time. In this test the test person is asked to breathe normally at rest, then to inhale as much as possible, and to exhale to the maximum. The corresponding air volumes are collected and measured. Figure 9.6 shows a schematic chart of such a recording.

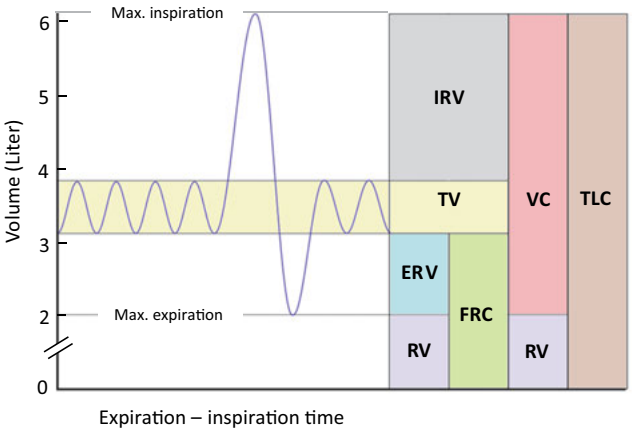


Fig. 9.6: Various volume capacities of the lung. TV = tidal volume; IRV = inspiratory rest volume; VC = vital capacity; ERV = expiratory rest volume; RV = residual volume; FRC = fractional rest capacity; TLC = total lung capacity.

At rest the inspiration-expiration volume is called the *tidal volume* (TV). This is on average about 0.5 l, independent of sex. When inhaling to absolute maximum capacity, the *inspiratory rest volume* (IRV) is filled. This is the volume beyond the tidal volume, which is about 3 l for men and 2 l for women. Conversely, upon expiration to the absolute minimum, the *expiratory reserve volume* (ERV) is reached (1.2 l for men, 0.7 l for women). The residual volume (RV) cannot be measured by inspiration-expiration, but must be either estimated by mixing air with a nontoxic gas such as He or can be imaged by special MRI methods discussed in Chapter 15. Typically RV is about 1.5–2 l. It is

important to maintain a residual gas volume to prevent the lungs from collapsing and for facilitating expansion of the lung during inspiration. These four main volumes, TV, IRV, ERV, and RV do not overlap. The remaining volume definitions follow from addition and subtraction, as can be seen by the colored areas in Fig. 9.6. These are as follows:

VC = vital capacity = ERV + TV + IRV; this is the capacity from maximum expiration to maximum inspiration. For an adult healthy person VC is about 3 l (women) to 4.5 l (men).

FRC = fractional residual capacity = ERV + RV.

TLC = total lung capacity = VC + RV.

These lung capacities are listed in Tab. 9.4 separately for men and women.

Tab. 9.4: Typical inspiratory-expiratory volumes of the lung. All numbers are in units of liter.

	Women	Men
TV	0.5	0.5
IRV	1.8	2.8
ERV	0.7	1.2
VC	3.0	4.5
RV	1.5	2.0
TLC	4.5	6.5

A healthy person can exhale 70 % of the vital capacity within 0.5 s, another 15 % in 1 s, and a total of 97 % within 3 s. During this exercise the volume flow rate is 6–8 l/s, a value that we will need later on. The flow velocity can be as high as 0.5–1 in units of the Mach number.

The *breathing frequency* (BF) of an adult is about 16 breathes per minute. For an infant the breathing rate is about 40/minute. Taking a tidal volume (TV) at rest of about 0.5 l, we determine the minute ventilation (MV) as MV = BF × TV = 8 l/min. This makes 11 500 l of inhaled air per day. It turns out that this is the same amount of inhaled air in liters as blood is pumped by the heart per day.

According to Chapter 4 the *caloric oxygen equivalent* is 1 l of O_2 for generating 20 kJ of energy. Per day the energy requirement is about 8 MJ at rest, or 400 l of oxygen/day.

From the 11 500 l air/day only 13 % arrives as oxygen in the alveoli, which amounts to about 1500 l/day. However, only about 30 % of the oxygen is really exchanged, which corresponds to 430 l oxygen/day. This agrees well with the metabolic requirement at rest. With increasing activity level there is sufficient oxygen reservoir to respond to needs.

9.5 Pulmonary volume and pressure changes

Lung, diaphragm, and thorax form a unity during breathing. The lung sits in the *thoracic cage* separated from the chest wall by an intrapleural space also called *pleural cavity* (see Fig. 9.3). As described above, this cavity consists of a folded membrane, where the visceral pleura covers the lung and the parietal pleura covers the inner thoracic wall. The interaction of the chest and the lung across the intrapleural space determines the lung volume. The volume change during respiration is dramatic, as we notice from the schematics in Fig. 9.7. During inspiration the diaphragm contracts and flattens and the intercostal muscles elevate the ribs and the sternum. The largest volume expansion of the thoracic cage results from movement of the diaphragm. During expiration the diaphragm and the intercostal muscles relax and the elastic properties of thorax and lungs allow a passive reduction of the thoracic cage volume. In addition, the expiratory contraction is supported by contraction of abdominal muscles.

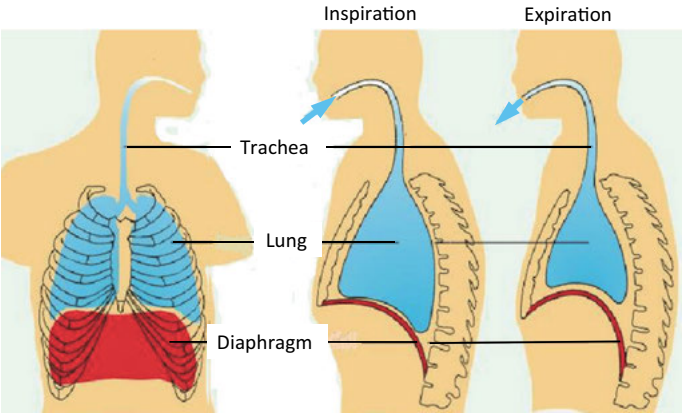


Fig. 9.7: Volume change of the thoracic cavity during respiration.

As the thoracic volume varies during respiration, the gas pressure changes in the lung. Figure 9.8 displays a simple mechanical model for the lung that represents all relevant pressures and forces. We start with the resting position, defined as the position where the gas pressure in the lung is the same as in the atmosphere, referred to as *barometric pressure*. Barometric pressure is the one present in the environment irrespective of the actual altitude above sea level. In the resting position the lung is expanded and would normally contract and eventually collapse if not attached to the thoracic wall. The thorax is compressed and would normally expand if not attached to the lung. The force F_L that pulls the lung together and the force F_T that expands the thorax point in opposite directions and keep a balance (red arrows in Fig. 9.8). This force pair pulls on either side of the membrane separating thorax and lung, thereby slightly in-

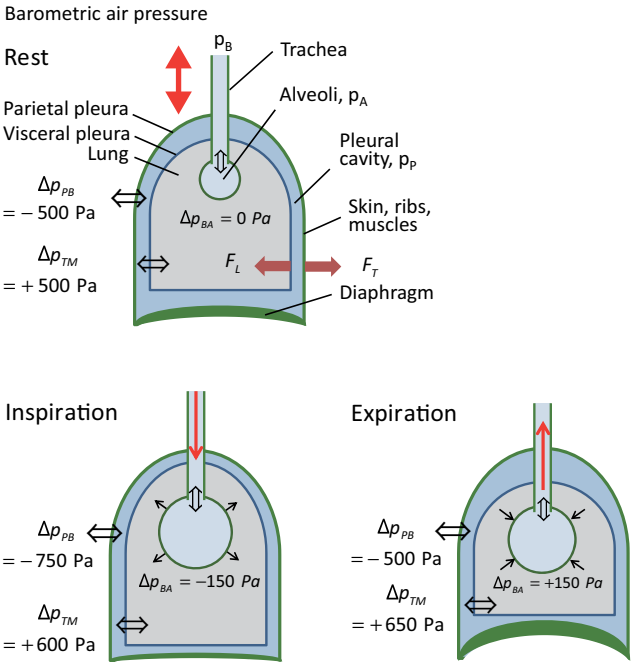


Fig. 9.8: Pressures and pressure differences at rest and during inspiration and expiration, demonstrated with a simple mechanical model. p_B = barometric pressure, p_A = gas pressure in the alveoli, p_P = pressure in the pleural cavity. Δp are the respective pressure differences.

creasing the intrapleural space and causing a reduction of the intrapleural pressure difference Δp_{PB} to −500 Pa with respect to barometric pressure. The transmural pressure difference between alveolar space in the lung and pleural cavity Δp_{TM} follows from $\Delta p_{TM} = \Delta p_{BA} - \Delta p_{PB} = +500$ Pa, where p_{BA} is the pressure difference between alveolar space and the environment. This is the pressure exerted on the visceral wall from the inside.

During inspiration the thorax expands due to the action of the intercostal muscles and due to the diaphragm pulling downwards. This increases the thoracic cage volume. As the lungs adhere to the thoracic wall, the lung volume will also expand. Furthermore, the elastic forces between lung and thorax slightly increase the intrapleural space such that the intrapleural pressure difference Δp_{PB} drops from −500 Pa to −750 Pa with respect to barometric pressure. As the pleural pressure decreases and the lung expands, so do the alveoli inside the lung. This expansion causes a pressure drop in the alveoli with respect to the environment to about $\Delta p_{BA} = -150$ Pa. The transmural pressure difference $\Delta p_{TM} = \Delta p_{BA} - \Delta p_{PB}$ goes up to about +600 Pa during inspiration. Because of this inspiratory pressure drop in the alveolar space and increased volume, air will flow into the lung until the pressure difference is balanced.

During expiration the volume of the thoracic cage decreases beyond the resting position. Therefore, in the alveoli an expiratory overpressure $\Delta p_{BA} = +150\,\text{Pa}$ develops and air flows out of the lung through the trachea into the environment. The other corresponding pressure differences are indicated in Fig. 9.8.

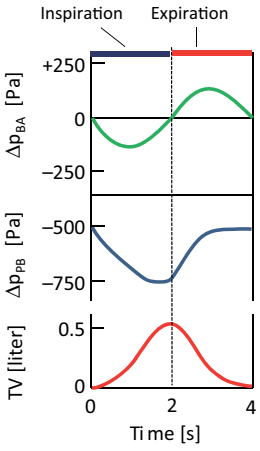


Fig. 9.9: Pressure changes and tidal volume during one inspiration-expiration cycle. Δp_{BA} is the pressure difference between the barometric pressure outside the body and alveolar air in the lung; Δp_{PB} is the pressure difference between the pleural cavity and barometric pressure. TV is the tidal volume.

The alveolar pressure difference Δp_{BA}, the interpleural pressure difference Δp_{PB}, and the tidal volume (TV) are plotted as a function of time in Fig. 9.9 for one inspiration-expiration cycle, assuming that the breathing frequency is $0.25\,\text{s}^{-1}$ or 15 breaths per minute. An overview of the three main pressure differences active during respiration is given in Tab. 9.5.

The volume work done by the thorax during respiration can be represented in a pV-diagram as shown in Fig. 9.10. The enclosed area corresponds to the volume work W_{Lung} during respiration at rest. This volume is equivalent to the tidal volume and the pressure refers to the intrapleural pressure difference Δp_{PB}. The work performed can be estimated as $W_{Thorax} = \frac{1}{2}(\Delta p_{PB} \times TV)$. But actually during expiration the thorax

Tab. 9.5: Overview of different pressure difference active during respiration.

Pressure difference	Main function
Δp_{BA}	Alveolar pressure difference, responsible for transport of gas into and out of the lung through the airways.
Δp_{PB}	Intrapleural pressure difference, responsible for maintaining an underpressure in the intrapleural space.
Δp_{TM}	Transmural pressure difference, responsible for inflating the alveolar sacs during inhaling.

recoils elastically from the expanded volume to the relaxed volume, which essentially costs no energy. Therefore, the volume work done by the thorax should be rather: $W_{\text{Thorax}} = \frac{1}{4}(\Delta p_{\text{PB}} \times \text{TV})$. This amounts to $W_{\text{Thorax}} = 0.25 \times 200\,\text{Pa} \times 0.5\,\text{l} = 0.025\,\text{J}$. With a breathing frequency BF = 15 per minute, the mechanical power consumption is only 6 mW. Even if we consider a mechanical efficiency of 20 %, the power consumption of the respiratory system is still negligibly low thanks to the elastic recoil properties of the thorax.

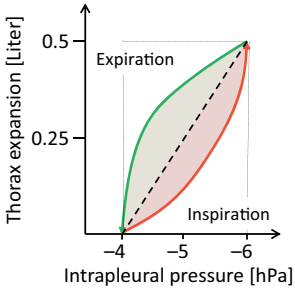


Fig. 9.10: Volume work of the lung.

9.6 Compliance

The ease of expanding the lung is also expressed by the high compliance of the lung C_{Lung}. The compliance of the lungs and the thorax is defined as the volume change ΔV by which the lungs and the thorax expand for each increment of pressure change Δp_{BA} in the alveoli. The relation between pressure change and volume change of elastic bodies is usually expressed in terms of Hooke's law (see Chapter 3):

$$\Delta p = B\frac{\Delta V}{V}.$$

Here B is the elastic modulus of the expandable medium. Rephrasing this equation we find:

$$\Delta V = \frac{V}{B}\Delta p = C\,\Delta p.$$

The ratio $C = V/B = \Delta V/\Delta p$ is called the elastic compliance. It can be considered as a mechanical susceptibility, where ΔV is the extensive property and Δp is the conjugated field to the volume change ΔV.

Figure 9.11 shows a plot of lung volume versus *transmural pressure difference* Δp_{TM}. The local first derivative refers to the compliance of the lung. Obviously the compliance is not a constant. The compliance is large if for a small pressure change the volume change is large. This is the case at the beginning of the green curve. However, for the same volume change an increasing pressure increment is required, i.e., the compliance decreases with inspiration beyond the TV region into the IRV region, as to be expected for saturation.

Figure 9.11 also shows the volume-pressure relationship for two frequent lung diseases. *Emphysema* is a disease where surfactants and cell membranes of alveoli are irreversibly destroyed, for instance by smoking (see next section). This leads to a saggy type of lung that lacks substantial elastic recoil (blue line). Conversely, *fibrosis* is a condition where the cell walls of the alveoli harden. Then the lung is much less elastic and compliance is strongly decreased (red line).

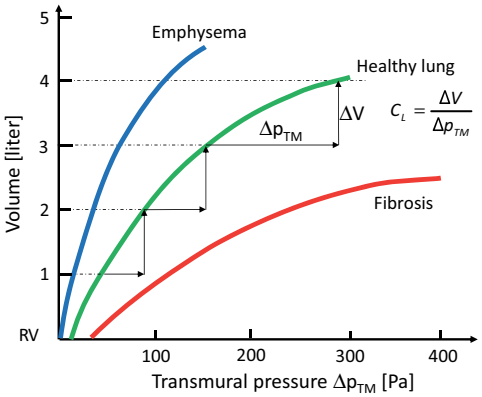


Fig. 9.11: Compliance of the lung for a healthy person (green line), for a patient with emphysema (blue line), and for a patient with fibrosis (red line).

The pressure volume relationship is different for the thorax alone and the lung alone, as can be seen in Fig. 9.12. Thorax and lung have opposite elastic properties. The compliance of the thorax C_T is small at the beginning and increases with expansion, similar to the compliance of an air balloon. In contrast, the compliance of the lung C_L is large to begin with but goes to zero when approaching saturation, as we have just seen.

The combined compliance of thorax and lung is shown by the green line in Fig. 9.12. In a mechanical model shown in Fig. 9.13 both lung and thorax can be represented by coupled springs with an equilibrium rest position corresponding to the fractional rest volume (FRC). When released from their coupling, the spring represent-

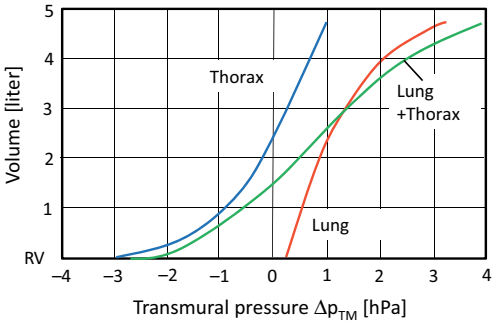


Fig. 9.12: Compliance of thorax and lung alone and of the combined organs.

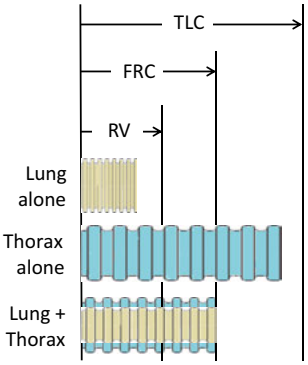


Fig. 9.13: Mechanical spring model representing lung and thorax.

ing the lung will contract to a volume less than the residual volume RV, whereas the spring representing the thorax will expand to a volume below the total lung capacity.

Because lung and thorax are arranged in parallel, the reciprocal values of the compliances, known as *elastances*, add up similar to parallel ohmic resistances:

$$\frac{1}{C_{\text{tot}}} = \frac{1}{C_{\text{L}}} + \frac{1}{C_{\text{T}}}.$$

Now we try to estimate both compliances independently. The compliance of the lung is defined by:

$$C_{\text{L}} = \frac{\Delta V}{\Delta \hat{p}_{\text{TM}}}.$$

Here we set $\Delta V = \text{TV}$. The transmural pressure difference was defined before as $\Delta p_{\text{TM}} = \Delta p_{\text{BA}} - \Delta p_{\text{PB}}$. However, for estimating the compliance of the lung we need to take the difference $\Delta \hat{p}_{\text{TM}} = \Delta p_{\text{TM,f}} - \Delta p_{\text{TM,i}}$, which refers to the transmural pressure at the plateau region (after inhaling is completed and breathing has stopped for a moment) and the transmural pressure at the beginning of inhaling. This transmural pressure difference from the beginning to the end of inhaling is 150 Pa. Therefore the compliance of the lung can be estimated as:

$$C_{\text{L}} = \frac{\Delta V}{\Delta \hat{p}_{\text{TM}}} = \frac{\text{TV}}{\Delta \hat{p}_{\text{TM}}} = \frac{0.5\,\text{l}}{150\,\text{Pa}} \approx 3.3 \times 10^{-6}\,\text{m}^3/\text{Pa}.$$

The compliance of the thorax, defined by

$$C_{\text{T}} = \frac{\Delta V}{\Delta \hat{p}_{\text{PB}}},$$

can be estimated in a similar fashion by considering the intrapleural pressure difference at the start and at the end of inhaling. The pressure difference is 250 Pa:

$$C_{\text{T}} = \frac{\Delta V}{\Delta \hat{p}_{\text{PB}}} = 2 \times 10^{-6}\,\text{m}^3/\text{Pa} = 2\,\text{l/kPa}.$$

The total compliance is therefore:

$$C_{\text{tot}} \approx 1.25 \times 10^{-6}\,\text{m}^3/\text{Pa} = 1.25\,\text{l/kPa}.$$

Often the total compliance is estimated by considering the thorax alone, setting:

$$C_{\text{tot}} \approx C_{\text{T}} = \frac{\Delta V}{\Delta \hat{p}_{\text{PB}}} \approx 2\,\text{l/kPa}.$$

Typical values for adults are 1–2 l/kPa. Measurements of compliance can indicate diseases of the lung, as we have already seen for the cases of emphysema and fibrosis.

9.7 Surface tension

Now we will take a more microscopic view on the alveoli and their compliances. A good model system for the alveoli is a soap bubble (Fig. 9.14 (a)). The force F_1 on the rim of a soap bubble tends to increase the surface (factor of two because of two surfaces, inside and outside):

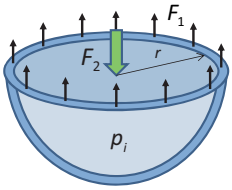
$$F_1 = 2\gamma \cdot 2\pi r.$$

Here r is the radius and the proportionality factor γ is the surface tension. This equation holds if the force required to increase the surface is constant and independent of the radius. On the other hand, the force F_2 on the equatorial plane with area πr^2 due to the inside gas pressure p_{i} is:

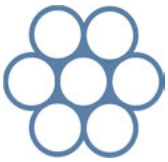
$$F_2 = p_{\text{i}} \pi r^2.$$

In equilibrium both forces must balance and therefore we obtain for the internal pressure p_{i}:

$$p_{\text{i}} = \frac{4\gamma}{r}.$$



(a)



(b)

Fig. 9.14: (a) Model of a soap bubble; (b) model of alveolar sacs. Alveoli surrounded by other alveoli have only one inside surface.

The lung of an adult has about $3–6 \times 10^8$ alveoli. Each alveolus has a diameter of 50 µm at expiration, which expands to about 250 µm at inspiration. The wall thickness is about 0.5 µm. These thicknesses can be obtained by microscopy of histological cuts through the lung tissue.

The surface tension of the alveoli γ tends to reduce the diameter and to contract the lung. During inspiration the gas pressure in the alveoli must equilibrate the pressure due to the wall tension:

$$p_{\text{i}} = p_{\text{wall}} = \frac{2\gamma}{r}.$$

The factor 2 missing is due to the fact that the alveoli have only one inside surface, unlike soap bubbles, as indicated in Fig. 9.14 (b).

The surface tension of water is $\gamma = 0.072\,\mathrm{N/m}$. For the transmural pressure difference we then obtain:

$$\Delta p_{\mathrm{TM}} = \frac{2\gamma}{r} = \frac{2 \times 0.072\,\frac{\mathrm{N}}{\mathrm{m}}}{25\,\mu\mathrm{m}} = 5.75\,\mathrm{kPa}.$$

This is by far greater than is actually observed. The actual transmural pressure difference is only a tenth of the calculated one: $p_{\mathrm{TM}} = 500\,\mathrm{Pa}$ instead of 5000 Pa. The reduction of the surface tension in the alveoli is due to surfactants in the wall membrane. The main component of this surfactant is the molecule dipalmitoylphosphatidylcholine (DPPtdCho), which can lower the surface tension from 72 mN/m to almost zero. Any damage to the pulmonary surfactant increases the transmural pressure and leads to dyspnea. In newborns the pulmonary surfactant is not completely developed, which may lead to an *infant respiratory distress syndrome*.

Alveoli with smaller radii have higher internal pressure than larger ones because the internal pressure is proportional to the reciprocal radius: $p \propto 1/r$. Air may then be pushed out of the smaller alveoli, as we experience this from soap bubbles, and the smaller alveoli would completely collapse (Fig. 9.15). However, the surfactants not only lower the surface tension but simultaneously hinder gas exchange. In the case that the surfactants are depleted or damaged, entire areas in the lung may collapse, which is known as *atelectasis*.

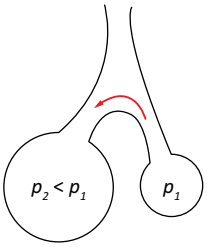


Fig. 9.15: Pressure in small bubbles is greater than in larger bubbles because $p \propto 1/r$.

One of the major dangers to surfactants is tobacco smoke. Tobacco smoke not only destroys the surfactants but also clogs up the airways and increases the airway's resistance. The result is potential emphysema, chronic bronchitis, and lung cancer.

9.8 Airway resistance

When breathing, air flows through the airway into and out of the lung. The volume flow rate of air is defined as:

$$\dot{V} = \frac{\Delta V}{\Delta t} = \mathrm{TV} \times \mathrm{BF},$$

where TV is the tidal volume and BF the breathing frequency at rest. During breathing we inhale and exhale the same amount. Therefore the tidal volume of 0.5 l has to be taken twice for determining the flow rate. Assuming BF to be 15/minute, the flow rate is about 0.25 l/s or 15 l/min (recall that the minute ventilation MV is half this amount). Knowing the flow rate, we can now determine the total airway resistance by applying Ohm's law:

$$\dot{V} = \frac{\Delta p_{BA}}{R_R}.$$

Here Δp_{BA} is the pressure difference between the alveolar sacs and barometric pressure outside. R_R is the total respiratory resistance, including trachea, bronchus and all further branches before reaching the alveolar space. For the total airway resistance we estimate:

$$R_R = \frac{\Delta p_{BA}}{\dot{V}} = \frac{150\,\text{Pa}}{0.25\,\text{l/s}} = 600\frac{\text{Pa}}{\text{l/s}}.$$

This is a rather small value confirming our impression that breathing half a liter of air in just two seconds with a mere pressure difference of 150 Pa is rather easy and therefore the flow resistance must be low. To set this number in perspective, we directly calculate the flow resistance of different tubes, representing trachea, bronchus, and the alveoli. Assuming laminar flow and cylindrical shaped tubes, the flow resistance according to Hagen–Poisseuille is:

$$R = \frac{8}{\pi}\eta\frac{\Delta L}{r^4} = 8\eta\pi\frac{\Delta L}{A^2}.$$

The viscosity of air at 20 °C is roughly 17×10^{-6} Pa s. Typical values for radius, cross sections, and lengths of trachea, bronchial tube, and alveoli are given in Fig. 9.16 and in Tab. 9.6. With this information we can calculate the flow resistance of different parts of the airway, the values are also listed in Tab. 9.6.

According to Tab. 9.6 the flow resistance of the trachea is very small. Also the bronchial tubes do not contribute much to the flow resistance. The main resistance, according to the table values, is due to the alveoli. However, these numbers are deceptive and it turns out that the interpretation of the flow resistance is too simple.

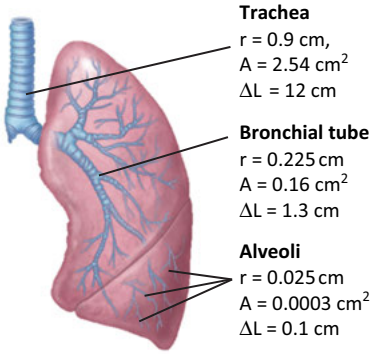


Fig. 9.16: Airway and lung with typical values for radius, cross section, and length of trachea, bronchial tube, and alveoli, respectively.

Tab. 9.6: Geometric data for parts of the airway in the respiratory system and respective flow resistances.

	Radius [cm]	Cross section [cm^2]	Length [cm]	R [Pa s l^{-1}]
Trachea	0.9	2.5	12	0.8
Bronchi	0.225	0.16	1.3	17
Alveoli	0.025	0.002	0.1	10^4

First, airflow can easily become turbulent because of the low viscosity. The trachea and even the bronchial tree have radii that could make the airflow turbulent. Therefore we calculate the Reynolds number for the airflow according to:

$$Re = \frac{\rho v d}{\eta}.$$

Here v is the flow velocity, ρ is the density, d the diameter of the conducting system, η is the viscosity. The dimensionless number should be smaller than 1000 to warrant laminar flow. A value above 2000 is indicative of turbulent flow. The density of air is roughly 1.2 kg/m^3, the flow velocity is $v = 1$ m/s and follows from the relation:

$$v = \frac{\dot{V}}{A},$$

where A is the cross section of the tube. Inserting the numbers listed above, we obtain for the trachea a Reynolds number of 1200, for the bronchial tree 320, and for the alveoli about 15. From this we conclude that in the trachea the flow is potentially turbulent, whereas in the bronchial tubes and in the bronchioles the flow can safely be assumed to be laminar. Turbulent flow dramatically increases airflow resistance.

Second, we have to consider that due to branching the total cross section substantially increases from the trachea to the bronchioles. The increasing flow resistance due to the decreasing tube diameter is more than compensated by an increasing number of divisions, reaching 100 million at the terminal bronchioles with an area of about 80 m^2. Accordingly, the flow resistance decreases because of parallel conductance of all branches. The situation is very similar to the flow resistance in the capillary bed discussed in Chapter 8.

Figure 9.17 shows a plot of airway resistance as a function of distance starting from the trachea. Contrary to our first impression, the resistance is highest in the area of the trachea and then continuously drops off with increasing cross-sectional area. The high resistance at the beginning is due to turbulent flow in the trachea and it even increases towards the first division into the primary bronchi. At this junction, turbulent flow dominates. As soon as the flow becomes laminar, the resistance drops off and quietens down. This area is therefore known as the “quite zone” of the lung.

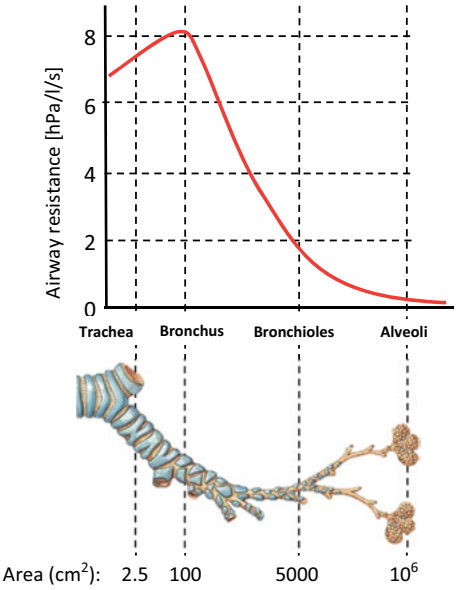


Fig. 9.17: Airway resistance in different parts of the respiratory system. At the bottom of the panel the cross-sectional area of the airways is indicated as it follows from the cross section of an individual branch times the number of branches.

To conclude, the airway resistance of the lung is mainly due to turbulent flow in the trachea. Less than 20 % of the total resistance is accounted for by the bronchioles and their terminal branches. Thus the airway resistance distribution is the opposite of what we initially concluded.

9.9 Cardiopulmonary bypass

Surgery on the heart, the pulmonary arteries, or the large vessels is ordinarily not possible with a beating heart. For such surgical procedures the heart and lungs are taken out of the circulator system and put to rest. Their functions are then temporarily taken over by an *extracorporal circulation* (ECC), also called *cardiopulmonary bypass* (CPB) or extrapulmonary ventilation.

Extracorporal circulation is required in the case of surgery to the surface or inside of the heart, in particular for repair or replacement of cardiac valves, correction of hereditary heart failures, fixing coronary artery bypasses, heart replacement, removal of pulmonary thrombosis, and for any surgery at the aorta.

For the duration of the surgery, which may take several hours, the cardiac and pulmonary activity is intentionally and temporarily immobilized. This artificial cardiopulmonary arrest is achieved by injecting cardioplegish substances, such as K^+, Mg^+ acetylcholine, neostigmine, and by injecting a sodium-free solution. Na^+ ion current is the most dominant ion current in the heart, compare Fig. 7.3. Reducing the Na^+ ion concentration will immediately stop the rapid depolarization starting at the sinoatrial node.

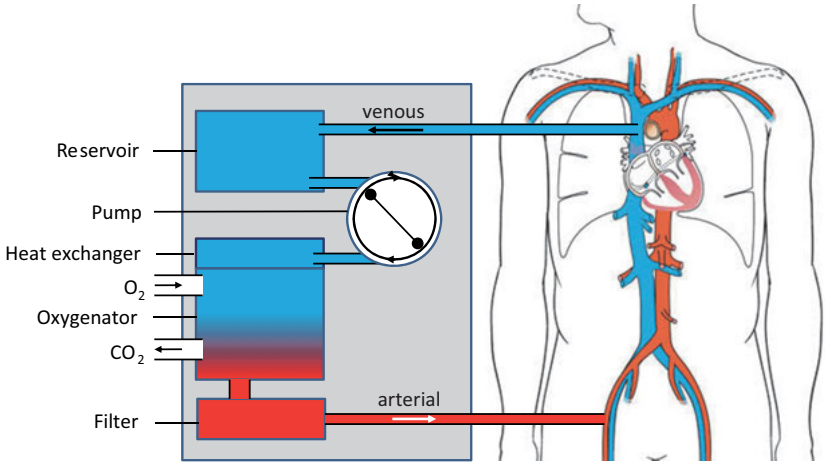


Fig. 9.18: Schematic overview of the main components of a heart-lung machine.

CPB is a method that bypasses either the entire blood circulation or part of it and diverts the blood into a machine, called a *heart-lung machine* (HLM) that takes over circulation at the proper frequency and blood pressure, provides gas exchange (O_2/CO_2), and filters the blood [1]. In the end the blood is returned at body temperature and volume rate to the arteries. The HLM was first used by the physician John Gibbon in 1953 and has since been applied successfully in numerous surgical procedures on the heart, at least half a million per annum worldwide.

The main components of the HLM as indicated in Fig. 9.18 are as follows:

1. Reservoir for collecting the blood from the veins;
2. Roller pump taking over the function of the heart;
3. Heat exchanger controlling the blood temperature;
4. Oxygenator taking over the function of the lung;
5. Filter for the blood.

These components are described in more detail as follows.

Before connecting the patient to a CPB circuit, the entire circuit must be expunged from all air in any parts and in particular in the arterial cannula that are connected back to the patient.

A reservoir is used to collect the venous blood from the vena cava through gravity or by vacuuming off the blood. Also blood that surfaces during surgery is sucked of and collected in the reservoir. Various filters and a defoamer to remove debris and gas bubbles are integrated in the reservoir. Also medication is added, such as anticoagulants to prevent clotting of the blood in the circuit, and cardioplegia to suppress heart beats and lower myocardial metabolism in order to protect heart tissue.

The pump in the HLM serves the purpose of maintaining circulation with the proper pressure and volume flow rate, and for circulating any medication. The requirements for pumps in the HLM are stringent. They must operate reliably and as simply as possible without damaging blood cells, while providing an exact pumping power and pressure. Two types of pumps are generally used: roller pumps and centrifugal pumps. Roller pumps push the blood in a pulsed fashion through flexible tubes. Centrifugal pumps produce blood flow by centrifugal forces instead of mechanically squeezing the blood. Therefore they are considered to be superior to roller pumps and believed to produce less blood damage. Both working principles are schematically shown in Fig. 9.19.

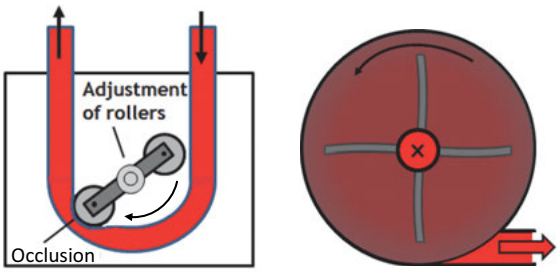


Fig. 9.19: Working principle of roller pump (left) and centrifugal pump (right). In the centrifugal pump blood flows in from the top and leaves on the side after being circulated by a wheel in a round box.

The heat exchanger has the task of lowering the temperature of the patient during surgery and of warming up after termination of hypothermia. Body cooling allows more time for the surgery without causing brain damage. The heat exchanger is also used for surgeries at normal temperatures to compensate for any heat losses.

During reheating of the blood the gaseous solution decreases and microbubbles are formed. Therefore in the CPB the heat exchanger comes first before the oxygenator, otherwise the formation of microbubbles during reheating would even be bigger.

The oxygenator is the most important part of the HLM. The oxygenator takes over the task of the lung: supply of oxygen and removal of CO_2. In the oxygenator the blood flows in the opposite direction to the gas flow separated by microporous fibers. The gas is an air-oxygen mixture. Gas exchange occurs at the fiber membrane because of a concentration gradient, which is the same principle as in the lung. The deoxygenated dark red blood enters the oxygenator and leaves the oxygenator with a bright red color after having been enriched with oxygen. The filter membrane consists of microporous polypropylene carbon fibers. The schematic of the gas exchanger is shown in Fig. 9.20.

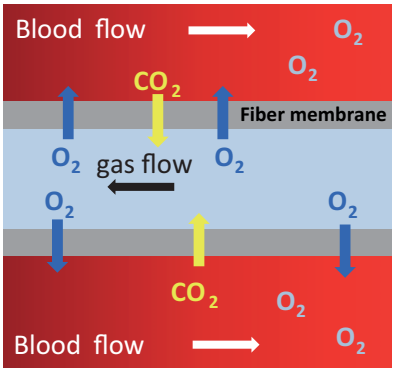


Fig. 9.20: Reoxygenation of blood is achieved by diffusion of oxygen across a porous membrane between a tube with O_2 enriched air and blood flowing in opposite directions.

Further filters after the oxygenator are inserted in the HLM to protect the patient from air bubbles and other corpuscular debris, which may cause embolism. However, the lower limit of pore size in the filter is determined by the size of the erythrocytes and by the maximum flow resistance that can be allowed for maintaining the required flow rate. Therefore the pore size is usually 20–40 µm.

Reanimation of the heart takes place when the heart again comes into contact with blood after opening the clamp at the aorta. It takes another 30–60 minutes for recovery and before the heart can again take over its self-stimulated pumping ability.

Surgeries requiring CPB are the most severe surgeries and can cause a number of associated problems, such as hemolysis, clotting of blood in the circuit, air embolism, and respiratory problems, etc. Problems with respect to the biocompatibility of tubes and walls within the HLM might lead to blood clotting. This problem could be partially overcome by using rough inside walls in all parts that have contact with blood. The adhesion and sticking of blood to the walls is enhanced by rough walls so that the circulating blood will only be in contact with a sacrificial layer of blood adsorbed at the rough surface.

In recent years so called miniaturized heart-lung machines (mini-CPB) or miniaturized extracorporeal circulation (mECC) have been developed for emergency applications and postoperative treatment. Studies have shown that these miniaturized machines also have a positive effect on reducing inflammatory, coagulopathic and hemodilutional side effects in comparison to normal HLMs for cardiopulmonary bypass operations [2]. Because of the positive experience gained so far, mCPBs have the potential to replace normal size HLMs in the long run. A comparative study of both techniques is published in [3].

9.10 Summary

1. The respiratory tract consists of pharynx, larynx, trachea, bronchi, bronchioles, and terminal alveoli.
2. The task of the lung is O_2-CO_2 gas exchange.
3. Gas exchange is controlled by perfusion, ventilation, and diffusion.
4. Gas exchange takes place in the alveolar space, which provides a surface of about 80 m^2.
5. CO_2 transport in blood is very different from O_2 transport.
6. Gas exchange occurs rapidly across the blood-air barrier and follows the concentration gradient via diffusion.
7. The gas exchange rate is about 250 ml O_2/min or 360 l/day.
8. The alveolar pressure difference is responsible for transport of gas into and out of the lung through the airways.
9. The intrapleural pressure difference is responsible for maintaining an underpressure in the intrapleural space.
10. The lung works in a rhythmic fashion via underpressure and overpressure for inhaling and exhaling air, respectively.
11. The mechanical power required by the lung is very small because of the elastic recoil properties of the thorax.
12. The tidal volume is 0.5 l, the breathing rate is about 15 breaths per minute.
13. Surfactants in the alveolar membrane decrease the tension by a factor of 9. Without surfactants breathing would not be possible.
14. Lung and thorax form a joint elastic system. At rest and in equilibrium the lung is expanded and would contract without coupling to the thorax, while the thorax is compressed and would expand without coupling to the lung.
15. The compliance of the lung is high but decreases with increasing lung volume.
16. The total compliance of thorax and lung is evaluated by adding their reciprocal values.
17. The airway resistance in the bronchial tree is lower than in the trachea in spite of smaller individual cross sections because of laminar flow and an increasing number of parallel airways.
18. The air flow in the trachea is partially turbulent but becomes laminar in the bronchus.
19. Heart-lung machines take over the pumping activity of the heart and the gas exchange of the lung during surgery on the heart or large vessels.

References

[1] https://en.wikipedia.org/wiki/Cardiopulmonary_bypass
[2] Harling L, Punjabi PP, Athanasiou T. Miniaturized extracorporeal circulation vs. off-pump coronary artery bypass grafting: what the evidence shows? Perfusion. 2011; 26: Suppl 40–47.
[3] Pereira SN, Balta Zumba I, Sulzbacher Batista M, Da Pieve D, dos Santos E, Stuermer R, Pereira de Oliveira G, Senger R. Comparison of two technics of cardiopulmonary bypass (conventional and mini CPB) in the trans- and postoperative periods of cardiac surgery. Braz J Cardiovasc Surg. 2015; 30: 433–442.

Further reading

Boron WF, Boulpaep EL. Medical Physiology. 2nd edition. Saunders W.B. Elsevier; 2012.

10 Kidneys

10.1 Introduction

A widespread misconception contends that whatever we drink will finally end up in the bladder and be disposed of as urine. However, this notion is much too simple. What we drink goes through the normal track of the digestive system, i.e., from esophagus to the stomach into the intestine. From there part of the fluid is cleared through the intestinal wall into the blood circulation and to the liver for processing, the rest continues through the intestines to the fecal exit. Blood is continuously filtered by the kidneys and water in the plasma that is no longer needed will finally be disposed of into the urinary tract. Therefore we can conclude that anything that is cleared by urinary excretion was previously in the blood. But the reverse conclusion does not hold, because water and other waste products also go out through the intestines and the lung. In fact, the kidneys are only one of three emunctories that the body maintains for disposal of waste. The other two are the lungs for the disposal of CO_2 and water, and the intestines/colon/anus for the disposal of feces.

The kidneys are a multitasking organ for maintaining an array of body functions:

1. Regulation of water level in the body;
2. Control of electrolytes and pH balance;
3. Removal of metabolites from blood and excretion of urine;
4. Synthesis of hormones to support the endocrine system (erythropoietin, renin, vitamin D3).

Some functions of the kidneys related to filtration, diffusion and perfusion, including discussions of malfunctions and remedies such as dialysis are treated in the next paragraphs. The main focus will be on renal clearance.

10.2 Global characteristics of kidneys

The kidneys are located left and right of the spinal column, below the diaphragm. Note that the kidneys are slightly asymmetrically positioned between the 11th thoracic vertebra and 2nd lumbar vertebra for the left kidney, but one vertebra lower for the right kidney. The kidneys have the form of a large bean and they weigh between 80–160 g each, the left kidney being slightly heavier than the right kidney.

The blood volume rate of both kidneys is 1 l/min called *renal blood flow* (RBF). This is about 20 % of the cardiac output (CO) (about 5 l/min). Since only plasma is filtered but not hematocrit, actually 0.6 l/min of plasma runs through the kidneys. This volume is called the *renal plasma flow* (RPF). From these 0.6 l a mere 20 % or just 120 ml/min = 172 l/day are actually filtered, which is called the *glomerulus filtration*

DOI 10.1515/9783110372830-012

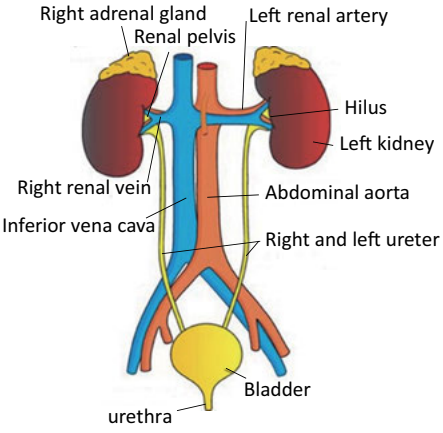


Fig. 10.1: Location of the kidneys and the ureter tract in the body (adapted from Wikimedia, © Creative Commons).

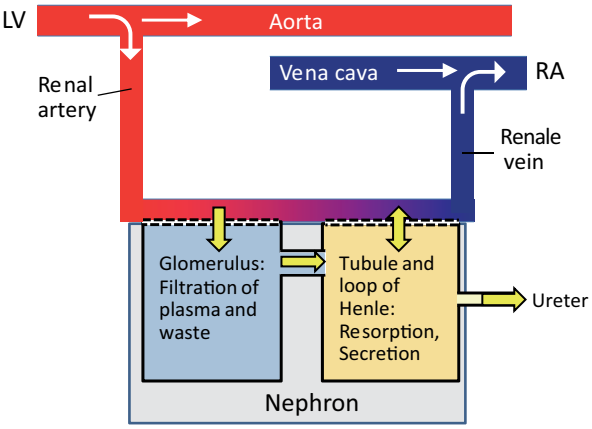


Fig. 10.2: Block diagram of the kidney parts and functions. LV = blood flow from left ventricle; RA = blood flow to right atrium.

rate (GFR). The ratio GFR/RPF is the *filtration fraction* (FF), which is 0.2 or 20 %. The production of *primary urine* is about 170–180 l/day. Most of this is, however, resorbed back into the venous blood stream and only 1–2 l/day of *secondary urine* is disposed of through the ureter to the bladder.

For understanding the complexity of the filtration process it is useful to consider first a simplified block diagram according to Fig. 10.2. Part of the arterial blood influx goes into the kidneys via the renal artery. The blood is then processed in *nephrons*. There are about a million nephrons in each kidney, all work in parallel. Each nephron consists of two main subsystems: *glomerulus* for filtration and *tubule* together with *loop of Henle* for reabsorption and secretion. The filtered and cleaned plasma is reabsorbed into the renal vein and from there to the vena cava. The end urine leaves the kidneys through the ureter tubes. The nephron processing of the plasma is highly

redundant. With a loss of 50 % of nephrons, full functionality is still guaranteed. The cleaning of blood plasma works according to the house cleaning principle: everything goes out of the house for cleaning and only useful items are allowed to be taken back in again, the rest is disposed of.

10.3 Structure of kidneys

Figure 10.3 shows a cross section through one of the kidneys. This image gives a first impression of the huge complexity of this organ. The main components are the influx of blood through the renal artery, the distribution through capillaries in the renal cortex and the medulla for filtration and reabsorption, and finally the outflow of filtered plasma via the renal vein on the one hand and urine through the renal papillae into the ureter on the other hand, leaving the kidneys at the renal pelvis. The most prominent features are the pyramidal shaped medullas which reach from the renal pelvis to the cortex. There are seven of those in the human kidney on either side.

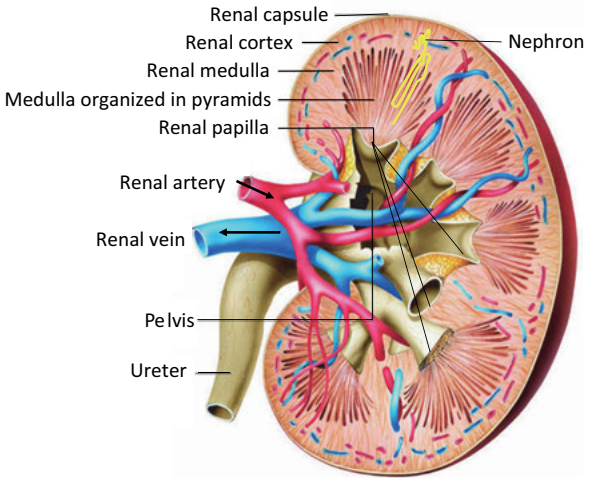


Fig. 10.3: Cross section of the kidney. Most prominent are seven pyramidal shaped medullas, containing nephrons for filtration of the blood. Only one nephron out of a million is sketched (reproduced from OpenStax Anatomy and Physiology, © Creative Commons).

The most important functional subunit in the renal medulla is the *nephron*. Nephron is the Greek word for kidneys, which is used nowadays only for the filtration subunit. It stretches from the cortex, where the filtration unit is located, to the lower part of the medulla, where reabsorption and secretion take place. The main parts of the nephron are schematically displayed in Fig. 10.4.

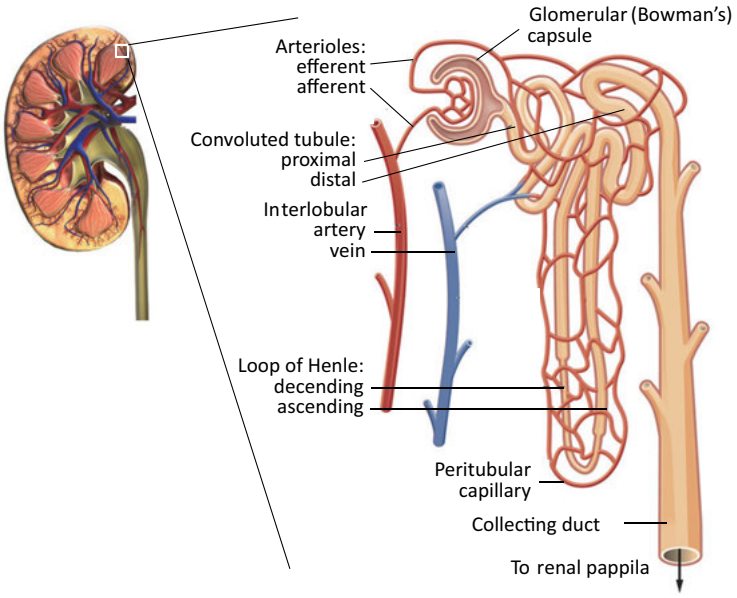


Fig. 10.4: Main parts of a nephron (adapted from Anatomy & Physiology, Connexions Web site. http://cnx.org/content/, © Creative Commons).

The nephron consists of a glomerular capsule also known as *Bowman's capsule*, proximal tubule, loop of Henle, distal tubule, and collecting duct for the urine. In the glomerular capsule the plasma is "pressed" out of the afferent arterioles, subsequently filtered as primary urine and resorbed in the proximal and distal tubules, the remaining filtrate being collected in the duct leading to the ureter. In the following we discuss these different parts separately.

10.4 Filtration

Filtration takes place in the glomerular capsule (Bowman's capsule), sketched in Fig. 10.5. The glomerulus located inside the capsules is a tiny ball-shaped structure composed of a dense network of arterioles. The arterioles have pores on the inside and outside for passing the plasma into the proximal tubule. For effective filtration a pressure difference between the renal afferent arteriole and the proximal tubule is required. If the *effective filtration pressure* p_{eff} is reached, the plasma is pressed out of the blood through narrow holes acting as a sieve.

Filtering is achieved by a three layer sieve system shown in Fig. 10.6: starting from the blood side there is first the *endothelial fenestration* (opening) of the glomerulus that prevents filtration of erythrocytes but allows all other components of the plasma to pass; then follows the *glomerular basement membrane* (GBM) stopping filtration

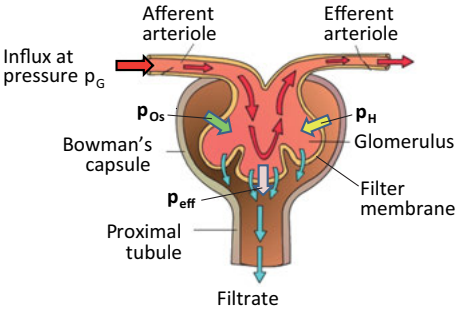


Fig. 10.5: Schematic of the Bowman's capsule containing arterioles and filters.

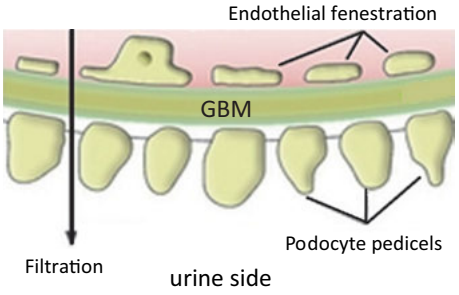


Fig. 10.6: Filter in the glomerulus consisting of three sieves with different pore holes.

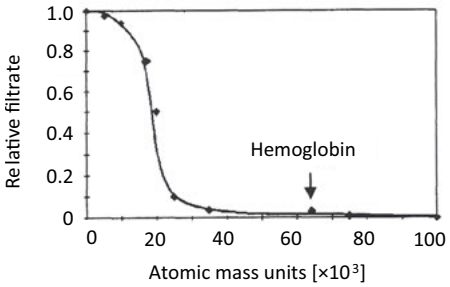


Fig. 10.7: Relative filtrate as a function of atomic weight.

of larger proteins. And finally there is a third membrane filter consisting of *podocyte pedicels* (little feet) that prevents filtration of medium-size proteins. Unlike ion channels in cell membranes, these three filters act as a mechanical sieve separating particles only by size but not by function or charge. In Fig. 10.7 the relative filtrate versus the molecular weight in atomic mass units (amu) is plotted. The *relative filtrate* is defined as the ratio of particle concentration in the plasma after filtration versus concentration in the plasma before filtration. Anything above 20 000 amu is not filtered out. In particular, the erythrocytes and hemoglobin do not pass the filter but smaller proteins can pass. For nanomedical applications discussed in Chapter 14/Vol. 2 it is important to know what the renal clearance of nanoparticles is with respect to size. Investigations

have shown that nanoparticles (NP) with a hydration diameter (HD) of 5 nm or less are filtered out whereas NPs with HD > 8 nm are not susceptible to glomerular filtration. In the intermediate size range filtration depends on the specific shape and surface composition of NPs [1]. In any case, the glomerular filtrate, also called primary urine, contains mainly water, glucose, salts, and urea ($CO(NH_2)_2$). About 180 l of glomerular filtrate is produced daily, but only 1–2 l is actually disposed of as urine. The other 99 % of the filtrate is resorbed, as we will discuss in the next paragraph. Table 10.1 gives an overview of the components of the filtrate in relation to the plasma.

Tab. 10.1: Components in the plasma in order of molecular weight, concentration, and filterability.

Component	Molecular weight [amu]	Concentration in plasma [mmol/l]	Filterability
Water	18	55 555	1
Sodium	23	135	1
Urea	60	3–7	1
Glucose	180	4–5	1
Inulin	5 200		1
Myoglobin	17 000		0.75
Hemoglobin	68 000		0.005

The *effective filtration pressure* is given by the blood pressure in the glomerular capillaries p_G, minus the hydrostatic pressure p_H in the Bowman capsule, and minus the osmotic pressure Π_{Os} from colloids in the plasma:

$$p_{eff} = p_G - p_H - \Pi_{Os}.$$

Using typical values for $p_G \approx 100$ hPa, $p_H \approx 20$ hPa, and $\Pi_{Os} \approx 40$ hPa, the effective filtration pressure p_{eff} is about 40 hPa. If p_{eff} dropped to zero, active filtration would cease and only diffusion through the filter would continue, which, however, is too slow by far. Both the *renal plasma flow* (RPF) and the *glomerular filtration rate* (GFR) depend critically on the blood pressure in the glomerulus p_G. If the pressure is too low, both rates will go down. If the pressure is too high, RPF and GFR will increase. In the intermediate plateau region from 100–230 hPa, RPF (≈ 600 ml/min) and GFR (≈ 120 ml/min) do not critically depend on p_G, see Fig. 10.8. This relative constancy of GFR and RPF is referred to as *autoregulation*. Autoregulation operates by changing the arterial vessel tension with a negative feedback system: increasing blood pressure will increase the tension thereby decreasing the vessel radius and the flow rate; with decreasing pressure the opposite reaction occurs. Furthermore, the kidneys are located at about the *hydrostatic indifference level*, where the blood pressure is independent of posture, standing up or lying down.

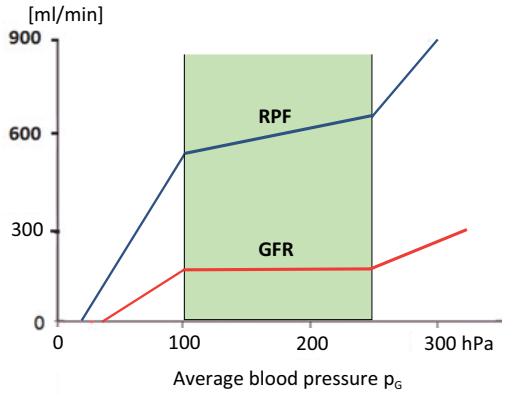


Fig. 10.8: Renal plasma flow (RPF) and glomerular filtration rate (GFR) as a function of blood pressure p_G in the glomerulus. Autoregulation occurs in the green area.

10.5 Reabsorption

In order to disentangle the complexity of the distal convoluted tubule and the peritubular capillaries depicted in Fig. 10.4, a simplified schematic is shown in Fig. 10.9 focusing on the functionality of the nephron system. As already discussed, much of the filtrate in the proximal convoluted tubule is reabsorbed into the blood within the proximal and distal tubular system. Figure 10.9 only shows the tubule for clarity, which, however, is strongly intertwined with arterioles for water and ion exchange as indicated in Fig. 10.4.

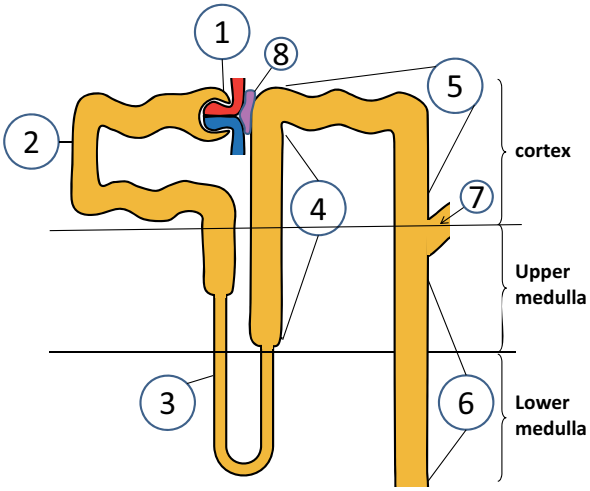


Fig. 10.9: Different parts of the nephron: (1) Bowman's capsule; (2) proximal convolute; (3) loop of Henle, descending part; (4) ascending part of the loop of Henle; (5) distal convolute; (6) collecting duct with connection to the ureter; (7) influx from neighboring nephrons. More details are explained in the text.

The different parts of the nephron have the following main functions, where numbers in the list refer to the numbers in Fig. 10.9:

1. *Bowman's capsule* contains the glomerulus with the arterioles that filter blood plasma into the proximal tubule, as shown in Fig. 10.5 and discussed already in the previous section.
2. *Proximal convolute tubule* is responsible for the reabsorption of most of the water and solvents into the efferent arterioles for delivering them back into circulation. In fact, 2/3 of the water in the filtrate or 80 ml/min is resorbed already in the proximal tubule. The rest is resorbed later, and only 0.1 % of the total filtration is finally excreted into the urine. Furthermore, glucose is completely resorbed by the capillary and also a large fraction of salt ions (Na^+, K^+, Ca^{2+}). The remaining filtrate is further processed in the distal tubule.
3. The *loop of Henle* consists of a thin water-permeable part, and a thicker water-impermeable part. The loop stretches needle-like from the cortex into the medulla and back again. The task of the thin part is the maintenance of a hypertonic environment in the medulla via opposing diffusional currents. This is a passive process that helps to stabilize an osmotic gradient in the medulla required for concentrating the urine.
4. In the thick part of the loop of Henle further reabsorption of the cations Na^+, K^+, and the anion Cl^- into the arterioles takes place, as well as reabsorption of divalent ions, such as Mg^{2+}. As ions leave the tubule while water cannot penetrate the wall, the osmotic pressure decreases leading to hypotonic conditions in the loop.
5. In the distal tubule reabsorption of ions into the arterioles continues. But with the absorption of K^+ ions into the tubule the conditions become less hypotonic and more isotonic.
6. In the collecting tube a final control of the water level, the urea concentration and the acid-base balance is performed before the remaining filtrate goes into the ureter.
7. Influx into the collecting duct from a neighboring nephron.
8. Macula densa determining and controlling the NaCl concentration in the distal tubule.

Figure 10.10 schematically shows the production of urine from the initial renal blood flow. After processing in the kidney a fraction of only 0.1 % of the initial blood volume is finally excreted in the urine.

This is a very brief overview and for further details we refer to standard textbooks on Biology or Medical Physiology. However, some remarks are warranted about the proximal convolute tubule, as different processes take place here, which are also of physical interest.

A schematic of the proximal tubule is shown in Fig. 10.11. The gray shaded area lists all mass transport processes that take place in the nephron. First there is the active filtration in the Bowman's capsule based on a pressure gradient between the

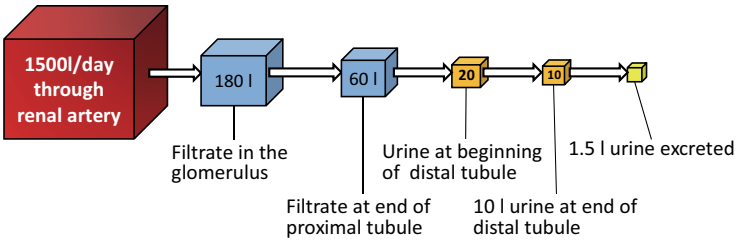


Fig. 10.10: Production of 1.5 l/day urine from a renal blood flow of 1500 l/day.

glomerular capillary and the capsule as already discussed. Next there is diffusion from the proximal tube to the peritubular capillary and vice versa due to a concentration difference for one specific substance. The diffusion process is followed, in general, by a shift of water in order to keep the osmotic pressure constant. Furthermore, ion transport in the proximal tubule is actively assisted by ATP ion pumps, like in depolarized cells for reestablishing a resting potential. To conclude, ion and water transport in the nephron can either be active with the help of pressure differences, or by an ATP pump, or may be diffusive according to concentration gradients and controlled by Fick's law. Both mechanisms work together for a most effective filtering and reabsorption of the plasma that serves the purpose of cleaning the blood, regulating the water level, and controlling the pH balance.

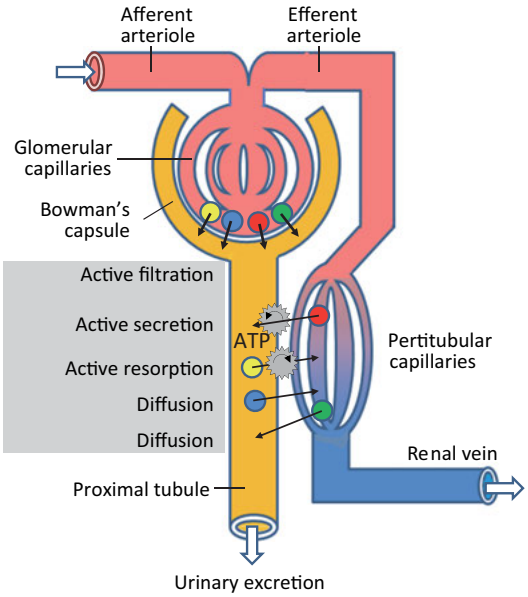


Fig. 10.11: Diffusion and reabsorption in the proximal tubule. Yellow, blue, red, and green circles symbolize different molecules passing the glomerular filter into the proximal tubule. From there the molecules diffuse back and forth or may actively be transported between proximal tubule and peritubular capillaries. Gear wheels indicate active transport.

10.6 Renal clearance

Renal clearance is an important concept not only for a basic understanding of the working principle of kidneys. Renal clearance is the volume of blood that is cleaned per unit of time. Therefore clearance has the unit of a volume rate [volume/time]. Clearance can tell us whether the kidneys work properly or whether there are signs of malfunction and disease. Clearance measures the overall function of nephrons: glomerular filtration, tubule reabsorption, and tubule secretion, related to the arterial input and venous or urinary output, as shown again in Fig. 10.12. Clearance is an integral measure of all individual processes that occur sequentially at different sites along the nephron. Furthermore, clearance integrates over all 2 million nephrons, working in parallel and highly redundantly. Therefore, clearance cannot localize individual damage to nephrons, but it gives overall information on the state of the kidneys. For more detailed microscopic studies, high resolution imaging techniques are needed as discussed in Chapter 15 and Chapters 5–7/Vol. 2, which can spectroscopically resolve different agents in the nanotubes and determine their distribution and flow velocities within the kidneys. For example, Fig. 6.5/Vol. 2 shows the time resolved perfusion of the radioisotope ^{99m}Tc-MAG3 via scintigraphy for investigating renal clearance.

Renal clearance is based on mass conservation for all substances that are neither metabolized in the kidneys nor synthesized. What goes into the renal artery must come out either through the renal vein or through the urine. Thus the arterial input must equal the sum of venous and urine output. The mass flow rate $\dot{m}$ is the product of mass density and volume flow rate:

$$\dot{m} = \rho \dot{V}.$$

Because of mass conservation, the following equation should hold for any substance S, where ρ_S is the mass density of S in units of [mg/ml] and $\dot{V}_S$ is the volume rate in units of [ml/min]:

$$r_{S,\mathrm{a}} \dot{V}_{S,\mathrm{a}} = \rho_{S,\mathrm{v}} \dot{V}_{S,\mathrm{v}} + \rho_{S,\mathrm{u}} \dot{V}_{S,\mathrm{u}}.$$

Here the subscripts a, v, and u refer to arterial, venous, and urine.

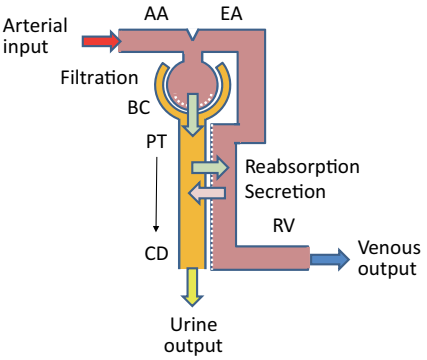


Fig. 10.12: Overview of the main functions of nephrons in kidneys: filtration, reabsorption, and secretion. The other acronyms stand for: AA = afferent arteriole; EA = efferent arteriole; BC = Bowman's capsule; PT = proximal tube; CD = collecting duct; RV = renal vein.

In physiology a different nomenclature is used for the same expression, which we want to adapt:

$$\underbrace{P_{S,\mathrm{a}} \times \mathrm{RPF_a}}_{\text{arterial input of } S} = \underbrace{P_{S,\mathrm{v}} \times \mathrm{RPF_v}}_{\text{venous output of } S} + \underbrace{U_S \times \dot{V}}_{\text{urine output of } S} .$$

In this expression $P_{S,\mathrm{a}}$, $P_{S,\mathrm{v}}$ are the densities of substance S in the renal artery and in the renal vein, respectively, $\mathrm{RPF_{a,v}}$ are the renal plasma flow rates in the renal artery and renal vein, respectively, U_S is the density of substance S in the urine, and $\dot{V}$ is the urine flow rate. The product $U_S\dot{V}$ is also known as the *urinary excretion rate*. All substances are carried by the plasma, and therefore the plasma flow rate is also the flow rate for the substances dissolved in the plasma. It is useful to remember that the RPF is 600 ml/min, GFR is 120 ml/min, and the urine flow rate $\dot{V}$ is about 1 ml/min.

Since we are only interested in the arterial input versus urine output, we rephrase the equation:

$$P_{S,\mathrm{a}}\left(\mathrm{RPF_a} - \frac{P_{S,\mathrm{v}}}{P_{S,\mathrm{a}}}\mathrm{RPF_v}\right) = U_S\dot{V}$$

or

$$P_{S,\mathrm{a}}C_S = U_S\dot{V}.$$

C_S is called renal clearance. Solving for C_S, we obtain:

$$C_S = \frac{U_S}{P_S}\dot{V}.$$

Here the subscript a in P_S for "arterial" is dropped, because the density of substance S in the renal veins no longer appears and P_S solely refers to the density of substance S in the arterial input. The clearance of different substances can be calculated from the concentration or density of substance S in the urine U_S, the concentration of the same substance in the blood plasma P_S, and the urine flow rate $\dot{V}$, which is the urine formed in a given time. These three quantities can easily be measured and hence it is not difficult to determine the clearance of substance S.

If, for instance, the nephrons completely clear substance S from the plasma in a single pass, then $P_{S,\mathrm{v}} = 0$ and the renal clearance becomes equal to the arterial plasma flow rate: $C_S = \mathrm{RPF_a}$.

Now we can rephrase the term *renal clearance* once more. Renal clearance C_S of a substance S is defined as the volume of plasma completely cleared of that substance by the kidneys per unit time. The unit is [ml/min]. Clearance compares a plasma volume per time in the afferent arterioles before filtering and the same plasma volume per time after filtering in the efferent arterioles. If the same substance is still present, the clearance is zero. If the substance is completely removed, the clearance is identical with the $\mathrm{RPF_a}$.

The solute para-aminohippuric acid (PAH) has the property that it is completely removed from the plasma by filtering, and therefore PAH can be used for determining the $\mathrm{RPF_a}$, as described in more detail later.

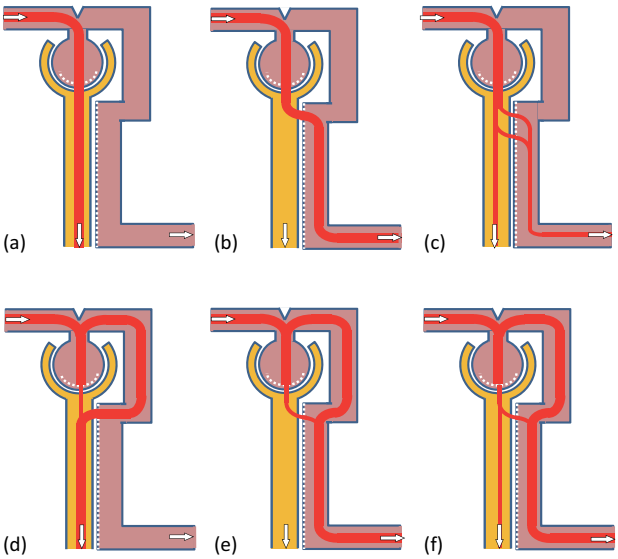


Fig. 10.13: Six different possibilities for clearance. (a) Complete filtering; (b) complete reabsorption; (c) complete filtering and partial reabsorption; (d) partial filtering and secretion; (e) partial filtering and reabsorption; (f) partial filtering and partial reabsorption.

We will now discuss some illustrative examples, which are schematically displayed in Fig. 10.13 for different cases of filtration, reabsorption, and secretion. Panels (a)–(c) show cases where a substance S is completely filtered, in panels (d)–(f) the filtration is incomplete.

In panel (a) the substance S_1 marked in red and dissolved in the plasma is completely filtered out in the Bowman's capsule. As no reabsorption from the tubule into the efferent veins or secretion from the veins into the tubule takes place, the density of substance S_1 is identical before and after filtration: $P_{S1} = U_{S1}$. Furthermore, the rate with which the substance S_1 appears in the urine ($U_{S1} \times \dot{V}$) must be the same as the glomerular filtration rate of that substance ($P_{S1} \times \mathrm{GFR}$):

$$P_{S1} \times \mathrm{GFR} = U_{S1} \times \dot{V}$$

or

$$\mathrm{GFR} = \frac{U_{S1}}{P_{S1}} \dot{V}.$$

Since $P_{S1} = U_{S1}$, as stated before, we have $\mathrm{GFR} = \dot{V} = C_{S1} = 120\,\mathrm{ml/min}$. Therefore with a substance that completely clears out by filtration, the glomerular filtration rate can be determined. In this case GFR is identical with the clearance C_{S1}. In fact, this first example is representative for the polysaccharide inulin. Inulin is completely filtered and is not reabsorbed. Therefore it serves as a "gold standard" for renal clearance. Clear-

ance of a substance S_x is often normalized by the clearance of inulin, called *fractional excretion* (FE). The FE for inulin is 1.

In case (b) another substance, S_2, is also completely filtered out in the Bowman's capsule. However, in contrast to the previous case, S_2 is completely reabsorbed from the proximal tubule to the efferent arterioles, so that after filtration the density in the tubule $U_{S2} = 0$. Therefore the renal clearance in this case is $C_{S2} = 0$ ml/min, although filtration has taken place. S_2 will not appear in the urine, all of it returns into circulation. This is in fact the case for glucose, which is reabsorbed completely. Here FE = 0.

In panel (c) a substance S_3 is completely filtered, but 50 % is reabsorbed in the peritubular capillaries and 50 % remains in the tubule. Therefore, $U_{S3}/P_{S3} = 0.5$, the clearance is: $C_{S3} = \text{GFR} \times 0.5 = 50$ ml/min, and the FE = 0.5. This example refers to the clearance of urea.

In panels (a)–(c), $P_{S,\text{v}} = 0$ and the filtration is 100 %. Panels (d)–(f), in contrast, show cases where the filtration is less than 100 %.

Panel (d) illustrates the situation when only a portion of the density P_{S4} of substance S_4 is filtered in the glomerulus. The rest bypasses the glomerular filter but is secreted from the efferent arterioles into the tubule. In the end the plasma in the efferent arterioles is free of substance S_4 and the total original density is excreted through the urinary tract. This situation occurs for para-aminohippuric acid (PAH). PAH is an organic acid that is not normally present in the body and must be administered through continuous intravenous infusion for testing. PAH is completely cleared from the plasma, therefore the clearance is equal to the total renal plasma flow (RPF), and the amount delivered to the kidneys is equal to the amount excreted. This is the same situation as for inulin although the pathways taken are different. To be specific, we assume that the PAH density in arterial blood plasma is $P_{\text{PAH}} = 0.1$ mg/ml. The renal plasma flow RPF is 600 ml/min, thus the amount RPF $\times$ P_{PAH} = 600 ml/min $\times$ 0.1 mg/ml = 60 mg/min will be loaded into the glomerulus for filtration, from which only 20 % or 12 mg/min will be filtered. The rest will remain in the plasma, but is finally secreted into the tubule, so that in the end the total amount of PAH in the urinary tract will be again 60 mg/min, which is the excretion rate $U_{\text{PAH}}\dot{V} = \text{RPF} \times P_{\text{PAH}}$. The clearance of PAH is therefore:

$$\begin{aligned} C_{\text{PAH}} &= \frac{U_{\text{PAH}}\dot{V}}{P_{\text{PAH}}} = \text{RPF} \\ &= \frac{60\,\text{mg/min}}{0.1\,\text{mg/ml}} = 600\,\text{ml/min}. \end{aligned}$$

Therefore, with the PAH test the RPF can be determined. The fractional excretion is FE = $C_{\text{PAH}}/C_{\text{Inulin}} = 600\,\text{ml}\,\text{min}^{-1}/120\,\text{ml}\,\text{min}^{-1} = 5$. This is the highest FE value known.

In panel (e) the filtration is less than 100 % and whatever has been filtered into the urinary track is reabsorbed back again into the arterioles. Therefore the clearance is zero as in panel (b).

Finally panel (f) is left to the reader for analysis.

Apart from very big proteins and erythrocytes, almost everything is filtered in the glomerulus. However, what will finally be excreted through the ureter duct is a question of reabsorption and secretion in the tubules and collecting duct. The amount of substance excreted in the ureter can be calculated as follows. The filtered part is

$$\mathrm{GFR} \times P_S.$$

From this we subtract the reabsorbed part with rate R_S and add the secreted part with rate S_S, yielding:

$$\underbrace{\mathrm{GFR} \times P_S}_{\text{filtrate}} - (R_S + S_S)U_S.$$

This sum equals the urinary excretion rate defined before:

$$\dot{V} \times U_S = (\mathrm{GFR} \times P_S) - (R_S - S_S)U_S.$$

In the case that $R_S = S_S = 0$, and if $U_S = P_S$ as is the case for inulin, then $\dot{V}_U = \mathrm{GFR} = C_S$. Therefore inulin can be used to determine the GFR, as noted before. The reabsorption rate R_S can also be determined in the case that the secretion rate $S_S = 0$:

$$R_S = \mathrm{GFR}\frac{P_S}{U_S} - \dot{V}.$$

Similarly we find for the secretion rate S_S if the reabsorption rate R_S is zero:

$$S_S = \dot{V} - \mathrm{GFR}\frac{P_S}{U_S}.$$

Now we come back once more to panel (e), which also represents the situation for the plasma flow itself: at the artery input plasma arrives with a PRF = 600 ml/min, but only 20 % is filtered, the remaining 80 % bypasses the glomerular filter. So the GFR is 120 ml/min. Most of the plasma filtered into the tubule is reabsorbed into the efferent arteriole. Only 1 % remains in the tubule. Without secretion we have for the excretion of plasma:

$$\dot{V} \times U_{\mathrm{Plas}} = \mathrm{GFR}_{\mathrm{Plas}} P_{\mathrm{Plas}} - R_{\mathrm{Plas}} U_{\mathrm{Plas}}.$$

In the case of plasma $P_{\mathrm{Plasma}} = U_{\mathrm{Plasma}} = 1\,\mathrm{g/ml}$. Then we have for the plasma clearance:

$$C_{\mathrm{Plas}} = \frac{U_{\mathrm{Plas}}}{P_{\mathrm{Plas}}}\dot{V} = \mathrm{GFR}_{\mathrm{Plas}} - R_{\mathrm{Plas}}\frac{U_{\mathrm{Plas}}}{P_{\mathrm{Plas}}},$$

or

$$C_{\mathrm{Plas}} = \dot{V} = \mathrm{GFR}_{\mathrm{Plas}} - R_{\mathrm{Plas}} = 120\,\mathrm{ml/min} - 119\,\mathrm{ml/min} = 1\,\mathrm{ml/min}.$$

The fractional excretion of plasma is:

$$\mathrm{FE}_{\mathrm{Plasma}} = C_{\mathrm{Plas}}/C_{\mathrm{Inulin}} = 1\,\mathrm{ml\,min^{-1}}/120\,\mathrm{ml\,min^{-1}} \approx 0.01.$$

This corresponds to the daily urinary volume of about 1.5 l.

Although inulin serves as a standard, a simpler test can actually be performed with creatinine. Creatinine is always in the body as a metabolic waste product of creatine. Creatine is produced by the body to supply energy mainly to muscles for anaerobic activity. Creatinine is removed from the body entirely via glomerular filtering in the kidneys similar to inulin. Thus the equation for creatinine clearance holds:

$$C_{\mathrm{Cr}} = \frac{U_{\mathrm{Cr}}}{P_{\mathrm{Cr}}}\dot{V} = \mathrm{GFR},$$

where $U_{\mathrm{Cr}} = P_{\mathrm{Cr}}$. This equation predicts that in a steady state situation, when metabolic production in muscles equals the urinary excretion rate $U_{\mathrm{Cr}}\dot{V}$ of creatinine, and assuming that creation and excretion remain fairly constant, the product $P_{\mathrm{Cr}}C_{\mathrm{Cr}} = \mathrm{const}$. Therefore, P_{Cr} should be inversely proportional to C_{Cr}. In Fig. 10.14 the plasma creatinine concentration P_{Cr} is plotted versus the glomerular filtration rate GFR. According to this figure, the prediction appears to hold. For instance, for a healthy kidney with a GFR of 100 ml/min the plasma creatinine density is approximately 0.01 mg/ml. The product GFR $\times$ P_{Cr} = 100 ml/min $\times$ 0.01 mg/ml = 1 mg/min corresponds to the production rate of creatinine. Thus the creatinine production rate and the creatinine excretion rate are equal. If by some reason the GFR drops to 50 ml/min, the creatinine level will rise to 0.02 ml/mg, while the product GFR $\times$ P_{Cr} remains constant. Creatinine tests are easy to perform. One just needs a sample of venous blood and a similar sample of urine that are analyzed with respect to creatinine density, which yields the desired ratio. If kidney function is not normal, the creatinine level in the blood will increase, since less creatinine is filtered in the glomerulus and therefore less creatinine is excreted through the urine.

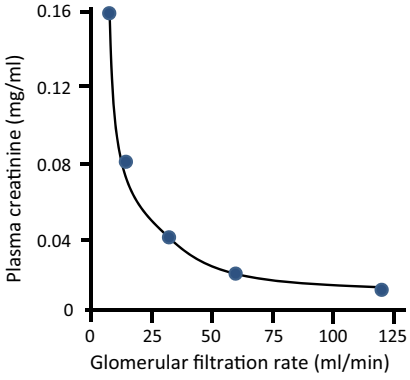


Fig. 10.14: Density of creatine in the plasma is plotted versus the glomerular filtration rate.

Table 10.2 lists different possible relations between filtration (F), clearance (C), and excretion (E) and conclusions that can be drawn from this. Table 10.3 is an overview of clearance and fractional excretion for a few standard solvents, and Tab. 10.4 summarizes different blood and urine tests discussed in the text.

Tab. 10.2: Overview of filtration, excretion, and clearance of substances in blood plasma. F_S = filtration of substance S; E_S = excretion of substance S; C_S = clearance of substance S.

Molecule S filtered at glomerulus	Renal processing
$F_S > E_S$	Net resorption of S
$F_S < E_S$	Net secretion of S
$F_S = E_S$	No exchange of S
$C_S < C_{inulin}$	Net reabsorption of S
$C_S > C_{inulin}$	Net secretion of S
$C_S = C_{inulin}$	No exchange of S

Tab. 10.3: Clearance and fractional excretion of different representative substances in the kidney.

Substance	Clearance [ml/min]	Fractional excretion
Glucose	0	0
Plasma	1	0.01
Urea	50	0.5
Inulin	120	1
Creatinine	120	1
PAH	600	5

Tab. 10.4: Overview of different renal tests. Samples may be taken from the blood and/or from the urine.

Test	Blood	Urine
Inulin	GFR	
PAH	RPF	
Creatine	GFR	GFR
Urea	Rest volume	
Glucose	Rest volume	

10.7 Artificial filtering: dialysis

If the creatinine level in the blood is too high compared to standard values of 0.007–0.012 mg/ml this may indicate a reduced GFR and therefore a malfunction of the kidneys.

One of the reasons for kidney malfunction is *diabetes mellitus*. Under normal conditions glucose in the plasma is completely removed from the plasma by glomerular filtration followed by a complete reabsorption from the tubule into the peritubular capillaries. Under normal conditions, the urine is free of glucose.

If the blood contains a rather high glucose concentration (> 1.80 mg/ml), this overloads the capability of the tubular system for reabsorption of glucose. Then more glucose remains in the urine. This will increase the osmotic pressure in the tubule and hinder water reabsorption into the blood. Then more urine will be excreted that contains more glucose than normal. This process leads to a thickening of the glomerular basement membrane (GBM) (see Fig. 10.6) and a swelling of the mesangial cells, which are located between the capillaries and the glomerulus and support the capillary walls. The glomerulus starts to leak proteins into the tubule, which clog up the filter, resulting in a reduced GFR. In the end normal kidney function stops completely, including disposal of waste products and control of water level and electrolytes in the body. In this final state the patient can only be helped with an artificial filtering and cleaning procedure, which is referred to as *dialysis*. Kidney diseases are discovered late. Serious clinical symptoms often do not show up until the number of functional nephrons drops to 70–75 % below normal. In fact, with the remaining 25–30 % relatively normal concentrations of most electrolytes in the blood and normal body fluid volumes can still be maintained. However, if the kidneys completely fail for whatever reason, and nothing is done to restore their function, the remaining lifespan is about 10 days.

Dialysis machines are designed to mimic the function of kidneys by removing waste products and maintaining water levels, electrolytes, and minerals within acceptable margins. Two dialysis methods are commonly used. *Hemodialysis* machines circulate blood outside the body, and *peritoneal dialysis* performs filtration within the body and uses the peritoneum as a natural filter for wastes and water. Peritoneum is a membrane covering most of the abdominal organs. In either case, dialysis is an incomplete replacement of lost kidney function, as only filtration of blood is achieved but no other tasks of a normal kidney are fulfilled, such as the control of electrolytes and synthesis of hormones.

We discuss here only the first version: hemodialysis. Figure 10.15 shows the schematic of a hemodialysis setup. It consists of two loops: a primary and a secondary loop. In the primary loop the blood is drained from an artery in the arm and after filtering returned back to a vein. Pumps, pressure control and air traps control the proper circulation of the blood in the primary loop. Heparin, a blood thinner, is injected to avoid blood clotting. The central part of the dialysis machine is a semipermeable membrane which removes excess water and waste that is drained in the secondary loop. During this process various solutions from the dialysate are mixed with the blood in order to remove specific contaminants or correct specific conditions according to the needs of the patient.

The semipermeable filter membrane between blood and dialysate contains pores of proper size that water, electrolytes, urea and other waste products can be passed through but not larger proteins and erythrocytes. Thus the semipermeable membrane has similar properties as the glomerular filter in the nephrons. The principle of the filter function is displayed in Fig. 10.16 (a) and in Fig. 10.16 (b) a sketch is shown of one of

the many possible technical realizations. The filter consists of a bundle of hollow fiber membranes. Blood flows into the hollow tubes of the fibers. Each fiber is embedded in dialysate fluid, which flows in opposite direction to the blood circulation. Tiny pores in the fibers filter the blood on the one hand and allow penetration of electrolytes from the dialysate on the other hand. Important for all designs is that blood and dialysate flow in counterstream fashion to avoid contamination.

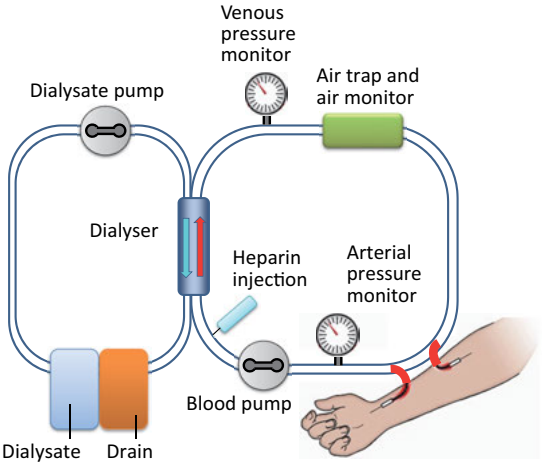


Fig. 10.15: Dialysis setup for cleaning blood.

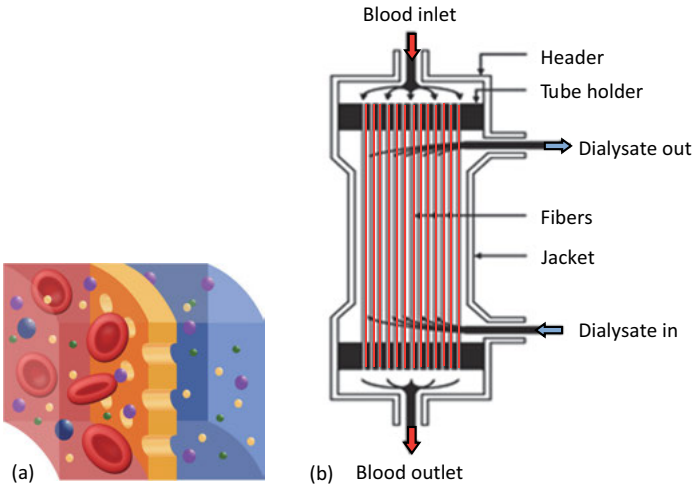


Fig. 10.16: (a) Schematic of a membrane filter in the dialyzer acting as a molecular sieve. Water and small molecules can pass, but not larger proteins and red blood cells (reproduced from https://en.wikipedia.org/wiki/Semipermeable_membrane, © Creative Commons); (b) sketch of fiber membrane filter. Blood flows from top to bottom through the hollow fibers while dialysate flows in counterstream fashion from bottom to top.

10.8 Summary

There are three main tasks of the kidneys: (1) Regulation of blood level; (2) Control of acid-base balance; (3) Removal of metabolites.

1. The nephron is the most important structure in the kidney that acts as a microscopic filtration unit.
2. The nephron consists of Bowman's capsule including glomerulus, proximal tubule, loop of Henle, distal tubule, and collecting duct.
3. There are three types of filtration/exchange of the nephrons: glomerular filtration, tubule reabsorption, and tubule secretion.
4. Filtration and exchange is achieved by active transport and diffusion.
5. The glomerulus is a dense mass of very fine blood capillaries in the Bowman's capsule that act as a filter.
6. Glomerular filtrate is the liquid removed from the blood by filtration in the kidney.
7. Filtration occurs only for molecules with molecular weight below 10 000 amu.
8. Renal clearance of a substance is defined as the volume of plasma completely cleared of that substance by the kidneys per unit time.
9. Clearance of Inulin is 1, clearance of glucose is 0.
10. Creatinine production rate and creatinine excretion rate are equal.
11. 1500 l of blood run through the kidneys daily. The production of primary urine is about 170–180 l/day, that of final secondary urine is 1–2 l/day.
12. Dialysis machines are designed to mimic the function of kidneys by removing waste products and maintaining water levels, electrolytes, and minerals within acceptable margins.

References

[1] Longmire M, Choyke PL, Kobayashi H. Clearance properties of nano-sized particles and molecules as imaging agents: Considerations and caveats. Nanomedicine. 2008; 3: 703–717.

Further reading

Layton AT, Edwards A. Mathematical modeling in renal physiology. Berlin, Heidelberg: Springer Verlag; 2014.

Boron WF, Boulpaep EL. Medical physiology. 2nd edition. Saunders W.B. Elsevier; 2012.

11 Basic mechanism of vision

11.1 Introduction

It has been estimated that 80 % of our sensory information from the environment is perceived through our eyes. The processing of visual perception occupies one-quarter of our brain. Visual intelligence of infants develops earlier and is completed before verbal skills advance. All these facts show how important the visual system is. Loss due to blindness is a severe disability.

Visual perception is an extremely complex procedure, consisting of three main parts: (1) optics of the eye, (2) photon detection and first image processing within the retina, (3) signal transmission and further processing in the visual cortex of the brain. This chapter covers the first two parts.

The eyeball as an optical instrument is the least impressive part of our visual system. Any compact camera provides sharper images than the human eye does. On the other hand, the photon detection in the retina is much more efficient than the photon detection with a CCD chip. The sensitivity of the retina covers 10 orders of magnitude from dim to bright sunshine. Although the bandwidth of photon wavelengths that we can recognize is rather limited from about 400 nm (violet) to about 700 nm (red), we can distinguish between millions of different hues within this wavelength range. The signal processing starting in the retina and continuing in the visual cortex is most remarkable. We perceive more with our brain than we see with our eyes.

We start with some anatomical aspects of the eye and physical aspects of image formation, and then progress to the retina. This can only be a short overview on the main features of the eye, for a more detailed description we refer to more specialized literature and textbooks listed at the end.

Figure 11.1 shows a cross section of the eyeball labeling the most important parts: *cornea* together with lens for refraction, zonular fibers for accommodation, iris for aperture action (pupil), retina for detection of optical signals, and optic nerve for transmission of action potentials to the brain. The *choroid* is filled with blood vessels for oxygen supply and for cooling the back of the retina which is warmed up by focused light. The eyeball is filled with a transparent glassy material called *vitreous humor*. The anterior chamber is filled with a watery fluid called *aqueous humor* and has the same refractive index as the vitreous humor. The lens, in contrast, has a higher refractive index produced by an arrangement of fiber-accumulating proteins known as *crystallins*.

The iris has the function of an aperture controlling the intensity of light focused on to the *retina* by increasing or decreasing the diameter of the *pupil*. The color of the eye is the color of the iris. Brown eyes contain melanin pigment in the fibers (stroma) of the iris, which absorbs the blue part of the spectrum, leaving a brown impression upon light reflection. However, in case of blue eyes there is no equivalent blue pigment. The

DOI 10.1515/9783110372830-013

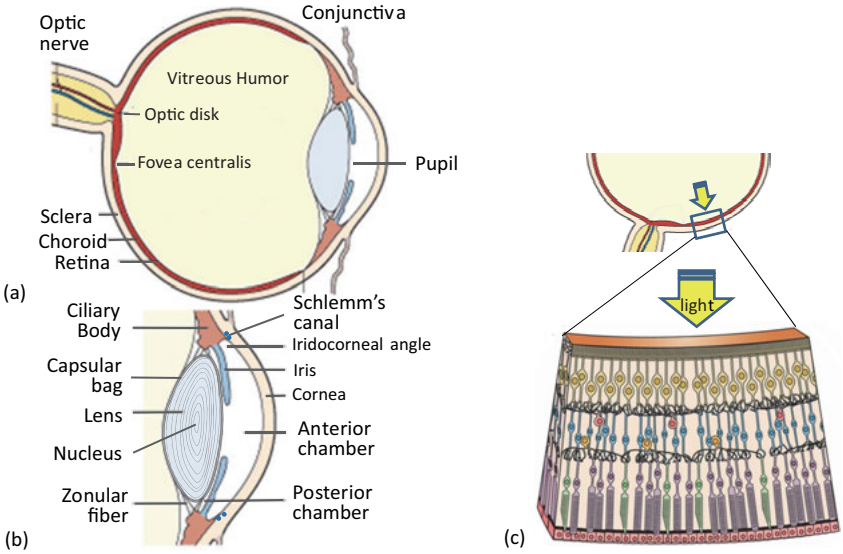


Fig. 11.1: (a) Cross section of the eyeball. (b) Enlargement of the lens system. (c) Cross section of the retina containing photoreceptors and various neurons for signal processing.

color blue is the result of light scattering in the tissue of the iris, until blue is left over, an effect known as Tyndall effect or Rayleigh scattering, which is also responsible for the blue color of the sky [1]. The lack of melanin in the iris is actually the result of a mutation that occurred some 10 000 years ago. The diameter of the pupil ranges from a minimum $d_{\min} = 1.5$ mm to a maximum $d_{\max} = 7.5$ mm, this corresponds to an area ratio of 25, or roughly one decade with respect to intensity. However, the *dynamic range* of the eye covers a much larger range of about 10 decades from 10^{-6}–10^{4} cd/m^2. Therefore, there must be an additional mechanism in the eye for adapting to light intensity apart from the iris/pupil, which we will learn about in later sections of this chapter.

The response time of the iris on intensity changes is very rapid, taking only 0.2 s to 0.5 s for contraction, which is important for protecting the retina against too high light intensity. The iris *contraction muscle* controlling the pupil (aperture) and the *ciliary muscle* controlling the curvature of the eye lens (accommodation) form one unit in the ciliary body.

Eyeball movement is controlled by a total of six muscles (Fig. 11.2), four of them are straight muscles: top (rectus superior), bottom (rectus inferior), left (rectus lateralis), and right (rectus medialis) controlling tilting movement, and two more oblique muscles, inferior and superior, controlling rotational movement.

In total, there are nine muscles per eye, six for controlling the eyeball movement, two muscles for adjusting the iris (pupil dilator muscle and pupil constrictor muscle), and one ciliary muscle for accommodation. The muscles for left and right eye movement are synchronized. If they move parallel in the same direction, eye movement is

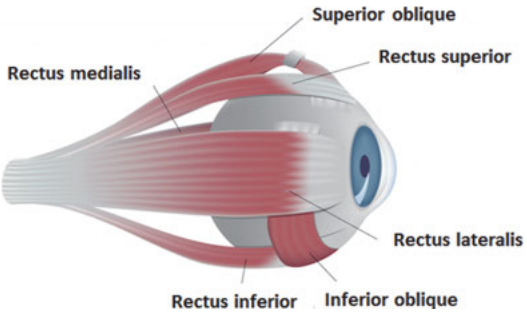


Fig. 11.2: Six exterior muscles of the eye control eye movement. Three more muscles are inside for control of aperture and accommodation.

called conjugated. If eye movement is tilted inwards, the movement is convergent. Eye movement allows a biaxial fixation on single objects that is necessary for obtaining distance information. All eye muscles are at constant tension, enabling fast reaction, much faster than any other body movement.

11.2 Optics of the eye

11.2.1 Refraction power of the eye

In first approximation the focal length f of the eye must be matched to the diameter of the eyeball, since objects at infinite distance should be focused on the retina. This distance can be estimated to be in the order of 25 mm. Usually we see objects sharply at distances greater than twice the focal length, i.e., at distances more than 50 mm. The near point, where we can still see objects sharp, is actually 100 mm at young age. Thus images of objects located at distances larger than twice the focal length are reduced in size and inverted on the retina.

Next we realize that the eye is completely filled with transparent material and all light rays entering through the cornea stay inside the glassy and jelly-like body of the eye, called vitreous humor. Therefore we have to treat the eye as a thick lens with a focal length f_2 on the image side within the eye and a different focal length f_1 outside in front of the eye. Using the lens equation for thick lenses we find:

$$\frac{1}{o} + \frac{n_2/n_1}{i} = \frac{1}{f_1}; \quad \frac{n_1/n_2}{o} + \frac{1}{i} = \frac{1}{f_2}.$$

Here the refractive index $n_1 = 1$ for air and $n_2 = 1.33$ for a watery substance, o is the object distance to the cornea and i is the image distance between cornea and retina. For an object at infinite distance these equations simplify to:

$$\frac{n_2}{f_2} = \frac{n_1}{f_1}.$$

Since $f_2 = 25$ mm we find for $f_1 = 18.8$ mm. The refraction power is defined as $1/f_1$ or n_2/f_2, and is in our case $53.2\,\mathrm{m}^{-1} = 53.2$ dpt. The unit is diopter (dpt), where $1\,\mathrm{m}^{-1} =$

1 dpt. Without taking any measurement we already arrive at a pretty good estimate for the focal length and the refraction power of the eye, which comes close to the real value of about 58 dpt.

The eye is in fact not a homogeneous spherical body but is composed of several parts with different refractive indices, which are indicated in Fig. 11.3. The cornea at the air/eye interface has the largest curvature and therefore causes the largest refractive effect. In fact, the thickness of the cornea at the center (550 µm) is less than at the rim (650 µm) and would therefore act as a diverging lens unless backed on the inside by a refractive index matching fluid (*vitreous humor*). The biconvex lens has the largest refractive index and the curvature can be adapted for accommodation. The average refractive index of the lens is 1.408. Closer inspection shows that there is a refractive index gradient in the direction normal to the lens. The refractive index is highest at the center of the lens and decreases outwards on either side [2]. The anterior chamber containing aqueous humor and the posterior chamber containing vitreous humor have refractive indices similar to water.

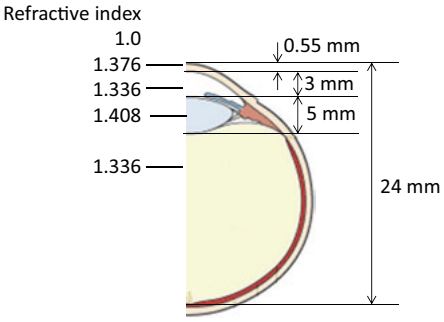


Fig. 11.3: Refractive indices and distances of different parts of the eye for a wavelength of 590 nm.

Now we are prepared to reconsider the optics of the eye by starting without the lens. For the standard eye the main refraction occurs at the air/cornea interface. From the curvature of the cornea ($R = 7.7$ mm) and using the lens maker equation:

$$f_1 = \frac{n_1}{n_2 - n_1} R,$$

we find $f_1 = 23$ mm assuming an average refractive index of $n_2 = 1.336$. Then $f_2 = f_1 \times n_2 = 31$ mm, and the refraction power is 43 dpt.

With these values the image of objects at infinite distance lies at 31 mm beyond the retina (see Fig. 11.4 (a)). The lens adds another 15 dpt to the refraction power to reach the focal point on the retina at a distance of 24 mm from the cornea in a relaxed state (Fig. 11.4 (b)). This corresponds to an average refractive index of 1.4, a refractive power of 58 dpt, and a focal length of 17 mm in front of the eye.

For accommodation the lens has to increase the curvature so that the total refractive power increases, the focal length f_2 moves in and f_1 also decreases. Then an

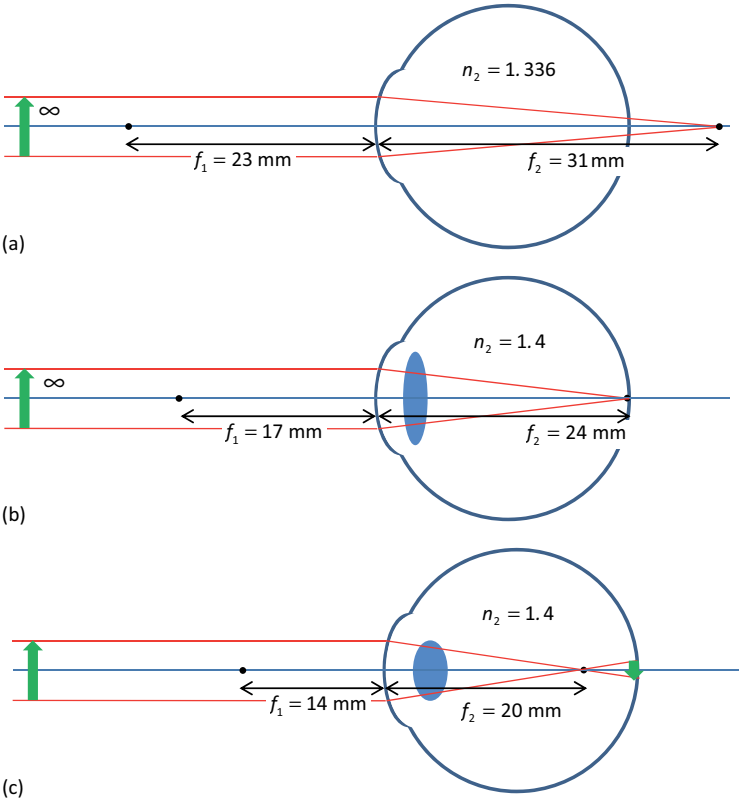


Fig. 11.4: Imaging with the eye as a thick lens. Note that the focal length before and in the eye are different. (a) Refraction only due to the cornea and assuming a homogeneous refractive index within the eye; (b) inserting a lens with a higher refractive index places the focus for infinite objects on the retina; (c) for objects that are closer to the eye, the lens has to accommodate the focal length by increasing its curvature.

inverted and reduced image occurs on the retina for objects located at more than twice the distance of the focal length f_1 (Fig. 11.4 (c)). This would be about 35 mm, but actually the shortest distance young people can accommodate to is about 80–100 mm.

11.2.2 Accommodation

The ability of the eye to produce a sharp image of objects at far distance up to the near point is called *accommodation*. Accommodation width ΔA is defined as the difference in the refraction power for objects at the near point and at the far point:

$$\Delta A = \frac{1}{d_{\text{near}}} - \frac{1}{d_{\text{far}}}.$$

For normal sighted (emmetrope) young individuals the far point is at infinite distance and the near point is about 0.08 m; the accommodation width is therefore:

$$1/0.08\,\mathrm{m} - 1/\infty = 12.5\,\mathrm{m}^{-1} - 0 = 12.5\,\mathrm{dpt}.$$

For short sighted older individuals the far distance is still infinite, the near point may be 0.33 m; then the accommodation width is:

$$1/0.33\,\mathrm{m} - 1/\infty\,\mathrm{m} = 3\,\mathrm{m}^{-1} - 0\,\mathrm{m}^{-1} = 3\,\mathrm{dpt}.$$

The age dependence of the accommodation width is shown in Fig. 11.5. The ability to accommodate strongly depends on age and decreases already at an age of 10 years, much before reaching adulthood. Accommodation is usually completely lost at the age of 50. This is one of the strongest age-dependent properties of the body. Reading glasses become necessary already beyond 40.

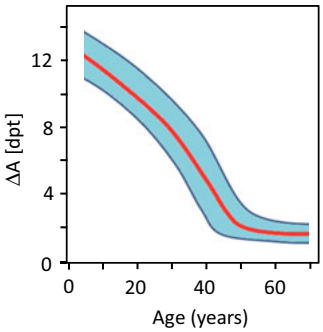


Fig. 11.5: Accommodation width as a function of age. The blue band indicates a one sigma standard deviation from the mean.

Accommodation is the cooperative result of the ciliary muscle and zonular fibers. In the relaxed state of the eye the ciliary muscle is relaxed and the zonular fibers are pulled straight, which flattens the lens (Fig. 11.6). In this state the lens contributes 15 dpt to the refraction power of the eye, as we have already seen. However, if we want to see objects at near distance, we need to accommodate the image on the retina. This is achieved by contracting the ciliary muscle, which releases the zonular fibers causing a rounding of the lens. The higher curvature of the lens increases the effective

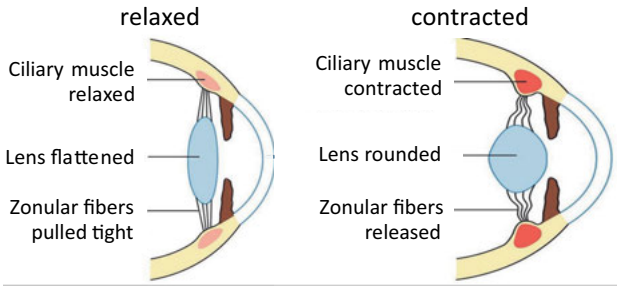


Fig. 11.6: Ciliary muscle in relaxed and contracted position.

refractive power, bringing the image onto the retina. The description of the accommodation mechanism goes back to Hermann von Helmholtz (1821–1894) and still holds today. With aging, the lens becomes increasingly stiffer and remains flat, progressively reducing the ciliary contraction effect. Different options for restoring the accommodation power by surgical correction of presbyopia are currently under investigation, for a review we refer to [3].

Presbyopia, the gradual age-related loss of accommodation, occurs primarily through a steady stiffening of the lens fiber material. There is no evidence that loss of ciliary muscle contractility would diminish with age and contribute to presbyopia. Stiffening of the lens fibers is surprising as they are continuously renewed throughout life. As new fibers are added from the outside and move towards the center, there is an age-related gradient of fibers, the oldest ones sitting in the nucleus of the lens thereby increasing the thickness of the lens over time. The age gradient is accompanied by a refractive index gradient, being higher in the central nucleus than in the shell, as mentioned before. The latter gradient is accounted for by a higher protein concentration in the nucleus [4]. Although stiffening of the lens has been confirmed in numerous studies, the origin remains unresolved.

11.2.3 Resolving power

According to the Rayleigh criterion for optical resolution, two objects a and b can be recognized as separate if the zeroth order intensity maximum of the diffraction pattern from object b falls into the first minimum of the diffraction pattern of object a. For circular apertures of radius r the diffraction pattern consists of airy discs. In this case the Rayleigh criterion is expressed in terms of a minimum angle $\alpha_{\min}$ at which two objects can be separated:

$$\alpha_{\min} = 1.22\frac{\lambda/n}{2r}.$$

For the eye we use average numbers: $2r = 5$ mm for the pupil, $\lambda = 550$ nm, $n_{\text{eye}} = 1.4$. The minimum angle is then (Fig. 11.7):

$$\alpha_{\min} = 1.22\frac{\lambda/n}{2r} = \frac{d_{\min}}{f},$$

where $f = 24$ mm. Inserting numbers, we find for two objects being recognized as separate a minimum distance on the retina of $d_{\min} = 2.3\,\mu$m. How does this match to the average distance of cones within the fovea of the retina? In the foveal region of 500 µm diameter the cones are rather small and their density is highest with an average separation of about 2.5–3 µm (Fig. 11.9). Therefore the optical resolution of the eye and the density of light receptors in the foveal region match very well. In fact, the physical resolution is slightly better than the physiological limit: we can distinguish two objects a and b on the retina as being different if the images of a and b fall on two

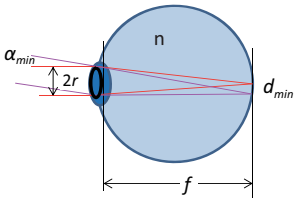


Fig. 11.7: Resolving power of the eye.

receptors separated by at least one receptor in between, which accounts for a spatial resolution of about 5 µm.

11.2.4 Visual acuity

At a distance g_0 of 25 cm we see objects with a psychophysical magnification of 1, i.e., our personal judgement tells us that at this distance objects appear to us by their proper size. At shorter distances objects appear enlarged; at larger distances they appear reduced. The distance g_0 is called the *distance of most distinct vision*. If G is the object height, we observe objects, in general, at a visual angle of

$$\varepsilon = \frac{G}{g},$$

and $\varepsilon_0 = G/g_0$ is the most distinct visual angle. $\varepsilon = 1°$ corresponds to an image height of 300 µm on the retina. The *visual acuity* is defined as the reciprocal value of the visual angle:

$$a = \frac{1}{\varepsilon}.$$

If a gap in the Landolt ring shown in Fig. 11.8 (a) can be recognized under a visual angle of 1′, then the visual acuity is 1. This corresponds to a separation on the retina by 5 µm. Visual acuity depends on the retinal location of the image projection and on the brightness. Visual acuity is highest in the area of the fovea with the highest density of cones and drops off outside with increasing eccentricity (see Fig. 11.9). In comparison we observe the sun and the moon under the same visual angle of 30′.

Visual acuity is often expressed in terms of a fraction, such as 20/40. The first number refers to the distance (in meters or feet) a person with impaired eyesight can correctly read the line with the smallest letters on a chart. The second number indicates the distance at which a person with normal eyesight could read the same line of letters correctly. The smaller the fraction, the worse is the eyesight of the patient.

Vernier acuity is the ability to recognize a shift in the contour of an object, as shown in Fig. 11.8 (b). Vernier acuity is about six times higher than visual acuity and better than what could be explained physically by the receptor density. Therefore, this high vernier acuity can be taken as a first hint of a physiological effect of enhancement and repression within receptor fields discussed further in Section 11.3.4.

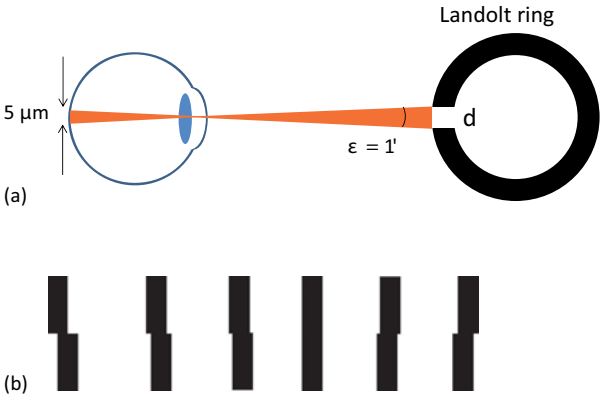


Fig. 11.8: (a) Visual acuity is defined by recognizing a 1′ gap in the Landolt ring.
(b) Vernier acuity is higher than visual acuity.

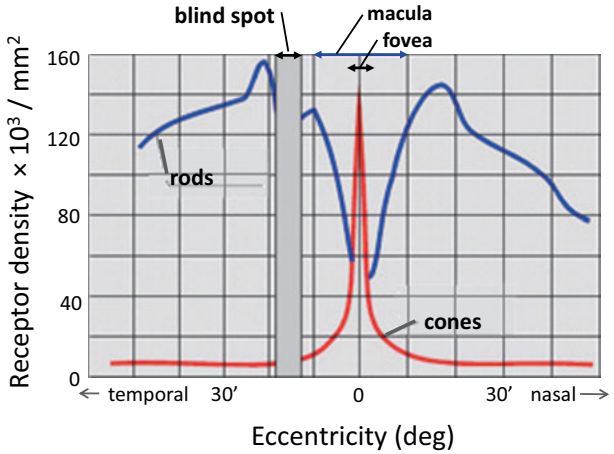


Fig. 11.9: Receptor density as a function of eccentricity. Note that the distribution of cones and rods is highly uneven in the retina.

11.2.5 Lens aberrations

The most frequent aberrations of the human lens are *myopia* and *hyperopia*. Both can be easily corrected with eye glasses.

Myopia or near-sightedness refers to the fact that the focal length of an object at infinite distance lies before the retina.

If $o_{\max}$ is the object distance at which a sharp image occurs on the retina as shown in Fig. 11.10, then the lens equation becomes:

$$\frac{1}{o_{\max}} + \frac{1}{i} = \frac{1}{f_e} = D_e,$$

where D_e is the refraction power of the eye. The goal is a sharp image on the retina of objects at infinite distance, which can be achieved with the help of a diverging lens. Then the lens equation changes to:

$$\frac{1}{\infty} + \frac{1}{i} = \frac{1}{f_e} + \frac{1}{f_{dl}} = D_e + D_{dl},$$

where f_{dl} and D_{dl} is the focal length and the refraction power of the diverging lens, respectively. In a lens system the refraction powers add up. Therefore we find for the required refraction power of the diverging lens:

$$D_{dl} = -\frac{1}{o_{max}}.$$

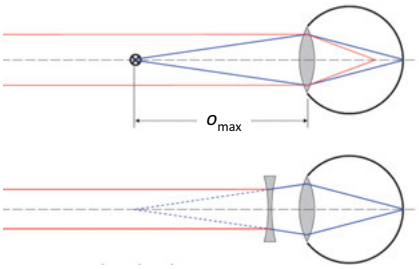


Fig. 11.10: Diverging biconcave correcting lens for myopia.

Similar arguments hold for imaging with hyperopia or far-sightedness (Fig. 11.11). At a minimum distance o_{min} a sharp image is projected on the retina, but for closer object distances the image lies behind the eye. Therefore a converging correcting lens is required to bring the image back onto the retina.

The lens equation without correcting lens for the minimum distance is:

$$\frac{1}{o_{min}} + \frac{1}{i} = \frac{1}{f_e} = D_e.$$

With correcting lens it is:

$$\frac{1}{g_0} + \frac{1}{i} = \frac{1}{f_e} = D_e + D_{cl}.$$

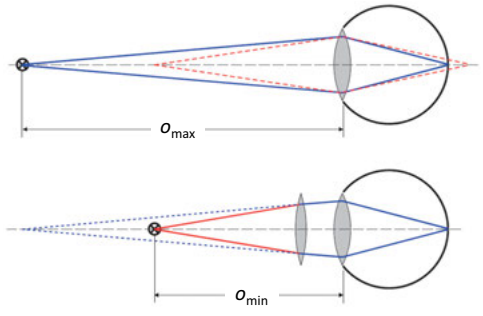


Fig. 11.11: Converging biconvex correcting lens in case of hyperopia.

For the refraction power D_{cl} of the converging correcting lens we therefore find:

$$D_{\mathrm{cl}} = \frac{1}{g_0} - \frac{1}{o_{\mathrm{min}}} = 4 - \frac{1}{o_{\mathrm{min}}}.$$

In the last equation we have set $g_0 = 0.25$ m for the distance of most distinct vision.

There are three more aberrations that are usually discussed in physics and optics textbooks: spherical aberration, chromatic aberration, and astigmatism. The first two play basically no role for the optics of the eye. But astigmatism is frequently observed because often the cornea has a nonspherical shape. Usually astigmatism occurs along with myopia or hyperopia and is corrected for simultaneously by properly crafting the correcting lenses. A critical review on all aspects of the eye's optics can be found in [5].

Corrections of myopia and hyperopia by laser applications, in particular photorefractive keratectomy (PRK) and laser-assisted interstitial keratomileusis (LASIK), are presented in Chapter 13/Vol. 2.

11.2.6 Cataract

Cataract is a disease of the visual system that causes the lens first to become opaque and at later stages to block the light completely from transmission towards the retina. It is one of several possible causes of blindness. Fortunately, it is the one which can nowadays be restored by replacing the lens with an artificial one.

Some cataracts are related to inherited genetic disorders. Others can be caused by medical conditions such as diabetes, trauma, past eye surgery, or eye injuries. Cataract is not necessarily age-related, although the fibers in the lens are responsible for both loss of accommodation and opacity [4].

Once cataract symptoms have started, they are progressive. In the opaque stage light becomes strongly scattered such that the image appears blurred, dimmed, and grayish. This type of scattering is known as Mie scattering. It is, for instance, responsible for the scattering of light in clouds and requires objects with a size in the order of the wavelength of light.

Recent research has shed light on the possible cause of cataract [6]. The lens requires a higher refractive index than the surrounding vitreous and aqueous humor, while remaining transparent for visible light. This is achieved by proteins called crystallins. Two proteins are mainly involved; we call them for simplicity *A* and *B*. The crystallins *A* and *B* tend to clump together unless kept separate by another "chaperone"-type protein *C*. There is only a finite amount of proteins *C* in the lens. When depleted by some reason or other, proteins *A* and *B* may aggregate causing the lens to become cloudy. Knowing the cause, there is hope to find a biochemical/pharmaceutical cure for cataract. Until then the only treatment is a surgical replacement of the lens.

Cataract can be treated by surgery in four steps as shown schematically in Fig. 11.12. The ambulant procedure taking less than 10 minutes starts with a small

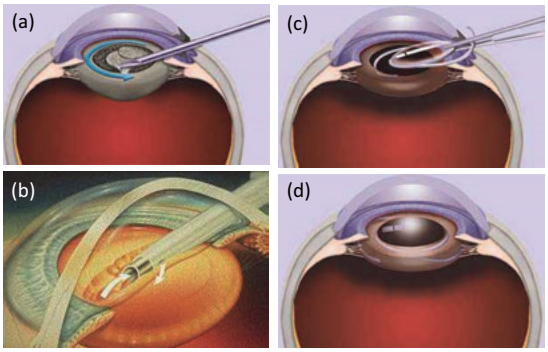


Fig. 11.12: Operational steps during a cataract lens replacement. (a) Opening the lens; (b) fragmentation of old lens and sucking off debris; (c) inserting new artificial lens; (d) closing the incision.

incision of 3 mm length through the cornea and into a pouch containing the vitreous material of the lens, called capsular bag (Fig. 11.12 (a)). The bag covering the lens is opened in a circular motion, which is known as a *continuous curvilinear capsulorhexis* (CCC). For the cut a sharp stainless knife is used and the CCC procedure is performed with tweezers.

In the second step (b) the lens material (nucleus) is fragmented using either ultrasonic techniques or sometimes also a femtosecond laser (see Chapter 13/Vol. 2). This procedure is called *phacoemulsification of the nucleus*. The remains are sucked off via a tiny pump.

In the third step (c) the original lens is replaced by an artificial lens. The refractive power of the artificial lens is preselected and cannot be changed after insertion. Plastic springs on the rim keep the lens in place.

During the fourth and final step (d) the cornea is closed again by self-sealing and without any stitches. After a short time, the tunnel through the cornea-sclera will close and heal and the procedure is finished. Sight is reestablished by this procedure, but accommodation is lost. Obviously it is extremely important to keep the eye free from any contaminants and in particular bacteria during surgery. Therefore cleanliness is the key to success. The entire procedure is well documented and explained in videos posted in the internet [7].

In Nepal a procedure has been developed and established that allows high volume, high quality, and low cost cataract surgery by a slightly different procedure [8]. First a sclerocornea tunnel to the lens is established with a small knife. Then the capsule is opened by the same CCC technique as described above. After the nucleus is extracted entirely from the capsular bag with the use of a tiny fishhook, the lens is replaced and the self-healing tunnel is closed again. The main difference to the standard technique is the fact that phacoemulsification is avoided, instead using for the entire procedure no special equipment other than a good microscope and a fishhook.

11.2.7 Intraocular pressure (IOP)

The intraocular pressure of eyes is in the order of 27 hPa. The static pressure is maintained by a balance of inflow and outflow of aqueous humor. A proper IOP is important for the spherical shape of the eyeball and for keeping all inner parts of the eye in place (Fig. 11.13).

Aqueous humor is a transparent fluid that consists mainly of water (98 %) similar to blood plasma, providing nutrition and oxygen to cornea, lens, retina, and vitreous humor, i.e., to all those parts of the eye that do not have access to the blood circulation. Aqueous humor is filtered and secreted from blood in the ciliary body behind the iris and the drainage goes through the trabecular meshwork into the *canal of Schlemm* and back into the circulatory system. The balance of in- and outflow is extremely important and usually controlled by stimulation of β-receptors. If the pressure is too low, the cornea will inflate; if the pressure is too high, there is a potential danger of damaging the retina and the optic nerve. Therefore it is important to measure the IOP during an eye examination. Normal IOPs are between 13 hPa and 27 hPa. *Hypertony* is defined as pressures beyond 27 hPa; *hypotony* is present for pressures below 6.5 hPa.

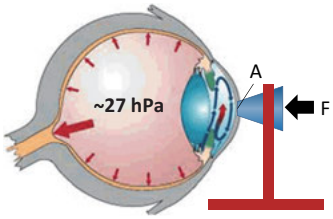


Fig. 11.13: Hydrostatic pressure of the eye is dynamically maintained by inflow and outflow of aqueous humor. If the pressure is too high, the retina and the optic nerve may be damaged. A tonometer can be used for measuring the intraocular pressure.

The IOP can be measured with a *tonometer*, the technique is illustrated in Fig. 11.13. The cornea is touched with a flat conical shaped prism after a topical anesthetic has been administered to the patient. Then the force F is measured that is required for flattening the cornea such that the area of the conical dip equals the indented area A on the cornea. A standard $A = 3.06$ mm is used. Under this condition the inside pressure equals the outside pressure $p = F/A$. However, the measured values can be affected by the thickness of the cornea and its rigidity. The normal and average thickness of the cornea is 555 µm. If for instance the corneal tissue is thicker than average, the tonometry reading is artificially high; whereas if the cornea is thinner than average, the tonometry reading is artificially low. Using an additional ultrasonic determination of the corneal thickness, the tonometry reading can be corrected for.

Another and contact free method measures the deflection of the cornea under pulsed air pressure (air-puff). This method is quick, but not error free, as the pressurized air pulse may be scattered at eye lids or damped by eyelashes. There are a number of others methods, but the two methods mentioned are the ones most frequently used by ophthalmologists.

Fig. 11.14: Perimeter for controlling the visual field.

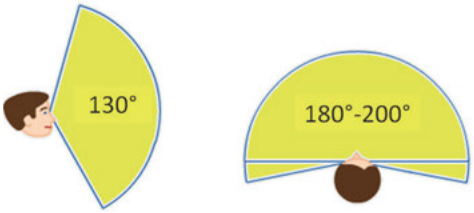


Fig. 11.15: Visual field of a normal eye in the vertical and horizontal direction.

In the case of hypertony the optic nerve and/or parts of the retina may have been damaged resulting in a reduced visual field. The visual field can be tested for both eyes independently with the use of a *perimeter*, shown in Fig. 11.14. A perimeter consists of a hollow white hemisphere. One eye is placed at the center, fixing on a light at the apex of the sphere, while the other one is blocked by an eye patch. A light is sequentially flashed at various positions and with various intensities all over the hemisphere. The patient, while not moving the eye, pushes a button whenever he/she recognizes a light point within the hemisphere. Several sweeps are performed for correcting errors. After the procedure is finished a visual map is established detailing areas of high and low light sensitivity.

In a normal eye the *visual field* spans 130° in the vertical direction and 180–200° in the horizontal direction (Fig. 11.15). There is only one *blind spot* at the location where the optic nerve and blood vessels enter the retina, called *optic disk*.

The visual field of a normal person mapped onto a circular disk and color coded is reproduced in Fig. 11.16 (a). The blind spot appears on the right side, implying that the map is from the right eye. The visual field of another person with *glaucoma* is demonstrated in Fig. 11.16 (b). The brown colored area indicates regions with reduced light sensitivity. With progressing glaucoma those areas may expand and turn black, like in the central part of the blind spot, unless measures are taken to reduce the IOP. It should be mentioned that glaucoma may also develop at normal IOP, called *normal tension glaucoma* (NTG). NTG damages the optic nerve rather than the retina. The reason for NTG is not clear presently. Perimeters are also used for distinguishing between glaucoma and any damage to the optic nerve as a result of stroke.

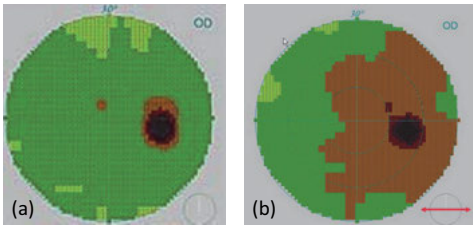


Fig. 11.16: Visual field mapped onto a circular disk shaped area. (a) Visual field of a normal person and of a person with glaucoma of the right eye (b). The area of the optic disk appears black.

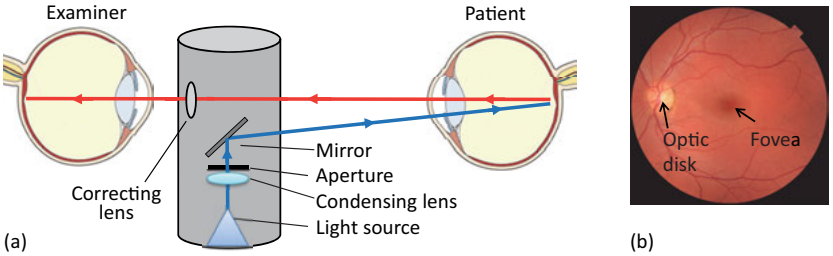


Fig. 11.17: (a) Schematic of a direct ophthalmoscope for the examination of the fundus; (b) image of a normal fundus of the left eye. The optic disk and the macula are visible (from Medical gallery of Mikael Häggström 2014. WikiJournal of Medicine 1(2). DOI: 10.15347/wjm/2014.008. Public Domain).

Degradations of the retina can also be examined with the use of an *ophthalmoscope* invented by Helmholtz. In the early version a hemispherical or parabolic mirror was used for focusing light through the eye lens of the patient onto the retina. The light reflected back is observed by the examiner through a hole at the center of the mirror. The schematic of a more modern version is outlined in Fig. 11.17 (a) where the hole is replaced by a mirror or prism. For inspection of a large area of the retina the pupil needs to be widened. There are three versions presently used for imaging the *fundus* of the eye, i.e., the interior back side of the eye: direct ophthalmoscope gives an enlarged, virtual and upright image of the retina with a magnification of about 15; indirect ophthalmoscope yields a real inverted image with a magnification of about 8; scanning laser ophthalmoscope provides good images of the fundus even without dilation of the pupil. Figure 11.17 (b) shows an example of a normal undamaged fundus.

In most cases damage to the retina or the optic nerve due to glaucoma is irreversible as neither ganglions in the retina nor nerve fibers regenerate even if the IOP has been lowered by surgical means. Glaucoma surgery therefore serves the purpose of preventing further damage rather than restoring the visual field that has already been lost. Medication may reduce the IOP and surgery may increase the drainage of aqueous humor either by extra filtration or an artificial implant [9]. Presently, biocompatible *microstents* are being tested for increasing the outflow of aqueous humor, as displayed in Fig. 11.18. Nevertheless, glaucoma remains one of the main reasons for developing blindness. Unfortunately, pharmaceutical treatments and/or various types of

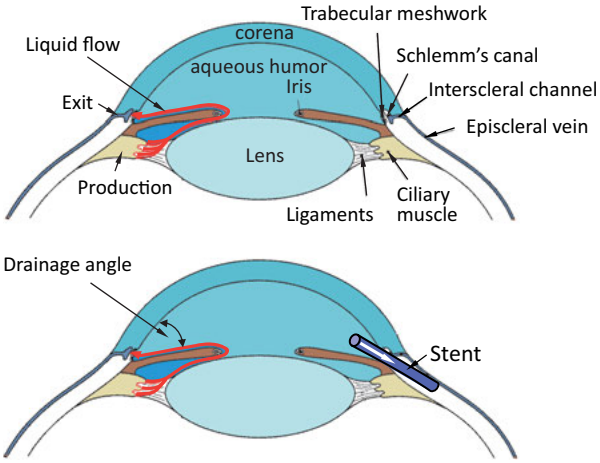


Fig. 11.18: Aqueous humor that fills the anterior chamber is produced by the ciliary body (muscle) and flows between the iris and lens, through the pupil and to the drainage angle at the junction of the iris and the cornea. Aqueous fluid exits the eye through the trabecular meshwork into Schlemm's canal, from where it goes back into the episcleral vein. Stent implants increase the outflow rate of aqueous humor to lower the intraocular pressure.

surgery are still by far less successful than treatments for cataract. Completely blind patients may gain some limited vision by implantation of epiretinal or subretinal microelectrode arrays discussed in more detail in Section 15.5.2/Vol. 2.

11.3 Photoreception and transduction

11.3.1 Structure of the retina

In this section we analyze how light that has reached the retina is detected by light sensitive receptors sparking an action potential that travels through the optic nerve to the visual cortex of the brain. Figure 11.19 shows a cross section of the retina consisting of several distinct layers. Starting from the proximal side and progressing to the distal layers, light first crosses a bundle of nerve fibers, then passes through a layer filled with ganglion cells (G); from there the light travels from the inner plexiform layer, a network of axons and dendrites, to the outer plexiform layer, passing three different types of cells: amacrine cells (A), bipolar cells (B), and horizontal cells (H) located in the inner nuclear layer. The outer plexiform layer contains nerve endings of bipolar cells, horizontal cells, and photoreceptor cells. Finally, the light is absorbed in two types of photoreceptors: rods and cones.

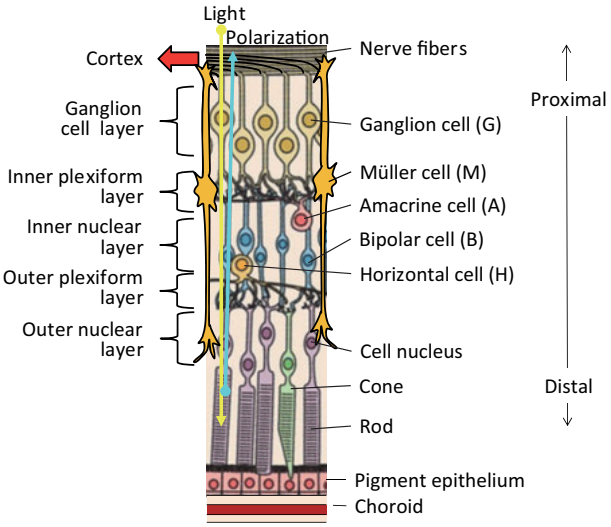


Fig. 11.19: Cross section of the retina showing different layers together with cell arrangement. Light comes from the top down (yellow arrow) and cell polarization travels from down up starting at cones and rods to the nerve fibers (cyan arrow).

The retinal *pigment epithelium* is a supply layer for retinal molecules used in photoreceptors for light absorption. Due to its black color caused by a high content of melanin, the epithelium layer absorbs most light that is not captured by the receptors and prevents it from reflecting back to the retina which would otherwise degrade image recognition. Thus, the pigment epithelium layer protects the photoreceptors from damaging levels of light intensity. Obviously the absorbance of the epithelium layer is not perfect since red eyes from back reflection can be seen on photos taken with a flash. The choroid contains blood vessels and capillaries for oxygen and nutrients, for collecting waste products, and for cooling the back side of the retina, which is exposed to high light intensity at the focal spot.

The retina is a highly complex arrangement of receptors and cells, all having their specific function to be detailed further below. In short, light is absorbed in cones with trichromatic color sensitivity and in rods with bright/dark sensitivity. Light causes a hyperpolarization in rods and cones that travels from the distal to the proximal side of the retina, and is processed by three different types of cells, horizontal cells (H), bipolar cells (B), and amacrine cells (A) in the inner nuclear layer, before arriving at the ganglion cells (G), which fire action potentials that are transmitted through the optic nerve to the visual cortex in the brain for final processing. Müller cells are intercalated between all previously mentioned cells, stretching from the outer nuclear layer to the ganglion layer. Müller cells support the metabolism of the retina by recycling neurotransmitters required for the synaptic activity of the cells in the inner retina. At the same time they lend structural support to the retinal network and guide light to the

photoreceptors, thereby reducing the S/N ratio and enhancing the light sensitivity of the receptors [10].

On the retina there are about 120×10^6 rods but only 7×10^6 cones, the ratio is almost 20 : 1. Rods and cones are unevenly distributed across the retina. Cones have a very high density in the region of the fovea (Fig. 11.20), spanning an area of about 1.5 mm in diameter. In the center of the fovea with a diameter of about 0.2 mm the cones are even slimmer than outside for enhanced density of about 150 000–200 000/mm^2. The fovea is the region of highest visual acuity and lies on the optical axis of the eye. Outside the fovea the cone density drops rapidly (compare Fig. 11.9). Rods, in contrast, are not present in the inner part of the fovea. Their density increases in the parafoveal region with increasing eccentricity up to a maximum and then drops again outside of the macula. Figure 11.20 shows once more a cross-sectional view of the retina in the area of fovea and macula. For increased light sensitivity the fovea is unobstructed by nerve fibers and blood vessels and can be recognized by a depression (fovea centralis) on the retina. Rods and cones converge into about 10^6 ganglion cells; each ganglion cell is connected to one nerve fiber, bundled together in the optic nerve that goes to the brain.

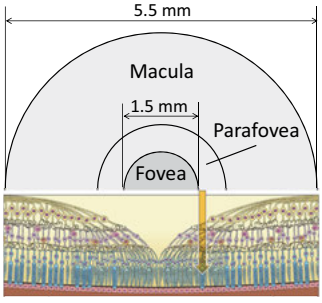


Fig. 11.20: Cross section of the retina in the area of the macula. The fovea is the region of maximum sensitivity where the density of cones is highest. The central rod-free region of the fovea has a diameter of about 1.5 mm.

It may appear odd that the nerve fibers are located at the proximal side of the retina and not on the distal side, which necessitates a blind spot on the retina. However, more important than the location of the nerve fibers is the contact of the rods and cones with their supply layer epithelium for light absorption and nutrients and with the choroid for oxygen uptake. Since the light absorption of the retinal epithelium is much higher than that of the nerve fibers, it must be located on the distal side.

11.3.2 Sensitivity and adaptation

Rods and cones form two independent visual systems, *scotopic* (darkness) system and *photopic* (brightness) system, respectively. The monochromatic vision of rods is adapted to work in dim light. In contrast, cones are adapted and optimized for trichromatic vision in daylight. The sensitivity regions for scotopic and photopic vision as a

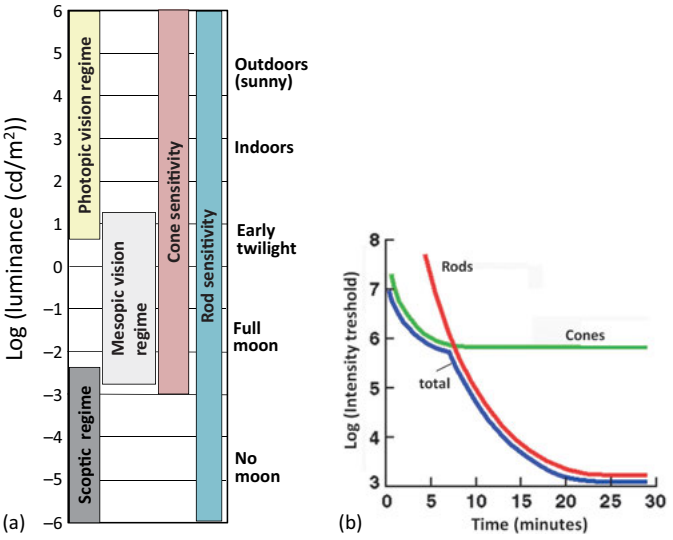


Fig. 11.21: (a) Dynamic range of photosensitivity of rods and cones. The regions for scotopic vision, photopic vision, and mesopic vision are also marked. (b) Dark adaptation of rods and cones as a function of time.

function of luminance are shown in Fig. 11.21 (a), the units are candela per square meter (cd/m^2). The total sensitivity range covers 10 to 11 orders of magnitude. This is an amazing dynamic range of light sensitivity, only matched by the auditory sensitivity, which also spans some 10–12 decades (Chapter 12). The lower limit of light sensitivity corresponds to single photon detection, whereas in the upper limit the retina is exposed to millions of photons per second.

In Fig. 11.21 (b) the *dark adaptation* of rods and cones is presented as a function of time. The threshold light intensity that a subject can recognize after switching off a bright light is plotted. Adaptation is fast in the beginning with sensitivity increasing by a factor of 500 after 5–8 minutes, followed by slower adaptation for the next 20–30 minutes during which sensitivity increases by another factor of 2000. The first fast drop is due to cones. However, their sensitivity levels off, and the additional sensitivity increase is due to rods. The combined curve shows a kink at the cross point, which is referred to as the *Kohlrausch kink*. This kink was an early indication for the existence of two visual systems in the eye. For enhancing the sensitivity three strategies are followed: increasing the pupil, regenerating visual pigment in the rods, and neural enhancement, which is discussed further below.

The sensitivity with respect to wavelengths is slightly different for rods and cones. Cones have a combined maximum sensitivity at 555 nm, whereas the sensitivity of rods is blue shifted to 500 nm, known as *Purkinje shift* (Fig. 11.22). Thus in the dark, blue objects can be better recognized than red objects, although both appear gray.

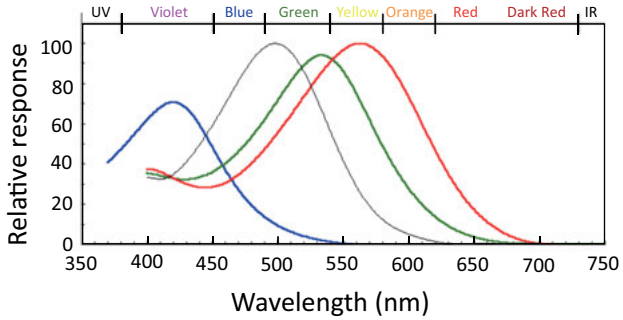


Fig. 11.22: Wavelength sensitivity of rods and cones. The sensitivity of rods and cones varies strongly and is highest for red cones.

Tab. 11.1: Main characteristics of the photopic and scotopic visual system.

Property	Photopic	Scotopic
Receptor	cones	rods
Color	tricromatic	monochromatic
Pigment	rhodopsin	rhodopsin
Sensitivity	low	high
Usage	daylight	twilight
Location	fovea	outside fovea
Acuity	high in fovea	low overall

The absorption maximum of blue cones is at 420 nm, of green cones at 535 nm, and of red cones at 565 nm. In Tab. 11.1 the main characteristics of the scotopic and photopic visual system are compared and summarized.

As previously mentioned and also shown in Fig. 11.22, the bandwidth of visual sensitivity covers the wavelength range from 400 nm to 700 nm at most. This bandwidth is provided by the sensitivity of rods and cones. If we could invent a method to increase the visual bandwidth of the receptors, would the optical system consisting of cornea, aqueous humor, lens, vitreous humor be transparent to this increased bandwidth? Rephrasing the question: what is the *spectral transmission* range of the eye's optical system? Measurements have shown that optical transparency stretches from 400 nm to about 1400 nm (see Fig. 11.23), but starting from 700 nm we are blind. So, another receptor in this "blind" region would be useful. Night vision devices have sensors with sensitivity in the infrared region that is then wavelength shifted by photocathodes and phosphorous screens to the visible range. Blindness in the infrared regime can be dangerous when working with infrared lasers because of the high intensity of lasers on the one hand and the lack of immediate retraction sensitivity on the other hand. Corneal burns and cataract development may result from laser exposure. Similarly, workers with exposure to high intensity infrared light may suffer from

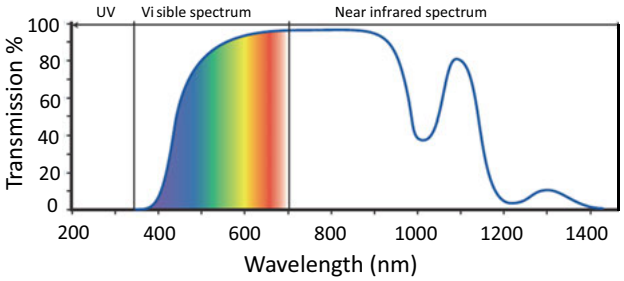


Fig. 11.23: Spectral transmission of the eye's optical system. The light sensitivity of rods and cones is shown in colors. The sensitivity band of the receptors matches the optical transmission band at small wavelengths in the UV region. In the infrared region the transparency of the optical system is much more extended than the sensitivity of the receptive field (adapted from [11]).

cataract, such as glass blowers and steel workers. The same potential danger also applies to UV radiation and beyond. However, in the UV range the eye lens is no longer transparent, but damage can nevertheless occur by absorption in the skin causing not only tanning but eventually skin cancer.

11.3.3 Phototransduction

Now we examine rods and cones in more detail and try to understand how the stimulus (photons) is converted into membrane potentials, called *phototransduction*. The essential parts of rods and cones are shown in Fig. 11.24. We distinguish an outer segment and an inner segment. The outer segment contains a stack of membrane disks for holding the visual pigment rhodopsin. The inner segment is made up of the essential cell components, mitochondria, nucleus, and dendrites. The synaptic terminal connects to bipolar cells across synaptic gaps. Light passes from the inner to the outer segment and is eventually absorbed by visual pigments in the disk membranes.

Each rod contains about 1000 disks. The disks are permanently renewed. The ones next to the pigment epithelium are being destroyed and fresh ones grow at the connecting cilium. About 10^6 visual pigments are embedded in the membrane of each disk (see lower right inset in Fig. 11.24). Therefore each rod contains about 10^9 pigments. Each pigment consists of two parts: the protein opsin and the light sensitive retinal; both together are called rhodopsin. The protein opsin is composed of seven transmembrane helices connected by loops. It belongs to the class of so called G protein coupled receptors that are also responsible for taste and smell. In rhodopsin the light sensitive receptor is an 11-cis retinal, which is an aldehyde of vitamin A_1, i.e., an H–C=O group attached to vitamin A. Rhodopsin attached to a membrane is sketched in Fig. 11.25. The membrane is a dense package of lipid molecules arranged in a double layer with hydrophobic tails inside and hydrophilic heads sticking out into the cytoplasm.

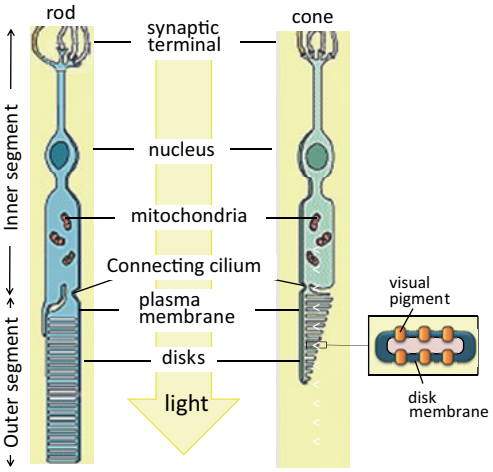


Fig. 11.24: Inner and outer segments of rods and cones. The outer segment holds a stack of disks which contain visual pigments.

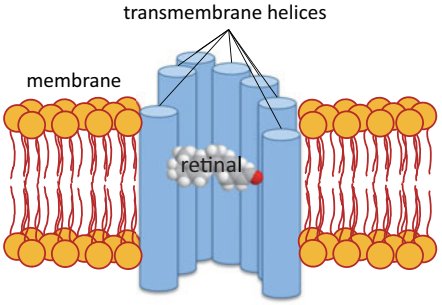


Fig. 11.25: Rhodopsin, composed of the protein opsin and retinal, is embedded in the disk membrane. The protein opsin is composed of seven transmembrane helices, connected by loops (here omitted for clarity). Each disk in rods and cones contains about 10^6 rhodopsin proteins.

When a photon is absorbed, *photoisomerization* of retinal takes place from the bent 11-cis conformation to the straight all-trans form (Fig. 11.26). This conformational change requires a rotation about one of the double bonds by 180°. The trans-state is also referred to as excited retinal R*. The cis-trans transition sketched in Fig. 11.26 takes less than 100 fs [12]. This ultrafast photochemical reaction initiates vision. The excited trans form remains in the opsin until transducin molecules are activated. Then the trans-retinal R* no longer fits into opsin; the weak covalent bonds between retinal and opsin are broken and all trans retinal molecules are released from the opsin protein, a process which is referred to as 'bleaching out'.

Opsin is now in an inactivated state and all parts must be recycled for the next photon absorption. Figure 11.27 shows the recycling process. Once opsin and retinal have split up, the retinal drops into the stretched ground state. To restore the kinked state for reassembling into the opsin, energy is required and delivered by an ATP-ADP enzymatic reaction, where ATP is supplied by mitochondria in the inner segment. Opsin and retinal rejoin and are ready for the next photon absorption. The energy landscape for retinal is shown in Fig. 11.28 for a complete cycle.

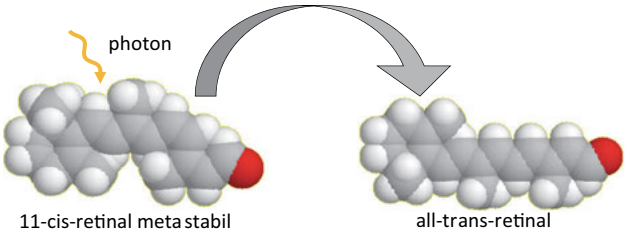


Fig. 11.26: Photoisomerization of retinal from the kinked 11-cis shape to the all-trans shape. The red ball indicates an oxygen atom.

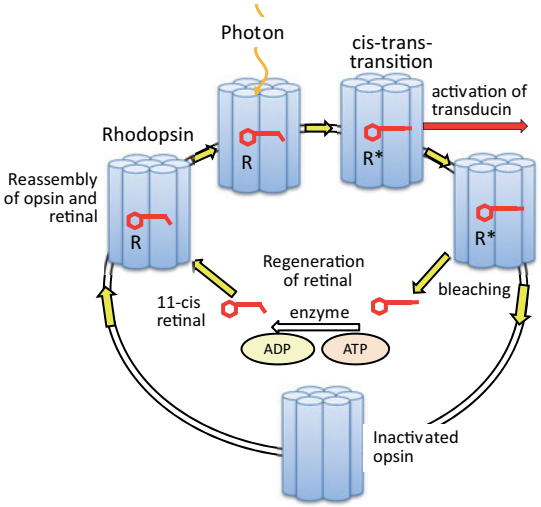


Fig. 11.27: Cycle of rhodopsin after photon absorption. Retinal goes from the cis to the trans form, activates transducin and is released from the opsin. Retinal is regenerated with the help of the ATP-ADP enzymatic process and then restored as rhodopsin.

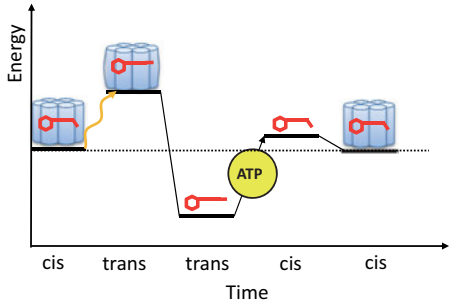


Fig. 11.28: Energy landscape for a complete retinal cycle. The ground state is an isolated and stretched retinal. Bending requires energy delivered by ATP. The energy is first lowered by the opsin-retinal reaction, followed by a strong increase during photoisomerization in a confined environment, causing the retinal molecule to stretch and the opsin molecule to strain. After the release of the retinal, it goes back into the ground state.

Usually in cells the K^+ channels are open and the Na^+ channels are closed, yielding a resting potential of about –70 meV (see Chapter 5). However, this is not so in photoreceptors. Here the Na^+/Ca^{++} channels in the outer segments are gated by cyclic guanosine monophosphate (cGMP) molecules, i.e., cGMP molecules attach to the Na^+/Ca^{++} channels and keep them open. Because of open Na^+/Ca^{++} channels the transmembrane potential is only –30 mV. Photoisomerization sets off three interlinked chemical reaction cycles in the cytoplasm of the receptor cells with the goal to reduce the cGMP concentration: (1) rhodopsin cycle (shown in Fig. 11.27); (2) transducin cycle, and (3) phosphodiesterase (PDE) cycle (Fig. 11.29). Altogether they result in a reduction of cGMP concentration by hydrolyzing cGMP to GMP. Once this is achieved, the cGMP-gated Na^+/Ca^{++} channels close, the K^+ channels in the inner segment of the receptor cell remain open, and the cell potential hyperpolarizes to –70 mV. Conversely, in the dark, cGMP binds back to Na^+/Ca^{++} channels and opens them again, accounting for dark current in rods and cones.

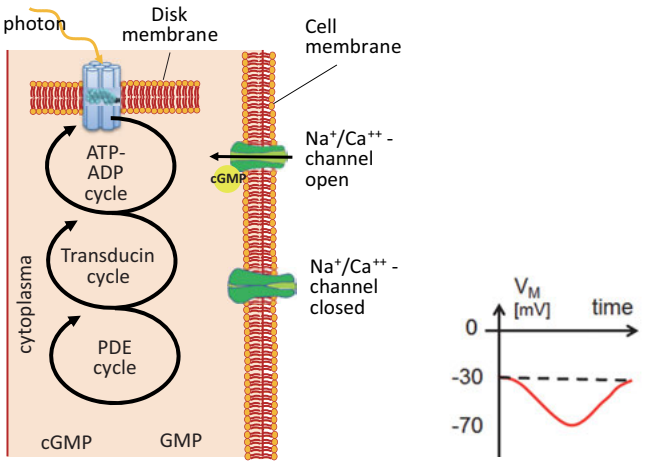


Fig. 11.29: Photoisomerization of retinal triggers three reaction cycles. The ATP-ADP cycle restores the retinal. The transducin cycle and the PDE cycle are responsible for reducing the cGMP concentration in the cytoplasm so that the cGMP-gated Na^+ channels close and the transmembrane potential V_M drops from –30 mV to –70 mV.

Phototransduction takes now the following main steps. In the dark state the rods are depolarized to –30 mV and the Na^+/Ca^{++} channels are open. Simultaneously at the synaptic terminals the transmitter glutamate is emitted and elicits inhibition that prevents one class of subsequent bipolar cells from depolarization. Upon light absorption the receptor potential drops to –70 mV, causing a reduction of transmitter emission and inhibition at the synapse. The bipolar cells can now depolarize and transmit the depolarization to the ganglion cells. Ganglion cells in contrast to all other A, B, H cells,

fire action potentials in response to depolarization whenever a threshold potential is exceeded. Thus hyperpolarization in the receptor cells will finally generate action potentials in ganglion cells after having passed a couple of intermediate steps, which are detailed in the next section. The transduction sequence from hyperpolarization to action potential is schematically summarized in Fig. 11.30. More light causes stronger hyperpolarization of receptors and stronger depolarization of bipolar cells, resulting in a higher frequency of action potentials in G cells. The omnipresent Müller cells are actively involved in all parts of phototransduction by regulating the synaptic activity via glutamate uptake and reprocessing [10].

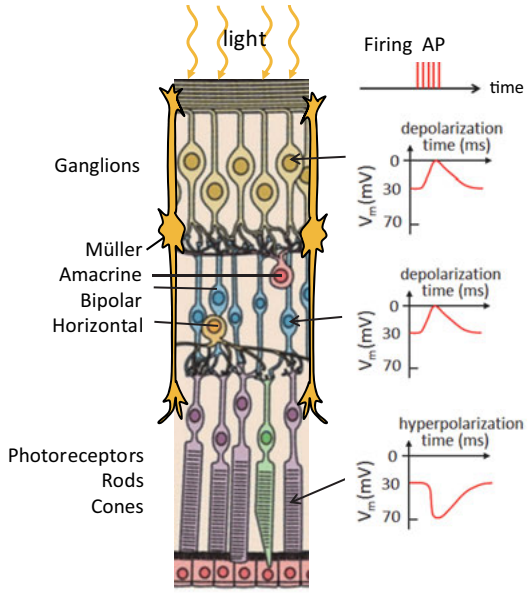


Fig. 11.30: From hyperpolarization to depolarization and action potential in the ganglion cells.

Two distinctive and remarkable features in the phototransduction process are different to all other receptor cells. First, stimulation of photoreceptors causes hyperpolarization instead of depolarization. And second, in the dark inactivated state the Na^+/Ca^{++} channels are open, generating dark current. Thus the inactivated state is in fact the most active state.

Now we want to discuss the amazing quantum efficiency of photon detection that our eyes achieve. The absorption of only one photon activates one retinal R to the excited state R* in a matter of 100 fs. The excited state activates about 1000 transducin molecules within 100 ms. Each transducin molecule interacts with one PDE, and each PDE deactivates about 1000 cGMP molecules. This results in an amplification factor of at least 10^6. The cGMP concentration in the cytoplasm then drops by about 8 %, closing about 250 Na^+ channels out of a total of 10^4 channels, i.e., closing some 2.5 % of all channels. This is sufficient to fire an action potential and recognize a single photon

in the visual cortex! [13, 14] The cones and rods of the retina have indeed been considered one of the best known quantum detectors [14]. However, only 8 out of 100 photons that cross the cornea are absorbed in retinal, the others are lost by scattering, reflection, or absorption in the pigment epithelium. A recent re-evaluation of the quantum efficiency of photons in the human eye shows that one has to distinguish between absorptive quantum efficiency of the receptors and light perception by a test subject [15]. Any photon that is absorbed causes a receptor response. However, the lowest number of photons crossing the cornea that a test subject perceives as a light flash is on average about 70 photons [15]. Assuming that more than 90 % of the photons are lost on the path to the photoreceptor, the perceived quantum efficiency is still about 20 %.

While the sensitivity of the retina is extremely high, data processing is rather slow, although the initial step is on the femtosecond time scale. The slowdown is due to the three reaction cycles shown in Fig. 11.29, such that the information processing rate is in the order of 10–50 Hz. Picture series with 20 frames per second are already perceived as a movie when rendered, but single shot pictures can eventually be recognized on shorter time scales.

Signal transduction in cones is similar to that in rods. There are three types of cones, named according to the color of light they absorb: blue, green, and red, the absorption spectra are shown in Fig. 11.22. The retinal in all three color cones is identical, but the binding sites in the opsin molecule are slightly different due to slightly different amino acid sequences in opsin. The response time of cones is faster than that of rods, but their overall sensitivity is lower. The density of red, green, and blue cones in the fovea is not equal and not homogeneously distributed. Red cones are most frequent, followed by green cones, blue cones have the lowest density in the foveal area [16]. The generation of intermediate hues is due to activation of more than one type of cone at the same time.

11.3.4 Retinal signal processing

The retina is not only a photoreceptor. It is also an integral part of signal processing that starts in the inner nuclear layer of the retina and is completed in the visual cortex of the brain. Figure 11.31 symbolically shows six different cells that participate in information processing, the Müller cells are left out for clarity.

Signal processing in the retina has its origin in the hyperpolarization of the receptor potentials. Receptor cells can only hyperpolarize from an intermediate negative transmembrane potential to a lower potential, for instance from –30 mV to –70 mV. In the retinal network the hyperpolarization is transmitted via synaptic connections to bipolar cells. Depending on the input they may either hyperpolarize or depolarize. The response of bipolar cells is communicated by synaptic connections to ganglion cells. The horizontal cells are responsible for lateral connections between bipolar cells and for lateral inhibition. The amacrine cells connect bipolar cells of rods with ganglion

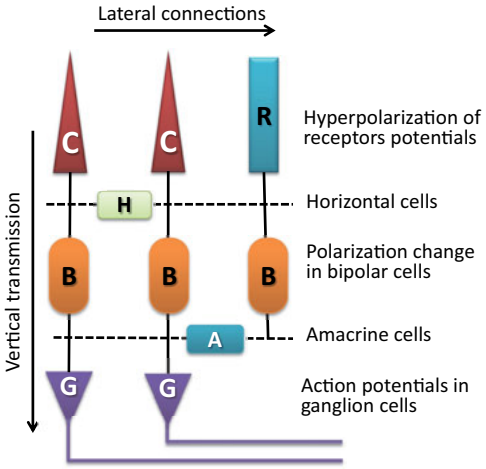


Fig. 11.31: Schematic of the six different cells in the retina with lateral and vertical connections.

cells. Like the horizontal cells, amacrine cells make lateral connections, and in most cases their action is inhibitory. The ganglion cells fire action potentials in response to depolarization of bipolar cells. One nerve fiber is attached to each ganglion cell that connects to the visual cortex.

Each receptor in the fovea attaches to a single bipolar cell and a single ganglion cell (compare Fig. 11.20). There are exclusively cones within the fovea. Each cone in the fovea has a direct line to the brain allowing an exact registration of the input location. Outside the fovea several receptors, mainly rods, converge to one ganglion cell as schematically shown in Fig. 11.32. With increasing eccentricity the number of rods connecting to one ganglion cell increases and reaches up to 150.

Summation of rod signals increases the sensitivity for bright/dark contrast, but the spatial resolution (acuity) suffers outside of the macula region. Vice versa, in the central rod-free region of the fovea with a diameter of about 500 µm the density of cones is highest and the 1 : 1 ratio between cones and ganglions supports this high spatial resolution. Indeed, the foveal region provides the highest visual acuity. Regions

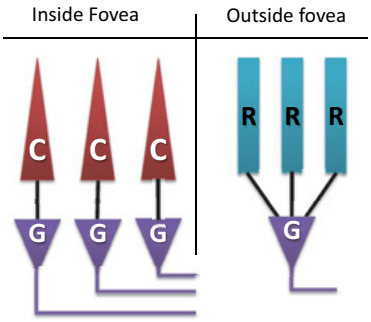


Fig. 11.32: Receptor fields inside and outside of the fovea in a simplified version without showing bipolar cells. Inside the fovea there is a 1 : 1 relation between cones and ganglions. Outside there is convergence of many rods to one ganglion.

on the retina with high spatial resolution require more space for image processing in the brain.

Now we will investigate the pathway of receptor potentials to ganglion action potentials in more detail. Bipolar cells receive synaptic input from either rods or cones, but not from both. They are designated as rod bipolar or cone bipolar cells, respectively. Bipolar cells communicate via graded potentials, rather than action potentials. Only ganglion cells produce action potentials when stimulated by bipolar cells.

The signal pathway and graded potential variations for cones is tabulated in Fig. 11.33. As already mentioned, after light is absorbed in rhodopsin, Na^+/Ca^{++} channels close and cause a hyperpolarization. The hyperpolarization, in turn, reduces the *glutamate* concentration in the synaptic space between cone and bipolar cell. Glutamate is a neurotransmitter that controls ion channels in the synaptic space (see Chapter 6). Cones connect to two antagonistic bipolar cells: metabotropic *ON bipolar cells*, and ionotropic *OFF bipolar cells*. They react oppositely with respect to glutamate concentration: if the glutamate concentration is high, ON bipolar cells close Na^+/Ca^{++} channels causing hyperpolarization whereas OFF bipolar cells open Na^+/Ca^{++} channels, causing depolarization. Hyperpolarization of cones activates ON bipolar cells by depolarization, depolarization of cones deactivates ON bipolar cells by hyperpolarization, and oppositely for the OFF bipolar cells. The ON (OFF) bipolar cells connect to ON (OFF) ganglion cells, which react in the same way. The various dependencies are summarized in Fig. 11.33 for future reference. We note that ON bipolar cells always reverse the potential with respect to receptor potential, while OFF bipolar cells keep the same potential as the receptor.

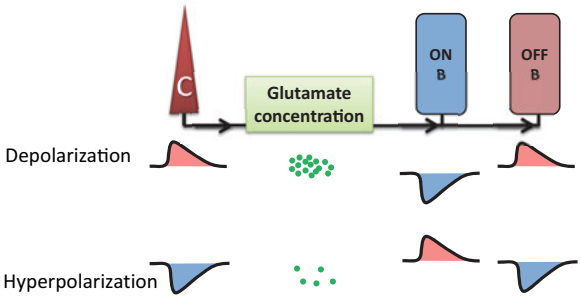


Fig. 11.33: Transmission of graded potential states from a cone to bipolar cells.

Now we are prepared to discuss the sequence of actions that takes place when light is turned on and off. We start with a bright state and then switch to the dark state. The dark response is shown in Fig. 11.34 (a). Na^+/Ca^{++} channels in the cone are open and open further, causing a depolarization of the cone. The glutamate concentration in the synaptic space is high causing Na^+ channels in ON-center cells to close and to turn off ON-center ganglion cells. Vice versa, the same receptor state causes the OFF bipolar

cell and the OFF ganglion to be activated. In the activated state the OFF ganglion cells fire a high frequency action potential, whereas the hyperpolarized ON ganglion cells are deactivated. When a light is turned on, the situation is reversed (Fig. 11.34 (b)): the receptor cell hyperpolarizes, ON cells become depolarized, and ON ganglion cells fire action potentials, while OFF cells are deactivated.

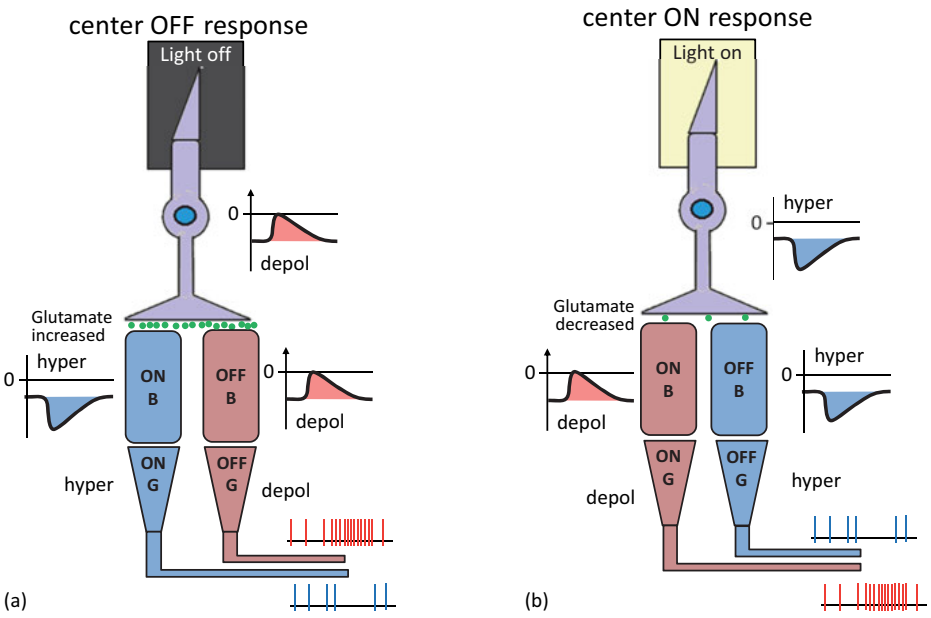


Fig. 11.34: Signal pathway from cones to ganglion cells (G) via two antagonistic bipolar cells, ON and OFF bipolar cells (B). (a) Bipolar potentials for darkness; (b) potential changes when a light is turned on.

Why do we have two antagonistic cells that communicate the same state to the brain? With a single light sensitive detector it should be possible to distinguish between light and dark, and any shades in between. However, the eye uses two detectors, one responsible for brightness and the other for detecting darkness. The dual system gives us a much more precise and weighted impression of the state of brightness or darkness. The two systems are compared in Fig. 11.35. A single detector measures intensity on an absolute scale, taken here as normalized from 0 to 100 in arbitrary units. The visual system with two antagonistic detectors, in contrast, determines brightness and darkness independently and processes the difference. This is a comparative and weighted measurement instead of a single and absolute measurement. It allows judging on which level contrast takes place, on a high or lower level of darkness versus brightness.

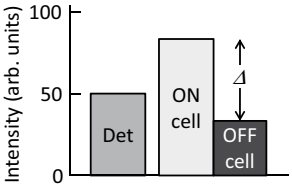


Fig. 11.35: Intensity measurement with a detector system (left) and with the antagonistic visual system (right).

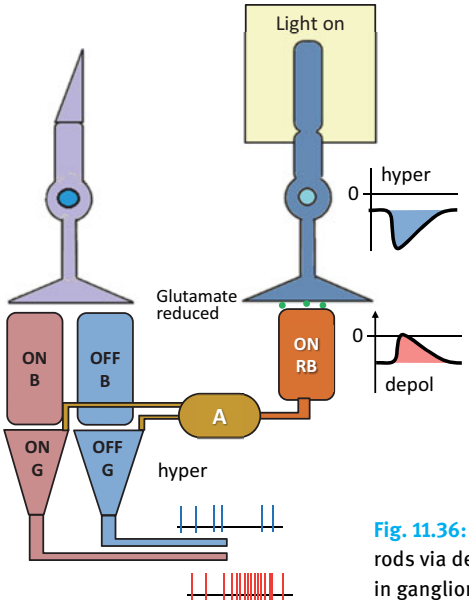


Fig. 11.36: Signal pathway in rods from hyperpolarization of rods via depolarization of bipolar cells to action potentials in ganglion cells taking a detour involving amacrine cells.

The signal pathway for rods is shown in Fig. 11.36. In rod bipolar cells (RB) reduced glutamate concentration in response to light leads to an opening of Na^+/Ca^{++} channels and consequently to depolarization. This excites an amacrine cell, which in turn, connects to ON and OFF synaptic connections at ganglion cells. With this detour, rods trigger the same on-off response to light as cones do. The detour is necessary as rods do not connect directly to ganglion cells.

11.3.5 Receptive fields

Each of the five neurons (photoreceptor, bipolar, amacrine, horizontal, ganglion) covers an area of vision in the retina. This spatial area, where an appropriate stimulus (light) modifies the activity of a particular neuron, is called the *receptive field* of this neuron.

The receptive field of a single photoreceptor cell may be limited to a tiny spot of light that corresponds to the precise location of this receptor on the retina. However,

with successive layers of the retina the receptive field increases in lateral extension and becomes increasingly complex because of numerous lateral and vertical interconnects. Furthermore, the size of the receptive field will depend on its location in the retina. In the foveal area it is small but increases with increasing eccentricity.

As an example we consider the receptive field of bipolar cells. The receptive field of each bipolar cell is approximately circular. But the *center* and the surrounding area of each circle have opposite responses: a light ray striking the center of the field has the opposite effect of one striking the surrounding area, called "*surround*". The difference is due to the ON and OFF bipolar cells.

A light stimulus applied to the center of a receptive field will cause the ON-center cell to be activated (see Fig. 11.34). The same light hitting the surround has the opposite effect on such an ON-center cell. In the surround, an ON-center cell will be deactivated while OFF-center cells are being excited. This is due to horizontal cells that connect the receptors in the lateral direction, as shown in Fig. 11.37. All signals become inverted by illumination of a spot in the surround: center receptors react by OFF response, surround receptors react by ON response. Conversely, light at the center turns off the OFF-center cells, while light on the surround turns on the OFF-center cells.

Just like bipolar cells, ganglion cells have concentric receptive fields with a center-surround antagonism. But contrary to the two types of bipolar cells, ON-center ganglion cells and OFF-center ganglion cells do not respond by depolarizing or hyperpolarizing, but rather by increasing or decreasing the frequency with which they dis-

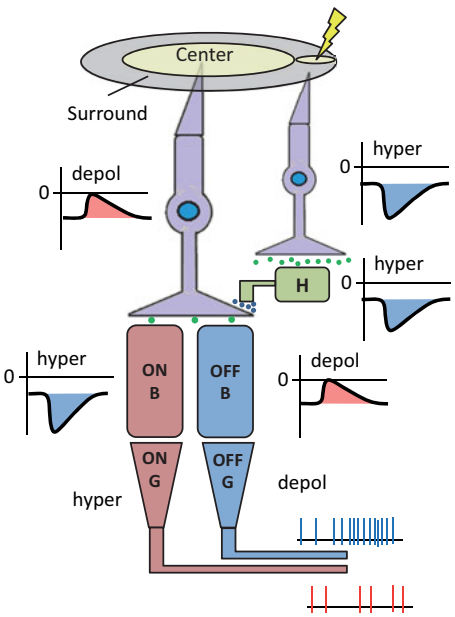


Fig. 11.37: Light in the surround of a receptive field has an inhibiting effect on ON-center bipolar cells due to the intervention of horizontal cells connecting receptors in the lateral direction.

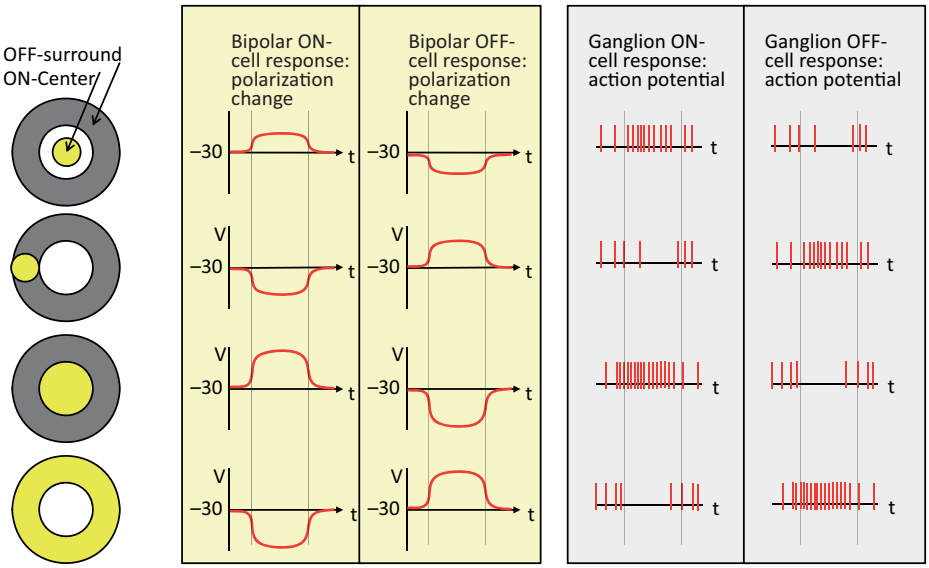


Fig. 11.38: Response of bipolar ON (OFF)-center cells and ganglion ON (OFF) cells for various situations from top to bottom: small light spot in the center of a receptive field; small light spot in the surround; complete illumination of the center; complete illumination of the surround.

charge action potentials. The response of bipolar cells and ganglion cells to light in the center of a receptive field and its surround is shown for several cases in Fig. 11.38. The antagonistic reaction of ON and OFF bipolar cells for center and surround illumination becomes quite obvious.

Center and surround are always connected by intervening horizontal cells that gradually change the potentials of the center. The response to stimulation of the center of the receptive field is always slightly inhibited by simultaneous stimulation of the surround.

Lateral inhibition serves the purpose of contrast enhancement. This is exemplified in Fig. 11.39. The receptor field 'observes' two connected areas with different intensities. Lateral inhibition leads to an overall lower response, but with an enhanced contrast at the border between two different areas. Without lateral inhibition the contrast in this example is 100 : 20 = 5. With a 10 % lateral inhibition of the receptor signal the contrast increases to 88 : 16 = 5.5. This is a 10 % contrast enhancement corresponding to the 10 % lateral inhibition. Although the effect is not dramatic, it is still extremely important for our visual perception, such as reading.

Lateral inhibition in the retina is a gentle process of modifying the receptor potentials. If we take it to the extreme and invent a horizontal cell that inverts all lateral potentials by 100 %, as shown schematically in Fig. 11.40, we could reach contrast enhancement by a factor of 4. Image processing can take advantage of the principle of lateral inhibition. A vivid example of contrast enhancement is shown in Fig. 11.41

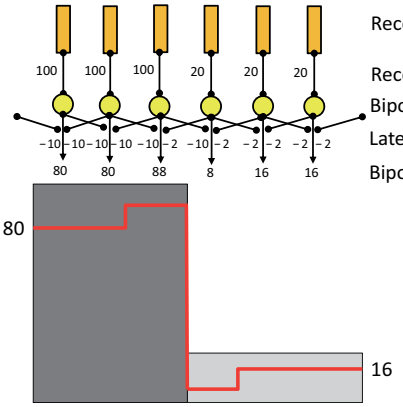


Fig. 11.39: Contrast enhancement by lateral inhibition.

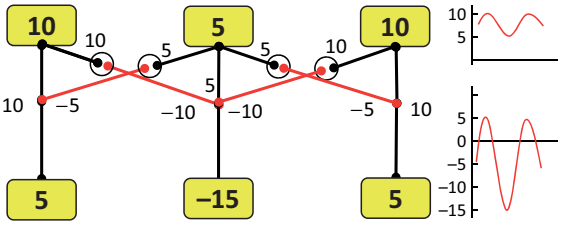


Fig. 11.40: Contrast enhancement by lateral inhibition via 100 % inversion of the receptor potential.

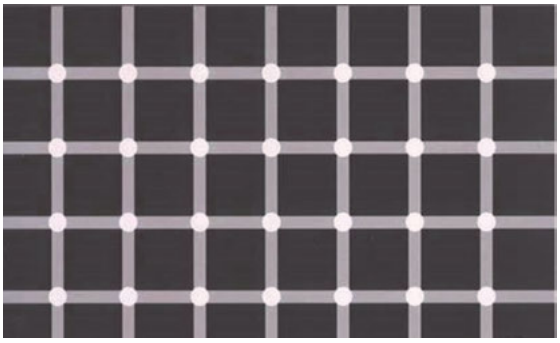


Fig. 11.41: A variant of the original Hermann lattice shows an optical illusion of disks at the center of cross points between black squares due to center enhancement and lateral inhibition. The dots switch between white and black.

where white and black disks flicker in the receptive field at the cross points because of switching between center and surround.

There are many more fascinating aspects of vision such as color perception, perception of space, modulation of visual signals by the thalamus, and the entire processing of action potentials in the visual cortex of the brain. There is not sufficient space in this volume to treat all these aspects. The interested reader is therefore referred to the respective books and publications listed under Further reading.

11.4 Summary

1. For the eye, the image distance is fixed and the focal length is variable.
2. The iris controls the aperture of the eye and thereby the intensity.
3. 75 % of the refraction power of the eye is provided by the refractive index and curvature of the cornea.
4. The lens contributes another 15 dpt to the refraction power of the eye, which is in total 58 dpt.
5. The lens has the ability to change the curvature for changing accommodation.
6. Age-related accommodation loss is due to an elastic stiffening of the lens.
7. The optical resolution determined by the opening of the pupil matches well with the separation of rods and cones on the retina.
8. The visual acuity of the eye is about 1′, corresponding to separation of images on the retina at a distance of 5 µm.
9. Cataract is caused by increasing opacity of the eye and can be well treated with an artificial lens.
10. Glaucoma is due to overpressure in the eye, which may lead to damage of the retina and optic nerve.
11. The retina contains two visual systems: photopic supported by cones, and scotopic supported by rods.
12. Rods provide bright–dark sensitivity, cones provide color sensitivity and brightness.
13. The dynamical intensity range of the eye covers 10 orders of magnitude.
14. The spectral band width of the receptors ranges from 400 nm to 700 nm.
15. In the foveal region cones with green/yellow and red sensitivity dominate.
16. The primary process of vision is the cis-trans conversion of retinal by photoabsorption.
17. Splitting of opsin and retinal triggers three reaction cycles. Their combined result is a hyperpolarization of the receptor potential and a recycling of retinal.
18. Cones are connected one by one to ganglion cells. In contrast, signals from many rods converge to one ganglion cell.
19. Each cone is connected to two antagonistic bipolar cells, which determine the difference between brightness and darkness.
20. Signals from bipolar cells to ganglion cells are modulated by horizontal and amacrine cells.
21. Signals from the surrounds act inhibitively on receptor potentials in the center.
22. Lateral connections are in most cases inhibitive and enhance contrast.
23. Only ganglion cells can fire action potential.
24. Any nerve fiber from a ganglion cell goes to the brain.

References

[1] Wang H, Lin S, Liu X, Kang SB. Separating reflections in human iris images for illumination estimation. Tenth IEEE International Conference on Computer Vision. 2005; 2: 1691–1698.
[2] Uhlhorn SR, Borja D, Manns F, Parel JM. Refractive index measurement of the isolated crystalline lens using optical coherence tomography. Vision Research. 2008; 48: 2732–2738.
[3] Glasser A. Restoration of accommodation: surgical options for correction of presbyopia. Clinical and Experimental Optometry. 2008; 91: 279–295.
[4] Michael R, Bron AJ. The ageing lens and cataract: a model of normal and pathological ageing. Phil Trans R Soc B. 2011; 366: 1278–1292.
[5] Navarro R. The optical design of the human eye: A critical review. J Optom. 2009; 2: 3–18.

[6] Kingsley CN, Brubaker WD, Markovic S, Diehl A, Brindley AJ, Oschkinat H, Martin RW. Preferential and specific binding of human alpha b-crystallin to a cataract-related variant of gamma s-crystallin. Structure. 2013; 3: 2221–2227.
[7] Cataract Surgery in 6 minutes. Narrated by Dr.Sibley, Florida Eye Center, https://www.youtube.com/watch?v=rUCoQzui704
[8] Hennig A, Kumar J, Yorston D, Foster A. Sutureless cataract surgery with nucleus extraction: outcome of a prospective study in Nepal. Br J Ophthalmol. 2003; 87: 266–270.
[9] Richter GM, Coleman AL. Minimally invasive glaucoma surgery: current status and future prospects. Clinical Ophthalmology. 2016; 10: 189–206.
[10] Bringmann A, Grosche A, Pannicke T, Reichenbach A. GABA and glutamate uptake and metabolism in retinal glial (Müller) cells. Front Endocrinol. 2013; 4: 48.
[11] Boettner EA, Wolter JR. Transmission of the ocular media. Invest Ophthal. 1962; 1: 776–783.
[12] Wang Q, Schoenlein RW, Peteanu LA, Mathies RA, Shank CV. Vibrationally coherent photochemistry in the femtosecond primary event of vision. Science. 1994; 266: 422–424.
[13] Hecht S, Schlaer S, Pirenne HP. Energy, quanta and vision. J Gen Physiol. 1942;25:819–840.
[14] Rieke F, Baylor DA. Single-photon detection by rod cells of the retina. Rev Modern Phys. 1998; 70: 1027–1036.
[15] Manasseh G, de Balthasar C, Sanguinetti B, Pomarico E, Gisin N, Grave de Peralta R, Gonzalez Andino SL. Retinal and post-retinal contributions to the quantum efficiency of the human eye revealed by electrical neuroimaging. Frontiers in Psychology. 2013; 845(4): 1–13.
[16] Wikler KC, Rakic P. Distribution of photoreceptor subtypes in the retina of diurnal and nocturnal primates. The Journal of Neuroscience. 1990; 10: 3390–3401.

Further reading

Kandel ER, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth AJ. Principles of neural science. 5th edition. McGraw Hill; 2013.
Purves D, Augustine GJ, Fitzpatrick D, Katz LC, LaMantia AS, McNamara JO, Williams SM, editors. Neuroscience. 2nd edition. Sunderland (MA): Sinauer Associates; 2001. Online textbook can be accessed but not browsed: www.ncbi.nlm.nih.gov/books/NBK11059/
Hubel D. Eye, Brain, and Vision. Online textbook: http://hubel.med.harvard.edu/book/bcontex.htm
Schwartz SH. Visual perception: A clinical orientation. 4th edition. McGraw-Hill Prof Med/Tech; 2009.
Kolb H, Nelson R, Fernandez E, Jones B, editors. Webvision. The organization of the retina and visual system. Free Webbook at http://webvision.med.utah.edu/
Pape H-C, Kurtz A, Silbernagel S, editors. Physiologie. 7th edition. Stuttgart, New York: Thieme Verlag; 2014.
Duane's Ophthalmology: www.oculist.net/downaton502/prof/ebook/duanes/index.html

12 Sound and sound perception

12.1 Introduction

Hearing is the ability of the ear to detect soundwaves, to convert pressure waves into receptor potentials, and to process the signals in the auditory cortex of the brain. This entire process is known as auditory perception. It is one of the most important senses of humans next to visual perception and one of the most intriguing from a neuroscience point of view. Similar to smelling or seeing, hearing is a remote sense, i.e., the source of information is at a certain distance from the body and requires a medium to be detected by specialized receptors. In the case of sound, a compressional pressure wave travels through air as the transmitting medium between source and ear. As a soundwave detector the ear consists of three main parts (Fig. 12.1): (1) the external ear with the pinna (auricle), the roughly 25 mm long ear canal (auditory meatus), and at the end the eardrum (tympanum or tympanic membrane); (2) the middle ear with the ossicles hammer (malleus), anvil (incus), and stapes (stirrup); (3) the inner ear with a bony spiral (cochlea) that contains the actual sound-sensitive nerve receptors. These three parts of the ear have different tasks: the outer ear acts as a sound resonator, the middle ear is responsible for impedance matching of soundwaves between air and body, and the inner ear contains receptors that are sensitive to pressure amplitudes and frequencies and convert these wave properties into electric signals transmitted via the auditory nerve to the auditory cortex in the brain.

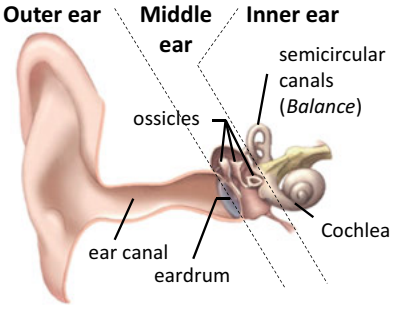


Fig. 12.1: Overview of the different parts of the ear (adapted from https://www.ncbi.nlm.nih.gov/pubmedhealth/).

Audibility implies coding frequencies and amplitudes of soundwaves into electrical signals that the brain "understands". Nature does not use Faraday's laws of induction. If it did, a simple microphone would suffice to convert pressure waves into a corresponding electrical signal output. Instead, nature works with receptors and receptor/action potentials. Therefore, another signal transduction has to be invented that converts mechanical motion into electrical signals, which indeed takes place highly efficiently in the ear.

DOI 10.1515/9783110372830-014

Before discussing the auditory pathway from the outer ear to the cochlea we will first derive some important equations of sound propagation in different media. At this point it should, however, not be omitted that the ear has another essential sensor attached to the inner ear: the vestibular apparatus consisting of semicircular canals which sense the body's balance. In contrast to the hearing organ, the vestibular apparatus is not in contact with the outside. From a physical point of view it is highly interesting how a sense of balance is achieved by fluid flow and tilting of nanocrystals. Unfortunately, there is not enough space to discuss this in any detail, but it is recommended as further reading in textbooks on medical physiology listed at the end under Further reading.

12.2 Soundwaves

The main task of the ear is to make soundwaves audible. Soundwaves are propagating pressure waves that displace molecules from their 'equilibrium' position in gases or liquids. We consider only longitudinal compression waves since transverse waves have no restoring force in gases or liquids and therefore cannot propagate. For longitudinal waves the propagation direction and the displacement amplitude are parallel. In the following we assume that the propagation direction, the displacement amplitude, and the velocity all point along an arbitrary x direction.

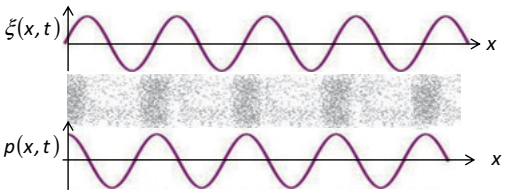


Fig. 12.2: Displacement and pressure of a propagating longitudinal soundwave.

The displacement $\xi(x, t)$ of molecules as a function of position x and time t can be described by a harmonic wave (Fig. 12.2):

$$\xi(x, t) = \xi_0 \sin(kx - \omega t),$$

where $k = 2\pi/\lambda$ is the wavenumber and ω is the frequency of the soundwave propagating along the x direction. The harmonic soundwave is a solution of the one-dimensional wave equation:

$$\frac{B}{\rho}\frac{\partial^2 \xi}{\partial x^2} = \frac{\partial^2 \xi}{\partial t^2}.$$

Here B is the compression modulus of gases or liquids and ρ is the respective density. The *phase velocity* of soundwaves is:

$$v_{\text{sound}} = \frac{\omega}{k} = \sqrt{\frac{B}{\rho}}.$$

For ideal gases the *compression modulus* follows from the ideal gas equation and is $B = p$ for isothermic processes and $B = \gamma p$ for adiabatic processes. Here γ is the ratio of the specific heat at constant pressure to the specific heat at constant volume, and p is the ideal gas pressure. For soundwaves in air the expression for adiabatic processes has to be used since for any frequency above 20 Hz thermal equilibrium by heat exchange can never be reached. Thus the sound velocity in liquids and gases is:

$$v_{\text{sound}}^{\text{liquid}} = \sqrt{\frac{B}{\rho}}, \quad v_{\text{sound}}^{\text{gas}} = \sqrt{\frac{\gamma p}{\rho}}.$$

In air at normal conditions of pressure and humidity the *speed of sound* is about 330 m/s, in water it is 1500 m/s. Note that the velocities in air and water are different by a factor of 5, which leads to an acoustic mismatch as we will see later. The construction of the ear has to deal with this difference!

Next we determine the velocity of particles in the soundwave from the first derivative of the displacement wave (not to be confused with phase velocity or group velocity):

$$u_x = \frac{\partial \xi(x,t)}{\partial t} = -\omega \xi_0 \cos(kx - \omega t).$$

u_x is also sometimes called amplitude velocity. The propagating pressure wave follows from Hooke's law via:

$$p_x = -B\frac{\partial \xi(x,t)}{\partial x} = -B\xi_0 k \cos(kx - \omega t).$$

The respective amplitudes are therefore:

$$u_0 = -\omega \xi_0; \quad p_0 = -B\xi_0 k.$$

The SI units are: $[u]$ = m/s and $[p]$ = Pa. To derive the power P of the wave and its intensity we need to consider the time average of the product of force F_x and particle velocity u_x:

$$\langle P \rangle = \langle F_x u_x \rangle.$$

When dividing the power by the area A through which the soundwave propagates we obtain the time averaged intensity:

$$\begin{aligned}\langle I \rangle &= \left\langle \frac{P}{A} \right\rangle = \left\langle \frac{F_x}{A} u_x \right\rangle = \langle p_x u_x \rangle \\ &= B\omega k \xi_0^2 \langle \cos^2(kx - \omega t) \rangle.\end{aligned}$$

The last term integrated over one period yields a factor $\frac{1}{2}$, such that the time average sound intensity is:

$$\langle I \rangle = \frac{1}{2} B\omega k \xi_0^2.$$

The unit of intensity is $[I] = \text{Watts/m}^2$. The time average intensity can be rephrased by various expressions using the previously derived definitions:

$$\langle I \rangle = \frac{1}{2} Z u_0^2 = \frac{1}{2} \frac{p_0^2}{Z},$$

where

$$Z = \rho v$$

is the *acoustic impedance* of the soundwave. The unit is: $[Z] = \text{kg m}^{-2}\,\text{s}^{-1}$.

The impedance is essential for evaluating transmission and reflection coefficients at interfaces between different media. The acoustic impedance should not be confused with the impedance of electric circuitries. The latter is connected with dissipative processes; acoustic impedance is a property that characterizes the sound intensity in different media. It is equivalent to the refractive index in optics. Typical values for Z of some relevant materials in the present context are listed in Tab. 12.1.

Tab. 12.1: Sound velocity, density, and acoustic impedance of different materials in the body.

	Sound velocity v [m/s]	**Density [kg/m^3]**	**Acoustic impedance Z [kg m^{-2} s^{-1}]**
Air	330	1.3	430
Water	1500	998	1.5×10^6
Fatty tissue	1470	970	1.38×10^6
Muscle	1570	1040	1.7×10^6
Bone	3600	1700	6×10^6

12.3 Crossing borders

Similar to optics, when soundwaves cross boundaries between media characterized by different impedances, then part of the intensity is transmitted and part of it is reflected as sketched in Fig. 12.3.

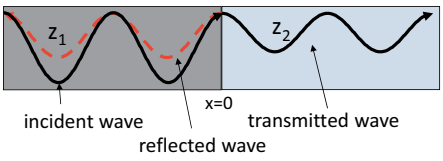


Fig. 12.3: Transmission and reflection of a soundwave at an interface between two materials characterized by different impedances Z_1 and Z_2 with $Z_1 > Z_2$.

For simplicity we consider normal incidence of soundwaves towards a boundary. Energy conservation requires that the intensities of incident (i), transmitted (t), and reflected (r) waves are related as:

$$I_\mathrm{i} = I_\mathrm{t} + I_\mathrm{r}$$
$$Z_1 u_\mathrm{i}^2 = Z_2 u_\mathrm{t}^2 + Z_1 u_\mathrm{r}^2.$$

Thus the waves in medium 1 and 2 can be expressed as:

$$\xi_1(x,t) = \xi_\mathrm{i} \sin(k_1 x - \omega t) + \xi_\mathrm{r} \sin(k_1 x + \omega t)$$
$$\xi_2(x,t) = \xi_\mathrm{t} \sin(k_2 x - \omega t).$$

At the boundary $x = 0$, continuity of the slope of the waves is required, which results in an additional condition for the particle velocities:

$$u_\mathrm{i} + u_\mathrm{r} = u_\mathrm{t}.$$

Both equations can be combined to yield for the particle velocities:

$$u_\mathrm{t} = u_\mathrm{i} \frac{2Z_1}{Z_1 + Z_2}; \quad u_\mathrm{r} = u_\mathrm{i} \frac{Z_1 - Z_2}{Z_1 + Z_2}.$$

Inserting these expressions into the equations for the intensities, we obtain:

$$I_\mathrm{t} = \frac{1}{2} Z_2 u_\mathrm{t}^2 = 4 I_\mathrm{i} \frac{Z_1 Z_2}{(Z_1 + Z_2)^2},$$
$$I_\mathrm{r} = \frac{1}{2} Z_1 u_\mathrm{r}^2 = I_\mathrm{i} \left(\frac{Z_1 - Z_2}{Z_1 + Z_2} \right)^2.$$

Now we are ready to rephrase the reflected and transmitted intensities normalized by the incident intensity in terms of acoustic impedances for the case of normal incidence:

$$R = \frac{I_\mathrm{r}}{I_\mathrm{i}} = \left(\frac{Z_1 - Z_2}{Z_1 + Z_2} \right)^2,$$
$$T = \frac{I_\mathrm{t}}{I_\mathrm{i}} = 4 \frac{Z_1 Z_2}{(Z_1 + Z_2)^2}.$$

The intensity R is known as *reflectivity* and T is called *transmissivity*. R and T fulfill the important relation confirming energy conservation:

$$R + T = 1.$$

In optics these equations are known as Fresnel equations for normal incidence if the impedance Z is replaced by the refractive index n.

Three limiting cases need discussion:

(a) For $Z_1 = Z_2$, $R = 0$ and $T = 1$, i.e., if the acoustic impedances are identical, although the materials may feature different densities and sound velocities, there will be no reflection.

(b) For $Z_2 \to 0$ or $Z_2/Z_1 \to 0$ we find $R = 1$ und $T = 0$, i.e., the wave is completely reflected.
(c) The same holds true if $Z_1 \to 0$ and $Z_1/Z_2 \to 0$. Also here we find $R = 1$ und $T = 0$, i.e., the wave is completely reflected.

In the case that reflection occurs at an interface to an acoustically less dense media characterized by $Z_1 > Z_2$, the amplitudes of the incident and reflected waves are in phase. Vice versa, for $Z_1 < Z_2$ the amplitudes of incident and reflected waves are out-of-phase by π or 180°.

As an example we consider the boundary between air ($Z_{\text{air}} = 430\,\text{kg}\,\text{m}^{-2}\,\text{s}^{-1}$) and water ($Z_{\text{water}} = 1.5 \times 10^6\,\text{kg}\,\text{m}^{-2}\,\text{s}{-}1$). With these impedances the reflectivity at the interface is $R = 0.9989$ or 99.89 %. Therefore the transmitted intensity is $T = 0.001$ or 0.1 %. Thus between air and water and vice versa there is a severe *impedance mismatch*. Sound can propagate easily in air or in water. But it cannot cross borders. Between air and water there is an acoustic wall.

Sound recognition requires that acoustic waves reach into the body from air. Assuming that the sound receptor is surrounded with a tissue with impedance similar to water, the transmitted wave has an intensity of 0.1 % of the incident intensity. With such a low transmittance we would be deaf. All three parts of the ear are constructed such that the transmittance is increased by a factor of roughly 625 so that the transmittance actually becomes 62 % instead of 0.1 %. The mechanism that provides this amplification is discussed in the next section. Sonography presented in Chapter 13 faces the same problem of coupling soundwaves into the body for imaging.

12.4 Sound intensity

The physical definition of *sound intensity*, also known as *acoustic intensity*, can be expressed in several equivalent forms:

$$I_{\text{sound}} = \frac{1}{2}\rho v_{\text{Phase}} u_0^2 = \frac{1}{2} Z u_0^2 = \frac{1}{2}\frac{p_0^2}{Z}.$$

For simplicity we have omitted here the bracket for the time average. The symbols are explained in Section 12.2. The usual range of auditory sound is:

$$10^{-10}\,\text{Watt}/\text{m}^2 < I_{\text{sound}} < 10^{-2}\,\text{Watt}/\text{m}^2.$$

The audible threshold is even lower at $10^{-12}\,\text{Watt}/\text{m}^2$ and the pain threshold is about $1\,\text{Watt}/\text{m}^2$. There are 12 orders of magnitude in between!

Using sound intensity and acoustic impedance we can calculate the pressure amplitude p_0 and the particle displacement ξ_0 for different intensities:

$$p_0 = \sqrt{2 Z_{\text{air}} \cdot I_{\text{sound}}},$$
$$\xi_0 = \frac{p_0}{Z_{\text{air}} \cdot \omega}.$$

The pressure amplitude only depends on the intensity and the impedance, whereas the displacement, in addition, is inversely proportional to the sound frequency. Typical values are given in Tab. 12.2. According to these values the pressure for the audible threshold in the frequency range of highest sensitivity from 1 kHz to 3 kHz is 2×10^{-7} Pa = 20 µPa = 2×10^{-9} mbar. This low pressure amplitude is comparable to pressures experienced in an ultrahigh vacuum environment. The corresponding particle displacement of only 0.01 nm corresponds to a diameter less than the size of an atom. The actual displacement amplitude at the basilar membrane (see Section 12.5) is about 1 nm at the threshold pressure level. This is an incredible sensitivity that nature has achieved. In fact the ear is the most sensitive receptor in the body.

Tab. 12.2: Typical values of pressure amplitudes and particle displacements from pain threshold to auditory threshold. Values are calculated for pressures in air.

I [Watt/m^2]	P_0 [hPa]	ξ_0
1	2×10^{-1}	16 µm
10^{-2}	2×10^{-2}	1.6 µm
10^{-10}	2×10^{-6}	0.1 nm
10^{-12}	2×10^{-7}	0.01 nm

Sound intensity is a linear function of the squared pressure amplitude. However, sound reception is not a linear function of sound intensity. Instead, auditory perception depends logarithmically on the intensity, which is known as the *Weber–Fechner law*. Because of the logarithmic perception it is advantageous to define a relative logarithmic scale for auditory intensity called *auditory level*. The auditory level is the logarithm to the base 10 of sound intensity I normalized by the threshold intensity I_0:

$$\log \frac{I}{10^{-12}}.$$

The result is multiplied by 10 and called *intensity level B*:

$$B = 10 \log \frac{I}{10^{-12}}.$$

Intensity levels B are expressed in Bel, named after Alexander Graham Bell. Although B is a pure number without unit, Bel is treated as a unit. One decibel, abbreviated dB, is a tenth of a Bel. The threshold intensity has 0 dB, the pain threshold corresponds to an intensity level of 120 dB. Sound intensity and sound level are compared in Fig. 12.4.

The sensitivity of the ear depends on the frequency. For the same intensity but different frequencies, the *loudness* of a tone judged by a person varies. The peak sensitivity is between 1000 to 3000 Hz and drops off for lower or higher frequencies. Young adults have an audible frequency range from 20 to 20 000 Hz. With age the audible frequency range at the upper end drops to about 8000–10 000 Hz.

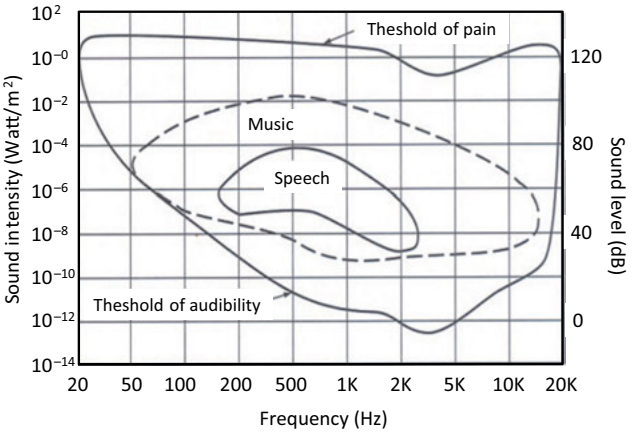


Fig. 12.4: Sound intensity–frequency map. The contours outline the threshold audibility and the threshold of pain. Typical intensities and frequencies for music and speech perception are encircled.

For quantifying the personal perception of sound intensity, a scale for loudness level is conveniently introduced by the definition:

$$L = 10 \log \frac{I}{10^{-12}}.$$

The unit of the loudness level is a phone. At 1000 Hz the scale of the sound intensity level (dB scale) and the scale of the perceived loudness level (phone scale) agree. At lower and higher frequencies these two scales deviate. Figure 12.5 is an isophone chart or equal loudness chart. It displays the sound pressure level in dB which is required for higher and lower frequencies to be perceived as the same loudness at the frequency of 1000 Hz. Since the sensitivity of the ear drops at lower and higher frequencies this is compensated for by a higher pressure level. Only between 2 kHz and 5 kHz in the region of highest sensitivity of the ear can the pressure level be reduced for the same loudness perception.

12.5 Outer and middle ear

In spite of the huge acoustic impedance mismatch between air and the body we are nevertheless capable of sensing soundwaves. How is this possible? More precisely: how can a soundwave with large displacement amplitude but small pressure amplitude in air be converted into a soundwave with small displacement amplitude but high pressure amplitude in fluids? To get an answer, we have to again look carefully at the construction of the ear.

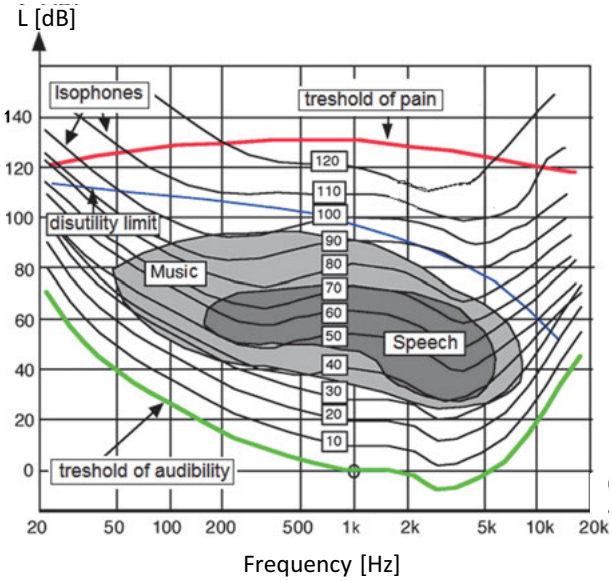


Fig. 12.5: Isophone chart comparing loudness with sound intensity. At 1 kHz both scales coincide.

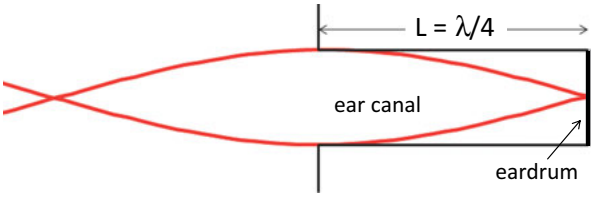


Fig. 12.6: Model of the outer ear canal in terms of an acoustic pipe with one end open. The displacement wave $\xi(x)$ is indicated, which is 90° out-of-phase with the pressure wave.

Outer ear

The outer ear has a length of about 25–30 mm and can be modeled as an *acoustic pipe* with one end closed by the eardrum and the other end open, Fig. 12.6. Resonance condition is found for multiples of $\lambda/4$ waves with frequencies:

$$f_n = (2n+1)\frac{v_{\text{sound}}}{4L}; \quad n = 1, 2, \ldots .$$

With $v_{\text{sound}} = 330\,\text{m/s}$ we find a first resonance at 3300 Hz and the next at 9900 Hz, etc. Indeed the peak sensitivity of the ear is at about 3000 Hz. What happens at other frequencies? The outer ear is not a perfect resonator but one with a poor quality factor that lets all other frequencies pass as well. But the sensitivity of the ear is highest between about 1000 Hz and 8000 Hz. The sensitivity is also determined by other parts of the ear that will be discussed later.

Middle ear

The middle ear is separated from the external ear by the eardrum. On both sides of the eardrum there is the same ambient pressure, since the middle ear is connected via the Eustachian tube with the upper part of the throat to the environment. The middle ear contains the ear bones (auditory ossicles) hammer, anvil, and stirrup, which together have the task of transferring sound pressure from the eardrum to the oval window of the cochlea (Fig. 12.7). Impedance mismatch between external ear and inner ear is superseded by the principles of hydraulics and levers.

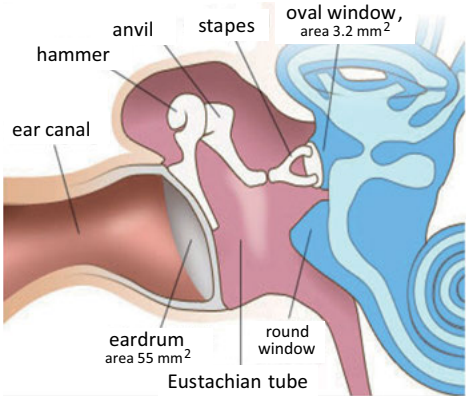


Fig. 12.7: Middle ear connecting the eardrum of the external ear with the oval window of the inner ear via the auditory ossicles hammer, anvil, and stirrup.

We first consider the hydraulic action. The *hydraulic principle* implies that two cylinders that are connected by an incompressible fluid in equilibrium experience the same pressure:

$$p_1 = \frac{F_1}{A_1} = p_2 = \frac{F_2}{A_2}.$$

Here $F_{1,2}$ are the respective forces and $A_{1,2}$ are the respective cylindrical areas. Using the hydraulic principle one obtains an amplification of forces when pushing with a small force on a small area, thereby gaining a large force on a larger area. The hydraulic principle implies that the force per area remains constant if the area is increased. If the incompressible fluid is replaced by a rigid rod of constant diameter, the force stays constant but the pressure can be amplified. To visualize this we assume that hammer, anvil, and stirrup are combined to a single bone that connects the eardrum with the oval window, as sketched in Fig. 12.8 (a). The eardrum transmits the sound pressure to the connecting bone by a force F_1, which subsequently acts on the oval window. For constant force we have: $F_1 = p_{\text{drum}} A_{\text{drum}} = p_{\text{oval}} A_{\text{oval}}$. From this we conclude that with constant force the pressure is increased if A_{oval} is smaller than A_{drum}. The pressure amplification is proportional to the ratio of the areas:

$$p_{\text{oval}} = \frac{A_{\text{drum}}}{A_{\text{oval}}} p_{\text{drum}}.$$

Using typical numbers for the areas of eardrum (55 mm^2) and oval window (3 mm^2) we gain a pressure enhancement by a factor of 18.

Next we consider the lever in the middle ear formed by the combined action of all three ossicles. The lever is sketched in Fig. 12.8 (b). The force F_1 acts from the eardrum via the hammer to the anvil. The anvil acts as a second class lever (see Section 2.4). This means that the load arm and the lift arm are on the same side of the pivot point; and the lift arm (l_F) is longer than the load arm (l_L). Second class levers have a mechanical advantage according to the ratio l_F/l_L. Again using typical values, the ratio is about 1.4. Thus the force F_2 on the oval window is greater than F_1 by a factor of 1.4 or roughly 40 %.

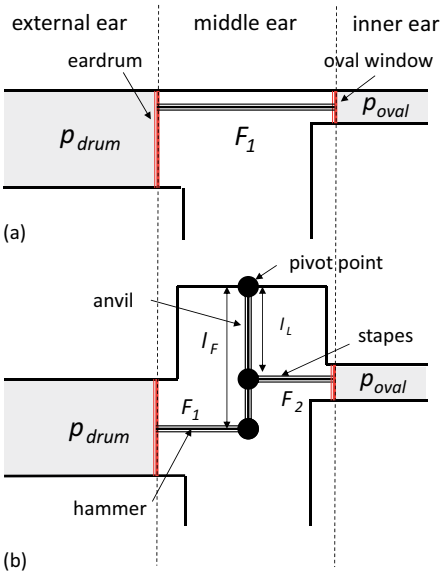


Fig. 12.8: (a) Mechanical analog of the hydraulic action of the middle ear; (b) schematic of the lever action, combining hammer, anvil, and stirrup.

Combining the hydraulic principle and lever action we obtain a pressure amplification of $1.4 \times 18 = 25$. Since the transmitted sound intensity is proportional to the square of the pressure amplitude: $I_t \approx p_0^2$, the actual amplification is 625, corresponding to an intensity increase by 28 dB. Therefore with the action of the middle ear, 62.5 % of sound intensity is transmitted instead of 0.1 % without impedance matching. This is an amazing amplification factor that our middle ear is capable of achieving solely by mechanical principles.

The middle ear fulfills two more tasks apart from sound amplification. The transmitted frequency band in the audible region is very broad, stretching from 20 Hz to about 20 000 Hz. The resonance frequency of the middle ear including eardrum and oval window is at about 1400 Hz. In order to transmit a broad frequency band, the resonance curve should be broad and flat, which is indeed the case, similar to the

low quality factor of the external ear. The other task is protection against too high intensities. This is achieved by an acoustic reflex triggering two sets of muscles, one tightening the eardrum, another pulling the stirrup away from the oval window with a latency of 10 ms to reduce the pressure transmittance to the fluid in the scala vestibuli.

Summarizing, the middle ear fulfills three main tasks: amplification of pressure amplitude, transmission of a large frequency band, and protection against destructive intensity levels.

12.6 Inner ear

Having solved the problem of impedance mismatch, sound detection in the inner ear must be responsive to at least two properties of the soundwave: intensity (loudness) and frequency (tone). In fact, it is also sensitive to sound location, as we will see later. Intensity is proportional to the squared pressure amplitude and some kind of pressure sensitive receptor could be easily imagined. However, discriminating sound frequencies may be regarded as the ability to recognize the "colors of sound". Visual perception has three separate receptors responsive to the RGB colors discussed in Chapter 11. How does the inner ear encode sound with 20 000 different frequencies? Furthermore, how can fast acoustic signals be recorded by relatively slow nerve conduction? In the following, we try to answer these questions.

12.6.1 Structure of the cochlea

The answer to the questions just posed lies in the complex structure of the inner ear (*cochlea*) displayed in Fig. 12.9. The cochlea is a 35 mm long conically shaped tube curled up in a snail-like fashion. When untwined we recognize three parallel tubes: two outer tubes (*scala vestibuli* and *scala tympani*) containing a sodium (Na^+) enriched liquid, and an inner cochlear duct (*scala media*) separating the outer tubes by the *basilar membrane* and *Reissner's membrane*. The inner tube is filled with a liquid rich in potassium ions (K^+). A cross section of the cochlea is shown in Fig. 12.12. The vestibular tube is closed at the base by the *oval window* and the tympanic tube is closed by the round window. The two outer fluid tubes are in contact at the apex of the cochlea at a point called the *helicotrema*. We have already seen that the pressure of the soundwave is transmitted to the oval window. The oval window acts as a drumhead setting up a traveling wave in the scala vestibuli moving back and forth in the rhythm of the pounding stirrup. It is important to realize that the traveling wave involves macroscopic fluid flow between the oval window and the round window, whereas soundwaves require only local particle movement about an equilibrium position. The round window at the base of the cochlea serves as a pressure release for the traveling and incompressible fluid wave. While the oval window moves inwards,

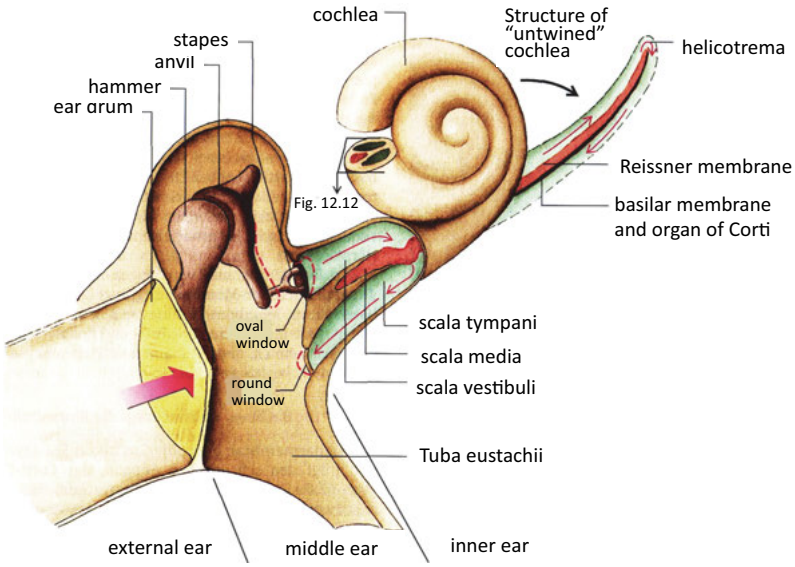


Fig. 12.9: Different parts of the ear in a cut-away view (adapted from Pape et al., Physiologie, with permission of Thieme Verlag).

the round window moves outwards, and vice versa, as outlined by the red dashed lines in Fig. 12.9. The fluid wave does not propagate all the way through the cochlea up to the apex. Instead it stops at locations along the basilar membrane specific to the frequency of the wave (Fig. 12.10). At the stopping point, the basilar membrane becomes bent like the buckling up of a rug when a runner suddenly stops. The bending of the membrane is essential for detection of soundwaves, as we will recognize further below.

In Fig. 12.10 the cochlea is schematically shown as a thin tube with a center line, representing the basilar membrane. Sound arrives at the oval window via the vibration of the stirrup. Low frequency traveling waves propagate towards the apex and become strongly absorbed close to the apex by the elastic properties of the membrane. High frequency waves travel in the liquid only a short distance and become absorbed close to the base of the cochlea. Intermediate frequencies are detected somewhere in the middle between base and apex. Complex soundwaves composed of different frequencies are absorbed at several positions along the basilar membrane.

The cochlea is a position-sensitive frequency detector similar to a harp when it is not being used as a musical instrument but as a resonator for different sound frequencies. High frequencies resonate with short strings at the knee and low frequencies resonate with long strings at the shoulder. The frequency dispersion along the basilar membrane is referred to as *tonotopy*. The tonotopic map of the basilar membrane from high to low frequencies is a fundamental feature of auditory coding and is preserved all the way up to the auditory cortex in the brain.

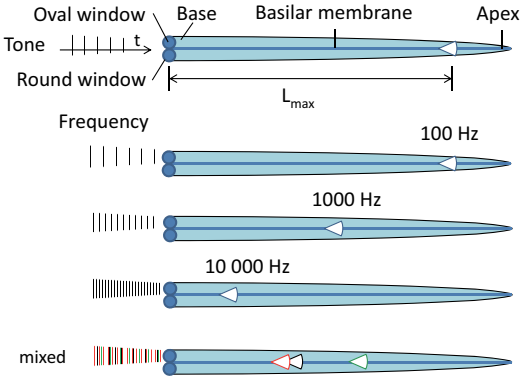


Fig. 12.10: Tonotopic absorption of soundwaves in the basilar membrane. The top panel schematically shows different parts of the cochlea, including base, apex, oval window, and round window. L_{max} is the distance from the base where the maximum absorption of the soundwave occurs. The frequency of different tones is indicated schematically by vertical bars on a time scale (not to scale). Sound consists of a mixture of tones, which are absorbed at different positions along the basilar membrane.

The tonotopic mapping of soundwaves is only possible if there is a gradient of the elastic properties along the basilar membrane. Indeed, the elastic modulus of the basilar membrane continuously decreases from base (stiff) to apex (soft). At the same time, the membrane becomes wider over the same distance from ≈ 0.12 mm to ≈ 0.5 mm (Fig. 12.11), in close analogy to the already quoted harp. The combined result of stiffness gradient and change of width yields a factor of about 100 in variation of elastic modulus from base to apex.

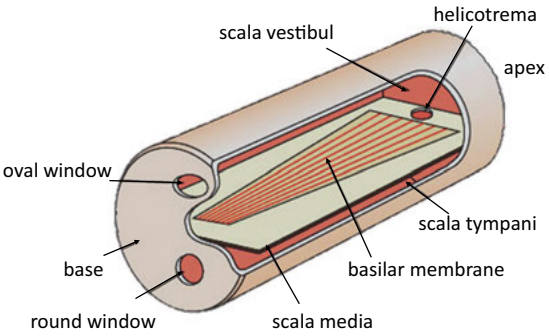


Fig. 12.11: Mechanical model of the basilar membrane featuring a gradient of width and stiffness from the base to the apex.

12.6.2 Organ of Corti

Now we concentrate on the fine structure of the inner duct, a cross-sectional view of the cochlea is shown in Fig. 12.12. Here we notice the basilar membrane together with a tectorial membrane enveloping fine hairs (*stereocilia*) connected to hair cells. Hair cells and tectorial membrane together are called the *organ of Corti* (in recognition of the Italian anatomist Alfonso Giacomo Gaspare Corti (1822–1876) who described it for the first time). Wherever the traveling wave is absorbed along the cochlea, the basilar membrane is bent, which inflects hair cells by moving against the tectorial membrane.

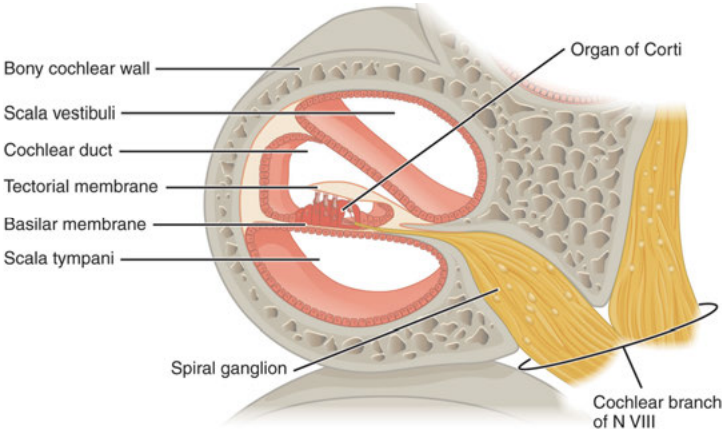


Fig. 12.12: Cross section through the cochlea showing the three main tubes: scala vestibuli, scala tympani, and scala media. Inner and outer hair cells are arranged between the basilar membrane and the tectorial membrane (reproduced from OpenStax Anatomy and Physiology, © Creative Commons).

Figure 12.13 provides an enlarged cross-sectional cut through the cochlea, showing in more detail how the bending of the basilar membrane and the inflection of the hair cells takes place [1]. In this cross-sectional view we recognize three outer hair cells (OHC) and one inner hair cell (IHC). OHC and IHD form rows along the cochlea, shown in Fig. 12.14. A bundle of tiny hairs (*stereocilia*) are attached to all hair cells. The outer hair cells together with their stereocilia touch the *tectorial membrane*, while the ones of the inner hair cell sit in a pocket without touching the membrane. The IOC are connected to afferent neurons going directly to the brain, whereas the OHC are connected to efferent neurons that provide sensory sensitivity and motor capabilities.

As the traveling wave is absorbed in an area specific to its frequency, the basilar membrane becomes bent at the Hinch point. The bending causes an inflection of the outer hair bundles against the tectorial membrane that depolarizes the outer hair cells. The depolarization activates motor proteins which contract the hair cells, similar

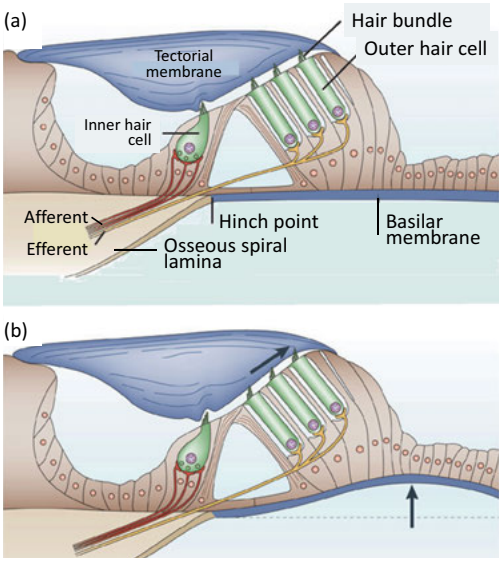


Fig. 12.13: (a) Cross section through the organ of Corti with basilar membrane and tectorial membrane eclipsing inner and outer sets of hair cells for sensing soundwaves. (b) The arrow indicates bending of the basilar membrane leading to transverse motion of stereociliary bundles against the tectorial membrane (reproduced from [1] with permission of Nature Publishing Group).

to muscle contraction. This contraction amplifies the basilar membrane motion, acting as a positive feedback on the original stimulus. By doing so, it allows fluid flow in the scala media passing by the inner hair cell bundle, bending them back and forth with the frequency of the soundwave. The positive feedback of the outer hair cells opposes the viscous damping of the stereociliar vibration by the cochlear fluid motion. The total amplification factor gained by this mechanism has been estimated to exceed 4000 [2]! The ciliary vibration stimulates a receptor potential which – in turn – is perceived within the auditory cortex as sound of a particular frequency, see Section 12.6.4.

12.6.3 Inner and outer hair cells

Now we go back and take a more global view on the cochlea. The frequency dependent sensitivity of the basilar membrane is provided by some ≈ 16 000 hair cells lined up in four rows along the basilar membrane: one inner row of the IOC and three outer rows of the OHC. The inner ≈ 4000 hair cells sense intensity and frequency *tonotopically*, i.e., each hair cell is responsible for the detection of one narrow frequency range, such that frequencies from 16 Hz to 20 000 Hz are mapped onto these ≈ 4000 inner hair cells. The outer ≈ 3 × 4000 hair cells serve as amplifiers of basilar membrane motion during resonance absorption. The highly ordered organization of both types of hair cells is shown impressively in the scanning electron micrograph of Fig. 12.14. Each V-shaped bundle of about 100 stereocilia is attached to one outer hair cell, and also each straight bundle of stereocilia is attached to one of the inner hair cells.

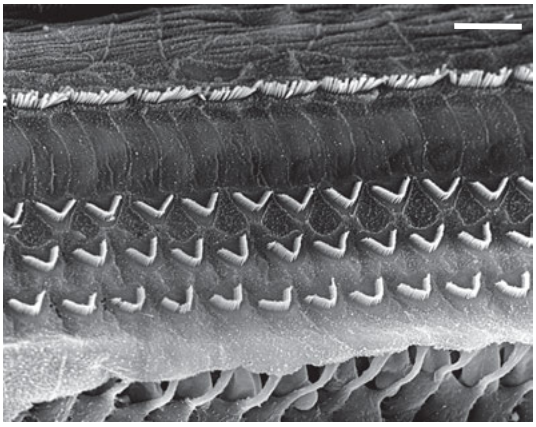


Fig. 12.14: Scanning electron microscopy image showing the highly ordered organization of hair cells in the organ of Corti. The single row of inner hair cells is seen at the top and three rows of outer hair cells are seen at the bottom. The scale bar corresponds to 15 µm (SEM image by Marc Lenoir, from the EDU website 'Journey into the World of Hearing', www.cochlea.eu, with permission of R. Pujol et al., NeurOreille, Montpellier).

The upright position of the stereocilia and their arrangement in bundles is very delicate. Too high a sound intensity can break them irreversibly and scramble up the proper alignment leading to severe hearing loss [3].

Summarizing we note that the perception of sound in the inner ear proceeds via three steps: first, the stirrup pounding on the oval window generates a mechanical wave; second, in regions where the traveling wave reaches maximum amplitude, the outer hair cells become activated and their active vibration amplifies the wave amplitude; third, the inner hair cells when stimulated respond by releasing a receptor potential, activating action potentials in the spiral ganglion cells on the way to the cochlear nucleus.

The analysis of the physical and physiological aspects of sound perception goes back to Hermann von Helmholtz in the nineteenth century, and in 1961 Georg von Békésy was awarded the Nobel Prize in physiology and medicine 'for his discoveries of the physical mechanism of stimulation within the cochlea'. Although much has been achieved in the understanding of sound detection, this field still yields new discoveries. In a recent study the role of the cochlea's curvature was investigated [4], showing that the increasing curvature redistributes wave energy density towards the cochlea's outer wall, affecting the shape of waves propagating on the membrane, particularly in the region where low frequency soundwaves are processed. Thus the shape also contributes to the amplification of sound, according to the 'whispering gallery' phenomenon observed in St. Paul's Cathedral in London or by holding an empty spiral-shaped shell to the ear.

12.6.4 From mechanical stimulus to receptor potential

Typical *inner* and *outer hair cells* are outlined schematically in Fig. 12.15. The cell body is submerged in the scala tympani, while the "hairs" punch through the membrane lamina reticularis into the scala media. The scala media is filled with endolymph fluid, the scala tympani contains perilymph fluid. The cation concentrations in both fluids are opposite. The endolymph fluid has a unique cation composition, which is not found anywhere else in the body. Usually the extracellular space has a low K^+ concentration and the cytoplasm is K^+-rich. Here the endolymph fluid is K^+-rich, and the perilymph fluid is Na^+-rich.

The average length of the stereocilia increases from base (high frequency) to apex (low frequency) to support the tonotopic frequency selectivity of the basilar membrane. Each hair cell has a characteristic frequency which correlates with its location on the basilar membrane. Within an individual bundle the stereocilia are slightly graded in length and inclined. The graded stereocilia are "mechanically" interlinked. If they lean towards the longer neighbors during basilar membrane motion, the receptor potential becomes depolarized; if they lean towards the opposite site, the receptor potential is hyperpolarized. Depolarization/hyperpolarization is gradual and proportional to the bending amplitude of the basilar membrane. Depolarization occurs due to K^+ channels at the tip of the cilia opening because of mechanical strain on connecting molecular chains. The potential change triggers the opening of Ca^{2+} channels further

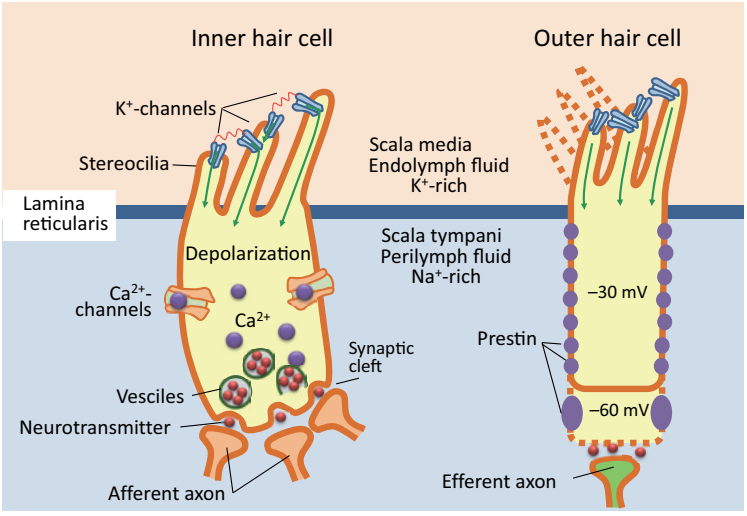


Fig. 12.15: Inner hair cell with stereocilia contain K^+ channels that open on bending towards the longer stereocilia, causing a depolarization of the hair cell. Simultaneously, Ca^{2+} channels open, stimulating the discharge of neurotransmitters. The outer hair cell acts similarly. In addition, depolarization causes a contraction of the longitudinal cell body, whereas hyperpolarization causes stretching, indicated by dashed lines and elongated prestin proteins.

down the cell body. The Ca^{2+} ions stimulate vesicles containing the neurotransmitter glutamate to discharge. The released neurotransmitter diffuses across the cell membrane into the synaptic gap. In response the axon conducts the receptor potential to the spiral ganglion, which fires an action potential traveling up to the auditory cortex. The Ca^{2+}-stimulated opening of vesicles and release of the neurotransmitter glutamate is the same procedure we have already seen at neuromuscular junctions (see Fig. 6.14) and in the retina between receptors and bipolar cells (Section 11.3.4).

The receptors of IHC and OHC act in a similar fashion. However, opening K^+ channels in OHC not only causes depolarization but also a longitudinal contraction of the cell body. This process, known as *mechanoelectric transduction* (MET), is accomplished by the protein *prestin*, which is integrated into the cell membrane and responds sensitively to the inner cell potential. In the depolarized state, the prestin molecules in the cell membrane are contracted; in the hyperpolarized state they are elongated. In resonance at a particular frequency the entire cell body will longitudinally oscillate. The oscillation modulates the liquid flow in the media scala passing by the tectorial membrane, which enhances or inhibits the bending of the stereocilia attached to the IHC. The MET resembles piezoelectric transducers used for sound generation discussed in Chapter 13. To associate IHC and OHC with "hair" is a bit of an understatement. These hairs in the organ of Corti are exceedingly functionalized and specialized *mechanoelectric* receptors.

There are about 30 000 afferent nerve fibers; the majority connect to the IHCs, 5 to 10 fibers to each inner hair cell, shown schematically in Fig. 12.16. There are only a few efferent nerve fibers arriving from the superior olivary complex that connect one by one to the OHCs. Their function is not well understood at present. However, it is known that the OHC modulate the sensitivity of the IHC mostly in an inhibiting fashion. In fact, we not only perceive sound from outside, we can also produce sound with our vocal cords. Therefore it is obvious and necessary that there must be a feedback system between sound generation and sound perception, which most likely has its origin in the efferent connection of the OHC.

Each auditory nerve fiber responds over a certain range of frequencies and sound pressures, but has a characteristic frequency f_c at which it has maximum sensitivity. By way of example we consider only one particular hair cell detecting one particular frequency. The depolarization of the hair cell is proportional to the deflection of the stereocilia in the direction towards the longer neighbor, whereas in the opposite direction the cell becomes hyperpolarized, as indicated in Fig. 12.17. The change in the receptor potential is extremely sensitive to the deflection direction and its amplitude. There is no threshold potential for complete depolarization. Instead, depolarization is gradual and proportional to the deflection, which again is proportional to the wave amplitude or loudness. At this level any mechanical motion of the basilar membrane is translated into a continuously varying and *graded receptor potential*. Thus we can conclude that the organ of Corti is a *mechanoreceptor* or more precisely a *mechanoelectric transducer* that transforms a mechanical stimulus into receptor potentials.

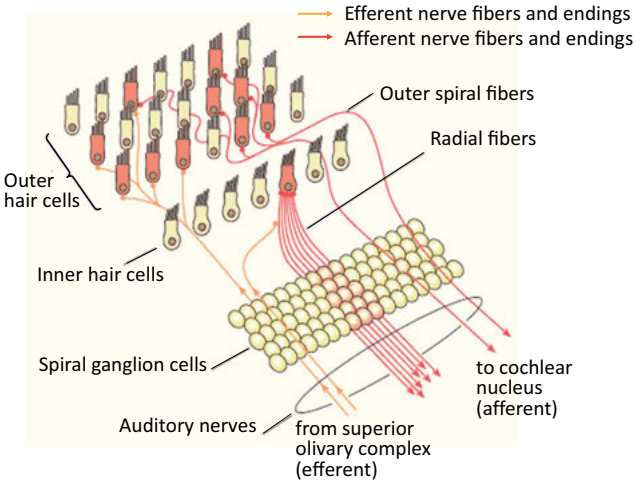


Fig. 12.16: Connection of afferent and efferent nerve fibers to inner and outer hair cells, respectively. Note that each inner hair cell is connected to many afferent nerve fibers (about 20) required for decoding frequency and intensity (adapted from www.open.edu/openlearn).

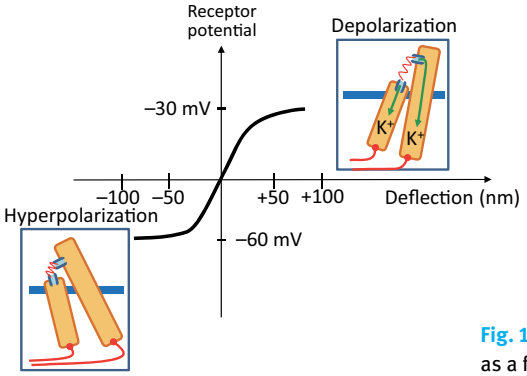


Fig. 12.17: Change of the receptor potential as a function of stereocilia deflection.

12.6.5 Frequency coding

On the way up from the receptor potential through the auditory nerve to the auditory cortex in the brain, sound frequency is coded by two different methods: tonotopic localization and phase locking.

Tonotopic localization of sound frequency is achieved by connecting some 30 000 nerve fibers to 4000 inner hair cells, as already described. The one-by-one correspondence between frequency and location is preserved and mapped onto the auditory cortex (Fig. 12.18). The brain analyzes action potentials arriving in the auditory cortex versus localization and derives a characteristic frequency. One may consider this process as a Fourier analysis of space.

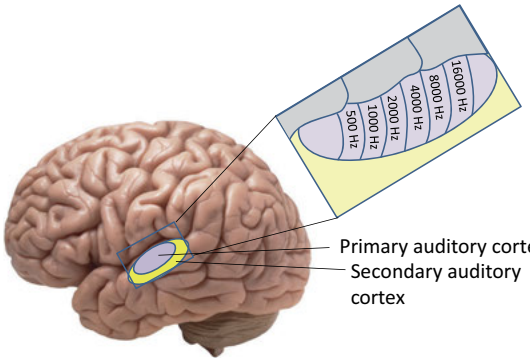


Fig. 12.18: Tonotopic mapping of sound frequencies in the cortex (adapted from www.open.edu/openlearn, © Creative Commons).

For frequencies up to about 5 kHz, in addition to tonotopic mapping the sound frequency is also encoded by phase locking, which can be considered a Fourier analysis of frequencies. The receptor potential has the same frequency as the soundwave simply by deflection of the stereocilia, Fig. 12.13. The postsynaptic action potential is phase locked to the same soundwave: upon each deflection in the rhythm of the sound an action potential is fired. However, intensity and frequency are not only encoded by action potentials of one fiber but by several afferent nerve fibers attached to the same hair cell as indicated in Figs. 12.16 and 12.19. While the first nerve fiber fires action potentials on any cycle in phase with the sound frequency, the second fiber fires – also phase locked – but only on any fourth cycle, the third on any eighth cycle, etc. With increasing amplitude the second fiber reacts more often, eventually further adjacent fibers follow. The frequency is encoded by the number of action potentials per second

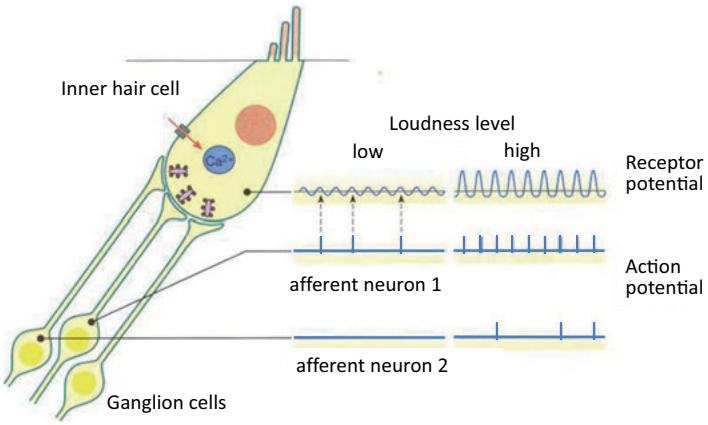


Fig. 12.19: Receptor potential and action potential of inner hair cells. The action potential is phase locked to the frequency of the receptor potential. With increasing loudness more fibers carry action potentials, all phase locked to the frequency of the receptor potential (adapted from Speckmann, *Physiologie*, 6th edition, 2013, with permission of Elsevier GmbH, Urban & Fischer, München).

carried by the first fiber and the intensity is encoded in the number of fibers that are stimulated in addition to the first one. From receptor potential to action potential we again observe the typical analog-digital conversion, which is at work for all senses in the body.

Phase locking works up to a frequency of about 5 kHz, i.e., 0.2 ms between action potentials. Considering that action potentials require a refractory period of about 1 ms, phase locking should stop beyond 1 kHz. However, for phase locking it is not necessary that an action potential is fired in any period. As long as the period is in phase with the vibrational frequency, the brain can encode a sequence of action potentials belonging to that same frequency.

12.6.6 Pathway to the auditory cortex

Finally, we want to briefly outline the pathway of the action potential to the auditory cortex in the brain (Fig. 12.20). All 30 000 nerve fibers that emerge from the cochlea carrying information about sound frequency and intensity and all nerve fibers that originate from the vestibular system transmitting information about balance are bundled up in the sensory VIII cranial nerve as the main channel to the brain. The first station on the path to the brain is the cochlear nucleus where neuronal processing of the digitized information from the inner ear occurs and crosslinking between fibers from the left and right cochlea takes place. Within the superior olive lie the *lateral superior olive* (LSO) and the *medial superior olive* (MSO). LSO is important for detecting interaural sound level differences while neurons in the MSO process microsecond time differences between left and right ear, the major cue for localizing low-frequency sounds (see next section). The inferior colliculus (IC) is a relay station in the ascending part of the auditory system, and most likely acts to integrate information specific for sound source localization from the superior olive before sending it to the medial geniculate body, which is part of the thalamic relay system. From there the nerve fibers go further up to the auditory cortex for final processing.

12.6.7 Sound localization

Localization of sound is part of survival strategy and therefore essential for all vertebrates. Pairs of ear can localize sound whether it comes from the front or the back, from the side or from sources above or below the azimuthal plane of the ear. Ears use different techniques to make these distinctions. For some techniques monaural hearing is sufficient, other techniques require a binaural capability. For frequencies f between 400 Hz to about 1400 Hz, the source of sound is primarily localized by the *interaural arrival time difference* (ITD) Δt of the pressure amplitude or equivalently by the phase difference $\Delta\phi = f\,\Delta t$ associated with the path difference $\Delta s = d \sin\alpha$, where d is the interaural distance (Fig. 12.21).

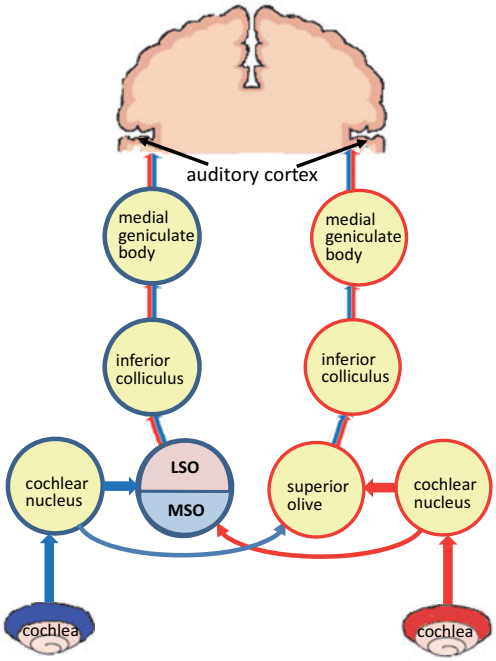


Fig. 12.20: Schematic outline of the auditory pathway from the left and right cochlea to the auditory cortex in the brain. Arrows indicate connections and crosslinking of the auditory nerve on the way to the brain. The superior olive on both sides is subdivided into a lateral superior olive (LSO) and a medial superior olive (MSO) (adapted from www.open.edu/openlearn, © Creative Commons).

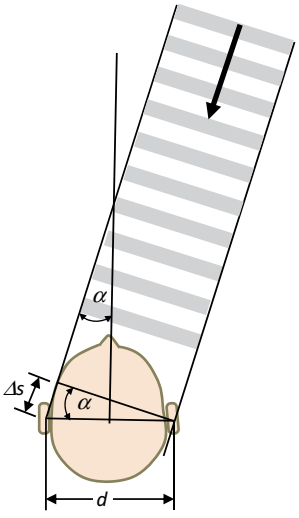


Fig. 12.21: In the far field, acoustic ranging is realized by interaural arrival time of soundwaves.

As an example we assume a sound source at an angle of $\alpha = 30°$, with a frequency of 1000 Hz traveling in air with sound velocity of 330 m/s. For the interaural separation d we take a typical value of 0.20 m, and we assume that the sound source is at a far distance compared to d, justifying the use of plane waves. This yields an arrival time difference of $\Delta t = 0.3$ ms. The maximum time delay of 0.6 ms is for sound arriving from the side at right angles to the ear. With decreasing angle the discrimination of location becomes increasingly challenging requiring higher time resolution. This problem is solved by a unique *coincidence detector* located in the *medial superior olive* (MSO, Fig. 12.20) that compares the arrival time of action potentials from the left and right ears. This unique principle is known as *interaural time difference* (ITD) *detection* and is shown schematically in Fig. 12.22. Neurons arriving from the left and right sides fan out in the MSO into a horizontal array of joined neurons forming a delay line. Only when excitatory inputs from both sides reach one of the joined neurons simultaneously will the sum potential stimulate an action potential. All other neurons remain quite. From the location of the firing neuron the brain can compute the delay time that yields the horizontal orientation of the sound. Only five joined neurons are shown in the schematic of Fig. 12.22. In reality the coincidence detector in the MSO contains thousands of them. The time resolution of this coincidence detector is of the order of 10 µs, corresponding to an angular accuracy of about 2°.

The ears are the only sensor in the body that measures time differences. This works well in the frequency range from 400 to 1400 Hz. Why ITD for humans fails at higher frequencies is not entirely clear, since it was observed that other mammals have a higher ITD frequency range [5]. Alternatively, at higher frequencies the location of sound sources is achieved by intensity level differences (ILD) rather than by ITD. At

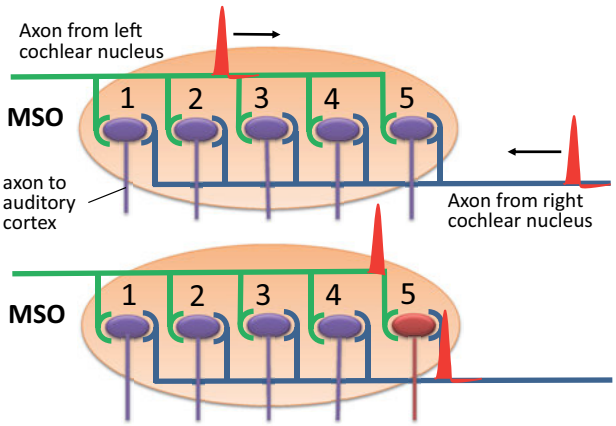


Fig. 12.22: Schematic of the coincidence detection of sound arriving at different times from the left and right cochlea for localization of the sound source in the azimuthal plane of the ears. An action potential is only fired in one of the joined neurons labeled 1 to 5 if the action potentials arriving from either side overlap, as indicated in the bottom panel at location 5.

high frequencies one ear is exposed to the soundwave intensity while the opposite ear is in the sound shadow receiving a lower intensity. ILD is processed in the lateral superior olive (LSO, Fig. 12.20) by integrating ipsilateral excitatory inputs and contralateral inhibitory inputs [6].

At low frequencies, in contrast, where the wavelength of sound exceeds the size of the head, diffraction effects distort the wavefront, making it more and more difficult to localize sound sources. On the other hand, distinction between front and back and up or down is realized by the asymmetry of the pinna (see Fig. 12.1). Sound becomes reflected from the rims that interfere with waves penetrating the outer ear canal without reflection. Interference and phase shift provide information that again allows acoustic ranging of the sound source.

Summarizing, at low frequencies below 400 Hz diffraction effects increasingly lead to confusion of sound localization. At intermediate frequencies between 400 Hz and 1400 Hz, sound localization in the horizontal plane is achieved by ITD in the MSO using contralateral excitatory inputs from both ears. At high frequencies, ILD is operational in the LSO, receiving ipsilateral excitatory inputs and contralateral inhibitory inputs.

ITD detection was proposed by Jeffress in a seminal publication from 1948 [7] and since then this intriguing mechanism has been reproduced in all major physiology textbooks. However, recently it has been debated whether coincidence detection is the sole explanation for ITD. Inhibitory input and interaction among synaptic activities play a similar role for ITD as for ILD [6]. Coincidence detection, presented above as an instantaneous process is oversimplified. Many more interactions causing internal delays appear to be operational [8].

The ears are indeed a unique and highly sensitive sensor for the detection of a wide range of sound frequencies, loudness levels, and spatial location of sound sources. In contrast to other sensors of the body, auditory perception follows different strategies to extend the range of sensitivity: phase locking for low frequencies and tonotopic mapping for higher frequencies; coincidence detection of sound localization at low frequencies and interaural level difference detection at high frequencies. Visual perception, in contrast, has a rather limited range of frequency sensitivity and no backdoor trick to extend it up to the infrared or ultraviolet regime.

12.7 Tone, sound, and noise

In general we distinguish between *tone*, *sound*, and *noise* (Fig. 12.23). *Tone* is a soundwave with a fixed frequency. Note that the frequency is responsible for the tone, not the wavelength, which changes from one medium to the other. The frequency of soundwaves is also sometimes called 'pitch'. For instance, the internationally accepted standard pitch for music instruments is 440 Hz. The ear has a very high frequency (*pitch*) resolution: it is in the order of 0.2 % which means that one can distinguish between

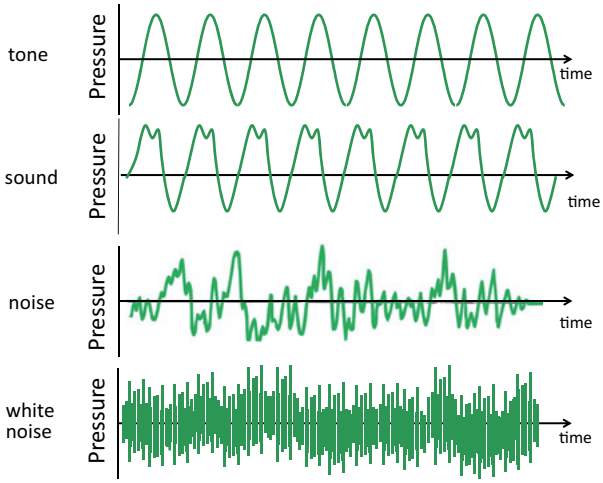


Fig. 12.23: Characteristic wave forms for a single frequency sinewave tone, for the sound of a musical instrument, for some noise, and for a white noise source.

a tone of 1000 Hz and one of 1002 Hz. Even though the auditory nerves of the left and right ears are connected, they keep a certain level of independence. If they are exposed to slightly different frequencies, they will recognize a beating effect in spite of the lack of any interference effect. Only with increasing frequency difference will both ears recognize two distinct tones.

By the term *'sound'* we generally understand a superposition of tones that are perceived as harmonic or consonant, like the sound of a violin. Usually these are overtones to a fundamental tone. In a Fourier analysis of a particular sound we will find only a few frequencies, which are in a rational relationship to the fundamental frequency and characteristic for the instrument generating the sound.

In contrast, *noise* is an inharmonic mixture of tones that is perceived as dissonant and often disturbing. Finally, we hear *'white noise'* from many different kinds of sound sources like traffic, machines, etc. The term 'white noise' is reminiscent of *'white light'* sources. In analogy, white noise contains the entire frequency spectrum with no correlation and phase relationship among the different frequencies.

12.8 Hearing aids

Loss of hearing can be tested by taking audiograms. The test measures the audible threshold as a function of frequency. If hearing ability is reduced by, for instance, 15 dB, then the auditory sensitivity is the same as if the sound level were increased by $\Delta L = 10 \times \log 10^{1.5}$ in comparison to normal sensitivity. With age, a decline in sound sensitivity is normal. A typical audiogram is shown in Fig. 12.24 for different frequencies. At high frequencies the loss is more severe than at low frequencies. Hearing loss with age is actually gender specific and more acute for men than for women.

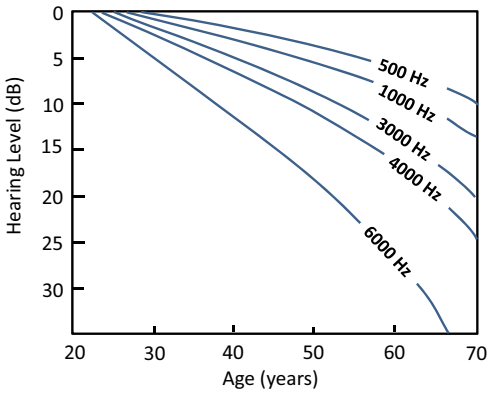


Fig. 12.24: Loss of hearing sensitivity with aging. Sensitivity loss is more severe at higher frequencies.

Loss of hearing ability is characterized by the following levels:
- 0 dB–20 dB: normal hearing;
- 20 dB–40 dB: mild hearing loss;
- 40 dB–55 dB: moderate hearing loss;
- 55 dB–70 dB: moderately severe hearing loss;
- 70 dB–90 dB: severe hearing loss;
- > 90 dB: profound hearing loss to deafness.

Most important for communication are sound frequencies in the range of 100–6000 Hz. If hearing loss is severe, a hearing aid may be able to compensate for the loss. But first the location of the loss should be identified. There are a number of potential locations for dysfunctionalities causing hearing loss:
1. occlusion in outer ear canal;
2. eardrum stiffening (tympanic membrane);
3. middle ear (otosclerosis);
4. cochlea, damage of hair cells;
5. degeneration of nerve cells in the auditory cortex.

The cause of hearing loss cannot be identified precisely with an audiogram, but likely causes can be isolated (Fig. 12.25). In the case of a conductive hearing loss of the middle ear the hearing level drops homogenously for all frequencies. In the case of damage to hair cells due to loud noise, usually a certain frequency band is affected while the other frequency regions behave normally. However, damage to hair cells may have additional adverse effects that reduce the ability to discriminate between sounds. This is often noticed as a reduced comprehension of speech, and simply amplifying the sound level as most hearing aids do, is often insufficient to improve speech perception.

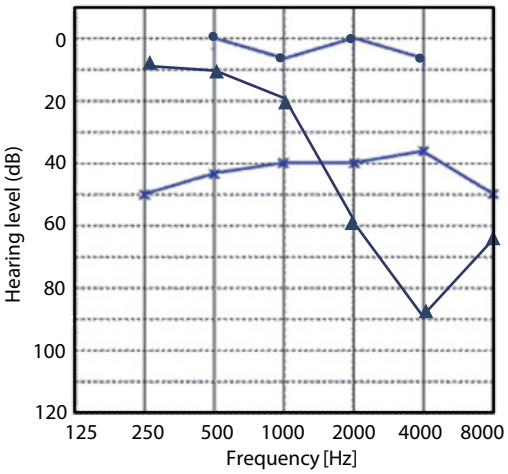


Fig. 12.25: Audiograms for different hearing losses. • normal hearing; × conductive hearing loss; ▲ damage in cochlea.

Presently there are three types of hearing aids available:

1. in-ear canal hearing aid
2. middle ear implants
3. cochlea implants

In all three cases we have to distinguish between an outer part and an inner part of the hearing aid. The outer part contains the sound receiver, sound processor and batteries. It can be hid behind the pinna, attached to eye glasses, anchored on the bone, or can be made small enough to fit into the ear canal.

The inner part of hearing aids is distinctively different depending on application. *In-ear canal hearing aids* contain a loudspeaker that amplifies the sound and transmits it to the tympanic membrane (Fig. 12.26). This is a rather simple device and the one most frequently applied. But improper fitting, occlusion, and compression effects often cause discomfort. Feedback between receiver and loudspeaker is another problem if fitting and adjustments are not proper. Modern electronics allows adapting the amplification factor to the individual audiogram.

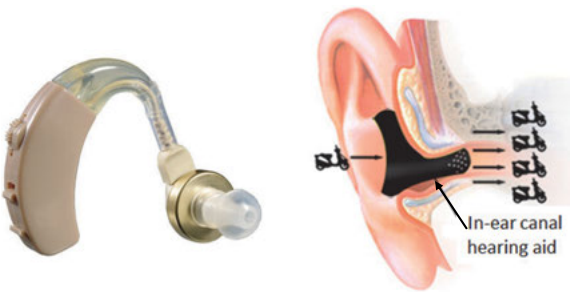


Fig. 12.26: In-ear canal hearing aid. External sound is amplified by the hearing aid and the drum head is exposed to an increased pressure amplitude.

Middle ear implants replace damaged ossicles, restoring transmission between the tympanum and the stapes footplate. Various mechanical implants are offered that take over the hydraulic action of the middle ear. They can be divided into ossicular replacement prostheses, which replace only the incus and the malleus, and those which replace all three ossicles including the stapes. With the advent of three-dimensional printers one may envision the fabrication of replacement prostheses that replicate more closely the shape and elastic properties of the original ossicular bones and their functionality. All this is still in developmental stage but may soon be on the market. Middle ear implants are further discussed in Section 15.5.1/Vol. 2.

Cochlea implants feed the output of the sound receiver into a wire implanted into the organ of Corti to stimulate the auditory nerve cells. Since these implants reach only the high frequency end of the cochlea close to the base and not the lower frequencies at the apex, and because of shortening problems of the wire immersed in the highly ionic and conducting fluid in the scala tympani, the stimulation of frequency-dependent receptor potentials is rather limited. Cochlear implants are still under development. Presently they are only recommended in cases of severe deafness due to damage to the sensory hairs in the organ of Corti. Electronic implants that directly connect to the auditory nerve may be more promising.

12.9 The making of sound

Speech is a sequence of sounds that carry meaning. Sound reception is developed early in newborns, while understanding the meaning takes more time. This example shows that there is an intimate intervention between the auditory cortex and other parts of the cortex. Hearing and speaking form one unity and they are the most important tools for communication among individuals. Without hearing there is no speaking. Deaf people without special training are mute as well, they are deaf-mute.

The human voice is generated during expiration. The vocal cords (glottis) in the larynx can close the tube between the oral cavity and the trachea in order to increase the air pressure in the chest during talking (Fig. 12.27). If the pressure is high enough, the glottis will open again.

The vocal cords approach each other due to muscle action in the larynx. The expiration air is then forced to flow through the narrow gap left by the glottis. According to the law of Bernoulli

$$p + \frac{1}{2}\rho v^2 = \text{const.},$$

air flow decreases the pressure and the vocal cords close completely. At a threshold expiration pressure of 400 to −500 Pa the glottis open again. The intermittent air flow generates the sound of a voice. The frequency depends on the length and the tension of the vocal cords.

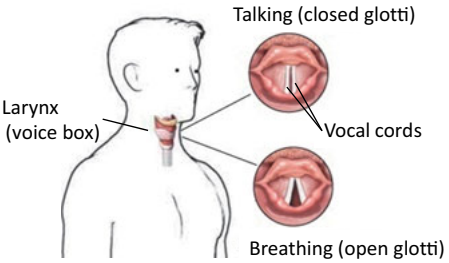


Fig. 12.27: Glottis is closed during talking but open during normal breathing.

Sound is further formed in the hollow space of the oral cavity acting as a resonance body of various shapes for different sounds (Fig. 12.28). Vowels (a, e, i, o, u) are formed in the oral space solely by changing the shape and volume of the cavity. Consonants are often fricative sounds that require the assistance of the tongue and lips to increase the friction; explosive sounds like 'p, t, k, q' etc. are formed by build-up and sudden release of pressure in the cavity; and hissing sounds like 'c, s, z' are formed by closing the teeth.

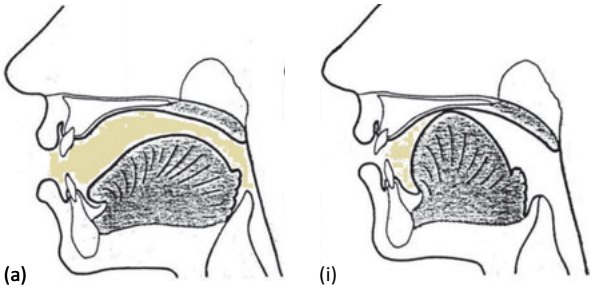


Fig. 12.28: Formation of sounds in the oral cavity with the help of the tongue.

12.10 Summary

1. Sound is a traveling longitudinal compression wave.
2. Sound intensity is characterized by pressure amplitude and acoustic impedance.
3. Acoustic impedance is determined by density of the media and sound velocity. It is the refractive index for sound propagation.
4. Sound velocity depends on the density of the medium and its compressibility.
5. At boundaries between two media characterized by different impedances, soundwaves are partially reflected and partially transmitted.
6. At boundaries to media with very low or very high impedance, soundwaves become completely reflected. Such boundaries have an impedance mismatch.
7. The boundary between air and water is acoustically mismatched.
8. The task of the ear is to funnel sound into the body by compensating the impedance mismatch and to convert sound into action potentials.
9. The ear consists of three parts: outer ear, middle ear, and inner ear.

10. The middle ear is a mechanical pressure amplifier. It serves the purpose of overcoming the impedance mismatch between air and watery solution.
11. The middle ear fulfills three main tasks: amplification of pressure amplitude, transmission of a large frequency band, and protection against destructive intensity levels.
12. The cochlea of the inner ear acts as a mechanoelectric transducer.
13. Hair cells in the basilar membrane of the cochlea together with the tectorial membrane form the *organ of Corti*.
14. Hair cells along the basilar membrane provide a tonotopic mapping of sound frequencies.
15. Inner hair cells are connected to afferent neurons. Their receptor potential initiates action potentials traveling to the brain.
16. The action potential of the inner hair cells is tonotopically frequency mapped and phase locked to the frequency of the receptor potential.
17. Outer hair cells are connected to efferent neurons. Their receptor potential causes longitudinal length changes of the cell, known as mechanoelectrical transduction.
18. The outer hair cells modulate the sound reception of inner hair cells.
19. Sound localization is achieved by coincidence detection in the medial superior olive that compares the arrival time of action potentials from the left and right ears.
20. The sensitivity of the ear is extremely high and extends over 12 decades of sound intensity.
21. Hearing loss is mainly due to loss of flexibility of the middle ear bones, in severe cases due to irreversible breakage of the sensory hairs in the organ of Corti.
22. Hearing aids can be placed in the external ear canal, or may involve a replacement of the middle ear bones, or may require a cochlea implant.
23. The human voice is produced during expiration by an intermittent opening of the vocal cords.
24. Sounds are formed in the hollow space of the oral cavity acting as a resonance body.

References

[1] Fettiplace R, Hackney CM. The sensory and motor roles of auditory hair cells. Nature Reviews, Neuroscience. 2006; 7: 19–29.
[2] Reichenbach T, Hudspeth AJ. Dual contribution to amplification in the mammalian inner ear. Phys Rev Lett. 2010; 105: 118102.
[3] Fitzakerley J. University of Minnesota Medical School Duluth. www.d.umn.edu/~jfitzake/Lectures/DMED/InnerEar/IEPathology/Fig.s/NormalHC.jpg
[4] Manoussaki D, Dimitriadis EK, Chadwick RS. Cochlea's graded curvature effect on low frequency waves. Phys Rev Lett. 2006; 96: 088701.
[5] Brughera A, Dunai L, Hartmann WM. Human interaural time difference thresholds for sine tones: The high-frequency limit. J Acoust Soc Am. 2013; 133: 2839.
[6] Grothe B, Pecka M, McAlpine D. Mechanisms of sound localization in mammals. Physiol Rev. 2010; 90: 983–1012.
[7] Jeffress L. A place theory of sound localization. J Comp Physiol Psychol. 1948; 41: 35–39.
[8] Franken TP, Roberts MT, Wei L, Golding NL, Joris PX. In vivo coincidence detection in mammalian sound localization generates phase delays. Nature Neuroscience. 2015; 18: 444–452.

Further reading

Tipler PA, Mosca G. Physics for scientists and engineers. Vol. 1: Mechanics, oscillations and waves and thermodynamics. London: W. H. Freeman; 2003.
Mullin WJ, George WJ, Mestre JP, Velleman SL. Fundamentals of sound with applications to speech and hearing. Boston: Allyn and Bacon; 2003.
Kandel ER, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth AJ. Principles of neural science. 5th edition. McGraw Hill; 2013.
Bear MF, Connors BW, Paradiso MA. Neuroscience: Exploring the brain. 5th edition. Wolter Kluwer; 2015.
Speckmann EJ, Hescheler J, Köhling R. Physiologie. 6th edition. Elsevier, Urban and Fischer; 2013.
Pape H-C, Kurtz A, Silbernagel S, editors. Physiologie. 7th edition. Stuttgart, New York: Thieme Verlag; 2014.

Useful website

www.open.edu/openlearn/science-maths-technology/science/biology/hearing/content-section-0

Part B: **Imaging modalities without ionizing radiation**

13 Sonography

13.1 Introduction

In medicine, sonography is a technique that uses ultrasound (US) for imaging internal organs. Ultrasound is also used in numerous nonmedical applications such as submarine navigation, seafloor mapping, food control, security screening, surface cleaning, nondestructive material testing, and bonding of different materials.

Medical sonography is an imaging modality that takes static images of organs and tissues, dynamic images of heart and lung movement, and kinetic images of blood flow. A well-known and common example of sonography is the imaging of fetuses as part of prenatal checkups.

US imaging is much easier to handle by a physician than any radiation-based method. It can be applied locally at the bedside or at the site of an accident and it does not require special safety procedures for the patient or for the examining staff. Direct communication with the patient is possible during examination, which is an important advantage compared to other imaging modalities such as MRI, CT, or PET. On the downside, conventional US images have lower resolution and require considerable experience to properly interpret them for useful diagnostics.

Figure 13.1 is a flow chart of the essential parts needed for medical sonography. First an ultrasound pulse generator is required and a coupling medium to transmit pulses into the body. The same transducer is also used as receiver of echo signals. After signal processing the result is displayed on a screen: x- and y-axes for the spatial coordinates and a gray scale for depth information. A typical frozen image of the heart

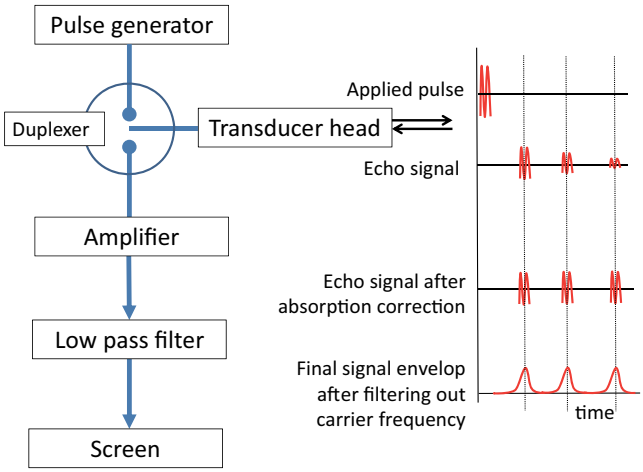


Fig. 13.1: Flow chart for application and signal processing of medical sonography.

DOI 10.1515/9783110372830-015

is shown in Fig. 13.2, but movement of the heart and the blood flow in the ventricles can be imaged as well in real time. Further medical applications of US imaging for diagnostics are in the following areas:

- cardiovascular system
- abdominal organs
- urology/prostate
- obstetrics/gynecology
- ophthalmology
- mammography

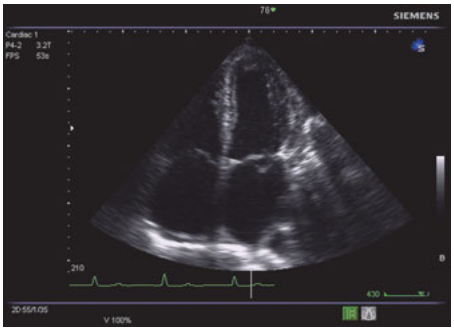


Fig. 13.2: Typical US image (B-mode) of the heart chambers (reproduced from https://openi.nlm.nih.gov/).

In medical sonography different imaging methods are distinguished:

- *A-mode* is a single line scan through the body, recording the amplitudes of returning echoes from interfaces between tissues having different impedances as a function of time.
- *B-mode* scans provide two-dimensional images, representing changes in acoustic impedance of the tissue within one section.
- *C-mode* scans are two-dimensional images formed by taking a sequence of slices in B-mode, repeated in a direction normal to B-mode images at a constant depth.
- *M-mode* or motion mode allows imaging of moving organs by A- or B-scans but with higher pulse repeat frequencies, adequate for recording videos.
- *Doppler mode* makes use of the frequency shift by moving reflectors, allowing visualization of blood flow.
- *Pulse inversion mode* uses two successive pulses with opposite sign whose difference displays body parts with nonlinear compressibility.

In the following we discuss some physical aspects required for sonographic imaging, starting with basic imaging conditions.

13.2 Basic physical conditions for ultrasound imaging

Sonography works with soundwaves. The basic properties and terms: pressure amplitude, sound velocity, particle velocity, acoustic impedance, sound intensity, reflection and transmission at interfaces, are presented in Sections 12.2–12.4. The frequencies used for sonography are far beyond audible sound, i.e., more than 20 kHz. In fact most sonographic systems use frequencies in the range from 2 to 20 MHz. Although these are much higher frequencies than discussed in Chapter 12, the physics of soundwaves remains the same.

We continue considering only longitudinal waves; transverse waves, although they exist in soft tissues and in bones, are not useful for imaging because of their high attenuation. Typical physical parameters for US imaging of body parts are as follows:

- *Sound velocity* (phase velocity) in soft matter is similar to water and is about $v \approx 1540$ m/s or 1.54 mm/µs. Throughout this text we use this average velocity for various estimates. More precise tissue specific values are listed in Tab. 13.1.
- *Echo signal* arrives after the time lapse $\Delta t = 2L/v_{\text{sound}}$, where L is the depth of a reflecting interface. For 1 µs ≡ 0.75 mm depth.
- *Wavelengths*: $\lambda = v/f = 1500$ µm (1 MHz)–150 µm (10 MHz).
- *Wave amplitude* or displacement amplitude $\xi_0 = u_0/\omega = p_0/(Z\omega)$. For $f = 3$ MHz and $p_0 = 10^5$ Pa the wave amplitude is $\xi_0 = 3$ nm.
- *Particle velocity* amplitude $u_0 = p_0/Z < 0.06$ m/s.
- *Pressure amplitude* $p_0 < 1$ MPa.

The (negative) peak pressure should not exceed 1 MPa = 10 bar to avoid any explosion of internal cavities. At a pressure of 1 MPa the particle velocity is accordingly 0.6 m/s and the intensity of the soundwave becomes:

$$\begin{aligned}\langle I\rangle &= 1/2 p_0 u_0\\ &= 1/2 \cdot 1\,\text{MPa} \times 6 \cdot 10^{-1}\,\text{m/s} = 3 \cdot 10^5\,\text{Pa}\cdot\text{m/s}\\ &= 3 \cdot 10^5\,\text{W/m}^2 \quad \text{or} \quad I = 30\,\text{W/cm}^2.\end{aligned}$$

From the quoted parameters we find intensities that are by far too high. The maximum power administered during sonographic imaging should not exceed 100 mW/cm^2. For estimating this limit, the following expression is used, known as the *mechanical index* (MI):

$$\text{MI} = \frac{p_0^-/1\,\text{MPa}}{\sqrt{f/1\,\text{MHz}}}.$$

Here p_0^- is the negative peak pressure and f is the frequency of soundwaves. The mechanical index (MI) is an attempt to estimate biological effects of ultrasound and to avoid formation and disruption of cavities (cavitation). The MI is found on most ultrasound display screens, along with other parameters. It is a dimensionless number and should be limited to values below 1.9 for safe application of US. The MI suggests

that higher pressures can be compensated for by increasing frequencies. However, one should be skeptical about such conclusions. For instance, a pressure amplitude of 1 MPa and a frequency of 10 MHz yields an MI of 0.3. However, for this 'safe' value the sound intensity is, as already estimated, 30 W/cm^2, by far more than what is considered safe. Therefore, one should be cautious when applying high intensities. Keeping the pressure level below 0.1 MPa is a safer bet.

13.3 Sound propagation and attenuation

In Section 12.2 we considered soundwaves at interfaces with perpendicular incidence. This is justified as the wavefront in the outer ear canal is indeed perpendicular to the tympanic membrane. However, sound propagating in the body usually hits an interface at an angle. Furthermore, interfaces of organs and bones are generally rough on the scale of US soundwave lengths, which are typically below 1 mm. Then at rough interfaces reflection, transmission, and scattering occurs, as indicated in Fig. 13.3. For sonographic imaging only the backreflected and/or backscattered intensity (echo) is of interest and detected by the transducer head now acting as receiver. The echo signal is attenuated by various intensity losses: scattering at rough interfaces in off-specular directions, transmission and absorption. Only a small fraction of intensity is reflected back and reaches the detector.

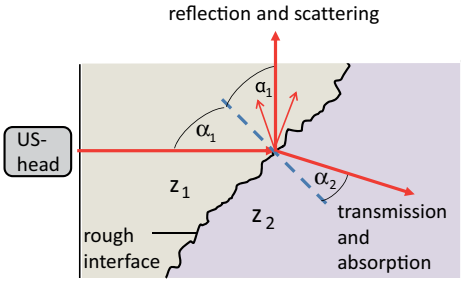


Fig. 13.3: Reflection, transmission and scattering at a rough interface between two materials characterized by different acoustic impedances z_1 and z_2.

Attenuation of the pressure amplitude has two main contributions: absorption and scattering. The time dependence of a propagating pressure wave with exponentially damped amplitude can be written:

$$p(x, t) = p_0 \exp(-\beta x) \cos(kx - \omega t).$$

Here $\beta = \beta_{\text{abs}} + \beta_{\text{scatt}}$ is the damping parameter with both contributions and unit $[\beta] = \text{m}^{-1}$; $k = 2\pi/\lambda$ is the wavenumber of the soundwave, unit $[k] = \text{m}^{-1}$; all other symbols have their usual meaning. Attenuation due to absorption is material specific and can be estimated. Damping due to scattering is interface specific and more difficult to estimate, although in most cases dominating.

Attenuation via absorption has two contributions: viscous damping and thermal conduction: $\beta_{\text{abs}} = \beta_{\text{vis}} + \beta_{\text{therm}}$. The *viscous damping* parameter can be expressed as:

$$\beta_{\text{vis}} = \frac{2}{3\rho v^3}\eta_{\text{s}}\omega^3.$$

Here ρ is the density and η_{s} the viscosity of the media in which the soundwave propagates. Obviously this damping becomes a severe factor with increasing frequency. Damping due to *thermal conduction* to neighboring tissues or organs is given by:

$$\beta_{\text{therm}} = \frac{1}{2\rho v^3}\frac{(\gamma-1)\kappa}{C_V}\omega^2.$$

In this equation C_{V} is the specific heat at constant volume, κ is the heat conduction coefficient, and γ is the adiabatic coefficient defined by the ratio C_p/C_V.

A simple rule of thumb for damping of soundwaves in biological tissue is as follows: 1 dB of damping occurs after traveling a distance z of 1 cm at a frequency f of 1 MHz, or:

$$\frac{\beta}{\text{dB}} = \frac{z}{\text{cm}}\frac{f}{\text{MHz}}.$$

Since β expressed in dB does not have a unit, z and f need to be divided by their units in order to get a simple number. The ratio β/zf is roughly 1 for most organs and the brain, but it is about 2 for muscles and 0.5 for fatty tissue.

Among all the other factors already considered, the backreflected intensity depends also on the contrast at the interface expressed in terms of the acoustic impedance:

$$R = \frac{I_{\text{r}}}{I_{\text{i}}} = \left(\frac{Z_1\cos\alpha_2 - Z_2\cos\alpha_1}{Z_1\cos\alpha_2 + Z_2\cos\alpha_1}\right)^2.$$

For backreflection, α_1 and α_2 should be within the opening angle of the transducer to be detected. In Tab. 13.1 some impedance values are listed and Fig. 13.4 schematically illustrates different idealized cases for reflection and scattering at interfaces. Case (a) of a smooth interface perpendicular to the incoming soundwave has the highest backreflected intensity. The same interface inclined by some angle as in case (b) will cause reflection and refraction, but the reflected beam may miss the detector. Rough interfaces as in (c) cause diffuse scattering, which drastically diminishes the intensity in the backward direction. Curved surfaces (d) widen the backscattered beam and only part of the intensity reaches the detector.

Apart from lung and bones, the impedance values vary only little for different organs. Two examples illustrate the challenges of US imaging. Let's assume that we want to image the heart in the surrounding body tissue. For the myocardium of the heart we adopt the impedance of muscles, and for its surroundings we take the impedance value of water. Then at normal incidence the backreflected intensity is 0.4 % of the incident intensity. This estimate neglects all other effects, such as scattering from rough interfaces and absorption. In fact, an echo signal of about 0.1 % is more realistic. The

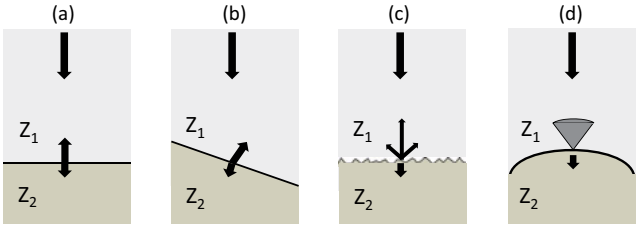


Fig. 13.4: Four different idealized cases for interfaces between tissues with different acoustic impedance.

Tab. 13.1: Sound velocities, densities and impedance values for some characteristic materials of relevance for ultrasound imaging.

	Sound velocity v [m/s]	Density ρ [kg/m^3]	Acoustic impedance Z [10^6 N s/m^3]
Air	330	1.3	0.000430
Water	1500	998	1.5
Blood	1530	1000	1.62
Fatty tissue	1470	970	1.38
Muscle	1570	1040	1.7
Bone	3600	1700	3–7
Lung	650–1160		0.3–0.4
PZT	4000	7500	30

second example concerns the echo signal from bones. The backreflected intensity is about 30 % of the incident intensity at the front side, and another 30 % of the transmitted intensity on the back side, which is about 21 % if scattering and absorption effects are neglected. For detecting an echo signal from tissues behind bones, not only the original intensity is reduced by more than 50 %, but the echo signal is again strongly backreflected from the bones. From this second example we can conclude that any organs behind bones are essentially invisible to US. Bones form sound walls and sound shadows; US imaging can only be applied to soft tissues, organs, and muscles that are in front of bones. The same considerations apply to air- or gas-filled organs such as the lung. At the air/tissue interface total reflection occurs. Therefore the lung cannot be penetrated by US and tissue behind the lungs remains invisible.

13.4 Ultrasound transducer

13.4.1 Piezoelectric effect

For generation and detection of US waves a piezoelectric head is used, also called a transducer. In general, transducers are devices that convert one form of energy into another. In the present case, US transducers convert electrical energy into vibrational energy of a crystal lattice using the piezoelectric effect. Piezoelectric crystals are ionic and insulating materials with a high electrical polarizability that is strongly coupled to the crystal lattice. Mechanical compressive or tensile strain generates electrical potentials, and vice versa electric potentials are converted into strain. In either case via strain or electrical potential, a polarization of electric dipole moments in the crystal is induced. The direct piezoelectrical effect relates the polarization P of these electric dipole moments to the stress σ applied, as illustrated in Fig. 13.5:

$$P = d\sigma,$$

where d is the piezoelectric constant with the unit $[d] = \mathrm{m/V}$. The converse piezoelectric effect relates the length change Δz of the crystal to the applied voltage change ΔV:

$$\Delta z = d\,\Delta V.$$

Note that the expansion of the crystal does *not* depend on its size but only on the voltage applied. To keep the discussion simple, the tensor property of the piezoelectric coefficient has been neglected. A more complete discussion can be found in [1]. The magnitude of the piezoelectric coefficient d is in the order of 500 pC/N or 0.5 nm/V.

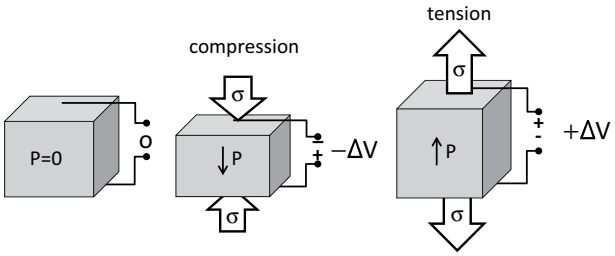


Fig. 13.5: Working principle of the direct piezoelectric effect. Application of a compressive or tensile stress induces an electric polarization and a voltage change. Conversely, application of a voltage changes the length of the piezoelectric crystal.

This transduction can be utilized in a static or dynamic fashion up to high frequencies. Piezoelectric crystals have numerous applications in physics and technology as actuators, sensors, and for nanopositioning. For instance, the sensing head of scanning tunneling microscopes and of atomic force microscopes is driven by piezoelectric crystals.

What are piezoelectric materials made of? Piezoelectric crystals are insulating ceramic materials that lack inversion symmetry. The best known example is the piezoelectric compound $PbTiO_3$ (short notation PT) or the Zr doped version $Pb(Zr_{1-x}Ti_x)O_3$ (short notation PZT). The movement of Zr/Ti^{4+} ions in and out of the oxygen plane (see Fig. 13.6) by application of an electrical field or by stress induces an electrical dipole moment. Therefore these piezoelectric materials can either be used as actuators by applying a voltage or as sensors of pressure or stress by measuring voltage changes.

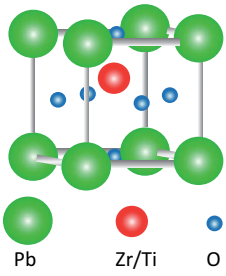


Fig. 13.6: Crystal structure of $Pb(Zr_{1-x}Ti_x)O_3$. By application of stress or voltage, the center Ti^{4+} or Zr^{4+} ions move out of the oxygen plane, creating an electric dipole moment.

13.4.2 US head

PZT crystals are the essential part of US transducers for US imaging and they are an integral part of US transducer heads. A schematic of such a US head is shown in Fig. 13.7 (a). In the transmitting mode, the energizing voltage is applied to the back side of a piezoelectric disk covered by a thin metal film connected to a coaxial cable, while the front side is grounded. Furthermore, on the back side of the piezocrystal there is some porous damping material, ensuring that the backward traveling sound-wave is not reflected forward where it would disturb the main signal.

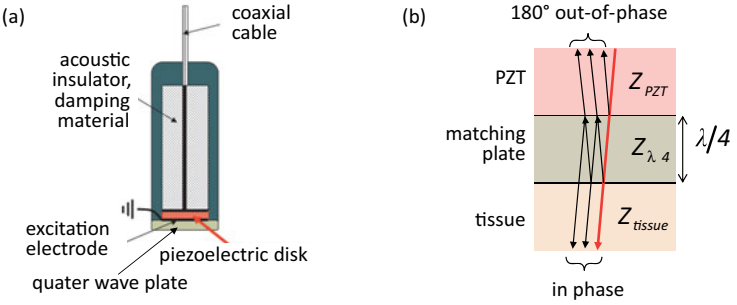


Fig. 13.7: (a) Section through an ultrasound transducer consisting of a piezoelectric disk connected to a coaxial cable, damping material on the back side and quarter wave plate as impedance bridge on the front side. (b) Reflection and transmission at the quarter wave plate.

Transducers can be excited by an AC voltage to emit soundwaves at any frequency. But transducers usually are excited at their resonance frequency, which is defined solely by the thickness of the piezoelectric disk. This can be seen as follows. Soundwaves emitted at the front face propagate in the forward and backward directions. The wave traveling backwards is reflected on the back side of the disk and overlaps with the soundwave emitted at the front side. For constructive interference of the waves from the front and the back, the path difference should be a multiple of the wavelength λ and therefore the thickness d of the disk should be $\lambda/2$. This condition assumes that the front and back sides of the transducer are in contact with materials characterized by lower impedances than that of PZT. Under these circumstances the resonance condition is:

$$f_0 = \frac{v_{\mathrm{PZT}}}{\lambda_{\mathrm{res}}} = n\frac{v_{\mathrm{PZT}}}{2d} \quad n = 1, 2, \ldots,$$

where v_{PZT} is the sound velocity in the PZT disk. For a resonance frequency of 1 MHz the thickness of the disk should be 2 mm. Using the transducer as a sensor, the sensitivity is highest at the resonance frequency. Changing the resonance frequency requires changing the disk's thickness.

The front face of the piezoelectric disk is protected by a thin plastic cover of thickness $\lambda/4$, which serves as an impedance bridge to the body. The impedance of the plastic piece is chosen as the geometric mean between the value of PZT ($Z_{\mathrm{PZT}} = 30 \times 10^6\,\mathrm{Ns/m^3}$) and of soft tissue ($Z_{\mathrm{tissue}} = 1.5 \times 10^6\,\mathrm{Ns/m^3}$): $Z_{\lambda/4} = \sqrt{Z_{\mathrm{PZT}} \cdot Z_{\mathrm{tissue}}}$. The thickness of the quarter wave plate is fixed for the soundwave length used and therefore matching is only ideal for one particular wavelength. The quarter wave plate enhances transmission of US in the forward direction and suppresses backreflection into the PZT according to the principle illustrated in Fig. 13.7 (b). Soundwaves transmitted into the $\lambda/4$ plate reverberate back and forth at both interfaces. As the $\lambda/4$ plate has an impedance value in between PZT and tissue, the first reflection from the front interface is in phase, whereas the reflection on the back side suffers a phase jump by 180°. By the time the reflected wave arrives again at the front interface, it has a path length difference of $\lambda/2$ and is in phase compared to an incoming and straight through US wave. Therefore these waves reinforce each other by constructive interference. On the other hand, the transmitted wave back into the PZT after first reflection at the interface to the tissue is 180° out-of-phase with the wave reflected at the PZT/matching plate interface and therefore they cancel each other out by destructive interference.

The quarter wave plate is covered with another piece of plastic that is impedance matched to the tissue and acts as an acoustic lens, discussed in the next section. Any air gap between the US head and the body is to be strictly avoided because of the huge impedance mismatch between US head and air and at air/body interface, causing total reflection of the soundwave instead of transmission into the body. Therefore a gel is used to cover the body part to be scanned and the US head is completely immersed in the gel. The gel has impedance similar to the plastic cover.

13.4.3 Time gain compensation

Transducers for US imaging are used in pulse mode. Once excited the duration of the US pulse depends on the damping constant. Resonators with a high quality factor Q exhibit a lower damping and therefore emit a longer pulse than resonators with low Q. The quality factor Q is defined as the ratio of the resonance frequency f_0 to the bandwidth (full width at half maximum, FWHM) of the frequency distribution Δf, which usually has a Gaussian form (Fig. 13.8):

$$Q = \frac{f_0}{\Delta f}.$$

Typically the length of a pulse wave train comprises about three wavelengths, corresponding to a pulse duration (PD) of $3T = 3/f = 3\,\mu s$ at 1 MHz. The Fourier transform of such a short pulse is rather broad, typically in the order of $0.8 f_0$, corresponding to a low Q factor of 1.25.

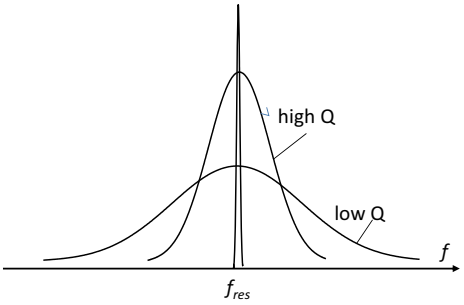


Fig. 13.8: Fourier transform of a US pulse for different quality factors of the resonator. The sharp line in the center corresponds to the intrinsic resonance frequency of the transducer.

The *pulse repeat frequency* (PRF) or probing frequency for US imaging is typically 1 kHz. Thus between any pulse of a few microseconds length there is almost a millisecond interval until the next pulse is fired. This rather long waiting time is utilized to switch the transducer from pulsing mode into sensor mode of echo signals after the time of echo TE. The switching is done by a duplexer (see Fig. 13.1). Duplexers, not to be confused with the duplex detection mode discussed in Section 13.7.2, are electronic devices that allow bi-directional communication via one joined signal path, such as an antenna for transmitter and receiver of microwaves in radio communication, or one US head for transmitting and receiving US waves. Duplexers can be based on frequency, time, or amplitude. In the case of the US duplexer, switching is achieved by pulse amplitude using nonlinear electronic devices (pn-junctions): high pulse amplitude for transmission and low pulse amplitude for detection.

While the pulse is easily produced by a pulse generator with optional variations of amplitude, pulse duration, pulse mean frequency and repeat frequency, the transducer in sensor mode has to deal with much lower and varying amplitudes that need to

be corrected for attenuation and shaped for further signal processing. The time scales are shown in Fig. 13.9 for a US frequency of 1 MHz and a repeat frequency of 1 kHz. As soon as the first echo pulse arrives, successive echo pulses are amplified on a logarithmic scale to ensure that all echoes from congenial interfaces at different distances to the transducer have equal signal amplitude, as indicated in the middle panel of Fig. 13.9. This amplification process is referred to as *time gain compensation* (TGC). Once the first echo signal has arrived, TGC is triggered to equalize the amplitude of echo signals. In the simplest case TGC is linear on the log dB versus time scale, but can be adapted to special circumstances.

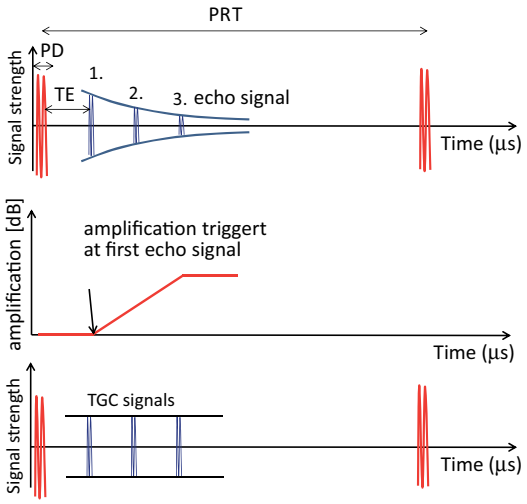


Fig. 13.9: Time sequence for pulsing and detecting US. Top panel: US pulses are emitted in time increments of the pulse repeat time (PRT). The pulse duration (PD) is short compared to the echo time TE and the PRT. After the first echo signal has been received, second and third echo signals will be detected with exponentially decaying amplitude. Middle panel: Once the first echo signal has arrived, time gain compensation (TGC) is triggered to equalize the amplitude of echo signals. Bottom panel: The result is roughly equal amplitude for all echo signals, as long as the original signal is still above noise level. Typical values are PRF = 1 kHz, i.e., PRT = 1000 µs; PD = 3 µs; TE = 100 µs.

13.4.4 Near field and far field

Next we consider the proper size, shape and diameter of transducers. If the transducer had the size in the order of the emitting wavelength, the wavefront would form a spherical surface. This is the opposite of what is required for US imaging. For US imaging a planar wavefront is preferred with potential focusing options. For this condition the size of the emitter should be much larger than the wavelength of the soundwave, at least 10 times as large. For a 1 MHz source this requirement implies a head size of at least 15 mm diameter.

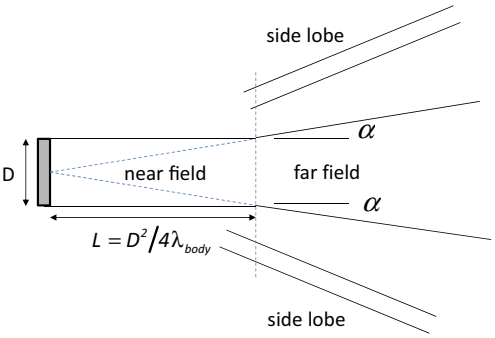


Fig. 13.10: Near field and far field of a soundwave emitter under the assumption that the source size $D \gg \lambda$.

In Fig. 13.10 the near field and the far field of an extended soundwave source are shown. In the near field the main propagation direction is in the forward direction, all rays spreading out to the side are extinct by destructive interference. The near field extends up to a distance of:

$$\frac{D^2}{4\lambda_{\text{body}}} = \frac{D^2 f_0}{4v_{\text{body}}},$$

where D is the source size and v_{body} is the sound velocity in body tissue. Within this distance known as the *Fresnel region*, the wavefront is nearly a plane wave. At larger distances in the far field, also known as the *Fraunhofer region*, interference effects are lost and the beam starts to diverge. At the edges of the sound field there are small areas of low intensity outside of the main beam, corresponding to first side maxima of a diffraction pattern, called side lobes. These intensities may cause artefacts in the US imaging. To give numbers, for a 1 MHz source with a source size of 15 mm, the near field in soft tissue extends up to about 36 mm. With higher frequencies the near field can be increased. As we shall see later, the near field region is the favorable distance for imaging organs. In the far field the angle of divergence α is $\alpha = 2\lambda_{\text{body}}/D$, i.e., low frequencies (large wavelengths) diverge more than high frequencies.

Obviously it is favorable to focus the ultrasound field in the region of interest, i.e., at specific organs. Focusing enhances the intensity of the echo signal and increases the lateral resolution, which otherwise is given by the source size in near field. Focusing can be achieved by several different methods; three are sketched in Fig. 13.11:

(a) focusing by a curved PZT transducer;
(b) focusing by a plastic lens in front of the PZT crystal;
(c) focusing by signal arrival time.

Time focusing can be realized if the transducer is subdivided into several smaller elements and element 1 emits first, followed by elements 2 and 3, etc. at time increments corresponding to the path differences. Time focusing requires precise electronic timing. Changing the time increments between the elements changes the focal length. Therefore, time focusing is very flexible.

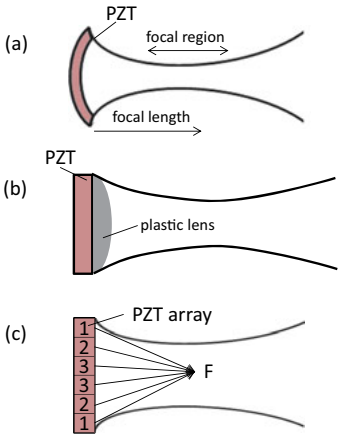


Fig. 13.11: Three different methods for realizing a focus of the US wave field. (a) Focusing by a curved transducer; (b) focusing by a convex plastic lens in front of a flat transducer; (c) electronically controlled time focusing.

13.5 Medical imaging

As already mentioned in the introduction, medical sonography is an imaging modality that takes static images of organs and tissues, real time images of periodically moving organs (heart and lung), and Doppler measurements from blood flow. We will first discuss static imaging, which is subdivided into A- and B-scans. A-scans are no longer used in practice. However, the A-scan illustrates rather well the principles of US imaging.

13.5.1 A-scan

In an A-scan (= amplitude mode echo ranging) the US head is kept stationary at one particular location. US pulses are transmitted into the body with the help of a gel between the US head and the skin to avoid any reflection at this first interface. Therefore an echo signal from position "0" in Fig. 13.12 is not expected. As soon as the soundwave travels further into the body and hits an organ at depth L (position 1) with impedance Z_2, different from the surrounding tissue with Z_1, part of the soundwave is reflected and the echo signal will arrive after a travel time:

$$\Delta t = 2L/v_{\text{body}}$$

at the detector. Although the velocity varies between different body parts, for signal processing an average and constant sound velocity in the body of $\bar{v}_{\text{body}} = 1540\,\text{m/s}$ is assumed. On the back side of the organ (position 2) the soundwave is again reflected, but the echo signal is weaker than from the top interface because of damping within the organ. Finally the echo signal arrives from the back side of the body at position 3. These three signals can be seen as "blips" on the screen of an oscilloscope, where the horizontal axis is the time axis (equivalent to the depth of the reflected soundwave),

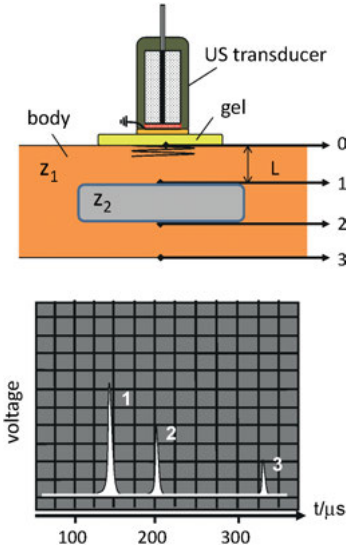


Fig. 13.12: Ultrasound transducer as source of soundwave pulses and sensor of echo signals from the interfaces at positions 1, 2, and 3. In the lower panel the echo signal is displayed on an oscilloscope screen.

and the vertical axis is the voltage output after signal amplification, which is proportional to the echo signal amplitude. Because of this display mode the A-scan is also known as *amplitude mode*.

In A-scans no further signal processing is performed. These unprocessed signals help locate a particular organ and optimize the echo signal with respect to pulse length, frequency, appropriate focal length, etc. Obviously the A-scan has no lateral resolution, but it has depth (axial) resolution, which is coupled to the time resolution. The time resolution depends on the FWHM of the pulse length or the bandwidth Δf of the pulse. For standard pulse signals the bandwidth is usually about $0.8 f_0$, where f_0 is the resonance frequency. From this an axial resolution ΔL can be estimated according to:

$$\Delta L = \frac{\bar{v}_{\text{body}}}{2\Delta f} = \frac{\bar{v}_{\text{body}}}{2 \times 0.8 f_0}.$$

Signals received within a time lapse of 1 µs (FWHM of the bandwidth) are from interfaces separated by 750 µm, which is the quoted depth (axial) resolution. With shorter pulses higher depth resolution can be achieved. However, as shorter pulses require higher frequencies, the damping will increase and also the attenuation due to scattering. Therefore, a compromise between resolution and signal strength has to be found for each application.

13.5.2 B-scan

The most frequently used scan in US imaging is the B-scan or brightness mode imaging. The B-scan can be regarded as a sequence of A-scans in time and space that

together represent a slice through the body. There are three basic types of B-scanners, which are presented in the following: sector scanner, array scanner, and phased array scanner.

Sector scanner

The basic principle of a sector array B-scanner is shown in Fig. 13.13. In this particular example a transducer head is rocked back and forth in an impedance matching fluid, emitting pulses perpendicular to its surface at different rocking angles. The backscattered and backreflected waves from interfaces between organ (Z_2) and surrounding tissue (Z_1) are detected by the transducer and are stitched together, yielding the contours of the imaged organ on a monitor screen as indicated in the lower panel by the dotted line. The gray scale of each dot represents the signal strength of the reflected soundwave at the interfaces. Backreflection from the far end of the body is usually so weak by absorption and scattering that it can be safely neglected. Here and in all other B-scans TGC is switched on for amplifying the signal. Then with TGC on, variations of echo signal amplitudes are ideally only due to variations in acoustic impedance and not due to depth. However, applying TGC also increases the noise and the signal-to-noise ratio (SNR) may worsen. Nevertheless, TGC provides a more balanced image.

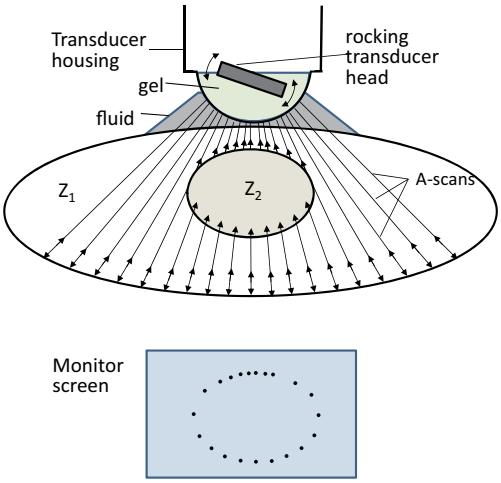


Fig. 13.13: Sector scan of an organ, consisting of a sequence of A-scans. The transducer head is rocked in an impedance matching fluid and soundwaves are emitted perpendicular to the transducer surface. Echo signals form the contours of an organ at one particular slice.

Array scanner

Instead of sweeping the transducer head mechanically, various other arrangements of emitter/receiver elements and of electronic sweep control are available in modern US heads. One such arrangement is sketched in Fig. 13.14. An extended transducer is subdivided into an array of narrow strips, each one about the width of a wavelength. A single strip would not produce a near field but a diverging far field, opposite to the

intention. The single elements are therefore activated in groups. Within a group, time focusing is applied for defining a focal length in point P. For this, elements 3 and 5 in the Fig. 13.14 are activated first, followed by element 4. After a short time delay, groups 4–6, then 5–7, etc. are successively activated, until a sweep is completed. In reality the number of elements in the array is much larger than indicated in the schematic.

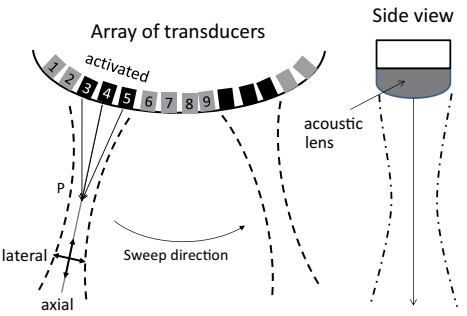


Fig. 13.14: Left panel: array scanner for an electronic sweep of US across an organ. Each transducer element is much smaller than in the section scanner. Right panel: side view also featuring an acoustic lens for defining the width of the imaged slice.

As previously mentioned, the focal length in the sweep direction (lateral or azimuthal plane) is adjusted by time focusing and can be electronically controlled. A larger time delay will result in a shorter distance of the focal point P. Focusing in the perpendicular direction is achieved by an acoustic lens, as shown in the side view of the array, defining the thickness of the slice imaged.

Phase array scanner

Similar to the array scanner, the phase array scanner consists of many transducers arranged in a linear array (Fig. 13.15). But there are two important modifications in comparison to the previous scanners. Each single element is shorter than in the array scanner and there are fewer of them. All elements are energized almost at the same time, however with a small time delay between neighboring elements. The time delay determines the phase difference and the extent of destructive and constructive interference of neighboring elements. The phase difference finally controls the sweep direction and the time focusing point. The sweep covers a sector field similar to the sector scanner but without any moving part.

Different types of scanners used for US imaging are shown in Fig. 13.16. The field of view (FOV) is the portion of organs or tissues that are intersected by the scanned slice. Depending on the probe used, the shape of this field can be a sector, a rectangular, a trapezoid, or a convex field.

Independent of which scan type is used for the display, the echo signal amplitude is encoded in the form of gray dots on a two-dimensional matrix, whose axes refer to the penetration depth versus the lateral sweep direction. The brightness of the gray dots after gain application encode the echo amplitude: bright dots correspond to

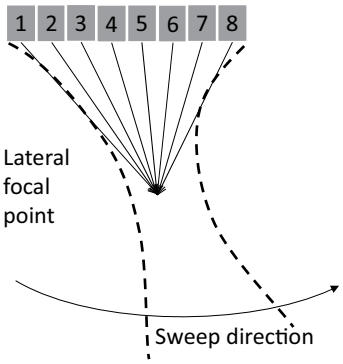


Fig. 13.15: Phased array scanners consist of a linear array of small transducer elements, which are activated simultaneously with minute time delays between neighboring elements defining sweep direction and focal length.

strong signals, dark dots represent weak signals. Because of this brightness scale the B-mode scan is also known as *brightness mode imaging*. Typically, strong intensities originate from near surface areas, whereas weaker signals are from deeper locations. But a number of artefacts can modify this simple interpretation, which is discussed in the next section. A typical example of a B-scan from the ventricles of the heart is shown in Fig. 13.2. The graph has to be read from top (close to the skin) to bottom (deeper inside). On a grey scale, high reflectivity (bone) appears in white; low reflectivity (muscle) has a grey tone, no reflectivity (water) shows up in black. Blood and blood vessels are dark. Many more examples can be found on webpages listed at the end of this chapter.

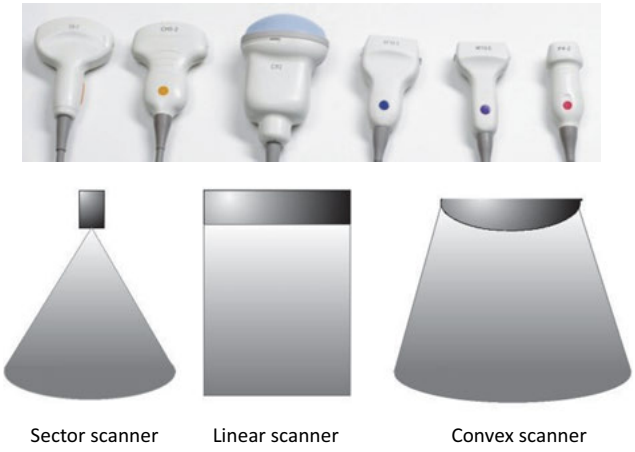


Fig. 13.16: Upper panel: various types of scanners used for US imaging.
Lower panel: depending on the scanner type, the imaged slice may have different shape.

13.5.3 C-mode

C-mode scans are collections of B-mode scans for reconstructing two-dimensional images of inner body organs at constant depth in the direction perpendicular to B-mode scans. All B-mode scans discussed so far take one slice of the body at a time. In C-mode, a sequence of B-mode scans is acquired in a direction normal to B-mode slices (Fig. 13.17). Then the pictures are stitched together and a plane of constant depth is selected, indicated by the red bordered area in Fig. 13.17. Three different scan types can be distinguished as sketched in Fig. 13.17: linear scan, sweep scan, or rotational scan. The rotational scan requires a transducer at the end of a stick, which is mainly used for rectal or vaginal screenings. In all cases, C-scans add information on a slice perpendicular to the scan direction.

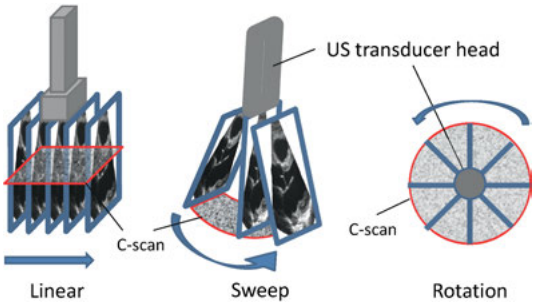


Fig. 13.17: Perpendicular cuts through B-mode scans at constant depth form two-dimensional images. The red bordered areas correspond to C-mode scans.

If properly stitched together, either B-mode or C-mode scans have the potential to generate tomographic images. However, in contrast to MRI and CT, in C-mode scans the coordinate system is not fixed. The manually held US head has to touch the skin and follow the contours of the body surface. This makes it much more difficult to reconstruct 3D images. In the past, practitioners had to mentally integrate sequences of pictures to get a feeling of the size and shape of an organ in the lateral direction. In recent years different methods have been proposed and are being used, all relying on fast data acquisition and computer processing. The simplest version is a handheld scan of predefined mode (linear, sweep, rotation) and preselected scan speed. The resulting image will give a 3D impression, although the geometry may not be precise but sufficient for a diagnosis. Sweep scans are easier to perform manually than linear scans since only a rotation at a constant position is required. However, the penetration depth varies with rotation angle because the distance between transducer head and organ changes by rotation. Manual scans can be improved by tracking the US head with a receiver attached to it in a gradient magnetic field setup around the patient, similar to techniques used for MRI imaging (see Chapter 15). Furthermore, linear scans can be acquired by mechanically scanning the transducer head using stepper motors, while the contour of the body surface is monitored capacitively and fed into a DC motor that

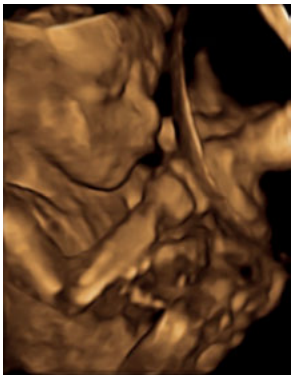


Fig. 13.18: Surface rendered 3D image of a twin fetus (head of one and foot of the other) taken by a 3D echo scan (private communication).

adjusts the elevation of the transducer. An overview of methods and techniques for 3D US imaging is provided in [1]. Impressive high quality 3D topography images can nowadays be taken, such as the one shown in Fig. 13.18 of a fetal face.

13.5.4 M-mode

M-mode or motion mode provides information about variations in signal amplitude versus time due to object motion. At a fixed position of the transducer head a line of data is acquired by an A-mode scan. The data is displayed as a series of dots or pixels with brightness level representing the intensity of the echoes. In repeating A-lines at a fixed position, the signal intensity variation is displayed on a horizontal time axis, while the y-axis indicates the distance of the echo from the transducer. An example of an M-mode image from the beating heart is shown in Fig. 13.19.

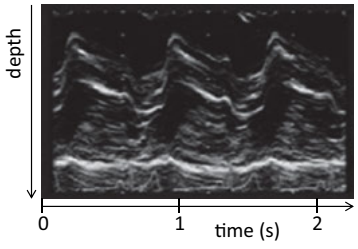


Fig. 13.19: M-mode display of mitral valve leaflet of a beating heart (reproduced from https://openi.nlm.nih.gov/).

The same concept can also be used for taking videos not just of a single A-line but of a full sector scan taken in B-mode. Present technology allows the recording of about 50 frames per second (FPS) with sufficient depth information and resolution. This is adequate temporal resolution for 2D visualization of normal heart action with a beat frequency of 70 beats per minute. A 3D echocardiogram of a beating heart can be seen on the webpage of [2].

US transducers can also be combined with endoscopy (see Chapter 14) to image the heart through the esophagus, the prostate through the rectum, and the fetus through the vagina.

13.6 Scan characteristics

After discussing the general concepts of A-, B-, C-, and M-scans, some additional considerations about focus, image formation, and resolution are in order, which is of concern to all scanning modes.

13.6.1 Focusing

Focusing improves the spatial resolution and image quality of objects within the focal depth, but it degrades the image quality beyond. To overcome this problem, multizone focusing has been introduced [3]. Along each sweep one focus per pulse emission is applied, which is fixed but can be adjusted to the depth of view. These pulses are emitted at the usual pulse repeat frequency (PRF). For receiving the echo signal, dynamic focusing is applied. This is achieved by a continuous phase control of the receiving elements within a phase array scanner. Starting from the near field region and continuing to deeper fields each time the number of detector elements is increased to receive echoes from corresponding focal zones. This is schematically illustrated in Fig. 13.20. For receiving signals from shallow regions, three detector elements (–1, 0, 1) in an array of 15 are activated, for receiving signals from deeper regions, further elements from –3 to 3 are added, etc.

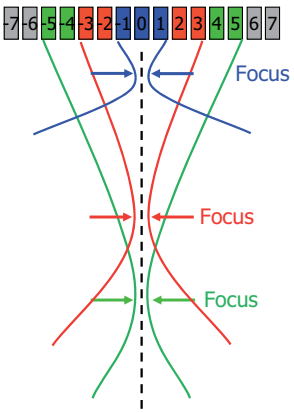


Fig. 13.20: Dynamic focusing for depth-dependent detection of echo signals (adapted with permission from [3]).

13.6.2 Line density

Whatever scan technique is used, the tissue of patients is always sampled by a number of line scans for completing one slice (section). The spatial resolution in the sweep direction is intrinsically limited to about 1 mm. Therefore, for a 10 cm wide slice 100 lines are necessary for acquiring a high quality high resolution image.

13.6.3 Scan frequency

Moving tissue requires a high pulse repeat frequency (PRF) in order to follow the movement. The PRF depends on the number of lines per frame and the frame rate:

$$\text{frame rate} \times \text{lines per frame} = \text{PRF}.$$

Typically, 30–50 frames/s are taken with 100 lines/frame, which yields a PRF of 3 kHz.

13.6.4 Depth of view

To image tissue at a particular depth, the sound pulse needs to have sufficient time to travel forth and back before the next pulse is emitted. This limits the PRF. When increasing the *depth of view* (DOV), the PRF has to be reduced correspondingly:

$$\text{DOV} = \frac{0.5 v_{\text{sound}}}{\text{PRF}}.$$

From this consideration we conclude that it is not possible to take high resolution images of moving tissues at large depth. A compromise is necessary either with respect to resolution or to depth. From the above relations follows the criteria:

$$\text{DOV} \times \text{PRF} \leq 0.5 v_{\text{sound}} = \text{const}.$$

Any choice of parameters has to obey this boundary condition. These considerations are independent of the actual operating frequency used. The choice of US frequency has, in addition, a large effect on the penetration depth of US, focus, and resolution.

13.6.5 Penetration depth

The penetration depth of US in the body is a complex function of frequency. It depends on viscous damping and heat conduction, on scattering at small particles and rough interfaces, and on already backreflected intensity at interfaces. When US waves hit bones, water, or gas-filled tissue, penetration into tissue on the back side drops dramatically. Because of this complexity it is difficult to give a general expression that

covers all eventualities. However, the half intensity depth $L_{1/2}$, i.e., the depth at which the intensity has dropped to half of its original value, as a function of frequency f can be estimated with the help of an empirical expression:

$$L_{1/2} = C/f.$$

The constant $C \approx 50$ cm MHz. Some examples are given in Tab. 13.2. One method to circumvent the attenuation problem is by so called (higher) *harmonic imaging*: the beam is transmitted at a fixed fundamental frequency f_0, whereas the received signal is analyzed at higher harmonics of $2f_0$ or $3f_0$. This increases the SNR of reflected signals particularly from the deepest parts of images without compromising resolution.

13.6.6 Spatial resolution

Spatial resolution is the ability to observe two objects, A and B, separated by a certain distance ΔL in space as two distinct images A' and B' on a screen. If the objects A and B are too close together, the echo pulses from A and B will overlap and will not be distinguishable. In US imaging the axial or depth resolution is very different from the lateral resolution for intrinsic physical reasons. Therefore we need to separate the discussion of both.

Axial resolution is a question of pulse length. The higher the frequency of soundwaves, the shorter the pulse length that can be chosen and the better the axial resolution that can be achieved. The axial resolution is the same as already quoted for A-scans, where Δf is the pulse bandwidth:

$$\Delta L = \frac{\bar{v}_{\text{body}}}{2\Delta f} = \frac{\bar{v}_{\text{body}}}{2 \times 0.8 f_0}.$$

Examples of axial resolution for different frequencies are listed in Tab. 13.2.

The lateral resolution δ_{lat} is given by the width of the focal point. The width of the focal point follows from the axial depth of the focal point L, the wavelength λ, and the width of the active aperture length D:

$$\delta_{\text{lat}} = \frac{L\lambda}{D} = \frac{\lambda}{2\alpha}.$$

Axial depth L and angle α are defined in Fig. 13.15. The last equation implies that lateral resolution depends linearly on the US wavelength, which is confirmed by the values listed in Tab. 13.2.

In general, the following conclusions can be drawn, which are tradeoffs between penetration depth and spatial resolution. *Low frequencies* provide high penetration depth at the expense of low axial and lateral resolution; *high frequency* US waves have low penetration depth, but high axial and lateral resolution.

Tab. 13.2: Penetration depth, lateral resolution and axial resolution for different US frequencies or wavelengths in body tissue with an average sound velocity of 1540 m/s (from PD Dr. Marc Kachelrieß, Erlangen).

f [MHz]	λ [nm]	Depth [cm]	Lateral [mm]	Axial [mm]
2.0	0.78	25	3.0	0.80
3.5	0.44	14	1.7	0.50
5.0	0.31	10	1.2	0.35
7.5	0.21	6.7	0.8	0.25
10	0.16	5.0	0.6	0.20
15	0.10	3.3	0.4	0.15

13.6.7 Artefacts

Artefacts are quite numerous in US imaging. This makes it rather difficult to interpret US images properly. The practitioner must know these pitfalls to avoid potentially wrong diagnoses. The most important artefacts are as follows:

- Speckle images are due to objects smaller than the length of soundwaves, giving rise to strong scattering in all directions (4π), also known as Rayleigh scattering.
- Reverberation occurs when the US wave is scattered several times between different organs before being received in the detector. The detector shows a series of delayed echoes which appear as spurious objects in the distance.
- Double reflection may occur at two parallel interfaces like the diaphragm. Structures in the liver can appear to lie in the lung when reflected twice.
- Acoustic shadowing occurs behind strongly reflecting objects such as bones and gas-filled hollow spaces.
- Acoustic enhancement takes place when the attenuation of sound is less than in normal tissue, such as in liquid-filled spaces like the bladder. Then the tissue behind such organs appears brighter than expected.
- Refraction occurs at inclined flat interfaces, which makes objects appear closer than they are, like the refraction effect of light observed in swimming pools.

13.7 Doppler Method

13.7.1 CW Doppler method

The Doppler effect is well known in physics and describes changes of frequencies if either source or observer move with a speed v_{Source} or v_{Obs}, respectively. The frequency shift occurs as a result of an artificial wavelength change whenever the source of sound or the receiver of sound have a finite velocity projection in the direction of source and receiver. The wavelength change can be translated into a frequency change as recog-

nized by the observer. The most general expression for the Doppler effect is:

$$f_{\text{Obs}} = f_{\text{Source}} \frac{v_{\text{sound}} \pm v_{\text{Obs}}}{v_{\text{sound}} \mp v_{\text{Source}}}.$$

Here f_{Obs} is the frequency received by the observer, f_{Source} is the frequency of the source, and v_{sound} is the sound velocity in the respective medium.

In medicine the Doppler effect is used for determining the velocity of blood in organs like the heart and kidneys or in blood vessels. This is done by emitting US soundwaves from a steady source into tissues containing moving blood cells (erythrocytes). Then in a first step the blood cells receive Doppler shifted US waves. The frequency shift is:

$$f_{\text{Source}} - f_{\text{Obs}} = \Delta f = f_{\text{Source}} \frac{v_{\text{blood}}}{v_{\text{sound}}}.$$

Here we have assumed that $v_{\text{blood}} \ll v_{\text{sound}}$. In a second step the blood cells become the moving source of frequency shifted US frequencies by scattering the US waves. These scattered waves are then detected by a receiver at rest. For this second step the same equation applies again, such that the total observed frequency shift by the transducer is:

$$\Delta f = 2 f_{\text{Source}} \frac{v_{\text{blood}}}{v_{\text{sound}}}.$$

Finally we need to consider that the transducer as source and as receiver of US waves encloses an angle θ against the blood flow direction (Fig. 13.21) and only the projection of the flow is detected. Therefore we have for the frequency shift:

$$\Delta f = 2 f_{\text{source}} \frac{v_{\text{blood}}}{v_{\text{sound}}} \cos\theta.$$

Notice that the shift Δf increases with the blood velocity, with the frequency of US waves, and that Δf is largest for an angle $\theta = 0$ but zero at perpendicular orientation. The sign of Δf indicates the direction of flow, towards or away from the transducer. The sign is often color coded on displays, red for flow towards the transducer, blue for flow away.

The operating frequencies for US Doppler measurements are between 2 and 10 MHz, according to the probing depth (see below). For precise measurements of frequency shifts a high Q of the transducer is required. Therefore the transducer is continuously energized at the resonance frequency and the usual damping block on the back side of the reducer is removed. In fact, continuous operation requires the action of two transducers, one for continuously emitting US waves and the other one for continuously receiving the US echo signal. Usually both transducers sit in the same housing with little separation between them as indicated in Fig. 13.21. The arrangement is similar to an A-scan with the important difference that Doppler applications require an inclination angle $\theta \neq 90°$.

A quick estimate tells us that the frequency shift is not large. Using a 5 MHz US source, an average sound velocity of 1540 m/s in tissue, and assuming a blood velocity

of 2 m/s the frequency shift observed at angle $\theta = 0$ is only

$$\Delta f = 2 \times 5\,\text{MHz}\frac{2}{1540} = 13\,\text{kHz}.$$

The frequency shift of US waves due to blood flow is in the order of 0.1 % of the emitted frequency and the difference is in the audible frequency regime! Filtering out this low frequency shift is done by *fast Fourier transform* (FFT). First the US wave with the source frequency f_{source} and amplitude A_1 (Fig. 13.21 (a)) is multiplied by the wave detected at the receiver having much smaller amplitude A_2 and Doppler shifted frequency $f_{\text{source}} \pm \Delta f$ (Fig. 13.21 (b)). The result is shown in Fig. 13.21 (c). After filtering the high frequency component, the wave with the Doppler beat frequency Δf is obtained (Fig. 13.21 (d)). This frequency can be made audible by a loudspeaker. The higher the pitch, the higher the velocity is. However, it cannot distinguish between positive or negative frequency changes, i.e., between flow directions towards the receiver or away from the receiver.

For converting Doppler shifts into velocities the inclination angle θ must be known. The operator needs to adjust a cursor on the display parallel to the flow direction and the operational system calculates the angle θ and the blood velocity.

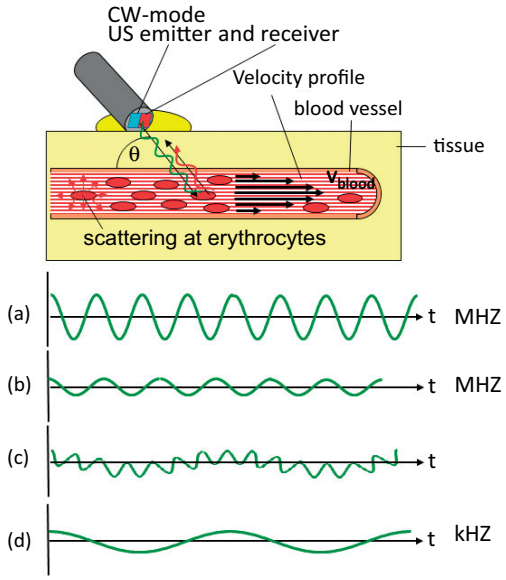


Fig. 13.21: The top panel shows the usual arrangement with a US wave emitter on the skin coupling a continuous wave (CW) into tissue that contains blood vessels. The MHz US wave is scattered by erythrocytes. The scattered and Doppler shifted wave at angle θ is detected in the receiver. Emitter and receiver are split. (a) Emitted US frequency; (b) Doppler shifted frequency received from backscattering; (c) product of original wave and Doppler shifted wave; (d) Doppler frequency in the kHz regime after filtering.

This is shown in a screen shot in Fig. 13.22. In the lower part the Doppler frequency is plotted versus time, which can be converted to velocity versus time via:

$$|v_{\text{blood}}| = \frac{\Delta f}{2f_{\text{sound}}} \frac{v_{\text{sound}}}{\cos\theta}.$$

The time structure expresses the pulsing frequency/velocity in the rhythm of the heartbeat. The peak maximum corresponds to the peak systolic velocity (PSV) and the minimum at the end of the tail is the end diastolic velocity (EDV). From these two values different quantities can be derived, characterizing proper blood flow versus potential stenosis independent of the inclination angle. The resistance index RI is defined as the ratio: RI = (PSV – EDV)/PSV. The pulsatile index PI is defined as: PI = (PSV – EDV)/MV, where MV is the mean velocity. PI is more difficult to evaluate than the RI, because the former involves the shape of the velocity profile and requires an integration of the profile. Often simply the ratio PSV/EDV is taken. For these indices reference values are known and tabulated for different blood vessels and organs, such as the carotid artery, the kidneys, the umbilical cord, etc.

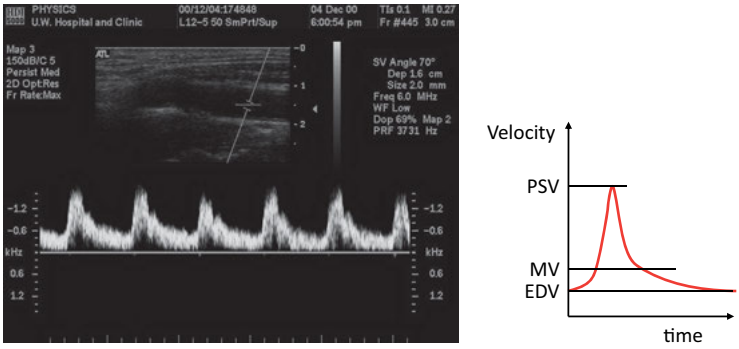


Fig. 13.22: Left panel: the source frequency (blood) versus time is plotted on the screen. The upper part shows a US image of the blood vessel taken in B-mode. The fine line inclined against the horizontal indicates the detector angle (70°). The fine horizontal line crossing the inclined line is the middle of the blood vessel indicating where the blood velocity is measured. Right panel: velocity profile with definitions of velocities: PSV = peak systolic velocity; EDV = end diastolic velocity; MV = mean velocity.

Actually, blood velocity not only changes in time according to the heartbeat. In addition there is a velocity profile across the diameter of the blood vessel. Assuming laminar flow, the velocity is highest in the center and drops to almost zero towards the walls. Therefore it is necessary to define a sample volume length over which the velocity is determined. If the volume is too large, the velocity spectrum spreads out, while the maximum velocity is still clearly visible for laminar flow. However, a better signal is obtained if the sample volume length is reduced to a narrow range of interest

(ROI), filtering out velocities close to the walls. Since blood is a non-Newtonian type fluid (see Chapter 8), the velocity profile is more flat in the center than expected from a parabolic profile. This helps focusing on the center and small deviations are not so severe.

By the same means the Doppler method can distinguish between laminar flow and turbulent flow. Turbulent flow may occur either due to a large diameter artery such as the carotid artery, or due to local constrictions. Deviations from laminar flow can be detected by the frequency or velocity distribution as a function of time as illustrated in Fig. 13.23. In the case of turbulence, both the velocity profile and the Doppler angle show a wide distribution.

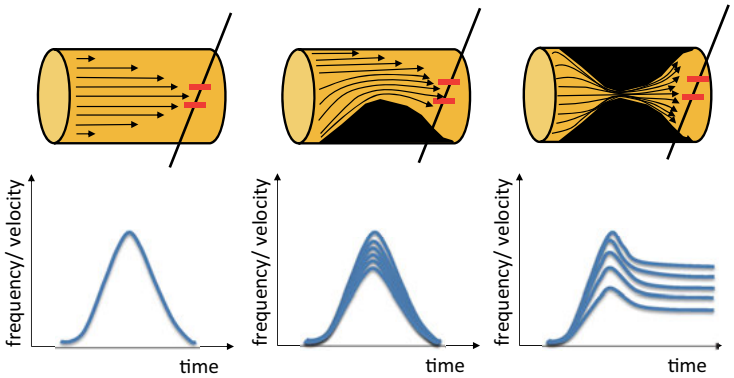


Fig. 13.23: Velocity distribution for laminar flow (left panel), some constriction (middle panel), and more severe constriction (right panel), causing turbulent flow. The region of interest selected by the detecting system is indicated by red bars.

A typical application is a check-up of the carotid artery blood supply to the brain. The artery goes through the base of the skull to the brain. Carotid means "providing sleep". Since any small disturbance in the blood flow can affect consciousness and may promote a stroke, testing the blood flow with the help of the Doppler effect can indicate a carotid artery stenosis (carotid stenosis), i.e., calcification of the carotid artery, already at an early stage. A typical Doppler test of the carotid artery is demonstrated in Fig. 13.24. The Doppler velocity test can even be combined with a simultaneous electrocardiogram (ECG) to determine the phase relationship between systolic ejection and peak velocity at the carotid artery from which the pulse wave velocity (PWV) is determined (see Section 8.5 for more details).

Now we discuss the optimum frequency for the US Doppler shift application. The scattering of US at the erythrocytes is of the Rayleigh type, since the extension of blood cells (6–8 µm) is much smaller than the wavelength of US (300 µm at 5 MHz). Rayleigh scattering is an isotropic scattering in a solid angle of 4π, which scales with the frequency according to f^4. Therefore it would be advantageous to use higher US

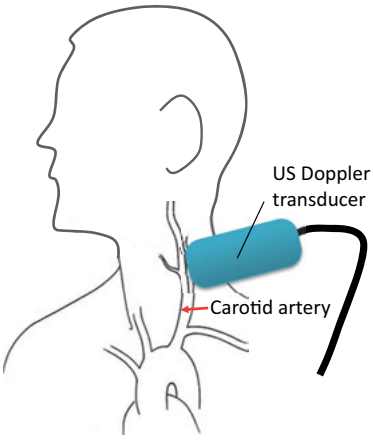


Fig. 13.24: Test of the blood flow through the carotid artery with the help of a US Doppler transducer.

frequencies to increase the intensity in the receiver. However, higher frequencies also experience higher damping, since viscous damping increases with f^3 and therefore the penetration depth becomes much reduced. A trade-off between signal strength and penetration depth has to be made. For blood vessels close to the surface high frequencies can be used, but for deeper vessels lower frequencies are preferred.

In CW operation it is not possible to locate the moving source or to distinguish between flows in overlapping blood vessels at different depths. On the other hand, with short pulses used for imaging, it is not possible to get accurate Doppler flow information. If information on both location and velocity is required, a compromise between accuracy of depth information (short pulse) and accuracy of velocity (CW) is required. Rephrasing, axial resolution requires large signal bandwidth, whereas velocity resolution requires signal duration. A compromise is a longer signal duration (longer wave train and shorter bandwidth) at higher PRF. This method is known as pulsed Doppler method or *duplex mode*, to be discussed in the next section.

13.7.2 Pulsed Doppler method

In pulsed mode the Doppler method is combined with B-mode scans. This combination is known as the *duplex mode*, which stands for B-mode plus pulse Doppler mode. In duplex mode, some transducers in the array used for B-mode scanning are saved for Doppler shift detection. Duplex mode has the advantage that focusing can be applied as in normal B-mode with inherent depth information. Furthermore, the ROI is adjustable for avoiding integration over all blood vessels in the beam as would be the case in CW-mode. Furthermore, the angle of inclination can be determined directly, as shown in Fig. 13.23. The price to pay for the extra depth information is a wider frequency distribution in pulsed mode and therefore a lower resolution for Doppler shifts compared to CW operation. To overcome this problem to some extent, a pulse length

of about 10 times the wavelength is used, much longer than in normal B-mode operation, where the pulse lengths are about 2–3 times the wavelengths. Another important difference concerns the pulse rate frequency (PRF). The PRF is in the order of 1 kHz for normal B-mode operation, but must be chosen much higher in duplex mode, as we will see below.

Apart from blood velocity measurements at some depth, in duplex mode in contrast to the CW-mode the flow direction can also be determined from the sign of the Doppler shift Δf usually shown in color on the screen, called color Doppler imaging. The color code is as follows: blue for negative Δf (flow away from the transducer) and red for a positive Δf (flow towards the transducer) (Fig. 13.25). This coding is the opposite of the natural color shift in astrophysics: the spectra of stars moving away from us are red shifted, the spectra of stars moving towards us are blue shifted. This natural color coding is unfortunately not transferred to medical equipment.

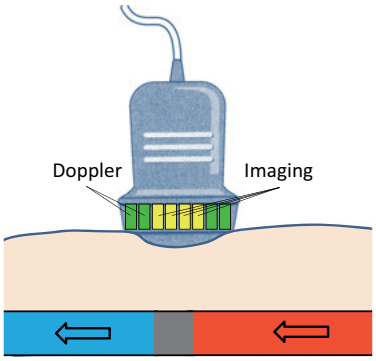


Fig. 13.25: Color coding for the Doppler shift. Flow towards the transducer with positive Δf coded red, flow away from transducer with negative Δf is coded blue. In the gray area under the transducer no Doppler shift is detected. The scanner contains separate elements for Doppler shift detection (marked green) and for imaging (marked yellow).

In pulse mode there are some boundary conditions that need to be considered. First, a new pulse cannot be fired before the first one has been scattered back and received by the detector. The travel time for the pulse is $\Delta t = 2L/v_{\text{sound}}$, where L is the depth of the blood vessel. Thus the maximum PRF is $f_{\text{PRF,max}} = 1/\Delta t = v_{\text{sound}}/2L$. For a depth of 5 cm, the PRF is 15.4 kHz. Second, the minimum PRF depends on the Doppler shift frequency Δf. If the PRF is less than $2 \times \Delta f$, *aliasing effects* occur and the detected frequency in the receiver will show wrong results. The aliasing effect is demonstrated in Fig. 13.26. If the probing frequency f_{probe} is lower than twice the source frequency f_{source}, the frequency of the source may be represented by a too low and false frequency. Thus the criterion for a correct representation of waves is $f_{\text{probe}} \geq 2f_{\text{source}}$, which is known as the *Nyquist criterion*.

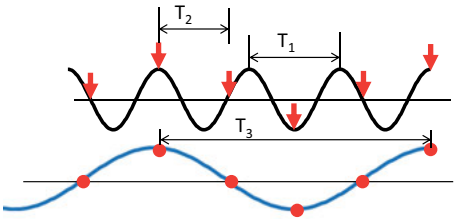


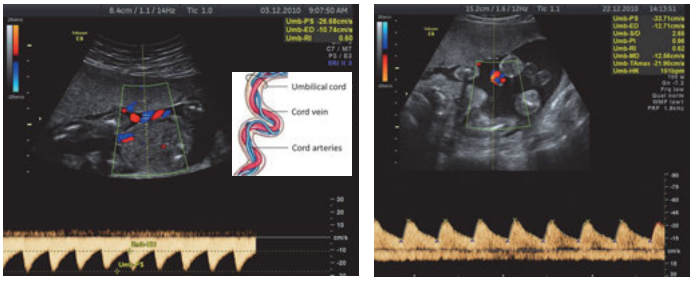
Fig. 13.26: Aliasing effect: the black line is the original wave with a period T_1, red arrows are probing events in intervals of $T_2 < T_1$. The blue line is the result of a probing measurement, yielding a period $T_3 \gg T_1$ and T_2. Only for $T_2 \leq T_1/2$ or for frequencies $f_2 \geq 2f_1$ is the correct frequency of the original wave represented.

If, for instance, the beating frequency $\Delta f = 4$ kHz corresponding to a velocity of 0.6 m/s at a transducer frequency of 5 MHz, the sampling PRF should be at least 8 kHz. The *aliasing effect* therefore sets a lower bound to the PRF: $f_{\mathrm{PRF,min}} = 2\Delta f$, where Δf depends on the blood flow velocity: $\Delta f = 2f_{\mathrm{source}}(v_{\mathrm{blood}}/v_{\mathrm{sound}})$, neglecting the inclination angle. For the given example the boundary condition is 8 kHz $\leq$ PRF $\leq$ 15.4 kHz. Setting these two limits equal, we obtain for the maximum blood velocity that can be detected at a depth L without aliasing effects:

$$v_{\mathrm{blood,max}} = \frac{v_{\mathrm{sound}}^2}{8f_{\mathrm{source}}L}.$$

For a source frequency of 5 MHz, a depth of 5 cm, and a sound velocity of 1540 m/s, we find for the maximum blood velocity that can be detected $v_{\mathrm{blood,max}} = 1.2$ m/s. Higher velocities would be displayed with false colors. In some blood vessels the velocity can be much higher. In order to measure these high velocities the source frequency needs to be reduced. Alternatively one may conclude that high velocities are unsuitable for detection in duplex mode. Low PRF in duplex mode and high blood velocities do not match. This immediately leads to *aliasing effects* and false color coding. The pulse mode obviously has a number of pitfalls and the operator needs to be very careful in choosing the proper parameters.

We conclude this chapter with an example of duplex scans from the umbilical cord of a fetus. The umbilical cord that connects the fetus with the placenta is an intertwined cord of two arteries and a vein shown schematically in the inset of Fig. 13.27 (a) with equal but antiparallel blood flow velocities. In panel (a) the focus is on the blue area at a depth of 8.4 cm, indicated by the dashed vertical line and two short horizontal lines. In this area the velocity is negative with umbilical peak systolic (Umb-PS) velocity of –26.7 cm/s and umbilical end diastolic (Umb-ED) velocity of –10.7 cm/s. It is per se not possible to decide whether this signal is from the vein or from the arteries. However, as the arteries come in a pair, and if two blood vessels next to each other have the same color, it must originate from the arteries. In the present case the blue color is indeed from the arteries. The velocity profile is shown in the lower part



(a) (b)

Fig. 13.27: Duplex mode imaging of the umbilical cord. The inset in panel (a) show a schematic of the umbilical cord with intertwined vein and artery blood vessels. Panels (a) and (b) show simultaneous depth resolution and frequency shifts. The frequency shift from antiparallel velocities in vein and arteries in the umbilical cord are color coded. The respective velocity profiles are shown at the bottom of the respective panels. In panel (a) the frequency shift is negative (blue color) and is from the arteries. In panel (b) the frequency shift is positive (red color) and is from the vein. The region of interest within which the velocity is measured is indicated by the vertical dashed line crossing two horizontal bars.

of the same panel. In panel (b) the focus is on a red part of the cord (vein) at a depth of 15.2 cm. Here we observe a positive velocity profile with Umb-PS = 33.7 cm/s and Umb-ED = 12.7 cm/s. Umb-S/D is the ratio PS/ED = 2.65, the umbilical resistance index is defined as Umb-RI = (PS – ED)/PS, which is here 0.62, whereas the umbilical pulsatile index is defined as Umb-PI = (PS – ED)/M, where M is the mean velocity. As the pulse shape is asymmetric, the mean velocity has to be determined graphically. In panel (b) the umbilical PI is 0.96. Note that here the short terms for the velocities taken from the display screen are slightly different from our previously defined notation. More examples on umbilical cord Doppler imaging can be found in [4].

13.8 Summary

1. Medical sonography has three main applications: (1) static imaging of tissues and organs, (2) dynamic imaging of the heart, and (3) measurement of blood flow velocity and direction.
2. Ultrasonic sonography uses soundwaves with frequencies in the range of 2–8 MHz.
3. Ultrasonic waves are produced by piezoelectric transducers.
4. The transducer also acts as receiver of soundwaves.
5. In medical sonography six modes can be distinguished: A-mode, B-mode, C-mode, M-mode, CW Doppler mode, and pulse Doppler mode.
6. The mechanical index is a criterion for the safe application of US to patients.
7. US propagating into tissue are partially reflected at interfaces separating tissues with different acoustic impedances.
8. The backreflected echo signal is used for imaging organs with different impedance compared to their surroundings.

9. Blood, air chambers, and bones strongly reflect US waves.
10. US waves propagating into tissue are attenuated due to dissipation in viscous media and due to thermal conduction.
11. Time gain compensation alleviates attenuation effects of echo signals.
12. The bandwidth of pulses is typically 80 % of the resonance frequency.
13. Pulse rate frequency is the probing frequency for US imaging, typically 1 kHz.
14. Imaging is usually performed in the near field or Fresnel regime.
15. A-scan is a line scan that probes the depth of reflecting interfaces.
16. B-scan is a sector scan composed of an angular sweep of A-scans. B-scans probe slices of the tissue within the probing depth of US.
17. C-scans consist of sequences of B-scans probed in the lateral direction normal to the B sectors providing two-dimensional images of constant depth.
18. By stitching together C-scans, 3D topographic images can be formed.
19. M-scans provide information on variations in signal amplitude to record moving objects, such as the heart. The M-scan technique is also used for making films of cardiac activity.
20. Low frequency US waves provide high penetration depth but low lateral resolution.
21. High frequency US waves have low penetration depth, but high axial and lateral resolution.
22. US scanning suffers from many artefacts due to double reflections and shadowing.
23. Using the Doppler effect the flow velocity of blood can be determined.
24. Pulsed Doppler wave methods add depth resolution and directional information on the blood flow.

References

[1] Fenster A, Parraga G, Bax J. Three-dimensional ultrasound scanning. Interface Focus. 2011; 1: 503–519.
[2] https://en.wikipedia.org/wiki/Echocardiography
[3] Ermert H, Hansen C. Ultraschall. In: Dössel O, Buzug TM, editors. Medizinische Bildgebung. Vol. 7. de Gruyter; 2014. p. 217–326.
[4] https://sonoworld.com/Client/Fetus/html/doppler/capitulos-html/chapter_01.htm

Further reading

Ermert H, Hansen C. Ultraschall. In: Dössel O, Buzug TM, editors. Medizinische Bildgebung. Vol. 7. de Gruyter; 2014. p. 217–326.
Dhawan AP. Medical image analysis. 2nd edition. Wiley-IEEE Press; 2011.
Bushberg JT, Seibert JA, Leidholdt EM Jr, Boone JM. The essential physics of medical imaging. 3rd edition. Lippincott Williams & Wilkins, Wolter Kluwer; 2012.
Kremkau FW. Sonography principles and instruments. 9th edition. Elsevier; 2015.
Hoskins P, Martin K, Thrush A, editors. Diagnostic ultrasound – physics and equipment. 2nd edition. Cambridge University Press; 2010.

Useful websites

Introduction to US imaging modes: www.criticalecho.com/content/tutorial-2-modes-ultrasound
Image gallery of US scans: www.medison.ru/uzi/eng/all/

14 Endoscopy

14.1 Introduction

Endoscopy is a method that allows examination of hollow (dark) spaces by visible light. The word “endoscopy” is derived from Greek by combining the prefix “endo”, meaning “within”, and the verb “skopein”, standing for “view” or “observe". Although one of the earliest medical imaging systems invented long before x-ray imaging, it played a minor role until recently, when major advances in fiberglass technology, light sources, and CCD sensors have made this technique an indispensable tool in everyday medical practice. Nowadays endoscopes are used in clinics not only for optical inspection of surface structures within hollow spaces in search of cancerous tissues, but also for assisting minimal invasive surgery, for taking biopsy samples, and for in vivo microscopic and spectroscopic investigations. Endoscopy is enormously versatile. Recent breakthroughs in micro-optics, light sources, and light detectors have already revolutionized the technology of endoscopes, and many more inventions can be foreseen. In this chapter we discuss the basic physics of endoscopes and their various advanced developments after a brief overview of their main uses.

14.2 Standard uses of medical endoscopes

Basic versions of medical endoscopes conduct light from an external white light source through a bundle of glass fibers to hollow spaces inside the body; a second bundle of glass fibers receives and transmits the backscattered light from the tissue to a light sensitive detector. Both bundles are wrapped into a flexible tube that is inserted into open ducts of the body. The main features of a standard traditional endoscope are sketched in Fig. 14.1.

Endoscopes, when inserted into the body through open channels (esophagus, rectum, urethra, vagina), provide visual evidence of problem zones, such as ulceration, inflammation, and cancerous tissue. Main applications of endoscopes are gastroscopy through the esophagus and colonoscopy through the rectum. They are sometimes also referred to as upper endoscopy and lower endoscopy, respectively (Fig. 14.2). Endoscopes can also be used for taking tissue samples for further laboratory examination (biopsy), for removing polyps, lumps etc., and for local application of pharmaceuticals. Furthermore, endoscopes support minimal invasive surgery (laparoscopy), such as hand, knee, gall bladder, and many more. For all these additional applications endoscopes are equipped with extra channels to insert and maneuver special instruments like snares and biopsy forceps, as indicated in Fig. 14.1. Other channels supply air or other gases and suck off fluids (blood and various debris).

DOI 10.1515/9783110372830-016

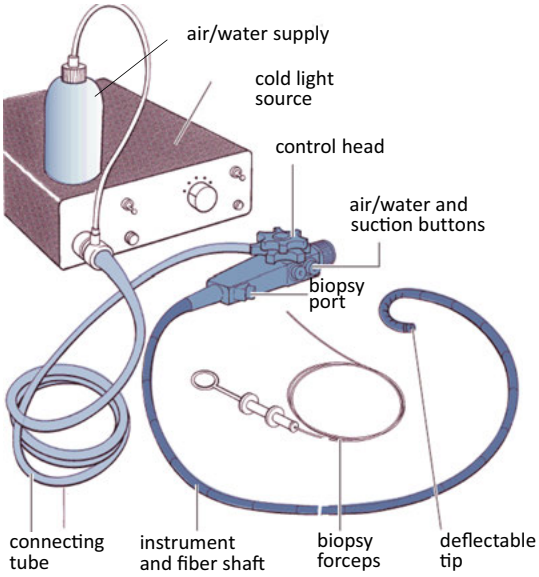


Fig. 14.1: Sketch of a traditional endoscope featuring the main components: light source, control head and tube containing the fiberglass bundles for illumination and imaging, air/water duct, and biopsy channel (adapted from [1] by permission of John Wiley and Sons Inc.).

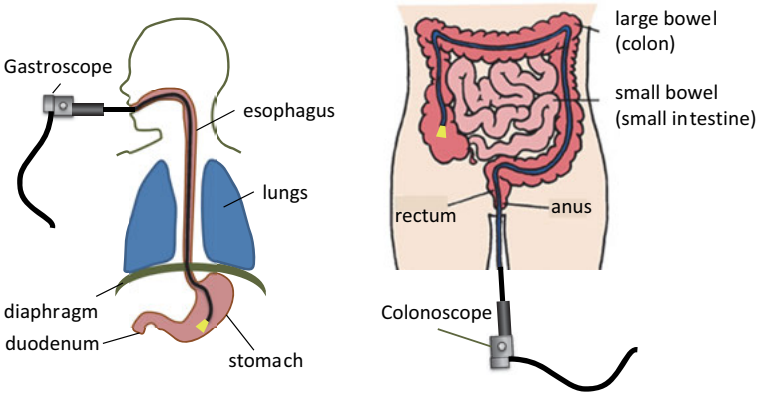


Fig. 14.2: The most common applications of endoscopy are gastroscopy through the esophagus (left panel) and colonoscopy through the rectum (right panel), both are used for early cancer recognition and removal of precancerous tissues.

14.3 Fiber optics

Endoscopes – as we know them today – use a point-by-point imaging method of an object. This is an entirely different optical imaging scheme than the usual pinhole type optics. The endoscopic view is similar to viewing through an array of straight drinking straws. Each individual straw has a very small *field of view* (FOV). But the collection of all overlapping FOVs yields a pixelated picture that is upright and not enlarged, in contrast to pictures from a pinhole camera. Both schemes are compared in Fig. 14.3.

Early endoscopes were straight like straws. But with the introduction of thin *glass fibers*, endoscopes can be bent and light can be made to go around "corners", just like the flexible fiber itself sketched in Fig. 14.4. Assuming that the fibers are well ordered in a coherent fashion instead of being scrambled up, the image corresponds pixel by pixel to the object, where the pixel size is given by the diameter of single glass fibers.

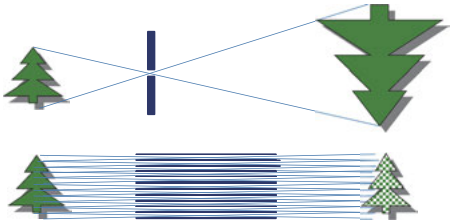


Fig. 14.3: Scheme of pictures generated with a pinhole camera as compared to a collection of long hollow cylinders.

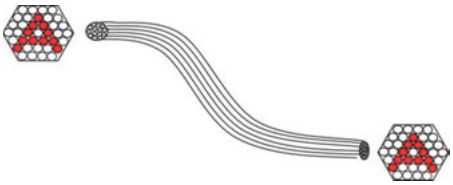


Fig. 14.4: Point to point imaging of objects by a close packed bundle of glass fibers.

Guidance of light through glass fibers is achieved by *total reflection* of light when it passes from transparent matter of higher optical density characterized by the refractive index n_1 to lower optical density, here air, with the refractive index $n_0 = 1$. Figure 14.5 (a) shows the case when light first enters from air into the fiber from the front end at an angle α. Then the light ray is refracted towards the optical axis of the fiber according to Snell's law:

$$n_0 \sin\alpha = n_1 \sin\beta,$$

and reflected at an angle γ at the interface to air. Total reflection occurs for an angle γ_c fulfilling the condition:

$$\cos\beta_c = \sin\gamma_c = n_0/n_1.$$

For all angles $\gamma \geq \gamma_c$, light rays are reflected back from the fiber/air interface to the inside, light rays incident at angles $\gamma \leq \gamma_c$ will pass through the fiber to the surrounding. The FOV for which total reflection occurs is 2α and depends on the refractive index of the fiber according to:

$$2\alpha = 2\sin^{-1}\left(\sqrt{n_1^2 - n_0^2}\right).$$

The aperture of the fiber is defined as:

$$A_{\text{fiber}} = n_0 \sin\alpha = \sqrt{n_1^2 - n_0^2}.$$

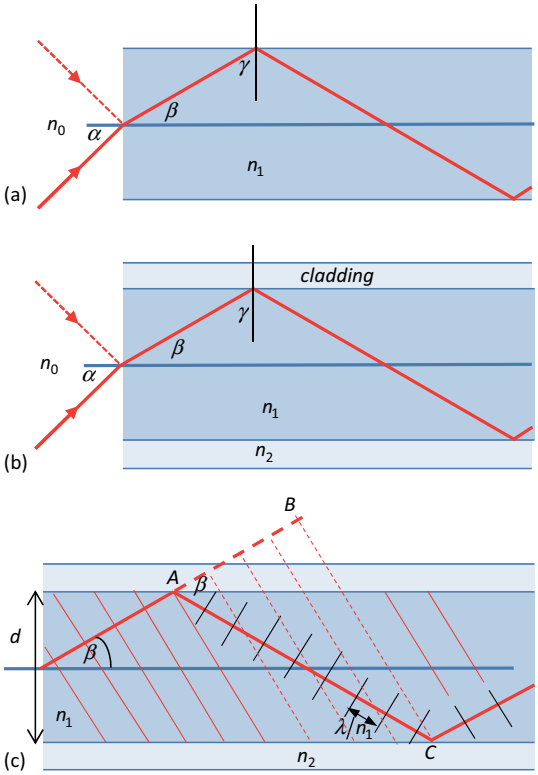


Fig. 14.5: Refraction and reflection in fiberglass. (a) Simple glass fiber without coating; (b) glass fiber with a core and a cladding material; (c) constructive interference in a waveguide.

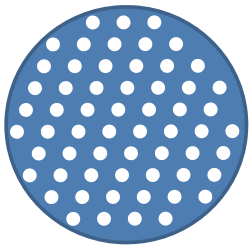


Fig. 14.6: Optical fibers embedded in cladding material.

Fibers are always coated with a cladding material characterized by a refractive index $n_2 < n_1$ (Fig. 14.5 (b)). The cladding has the purpose of suppressing cross communication between fibers. Total reflection then occurs at the core/cladding interface. In endoscopes fibers are in fact embedded in a cladding material shared by all fibers, as indicated in Fig. 14.6. With cladding the condition for total reflection at the internal interface is modified to:

$$\sin \alpha = \sqrt{n_1^2 - n_2^2}.$$

Thus the FOV and the aperture increase with increasing difference of the refractive indices $\Delta = n_1^2 - n_2^2$.

Endoscopes typically have an outside diameter ranging from 0.5 mm for very narrow channels up to 9 mm for wider channels like the esophagus and the colon. As already stated, endoscopes house two bundles of optical fibers: one for illumination at the distal end of an endoscope, and one for guiding the scattered light back to the proximal end of the instrument. Both fiber bundles contain 20 000–40 000 fine glass fibers, each about 5–10 µm thick. Endoscopes have a conflicting design problem to solve. The outside diameter should be small, yet the number of fibers inside should be as large as possible for high resolution images. When the coherence length of light, i.e., the length of the wave train, becomes larger than the diameter of a single fiber, then wave optics needs to be taken into account. After two reflections at the interface, the wavefront of the light has to interfere constructively with the part of the same wave train that has not yet been reflected. In Fig. 14.5 (c) the incoming wave travels the distance AB, while the reflected wave travels the distance AC. After reflection, both waves should match in phase. Therefore the difference in path length $AC - AB$ should be a multiple of the wavelength λ/n_1 in the medium with refractive index n_1. With $AC = d/\sin\beta$ and $AB = AC\cos(2\beta)$ the condition for constructive interference is

$$\sin\beta = \frac{m\lambda}{2n_1 d},$$

where m is the order of interference. Phase jumps at the core/cladding interface do not occur at the boundary to a medium with lower refractive index. In terms of the aperture the condition then reads:

$$n_1 \sin\beta = \frac{m\lambda}{2d} \leq A = \sqrt{n_1^2 - n_2^2}$$

and

$$m \leq \frac{2d}{\lambda}\sqrt{n_1^2 - n_2^2}.$$

The largest number $m_{\max}$ that fulfills the condition of constructive interference before reaching the critical angle for total reflection defines the number of allowed light ray directions. If $m_{\max} = 0$, the waveguide is called a *single mode guide* allowing only the fundamental mode to go through. Single mode and *multimode fibers* are compared in Fig. 14.7. Narrow fibers with high packing density for high resolution images are usually single mode fibers.

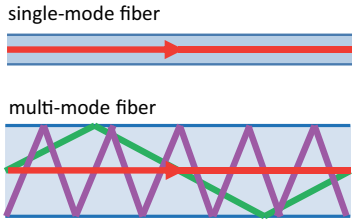


Fig. 14.7: Single and multimode fibers. Note that the colors do not indicate different wavelengths but different directions of the light beam for the same wavelength.

14.4 Endoscope optics

In early endoscopes the backscattered light was directly viewed through an ocular eyepiece by the examiner. However, modern technology has replaced direct viewing by a light sensitive chip, such as a charge coupled device (CCD) or a complementary metal-oxide-semiconductor (CMOS) sensor for higher image quality, schematically shown in Fig. 14.8 (a). The electrical output is fed into a PC for image processing and displayed on an LCD screen. The CCD chip measures only light intensity but is not color sensitive. For color images the sensor is covered by three color filters according to the RGB color code as in digital cameras. A lens in front of the sensor may either focus the light on the chip or magnify the pixel picture at the proximal end of the fiber bundle. Endoscopes designed according to this scheme are known as *fiber optic endoscopes*.

Viewing with a CCD chip has the advantage that more than one person can simultaneously watch images and videos taken during an examination and even remote viewing is possible. Furthermore, the pixel density of a modern CCD chip matches well with the density of fiber bundles, so that there is no loss of information or effect on the high definition of the recorded image.

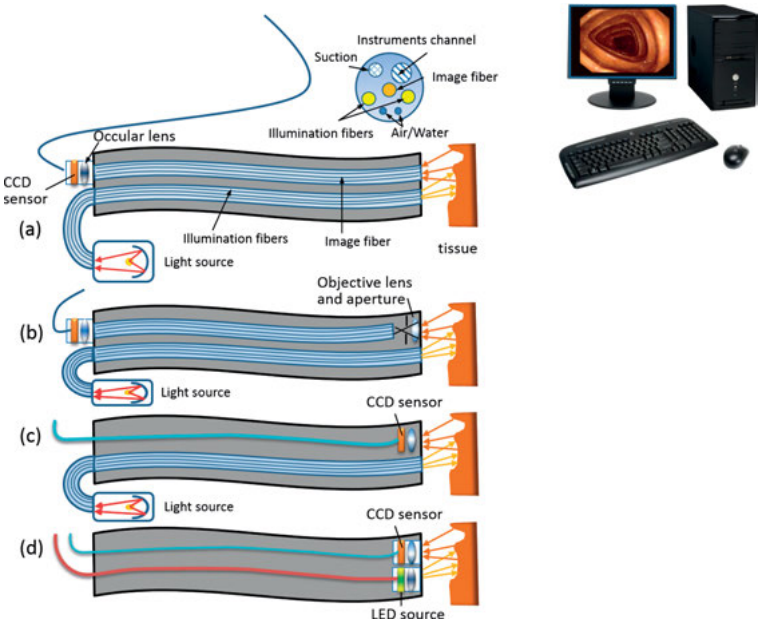


Fig. 14.8: (a) Essential parts of endoscopes consisting of fiberglass bundles for illumination and for transmitting backscattered light. A cross section of the endoscope features channels for surgical instruments and for air/water inlets; (b) lens and aperture at the distal end in front of the fiberglass bundles improves the optics; (c) a CCD chip at the distal end may completely replace the fiber bundle for imaging; (d) an LED light source may replace the illuminating fiber bundle.

While keeping the functionalities described above, the simple optics of a standard endoscope has room for improvements. Although endoscopes do not have an intrinsic focus, modern micro-optics allow the rays on the object side to be collected and focused on the fiber bundle. By inserting an aperture and an objective lens at the distal end, the *focal length* and the *depth of field* can be adjusted (Fig. 14.8 (b)). The depth of field is the distance range from the objective lens to the tissue over which images remain in focus. This is typically 5–100 mm in modern video endoscopes to be presented next.

With miniaturization of CCD chips it became possible to remove the receiving fiberglass bundle entirely and replace it by a CCD sensor at the distal end, while retaining the optics environment (Fig. 14.8 (c)), which is referred to as '*chip in the tip*'. This design, known as *video endoscope*, offers flexibility with respect to positioning the CCD chip either at the distal end (distal tip) or on the side of the tube. It has been shown that imaging with distal sensor endoscopes provides better images with respect to resolution, contrast, and color discrimination than standard ones. Most currently used endoscopes in clinics use this version of video endoscope.

The next step in endoscopic development is the substitution of the illuminating fiberglass bundle by a tiny but very bright solid state light source such as a light emitting diode (LED), indicated in Fig. 14.8 (d). In this design, endoscopes are realized without any fiber bundles and without the need for a very expensive xenon white light source. In the past the development of fiberglass technology has paved the way to the successful use of endoscopes. But simultaneous advances in semiconductor technology and microelectronics have made optical fiber bundles obsolete. Apart from better illumination and higher image quality another important improvement is weight reduction. The lighter endoscopes are, the better the additional instruments can be manipulated. Furthermore, optical fiber bundles can no longer break by overbending. These developments are still in progress.

Although endoscopy is doubtless an enormously successful procedure for visual inspection and diagnoses of early stages of cancer, some complications should be addressed. Endoscopy can only be executed in empty hollow spaces where any liquids including blood have been removed. In the case of minimal invasive surgery with the help of endoscopes this constitutes additional difficulties. Often the body parts to be examined or treated need first to be inflated with air or carbon dioxide. Additional fluids and debris have to be sucked off. Hygiene is another issue of concern. All equipment parts that come into contact with patients must be sterilized. Alternatively, but more costly than reconditioning, are endoscopes with one-time-use components that are delivered and stored sterilized. For further practical procedures of endoscopy we refer to [1].

14.5 Resolution and magnification

Resolution is the ability to distinguish two closely spaced objects. In the diffraction limit, resolution is defined by the Abbé or Rayleigh criterion and is mainly limited by the wavelength used. However for endoscopes, which operate far away from the diffraction limit, the resolution is defined by the ratio of illuminated area $A = \pi r^2 = \pi(d \tan \alpha)^2$ provided by the FOV (see Fig. 14.9) divided by the number N of fibers in the case of fiber endoscopes, or the number of pixels N in the case of video endoscopes:

$$R = \sqrt{\frac{\pi(d \tan \alpha)^2}{N}}.$$

With a working distance of 10 mm and an opening angle $\alpha = 70°$, assuming $N = 10^5$, a resolution of 0.15 mm is achieved. For even higher resolution and lower signal-to-noise ratio (SNR) more costly CMOS sensors are used instead of CCD chips.

Currently, *high definition* (HD) *video endoscopes* with 9 mm outer tube diameter, suitable for gastroscopy and colonoscopy, contain sensors allowing images with 1–2 megapixels to be taken and provide a spatial resolution of about 100 µm at a working distance of about 10 mm [2]. The spatial resolution decreases with increasing working distance.

With higher resolution more details are visible. In contrast, *magnification* enlarges the images without improving the resolution. Standard instruments magnify up to about ×35. High magnification instruments using special optics reach magnifications of about ×150. This high magnification is achieved by using movable lenses in the tip of the endoscope, as indicated in Fig. 14.9. Figure 14.10 shows an example of a high resolution endoscopic image of the colon.

Summarizing, standard but state-of-the art video endoscopes have an FOV of about 140°, a depth of field variable between 5 to 100 mm, spatial resolution of about 0.1 mm at a working distance of 5–10 mm, and in some cases a magnification up to 150. The information depth of the backscattered light spans from the surface to about 5 µm depth into the skin depending on the wavelength used. The physical length is about 1 m for gastroscopy and 1.6 m for colonoscopy. These lengths are sufficient for reaching the stomach on one side and the full length of the colon on the other side, respectively. But it is not sufficient to image the small intestines. For the latter examination different methods are applied, such as the capsule endoscope presented further below.

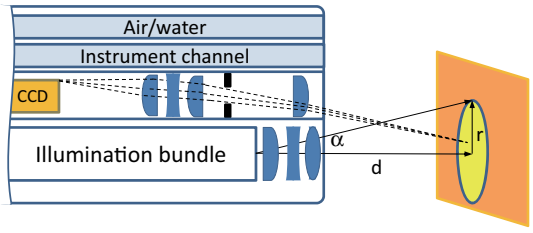


Fig. 14.9: Distal end of a video endoscope. The surface of the illuminated tissue is at a distance d and the field of view has a radius r.

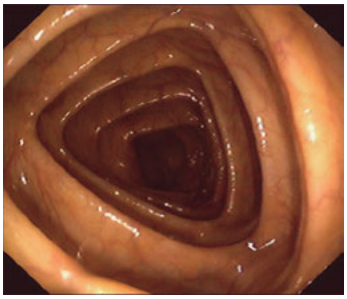


Fig. 14.10: High resolution image of the colon taken with a modern endoscope (reproduced from www.gastrolab.net/ni.htm).

14.6 Specialized endoscopes

Endoscopes in medicine are generally used for optical visualization and diagnostics of inner hollow body parts and for visual support of minimal invasive surgery. A number of endoscopes have been developed for additional special tasks, such as microscopy and spectroscopy and for overcoming some limitations of standard endoscopy. Some of the developments are briefly presented here. Lasers, which are used in most of these special endoscopic applications, are presented in Chapter 13/Vol. 2.

14.6.1 Narrow band imaging

Narrow band imaging (NBI) refers to an endoscopic imaging technique that uses only a narrow band of wavelengths for imaging and testing. Standard endoscopes use the full wavelength band of visible light from 450 nm to 650 nm. The color of an object is then a question of absorption versus scattering. Wavelengths that are scattered mix and yield the specific color of a body. For instance, if the wavelength band for blue ($\approx$ 440–460 nm) and green ($\approx$ 540–560 nm) are absorbed, but the wavelength band for red ($\approx$ 580–600 nm) is scattered, the color of this body part will appear red. However, if a red body is illuminated with blue light, it will appear black. In standard endoscopy blood vessels give images a reddish color since hemoglobin has absorption bands in the blue and green region, but scatters red light. If a filtered light source is used that blocks out the red wavelength band, blood vessels appear dark. This allows concentrating on other surface structures, such as the mucosa in the stomach or colon, which then stand out in greenish color when illuminated with a green filter. A comparison of a normal image taken with a white light source and one with a filter inserted is schematically shown in Fig. 14.11. Detection of fluorescent light is also possible if the tissue has been stained with an appropriate fluorophore. Alternatively, instead of using an optical transmission filter for one particular color, one may read out just one of the three pixel sets of the CCD sensor, or for illumination one may use LEDs with preselected wavelength bands.

Using a narrow color band for illumination has another advantage. The penetration depth of light into tissue depends on the wavelength. Wavelengths in the blue color regime absorb and reflect near the surface, whereas wavelengths in the red color band penetrate deeper into the tissue and reflect from subsurface structures. NBI is an endoscopic option that can be easily added on to any type of endoscope as it requires only the use of one or two color filters.

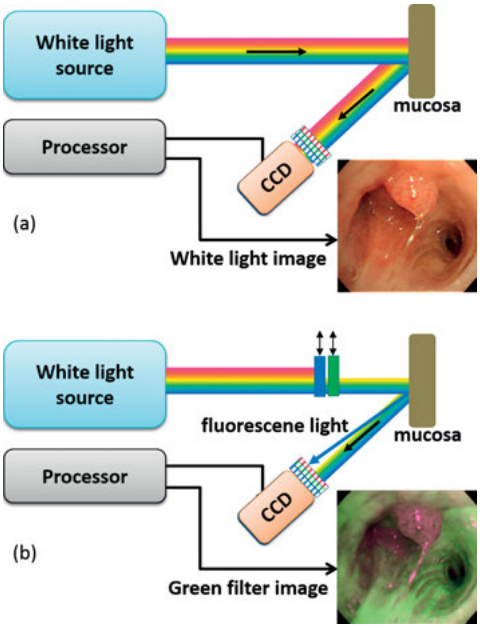


Fig. 14.11: Comparison of endoscopic imaging with white light (a) and with filtered light (b). In both cases the backscattered light is detected by a CCD sensor. When red light is blocked by a green filter, the backscattered light mainly contains information from structures scattering green light. The same also holds for a blue filter. At the same time fluorescent light becomes visible.

14.6.2 Chromoendoscopy

Chromoendoscopy is an alternative endoscopic procedure where dyes or stains are used to enhance contrast. They are infused into the gastrointestinal tract just before inspection with an endoscope. The dye enhances characteristic features of the tissue and facilitates identification of different tissue types or pathologies based on the pattern recognized. Several specific dyes are available for recognition of different diseases.

14.6.3 Endomicroscopy

Endomicroscopy is a technique that not only enlarges the image as the name suggests, but is also used for obtaining histology-like images from inside the human body in real time, a process known as '*optical biopsy*'. In medicine, biopsy implies the removal of some tissue for external histological examination. Endomicroscopy allows histological tests in vivo and in real time. The magnification is achieved either by *confocal microscopy* or *optical coherence tomography* (for explanations see below). Clinical endomicroscopes achieve a resolution in the order of 1 µm which is 100 times higher than for high definition video endoscopes. At this high resolution the field of view is reduced to several hundred micrometers. Endomicroscopy is mainly used for imaging the gastrointestinal tract and in particular for diagnosis and characterization of *Barrett's esophagus* (heartburn) together with other types of precancerous lesions.

One of the characteristics of precancer conditions is an enlarged cell nucleus and an increased size ratio of nucleus to cell body. Therefore the ability to image nuclear cell morphology in vivo represents a major step forward towards detection of epithelial precancers. Epithelia cells cover the inner surface of all tubes and ducts of the body. They are the equivalent of epidermis for inner body surfaces. Early observation of structural changes in the epithelial structure can be treated and can save lives.

14.6.4 Confocal laser endoscopy

Confocal laser endoscopy (CLE) is used in endomicroscopes for achieving very high resolution that allows in vivo histological diagnostics. Conventional microscopes are so called wide field microscopes, i.e., an image of an object is taken at once over the entire field of view. This is the usual procedure for video endoscopes. However, wide field microscopes are generally unsuitable for imaging thick tissue with very high resolution because the images are blurred by out-of-focus background signal. In confocal microscopy this problem is remedied by scanning a fine laser beam point-by-point across the field of view similar to a scanning electron microscope.

A schematic diagram of a fiber optic confocal reflectance microscope (FCRM) is shown in Fig. 14.12 [3]. A fiber optic bundle is located between the scanning mirrors and the objective lens. The illuminating laser beam is coupled into only one fiber at a time on the proximal end of the fiber bundle using an X-Y mirror scanner. On the distal end of the fiber bundle each fiber serves both as a point light source and a detection pinhole. This illuminated fiber is imaged onto the tissue by a miniature objective lens; backscattered light from the tissue is imaged by the objective lens and relayed through the fibers to a beam splitter and photodetector. Backscattered light from out-of-focus regions in the tissue is distributed over multiple fibers and mostly rejected by a pinhole aperture in front of the detector. Surface confocal images are produced by raster scanning the illumination spot across the proximal face of the fiber bundle. The field

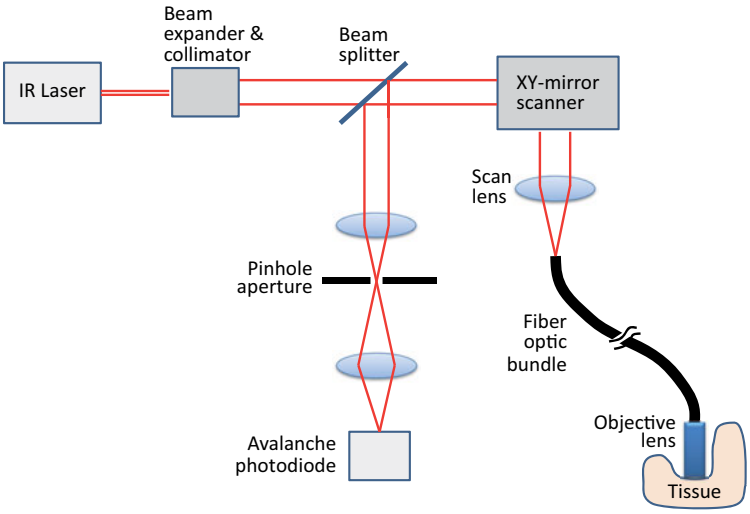


Fig. 14.12: Schematic diagram of the fiber optic confocal reflectance microscope (adapted from [3]).

of view is essentially identical to the scan range of the fiber bundle. Typical values are 150 µm for the lateral scan range at a resolution of 1.5 µm. Confocal microscopes work in fluorescence or in reflectance mode. For fluorescence imaging, the tissue has to be stained with an appropriate fluorophore.

If not used internally in hollow spaces, the fiber's optics is not required and the objective lens of the confocal reflectance microscope (CRM) can be directly focused on the area of interest. This is indeed done in skin oncology [4]. There are two versions of CRM, fixed and movable. In the fixed version the laser beam is scanned over the area of interest with a lateral resolution of about 1 µm and a depth resolution of 3–5 µm to a depth of about 250–300 µm. A sequence of scans in steps of 5 µm allows an in vivo "optical biopsy", providing an analysis of the epidermal structure at nearly histologic resolution including images of melanoma in the skin. The movable CRM version can be used to probe parts of the skin that are otherwise difficult to reach and allows a simple and fast full body skin examination. The handheld device is used in a manner similar to US scanners and the image can be seen live on the screen.

14.6.5 Optical coherence tomography endoscopes

Optical coherence tomography (OCT) endoscopes add the third dimension to confocal microscopy. They have a lateral resolution comparable to CLE and enable cross-sectional depth-resolved views in the third dimension by using interferometry. OCT works as an optical biopsy tool for in vivo microstructural information about tissue in three dimensions (3D).

OCT is an optical analog to B-mode ultrasonographic imaging, which is performed by measuring the echo time delay and intensity of backreflected soundwaves discussed in Section 13.5.2. In OCT the depth of backscattered light is determined by the echo time delay of light and measured interferometrically with the help of a reference beam in two arms of a Michelson interferometer. The basic principle of a Michelson interferometer is shown in Fig. 14.13. In the detector, intensity oscillations are observed with maxima that correspond to the constructive interference of two monochromatic waves, one traveling to the fixed mirror and back while the other one goes to the scanning mirror and back. Constructive interference is observed whenever the total travel distance of these split beams, which are recombined on the return path in the detector, is a multiple of their wavelengths:

$$L_{\text{fixed}} - L_{\text{scan}} = c\tau = m\lambda.$$

Here L_{fixed} and L_{scan} are the respective travel lengths of light, τ is the delay time, and c is the speed of light. In OCT instruments the fixed arm is replaced by the probing arm of the sample to be investigated.

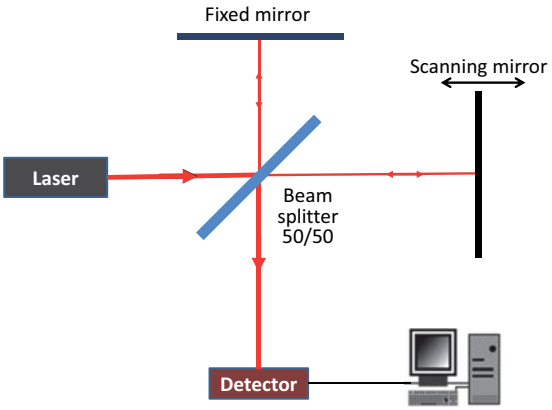


Fig. 14.13: Basic design of a Michelson interferometer.

OCT devices use low power infrared light with a wavelength of 750–1300 nm, which penetrates deeper into the tissue than visible light. The depth of penetration is typically 1–3 mm, depending on tissue structure and depth of focus. The light is either from a laser source with artificially shortened coherence length, or from an LED, also with short coherence length. Long coherence length is good for interferometry but reduces the image quality due to formation of speckle patterns. Therefore a short coherence length is a compromise between the needs of an interferometer and the demands for high image quality. One arm of the interferometer goes into the fiber endoscope for 3D scanning, whereas the other arm is used for determining the temporal delay from which the depth information is gained. A schematic diagram of a typical OCT system is shown in Fig. 14.14. The lateral scanning unit can either be on the proximal or the distal end of the endoscope. Proximal scanning is done with a relay lens across the fiber

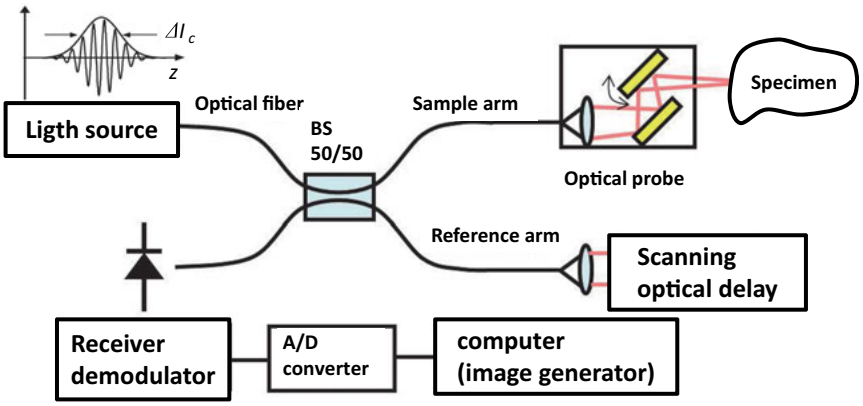


Fig. 14.14: Schematic block diagram of a typical OCT system using light with a short coherence length (adapted from [5]).

bundle, distal scanning is performed by rotation of the laser beam sideways about the optical axis or by moving mirrors for forward scanning. A recent review of the OCT technique and its various clinical uses can be found in [5].

3D-OCT has not only successfully been applied to tissues of inner organs, but also for scanning the outer skin. Another extensive and successful application of OCT is in ophthalmology, where the thickness of the retina and in particular of the macula is determined, including imaging of any lesions.

Tab. 14.1: Comparison of different imaging modalities. Nomenclature: CE = chromoendoscopy, NBI = narrow band imaging, CLE = confocal laser endoscopy, LCI = low coherence interferometry, OCT = optical coherence tomography, US = ultrasound, 3D = three dimensional.

Imaging modality	Depth sensitivity range	Depth resolution	Pros	Cons
CE	N/A	N/A	Enhanced contrast	Superficial imaging Use of dyes
NBI	N/A	N/A	Enhanced contrast – no dye required	Superficial imaging
CLE	< 250 µm	1–5 µm	Cellular resolution	Limited imaging depth Limited FOV
LCI	≈ 250 µm	N/A		Limited detection depth
OCT	3 mm	5–30 µm	3D imaging No contrast agent	No fluorescence imaging 3D not in real time
US	1–10 cm	50 µm	Broad field	Low resolution

In Tab. 14.1 the different imaging methods that have been discussed in this chapter are compared, including their pros and cons. The last line contains a comparison with sonographic techniques.

14.6.6 Capsule endoscopy

Capsule endoscopy is a method to record images of sections of the digestive tract that endoscopes cannot reach. The capsule has the size and shape of a pill and contains a tiny light source, a camera, a video transmitter, and a battery to power the device. After patients have swallowed the capsule, it takes pictures of the inside of the gastrointestinal tract at constant time intervals, typically 2 frames per second, that are received and stored by a recording device worn by patients externally on the body. Patients can pursue their usual activities for the duration of the examination. The capsule moves by normal contraction of surrounding muscles in the intestines. At present, neither the speed nor the orientation of the capsule is controllable from the outside. The average transition time through an empty gastrointestinal tract is about 80 min before capsules are expelled through the anus. The transit time is higher for patients with diabetes. The primary use of capsule endoscopy is to examine the small intestine that cannot be seen by other types of endoscopy. In the future it may be possible to maneuver the capsule externally by adding a small magnet or other robotic features. This would also open the door to a transition from passive video recording to more active but minimally invasive procedures. The capsule may also be loaded with additional sensors for scanning pressure, temperature, and pH values along the gastrointestinal tract. The size of a capsule endoscope and its relation to the small intestines is shown in Fig. 14.15.

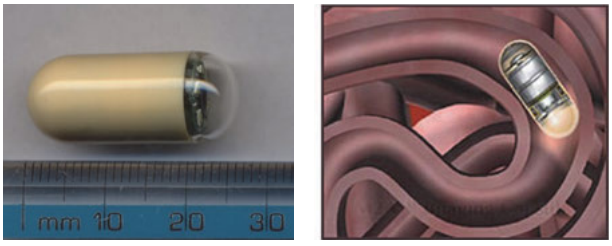


Fig. 14.15: Pill sized capsule for endoscopy of the intestinal tract. The capsule contains a light source and a camera. Recorded images are sent to a storage device worn by the patient. The capsule moves by intestinal muscle activity (reproduced from https://en.wikipedia.org/wiki/Capsule_endoscopy, © Creative Commons).

An interesting new development in capsule endoscopy is to attach a capsule to a thin tether containing optical fibers that connect the tiny light sensors to an imaging console [6]. After swallowing, the capsule makes its way through the esophagus into the stomach, and after taking pictures it can be pulled back out again. This procedure takes about two minutes and is mainly used for examining Barrett's esophagus syndrome, which is a gastrointestinal reflux of acidic fluid from the stomach into the esophagus (heartburn), a major precursor to esophageal cancer. A quickly rotating laser tip emits near-infrared (NIR) light, while tiny sensors record light reflected back from the esophageal lining that is sent back through the fiber bundle to the image processor. Very sharp microscopic images have been recorded, generated by the optical frequency domain imaging (OFDI) method, revealing subsurface structures not easily seen with standard endoscopy [7]. A picture of the tethered and retractable endoscope is reproduced in Fig. 14.16.

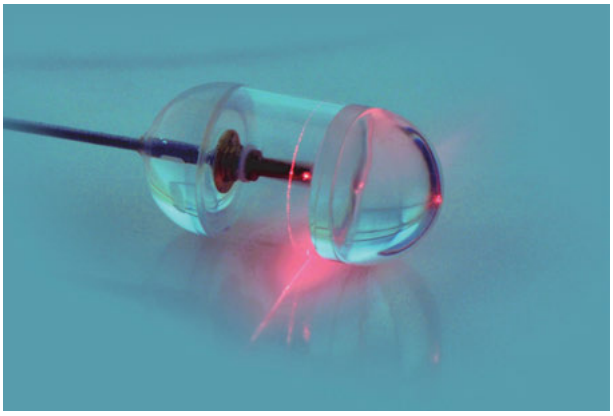


Fig. 14.16: Tethered, one inch long endomicroscopy capsule containing rotating infrared laser and sensors for recording reflected light (courtesy Michalina Gora, PhD, and Guillermo Tearney, MD, PhD, Wellman Center for Photomedicine, Massachusetts General Hospital).

14.7 Future directions

The various types of endoscopy have benefitted strongly from rapid developments in nanotechnology, particularly in the field of optics including lasers and electronics. This trend will most likely continue in the future, although the lateral and in-depth resolution achieved today is already at its optimum for in vivo diagnostics. At present the focus is on high resolution imaging for detecting pathological anomalies. For enhanced diagnostics, spectroscopic capabilities may be added to the arsenal of endoscopic tools. Further advances may be seen in decreasing size (micro-endoscopy), such that ducts not presently accessible can be examined in the future. Miniaturized

capsules may be constructed from biodegradable materials and could be injected into the blood stream for patrolling potential problem areas like arteriosclerosis. In the next step the capsule would be equipped with tools that remove calcifications and provide drugs to lesions. Miniaturized capsules may be equipped with self-propelling or self-steering capabilities, or they may be controlled by an outside console. In general, one may foresee an endoscopic development that carries imaging and diagnostics forward to more spectroscopy and treatment in more narrow and localized areas. Another area of potential development is the combination of different imaging techniques or medical treatments. For instance, an endoscope may contain a transducer for US imaging, or may carry radioisotopes for local exposure. A third emerging field is the combination of endoscopy with surgical robotics, where the demands for high definition and fast imaging are particularly stringent.

14.8 Summary

1. Endoscopy is a method for remote optical inspection of hollow dark spaces in the body.
2. Traditionally endoscopes consist of a flexible tube that contains two bundles of glass fibers, one for illumination and one for guiding scattered light back to the observer.
3. Endoscopes can be diagnostic for observation only, or operative, having channels for irrigation, suction, and the insertion of accessory instruments for surgical procedures.
4. The most common applications of endoscopy are gastroscopy through the esophagus and colonoscopy through the anus. In both cases, endoscopy supports early cancer recognition.
5. Fiber optics is based on the principle of total reflection. For all angles of incidence that are larger than a critical angle of incidence, light is reflected back into the material with the higher refractive index.
6. In modern endoscopes the fiber optic bundle for backreflected light is replaced by a CCD chip at the distal end.
7. Endoscopes have been developed for special tasks beyond the capabilities of standard endoscopes. These include imaging with one color (narrow band imaging), imaging of stained tissue (chromoendoscopy), taking microscopy pictures (endomicroscopy), high resolution confocal imaging (confocal laser endoscopy), or three dimensional volume imaging (optical coherence tomography endoscope).
8. Confocal and OCT endomicroscopy provides morphological information with subcellular resolution for in vivo and real time biopsy.
9. Fluorescence spectroscopy using fluorophores reveals molecular information.
10. NBI visualizes the upper epithelium and mucosa by blocking out scattered light from blood vessels.
11. Chromoendoscopy is sensitive to infused dyes and stains.
12. Capsule endoscopes contain miniaturized optics for illumination and image recording. The capsule is swallowed and images from the gastrointestinal tract are transmitted to an external recording device.
13. One of the main advantages of endoscopy is its versatility.

References

[1] Cotton PB, Williams CB. Practical gastrointestinal endoscopy: The fundamentals. 6th edition. Blackwell Publishing; 2008.
[2] Bhat YM, Dayyeh BK, Chauhan SS, Gottlieb KT, Hwang JH, Komanduri S, Konda V, Lo SK, Manfredi MA, Maple JT, Murad FM, Siddiqui UD, Banerjee S, Wallace MB. High-definition and high-magnification endoscopes. Gastrointest Endosc. 2014; 80: 919–927.
[3] Sung KB, Richards-Kortum R, Follen M, Malpica A, Liang C, Descour MR. Fiber optic confocal reflectance microscopy: a new real-time technique to view nuclear morphology in cervical squamous epithelium in vivo. Optics Express. 2003; 11: 3171.
[4] Meschieri A, Pupelli G, Pellacani G, Rajadhyaksha M, Longo C. Reflectance confocal microscopy: A new tool in skin oncology. Photon Lasers Med. 2013; 2: 277–285.
[5] Tsai TH, Fujimoto JG, Mashimo H. Endoscopic optical coherence tomography for clinical gastroenterology. Diagnostics. 2014; 4: 57–93.
[6] Seibel EJ, Carroll RE, Dominitz JA, Johnston RS, Melville CD, Lee CM, Seitz SM, Kimmey MB. Tethered capsule endoscopy, a low-cost and high-performance alternative technology for the screening of esophageal cancer and Barrett's esophagus. IEEE Trans Biomed Eng. 2008; 55: 1032–1042.
[7] Gora MJ, Sauk JS, Carruth RW, Gallagher KA, Suter MJ, Nishioka NS, Kava LE, Rosenberg M, Bouma BE, Tearney GJ. Tethered capsule endomicroscopy enables less invasive imaging of gastrointestinal tract microstructure. Nature Medicine. 2013; 19: 238–240.

Further reading

Maitland KC, Wang TD. Endoscopy. In: Zouridakis G, Moore JE Jr, Maitland DJ, editors. Biomedical technology and devices. 2nd edition. CRC Press; 2013. Chapter 9.
Choi Y, Yoon C, Kim M, Yang TD, Fang-Yen C, Dasari RR, Lee KJ, Choi W. Scanner-free and wide-field endoscopic imaging by using a single multimode optical fiber. Phys Rev Lett. 2012; 109: 203901.

Useful website

www.endoatlas.com/atlas_1.html

15 Magnetic resonance imaging

15.1 Introduction

When *nuclear magnetic resonance* (NMR) was first observed in atomic beams by Isidor Rabi in 1938 [1] (Nobel Prize 1944), it was a scientific curiosity and a far cry from present day medical imaging applications. The principle of NMR applied to condensed matter was demonstrated later by Purcell and Bloch and published in 1946 [2, 3], both were Nobel Prize winners in Physics 1952. The next breakthrough came in 1973 when Lauterbur and Mansfield demonstrated independently that local resonance conditions set up by a magnetic field gradient can be used for imaging objects in space [4, 5], both Nobel Prize winners in Medicine in 2003. Since these early demonstrations, magnetic resonance imaging (MRI), also called *magnetic resonance tomography* (MRT) has developed immensely. MRI was introduced to clinical practice in 1984 for imaging inner organs and visualizing tumorous tissues. MRI has since progressed into many different sub-branches, such as angio-MRI, multiparameter MRI, functional MRI, gated MRI, diffusion-weighted MRI, or hyperpolarization MRI.

Nuclear magnetic resonance is an experimental method that determines the resonance frequency (energy) of nuclei with finite *nuclear magnetic moment* in a magnetic field. Protons and neutrons are Fermi particles with an intrinsic nuclear spin $S = 1/2$. The magnetic moments associated with their spins are in terms of nuclear magnetons μ_N:

$$\mu_p = 2.793\mu_N,$$
$$\mu_n = -1.913\mu_N,$$

where the nuclear magneton for a Fermion particle is:

$$\mu_N = \frac{e\hbar}{2m_p} = 5.05783 \times 10^{-27}\,\text{J/T}.$$

e is the elementary charge and m_p is the proton mass. The fact that $\mu_p \neq 1\mu_N$ and $\mu_n \neq 0$ shows that the proton is not a simple Fermi particle and that the neutron contains a charge current although it should be neutral. In fact, both particles are composed of quarks and their quark structure explains the unexpected magnetic moments. When protons and neutrons combine in nuclei, spins of neutrons pair up antiparallel and spins of protons do the same. Even number isotopes therefore have zero *angular momentum* and zero magnetic moment. For instance, in ^{12}C there are 6 protons and 6 neutrons and these pair up as in atomic shells: 2 neutrons and 2 protons fit in the $1s_{1/2}$ nuclear shell. The next subshell $1p_{3/2}$ is occupied by 4 neutrons and 4 protons, resulting in a total nuclear angular momentum $I = 0$. Nuclei with total angular momentum $I = 0$ cannot be used for NMR. However, ^{13}C has one unpaired neutron spin which goes into the subshell $1p_{1/2}$ with a total nuclear angular momentum $I = L + S = 1 - 1/2 = 1/2$, where L and S are the orbital and spin angular momentum of the nucleus in units of $\hbar$,

DOI 10.1515/9783110372830-017

respectively. An interesting example is the case of oxygen. ^{16}O has a total nuclear momentum $I = 0$, as expected for an even-even nucleus. Adding one neutron to get the isotope ^{17}O, one would expect $I = 1/2$, but instead it is $I = 5/2$. This is due to the fact that the extra neutron occupies the nuclear subshell $1d_{5/2}$ with $L = 2$ and $S = 1/2$. ^{13}C and ^{17}O are used for MRI of carbon and oxygen perfusion studies in the body. Although the signal is rather small and the spatial resolution is comparatively poor, it is one of the new trends that will be discussed in later sections. Nuclei with total angular momentum $I > 1$ have a nonspherical magnetization distribution contributing a quadrupole moment to the resonance condition. In 98 % of cases MRI is performed on single protons with spin $S = 1/2$ and nuclear magnetic moment $\mu_p = 2.793\mu_n$ with spherical magnetization distribution. Most of our discussion in the next sections is concentrated on proton MRI, more exotic procedures will be discussed at the end. The intention of this chapter is to provide a basic introduction to MRI methods. This imaging technique is intricate with many facets to be considered, not all of which can be discussed here. For a more extensive treatment literature is provided at the end.

15.2 NMR basics

15.2.1 Zeeman splitting

The magnetic moment of a nucleus with total angular moment I is:

$$\vec{\mu}_I = g\frac{e}{2m_p}\vec{I} = \frac{g\mu_N}{\hbar}\vec{I} = \gamma\vec{I}.$$

Here μ_N is the *nuclear magneton* and γ is the *gyromagnetic ratio* that relates angular momentum to magnetic moments, defined by

$$\gamma = \frac{g\mu_N}{\hbar} = 2.675 \times 10^8\,\mathrm{T^{-1}\,s^{-1}}.$$

g is the nuclear g factor. For isolated protons with $I = S = 1/2$ it is $g = 5.585$.

Magnetic moments $\vec{\mu}_I$ interact with the *magnetic induction* $\vec{B}$ according to:

$$E = -\vec{\mu}_I \cdot \vec{B} = -\gamma\vec{I} \cdot \vec{B}.$$

This expression is known as the *Zeeman energy*. The unit of magnetic induction is Tesla or T. The energy is lowered when magnetic moment $\vec{\mu}_I$ and magnetic induction $\vec{B}$ are parallel instead of antiparallel. If the magnetic induction has only one component in the z direction, the Zeeman energy for the z component is accordingly.

$$E_z = -\mu_{I,z}B_z = -\gamma\hbar m_z B_z$$

Here m_z is the z component of the angular momentum $\vec{I}$ with $2I + 1$ degenerate energy eigenstates, which split up in a magnetic field.

In the case of protons with $I = S = 1/2$ having two eigenstates for $m_z = \pm 1/2$, the energy splitting between the eigenstates is (Fig. 15.1):

$$\Delta E = \gamma\hbar\frac{1}{2}B_z - \gamma\hbar\left(-\frac{1}{2}\right)B_z = \gamma\hbar B_z.$$

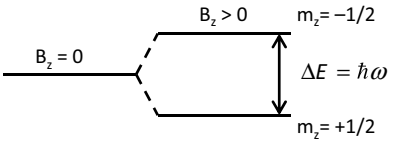


Fig. 15.1: Zeeman splitting of $S = 1/2$ nuclear level in an external magnetic induction B_z.

Expressing the energy splitting in terms of frequencies:

$$\Delta E = \hbar\omega_{\mathrm{L}} = \gamma\hbar B_z$$

we find for the frequency:

$$\omega_{\mathrm{L}} = \gamma B_z.$$

ω_{L} is called the *Larmor frequency* of the proton. For the Larmor frequency of protons we find the relation where the units of B_z are in Tesla:

$$\omega/2\pi = f(\mathrm{MHz}) = 42.58 B_z(\mathrm{MHz}).$$

This frequency is in the range of radio frequencies and sets the stage for MRI experiments, which usually operate in fields of 1 to 7 T corresponding to frequencies from 40–300 MHz. The resonance frequency of 42.58 MHz corresponds to 0.2 μeV on the energy scale or 2 mK on the temperature scale. Gyromagnetic ratio and Larmor frequencies of some isotopes used for MRI are listed in Tab. 15.1.

Tab. 15.1: Gyromagnetic ratios and Larmor frequencies of isotopes used for MRI. Positive signs indicate that nuclear magnetic moment and angular momentum are parallel (proton like), negative signs indicate they are antiparallel (neutron like). The precession is counterclockwise for positive sign and clockwise for negative sign.

Isotope	Gyromagnetic ratio γ_n [10^6 rad s^{-1} T^{-1}]	Larmor frequency f [MHz T^{-1}]
^{1}H	267.513	42.576
^{3}He	−203.789	−32.434
^{13}C	67.262	10.705
^{15}N	−27.116	−4.316
^{17}O	−36.264	−5.772
^{19}F	251.662	40.052
^{31}P	108.291	17.235

15.2.2 Equation of motion

Because of *angular momentum* conservation, changes of angular momentum $\vec{I}$ in space and time is only possible by application of a torque $\vec{T}$:

$$\frac{\mathrm{d}\vec{I}}{\mathrm{d}t} = \vec{T}.$$

Any magnetic induction which is not exactly parallel to the magnetic moment will cause a torque:

$$\vec{T} = \vec{\mu}_I \times \vec{B},$$

such that the magnetic moment will precess about the external field. The *equation of motion* for the angular momentum $\vec{I}$ is then:

$$\frac{\mathrm{d}\vec{I}}{\mathrm{d}t} = \vec{\mu}_I \times \vec{B}$$

or:

$$\frac{\mathrm{d}\vec{\mu}_I}{\mathrm{d}t} = \gamma\vec{\mu}_I \times \vec{B}.$$

The *precession* is counterclockwise for a positive particle like the proton (Fig. 15.2).

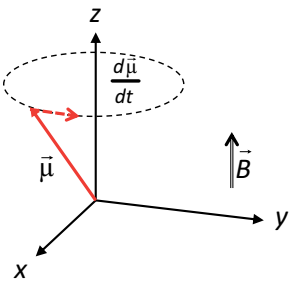


Fig. 15.2: Counterclockwise precessional motion of a nuclear spin about the external magnetic induction B.

In the following we consider an ensemble of N nuclear moments in a volume V, which together yield a *magnetization*:

$$\vec{M} = \frac{1}{V}\sum_{i=1}^{N}\vec{\mu}_i.$$

Therefore the magnetization is the sum vector of all nuclear moments in a unit volume. As these nuclear moments are in random motion, the magnetization will fluctuate over time and without an external field the time average magnetization $M = 0$. Nevertheless we take a statistical view and define the magnetization as the ensemble average over all magnetic moments in a defined volume:

$$\langle\vec{M}\rangle = \frac{1}{V}\sum_{i=1}^{N}\vec{\mu}_i = \frac{N}{V}\langle\vec{\mu}\rangle.$$

The brackets indicate an ensemble average over all configurations in equilibrium at a temperature T. The temperature is defined by contact with a thermal bath of temperature T. In a magnetic field the moments partially line up and yield a finite magnetization $\vec{M}$. Here and in the following we neglect the brackets, since $\vec{M}$ is always understood as a statistical average over individual moments. Now we assume that $B = B_z$, i.e., parallel to the z direction. Then in thermal equilibrium at temperature T the magnetization components are:

$$M_x = 0; \quad M_y = 0; \quad M_z = \chi B_z,$$

where χ is the *nuclear spin paramagnetic susceptibility*. In the simplest form, χ is a scalar and therefore M_z is linearly proportional to B_z.

In a two level system of protons the average magnetization in thermal equilibrium follows from the occupational number N_2 of magnetic moments in the upper energy level, minus the number of moments N_1 in the lower energy level (Fig. 15.3):

$$M_z = \frac{N_2 - N_1}{V}\mu_z$$

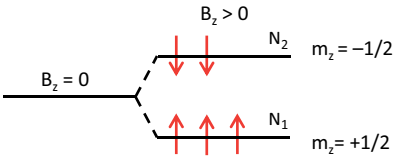


Fig. 15.3: Occupation of energy levels in thermal equilibrium.

In this *two level system* with an energy difference of $\Delta E = \hbar\omega_L = \gamma\hbar B_z$, the ratio of the occupational numbers in thermal equilibrium is thus:

$$\frac{N_2}{N_1} = \exp\left(-\frac{\gamma\hbar B_z}{k_B T}\right),$$

where k_B is the Boltzmann constant. Inserting into the average magnetization yields

$$M_z(T) = \frac{N}{V}\langle\mu_z\rangle \tanh\left(\frac{\mu_z B_z}{k_B T}\right).$$

Using realistic numbers for a 1 T field and a temperature of 300 K, the ratio $\Delta N/N = 6 \times 10^{-6}$. Thus out of 10^6 spins the occupational difference between the upper and lower state will only be 6 spins, a negligible amount! However, in one mole of protons, the occupational difference is 4×10^{18} protons, which becomes detectable as we will see later.

In the case that the magnetization is not in equilibrium, it will change in time proportional to the deviation from equilibrium until it again reaches equilibrium:

$$\frac{dM_z(t)}{dt} = \frac{M_{equil} - M_z}{T_1}.$$

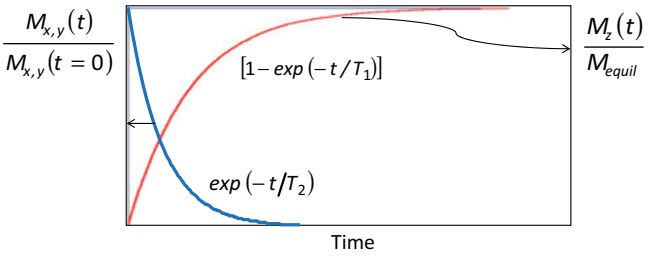


Fig. 15.4: Longitudinal and transverse relaxation.

Time integration of this equation yields an exponential approach to equilibrium according to:

$$M_z(t) = M_{\text{equil}}[1 - \exp(-t/T_1)].$$

T_1 is called the *longitudinal relaxation time*, and M_z is the *longitudinal magnetization*. In MRI literature T_1 is often written as *T1*. Note that T_1 is not a temperature but a characteristic time, while t is the actual laboratory time. Longitudinal means that the modulus of $M_z(t)$, i.e., $|M_z(t)|$ changes with time during relaxation along the z direction, i.e., parallel to the applied magnetic induction. For instance, $M_z(t)$ may grow from zero to a finite value corresponding to thermal equilibrium after switching on a magnetic field. The time dependence is shown in Fig. 15.4.

A simple toy model for the relaxation is shown in Fig. 15.5 together with definitions in panel (a) of magnetic moment components parallel (m_z) and perpendicular to the z axis (m_{xy}). In panels (b) and (c) the protons are omitted for clarity. At the beginning (panel (b)) there are 6 magnetic moments equally distributed over two energy levels, the splitting is taken to be very small but sufficient to define a z axis in space. The magnetization is zero, as all components of the magnetic moments cancel out. When a magnetic induction is turned on, the magnetic moments have to redistribute over these two energy levels and in the end one magnetic moment will turn from the higher energy level (moment antiparallel to field) to the lower level (parallel to field). In our simple example the magnetization M_z versus time after switching on the field would be a step function. But assuming billions of magnetic moments in a field, the average magnetization will change continuously in time until saturation is reached, as seen in Fig. 15.4 and characterized by the relaxation time T_1. $1/T_1$ is the *rate of energy transfer* to the environment. The *transverse magnetization* M_{xy} will remain zero during the entire relaxation process, since the in-plane components of the magnetic moments are randomly distributed and not in phase.

Taking together the time dependence of the magnetization due to precession and due to longitudinal relaxation, the equation of motion with relaxation term is now:

$$\frac{\mathrm{d}M_z}{\mathrm{d}t} = \underbrace{\gamma(\vec{M} \times \vec{B})_z}_{\text{Precession}} + \underbrace{\frac{M_{\text{equil}} - M_z}{T_1}}_{\text{Relaxation}}.$$

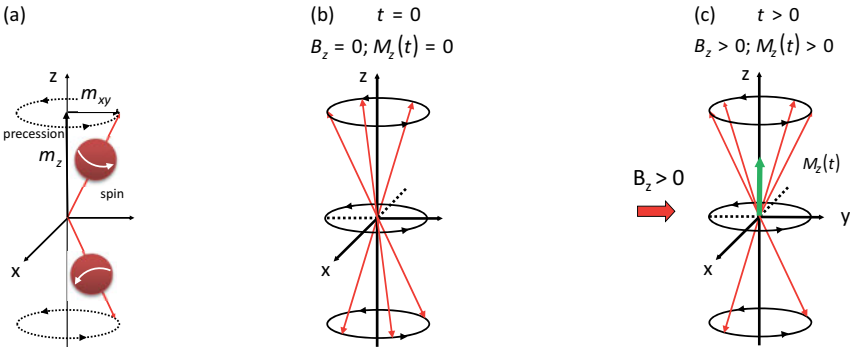


Fig. 15.5: Magnetic moment arrangement before and after turning on a magnetic induction. (a) Definition of spin and precession of individual magnetic moments with projections parallel and perpendicular to the z axis; (b) magnetic moment distribution before turning on a magnetic field; (c) redistribution of magnetic moments in a magnetic field. One moment flips from the higher energy level (antiparallel) to the lower energy level (parallel). The green arrow represents the magnetization M_z.

During longitudinal relaxation of the z component the transverse components of the magnetization $M_x(t)$ and $M_y(t)$ may temporarily differ from zero. In this case the transverse xy components will relax to zero during a *transverse* relaxation time T_2.

$$\frac{\mathrm{d}M_x}{\mathrm{d}t} = \gamma(\vec{M} \times \vec{B})_x - \frac{M_x}{T_2}$$
$$\frac{\mathrm{d}M_y}{\mathrm{d}t} = \gamma(\vec{M} \times \vec{B})_y - \frac{M_y}{T_2}.$$

These three equations of motion are known as the *Bloch equations*. The transverse relaxation time T_2 and the longitudinal relaxation time T_1 are unrelated and may have very different values. Usually $T_2 \ll T_1$. The longitudinal relaxation concerns the modulus of $|M_z(t)|$ and any change indicates flipping of proton spins. T_1 may therefore be considered the *lifetime of antiparallel spins*. It is mainly controlled by *spin-lattice interaction*, i.e., the interaction of local proton spins with the thermal motion of their environment. The transverse relaxation time T_2 relates to a desynchronization of spin components in the xy plane due to stochastic processes and random interactions, also referred to as *spin-spin relaxation*. T_2 cannot be directly measured. But with special techniques described in the following, some phase coherence of spins can be created in the transverse plane by a field pulse. The dephasing time T_2, in MRI literature often written as *T2*, can then be detected by spin-echo techniques. Both relaxation times are illustrated with our toy model in Fig. 15.6. In contrast to the previous case, flipping of the spins has caused a finite in-plane magnetization M_{xy}, which vanishes quickly by dephasing, and only M_z remains. The corresponding relaxation times are plotted in Fig. 15.4.

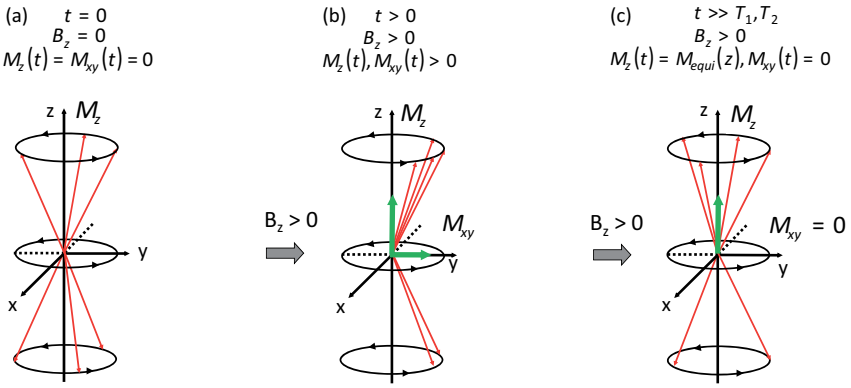


Fig. 15.6: Comparison of T_1 and T_2 relaxation times. (a) Starting situation at $t = 0$ has all moments equally distributed, yielding zero magnetization in the xy plane and in the z direction. (b) A pulse field B_z turns one magnetic moment from down to up and bunches up all moments along the y direction. This causes a finite magnetization $M_z > 0$ and a transverse magnetization $M_{xy} > 0$. (c) While the magnetic field B_z is still turned on, the transverse components quickly dephase within the transverse relaxation time T_2, yielding $M_{xy} = 0$, whereas $M_z > 0$.

15.2.3 Resonance absorption

We are now ready to design a simple nuclear magnetic resonance experiment as illustrated in Fig. 15.7. A magnet generates a magnetic field in the z direction and the energy splitting of protons in the sample increases with increasing field. A *radio frequency* (RF) coil provides an oscillating magnetic field in the x direction with a frequency ω_0 referred to as *RF field*:

$$B_x(t) = B_{x0}\cos(\omega_0 t).$$

The x component exerts a torque that flips the spins from the lower energy level to the upper energy level. This is most efficient if the energy of the oscillating magnetic field matches the energy splitting:

$$\Delta E = \hbar\omega_0 = \hbar\omega_{\mathrm{L}}.$$

Then the RF power is strongly absorbed from the coil into the proton system and the absorption can be measured by a detector that is sensitive to the field amplitude in the RF coil. The NMR experiment can be performed in two different ways: either keeping the RF field frequency constant and sweeping the magnetic induction in the z direction, or vice versa, keeping the field constant and sweeping the frequency. Because of technical reasons the first option is usually realized and schematically shown in Fig. 15.7.

Often in NMR experiments the resonance field $B_x(t)$ is applied in the form of a short pulse. The *field pulse* is chosen such that the magnetization is dipped into the xy plane, corresponding to a 90° rotation about the x axis. M_z is then reduced or may even be zero during the pulse time. The spins are dragged into synchronization within

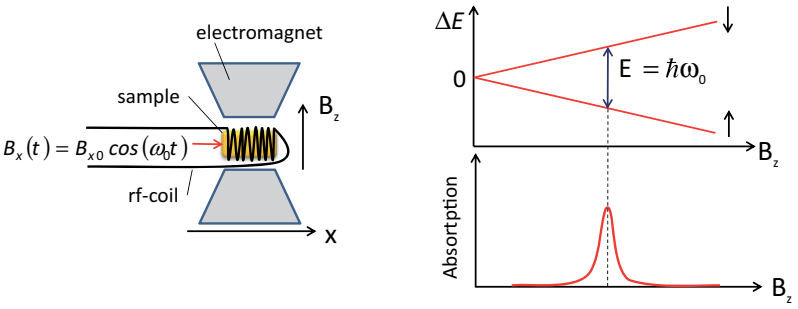


Fig. 15.7: Nuclear magnetic resonance experiment. Resonance absorption occurs at the energy of the RF field that corresponds to the Zeeman energy splitting.

the xy plane and therefore create an M_{xy} component that rotates with the Larmor frequency (see Fig. 15.6). While rotating, the moments produce a flux through the coil which alternates as the spins precess. The resultant induced voltage can be recorded as a damped oscillation, known as *free induction decay* (FID). The corresponding transverse magnetization relaxation is described by (Fig. 15.8):

$$M_{xy} = M_0 \sin(\omega_0 t) \exp(-t/T_2^*).$$

Usually the pulses are repeated many times to collect sufficient signal.

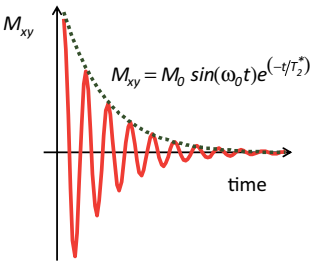


Fig. 15.8: Free induction decay of transverse magnetization after pulse excitation.

The M_{xy} component decays quickly according to the T_2^* relaxation time and the spins will switch back until M_z is fully recovered according to the T_1 relaxation time. The magnetization makes a spiral motion about the z axis where M_z increases continuously while M_{xy} decreases, as illustrated in Fig. 15.9. Both have their own and independent relaxation times. The relaxation time T_2^* is not intrinsic but affected by many external factors, such as field inhomogeneities. Therefore T_2^* is physically speaking not of interest, but very important for recording magnetic resonance images (MRI), as we will see later. T_2 and T_2^* are related through the equation:

$$\frac{2\pi}{T_2^*} = \frac{2\pi}{T_2} + \gamma B_{\text{in}},$$

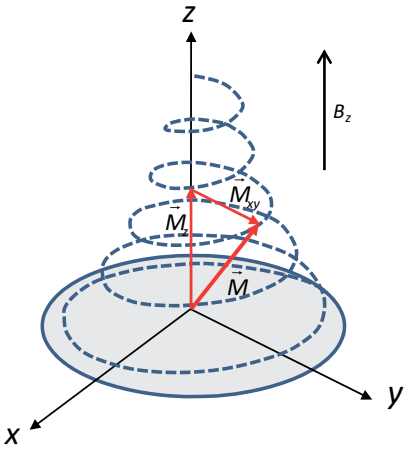


Fig. 15.9: Relaxation of M_z and M_{xy} after a 90° pulse. The relaxation is spiral-like about the z axis. M_{xy} vanishes completely after M_z is fully recovered.

where γB_{in} represents the relaxation by field inhomogeneities, referred to as susceptibility effect. The pulse technique allows determining T_1 indirectly, but is not suitable for measuring T_2. Therefore, another method is required that is sensitive to the intrinsic T_2 relaxation time, and this is discussed in the next section.

It is important to note that T_1 cannot be directly determined from the M_z relaxation as this relaxation does not produce a magnetic resonance (MR) signal. The only MR signal that can be recorded with an induction coil is the FID. Therefore, for measuring T_1 the 90° pulse has to be repeated several times with increasing delay time. The measurement scheme is shown in Fig. 15.10. After starting with a 90° pulse, M_z relaxes up to the delay time t_1. Then another 90° readout pulse is applied and the subsequent FID from the M_{xy} decay is proportional to the amplitude $M_z(t_1)$ at the time t_1. This procedure is repeated after *time of repetition* (TR), starting again with an initial 90° pulse followed by a readout 90° pulse now after delay time $t_2 > t_1$. The recorded FID is then representative for $M_z(t_2)$ at time t_2, etc. After a couple of delay times, the relaxation time T_1 can be extracted.

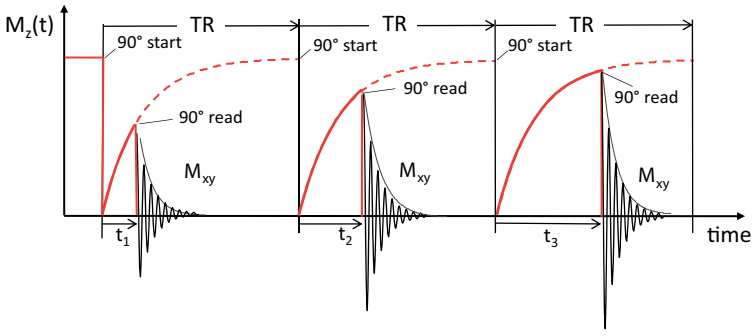


Fig. 15.10: Measurement procedure for determining T_1.

15.2.4 Spin-echo techniques

For the following discussion it is useful to change the coordinate system. So far we have observed the precession in a Cartesian coordinate system (x, y, z) from the outside, referred to as laboratory frame. Now we take a position in the rotating frame of the spins (x', y', z'), assuming that we rotate with the same Larmor frequency ω_L as the spins do. In the rotating frame z and z' are parallel and the xy plane is parallel to the $x'y'$ plane, but the latter rotates with the Larmor frequency ω_L. From this perspective, the magnetization component M_z is at rest and parallel to the z' axis in the rotating frame, and the magnetic induction in the z' direction B_z vanishes (Fig. 15.11 (a)). Now we apply an RF field B_1 parallel to the x' axis in the rotating frame. This field exerts a torque on the magnetic moments and rotates the magnetization M' in the $y'z'$ plane according to the equation of motion:

$$\frac{d\vec{M}'}{dt} = \gamma \vec{M}' \times \vec{B}_1,$$

with the frequency $\omega_1 = \gamma B_1$. The sequence of events is illustrated in Fig. 15.11. At time $t = 0$, $M' = M_z$ (panel (a)). In order to turn the magnetization by a finite angle α, a short pulse is required of duration (panel (b)):

$$\Delta t_{\text{pulse}} = \frac{\alpha}{\gamma B_1}.$$

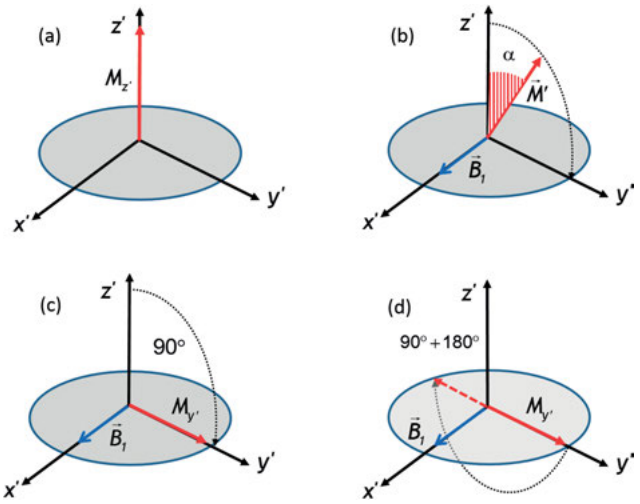


Fig. 15.11: Nuclear magnetic moments in the rotating frame $x'y'z'$. (a) Starting situation with $M' = M_z$ parallel to the z' direction; (b) a 90° magnetic field pulse is applied parallel to the x' direction, turning the magnetization M' into the y' direction; (c) the magnetization now only has a y' component; (d) another field pulse is applied parallel to the x' direction, turning M'_y by 180° to the $-y'$ direction.

We assume that Δt_{pulse} is long enough to cause a 90° *flip* (panel (c)). Following the pulse, an excess magnetization $M_{x'y'}$ will appear in the $x'y'$ plane parallel to the y' axis and will precess with angular frequency $\omega_L = \gamma B_z$ as we have already discussed in Section 15.2.3. However, as we rotate with the same frequency, the magnetization $M_{x'y'}$ will remain constant along the y' axis, neglecting relaxation. Another pulse for a duration $\Delta t_{\text{pulse}}^{180°} = \pi/\gamma B_1$ will make a 180° flip, bringing the magnetization to the $-y'$ axis (panel (d)). This shows that the transverse magnetization can be flipped by successive pulses. We recall that physically only magnetic moments can be flipped and the magnetization is the sum of all moments. The moments can only be flipped between the upper and lower state, but the resultant magnetization may have components parallel to the z' or y' direction. Here and in the following we will only discuss the resultant effective magnetization.

Now we take a step back and reconsider the transverse magnetization $M_{x'y'}$ after the first 90° flip, the sequence of events is sketched in Fig. 15.12.

The first 90° pulse sets the clock to $t = 0$. The transverse magnetization is composed of individual magnetic moments $m_{x'y'}$ all sitting in different environments and all precessing at slightly different frequencies. The ones precessing exactly at the Larmor frequency appear not to move, the faster ones move forward and the slower ones appear to move backwards.

The time period during which the magnetic moments fan out is called the *dephasing time*. It is characterized by a free *induction decay* (FID) with a relaxation time T_2^*, as mentioned before. Now we apply at time $t = t_0$ after the first 90° pulse a second pulse turning the spins by 180°. Since the rotation axis is parallel to the x' axis, the moments at position a switch to position a', and moments at b switch to b'. The same applies to all moments in the fanned out red area (Fig. 15.12 (b, c)). During the time from t_0 to $2t_0$ we watch the flipped magnetic moments moving: the faster ones move clockwise, the slower ones move counterclockwise, i.e., the fanned out triangle closes at $t = 2t_0$, as indicated in Fig. 15.12 (d, e). All stray moments are reassembled (neglecting loss due to other relaxation processes) and give a large $M_{x'y'}$ echo signal. For this reason the 180° pulse is also referred to as the *refocusing pulse*. At times $t > 2t_0$ the moments again dephase. This ingenious spin-echo method was invented by Hahn [6].

Figure 15.13 shows the scheme of a T_2 measurement. The first 90° pulse sets the clock. The $M_{x'y'}$ component rapidly decays from a maximum value $M_{x'y'(0)}$ via FID with the relaxation time T_2^*. After a waiting time t_0 a refocusing 180° pulse is applied. Then an echo signal appears in a pick-up coil after the *spin-echo time* $\text{TE} = 2t_0$ with strength proportional to $M_{x'y'}(2t_0)$. The signal strength of this first echo is a measure of the *proton density* (PD) in the sample. From the ratio $M_{x'y'}(2t_0)/M_{x'y'}(0)$ one could determine the intrinsic relaxation time T_2. However, the initial amplitude $M_{x'y'}(0)$ is not well defined. Therefore, T_2 is generally determined by two different methods. (1) Referring to the top panel of Fig. 15.13: the spin-echo experiment is repeated with increasing waiting times t_0 and accordingly echo time $2t_0$. Then the echo signal will become smaller and smaller with increasing TE, from which T_2 can be determined.

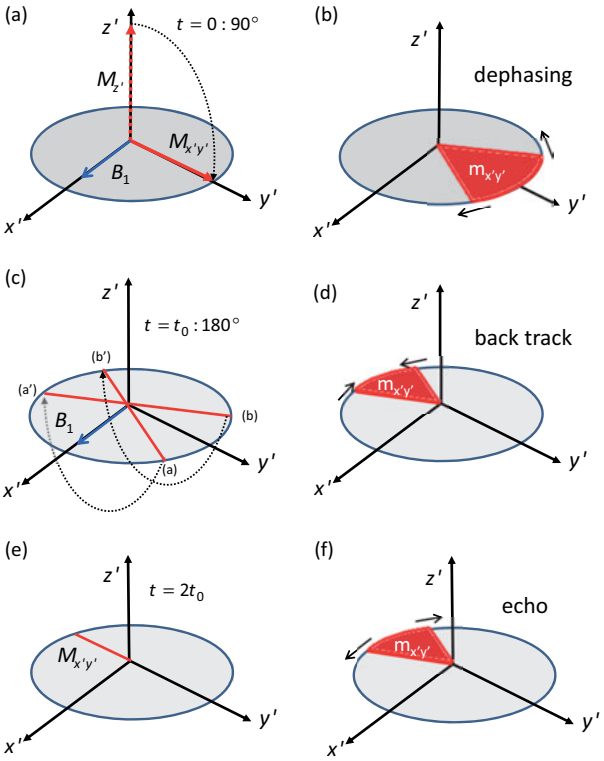


Fig. 15.12: Spin-echo procedure in the rotating frame. (a) Starting at $t = 0$ with a 90° flip of the magnetization M' into the $x'y'$ plane; (b) dephasing of the magnetic moments $m_{x'y'}$ and fanning out. In the laboratory frame the magnetization M_{xy} rotates with the Larmor frequency and is strongly damped with a decay time T_2^*; (c) at $t = t_0$ a 180° pulse turns the magnetic moments about the x' axis; (d) the fanned out magnetic moments $m_{x'y'}$ move back together; (e) the magnetic moments pile up in phase at $t = 2t_0$ and again yield a large $M_{x'y'}$, called an echo signal; (f) dephasing starts again for $t > 2t_0$.

The procedure is similar to the one discussed for determining T_1. (2) Referring to the bottom panel of Fig. 15.13: after the first spin echo at time TE, the 180° refocusing pulse is repeated several times, always resulting in a spin echo with decreasing amplitude. From the decaying amplitudes, T_2 can again be determined. In both cases there should be sufficient time for recovering M_z with relaxation time T_1 before the spin-echo experiment is repeated after TR. The MR *signal strength* S produced by M_{xy} at the maximum of the echo is:

$$S \approx \rho_{\mathrm{H}} \left(1 - \exp\left(-\frac{\mathrm{TR}}{T_1}\right)\right) \exp\left(-\frac{\mathrm{TE}}{T_2}\right).$$

Here ρ_{H} is the proton density (PD) in a particular tissue. S can be an induced voltage in a pick-up coil or any other suitable sensor such as a giant magnetic resistance (GMR)

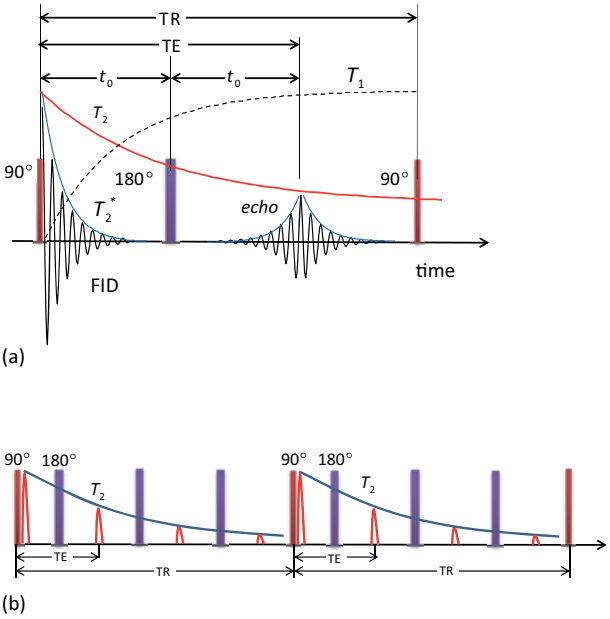


Fig. 15.13: Procedures for determining T_2. (a) The spin-echo time TE $= 2t_0$ can be successively increased for determining the exponential decay of T_2. The dashed line is the M_z recovery curve. (b) After the first 90° pulse, the rephrasing 180° pulse is repeated several times within one TR, and can be repeated again in the next TR time span.

sensor. The pick-up coil is usually identical with the RF coil used as receiver in between two pulses. The first bracket describes the M_z recovery during one TR time, from the second exponential term the transverse and intrinsic relaxation T_2 can be determined.

15.2.5 Autocorrelation and spectral density

Protons in their environment such as in water have their own characteristic wobbling frequency ω_p, which may be a combination of rotational, vibrational, or diffusional motion. Their characteristic relaxation rate $1/\tau_p \approx \omega_p$ depends on whether they are bound in molecules or freely diffusing. This frequency may not be well defined and may span a large frequency range. In any case, energy transfer to the environment is most effective if the Larmor frequency is close to the characteristic frequency of the system: $\omega_L \approx \omega_p$. If the frequencies match, T_1 is short. If there is a large mismatch, energy transfer is inefficient and T_1 becomes large.

In order to get an understanding for different relaxation times of protons in various environments we need to know their frequency distribution. This information is gained by considering the *spectral density* $J(\omega)$, i.e., the weighted intensity of a particular motion with an associated frequency.

As an example we consider the random diffusion of particles in liquids or gases referred to as *Brownian motion*, Fig. 15.14 illustrates a random walk of a particle in a liquid. At time t the particle is at $x(t)$ and at a later time $t + \tau$ it is at another position $x'(t + \tau)$. According to Einstein the *mean square displacement* of a randomly diffusing particle during the time increment τ is:

$$\langle \Delta x \rangle^2 = \langle x'(t + \tau) - x(t) \rangle^2 = 6D\tau,$$

where D is the diffusion coefficient.

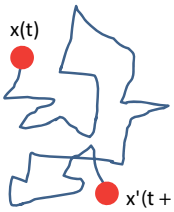


Fig. 15.14: Brownian motion of a particle.

For the diffusive motion we define an *autocorrelation function* that relates the spatial coordinates of the same particle at times t and $t + \tau$:

$$R(\tau) = \lim_{\tilde{t} \to \infty} \frac{1}{2\tilde{t}} \int_{-\tilde{t}}^{+\tilde{t}} x'(t + \tau) x(t) \, \mathrm{d}t.$$

Here the integral is taken over all times and is properly normalized. The autocorrelation function (also known as self-correction function) in liquids shows an exponential decay determined by the diffusivity of particles, expressed in terms of a *diffusional correlation time* τ_c:

$$R(\tau) = R_0 \exp\left(-\frac{\tau}{\tau_c}\right).$$

For times $\tau < \tau_c$ partial correlation exists, for $\tau > \tau_c$ correlation is lost.

Next we take the Fourier transform of the autocorrelation function $R(\tau)$ by integrating over all possible relaxation times τ, which carries relaxation processes from time space over into frequency space. This yields the *spectral density*, defined by:

$$J(\omega) = \int_{-\infty}^{\infty} R(\tau) \exp(\mathrm{i}\omega t) \, \mathrm{d}\tau.$$

For diffusing particles the spectral density has a Lorentzian form:

$$\begin{aligned} J(\omega) &= R_0 \int_{-\infty}^{\infty} \mathrm{e}^{-\tau/\tau_c} \mathrm{e}^{\mathrm{i}\omega t} \, \mathrm{d}\tau \\ &\approx R_0 \frac{\omega}{\omega^2 + (1/\tau_c)^2}. \end{aligned}$$

If there is a broad distribution of correlation times, then the spectral density has a broad plateau as a function of frequency and a cut-off for frequencies that are much higher than any internal characteristic relaxation times: $\omega \gg \tau_c^{-1}$ or $\omega\tau_c \gg 1$.

The spectral density for three representative systems is shown qualitatively in Fig. 15.15 for solids, viscous liquids, and watery liquids. Note that the area under each curve is identical. This is required by the condition that at the same temperature each system has the same total kinetic energy distributed over various excitations.

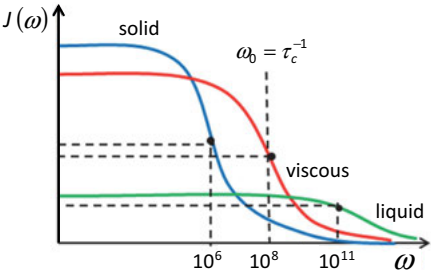


Fig. 15.15: Spectral density of different systems as a function of frequency.

In solids the diffusion is very slow, therefore the spectral density is shifted to small frequencies, corresponding to large correlation times τ_c. In liquids the diffusion is distributed over a wide range up to high frequencies. Viscose liquids have a spectral density between solids and liquids.

For the viscous curve a correlation time is selected to maximize $J(\omega)$ at $\omega_0 = \tau_c^{-1}$. Any curve with a larger or smaller τ_c has a smaller $J(\omega)$ at the same frequency ω_0. If we choose ω_0 to be the Larmor frequency ω_L, the magnitude of $J(\omega_0)$ will be proportional to the relaxation efficiency: the greater the intensity of the fluctuating moments, the more effective the relaxation is, and the shorter the relaxation time T_1 is. In the example shown in Fig. 15.15 this is the case for the viscous fluid, i.e., for this component the probability is highest to absorb the energy from the proton spins and thus the relaxation time for the viscous fluid is shortest. At the same frequency the solid has barely any spectral density as $\omega_0\tau_c \gg 1$, and the liquid is in the plateau region. In the plateau region $\omega_0\tau_c \ll 1$ and therefore absorption of proton spin energy during relaxation is unlikely. Maximized $J(\omega)$ and short relaxation times are valid for one particular Larmor frequency and corresponding field B. When changing the field, the relaxation times will also change, and concurrently the contrast between different tissues, as we will see later.

In a mixed system consisting of solid, viscous fluid and water, it will be the viscous fluid that shows the shortest T_1. In Fig. 15.16 the relaxation times for T_1 and T_2 are plotted as a function of the diffusivity of the system. The dashed lines correspond to the optimal frequencies indicated by solid dots in Fig. 15.15. In the blue shaded area the relaxation times are those which can be found in tissues of the body. Typical relaxation times for different tissues are listed in Tab. 15.2. These are important for generating

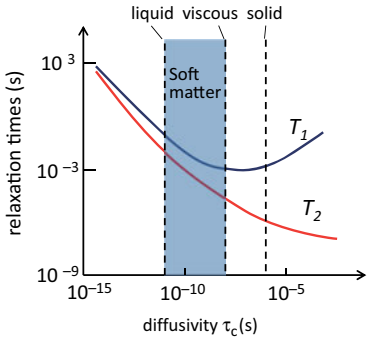


Fig. 15.16: Relaxation times T_1 and T_2 in relation to diffusivities in different materials. Relaxation times may span many orders of magnitude.

Tab. 15.2: Relaxation times for different tissues listed for a magnetic induction field of 1 T.

Tissue	T_1 [ms]	T_2 [ms]
Fat	210	80
Liver	400	40
Kidney	550	60
Muscle	750	45
White matter	650	90
Gray matter	800	100
CSF	2000	150
Bulk water	3000	3000

CSF = cerebrospinal fluid

contrast in MRI. In general, $T_1 > T_2$ and furthermore $T_2 > T_2^*$ for all tissues. T_2^* is not listed in Tab. 15.2 as it is an extrinsic parameter and may vary depending on machine settings.

According to Tab. 15.2 bulk water has the longest relaxation time because of a mismatch between ω_L and ω_0. However, when protons are attached to proteins, their mobility and diffusivity is much reduced, relaxation becomes more effective and the relaxation time T_1 shortens.

15.2.6 Final notes

In NMR experiments two more issues need attention.

(a) Amplitude

The amplitude of the RF field should not be too high. Otherwise the absorption saturates as soon as the number of spins in the upper level and in the lower level are equal. In a saturated state (red line in Fig. 15.17 (a)) contrast is lost, while the signal strength indicated by the blue line still has variability.

(b) Chemical shift

Protons sit in a chemical environment with a *diamagnetic screening*. Therefore the effective magnetic induction at the location of protons may be different from protons in vacuum:

$$B_{\text{eff}} = B_0(1 - \mu_r),$$

where μ_r is the *relative magnetic permeability*. Due to diamagnetic screening, protons at different sites in molecules have different resonance fields and can be distinguished according to their chemical environment (Fig. 15.17 (b)). This effect is known as chemical shift. With increasing field the chemical shift becomes more noticeable and may cause artefacts and blurring of MR images. For instance, protons in water and in fat have slightly different frequencies. When adding field gradients for spatial encoding of protons, chemical shifts can interfere with frequency encoding gradients. If, for example, fatty tissues and watery tissues are spatially located behind each other, they may appear laterally shifted. Conversely, the chemical shift can be used for spectroscopy purposes, which is indeed used in spectroscopic MRI.

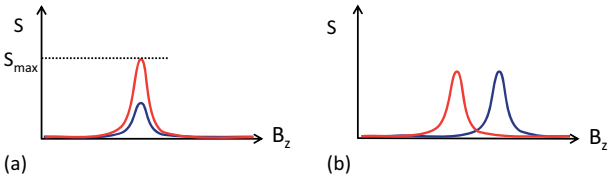


Fig. 15.17: (a) Saturation effect in case of too high RF power. S refers to signal strength due to absorption of the microwave; (b) chemical shift of resonance fields depending on the chemical environment and magnetic field strength.

15.3 Acquisition parameters and contrast

MRI offers a number of schemes for generating contrast between different tissues that need to be tested and optimized. The main contrast "tools" are differences in proton density (PD) and differences in relaxation times T_1 and T_2. In this section contrast schemes between tissues are discussed without considering their location. MRI procedures use a certain amount of jargon; a brief summary is provided here and used in later sections. Moreover, as we approach the core of MRI methods, we will switch over to the nomenclature that is used in standard MRI literature: $T_1 \rightarrow$ T1 and $T_2 \rightarrow$ T2.

15.3.1 Standard terms

TR: the acquisition of data requires repetition of a well defined course of actions within a time to repeat (TR), starting at an initializing B_1 field pulse turning the magnetization by 90°, followed by additional 90° and/or 180° pulses. During TR, T2 decay and full or partial T1 recovery takes place. Therefore TR is longer than pulse times but can be shorter than relaxation times.

SE: spin-echo technique is a nuclear resonance procedure, which uses a sequence of 90° and 180° B_1 field pulses. The first pulse at time $t = 0$ converts M_z into M_{xy}. The second pulse after time $t = t_0$ inverts all moments m_{xy} in the transverse plane such that dephasing of magnetic moments is reversed and a full M_{xy} amplitude is recovered at the SE time $t = 2t_0$. Here and in the following we neglect the prime referring to the rotating frame.

TE: the time to echo (TE) is the time t_0 between the first initializing B_1 90° field pulse and the 180° refocusing pulse, plus the time t_0 from the inversion pulse to the echo. In total, TE = $2t_0$.

MR: the magnetic resonance (MR) signal detected in a pick-up coil is solely produced by M_{xy} dephasing and decay by FID. M_z recovery does not produce a MR signal. In MRI maps MR signal strength of a voxel is usually encoded in terms of pixel brightness.

M_z: this is the magnetization in the z direction proportional to the magnetic induction (applied magnetic field). In a two level system like the one for protons, the magnetization follows from the difference of the occupational numbers of the lower and upper energy states. In the lower energy state, the magnetic moments are parallel to the external field and antiparallel in the upper energy state. The energy splitting is linearly proportional to the applied magnetic field. At constant temperature, the magnetization increases with increasing field because more moments go into the lower energy state.

M_{xy}: this is the transverse magnetization, which is obtained by rotating M_z via a 90° pulse into the transverse plane. M_{xy} is in an indifferent state and decays quickly by relaxing back to M_z through a precessing spiral motion. But even if M_{xy} could be kept in the transverse plane, it would decay by dephasing of the magnetic moments m_{xy}, which sum up to the magnetization M_{xy}. The in-plane relaxation from a maximum value immediately after the 90° pulse to zero is characterized by the relaxation time T2*.

FID: free induction decay is the oscillatory signal recorded by a pick-up coil that originates from precessing M_{xy} in the transverse plane. The precession is damped by dephasing with a relaxation time T2*.

PD: proton density is the number of protons (or hydrogen atoms) per unit volume. This greatly depends on the local tissue of interest.

T1: the relaxation time T1 is the (1 − 1/e) recovering time of M_z, i.e., at t = T1, 63.2 % of the M_z magnetization has been recovered. T1 values range from 200 ms (fat) up to 2000 ms (CSF) and are an indication of how effective energy can be dissipated from protons to surrounding tissue at the resonance frequency. T1-weighted images emphasize differences in T1 in different tissues and de-emphasize differences due to T2. T1-weighted images are best for illustrating anatomic details.

T2: the relaxation time T2 is the 1/e decay time of the transverse magnetization M_{xy}, i.e., after T2, 36.7 % of the original M_{xy} remains, after 2 × T2 13.5 % remains, etc. T2 is the intrinsic transverse relaxation time due to spin-spin interaction, in contrast to T2*, which is related to extrinsic dephasing sources. In general T2 > T2*. T2-weighted images emphasize differences in T2 in different tissues and de-emphasize differences due to T1. T2-weighted images highlight alterations in water content, which appears black in T1-weighted images.

15.3.2 Contrast generation

Having clarified once more the standard terms, we will now discuss some techniques for contrast generation that are available for distinguishing different tissues in the body according to their relaxation times T1 and T2 and proton densities PD. In general the tissue with the higher $M_z(t_0)$ value at time t_0 just before turning on a 90° pulse has a higher brightness on MR images.

(a) T1 contrast

Using SE techniques, T1-weighted images are obtained by de-emphasizing T2 in order to maximize contrast between tissues of different T1. An example is shown in Fig. 15.18. The top panel shows the M_z relaxation to almost saturation after about 2000 ms. In the relaxed state there is no contrast between different tissues. The highest contrast is seen in the graph after about 500 ms. Therefore T1 contrast is achieved by using a fairly short M_z recovery time TR, which is in the order of the average $\langle T1 \rangle \approx 300–800$ ms for the tissue of interest. The middle panel shows the sequence of pulses: after 500 ms a 90° pulse warrants maximum contrast, followed by a 180° refocusing pulse after very short time t_0 of about 15 ms, resulting in a TE of about 30 ms. During such a short TE, M_{xy} contrast is weak. Ideally M_{xy} contrast is lost at the crossing point of both T2 curves. In any case, the MR signal detected is then proportional to $M_{xy}(t = \text{TR}) \approx M_z(t = \text{TR})$. During the time TE, relaxation of M_z has already started and therefore the recovering time of M_z is simultaneously the time to repeat TR. The bottom panel of Fig. 15.18 shows a sequence of M_z recovering curves after each 90° pulse and M_{xy} relaxations.

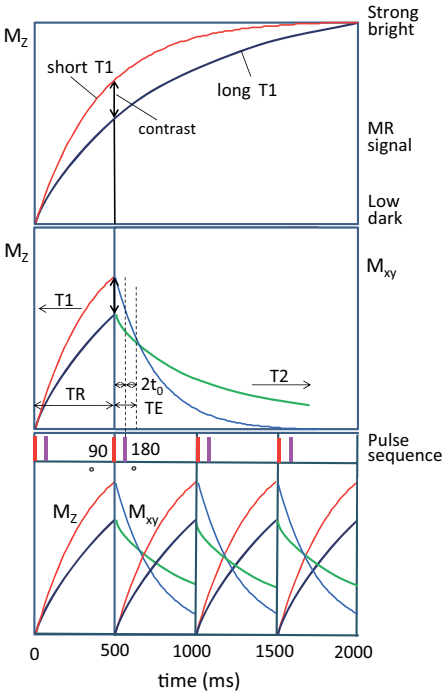


Fig. 15.18: T1-weighting. Top panel: different T1 relaxation times for different tissues. Middle panel: Short TE for eliminating T2 contrast but maintaining T1 contrast. Bottom panel: Pulse sequence and sequence of M_z recovery with simultaneous M_{xy} decay.

(b) T2 contrast

Assuming that there is little difference in T1 for different tissues but sufficient difference in T2, contrast can be achieved by using a long TR and a TE that is optimized for T2 contrast, usually 90–140 ms (Fig. 15.19). After a long TR (≈ 1000–2000 ms), differences between T1 become less pronounced. Differences in T2-weighted images are mainly due to proton density and T2 differences.

(c) PD contrast

In order to achieve contrast based on proton density, T1 and T2 contrast is eliminated. T1 contrast can be reduced by choosing a TR in the order of 2–3 times T1. Similarly, T2 contrast is lost by selecting a very short TE lower than typical T2 values. Then the final contrast is just due to differences in PD (Fig. 15.20). Under these conditions, fat, CSF, and lipids with a high proton content appear bright.

These three different weighting schemes are summarized in Tab. 15.3. Figure 15.21 shows MRI cross sections of the brain using these weighing schemes discussed before. Changes in the contrast of the white and gray matter and in particular for the CSF are very pronounced.

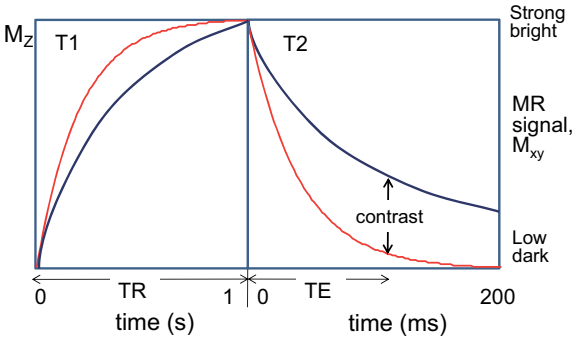


Fig. 15.19: T2 contrast is achieved by using a long TR and TE optimized for T2 contrast. Note the different time scales for T1 and T2.

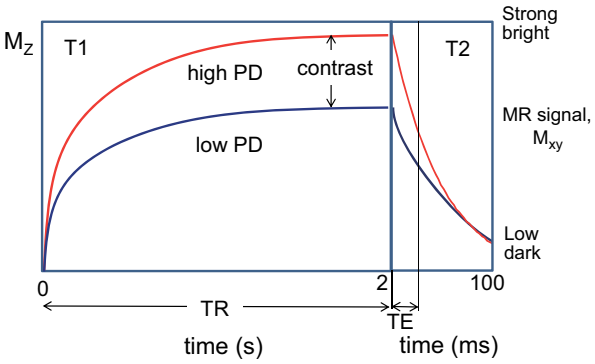


Fig. 15.20: Proton density-weighting by using long TR and short TE. Note the different time scales for T1 and T2.

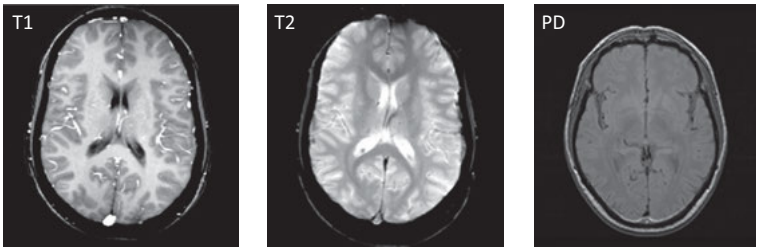


Fig. 15.21: MRI cross sections of the brain with different weighting factors. T1-weighting: TR = 500 ms; TE = 20 ms, shorter T1 areas appear brighter. T2-weighting: TR = 6000 ms; TE = 70 ms, longer T2 areas appear brighter. PD-weighting (TR = 2600 ms; TE = 20 ms): higher PD areas appear brighter. Note that CSF appears black in T1-weighting since it has a longer T1 than white or gray matter and therefore it has a lower M_z within the TR of 500 ms. In T2-weighting CSF appears bright because it has a longer T2 than the surrounding white and gray matter (reproduced from https://openi.nlm.nih.gov/).

Tab. 15.3: Summary of TR and TE times for T1-, T2-, and PD-weighted contrast.

Weighting	TR [ms]	TE [ms]
T1 contrast Shorter T1 appears brighter	short 400–600	short 5–30
T2 contrast Longer T2 appears brighter	long 2000–6000	long 60–150
PD contrast Higher PD appears brighter	long 2000–6000	short 5–30

Apart from the standard weighting procedures T1, T2, and PD, many more imaging modalities with special pulse sequences are available, but they are less frequently applied. Two of these, labeled IR and STIR will be briefly described.

(d) Inversion recovery (IR)

Inversion recovery is a pulse sequence that separates two different T1 relaxation times, in case the usual T1-weighting is not sufficient. The procedure is illustrated in Fig. 15.22. First a 180° pulse is applied at time $t = 0$, turning $+M_z$ to $-M_z$. Both systems relax back to $+M_z$ passing through $M_z = 0$ at different time spans t_i. The relaxation of $M_z(t)$ has the form:

$$M_z(t) = M_{z,\text{sat}}[1 - 2\exp(-t/\text{T1})],$$

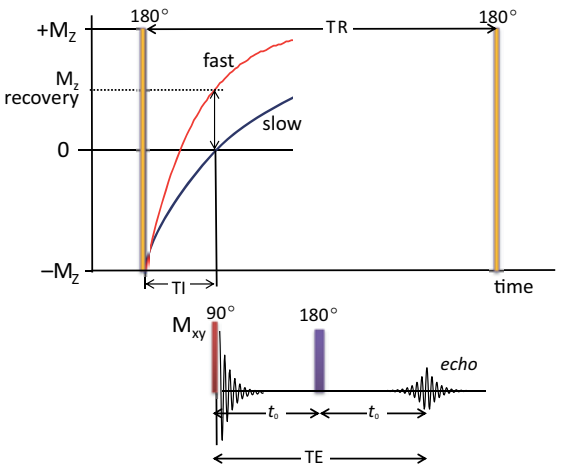


Fig. 15.22: Time and pulse sequence for inversion recovery technique.

such that immediately after the 180° pulse $M_z(t = 0) = -M_z$. For $M_z(t = t_i) = 0$ we find the condition: $t_i = \mathrm{T1} \ln 2$. The time t_i for the slower system with the longer T1 is called the *time of inversion* TI. Therefore, if at TI a 90° pulse is applied, only those spins will return from recovered $+M_z$ to the xy plane that have already passed through the $M_z(t_0) = 0$ point and that have attained a $+M_z > 0$. This is the faster relaxing system with the shorter T1, which can be flipped back from recovered $+M_z$ into the transverse plane, generating an M_{xy} signal. With the M_{xy} component of the faster system the usual SE procedure is followed, i.e., a 180° pulse is applied at time t_0 after TI, inverting M_{xy} to $-M_{xy}$, followed by spin echo at $\mathrm{TE} = 2t_0$, which can be detected in the usual way. The repeat time TR > TI + TE. The IR sequence is also referred to as 180°-90°-180°-sequence.

(e) Short TI inversion recovery (STIR)

STIR is a variation of the IR sequence. IR emphasizes the faster relaxing system. In contrast, STIR emphasizes the slower relaxing system. This is demonstrated in Fig. 15.23. Here at t_i for the fast system $M_z(t_i) = 0$ and the slower system still has some negative magnetization. This partially recovered negative magnetization is then turned by a −90° pulse into the xy plane and usual SE procedures start again, now on the slower system. STIR procedure is used to cancel out, for instance, fast relaxing fatty tissue so that slower relaxing parts nearby become better visible.

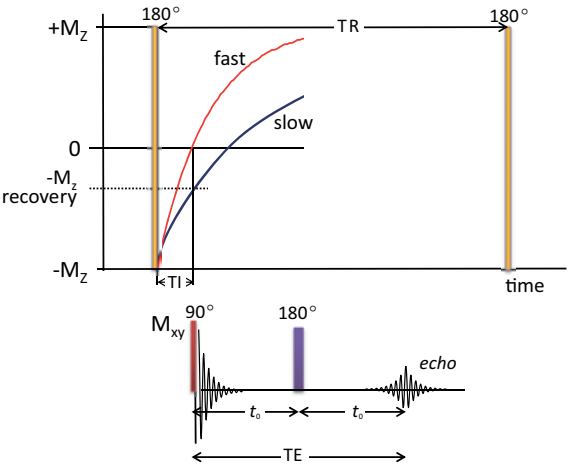


Fig. 15.23: Time and pulse sequence for the short time inversion recovery (STIR) pulse structure.

15.4 MR signal localization

So far we have discussed various signal generating procedures, but we have neglected the location of such signals in the body. For imaging, 3D localization of individual voxels from where the signals originate is essential. This can be realized in three steps, as illustrated in Fig. 15.24: (1) Slice selection in the Z direction (usually head-to-toe); (2) column selection within the slice in the X direction (usually left-to-right); (3) point selection within the column in the Y direction (usually back-to-front). For the coordinate system of slices, columns, and voxels we use capital letters in order to distinguish the spatial coordinates of the MRI machine and the patient's body from the coordinate system of the nuclear spins. However, the main axial field direction in MRI machines coincides with the z direction defined before, the XY plane of a slice coincides with the xy and $x'y'$ planes in previous sections.

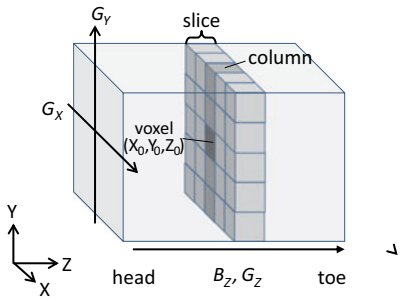


Fig. 15.24: Selection of a voxel within a column of a slice. G_X, G_Y, G_Z refer to field gradients in the respective directions.

15.4.1 Slice encoding gradient

The slice selection is performed as follows. First, a constant magnetic induction B_Z is applied in the horizontal Z direction by a large DC solenoid that fits a patient (see Fig. 15.30). The technical realization is discussed in the next section. Superimposed on the magnetic induction B_Z is a gradient field $G_Z = \mathrm{d}B_Z/\mathrm{d}Z$. With the gradient field each position along the Z direction has a slightly different resonance frequency according to:

$$\omega_\mathrm{L}(Z) = \gamma(B_Z + G_Z Z).$$

Thus the Larmor frequency depends on the position in the Z direction, like the keyboard of a piano. A narrow band RF transmitter is tuned to generate RF pulses (90°, 180°, etc.) at the local resonance frequency $\omega_\mathrm{L}(Z_0)$ for a specific location Z_0 along the field gradient for just a small frequency range Δf, such that the protons are only excited in a thin slice ΔZ, as illustrated in Fig. 15.25.

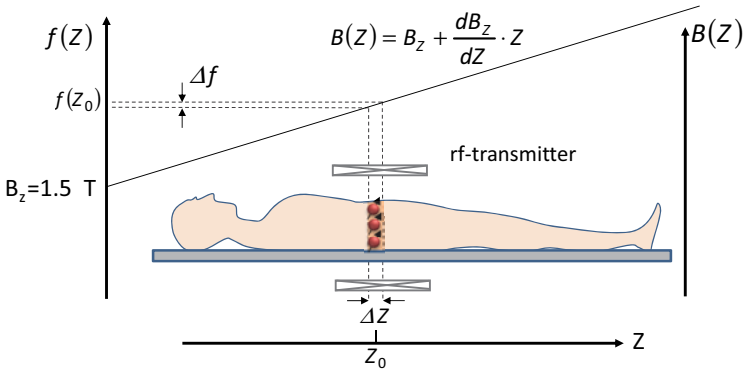


Fig. 15.25: Application of a gradient field in the Z direction.

As can be seen in Fig. 15.25, the thickness of the slice depends on the field gradient G_Z and on the band width Δf of the RF transmitter. Both can be tuned to the desired spatial resolution, which typically ranges from 1 mm to 3 mm. The *slice selection gradient* (SSG) in the Z direction is switched on only during the application of an RF pulse, while B_Z is permanently turned on.

The slice can now be subdivided into a matrix of 256×256 voxels, generating an image of the slice with a similar number of pixels representing the voxels in the slice. If the *field of view* (FOV) has the size $250 \times 250\,\text{mm}^2$, then the pixel represents an area of about $1 \times 1\,\text{mm}^2$ in the slice. To identify the signals from individual voxels in the slice, additional field gradients in the X and Y directions are required, as indicated in Fig. 15.24. By superposition of all gradients, the voxels will fulfill the local resonance condition at (X, Y, Z):

$$\omega_L(X, Y, Z) = \gamma(B_Z + G_X X + G_Y Y + G_Z Z).$$

However, this scheme does not work since within a slice the sum of the local fields $G_X X + G_Y Y$ is not unique. Points on both sides of the diagonal $X = Y$ experience the same gradient field and therefore the same frequency. For encoding columns in slices and voxels in columns different MRI procedures are needed and are applied. We distinguish between a *frequency encoding gradient* (FEG) in the X direction and a *phase encoding gradient* (PEG) in the Y direction.

15.4.2 Frequency encoding gradient

First we discuss the selection of columns by applying a field gradient in the X direction. While the Z gradient is turned on during emission of the RF transmitter to generate an M_z 90° flip and a 180° M_{xy} refocusing pulse, the X gradient is turned on for a few milliseconds during reception of the MR echo signal. Protons in differ-

ent columns will emit MR signals at slightly different frequencies and the RF coil, now used as receiver instead of transmitter, will record all of them simultaneously. As this RF coil has a double purpose, it is called *transceiver*. The MR signal received by the transceiver therefore consists of a spectrum of RF frequencies instead of a single frequency, as indicated in Fig. 15.26. The lower frequencies are, for instance, from columns on the left side of the slice and the higher frequencies from columns on the right side of the slice. Furthermore, the amplitude of the received signals depends on the local proton density. By Fourier analysis of the spectrum these different frequencies and amplitudes can be identified and assigned to the 256 columns in the slice.

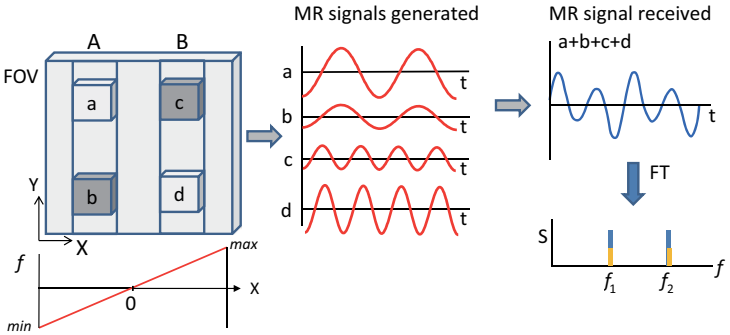


Fig. 15.26: Columns A and B generate MR signals with different frequencies and different amplitudes. The receiving coil senses a superposition of all frequencies from different columns. Fourier transform of the received signal yields the contributing frequencies and amplitudes. Pixels a, b and c, d can only be separated by an additional gradient in the *Y* direction. Note that the darker voxels b and c have lower amplitude.

The field of view (FOV) of the receiver coil depends on the bandwidth of the receiver and the field gradient. Since

$$\mathrm{d}\omega = \gamma G_X \,\mathrm{d}X,$$

and setting $\mathrm{d}X = \mathrm{FOV}$, we find:

$$\mathrm{FOV} = \frac{2\pi\,\mathrm{d}f}{\gamma G_X}.$$

For instance, a bandwidth of $\mathrm{d}f = 19\,\mathrm{kHz}$ and a field gradient of $G_X = 3\,\mathrm{mT/m}$ produces an FOV of 150 mm. The pixel width equals the FOV divided by the number of frequency components by which the frequency band width is sampled. In our example this is $150\,\mathrm{mm}/256 = 0.58\,\mathrm{mm}$ across. The frequency separation of each pixel is about 75 Hz.

15.4.3 Phase encoding gradient

Finally, we need to identify the voxels within a column. The standard procedure is to use a *phase encoding gradient* (PEG), see Fig. 15.27. After inversion from M_z to M_{xy} and before time t_0 for the 180° inversion pulse, a field gradient in the vertical Y direction is applied for a few milliseconds. Because of the gradient, the protons in all columns will precess with slightly different frequencies from top to bottom. When the Y gradient is turned off, within a particular column the protons precess again with the same frequency. But they will be out of phase. The ones that were precessing faster are still ahead in phase, and the ones that were precessing slower are still lagging behind. Thus the short Y gradient has imposed a phase gradient across all columns, which can be detected.

15.4.4 K-map

The MR signal from each slice ΔZ is therefore comprised of a frequency spectrum of bandwidth $\mathrm{d}\omega$ due to the X gradient, and a phase spectrum due to the G_Y gradient. There are two possibilities to gain information on both: either frequency and phase are detected by two coils arranged at 90° to each other in order to receive two independent projections of the precessing protons: k_X and k_Y; or only one coil is used for the MR frequency detection and the step-by-step phase angle variation is pre-adjusted via the G_Y coil. In any case, the (k_X, k_Y) points fill a 256×256 pixel K-map as illustrated in Fig. 15.27. k_X represents the frequency in a column, and k_Y represents the phase. k_Y goes from maximum negative values to maximum positive values, and zero is at the center. Frequencies cannot be negative. What is measured here is actually the frequency difference with respect to the precessional frequency of the slice at position Z and the center $X = 0$:

$$\Delta\omega = (\gamma B_Z + G_Z Z) \pm G_X X.$$

Same arguments hold for the k_Y values.

Figure 15.27 shows exemplarily how the K-map is filled with data points. First we concentrate on one particular column, which represents one particular frequency according to the G_X gradient, i.e., protons in voxel a, b, and c precess with the same frequency, but with different amplitudes due to differences in PD or T1–T2 relaxation times. The precessional frequency, indicated by yellow color in the K-map, is not the highest possible one, because the column chosen is not the last one in the FOV. Now we turn on a G_y gradient causing a maximum phase shift of +90° between the wave trains received from voxels a, b, and c. Frequency and phase shift go into the top row of the K-map marked in yellow. In the next round after completing one cycle within time TR, the phase shift is set to zero, providing one point in the middle row; in the third cycle the phase is set to −90° filling one yellow point in the last row.

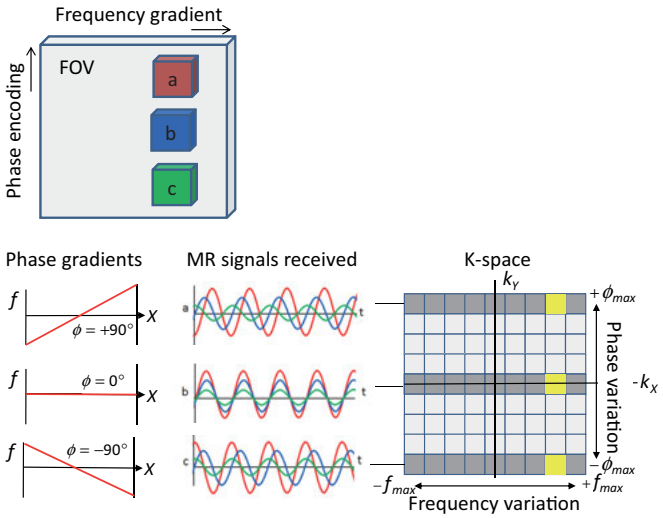


Fig. 15.27: Procedure of phase encoding. The top panel shows three voxels within one column. They all have the same precessional frequency. Excitation with different G_y gradients provides phase information, shown in lower left panel. The MR signal received is a superposition of all waves, including frequency and phase (lower middle panel). Fourier analysis reveals their frequency value (k_x) and phase (k_y).

In reality, the K-map consists of 256 rows, requiring 256 TR cycles for scanning all intermediate phase angles. Furthermore, in each cycle not one frequency of one column is recorded, but all 256 columns simultaneously, each assigned to a specific frequency, spanning the entire FOV. The received MR signal during one cycle is a superposition of all 256 different frequencies including their phase shift. Fourier analysis allows separating and assigning them to the different k_X points within one row. During the next cycle with a different phase angle, the next row is filled in, until the complete K-map is filled after 256 repeats.

The K-map is a Fourier representation of the real space from scanning one slice. The same argument holds for all slices, yielding a K-map in 3D. Back transformation should provide real space images, which is indeed the case as shown in Fig. 15.28.

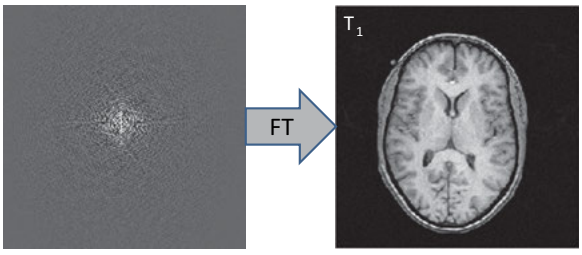


Fig. 15.28: Fourier transform of a K-space map into a real space image.

15.4.5 Fourier transform

Mathematically the Fourier transform can be shown as follows. First we realize that any voxel within a slice at position (X, Y) with local proton density $\rho(X, Y)$ has as a specific resonance frequency:

$$\omega(X, Y) = \omega_0(Z) + \gamma G_X X + \gamma G_Y Y,$$

where $\omega_0(Z)$ is a constant frequency that depends on the slice location Z. The frequency shift in the XY plane at constant Z can be identified with a phase change $\phi = \omega t$:

$$\phi(X, Y) = \gamma \int G_X X \, \mathrm{d}t + \gamma \int G_Y Y \, \mathrm{d}t.$$

The signal strength recorded for a proton density $\rho(X, Y)$ at constant Z is then:

$$\mathrm{d}S(t) = \rho(X, Y, t) \exp[2\pi \mathrm{i} \phi(X, Y, t)] \, \mathrm{d}X \mathrm{d}Y.$$

Converting phases into wavenumbers we can rephrase in vectorial notation:

$$\vec{K} = k_X \vec{e}_X + k_Y \vec{e}_Y;$$
$$k_X = 2\pi\gamma \int G_X \, \mathrm{d}t; \quad k_Y = 2\pi\gamma \int G_Y \, \mathrm{d}t;$$
$$\vec{K} \cdot \vec{r} = k_X X + k_Y Y,$$

where $\vec{r}$ is a spatial vector in the (X, Y) plane and $\vec{e}_x$, $\vec{e}_y$ are unit vectors. Integrating over the slice we find for the signal strength in K-space:

$$S(k_X, k_Y, t) = \iint \rho(X, Y, Z, t) \exp[\mathrm{i}(\vec{K} \cdot \vec{r})] \, \mathrm{d}X \mathrm{d}Y.$$

This is a standard 2D Fourier transform of an object with density distribution $\rho(X, Y)$ in real space. Therefore we conclude that MRI measurements in K-space represent 2D Fourier transforms of the proton density distribution $\rho(X, Y)$ in real space by providing information on phase and frequency within a slice. Back Fourier transformation into real space yields the desired image:

$$\rho(X, Y) = \iint S(k_X, k_Y) \exp[-\mathrm{i}(k_X X + k_Y Y)] \, \mathrm{d}k_X \mathrm{d}k_Y.$$

15.4.6 Data acquisition

Summarizing the spatial encoding for magnetic resonance imaging: the Z gradient defines the slice thickness, which is usually a transverse cut through the body starting from the head to the toe in the case of a full body scan. The Z gradient is applied as soon as the RF pulse is turned on. The X gradient produces a frequency change and

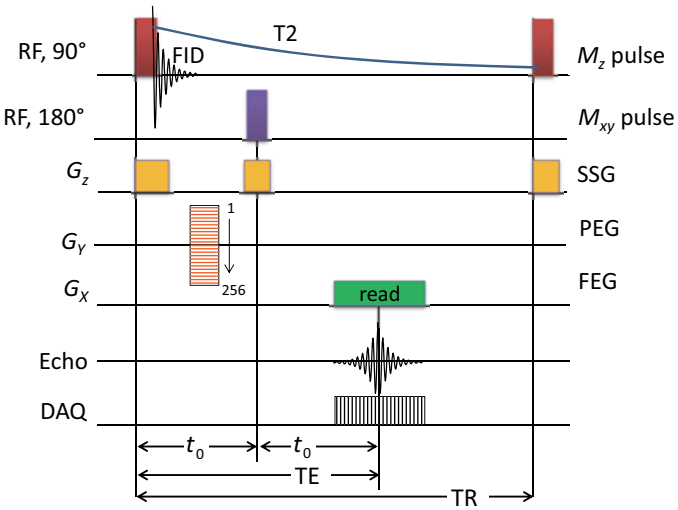


Fig. 15.29: Time sequence of different field gradients applied. SSG = slice selection gradient; PEG = phase encoding gradient; FEG = frequency encoding gradient; DAQ = data acquisition.

is applied during reception of the MR echo signal. The Y gradient produces a phase shift and is applied just before the 180° M_{xy} pulse reversal. The time relations of the different gradients applied are shown in Fig. 15.29. Data acquisition takes place during the X scan.

The gradients can be used for controlling the thickness of slices, FOV, and pixel size. The steeper the Z gradient, the thinner the transverse slices are. The steeper the frequency gradient and the phase gradient, the smaller the FOV is. The FOV is also controlled by the bandwidth of the transceiver. The bigger the bandwidth, the larger the FOV is.

Frequency and phase encoding operate on very different time scales. While frequency encoding is obtained in matters of 10–20 ms during recording of the echo signal, phase encoding takes much longer since the gradient has to be changed in a sequence of TR scans for each column. This can take as many as a few seconds up to minutes for one slice. During this time the examined person has to be immobilized, otherwise motion artefacts show up on the image.

15.5 Magnets and coils

Figure 15.30 shows a cutaway view of an MRI scanner displaying the most important components: main coil for the constant field B_Z, gradient coils for the X, Y, and Z gradients, and an RF transceiver. The patient lies on a coach, which can be slid into a borehole of about 70 cm diameter. All parts are cylindrically arranged about the Z axis.

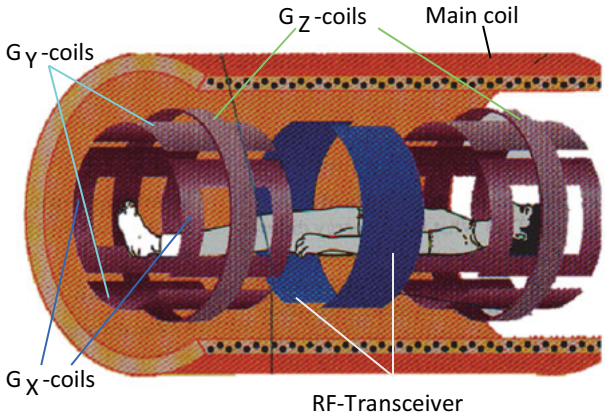


Fig. 15.30: Constant field and gradient field coils for MR imaging.

15.5.1 Main coil

The most expensive component of an MR imaging system is the magnet for the constant B_Z field. As the field required is above 1 T, conventional electromagnets cannot be used. The combination of a high magnetic field and a large borehole to fit a whole body is a challenging task requiring the use of *superconducting solenoids*. They are constructed like a long Helmholtz coil but with the difference that the wires are made of metals that become superconducting below a critical temperature T_c. Usually NbTi alloy wires are used, which have a transition temperature of about 9 K. This implies that the wires have to be cooled to temperatures below 9 K, which is done by a combination of liquid N_2 and liquid He cooling. The wires are bathed in liquid He at 4 K, surrounded by a heat radiation shield cooled by liquid N_2 to 77 K. Depending on specifications, *superconducting magnets* between 1–7 T are used; the standard ones for clinical applications have B_Z fields in the range of 1.5–3 T; for research often higher fields are applied. With larger fields, larger M_z and consequentially larger MR can be accomplished. However, larger fields also imply a longer recovery time T1 and therefore a longer repeat time TR, which contributes to an increased total time for taking images. Also artefacts from chemical shifts increase with increasing field.

One of the most important requirements for MRI is the field homogeneity along the Z direction and over the FOV with tolerance as low as 1 ppm. For this purpose permanent ferromagnets or additional electromagnets are used for fine tuning the main magnet.

Once current flows in the superconducting coil, it will continue to flow as long as the coil is kept at liquid helium temperatures. In modern MRI machines the boil-off of liquid He can be kept to a minimum, greatly reducing the operating cost of a scanner. Danger occurs when the superconducting coils quench by unplanned warming above

the critical temperature. Then the stored energy of some 20 kWh is transferred to the surrounding liquid gases, which rapidly vaporize.

It would certainly be more cost-effective if high temperature superconducting (HTS) wires with T_c above liquid N_2 temperature were used instead of conventional low temperature superconducting (LTS) wires. This would not only abolish expensive liquid He consumption, but also considerably simplify the cooling system of MRI machines. So far the brittle ceramic high T_c materials have hindered fabrication of wires needed for winding coils. However, this obstacle has been overcome recently by using second generation HTS wires. The major manufacturers of MRI machines are now implementing this new technology, which will likely soon be on the market.

15.5.2 Gradient coils

For generating field gradients, resistive wires are used which are operated at room temperature. A gradient G_z in the Z direction is achieved with an *anti-Helmholtz type coil* sketched in Fig. 15.31 (a). The current in both coils flows in opposite directions creating a magnetic field gradient between them. At the center between both coils the B_z field goes through zero. Anti-Helmholtz coils are also known as magnetic quadrupole coils. With such gradient coils field gradients in the order of 30 mT/m can be produced. Fast switching of gradient coils causes a loud knocking noise in MRI scanners, because of the exertion or release of Lorentz forces on the wires.

The X and Y field gradients are generated by pairs of so called *butterfly coils*, also known as Golay coils. Because of their half-circle shape and current flow direction, the field is oriented parallel to the Z axis for each pair of coils on one side and in the opposite direction on the other side. The immediately opposing loops (marked in red in Fig. 15.31 (b)) can be considered as half Helmholtz coils, the rest of the loop is simply returning the current. Therefore a field gradient develops in between, similar to the one in the Z direction. It is important to realize that all gradient coils produce

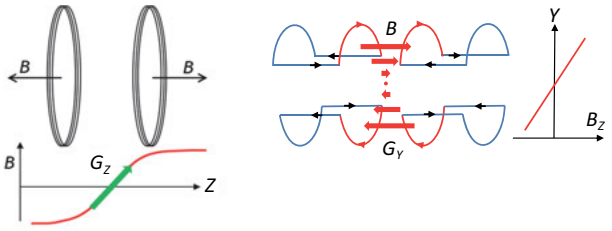


Fig. 15.31: Gradient magnetic field coils. For G_Z anti-Helmholtz coils are used (top panel), for the G_X and G_Y gradients pairs of half-circle Helmholtz coils generate field gradients in the X and Y direction (bottom panel). Only the G_Y gradient coil is shown, the G_X gradient coil works according to the same principle.

magnetic fields in the Z direction, only the field gradients have varying directions in the X, Y, and Z direction. More information on gradient coils and designs features can be found in [7].

15.5.3 RF coils

RF coils can be tuned like a radio to the resonant frequency. In return, the coil acts as a Faraday coil for receiving electromagnetic waves via measurable induction. Three different types of transceivers are distinguished. (1) *Standard body coils* for transmitting RF pulses and for picking up MR signals when imaging large parts of the body such as the chest or the abdomen. (2) *Head receiver coils* included in a helmet specifically used for brain imaging. (3) *Surface coils*, designed to be used locally for small area scans such as lumbar spine, knee, etc. These additional coils are fixed before the patient is slid into the machine. They allow smaller voxels and give better resolution over a smaller FOV.

15.5.4 MRI machine specifications

Most common MRI systems are specified with main magnetic fields of either 1.5 T or 3 T. Most applications can be performed with high quality using 1.5 T systems. However, the most advanced applications such as functional imaging, diffusion-weighted imaging, and time-resolved imaging require 3 T models. Some typical machine specifications are listed in Tab. 15.4. A 1.5 T commercial MRI scanner is shown in Fig. 15.32. An eight channel head scanner for high-resolution and high-speed brain imaging in shown in Fig. 15.33.

Tab. 15.4: Some parameters of standard 1.5 T and 3 T MRI systems. DSV = defined spherical volume.

	1.5 T	**3 T**
Field homogeneity 40 cm DSV ppm	0.2–0.4	0.1–0.5
Max FOV isotropic [mm]	500	500
Min FOV isotropic [mm]	5	5
Bore diameter [cm]	70	70
Bore length [cm]	150	150
Field gradient [mT/m]	33	50
He refill [years]	3	1

Fig. 15.32: Clinical 1.5 T MRI system with an RF power of 1 kW, field gradient of 33 mT/m, FOV of 53 cm, and a power consumption during scanning of 22 kW. Vendor: Philips, The Netherlands.

Fig. 15.33: Head scanner for brain imaging in a MRI system. Vendor: Siemens, Germany.

One of the main applications of MRI is the inspection of diseases and injuries in the area of joints at the extremities, such as knee, elbow, hand, wrist, and ankle. For such screenings full-body MRI scanners are not needed, but much smaller scanners in length as well as borehole size are sufficient. This has a dramatic impact on the design of such systems and finally on the operational cost. With a smaller borehole maintenance free permanent magnets can be used which provide a B_Z field of about 0.3 T, field gradients of 20 mT/m, and the total power consumption may be as low as 1 kW in contrast to 60 kW for a standard 1.5 T machine. A small unit extremity scanner is shown in Fig. 15.34.

Fig. 15.34: Extremity scanner using a permanent magnet of 0.3 T, maximum RF power of 1.5 kW, field gradient of ±20 mT/m, and FOV of 14 cm. Vendor: Esaote, Spain.

15.6 Applications of MRI

Main applications of MRI scanners are for imaging joints, brain and the entire head, mammary, and prostate. Imaging of moving organs like the heart was challenging in the past, but methods have been developed to overcome these problems. A number of specialized MRI applications have been developed over recent years, such as diffusion-weighted MRI, angio-MRI, multiparameter MRI, functional MRI, gated MRI, or hyperpolarization MRI. MRI is also greatly beneficial for detecting brain impairments via tumors, dementia, stroke, etc. For localizing brain activity in response to outside stimulus, functional MRI (fMRI) has been developed, which is also of great interest for brain research, cognitive sciences, and psychiatry. Some of the many MRI applications are discussed in the following.

15.6.1 Joints

There is a high MRI contrast between bones and tissue, which allows inspection of injuries in the area of joints. The contrast is mainly due to differences in T1 and PD of bones and surrounding tissues. Distinction of muscles, tendons, cartilage, fat, ligaments, and fluids is clearly possible. Therefore MRI is of benefit to disciplines such as sports medicine and orthopedics. In Fig. 15.35 some examples are shown. Recording of these images is done with standard T1- and T2-weighting procedures.

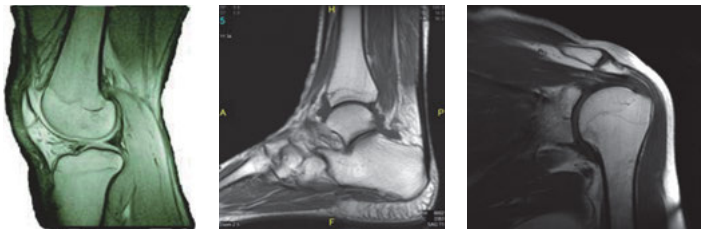


Fig. 15.35: MRI scans of joints. From left: sagittal scan of knee, foot, and coronal scan of shoulder, T1-weighted images (reproduced from https://openi.nlm.nih.gov/).

15.6.2 Dynamical contrast enhancement

In the case that T1- and T2-weighting does not produce sufficient contrast in soft tissue, such as the brain, there is the possibility to enhance the contrast with the use of paramagnetic ions. Paramagnetic ions generate magnetic dipole fields which are five orders of magnitude higher than magnetic fields from proton spins. The main effect of *contrast agents* (CA) is a shortening of T1 and/or T2 relaxation times.

When applying a contrast enhancing agent, the time dependence of CA distribution and washout has an important impact on the *dynamical contrast enhancement* MRI analysis (DCE-MRI). In Fig. 15.36 three types of time dependencies are illustrated. After administering the CA at time t_0, the distribution may be very slow and does not level off (type I), or may be fast in the beginning and very slow after reaching a plateau (type II), or fast in the beginning and washing out fast after reaching a maximum (type III). From an MRI point of view, type II is ideal. However, from a biological point of view, type III is more favorable. The curve shape may not only depend on the CA used but also on the type of tumor investigated. For instance, it has been shown that by curve shape analysis it is possible to discriminate between benign and malignant breast tumors since the residence time of CAs in malignant tumors is higher than in benign tumors.

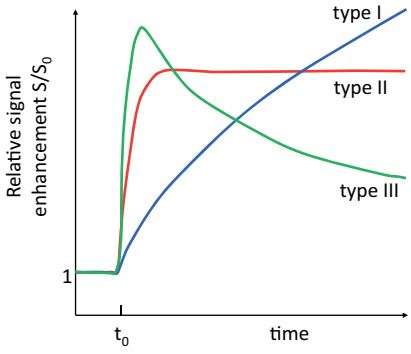


Fig. 15.36: Three types of time dependencies for contrast agent distribution and washout in the body after administering a CA at time t_0.

CAs have the purpose to enhance the relaxation, which is referred to as "relaxivity". *Relaxivity* is described by the relaxation rate $R = 1/\mathrm{T}$, where T is the relaxation time. For instance, $R_1 = 1/\mathrm{T1}$ is the relaxivity for the longitudinal relaxation time T1. This may vary between different tissues. However, assuming constant magnetic induction B_z and temperature it is generally assumed that the relaxivity is linearly proportional to the concentration x_{CA} of CA administered:

$$R_1 = r_1 x_{\mathrm{CA}} + R_{10}.$$

Here r_1 is the specific relaxivity (units: $\mathrm{mol}^{-1}\,\mathrm{s}^{-1}$) and R_{10} is the baseline relaxivity without CA. Using dynamic acquisition with different rates R_1 before, during, and after CA administration allows for a qualitative or quantitative characterization of specific tissues with typical differences between healthy and malignant behavior.

The rare earth metal ion Gd^{3+} is most often used as a CA. Gd^{3+} has a half-filled 4f shell with a spin state of $S = 7/2$ and a total magnetic moment of nearly 8 Bohr magnetons. Only the outer $5d^16s^2$ electrons take part in chemical bonding while the 4f shell and the full magnetic moment remain preserved. Gd^{3+} packed in a *chelate complex* such as *diethylenetriamine penta-acetic acid* (short DTPA) is biocompatible, as the cytotoxic heavy metal ion is strongly bound and thus hindered from leaching into the cellular environment. The chemical structure is shown in Fig. 15.37. It is one of a family of chelates that are used for contrast enhancement. Gd chelates belong to type III CAs: after intravenous injection they are distributed within 12 minutes and cleared out from the plasma within a short blood circulation half-life of about 100 minutes, sufficient for taking MRI images. Excretion in an unchanged form takes place via the kidneys.

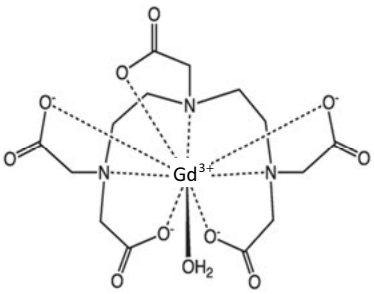
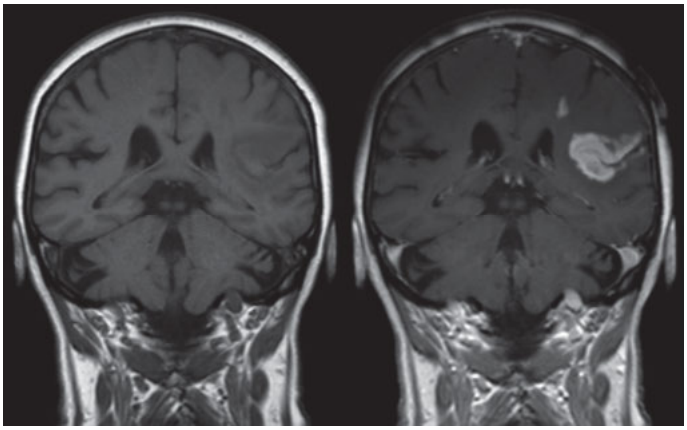


Fig. 15.37: Chemical structure of the Gd^{3+} chelate diethylenetriamine penta-acetic acid (DTPA), used for T1 contrast enhancement of cranial and spinal MRI.

Gd chelates preferentially shorten T1 values in tissues where they accumulate, rendering them bright on T1-weighted images. Gd^{3+} chelates do not pass the intact *blood-brain barrier* (BBB) because of their hydrophilic properties. However, in the case of a BBB breakdown due to tumors or stroke, Gd chelates can enter the brain tissue and support contrast enhancement. In Fig. 15.38 an example is shown of a T1-weighted coronal section through the brain taken with and without Gd-CA. The Gd-enhanced image shows clear changes in the region of the BBB breakdown.



(a) (b)

Fig. 15.38: T1-weighted MRI scans of the brain without (a) and with Gd chelate enhancement (b). The T1-weighted coronal sections show in comparison defects of the blood–brain barrier that occurred after stroke. Without these defects Gd chelate contrast agent would not be able to penetrate the blood-brain barrier (reproduced from Wikipedia, © Creative Commons).

Some tumors have little contrast to their surrounding healthy tissue, making differentiation difficult. In those cases Gd-CAs are often used for better identification and delineation of the tumor volume. A discussion of the identification of tumors in the abdomen, pelvis, brain and spine using Gd-CAs is provided in [8].

Gd chelates are called positive image CAs, as the shortening of the T1 relaxation time means that those tissues where the agent accumulates appear brighter. This method has played a vital role in early tumor recognition.

Superparamagnetic iron oxide nanoparticles (see Chapter 14/Vol. 2) are more effective for shortening the T2 relaxation time, producing negative image contrast as shorter T2 appear darker (see Tab. 15.3).

There are many more contrast CAs in use and still being tested for specific tasks and targets. Nevertheless, the majority of MRI CAs used in clinics are Gd^{3+} chelates. This is due to their favorable properties in terms of contrast enhancement, high chemical stability, short biological half-life, and inertness in the body. A recent overview on contrast enhancement agents is provided in [9]. However, there are also some shortcomings with respect to the very short circulation half-life and decreasing r_1 values in high fields that make T2 CAs attractive.

15.6.3 Angio-MRI

Venous blood consists of 70 % paramagnetic deoxyhemoglobin. In contrast, arterial blood is 95 % nonmagnetic oxyhemoglobin. This provides a contrast mainly in T2-

weighted images. However, since blood flows, images of blood vessels appear blurred. With the help of Gd chelates the contrast can be dramatically increased, but the distinction between venous and arterial blood is lost [10]. An impressive angio MRI image of the blood vessels taken by full-body scan in several sections is shown in Fig. 15.39. The distinction between oxy- and deoxyhemoglobin is essential for functional MRI (fMRI), which is discussed in Section 15.6.7.

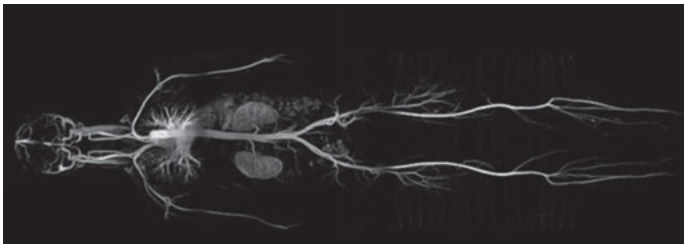


Fig. 15.39: Angio-MRI of the blood vessels. The contrast is enhanced by Gd chelate (reproduced from https://www.healthcare.siemens.de/magnetic-resonance-imaging).

15.6.4 Hyperpolarization MRI

Hyperpolarization MRI (hMRI) is an imaging modality similar to DCE-MRI. It also uses a contrast enhancing agent, but in distinction to DCE-MRI the CA is not paramagnetic. Instead, stable isotopes are used with an odd number of nuclei and therefore unpaired nuclear spins, which are themselves NMR active. Examples are: ^{3}He, ^{13}C, ^{15}N, ^{17}O, ^{19}F, and ^{31}P. Since these nuclei are not as abundant as ^{1}H in the body, their lower density ρ is compensated for by an artificially induced higher polarization p in order to achieve high signal strength $S \approx p\rho\mu$ and a good signal-to-noise ratio (SNR), where μ is the respective nuclear magnetic moment. Hyperpolarization refers to the fact that the artificial polarization of the nuclei is beyond thermal equilibrium and therefore unstable with lifetimes ranging from seconds to hours. The nuclear hyperpolarization is always performed ex situ. Three methods are commonly used: *laser excitation*, *dynamic nuclear polarization* (DNP), or *parahydrogen induced polarization* (PHIP).

The isotope ^{3}He can be polarized by laser excitation, a process referred to as metastability exchange optical pumping (MEOP) [11]. In this method, optical pumping is first performed on the metastable 2^3S triplet state of ortho-^{3}He atoms (see Chapter 13/Vol. 2). By optical pumping with a polarized laser the 2^3S state becomes polarized. Due to an efficient hyperfine coupling between electrons and nuclei of 2^3S He atoms, this electronic polarization induces nuclear polarization as well. Finally, collisions between metastable 2^3S and ground state 1^1S He atoms transfer the nuclear polarization to the ground state. This process requires only low magnetic fields and room temperature conditions. Polarization of more than 90 % can be achieved.

DNP implies a transfer of high electron spin polarization of paramagnetic ions to nuclear spins via electron-nuclear spin-spin interaction. For this purpose the target containing the isotope in question is doped by paramagnetic ions. Polarization transfer to the isotopes is achieved by microwave irradiation with frequencies close to the electron spin resonance in high fields ($\approx 3\,\mathrm{T}$) and low temperatures ($\approx 1\,\mathrm{K}$) [12].

Similarly, PHIP uses the low temperature antiparallel proton spin state of parahydrogen that is transferred to molecules in contact, causing a hyperpolarization of sublevels much beyond Boltzmann distribution. By this process the NMR signal strength becomes dramatically enhanced. In all cases, following hyperpolarization, the respective agents are either inhaled or intravenously injected for distribution in the body via the circulatory system (see Chapter 8).

Unlike CAs, hyperpolarized isotopes do not contribute to changes of T1 or T2 of ^{1}H-MRI. Instead these isotopes, aside from ^{3}He, take part in the metabolic process. *Magnetic resonance spectroscopy imaging* (MRSI) at specific resonance lines of these isotopes provides metabolic information in addition to morphologic contrast. In this respect hMRI has similarities with functional MRI (fMRI) presented in Section 15.6.7.

In the following a few examples for hMRI are discussed.

(a) ^{3}He

The respiratory system is difficult to image using ^{1}H-MRI as the proton density is rather low. Nevertheless this image modality has advantages over x-ray projection radiography or CT in those cases when ionizing radiation is not tolerable, for instance during pregnancy, for small children, or when multiple radiographs are needed and the total accumulated dose should be kept as low as possible. Because of the low contrast, contrast enhancement methods are needed. One possibility is filling the respiratory volume with Gd-containing nanoparticles. But the risk of this procedure is quite high, as the nanoparticles similar to natural fine dust can impact on the immune system and enhance toxicity, in particular when the heavy metal ions penetrate into the tissue. An interesting alternative has been proposed, using polarized Helium for inhalation [13]. Most people have already experienced by themselves or by demonstration the impact of He gas on the voice when inhaling and exhaling. ^{3}He is an isotope of standard ^{4}He with one missing neutron and therefore with an uncompensated neutron spin that can be polarized by laser pumping, as described above. The hyperpolarization is performed in a magnetic field external to the patient, but even if taken out of the field and inhaled, the polarization remains sufficiently high for acquiring MRI images. Using fast spin-echo techniques a sequence of images can be taken for studying pulmonary ventilation during inhaling and exhaling, as shown in Fig. 15.40. One can easily recognize the flow of ^{3}He gas through the bronchial branches, distributing over the entire tidal volume, and retracting again during exhaling. From such images damage to the lung tissue through emboli, tumors, smoking, or asthma can be recognized and analyzed [14]. Turning to the signal strength: $S \approx p\rho\mu$. In the case of protons the

polarization in a field of 1 T is in the order of 10^{-6}, for ^{3}He it is nearly 1 after hyperpolarization. The ratio of the densities is about $\rho_H/\rho_{He} \approx 2500$. The magnetic moments have different signs but otherwise are almost equal (see Section 15.1). Taking all these factors into account the ratio of the signal strengths for ^{3}He imaging versus proton (^{1}H) imaging is about $S_{He}/S_H > 10$. This is indeed a sizeable enhancement factor. The only disadvantage is the extreme rarity of ^{3}He and therefore its high price. To overcome the cost barrier other hyperpolarized rare gases have been tested, such as ^{129}Xe and ^{83}Kr. A discussion of the present status of hyperpolarized gases for pulmonary MRI studies can be found in [15].

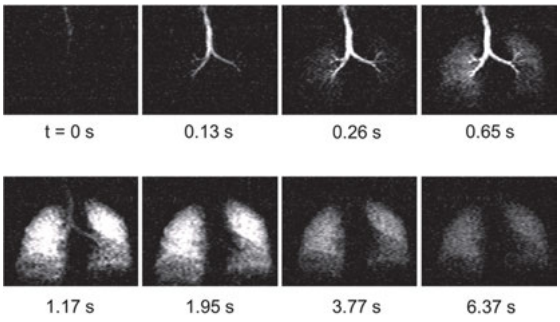


Fig. 15.40: MRI sequence of images taken during inhaling and exhaling polarized ^{3}He gas (courtesy Werner Heil, Johannes Gutenberg Universität Mainz, Germany).

(b) ^{13}C

For assessing prostate cancer ^{13}C hMRI has been used [16]. ^{13}C has a natural abundance of 1.1 %. Because of its larger mass the gyromagnetic ratio is smaller than for protons: $\gamma_{^{13}C} = 0.673 \times 10^8\ T^{-1}\,s^{-1}$ and the resonance frequency is accordingly lower: $\omega/2\pi = f(\text{MHz}) = 10.7 B_z(\text{MHz})$. Therefore higher magnetic fields are required for inducing a Zeeman splitting similar to the one for protons. Furthermore, molecules used for studying metabolic pathways, such as pyruvate ($C_3H_4O_3$), are enriched with ^{13}C to increase the isotopic density. In addition, DNP or PHIP is applied for hyperpolarization. DNP increases the SNR by 4 to 5 orders of magnitude, rendering in vivo imaging of various metabolites and their enzymatic conversion into other species possible. The main limitation of this imaging technique is the short half-life of the polarization, which is only 30–40 s, meaning that the hyperpolarized signal is useful only for a mere 2–3 minutes. Therefore, administering enriched and hyperpolarized ^{13}C containing molecules to the target has to proceed very fast.

(c) ^{17}O

^{17}O-hMRI is predestinated for performing studies of oxygen metabolism in general and in particular of the brain [17]. Similar to ^{13}C, ^{17}O also has a small gyromagnetic ratio of $\gamma_{^{17}O} = -0.362 \times 10^8\ T^{-1}\,s^{-1}$ and a correspondingly low resonance frequency $f(\text{MHz}) =$

$-5.77B_z$(MHz). The minus sign indicates that angular momentum and nuclear magnetic moment are antiparallel, which, however, does not impact on MRI protocols. The natural abundance of ^{17}O is only 0.037 %, which is a severe limitation and makes the method costly. Enrichment of the oxygen is absolutely necessary as well as recycling, similar to ^{3}He. When ^{17}O is used in water ($H_2{}^{17}O$) hyperpolarization is possible via DNP, but presently not in the gaseous state.

(d) ^{19}F

Among the above listed isotopes, ^{19}F features the most favorable NMR properties. The natural abundance of ^{19}F is 100 %, the spin 1/2, gyromagnetic ratio $\gamma_{^{19}F} = 2.51 \times 10^8\,T^{-1}\,s^{-1}$ and resonance frequency f(MHz) = $40.08B_z$(MHz) are very close to ^{1}H-MRI. Only trace amounts of ^{19}F can be found in the body in solid form (bones and teeth). Therefore the signal is naturally background free and MRI signal strength can be directly related to local ^{19}F concentrations. Hyperpolarization with PHIP has been demonstrated. The potential for ^{19}F MRSI studies is high. Many drugs contain fluor and MRI can be used for studying the pharmakinetics of drug delivery. Furthermore many fluor containing molecules, polymers, and nanoparticles have been designed for targeting specific tissues. ^{19}F-MRSI also has the potential to replace ^{18}FDG-PET, which is presently the prime modality for imaging metabolic alterations in cancerous tissues. ^{18}FDG-PET, discussed in Chapter 6/Vol. 2, requires an extensive infrastructure for producing and handling short lived ^{18}F-radioisotopes that would become obsolete with the introduction and routine use of ^{19}F-MRSI. In spite of all these favorable properties, in particular low background, high sensitivity, and zero radiation risk, clinical applications of fluorinated CAs are still limited. It is likely that this will change in the near future. A recent review of this topic can be found in [18].

15.6.5 Diffusion-weighted imaging MRI (DWI)

Diffusion-weighted imaging MRI (DWI) can be applied to any body part for studying diffusion and perfusion of liquids and gases through tissue. DWI is particularly valuable for investigations of *neuroactivities*.

Randomly diffusing particles have a mean square displacement $\langle x \rangle^2 = 6D\tau$, where D is the diffusion constant and τ is the incremental time during which the displacement occurs. When protons diffuse, this will lead to an additional damping of the SE signal, i.e., the T2 time will be shortened. Diffusion can be tested by applying a gradient field pulse G_D just before the 180° refocusing pulse and another one of the same magnitude but opposite in direction just after the 180° pulse. The first G_D pulse marked in yellow in Fig. 15.41 will dephase the proton spins in the xy plane in addition to the already existing dephasing due to T2* relaxation, while the second opposing pulse will reverse the dephasing of the first pulse. If the system is static, these two G_D pulses can-

cel each other out and no additional change will occur. However, if the protons are in diffusional motion, the cancellation is incomplete and the MRI signal strength will experience an additionally damping. The attenuation due to diffusion is expressed as:

$$S_{\mathrm{DWI}} = S_0 \exp\left[-(\gamma G_{\mathrm{D}}\delta)^2\left(\Delta - \frac{\delta}{3}\right)D\right] = S_0 \exp[-bD].$$

S_0 contains the usual relaxations due to T1 recovery and T2 decay:

$$S_0 = K[\rho_{\mathrm{H}}]\left(1 - \exp\left(-\frac{\mathrm{TR}}{T_1}\right)\right)\exp\left(-\frac{\mathrm{TE}}{T_2}\right).$$

Combined we have:

$$S_{\mathrm{DWI}} = K[\rho_{\mathrm{H}}]\left(1 - \exp\left(-\frac{\mathrm{TR}}{T_1}\right)\right)\exp\left(-\frac{\mathrm{TE}}{T_2}\right)\exp[-bD].$$

This equation is known as the Stejskal–Tanner equation [19]. γ is the gyromagnetic ratio, G_{D} is the pulse amplitude, δ is the pulse length and Δ is the pulse separation. $K[\rho_{\mathrm{H}}]$ is a constant proportional to the proton density. The symbols are also explained in Fig. 15.41. Usually the prefactors in the exponent are combined in a b factor: $b = (\gamma G_{\mathrm{D}}\delta)^2(\Delta - \delta/3)$, which has the units of $\mathrm{s/m^2}$. b is a machine parameter that can be controlled by the operator, D is the intrinsic diffusion constant of the system. Because of the complexity of all contributing factors, D is also called *apparent diffusional constant* (ADC). The gradient can be applied in any direction X, Y, or Z to analyze the main diffusional direction of fluids, and diffusional MRI can be combined with any of the other already discussed protocols for time-resolved and fast MRI methods. However, the Stejskal–Tanner equation is only valid for tissues in which diffusion shows no spatial anisotropy, i.e., purely Brownian type motion.

If only one measurement is performed with one b value, it is difficult to distinguish whether the attenuation is due to diffusion or additional T2* relaxation. However, using two different b values, the diffusional contrast can be separated from all other effects:

$$\ln\frac{S^1_{\mathrm{DWI}}}{S^2_{\mathrm{DWI}}} = -(b_1 - b_2)D.$$

The diffusional constant D or better ADC can then be characterized for specific locations, for instance in the white or gray matter of the brain, and differences can be identified for healthy tissues compared to those which have suffered damage or injury.

While DWI methods have successfully been applied in clinical studies, it turned out that the diffusional behavior of bound and unbound water is so complex, in particular in the brain, that it is necessary to use a full anisotropic diffusional model, where D and b are treated as tensors. The ratio of signal strength with and without G_{D} turned on is then:

$$\ln\frac{S_{\mathrm{DWI}}}{S_0} = -b_{ij}D_{ij},$$

where according to tensor notation sums are taken over double indices. Applications of this method are named *diffusion tensor imaging* (DTI). The idea is that in fibers such

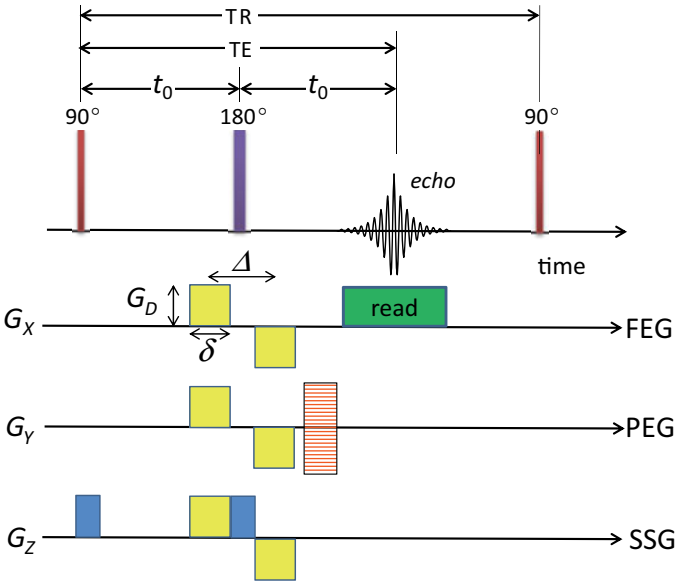


Fig. 15.41: Time sequence for diffusion-weighted imaging. An additional gradient field pulse G_D is applied shortly before and shortly after the 180° refocusing pulse. The G_D pulse can be applied in any direction in addition to the usual pulse sequence.

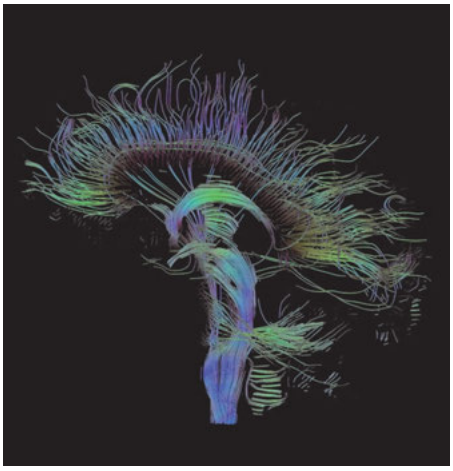


Fig. 15.42: Wiring of the human brain visualized by tracking the movement of water molecules using diffusion tensor imaging (DTI). The nerve fibers run through the mid-sagittal plane. Particularly noticeable are fibers that connect the two hemispheres through the corpus callosum and those which descend toward the spine (blue, within the image plane) (reproduced from Wikipedia https://en.wikipedia.org/wiki/Diffusion_MRI, © creative Commons).

as nerve fibers the diffusion is highly anisotropic and that this anisotropic and directional diffusion can still be visualized by DTI, an example is shown in Fig. 15.42. For a review and further details we refer to [19].

15.6.6 Multiple parameter MRI (mpMRI)

Multiparameter MRI (mpMRI) is becoming an increasingly important modality for early cancer recognition with superior reliability as compared to PET or US screening. mpMRI combines diffusion-weighted imaging (DWI), dynamic contrast enhanced MRI (DCE-MRI), and magnetic resonance spectroscopy imaging (MRSI) in addition to T2-weighted contrast imaging. The combination of these methods allows for determining not only the size and morphology of potential tumors, but also their physiological response, for instance in cases of prostate carcinoma. Therefore a distinction is possible between clinically not relevant tumors and malignant tumors with much enhanced sensitivity and specificity. Employing mpMRI, locations for biopsy that are required for confirmation of tumors can be placed much more precisely and unnecessary surgery can be avoided. mpMRI is comparable to functional MRI, discussed in the next section, as it provides morphological and functional information on specific tissue volumes. mpMRI is also important after radiation treatment of prostate carcinoma to evaluate tumor response without accumulating further radiation.

15.6.7 Functional MRI (fMRI)

Functional MRI (fMRI) is based on changes of T1 and T2 relaxation times (R1 and R2 relaxivity) of those protons which are close to hemoglobin in the brain. Neural activity enhances oxygen-rich blood flow. Without oxygen bonding (deoxyhemoglobin) Fe^{2+} is in a high-spin $S = 2$ paramagnetic state associated with a magnetic moment of 5.4 μ_B. In contrast, with O_2 bonding Fe^{2+} is in a low-spin $S = 0$ state with zero magnetic moment. Further details on the high spin-low spin transition of Fe^{2+} in hemoglobin are discussed in Section 8.6.2. In the high spin deoxy state the spin-spin interaction is large, similar to the case of Gd^{3+} discussed in Section 15.6.2, and the T2* relaxation time is accordingly short, whereas in the oxy state T2* is normal. From this we notice that Fe^{2+} acts as an endogenous CA that can, in addition, differentiate between oxy- and deoxyhemoglobin, which, in turn, is related to brain activity. Therefore the T2* relaxation time can be utilized for mapping out the brain and correlating locations of brain activity with enhanced blood flow. The amount of oxygen delivered by the blood flow to neuronal activity centers exceeds the amount required by the surrounding tissue, thus increasing the ratio of oxy- to deoxyhemoglobin. The signal change is small, but is reliably detectable by subtracting images collected at rest from images collected during activity.

In short, fMRI relates high metabolic activity of the brain with external stimuli for locating responsive centers in the brain. This scanning technique is named *blood oxygen level dependent* (BOLD) fMRI acquisition. The stimuli may be visual, audible, physical (tapping a finger), or cognitive (associative, problem solving, etc.). An example for an associative stimulus is shown in Fig. 15.43. Children were given the task of associating objects with a physical activity, such as "ball" and "throw" [20]. A high correlation between different locations in the brain needed for this task is clearly visible in the T2-weighted MRI cross sections using fast EPI scanning techniques, discussed in the next section. fMRI has been used for many different tasks, mapping out most of the brain and localizing the centers for language, music, different peripheral movements, etc. In fact the resolution has become so high that single words can be identified and "read" by fMRI mapping. Another active research branch of fMRI are studies of neural interconnectivity of the brain in the resting state, i.e., in a state without specific tasks. An overview on current brain research with fMRI is provided in [21].

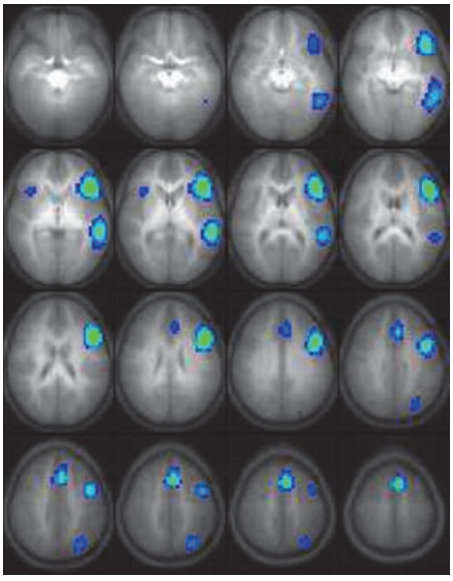


Fig. 15.43: Sequence of fMRI maps of the brain after stimulation through word association (reproduced from [20] by permission of John Wiley and Sons Inc.).

One of the most intriguing problems in brain research is the question of plasticity. Nerve cells in the brain develop at an early age but after reaching adulthood neurons and neural networks are never refurbished. This at least is the traditional view. In Fig. 15.44 from [22] we see fMRI images upon light response of a blind dog and a healthy dog for control. The blind dog has a gene defect and was therefore blind since birth. After gene manipulation retinal visibility could be re-established. However, the ques-

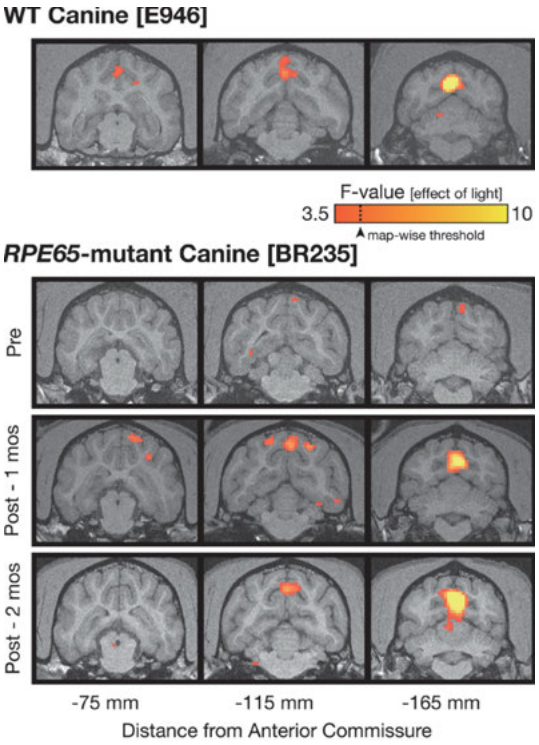


Fig. 15.44: fMRI responses of a blind dog before and after gene therapy. Three coronal slices through the brain are shown at different distances from the anterior commissure, red and yellow colors indicate the location of significant responses to light stimulation. Top row: visual responses of a healthy control dog labeled WT Canine [E946]. Lower three rows: pre- and post-treatment data from a blind dog. Post-treatment data were taken during two separate sessions separated by one month. The MRI scans confirm normal response in both sessions (reproduced from open access [22]).

tion was whether the blind dog would ever develop optic nerves and a visual cortex for recognizing light. Can blindness be cured even if born blind? Figure 15.44 demonstrates that the blind dog regained his visual perception after only two months since gene therapy. fMRI BOLD scans shows visual sensitivity at the same location of the visual cortex as the healthy dog. The images are taken at three different coronal slices behind the anterior commissure, which is a reference point in the brain.

15.6.8 Real time MRI

Standard MRI has a time resolution of 1–5 s for scanning a complete slice. Rhythmically moving organs such as the heart with a cycle time of about 700–900 ms at rest cannot be imaged without additional procedures. Two techniques are applied for these cases:

(a) *Stroboscopic imaging* by gating the transceiver with the ECG signal. This has the advantage of a high spatial and time resolution. However, the method can only be used in case of periodic sequences. The method fails in case of cardiac arrhythmia, which are typical for heart diseases. Recent further developments have shown that arrhythmic heart beats can still be imaged by self-gating methods.

(b) *Echo planar imaging* (EPI), also named *turbo-MRI*, is an imaging modality in real time with a time resolution of about ca. 10–20 frames/s. With such a high time resolution aperiodic processes can also be recorded, such as perfusion studies of heart, lung or kidney. The disadvantage is a poor spatial resolution of the frames. The technique is based on the *fast spin-echo* (FSE) technique. In FSE the 180° refocusing pulse is applied several times before reaching TR. Each time the pulses will be refocused but consecutive amplitudes of the SE signal will decrease while still detectable, as shown in Fig. 15.13. Furthermore, for each 180° pulse the PEG is changed between echoes, which allows filling up several lines in the K-map within one TR interval. The FSE pulse sequence shown in Fig. 15.45 is a combination of the MRI protocols displayed in Fig. 15.13 and Fig. 15.29. The number of echoes recorded in a given TR interval is labeled *echo train length* (ETL) or *turbo factor*. The ETL typically ranges from 4 to 32 for routine imaging, in Fig. 15.42 it is only 3. In the case of ETL = 32 only 8 TRs are required for scanning an entire K-map, which is a remarkable time saving.

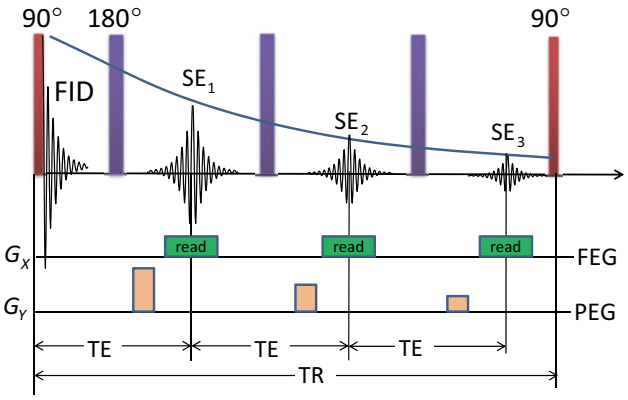


Fig. 15.45: Time sequence for fast spin-echo (FSE) scans. SE is repeated many times within one TR, each time the phase gradient is changed for acquiring multiple lines of the K-map within one TR interval.

EPI is a variant of FSE. A sequence of consecutive spin echoes are generated by 180° refocusing pulses similar to FSE (see Fig. 15.46). However, after each echo the sign of the readout FEG G_X gradient is alternated and the initial PEG is reduced sequentially by opposing smaller G_Y gradients. The consequence is a faster filling of the K-map,

by running from left to right in the topmost row, then right to left in the next row, etc. Finally the K-map is completely filled within just one or a few TR intervals. For time-resolved measurements, only one TR interval with an ETL of 128 is used to fill a 128 × 128 K-map. There are several other sequences that are being used for time-resolved MRI. MRI movies of the beating heart can be seen in the internet, for instance on the webpage of [23].

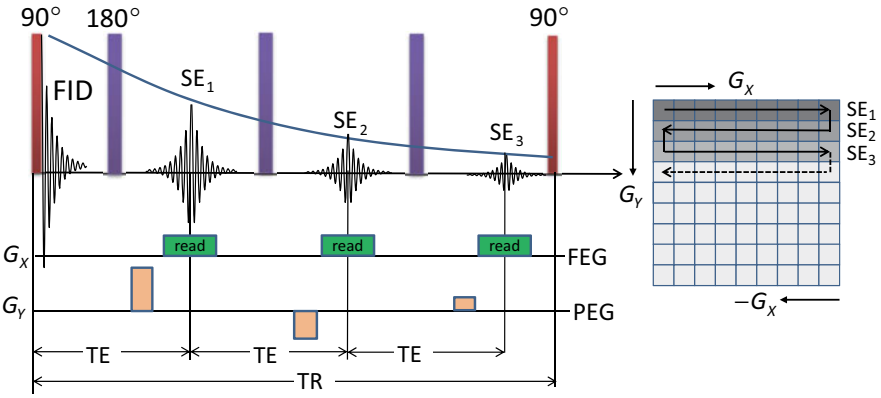


Fig. 15.46: Echo planar imaging method is a fast spin-echo sequence for time-resolved imaging of moving organs such as the heart.

15.7 New trends

MRI methods and equipment have reached a very high standard. There are many procedures and protocols available for enhancing signals, increasing time resolution, determining diffusion, perfusion, and brain activity. 1.5 T and 3 T MRI scanners are established as the most useful devices for clinical applications. For research, scanners up to 9 T are being used, but the benefit of these high fields is not obvious. The high resolution is counterbalanced by a number of artefacts due to susceptibility and chemical shift problems. Smaller units specialized on scanning extremities (hands, elbows, feet, knees) are entering the market, which have decisive advantages over their larger sized sisters, such as regarding maintenance, room temperature operation, and investment cost as well as operational cost. Clearly, the versatility of full-size scanners is not available, but high contrast imaging of static and anatomic lesions is still possible. It can be foreseen that in the near future almost any medical practitioner will have a small MRI scanner, as is already the case for US scanners. It may also be foreseen that in the long run full size MRI scanners will take over SPECT and PET. MRI has the compelling advantage that no or negligible radiation hazards are imposed on pa-

tients and that no radioisotopes are required, including infrastructure for production and disposal of short lived isotopes.

MRI development could do even better than already impressively demonstrated if certain paradigms were called into question. The two main paradigms are: (1) high fields, and (2) inductive detection. They go hand-in-hand. However, brain imaging has been demonstrated with only 130 μT using SQUID detectors [24]. More generally, new emerging ultralow field MRI techniques (ULFMRI) may even be applicable to imaging of tumors without the need for contrast enhancing agents and imaging protocols that require fields of at least 1.5 T. The message of these promising developments is the following: as long as inductive coils are used for detecting MR signals, high frequencies are required since the induced voltage increases with frequency. High frequencies require high fields. Using pick-up sensors based on SQUIDs or atomic field sensors, MRI could be performed at much lower fields and may become more affordable than present day units. There is still much room for further developments.

15.8 Advantages, hazards, and disadvantages

Because of the high magnetic field operation, extreme precautions must be taken to keep metallic objects out of the room where the MRI machine is operating. The same applies to patients. Any metal wristwatch, jewelry, pacemaker, glasses, hearing aids, metal implants unless diamagnetic must stay outside.

The RF coils act like a microwave oven. Care has to be taken to keep the power at a level that the body temperature does not increase more than one centigrade.

Some people suffer from claustrophobia and find the confinement in a closed tube frightening. Open C-frame machines have been designed for these cases. However, magnetic field homogeneity, which is essential for resolution, is naturally not as perfect as in a closed cylindrical system.

MRI machines tend to make very loud and continuous hammering noise when operating because of the fast switching of the coils. The fast switching of the gradient coils may also generate eddy currents in electrolytes. However, negative effects have so far not been reported.

Some patients may be too big to fit inside the tube. The present standard size is 70 cm diameter. Open C-frame machines may be a solution also in those cases.

Some patients may be allergic to Gd chelate contrast agent. In those cases alternative CAs have to be used.

MRI scans require patients to hold very still for long periods of time. But using fast scanning techniques, this condition has been loosened considerably.

MRI systems are expensive to buy and to run. Accordingly MRI scans are expensive. This will change in future as smaller units with much reduced investment and operational cost become available.

There are no radiation hazards to patients or to the staff. But potential dangers occur due the high magnetic fields and in the case that the superconducting coils quench.

15.9 Summary

1. Nuclei with odd atomic numbers exhibit weak nuclear paramagnetism.
2. In an external field nuclear moments precess with a characteristic frequency, i.e., the Larmor frequency.
3. Applying an RF field perpendicular to the main field, nuclear magnetic moments can be rotated by various angles, most prominently by 90° or 180°.
4. Spin-echo refocusing is the single most important feature for image generation.
5. Transverse slices are defined by field gradient in the *z* direction.
6. The columns in a slice are encoded by a frequency gradient; rows are encoded by a phase gradient.
7. Frequency and phase are recorded in a K-map, which delivers a real space picture upon Fourier transformation.
8. The longitudinal relaxation time T1 into the equilibrium state depends on the spin-lattice interaction, i.e., the interaction of the nuclear spin with its surrounding.
9. The transverse relaxation time T2 for dephasing of spins in the plane perpendicular to the main field depends on spin-spin interactions.
10. Magnetic resonance imaging (MRI) offers a high resolution noninvasive and radiation-free three-dimensional imaging modality of tissues with contrast generated by differences in T1, T2, or proton density.
11. Contrast enhancement can be achieved by specific T1-, T2-, and proton density-weighting schemes, or by the use of paramagnetic contrast enhancing agents.
12. Fast spin echo for moving organs, functional MRI for stimulated brain activity, and diffusional MRI for diffusion or perfusion studies are additional special MRI applications.
13. MRI with protons (^{1}H-MRI) is supplemented by MR spectroscopic imaging using a variety of stable isotopes: ^{3}He, ^{13}C, ^{15}N, ^{17}O, ^{19}F, ^{31}P, ^{83}Kr, and ^{129}Xe for specific tasks.

References

[1] Rabi II, Zacharias JR, Millman S, Kusch P. A new method of measuring nuclear magnetic moment. Physical Review. 1938; 53: 318.
[2] Bloch F, Hansen W, Packard M. The nuclear induction experiment. Physical Review. 1946: 70: 474–485.
[3] Purcell EM, Torrey HC, Pound RV. Resonance absorption by nuclear magnetic moments in a solid. Physical Review. 1946; 69: 37–38.
[4] Lauterbur P. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. Nature. 1973; 242: 190–191.
[5] Mansfield P, Grannell PK. NMR 'diffraction' in solids? J Phys C: Solid State Phys. 1973; 6: L422–L426.
[6] Hahn E. Spin echoes. Phys Rev. 1950; 80: 580.

[7] Hidalgo-Tobon SS. Theory of gradient coil design methods for magnetic resonance imaging. Concepts in Magnetic Resonance Part A. 2010; 36A: 223–242.
[8] Pruvo J, Vilgrain V, Roy C, Peretti P, Halimi P, Ernst O, Valette P, Matos C, El-Khoury C. Characterisation of central nervous system, liver, and abdomino-pelvic tumours using meglumine gadoterate: Pooled phase III studies. The Internet Journal of Radiology. 2010; 13: 2.
[9] Zhou Z, Lu ZR. Gadolinium-based contrast agents for MR cancer imaging. Wiley Interdiscip Rev Nanomed Nanobiotechnol. 2013; 5: 1–18.
[10] Fenchel M, Nael K, Seeger A, Kramer U, Saleh R, Miller S. Whole-body magnetic resonance angiography at 3.0 Tesla. Eur Radiol. 2008; 18: 1473–1483.
[11] Batz M, Nacher PJ, Tastevin G. Fundamentals of metastability exchange optical pumping in helium. Journal of Physics: Conference Series. 2011; 294: 012002.
[12] Goertz S, Meyer W, Reicherz G. Polarized H, D and He-3 targets for particle physics experiments. Prog Nucl Phys. 2002; 49: 403.
[13] Kauczor HU, Ebert M, Kreitner KF, Nilgens H, Surkau R, Heil W, Hofmann D, Otten EW, Thelen M. Imaging of the lungs using 3He MRI: preliminary clinical experience in 18 patients with and without lung disease. J Magn Reson Imaging. 1997; 7: 538–543.
[14] Fain S, Schiebler ML, McCormack DG, Parraga G. Imaging of lung function using hyperpolarized helium-3 magnetic resonance imaging: Review of current and emerging translational methods and applications. J Magn Reson Imaging. 2010; 32: 1398–1408.
[15] Lilburn DML, Pavlovskaya GE, Meersmann T. Perspectives of hyperpolarized noble gas MRI beyond ^{3}He. J Magn Resonance. 2013; 229: 173–186.
[16] Serrao EM, Brindle KM. Potential clinical roles for metabolic imaging with hyperpolarized [1-^{13}C]pyruvate. Frontiers in Oncology. 2016; 6: 59.
[17] Hoffmann SH, Begovatz P, Nagel AM, Umathum R, Schommer K, Bachert P, Bock M. A measurement setup for direct ^{17}O MRI at 7 T. Magnetic Resonance in Medicine. 2011; 66: 1109–1115.
[18] Tirotta I, Dichiarante V, Pigliacelli C, Cavallo G, Terraneo G, Baldelli Bombelli F, Metrangolo P, Resnati G. ^{19}F Magnetic Resonance Imaging (MRI): from design of materials to clinical applications. Chemical Reviews, Special Issue 2015, Fluorine Chemistry. 2014: 1106.
[19] Winston GP. The physical and biological basis of quantitative parameters derived from diffusion MRI. Quant Imaging Med Surg. 2012; 2: 254–265.
[20] Szaflarski JP, Holland SK, Schmithorst VJ, Byars AW. An fMRI study of language lateralization in children and adults. Hum Brain Mapp. 2006; 27: 202–212.
[21] Papageorgiou TF, Christopoulos GI, Smirnakis SM, editors. Advanced brain neuroimaging topics in health and disease – methods and applications. InTech; 2014. Open access available at: www.intechopen.com/books/advanced-brain-neuroimaging-topics-in-health-and-disease-methods-and-applications
[22] Aguirre GK, Komaromy AM, Cideciyan AV, Brainard DH, Aleman TS, Roman AJ, Avants BB, Gee JC, Korczykowski M, Hauswirth WW, Acland GM, Aguirre GD, Jacobson SG. Canine and human visual cortex intact and responsive despite early retinal blindness from RPE65 mutation. PLoS Medicine. 2007; 4: e230.
[23] www.biomednmr.mpg.de/index.php
[24] Inglisa B, Buckenmaier K, SanGiorgio P, Pedersen AF, Nichols MA, Clarke J. MRI of the human brain at 130 microtesla. PNAS. 2013; 110: 19194–19201.

Further reading

Liang ZP, Lauterbur PC. Principles of Magnetic Resonance Imaging: A signal processing perspective. Wiley-IEEE Press; October 1999.
Hornak JP. The basics of NMR. Available online at: www.cis.rit.edu/htbooks/nmr/bnmr.htm
Westbrook C. Handbook of MRI technique. 3rd edition. Wiley-Blackwell; 2008.
Bushberg JT, Seibert JA, Leidholdt EM Jr, Boone JM. The essential physics of medical imaging. 3rd edition. Lippincott Williams & Wilkins, Wolter Kluwer; 2012.

Useful websites

Image gallery of MRI slices using a variety of parameters: http://health.siemens.com/mr/image-gallery/#/search/all/
Image gallery of various imaging modalities: https://openi.nlm.nih.gov/
Questions and answers on MRI: http://mriquestions.com/index.html

16 Questions & answers

Questions

Chapter 1

1. What is the function of cell membranes?
2. What are cell membranes made of?
3. How many cells are in the body?
4. How do cells organize themselves?
5. What is the difference between organs and systems?
6. Body temperature is controlled by a negative feedback system. Which other negative or positive feedback systems are operational in the body?
7. What keeps the circulatory system going?
8. Do the lungs and the heart work synchronously?
9. The heart pumps about 6 l/min blood through the left ventricle. How much blood does the heart pump through right ventricle?
10. Why is oxygen-rich blood more reddish than oxygen-poor blood?
11. How many muscles are in the heart?
12. What are the three main tasks of the kidneys?
13. When does blood have a blueish color and when a more reddish color?
14. Why is blood red?
15. What is the purpose of the lungs?
16. Does the expiration of carbon dioxide hinder the uptake of oxygen during inhalation?
17. Why do we have two different nervous systems, efferent and afferent?
18. Which organs belong to the digestive system?
19. Which body system removes solid waste and which removes liquid waste?
20. How many sensors does the body contain? List at least six.
21. Where does the information processing of the sensors take place?
22. How does the receptor information of the sensors get to the processing location?
23. If muscles can only contract, how can we move our extremities back and forth?
24. What are the two main pigments in the skin, and how do they determine the color of the skin?
25. The skin is the largest organ of the body. What functions does it have?
26. Is ejaculation a sympathetic or parasympathetic reflex coordinated by the lumbarportion of the spinal cord?
27. Where are the ovaries of the female reproductive system located, inside or outside of the pelvic cavity?

DOI 10.1515/9783110372830-018

Chapter 2

1. The density of the body is an important physical parameter. Why? What depends on the density?
2. Why do bones have a higher density than muscles and fatty tissue?
3. The coronal plane divides the body into which parts?
4. The transverse plane divides the body in which parts?
5. Discuss the movement of the center of mass of a long jumper. What trajectory does it follow?
6. Why do most levers in the body have a mechanical advantage lower than 1?
7. Identify levers in the body that have a mechanical advantage higher than 1.
8. It was determined that the maximum force that the biceps can exert is about 2600 N. Why is it then not possible to lift a weight of 2600 N with one hand?
9. What is the reason for a permanent plastic deformation of the femur that changes the CCD angle and increases the torque?
10. Give examples for ball and socket joints in the body.
11. How can the joint of the knee be characterized?
12. How many phases can you identify for normal walking?
13. Name the antagonistic muscles which are responsible for bending the knee.
14. How many hierarchy levels does a muscle have? Name them.
15. On which level does muscle contraction take place?
16. How is muscle contraction achieved?
17. Does the force that a muscle can produce depend on its length or on its cross section?
18. What is the difference between isometric and isotonic tension?
19. How are muscles activated to contract?
20. Which ion exchange is important for contraction? What is required for release?
21. What do you understand a twitch to be?
22. What is a fused or unfused tetanus?
23. Is the formation of a tetanus possible for the myocardium?

Chapter 3

1. What do you understand by Hooke's law?
2. How can a hydrostatic pressure be exerted on a body?
3. How can elastic behavior be distinguished from plastic behavior?
4. What is the difference between ductile and brittle materials?
5. Which materials are usually ductile, and which ones are usually brittle?
6. How can you recognize whether a material is ductile or brittle?
7. What does toughness mean? When is a material tough, and when is it not?
8. What type of fractures occurs mostly in bones?

9. What is the basic structural unit of bones?
10. Describe the hierarchical structure of bones.
11. Which antagonistic cells are present in bones? What are their tasks?
12. What is the microscopic reason for osteoporosis?
13. What makes bones lightweight but strong?
14. What is missing in the spongy part of the bone which is present in the cortical part of bones?
15. Explain the term "viscoelasticity"?
16. Why do bones show a viscoelastic behavior?
17. Why is the stress-strain curve different for cortical and for trabecular bone structure?
18. What is the microscopic mechanism of plastic deformation in bones?

Chapter 4

1. How is BMR defined?
2. What is the difference between BMR, RMR, and MHR?
3. How is the caloric oxygen equivalent defined?
4. What is the daily BMR for an average person?
5. How many conventional and unconventional types of heat loss are important for the body to control the body temperature?
6. When does Newton's law for heat conductivity apply?
7. How is energy produced during anaerobic and aerobic exercise?
8. Is the homeostatic control of the body temperature a positive or negative feedback system?

Chapter 5

1. What does the cell membrane consist of?
2. Which ion channels can open and close?
3. Which ion channels are always open?
4. Which cation is predominately in the cytoplasm, and which one is in the extracellular space?
5. What determines the resting potential?
6. What is the sign and value of the resting potential?
7. What causes an action potential?
8. How long does an action potential typically last before it returns to the resting potential?
9. How is the strength of a stimulus translated into action potentials?

10. Explain the working principle of an ATP pump? What is pumped, and how is this achieved?
11. Which is correct: the ATP pump supports active ion transport, or it supports passive ion transport?

Chapter 6

1. What are the four main parts of a neuron?
2. What is the difference between efferent and afferent nerve fibers?
3. Why are some axons myelinated and others are not?
4. How are the three main receptor neurons distinguished?
5. What are the main differences between receptor and action potentials?
6. How does the analog–digital conversion (ADC) from receptor potential to action potential work?
7. Is there also a DAC in neuron signal transmission?
8. What is the difference between automatic neuron reflexes as compared to voluntary neuron action?
9. What can be tested with EMG, EEG, and MEG?
10. Describe how an action potential crosses a synaptic cleft.
11. What makes action potentials move in only one direction?
12. What do you understand saltatory conduction to mean?
13. What is the difference between electric conduction of a metal wire and a polarization current?

Chapter 7

1. What do you understand by an absolute refractory time and a relative refractory time?
2. What is the role of Ca^{2+} ions for the cardiac action potential?
3. How is the action potential of the SA node different from that of the myocardium?
4. What keeps the heart beating, and what is the natural pacemaker?
5. How is the action potential distributed over the myocardium?
6. How is the Eindhoven triangle arranged?
7. What is the purpose of the Eindhoven triangle?
8. What is the difference between the Eindhoven leads and the leads according to Goldberger?
9. Name the three characteristic intervals of the heart cycle.
10. Describe the potential variations measured at lead I with respect to the cardiac cycle.

11. If lead II reads positive values but aVL reads negative during the QRS complex, which orientation does the heart axis have?
12. What is recorded during an ECG examination?
13. What type of artificial pacemakers are available?
14. When should a defibrillator be used?

Chapter 8

1. The circulatory system can be subdivided In which main blood circulations?
2. What is the main task of the circulatory system?
3. How can veins and arteries be distinguished?
4. Can veins and arteries be distinguished by their oxygen level?
5. How many valves does a heart contain?
6. How can the four main phases of the heart cycle be characterized?
7. What is the ejection fraction in percent?
8. What is cardiac output?
9. What is the total power consumption of the heart and its efficiency?
10. How big is the pressure difference between the arterial and venous side of the circulatory system?
11. Which blood vessels have the higher compliance, veins or arteries?
12. Is the circulatory system a good example of classical hydrodynamics? If not, why?
13. How can the rather low total peripheral flow resistance be understood if the capillaries deliver a high flow resistance?
14. What is the purpose of the windkessel?
15. Which information can be gained from the pulse wave velocity?
16. Why is it important to keep an isotonic osmotic pressure in blood?
17. What is the main structure of hemoglobin?
18. How does the high-spin low-spin transition come about, and what is it important for?
19. Why do myglobin and hemoglobin have different oxygen binding curves?
20. What is the task of ferritin?
21. What are the main absorption bands of hemoglobin?
22. Which spectral region determines the color of blood?

Chapter 9

1. What are the three main organs involved in respiration?
2. Where does the gas exchange take place?
3. Which gases are exchanged?

4. What is the function of the pleural cavity?
5. What happens if the pleural cavity is ruptured?
6. What is more important for oxygen transport in blood, the physical solution or the chemical binding?
7. How is CO_2 transported in blood?
8. The vital capacity of the lung is composed of which lung volumes?
9. How big is the tidal volume?
10. What is the breathing frequency at rest?
11. The pressure in the alveoli varies sinusoidally during inhalation and expiration. What is the pressure amplitude?
12. Why is the volume work of the lungs during breathing negligible?
13. How is the compliance of the lung/thorax affected by diseases?
14. Why are surfactants of the alveoli important?
15. What is the main contribution to the airway flow resistance during respiration?
16. What is the task of a cardiopulmonary bypass.

Chapter 10

1. What are the main tasks of the kidney?
2. What is the most important unit in kidneys?
3. Which part of the nephrons is responsible for filtration?
4. Which part of the blood is filtered, plasma or hematocrit?
5. What is the difference between the renal plasma flow rate and the glomerulus filtration rate?
6. Describe the different parts of a nephron.
7. Which feature of the glomerulus helps the process of filtration?
8. Name the five components of unfiltered blood which appear in the glomerular filtrate.
9. Why do blood cells and protein molecules not appear in the glomerular filtrate?
10. Which three components of the glomerular filtrate are reabsorbed?
11. Why is it important for them to be reabsorbed?
12. Which substances are present in the final urine?
13. What is the fractional excretion of plasma, and what is the total excretion of plasma per day?
14. Which physical processes take place for secretion and reabsorption in nephrons?

Chapter 11

1. What does the optical system of the eye consist of?
2. What is the reason for a gradual loss of accommodation capability?

3. How can intraocular pressure be determined?
4. How can myopia be corrected?
5. What is the difference between cataracts and glaucoma?
6. How many different kinds of cells are in the retina? Name them.
7. What is the function of Müller cells?
8. Why is the retinal epithelium layer located on the distal side of the retina?
9. What is the difference between fovea and macula?
10. How many visual systems are in the retina, and how are they supported?
11. What is the Kohlrausch kink, and what is it an indication for?
12. What is the dynamical width of the visual system from bright to dark?
13. What is the spectral band width of all receptors?
14. Where on the retina is visual acuity the highest?
15. Which protein is responsible for visual sensitivity?
16. Which primary process takes place when a photon is absorbed, and what is the time scale of this process?
17. Does light cause an action potential in rods and cones?
18. Where is the action potential generatated which goes to the visual cortex?
19. What is the difference between a CCD chip and the retina?
20. Which cells are activated in the dark state? Which ones in the bright ON state?
21. What is the main working principle of the receptive field?
22. How is contrast enhancement achieved?

Chapter 12

1. Which three pieces of physical information does a soundwave carry?
2. Describe the main parts of the ear as a sound detector.
3. How is the acoustic impedance defined?
4. Which impedance is higher, water or air?
5. When do soundwaves totally reflect at an interface to another medium?
6. How big is the acoustic mismatch between air and water?
7. What is the unit for the auditory level?
8. What is the difference between auditory level and loudness level?
9. What is the unit for the loudness level?
10. Isophones are equal loudness curves. Do isophones depend on the frequency?
11. What are the names of the three ossicles in the middle ear?
12. By what action is the acoustic mismatch overcome?
13. What is the amplification factor of the middle ear?
14. What are the three main tasks of the middle ear?
15. What is the task of the oval window?
16. What is the task of the round window?
17. The basilar membrane is a center line in the cochlea. What does it separate?

18. Along the basilar membrane, where are the high and low frequencies detected?
19. Frequency detection is achieved by a spatial separation. What is the technical term for such a separation?
20. The organ of Corti contains which two types of hair cells?
21. Are the Ca^{2+} channels of the inner hair cells potential gated or mechanically gated?
22. Are the K^{+} channels of the inner hair cells potential gated or mechanically gated?
23. What is the task of the outer hair cells?
24. What is the task of the inner hair cells?
25. How many inner hair cells are responsible for detection a frequency range from 20 to 20 000 Hz?
26. The outer and inner hair cells are connected to afferent and efferent nerve fibers. Which ones are connected to which nerve fibers?
27. What are the unique properties of the outer hair cells, and how is this achieved?
28. When do the inner and outer hair cells depolarize or hyperpolarize by leaning to their longer or shorter neighbors?
29. How can the receptor in the organ of Corti be characterized?
30. How is the frequency and the sound intensity encoded on the way up to the auditory cortex?
31. How is sound in the equatorial plane spatially localized?
32. What is the difference between tone, sound, and noise
33. Is age-dependent hearing loss more severe for high or for low frequencies?
34. Hearing aids can be distinguished according to their location. Name them.
35. Where does sound produced by the body originate from?

Chapter 13

1. In which fields of science and technology is ultrasound used?
2. What is ultrasound used for in medicine?
3. Which frequencies are typically used for US imaging?
4. What is the definition of acoustic impedance Z?
5. What is meant by the term “impedance mismatch”?
6. Soundwaves are attenuated in matter. What are the physical reasons?
7. At interfaces between tissue of different acoustic impedance the sound wave is either …?
8. What is required for receiving a strongly reflected echo signal?
9. What is a transducer?
10. How is ultrasound generated?
11. What are the essential parts of a transducer?
12. Why is a gel required between transducer head and skin?

13. Why is ultrasound transmitted into tissue in pulsed form instead of continuously?
14. What is time gain compensation?
15. What distinguishes the near field region and the far field region of an extended sound wave source?
16. Is the focusing of ultrasound achieved in the near field or far field region?
17. How is focusing achieved?
18. The echo time, i.e. the time between emission of a sound wave and receiving the echo signal, is proportional to the distance between transducer and interface. What determines the depth or axial resolution of the reflecting interface?
19. What is a B-scan?
20. What type of scanners are used for B-scans?
21. What is the pulse repeat frequency and how should it be chosen?
22. What is a C-scan?
23. What kind of artefacts may occur during US imaging?
24. How can the fluid velocity of blood be determined?
25. Is the Doppler frequency shift recorded in pulse or continuous mode?
26. When imaging and flow velocity measurement has to be combined, what are the limiting considerations?
27. What are the boundary conditions for the pulse repeat frequency PRF when applying pulsed Doppler methods?
28. How are the resistance index RI and the pulsatile index PI defined?

Chapter 14

1. What is an endoscope in simple terms?
2. What are the main components of an endoscope?
3. Traditionally endoscopes use glass fibers for illumination and for image formation. Which optical boundary condition have to be considered?
4. What are the most common endoscopic applications in medicine?
5. What is the typical lateral resolution of an endoscope?
6. What is the probing depth of an endoscope?
7. How does a confocal endoscope work?
8. What is meant by an OCT endoscope?
9. How can the small intestines be examined?
10. What is a tethered endoscopic capsule?

Chapter 15

1. Which nuclei are suitable for MRI?
2. Which type of MRI is most prominent?
3. How is the Larmor frequency defined?
4. What is the typical precessional frequency range of protons in a magnetic field?
5. The nuclear spin precession is damped by which two independent processes?
6. What is the free induction decay FID good for?
7. When does FID occur?
8. How can magnetization be turned from the z direction to the in-plane xy direction?
9. Which relaxation time controls the return of the in-plane component M_{xy}?
10. Which relaxation time controls the magnetization component M_z?
11. What is meant by spin-echo?
12. What is the pulse sequence in a spin-echo detection?
13. At what time does a spin-echo occur, if the 180° pulse was applied at time t_0 after the 90° pulse?
14. What is the time sequence of RF pulses applied for T1 and T2 contrast?
15. What is the time sequence of RF pulses applied for PD contrast?
16. What is the idea of inversion recovery IR, and how is it applied?
17. Discuss why the inverse relaxation time 1/T1 should be on the same order as the resonance frequency f_0 for a maximum signal in NMR experiments.
18. Why does the magnetic induction B_z vanish in the rotating frame?
19. How is a slice in the body selected by MRI methods?
20. How are the X and the Y directions selected to define a voxel ΔX, ΔY, ΔZ in the body by MRI methods?
21. How is the field of view determined?
22. What is a K-map?
23. How is a real space image generated from the K-map?
24. When is the phase gradient applied during TR?
25. What are typical MRI field specifications?
26. How is the main magnetic field of 1 T to 3 T generated?
27. How are field gradients generated?
28. What is the typical power consumption during MRI scanning with a conventional MRI system?
29. What is the effect of a contrast enhancing agent?
30. What kind of ion do contrast agents always contain?
31. What is meant by hyperpolarization MRI?
32. What is the advantage of hMRI compared to conventional MRI? What is the main disadvantage?
33. What is meant by diffusion weighted imaging, and how is it realized?
34. What is the main application of diffusion weighted imaging?

35. What is the basic principle of functional MRI (fMRI)?
36. How can moving organs like lung and heart be imaged with MRI?
37. How many frames per second can be achieved by application of turbo MRI?
38. What are the new trends in MRI technology?
39. What are the three main hazards which need to be considered for the application of MRI scanning?

Answers

Chapter 1

1. The main function is to create an environment inside, which is different from the outside. Cell membranes contain ion channels that control the passage of ions and small molecules from cytoplasm to the extracellular space and vice versa.
2. Double lipid layers.
3. About 60 trillion (60×10^{12}).
4. Cells organize themselves to form tissues, tissues form organs, and organs are assembled to systems.
5. An organ has a specific function such as the lung. Several organs need to work together in order to fulfill a certain body task, such as the respiratory system for the gas exchange.
6. Negative feedback system: temperature control; filtration rate in the kidneys; water level and pH-value. Positive feedback system: oxygen uptake in hemoglobin, contraction of outer hair cells in the cochlea.
7. The heart acts as a pump for the circulatory system, activated by self-excitation.
8. No. Heart and lungs work independently and have different rhythms.
9. The heart pumps through the right ventricle the exact same amount.
10. Because in oxygen rich blood the red color is less absorbed and more scattered than in oxygen-poor blood.
11. The heart contains just one muscle, the myocardium.
12. (1) Filtering of blood and removal of toxic substances; (2) control of ion concentrations and the pH-value; (3) control of water balance, i.e. the osmotic pressure.
13. Deoxygenated blood is blueish, oxygenated blood is reddish.
14. Blood has low absorption in the wavelength band from 600 to 700 nm, and the scattered light is perceived as red.
15. The lung has the purpose of gas exchange, inhaling oxygen, and exhaling carbon dioxide.
16. Yes. The oxygen concentration during inspiration is not as high as it could be if carbon dioxide were not present.
17. Afferent nerve fibers transmit information from the periphery to the CNS; efferent nerve fibers are required to move body parts after processing in the brain or spinal cord.
18. The digestive system encompasses the following organs: oral cavity including teeth and tongue, esophagus, stomach, liver, spleen, gallbladder, pancreas, intestines, colon, and rectum.
19. The small intestine together with the colon removes solid waste; the kidneys remove liquid waste.

20. The body contains sensors for light, sound, temperature, pressure, balance, smell, and taste.
21. Information processing takes place in the central nervous system, i.e. the brain and the spinal cord.
22. The information of the receptors is transmitted to the CNS via nerve fibers.
23. Each extremity has two antagonistic muscles, one for moving up, the other one for moving down, such as the biceps and triceps of the forearm.
24. Melanin and blood (hemoglobin); melanin develops a brownish color under sunshine, blood can turn the skin color more reddish or more blueish.
25. The skin protects the organs, contains a number of sensors, controls the surface temperature, excretes waste products, and produces vitamins.
26. Ejaculation is a parasympathetic reflex.
27. Inside of the pelvic cavity.

Chapter 2

1. Density of a body in relation to the density of water determines swimming, floating, or sinking. Physiologically the proportion of fat can be estimated from the density.
2. Because of a higher calcium content.
3. Anterior and posterior, or ventral and dorsal.
4. Superior and inferior, or cranial and caudal.
5. A parabolic trajectory, like throwing a ball.
6. Because levers in the body are mainly constructed for mobility and agility but less so for lifting heavy weight.
7. Only the foot is a lever with mechanical advantage bigger than 1.
8. For lifting the biceps starts from a relaxed state. In the relaxed state the force that can be developed is much less than in the partially contracted state.
9. Bones are not a rigid system. Osteoclast cells can under load eliminate old bone structures and build new ones, which are better adapted.
10. Femur (hip), shoulder (humerus).
11. The knee is a modified hinge joint.
12. Two main phases: stance and swing phase, and up to 16 subphases.
13. Knee flexors: hamstrings, biceps femoris; knee extensor: quadriceps, rectus femoris.
14. Bundles → fibers → myofibrils → myofilaments.
15. On the sarcomere level.
16. By movement of the myosin filament against the actin filament.
17. On the cross section. The cross section increases with the number of fibers, each fiber contributing to the total force.

18. Isometric tension is muscle activation without changing the muscle length; isotonic tension is muscle contraction without change of tension.
19. Muscles are activated by somatic motor neurons.
20. Ca^{2+} is required for contraction, ATP for release.
21. A twitch is a short contraction of muscle fibers activated by just one motor unit in response to one action potential arriving at a somatic motor neuron.
22. Low frequency twitches, which do not overlap in time, cause an unfused tetanus. High frequency twitches, which overlap in time, cause a fused tetanus.
23. No, because the refractory time is rather long and does not allow a temporal overlap of two muscle contractions.

Chapter 3

1. Linearity between strain and stress; no plastic deformation.
2. For hydrostatic pressure a medium is required to transmit the same pressure to all surfaces. The medium can be gas or a liquid like water. Therefore the term "hydrostatic pressure".
3. Elastic and plastic behavior can be distinguished by their hysteresis. In an elastic response there is no opening of a hysteresis. For plastic behavior the hysteresis encloses an area that corresponds to the energy required to cause the plastic deformation.
4. Ductile materials deform but do not break. Brittle materials break without deformation.
5. Metals are usually ductile, ceramic materials are usually brittle.
6. By bending. A ductile material deforms, a brittle material breaks.
7. Toughness implies material deformation beyond the yield point. A tough material accepts much plastic deformation until it breaks, a less tough material already breaks at a little beyond the yield point.
8. Mainly shear type of fracture.
9. Collagen and HA nanocrystals forming mineral reinforced fibrils.
10. Collagen → fibrils → osteons → compact bones
11. Osteoblast and osteoclast cells. Osteoblasts build new cells, osteoclasts eliminate old bone structures.
12. At a greater age the osteoclast cells may dominate over osteoblast cells, diminishing the mineral content of bones.
13. The open spongy structure. In long bones also the medullary cavity.
14. Osteons, Haver's canal, transverse Volkmann canals, nerve fibers.
15. Viscoelastic materials show elastic and viscous characteristics upon deformation. Part of the deformational energy is absorbed by viscous flow. The stress-strain curve depends on the rate of the applied stress.

16. Because bones are composite materials with hard nanocrystals embedded into softer fibrils, the latter ones can move against each other.
17. The compact cortical bone is strong but not tough. In contrast, the trabecular part of the bone is less strong, but much tougher. The difference is due to the compact versus open meshwork of fibrils in the bone structure.
18. Plastic deformation of bones is transformed into shear strain by cross links in the interfibrillar matrix; glue filaments between fibrils help absorbing strain energy suppressing fracture.

Chapter 4

1. Energy consumption per kilogram body weight per hour required for maintaining all functions of the inner organs without performing any physical work.
2. BMR is the energy consumption required for maintaining all functions of the inner organs at rest. RMR is the resting metabolic rate and is defined similarly to BMR. MHR is the sum of BMR and heat produced by physical activity.
3. 1 liter of oxygen is required for the production of 20 kJ of energy.
4. 8 MJ.
5. 3 conventional, 1 unconventional.
6. For small temperature differences.
7. Anaerobic: energy is produced by splitting creatine phosphate (CP) into creatine and phosphate and by conversion of glucose to lactate. Aerobic: energy is produced by normal metabolism, i.e. combustion of food with oxygen into its constituents carbon dioxide CO_2 and water H_2O.
8. Negative feedback system.

Chapter 5

1. Double lipid layer.
2. Na^+ channels.
3. K^+ channels.
4. In the cytoplasm: K^+; in the extracellular space: Na^+.
5. The resting potential is determined by the K^+ ion concentration difference between cytoplasm and extracellular space.
6. −75 mV to −90 mV.
7. Any kind of stimulus that surpasses the threshold potential.
8. 5 ms.
9. The strength of a stimulus is translated into the sequence and frequency of action potentials.

10. ATP is a rocking two-way ion channel for K^+ going into the cytoplasm and Na^+ going out. The ion channel is controlled by the molecule ATP containing three phosphate groups.
11. Correct is: active ion transport.

Chapter 6

1. Soma, dendrites, axon, axon terminal.
2. Afferent neurons are sensory neurons, conducting receptor signals to the CNS, efferent neurons are motor neurons conducing action potentials from the CNS to the periphery.
3. Myelinated axons support saltatory conduction of polarization current, speeding up the signal transmission in long nerve fibers. In the brain the nerve fibers are not myelinated because of the short distances and limited space.
4. They are distinguished by their response to external stimulus: temperature, pressure, or chemicals.
5. Receptor potential is graded, action potential is all or none.
6. The conversion takes place by a change from graded potentials to action potentials. Action potentials occur when the nerve fibers contain an increasing number of ion channels for fast depolarization.
7. Yes, at all chemical synapsis a DAC takes place via neurotransmitters.
8. Automatic neuron reflexes take place by connecting afferent and efferent neurons via interneurons. In voluntary neuron activity the afferent neuron is connected to the CNS (brain) for processing before an efferent neuron can cause any action.
9. With EMG muscle activity and conductivity is tested. With EEG the brain activity is tested. With MEG also the brain activity is tested but with magnetic fields instead of electric fields.
10. An action potential crosses a synaptic cleft by first converting the frequency of the action potential to a proportional amount of neurotransmitters. The neurotransmitter is converted back to an action potential by activating ligand gated ion channels.
11. The depolarized tail does not allow propagation into a zone that is already depolarized. Therefore propagation of action potentials can only proceed in the direction that is not yet depolarized.
12. Saltatory conduction is jump-like conduction of the polarization current along nerve fibers.
13. Conduction in metal wires follows an electric potential gradient carried by a viscous flow of charge (electrons). Polarization current starts locally by an action potential across a cell membrane and propagates laterally due to electric dipole fields that opens neighboring gated ion channels.

Chapter 7

1. Absolute refractory is the time during which the heart is insensitive to a new action potential. The relative refractory period allows depolarization, but only with an enhanced threshold potential.
2. Ca^{2+} ions prolong the plateau phase.
3. The action potential of the SA node has a lower resting potential, lower threshold potential, no plateau phase and repolarization that starts immediately after depolarization.
4. The natural pacemaker is the sinoatrial node.
5. The action potential is first distributed from the atrium to the AV node and from there along the highly conducting Purkinje fibers to the lower part of the myocardium.
6. The Eindhoven triangle has the base from right to left arm and the tip pointing down to the pelvic.
7. The purpose of the Eindhoven triangle is the measurement of potential differences between the extremities, which are representative for the cardiac action potential during a heart cycle.
8. Eindhoven leads provide potential differences between two extremities; Goldberger leads provide potential differences between one extremity and a neutral point.
9. P-wave, QRS-complex, T-wave.
10. During the P-wave the mitral valve is open. The P-wave causes a contraction of the atria that leads to a rapid filling of the ventricles. During the QRS complex the ventricle becomes depolarized, causing a contraction of the ventricle and an increase of pressure. At the end of the S-phase, ejection of blood into the artery takes place. During the T-wave the semilunar valves close, defining the diastole phase, which is the refractory period of the heart.
11. Left-oriented heart.
12. ECG examination records 12 potential, 3 according to Eindhoven, 3 according to Goldberger, and 6 according to Wilson.
13. Single lead and wireless pacemakers.
14. In case of a cardiac arrhythmia that causes cardiac fibrillation.

Chapter 8

1. Pulmonary circuit and systemic circuit.
2. Delivery of oxygen, deposition of carbon dioxide.
3. They can be distinguished according to their blood pressure. Veins have low pressure, arteries have high pressure. Accordingly the vessel walls are thin for

veins and thick for arteries. Furthermore all veins lead to the heart, all arteries lead away from the heart.
4. No. Veins from the capillary bed to the heart have a low oxygen concentration; veins from the lung to the heart have a high oxygen concentration.
5. Four, two for filling and two for ejection.
6. (1) Inflow, (2) contraction, (3) ejection, (4) relaxation.
7. 60 %.
8. 5–6 l/min.
9. 5.6 Watt, efficiency of 25 %.
10. 120 hPa.
11. Veins.
12. The circulatory system is not a good example for classical hydrodynamics. The reasons are: (1) blood is a non-Newton fluid; (2) vessels are not rigid; (3) the blood flow is pulsatile, and not continuous; (4) the flow can be turbulent and not laminar.
13. Because of Kirchhoff's second law of parallel resistances.
14. Suppression of turbulent flow.
15. Information on the cardiac performance.
16. Hypotonic RBC may rapture, hypertonic RBC cannot carry oxygen.
17. Hemoglobin consists of four side chains, each containing one heme molecule with Fe^{2+} at the center.
18. The high-low spin transition is due to Fe^{2+} sitting in a crystal electric field that changes with O_2 absorption. It is not important for oxygen transport, but it signifies high oxygen consumption in active parts of the brain during MRI imaging.
19. O_2 binding in myoglobin can be described by a simple Langmuir equation, whereas O_2 binding in hemoglobin has an S-shape. The S-shape signals a cooperative O_2 uptake with a positive feedback system.
20. Ferritin stores iron as a buffer.
21. Oxy-hem: 535 and 575 nm; De-oxy-hem: 560 nm.
22. The color of blood is determined by the difference in absorption for oxy- and deoxy-hemoglobin in the spectral range from 600 nm to 700 nm.

Chapter 9

1. Lung, thorax, diaphragm.
2. In the alveoli.
3. O_2 versus CO_2.
4. The pleural cavity maintains a pressure lower than atmospheric pressure, which couples the thorax to the lung.
5. The lung will collapse (pneumothorax).

6. Chemical binding.
7. By chemical reaction to bicarbonate.
8. The vital capacity is composed of expiratory reserve volume, tidal volume, and residual volume.
9. 0.5 l.
10. 15/min.
11. 150 Pa.
12. Because of the elastic properties of the lung and thorax acting like a combined elastic spring.
13. The compliance increases in case of emphysema, the compliance decreases in case of fibrosis.
14. Surfactants lower the surface tension and enable the alveoli to expand.
15. Turbulent flow in the trachea.
16. Cardiopulmonary bypass takes over the tasks of heart and lung during open-heart or lung surgery.

Chapter 10

1. Control of water level and electrolytes, removal of metabolites, in particular urine.
2. Nephron.
3. Glomerular capsule or Bowman's capsule.
4. Plasma.
5. The renal plasma flow rate is the total plasma flow through the kidneys, glomerulus filtration rate is that part of the RPF which is filtered, amounting to 20 % of PRF.
6. The nephron consists of Bowman's capsule including glomerulus, proximal tubule, loop of Henle, distal tubule, and collecting duct.
7. Sieve-like filter, allowing only small molecules to pass.
8. Water, urea, creatinine, salts, glucose.
9. Their size is too big.
10. Water, salts, glucose.
11. For balance of water and electrolytes controlling osmotic pressure.
12. Water, urea, creatinine.
13. Fractional excretion: 0.01; total excretion: 1.5 l.
14. Active secretion and resorption via ATP pump, diffusion, and active filtration via overpressure.

Chapter 11

1. Cornea, aqueous humor, lens, vitreous humor, retina.
2. A stiffening of the lens.
3. With a tonometer pressing against the cornea.
4. With a diverging lens.
5. Cataract is a disease of the lens, glaucoma is a disease of the retina.
6. Seven, cones, rods, horizontal, bipolar, amacrine, Müller, and ganglion cells.
7. Support of metabolism and recycling of neurotransmitters.
8. Because it is highly absorbing and a supply layer for cones and rods, which would not function on the proximal side of the retina.
9. Fovea is the center of the macula with the highest density of cones. Macula is the area on the retina with the highest sensitivity.
10. Two systems: photopic and scotopic. The photopic system is supported by cones, the scotopic system by rods.
11. The Kohlrausch kink occurs in a plot of light sensitivity versus time. It signifies the crossover from cone to rod sensitivity in the dark. It is an indication of two visual systems: scotopic and photopic.
12. 10 orders of magnitude in intensity.
13. 400–700 nm.
14. In the fovea.
15. Rhodopsin.
16. Photoisomerization of the retinal from cis to trans within 100 fs.
17. No, it leads to a graded potential, in fact to a hyperpolarization from –30 mV to –70 mV.
18. In ganglion cells.
19. The CCD chip takes a single response to intensity, the retina has a dual system that measures both, brightness and darkness.
20. In the OFF state: off-bipolar and off-ganglion cells. In the ON state the respective ON bipolar and ON ganglion cells.
21. Receptive field distinguishes between center and surround. The surround has an inhibitive effect on the center for contrast enhancement.
22. By horizontal connection of bipolar cells.

Chapter 12

1. Frequency, amplitude or loudness, and direction.
2. Outer ear canal as transducer of sound waves in air to mechanical oscillations of the ear drum, middle ear for impedance matching and amplification; the inner ear for mechanoelectrical transduction and generation of action potentials.

3. The acoustic impedance is defined as product of density and sound velocity in which the sound wave travels.
4. Water has an impedance higher by a factor of 3500 compared to air.
5. Whenever the other medium has an acoustic impedance that is much higher or much lower than the medium where the acoustic wave arrives from.
6. The mismatch is such that 99.9 % of the sound intensity is reflected at the interface.
7. Bel or decibel.
8. The auditory level is a physical measureable quantity derived from the sound intensity. The loudness level takes the individual perception of sound intensity into account.
9. Phone.
10. Yes they do. Only at 1000 Hz do the auditory level and the loudness level overlap.
11. Hammer, anvil, stapes
12. By hydraulics and levers.
13. About 600.
14. Acoustic impedance matching, transmittance of a wide frequency band, and protection against destructive noise levels.
15. Transmission of mechanical vibration into fluid flow.
16. The task of the round window is a pressure release window and enables the fluid in the cochlea to move back and forth.
17. The scala vestibule and the scala tympani.
18. High frequencies are detected at the base and low frequencies at the apex.
19. Tonotopic mapping of sound frequencies.
20. Inner hair cells and outer hair cells.
21. Potential gated.
22. Mechanically gated.
23. The outer hair cells amplify the fluid flow.
24. The inner hair cells are responsible for the tonotopic detection of sound frequencies.
25. 4000 inner hair cells.
26. The inner hair cells are connected to afferent nerve fibers, the outer hair cells to efferent nerve fibers.
27. The outer hair cells can stretch and contract with the frequency of the sound wave. This is achieved by prestin proteins that can change length.
28. Depolarization by leaning to the longer neighbors, hyperpolarization by bending towards the shorter neighbors.
29. The receptor is a mechanoelectric transducer transforming mechanical stimulus into receptor potentials.
30. The frequency is encoded tonotopically and the intensity is encoded by stimulating additional nerve fibers phase locked to the primary excitation.

31. There are two systems that work at different frequencies. At low frequencies localization is achieved by interaural time difference, at high frequencies by intensity level differences.
32. Tone is a soundwave of single frequency, sound may consist of several frequencies, noise refers to a continuous distribution of frequencies.
33. For high frequencies.
34. Ear canal hearing aids, middle ear implants, and cochlea implants.
35. From the glottis in the larynx being responsible for the frequency and the oral cavity for the color of the sound.

Chapter 13

1. Materials science, engineering, geosciences, bonding, surface cleaning.
2. Static and dynamic imaging, fragmentation of bladder stones, and lens fragmentation during cataract surgery.
3. 2–20 MHz.
4. Acoustic impedance is defined as the product of density and sound velocity.
5. Two materials with very different impedance values that share a common interface.
6. Viscous damping, thermal conduction, and scattering.
7. Reflected, transmitted, or scattered.
8. A large impedance mismatch at the interface between two different materials and normal incidence to a smooth and flat interface.
9. A transducer is a converter of one form of energy into another.
10. By the use of a piezoelectrical transducer.
11. Essential parts are a piezoelectrical crystal, a damping material in the back, and a quarter wave plate in the front.
12. For impedance matching, i.e. to avoid reflections at the air/skin interface.
13. Because imaging is achieved by detecting the echo signal, which is only possible when the US wave is pulsed.
14. TGC is an amplification of echo signals arriving later than the first echo compensating for the decayed echo amplitude.
15. In the near field region the wavefront is nearly a plane wave, in the far field region diffraction effects occur.
16. At the border between near field and far field.
17. By shaping the transducer, or using a plastic lens in front of the transducer, or by controlling the signal arrival time electronically.
18. The depth resolution depends on the pulse length of the emitted sound wave. At higher frequencies the pulse is shorter and therefore the resolution increases.
19. B-scan is brightness mode scan. It consists of a sequence of A-scans which combined images a slice of the body.

20. Sector scanner, array scanner, phase array scanner.
21. The PRF is the reciprocal time separation between two US pulses emitted. The minimum PRF is given by the depth of view. The larger the DOV, the lower the PRF.
22. A C-scan probes a plane at a certain depth taken from a sequence of B-scans.
23. Double reflection, shadowing effects, refraction effects.
24. By US scattering from erythrocytes, which act as moving receiver and emitter for sound waves that cause a frequency shift proportional to the speed, also called Doppler shift.
25. For high resolution velocity profile determination the probing sound source is continuous.
26. For good imaging short pulses are required, whereas high resolution velocity determination requires a continuous wave. The compromise are long pulses.
27. Imaging the depth of object sets an upper limited on PRF that depends on the reciprocal echo time. Determination of the velocity sets a lower limit on the PRF in order to avoid aliasing effects.
28. RI is defined by the difference of the peak systolic velocity and end diastolic velocity normalized by the peak systolic velocity. PI is defined by the same velocity difference, however normalized by the mean velocity.

Chapter 14

1. A flexible tube that shines light into dark hollow spaces and channels out an image from the illuminated area.
2. A white light source, a fiber glass bundle, and a receiving optics of the reflected light.
3. Total reflection at the inner walls of the glass fiber and constructive interference of the head and tail of a traveling light wave train.
4. Gastroscopy and colonoscopy.
5. 100 µm.
6. Depends on the wavelength and penetration of light into tissue, which can be as much as a few mm.
7. Confocal endoscopes are scanning endoscopes. A fine light spot is scanned via a mirror scanner across the field of interest.
8. Optical coherence tomography endoscope is an instrument that carries depth information by using interference of laser light.
9. By a capsule endoscope.
10. A capsule endoscope connected to a string containing optical fibers for data transfer that can be swallowed and pulled back out again.

Chapter 15

1. Light nuclei with an odd number of nucleons.
2. Proton based MRI.
3. The Larmor frequency is the precessional frequency of spins in an external magnetic field. In case of nuclei it is the product of the nuclear gyromagnetic ratio γ and the magnetic induction B.
4. About 50 to 150 MHz, depending on the field.
5. The precession is damped by the longitudinal relaxation time T1 and the transverse relaxation time T2. T1 is the spin lattice relaxation process, T2 is the spin-spin relaxation process.
6. FID is essential for recording an induced voltage in pick-up coils from precessing protons.
7. FID occurs only upon relaxation of in-plane magnetization M_{xy} to magnetization along the z direction M_z.
8. The magnetization can be turned by a resonating Larmor precession over a finite time that turns the spins by 90°.
9. The transverse relaxation time T2.
10. The longitudinal relaxation time T1.
11. Spin-echo implies that the dephasing of the m_{xy} spins and the decay of the M_{xy} magnetization is partially recovered by a refocusing 180° field pulse.
12. First a 90° pulse turning the spins in plane, followed by 180° pulse for spin reversal after a time span of t_0.
13. At $2t_0$.
14. For T1 short TR and short SE. For T2 long TR and long SE.
15. Long TR and short TE.
16. IR is applied when the T1 weighting procedure is not sufficient. Then first a 180° pulse is applied and spin echo is followed as soon as the fast system has passed the zero line of M_z and the relaxation of the slower system is at M_z.
17. Because if the relaxation time of the spin system matches the characteristic frequencies of the embedding system, the spin-lattice coupling is strongest, and the relaxation times are shortest.
18. Because in the rotating frame the Lamor frequency is zero and therefore also the field in the z direction.
19. Slice selection is achieved by magnetic field gradient in the z direction.
20. z direction by field gradient, x direction by a frequency gradient, and the y direction by a phase encoding gradient.
21. The FOV is mainly determined by the field gradient in the x direction and the bandwidth of the receiving coil.
22. A K-map is a two-dimensional map of with 256×256 pixels, each one containing the frequency in x direction and the phase in the y direction.
23. By Fourier transformation of the K-map.

24. Just before the 180° refocusing pulse.
25. 1.5 T and 3 T.
26. By a solenoid with superconducting wires.
27. Gradient in the *z* direction is generated by anti-Helmholtz coils, gradients in the *x* and *y* direction by half circle Golay coils.
28. About 20 kW.
29. CAs shorten the relaxation time, preferentially either T1 or T2.
30. They always contain a magnetic ion, such as Gd^{3+}.
31. In hyperpolarization MRI isotopes are used with an odd number of nuclei and unpaired nuclear spin that can be polarized externally before administering for imaging.
32. With polarized ^{3}He the respiration can be studied, with ^{17}O and ^{19}F metabolic processes and pharmakinetics can be investigated. The main disadvantage of ^{3}He and ^{17}O application is the rarity of the isotopes and therefore the high cost.
33. With diffusion weighted MRI the flow of liquids can be studied using two opposing short field gradients applied just before and just after the 180° refocusing pulse.
34. Investigations of neural activity, mainly in the brain.
35. Functional MRI is based on high-spin deoxyhemoblobin and low-spin oxyhemoglobin. In the high spin state the T1 relaxation time is shorter than in the low spin state. Active areas in the brain consume more oxygen than neighboring areas, which can be used for imaging active areas.
36. Moving organs can be imaged with the use of turbo MRI. Turbo MRI uses a sequence of 180° refocusing pulses together with phase encoding pulses within one TR.
37. About 20 frames/s.
38. Use of smaller scanners with conventional magnets or magnets with wires of high temperature superconductors. Use of new and more sensitive detectors compared to induction coils. Further development of multiparameter MRI, which could replace radiographic imaging with x-rays, γ-rays like PET and SPECT.
39. (1) The high magnetic fields applied exclude the use of any metals nearby. (2) The power consumption has to be adjusted not to heat up the body by more than one centigrade. (3) Contrast agents such as Gd-chelates may be incompatible for some patients.

List of acronyms used in this book

ADC	analog digital converter
ADC	apparent diffusional constant
ADP	adenosine diphosphate
ALARA	as low as reasonably achievable
amu	atomic mass units
ATP	adenosine triphosphate
AV	atrioventricular node
AVV	atrioventricular valve
BBB	blood-brain barrier
BF	breathing frequency
BMD	bone mineral density
BMR	basal metabolic rate
BMS	bare metal stents
BOLD fMRI	blood oxygen level dependent fMRI
BPS	biodegradable polymeric stents
BSA	beam-shaping assembly
BSA	body surface area
BW	body weight
c.m.	center of mass
CA	contrast agent
CAP	cardiac action potential
CBF	cerebral blood flow
CCC	continuous curvilinear capsulorhexis
CCD	caput-collum-diaphyseal angle
CCD	charge coupled device
CIRT	carbon ion radiation therapy
CLE	confocal laser endoscopy
CO	cardiac output
COE	caloric oxygen equivalent
CPA	charged particle activation
CPB	cardiopulmonary bypass
CS	Compton scattering
CSF	cerebrospinal fluid
CT	computed tomography
CTV	clinical Target Volume
cw	continuous wave
dB	decibel
DBT	digital breast tomosynthesis
DCE	dynamic contrast enhancement
DES	drug elusion stent
DNP	dynamic nuclear polarization
DOV	depth of view
DR	digital recording
DSA	digital subtraction angiography
DSB	double strand break
DTI	diffusion tensor imaging

DOI 10.1515/9783110372830-019

DTL	drift tube linac
DWI	diffusion weighted imaging
EBRT	external beam radiotherapy
ECC	extracorporal circulation
ECG	electrocardiography
EDP	end diastolic pressure
EDV	end diastolic volume
EEG	electroencephalography
EF	ejection fraction
EM	electromagnetic
EMG	electromyography
EPI	echo planar imaging
EPP	end plate potential
EPR	enhanced permeation and retention
ERBT	external radiation beam therapy
ERV	expiratory rest volume
ESP	end systolic pressure
ESV	end systolic volume
ETL	echo train length
FCRM	fiber optic confocal reflectance microscope
FE	fractional excretion
FF	flattening filter
FF	filtration fraction
FFDM	full-field digital mammography
FFF	flattening filter free
FFT	fast Fourier transform
FID	free induction decay
fMRI	functional magnetic resonance imaging
FOV	field of view
FRC	fractional rest volume
FSE	fast spin echo
GFR	glomerular filtration rate
GTV	gross tumor volume
Hct	hematocrit value
HD	hydrodynamic diameter
HDR	high dose rate
HEP	hemi-endoprosthesis
hMRI	hyperpolarization magnetic resonance imaging
HSLS	high spin - low spin
IAEA	International Atomic Energy Agency
ICRP	International Commission for Radiological Protection
IGRT	image guided radiotherapy
IHC	inner hair cell
IMRT	intensity modulated radiation therapy
IRV	inspiratory rest volume
ITD	interaural time difference
Kerma	kinetic energy release in matter
kHz	kilohertz
kV	kilovolt

kVp	peak kilovoltage
kW	kilowatt
LASEK	laser epithelial keratomileusis
LASER	light amplification by stimulated emission of radiation
LASIK	laser-assisted interstitial keratomileusis
LCI	low coherence interferometry
LDR	low dose rate
LED	light emitting device
LET	linear energy transfer
Linac	linear accelerator
LSO	lateral superior olive
MEG	magnetoencephalography
MEI	middle ear implants
MET	mechanoelectric transduction
MeV	mega-electronvolt
MHR	metabolic heat production
MHT	magnetic hyperthermia
MHz	megahertz
MI	mechanical index
MLC	multileaf collimator
MNP	magnetic nanoparticle
mpMRI	multiparameter MRI
MRI	magnetic resonance imaging
MRSI	magnetic resonance spectroscopy imaging
MRT	magnetic resonance tomography
MSFP	mean systemic filling pressure
MSO	medial superior olive
MUAP	motor unit action potential
MV	minute ventilation
NBI	narrow band imaging
NIR	near-infrared
NMJ	neuromuscular junction
NP	nanoparticle
NRT	neutron radiation treatment
NTD	nontarget dose
OCT	optical coherence tomography
OER	oxygen enhancement ratio
OHC	outer hair cell
PAH	para-aminohippuric acid
PCI	percutaneous coronary intervention
PCI	phase contrast imaging
PCV	packed cell volume
PD	proton density
PDR	proliferative diabetic retinopathy
PDR	pulse dose rate
PDT	photodynamic therapy
PE	photoelectric effect
PES	photoelectron emission spectroscopy
PET	positron emission tomography

PHIP	parahydrogen induced polarization
PI	pulsatility index
PPD	percentage photon dose
PRF	pulse repeat frequency
PRK	photorefractive keratectomy
PRP	panretinal photocoagulation
PRT	proton radiotherapy
PRT	pulse repeat time
PTV	planning target volume
PWV	pulse wave velocity
PZT	$PbTiO_3$
Q	quality factor
RBC	red blood cell
RBE	relative biological effectiveness
RBF	renal blood flow
RC	respiratory coefficient
Re	Reynolds number
RES	reticuloendothelial system
RF	radio frequency
RF	respiratory fraction
RMR	resting metabolic rate
RPF	renal plasma flow
RT	radiotherapy
RTR	real-time radiography
RV	residual volume
SA	sinoatrial node
SAD	source to axis distance
SATP	standard ambient temperature and pressure
SAXS	small angle x-ray scattering
SCI	spinal cord injury
SE	spin echo
SERS	surface enhanced Raman scattering
SLAC	Stanford linear accelerator
SNR	signal-to-noise ratio
SOBP	spread-out Bragg peak
SPE	single photon emission
SPECT	single photon emission computed tomography
SPIO	superparamagnetic iron oxide
SPR	surface plasmon resonance
SSB	single strand break
SSD	source-to-surface distance
SV	stroke volume
TE	time of (spin, acoustic) echo
TEP	total endoprosthesis
TGC	time gain compensation
THz	terahertz
TIPPB	transperineal interstitial permanent prostate brachytherapy
TLC	total lung capacity
TMR	targeted muscle reinnervation

TPFR	total peripheral flow resistance
TR	time of repetition
TT	transfer time
TV	tidal volume
UHMWPE	ultrahigh molecular weight polyethylene
ULFMRI	ultralow field magnetic resonance imaging
US	ultrasound
VC	vital capacity
VEGF	vascular endothelial growth factor
XRT	x-ray radiotherapy
YAG	yttrium-aluminum garnet
ZFC	zero field cooling

Index