CRC Press
Taylor & Francis Group
A GARLAND SCIENCE BOOK

The Human Genome in Health and Disease

A Story of Four Letters

Tore Samuelsson

The Human Genome in Health and Disease

A Story of Four Letters

To my parents

The Human Genome in Health and Disease

A Story of Four Letters

Tore Samuelsson

CRC Press
Taylor & Francis Group
A GARLAND SCIENCE BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-0-367-07633-7 (Hardback)
978-0-8153-4591-6 (Paperback)

Library of Congress Cataloging-in-Publication Data

Names: Samuelsson, Tore, 1951- author.
Title: The human genome in health and disease : a story of four letters / Tore Samuelsson.
Description: Boca Raton : Taylor & Francis, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2018040829| ISBN 9780815345916 (pbk. : alk. paper) | ISBN 9780367076337 (hardback : alk. paper) | ISBN 9780429021732 (e-ISBN)
Subjects: | MESH: Genome, Human | Disease--genetics
Classification: LCC QH447 | NLM QU 460 | DDC 611/.0181663--dc23
LC record available at https://lccn.loc.gov/2018040829

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

eResource material is available for this title at https://www.crcpress.com/9780367076337

Contents

Preface

The DNA molecule contains genetic information in the form of sequences of nucleotides abbreviated A, T, C, and G. Therefore, genetic information is essentially a very long sequence of these four letters. In a single human individual, the genetic content amounts to about 3 billion letters. Furthermore, information in DNA is used to direct the production of proteins and RNAs. Sequences of nucleotides in DNA and RNA, as well as sequences of amino acids in proteins, are all examples of *molecular sequences*. This book attempts to present an accessible account of the molecular sequence information contained in the human genome and how it is being used to direct the production of RNA and protein. It addresses the question of what important biological signals are found in the linear sequence of nucleotides in DNA and shows how specific DNA sequences have distinct functions. Such functional sequence elements, typically in a small size range like 3–20 nucleotides, may be classified into a number of different functional categories. Examples are the three-nucleotide sequences that specify amino acids, or short DNA sequences that are targets for proteins that regulate transcription.

To provide biological motivation, the importance of the DNA sequence for biological function is illustrated with a variety of inherited disorders or with other genetic disorders such as cancer. With such examples, different functional elements of a gene, as well as different aspects of genetic information transfer within the cell, are introduced. For instance, a point mutation in a coding sequence of a globin gene gives rise to sickle cell anemia, a splicing mutation results in a form of hemophilia, and a mutation in an untranslated region of an mRNA leads to an iron metabolism disorder.

When discussing the functional consequences of DNA sequence, it is also of interest to consider the molecular impact of mutations and sequence variation. For instance, changes in DNA sequence may affect the recognition by a nucleic acid, such as in the case of a codon being read by a tRNA anticodon. Alternatively, a DNA variant will affect the recognition by a protein, such as a protein regulating transcription. There are indeed throughout the book several examples showing in structural detail the interaction of a nucleic acid with another nucleic acid or with a protein, illustrating the consequences of mutations at the level of molecular interactions.

In essence, therefore, this book has a *molecular sequence perspective*, and recurring themes are *functional DNA sequence elements*, illustration of *functional impact* with genetic disorders, and *molecular interactions* affected by sequence variation.

Why is it important to know about the different DNA sequence elements of the human genome and their functional impact? There are many medical areas where such knowledge is critical, as shown in this book. Examples include pharmacology, gene therapy, and the diagnosis of rare inherited disorders and cancer.

The book is organized such that a major part (Chapters 6 through 12) discusses specific functional sequence elements of the human genome. The preceding chapters offer an introduction to basic concepts with regard to the human genome. Chapters 2 and 3 introduce human genes and their relationship to human disease using sickle cell anemia as an example. Proteins and their structure are introduced (Chapter 2), and the focus is then on DNA structure, the flow of genetic information from DNA to protein, as well as the basic principles of translating an mRNA using the genetic code (Chapter 3). Human genome sequencing and the organization and overall

content of the human genome are topics of Chapter 4. Individual genetic variation, the phenotypic effects of mutations, as well as cancer are then discussed in the following chapter.

Chapters 6 through 12 provide details about functional sequence elements of the genome and their role in the flow of genetic information. The importance of these elements is illustrated with inherited disorders or cancer. First, the role of the coding regions of mRNA molecules is discussed in Chapter 6. Chapter 7 covers repeat expansions in coding regions. Then the role of mRNA untranslated regions (Chapter 8) and the process of splicing to form mature mRNA (Chapter 9) is considered. The complex but important area of transcriptional control is then discussed in Chapter 10. The noncoding RNAs are the subject of Chapter 11. The analysis of DNA and protein sequences requires a substantial amount of computing. Therefore, in Chapter 12, computational methods used with biological sequences are briefly reviewed.

Chapters 13 through 15 provide additional biological motivation as they further illustrate why careful studies of the relationship between sequence and function are essential. These chapters deal with important medical applications of human genome sequencing. Thus, experimental methods to diagnose errors in DNA sequences are described, including successful efforts to reveal causative mutations in rare inherited disorders (Chapter 13). Furthermore, there is recent progress in the area of gene therapy. Some of the most important methods are covered, as well as a few cases of successful therapy for inherited disease (Chapter 14). Finally, some of the applications of human genome sequencing raise a number of ethical concerns, and these are discussed in the final chapter.

When it comes to level of difficulty, I have attempted to present the material at a basic level to make it understandable for a reader without previous studies of genetics and molecular biology. However, the reader will benefit from a basic knowledge about chemistry and biochemistry. Following the principle that a picture says more than a thousand words, the book is richly illustrated. Furthermore, a web supplement to the book that includes scientific updates and answers to selected chapter questions is available at http://toresamuelsson.se/hg.

Why did this book come about? We currently see a dramatic development in terms of human genome sequencing. Research laboratories around the world generate a wealth of genomic sequence data. Sequencing is becoming widely used in the clinic to analyze a variety of genetic disorders, including cancer. In addition, you can order your own genome sequence from a company ("direct-to-consumer" sequencing). With this wealth of genetic information, it becomes increasingly important to know what the human genome sequence is all about, how the sequence should be understood in terms of biological function, and how particular variants in the genome should be interpreted. I wanted to write a book providing an introduction to these topics. In the early days of my scientific career, I worked as a molecular biologist. Eventually, I turned to bioinformatics with a focus on molecular sequences. For a long time I have been intrigued by the digital nature of genetic information, that the genome sequence can be handled with computers as a long string of letters, and that computing with molecular sequences may be used to address a variety of biological problems. There are already very good books dealing with the human genome, but this book has a focus on molecular sequences and it examines in a systematic manner the functional role of DNA sequence elements as illustrated with human genetic disorders.

ACKNOWLEDGMENTS

Illustrations of the book were typically created with CorelDRAW version X4. All images of protein or nucleic acid structures were made with the UCSF Chimera package and structures available in the Protein Data Bank (www.rcsb.org; Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* 28:235–242). Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). Chimera is described in Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612. I am indebted to the KEGG/GenomeNet Support Team for permission to use a metabolic pathway image.

I am grateful to a number of people for their help in the context of this book. Jan Korbel, European Molecular Biology Laboratory, Heidelberg, and Evan Eichler, University of Washington, Seattle, helped out with comments about human structural variation. Johanna Rommens at the Hospital for Sick Children, Toronto, Canada, provided information about the Cystic Fibrosis Mutation Database, and Catherine Porcher, University of Oxford, United Kingdom, about the ELK/GATA sequence logo. Oxford Nanopore Technology allowed the use of a photograph of the MinION apparatus. Aravinda Chakravarti and Sumantra Chatterjee, Johns Hopkins University School of Medicine, Baltimore, Maryland, informed me on the role of transcription factor binding sites with regard to Hirschsprung disease. Eric Ottesen, Iowa State University, Ames, provided information regarding SMN2 exon 8 and nusinersen. I am grateful to Stefan Mundlos, Max Planck Institute for Molecular Genetics, Berlin, Germany, for providing clinical photographs in the context of developmental disorders resulting from genomic rearrangements in the EPHA4 locus. Sally Heywood, Cardiff University, United Kingdom, provided statistics of the collection of mutations in the Human Gene Mutation Database (HGMD). Mark Johnson, reporter at the Milwaukee *Journal Sentinel*, helped out with questions regarding the Nicholas Volker case. The National Center for Biotechnology Information (NCBI) User Services provided help regarding human single-nucleotide variants that are part of the NCBI dbSNP. I am grateful to Retta Beery for permission to use a photograph of her family. Niclas Juth, Karolinska Institute, Sweden, commented on the ethical topics of Chapter 15.

Colleagues at my department were also very helpful. Gunnar Hansson helped out with the parts of the book dealing with cystic fibrosis, and the noncoding RNA chapter was examined by Chandrasekhar Kanduri. I am also grateful to Per Elias for discussions on a variety of topics. Erik Larsson carefully examined specific parts of the manuscript and had very useful suggestions regarding the contents of Chapter 10.

A number of anonymous reviewers recruited by the publisher provided important comments to the manuscript. Furthermore, I am indebted to staff at Garland Science/CRC Press. Elizabeth Owen, Senior Editor at Garland Science, carefully read all of the text and offered very helpful comments in an early phase of the book writing. I am also grateful for the work and support of Developmental Editor Jordan Wearing and Senior Editor Chuck Crumly at CRC Press.

Finally, I'm very grateful for a stipend from The Royal Society of Arts and Sciences in Gothenburg, Sweden, that allowed me to work on this book during one month in 2017 at Hotel Chevillon, Grez-sur-Loing, France.

Tore Samuelsson

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1 Introduction

DNA, short for deoxyribonucleic acid, is a universal carrier of hereditary information. In all life forms—viruses, bacteria, fungi, plants, and animals—it carries important instructions for the design of the organism. And not only does it carry information—it is also a molecule designed so that it may be accurately copied to the next generation. DNA is built from simple units, referred to as nucleotides, that are joined to form very long molecules. Each nucleotide contains any of four different nitrogenous bases: adenine, thymine, cytosine, or guanine, abbreviated A, T, C, and G, respectively. It is the sequence of these bases that forms the actual genetic message. Thus, the information in DNA may be expressed as a long sequence of the letters A, T, C, and G—for an example, see **Figure 1.1**.

We refer to the complete genetic material of an organism as its **genome**. The human genome is an astounding three billion letters. An important milestone was reached in biomedical research in 2001 when, for the first time, a draft of the human genome was presented and the complete sequence of letters could be read. A small fraction of the human genome is shown in Figure 1.1. Consider the whole genome printed as a physical book. A total of 6,400 bases are in Figure 1.1. You would need in the order of 500,000 pages like this to cover the full human genome. That would correspond to more than 1,600 books, each with 300 pages. For more on printing the human genome on paper, see **Figure 1.2**.

The issues addressed by this textbook are related to the three billion letter sequence of the human genome. How are we to make sense of and understand this vast information? What different biological signals are contained in the DNA? How important are different regions of the sequence? Are some regions more important than others? What are the effects in the event the sequence of letters in DNA is changed? In molecular biology laboratories, scientists have carried out experiments to address these questions. In addition, as changes or mutations in DNA are natural components of evolution, nature has by itself carried out experiments during billions of years that may guide us in understanding the relationship between genetic information and biological function. For instance, mutations in DNA can

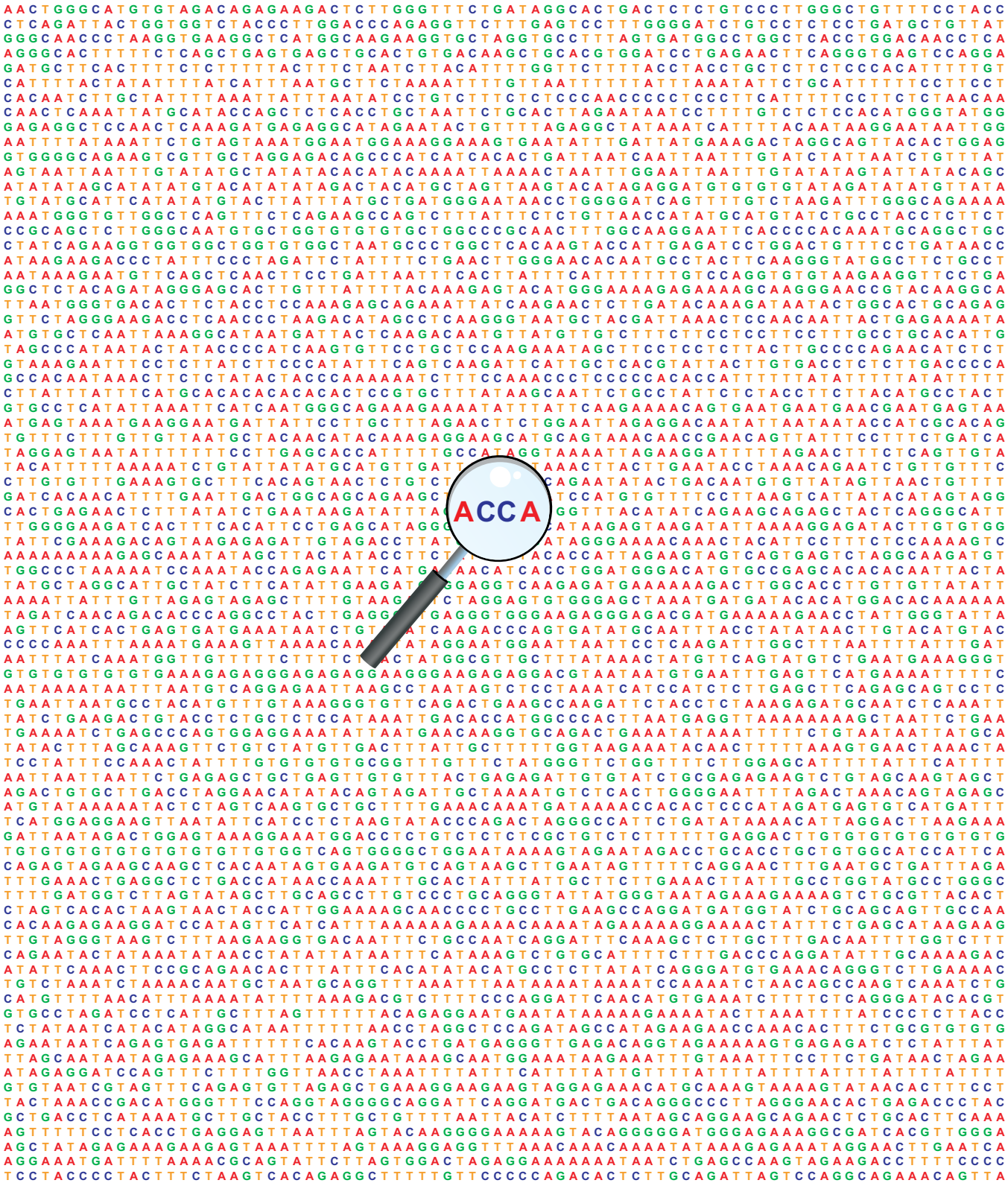

Figure 1.1 **Portion of the human genome.** Letters A, T, C, and G represent the DNA bases adenine, thymine, cytosine, and guanine, respectively. This page has 6,400 bases. It would take 500,000 pages like this to cover the full human genome. This would correspond to more than 1,600 books assuming one book contains 300 pages. The magnifying glass is to indicate that the object of research concerning the human genome is to associate sequence elements with biological functions. Scientists use experimental methods, as well as computational methods, to do this—a work in progress.

give rise to specific inherited diseases as well as cancer. What are the changes in the DNA sequence that cause such deleterious effects?

To answer these questions, we need to understand the organization of the human genome, as well as the different functional sequence elements in that genome. The flow of genetic information is crucial. Hence, DNA specifies what RNA molecules are to be made. One subclass of these RNAs is subject to processing to form messenger RNA (mRNA) molecules. These mRNA molecules in turn act as templates for the production of proteins. Another abundant class of RNA molecules has functions other than to specify proteins. Throughout the elaborate flow of genetic information that includes copying of DNA sequences to RNA, RNA processing, as well as the synthesis of protein using mRNA, specific nucleotide sequences have distinct functions.

In essence, this book explores the information in the human genome and all of the important biological signals that are present. It illustrates various functions of DNA sequences. Examples include protein coding sequences and sequences that regulate the flow of genetic information. For all of the different sequence elements, the relationship between sequence and function is illustrated with disorders of a genetic background.

Figure 1.2 Human genome printed on paper. Scientists at the University of Leicester printed the whole human genome on paper. It resulted in 130 book volumes that would take 95 years to read. (Published under CC BY-SA 2.0.)

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

A Molecular Disorder

2

The theme of this book is the information contained within the human genome as outlined in Chapter 1. As a first element of information, we consider regions in the genome that specify the proteins to be made. Proteins are molecules built from amino acids, and the sequence of amino acids is determined by the sequence of nucleotides in the genome. Regions specifying proteins make up only a minute portion of the entire genome but are nevertheless significant.

We first consider how amino acid sequences are related to inherited disorders. As an example, we discuss the disorder **sickle cell anemia**. It is caused by a mutation that gives rise to a replacement of the amino acid glutamic acid to valine in the protein hemoglobin. There are multiple reasons why we discuss this particular disorder in some detail in the first book chapters. Early studies of sickle cell anemia were based on nonmolecular clinical observations, and it seemed likely that the disease is inherited. But as research progressed, we eventually obtained a detailed molecular understanding of sickle cell anemia from the changes in DNA to detailed structural information about the hemoglobin protein. Sickle cell anemia is historically significant as it was the first disease to be characterized where a genetic change is associated with a well-defined change in a protein molecule. This finding gave an early clue as to the power of molecular medicine. In addition, very few inherited disorders have been so thoroughly examined as sickle cell anemia, and information about the disease is still being collected today. There is a significant medical impact from this knowledge, which is why we return to this disorder also in other contexts such as gene therapy (Chapter 14).

THE DISCOVERY OF SICKLE CELL ANEMIA

Abotutuo. Chwechweechwe. Nwiiwii. Nuiduidui. These are all names of a disease common in Western Africa—a disease we now know also as sickle cell anemia. Its history in Africa may be tracked as far back as the seventeenth century. Classification of the disease was difficult on this continent, because the symptoms were closely related to those of other diseases in tropical areas. It was not until the early twentieth century that sickle cell

anemia was first described in a medical publication. The affected individual was Walter Clement Noel.

Noel was born in 1884 on a large estate on Grenada. At this time, this island was a British colony. Noel was from a wealthy black family. He suffered from sickle cell anemia but was still able to attend school, and he completed his undergraduate studies in 1904. The same year, he sailed to New York. During this week-long journey, he developed a leg ulcer, a common complication of sickle cell anemia. When he arrived in New York, he immediately sought medical attention. His ulcer was treated with iodine—this chemical had the effect of killing the bacteria of the ulcer. Noel was thus cured and then traveled to Chicago, Illinois, where he was to study dentistry at the Chicago College of Dental Surgery. Noel was an unusual student at this time in the sense that it was uncommon for students of African descent to reach higher studies. In November 1904, his disease unfortunately got more severe as he developed respiratory problems. These problems persisted for more than 1 month. He finally sought medical attention at the Presbyterian Hospital in Chicago and was examined by an intern, Ernest Irons. Among other tests, Irons examined a blood sample from Noel. Under the microscope he could see that the sample contained, as he phrased it, "many pear-shaped and elongated forms—some small." Iron discussed these findings with his supervisor James Brian Herrick. Despite further investigations, Herrick and Irons could not reveal the cause of these unusual cells.

Figure 2.1 Sickle or crescent shape of red blood cells characteristic of sickle cell anemia. A sickle cell is shown (to the left) along with normal red blood cells. (Published under CC BY 3.0.)

Noel eventually recovered from his respiratory problems and returned to continue his studies at the dentistry school. However, he experienced additional illnesses during a period of more than two years of studies. Thus, he once was hospitalized for bronchitis and once for painful muscular crises and gallstones. He was then under the care of Irons who kept dutiful notes. When Irons was done with his training, he gave these notes to Herrick. Herrick then took care of Noel for two years.

Herrick was at a national meeting in 1910 and there presented the case of Noel—incidentally without giving credit to Irons. He published a report the same year. To describe the blood cells of the disease, Herrick used the term "sickle shaped cells" (**Figure 2.1**).

Despite his illnesses, Noel graduated from dental school. He then went back to Grenada where he set up a general dentistry practice in St. George's, the capital of Grenada. Not much is known of him from then on. However, in 1916, he overexerted himself. He attended a horse race on Grenada a long way from home and traveled home, all on the same day. As a result of this, he developed a serious respiratory infection. Only a few weeks after the horse race, he died from pneumonia at 32 years old. Noel is buried in a churchyard with a view of the Caribbean Sea. He is next to his sister Jane, who also died young of respiratory problems.

Only three months after Herrick published the case of Noel in 1910, a second case of the same disorder was described. Blood samples from a 25-year-old woman, Ellen Anthony, a resident of Virginia, showed the same strange shape of red blood cells as was observed in Noel. As more cases were identified in the 1920s, it was noted that all individuals with the disease were of African origin. The disease was eventually to be named *sickle cell anemia* (SCA).

A RECESSIVE INHERITED DISORDER

Significant advances in understanding sickle cell anemia were made by the end of the 1940s—both with respect to genetic inheritance and as to the molecular basis of the disease. Already in 1923, John Huck studied families with sickle cell anemia and noted that the disease was probably inherited, although his studies did not provide any firm evidence of this theory. Studies were complicated

by the fact that some individuals had symptoms related to sickle cell anemia but had a much milder form of the disease that did not shorten their lives. These individuals were said to have the **sickle cell trait**. It was often difficult to distinguish between these two categories of patients. But in 1949, James Neel carried out a careful examination of families affected by sickle cell anemia and was able to conclude that the disorder is indeed hereditary.

How is sickle cell anemia inherited? We turn to basic principles of genetics and inheritance, first elucidated by Gregor Mendel in the nineteenth century. The genetic makeup of an individual is referred to as the **genotype**, whereas the **phenotype** is the collection of observable characteristics. The phenotype is determined by the genotype and/or environmental factors. Human individuals—like all other animals—have two copies of each gene, one of paternal and one of maternal origin. A gene may have two or more variants—in the language of genetics, such variants are referred to as **alleles**. If two individuals have the same allele, they are said to be **homozygous** for that allele, and if they have two different alleles, they are **heterozygous**. From the perspective of an allele, one basic principle of inheritance is outlined in **Figure 2.2**. In this example, two different alleles "A" and "a" are considered, and the two parents are heterozygous, since they both have the allele configuration (genotype) Aa. During the formation of sperm and egg cells that occurs during **meiosis**, each cell ends up with any one of the two alleles. During fertilization, alleles are combined, and a child of the two parents may in this case have any of the genotypes AA, Aa, and aa.

For the discussion of sickle cell anemia, the two different alleles we consider are the sickle cell variant and the normal form not associated with disease. Individuals with two copies of the sickle gene develop sickle cell anemia, whereas patients with one copy of the sickle gene and one normal copy of the hemoglobin gene have milder symptoms and express the sickle cell trait. This is illustrated by the pedigree (family tree) with members affected by the disorder in **Figure 2.3**.

The inheritance of sickle cell anemia follows the rules of a **recessive** disorder. In such a disorder, two copies of the disease allele are required to develop the disease. The rules of inheritance also inform us on probabilities on inheriting disorders as explained in the diagram in **Figure 2.4**,

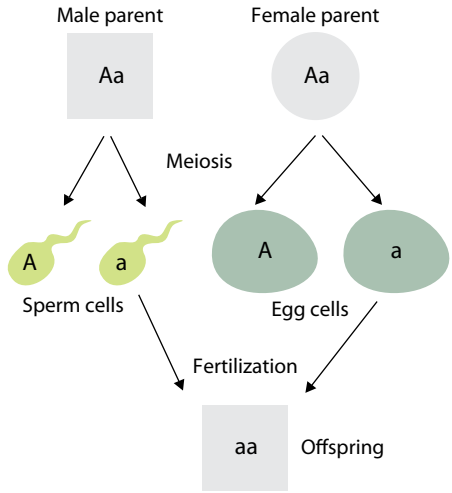

Figure 2.2 Inheritance of alleles. Two different alleles (gene variants) "A" and "a" are considered. In this example, both parents have the same setup of these alleles. The sperm and egg cells formed during meiosis have only one copy of each allele. The probability that a certain sperm or egg cell has a specific allele is about 50%. The fertilized egg will have two copies of each autosomal chromosome, and in the example shown here, "a" from the father is combined with the same allele from the mother. However, other outcomes are possible given the parent genotypes. Thus, offspring may be of three different genotypes: AA, Aa, or aa.

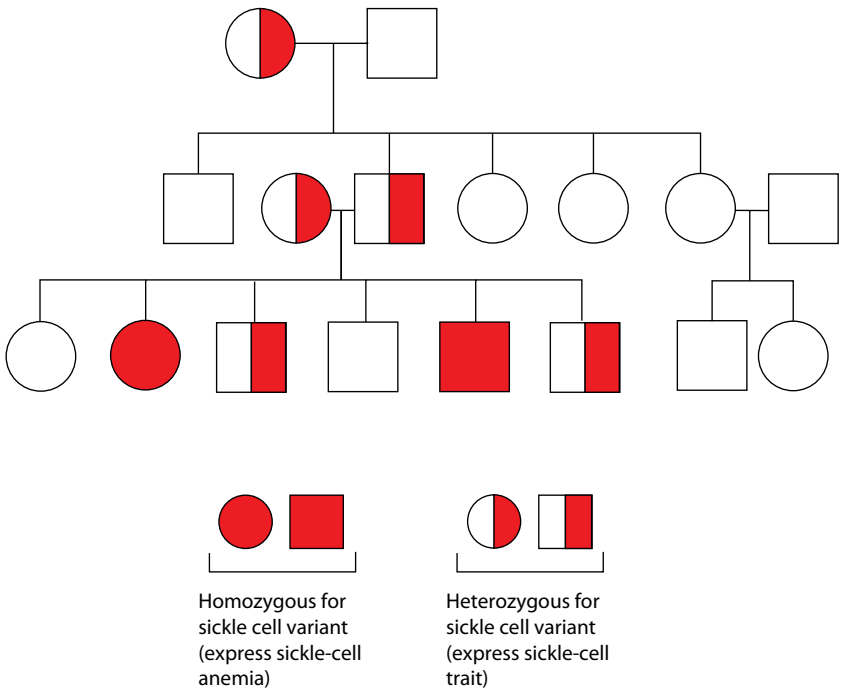

Figure 2.3 Sickle cell anemia is a recessive inherited disorder. The pedigree illustrates inheritance of sickle cell anemia. Males are represented by squares, females by circles. The disease gene is indicated in red. In recessive disorders, two copies of the affected allele are required to develop the disease.

(a)

	N	S
S	NS	SS
S	NS	SS

(b)

	N	S
N	NN	NS
S	NS	SS

Figure 2.4 Probabilities related to sickle cell anemia. Squares have been arranged to discover all possible genotypes that occur in children, given the genotypes of their parents. From this diagram, we may also infer the probability of each offspring genotype. The sickle cell allele is represented by "S" and the normal allele as "N." The top row has the genotype of one parent, and the leftmost column the genotype of the other parent. The other boxes are obtained by copying the parent letters across or down. In this way, we get the predicted frequencies of all the potential genotypes. (a) This is a case where one parent has sickle cell anemia (genotype SS) and the other parent is heterozygous. As the SS genotype occurs in two of the four cells, we estimate that the probability that a child has sickle cell anemia is 50%. In (b) both parents are heterozygous. Here, the SS genotype occurs in one of the four cells, and we estimate that the probability is 25% that a child is affected by the disease. The example shows sickle cell anemia, a recessive disorder, but the method of inferring probabilities may be used to examine just about any case of recessive or dominant inheritance.

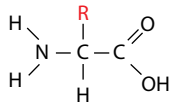

Figure 2.5 Chemical structure of an amino acid. All amino acids contain a carbon atom to which is connected an amino group, a carboxyl group, a hydrogen atom, and a side chain (R) specific to the amino acid.

known as a Punnett square. For instance, if one parent is homozygous for the sickle variant and the other parent is heterozygous, the probability that a child of these parents develops sickle cell anemia is 50%. If both parents are heterozygous, the probability is instead 25%.

As opposed to recessive disorders, in **dominant** disorders, only one disease gene is sufficient to develop the disease. Recessive and dominant diseases may be distinguished because they have different inheritance patterns (for an example of a dominant disorder and its inheritance, see Chapter 7).

SICKLE CELL ANEMIA AND MALARIA

Sickle cell anemia affects many hundreds of thousands around the world. In particular, it is common among individuals of Sub-Saharan African descent. In the United States, it occurs among about 1 out of every 365 black American births, and about 1 in 13 black American babies is born with sickle cell trait. In Africa, sickle cell disease is even more common—in some parts of Sub-Saharan Africa, up to 1 in 30 of all newborns are affected by the disease.

Why is sickle cell anemia such a common inherited disorder in parts of Africa? It was demonstrated in the 1950s that people with the sickle cell trait are more resistant against malaria caused by the protozoan parasite *Plasmodium falciparum*. Therefore, in areas of malaria, the sickle cell gene has a selective advantage. It is still not clear why the sickle cell gene protects against malaria.

CHARACTERIZING SICKLE CELL ANEMIA FURTHER: THE ROLE OF HEMOGLOBIN

What is actually the cause of sickle cell anemia? Research eventually zoomed in on **hemoglobin**, a protein.

Proteins are important molecules in biology, as they carry out many critical functions. For instance, they act as enzymes, transporters, and receptors mediating hormonal response. By the turn of the twentieth century, it was shown that proteins are composed of **amino acids**. All amino acids have a carboxyl group and an amino group in addition to a side chain specific to the amino acid (**Figure 2.5**). The chemical structures of selected amino acids are shown in **Figure 2.6** (see Appendix Figure A.1 for the structure of all 20 amino acids). In proteins, the amino acids are joined through **peptide bonds** (**Figure 2.7**).

Hemoglobin belongs to the family of transporter proteins. It is an important protein responsible for transporting oxygen from the lungs to the tissues; in addition, it carries carbon dioxide back to the lungs. The oxygen-binding properties of hemoglobin had already been discovered in the nineteenth century. Could the amino acid composition of hemoglobin somehow

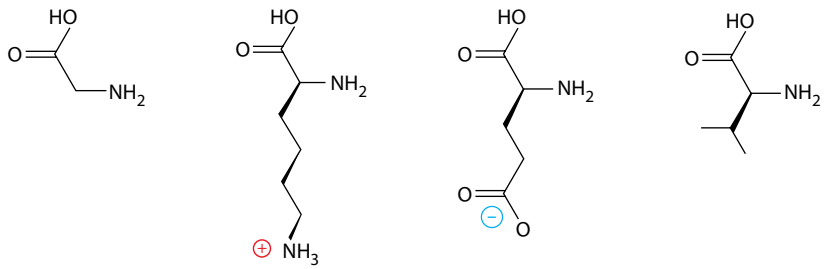

Figure 2.6 Chemical structures of selected amino acids. Glycine is the smallest amino acid with a hydrogen as its side chain. Lysine and glutamic acid are amino acids that are charged because of their amino and carboxyl groups, respectively. Valine belongs to a group of nonpolar amino acids.

be changed in sickle cell anemia? In one classic experiment in 1949, Linus Pauling compared hemoglobin from healthy individuals to that of sickle cell patients. It turned out that sickle cell globin had a different electrophoretic mobility—that is, when the proteins were subjected to an electric field, the sickle cell hemoglobin moved as a positively charged protein, in contrast to the normal hemoglobin (**Figure 2.8**). Pauling suggested that the difference was due to a change in the number of charged amino acids in the sickle cell version of the protein. The amino acids glutamic acid and aspartic acid are negatively charged because they carry a carboxyl group, and lysine and arginine are positively charged as they have an amino group (see Appendix Figure A.1). It seemed likely that sickle cell hemoglobin was affected in one or more of these amino acids. In the context of this finding, Pauling also used the term "molecular disease." This expression referred to the fact that for the first time the molecular background to a disease was identified.

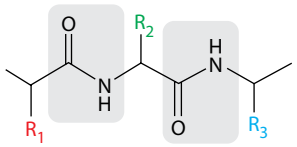

Figure 2.7 Amino acids are connected through peptide bonds. A unit with CONH represents the peptide bond. R1, R2, and R3 represent amino acid side chains.

MAPPING A DELETERIOUS CHANGE IN HEMOGLOBIN

Further details as to the molecular basis of the disease were elucidated in the 1950s. Vernon Ingram did one crucial experiment in 1956. A protein may be cut up into smaller pieces—peptides—using an enzyme that is able to break up peptide bonds. One such enzyme is trypsin that cuts on the carboxyl-terminal side of lysine and arginine residues. Ingram used this enzyme to fragment hemoglobin and was able to separate the resulting peptides with a newly developed technique of two-dimensional separation by electrophoresis and chromatography. When comparing normal hemoglobin to the sickle cell variant, it turned out that one single peptide was different (**Figure 2.9a**). The peptide in sickle cell hemoglobin was more positively charged than that of normal hemoglobin. Analysis of its amino acid composition showed that it had less glutamic acid and more of valine, suggesting that in sickle cell anemia, glutamic acid had been replaced by valine. The exact sequence of amino acids in the peptide was soon after determined. In the sickle cell protein, the peptide had valine instead of glutamic acid in one position (**Figure 2.9b**).

The analysis by Pauling and Ingram of sickle cell hemoglobin in the 1940s and 1950s was an important milestone in molecular biology. For the first time, a genetic inherited disorder was explained in terms of a specific amino acid substitution in a protein. We now know many more examples of inherited disorders associated with amino acid replacements as will be apparent in forthcoming chapters.

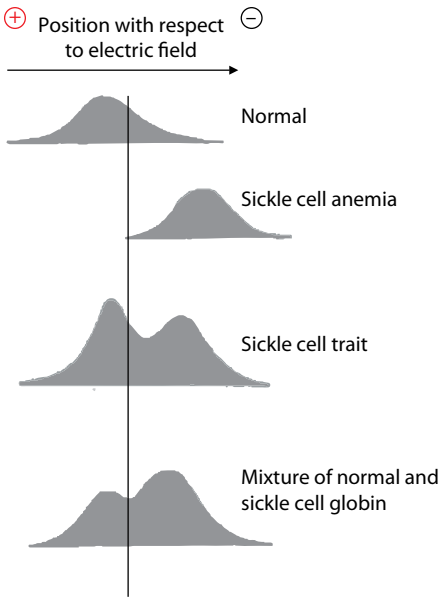

Figure 2.8 Electrophoretic mobility of normal hemoglobin and its sickle cell variant as studied by Linus Pauling. Pauling used a method of electrophoresis where the protein samples to be separated are in solution. By applying a current, the proteins will move in the solution dependent on their charge. The position with respect to the electric field is on the *x*-axis, and the peak height (*y*-axis) is proportional to protein concentration. The vertical line represents the position of an electrically neutral molecule. The results of Pauling's experiment demonstrated that the sickle cell hemoglobin moved as a positively charged molecule, whereas the normal protein migrated as a negatively charged molecule. The four different samples are (1) normal hemoglobin, (2) sickle cell hemoglobin, (3) hemoglobin prepared from individuals with the sickle cell trait (such individuals have approximately equal proportions of sickle and normal protein), and (4) a synthetic mixture with equal amounts of normal and sickle cell protein.

A PROTEIN FOLDS INTO A THREE-DIMENSIONAL SHAPE BASED ON ITS AMINO ACID SEQUENCE

To understand properly the molecular basis of sickle cell anemia, we also need to know the structural consequences of the amino acid substitution. Proteins are built from one or more **polypeptide** chains—each such chain is a string of amino acids joined by peptide bonds. The sequence of amino acids in each of the polypeptide chains will determine the three-dimensional shape of the protein. We commonly refer to four different levels of protein structure: **primary**, **secondary**, **tertiary**, and **quaternary** (**Figure 2.10**). The primary structure refers to the amino acid sequence of the polypeptide chain (**Figure 2.10a**). Noncovalent interactions between amino acids in the same polypeptide chain give rise to structures known

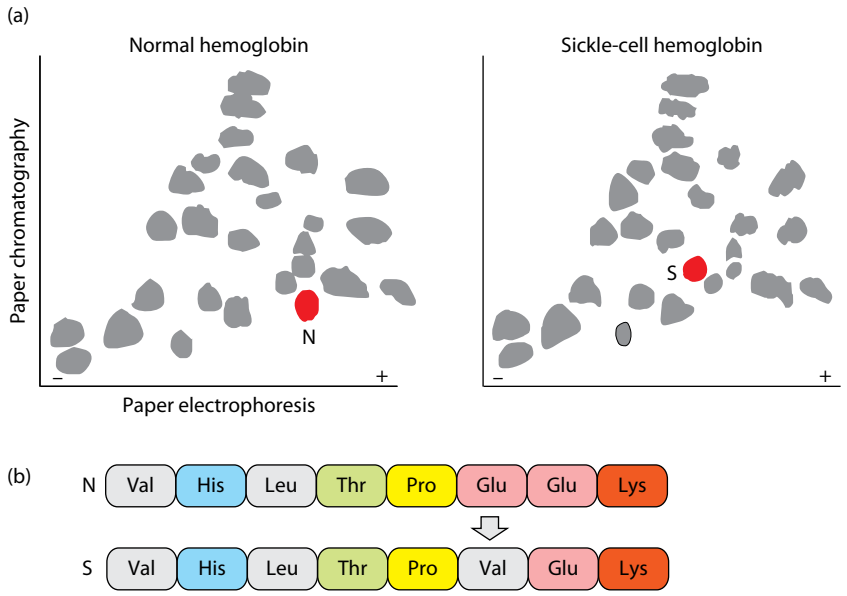

Figure 2.9 Two-dimensional separation of hemoglobin peptides. (a) Normal hemoglobin and sickle cell globin were treated with trypsin to generate a number of peptides. These peptides were separated in two dimensions using paper electrophoresis and paper chromatography. The patterns obtained are identical except for one peptide (N and S, for normal and sickle cell hemoglobin, respectively). (b) Determination of the peptide sequences showed that glutamic acid in position 6 of the normal protein had been replaced by valine in the sickle cell protein.

as **α-helices** and **β-sheets**. These are secondary structure elements (**Figure 2.10b**). The tertiary structure of a protein refers to the global folding of the entire polypeptide chain of a protein. This structure may contain α-helices and/or β-sheets, as well as loop regions in between (**Figure 2.10c**). In the event the protein has at least two polypeptide chains, also known as **subunits**, we also need to consider the number and arrangement of these polypeptide chains—the quaternary structure (**Figure 2.10d**).

There are three major experimental methods whereby the three-dimensional structure of a protein may be inferred. In **x-ray crystallography**, the protein is in the form of a crystal. The crystal is irradiated with x-rays, and the diffraction pattern obtained is used to calculate the electron density of the crystal. The electron density is finally used to elucidate the coordinates of the different atoms in the protein. Most available protein structures have been determined with x-ray crystallography. Second, **nuclear magnetic resonance** (**NMR**) is a method for protein structure determination where the protein is in solution. Third, **cryo-electron microscopy** is a more recent method used for structural studies of large proteins or complexes involving proteins and other large molecules. Structure may also be predicted from the sequence of amino acids using computational tools, although this is much less reliable than the experimental methods.

SICKLE CELL DISEASE MAY NOW BE UNDERSTOOD IN THE CONTEXT OF PROTEIN STRUCTURE

In 1959, the structure of hemoglobin was elucidated by Max Perutz using the technique of x-ray crystallography (**Figure 2.11**). The hemoglobin molecule is built from two different polypeptides, α-globin and β-globin. In one molecule of hemoglobin, there are two α-chains and two β-chains. It is the β-chain that is affected in sickle cell anemia.

Hemoglobin and a related protein myoglobin were in fact the first protein structures ever to be presented. The structure of hemoglobin, as well as that of the sickle cell variant, have been further refined in later work.

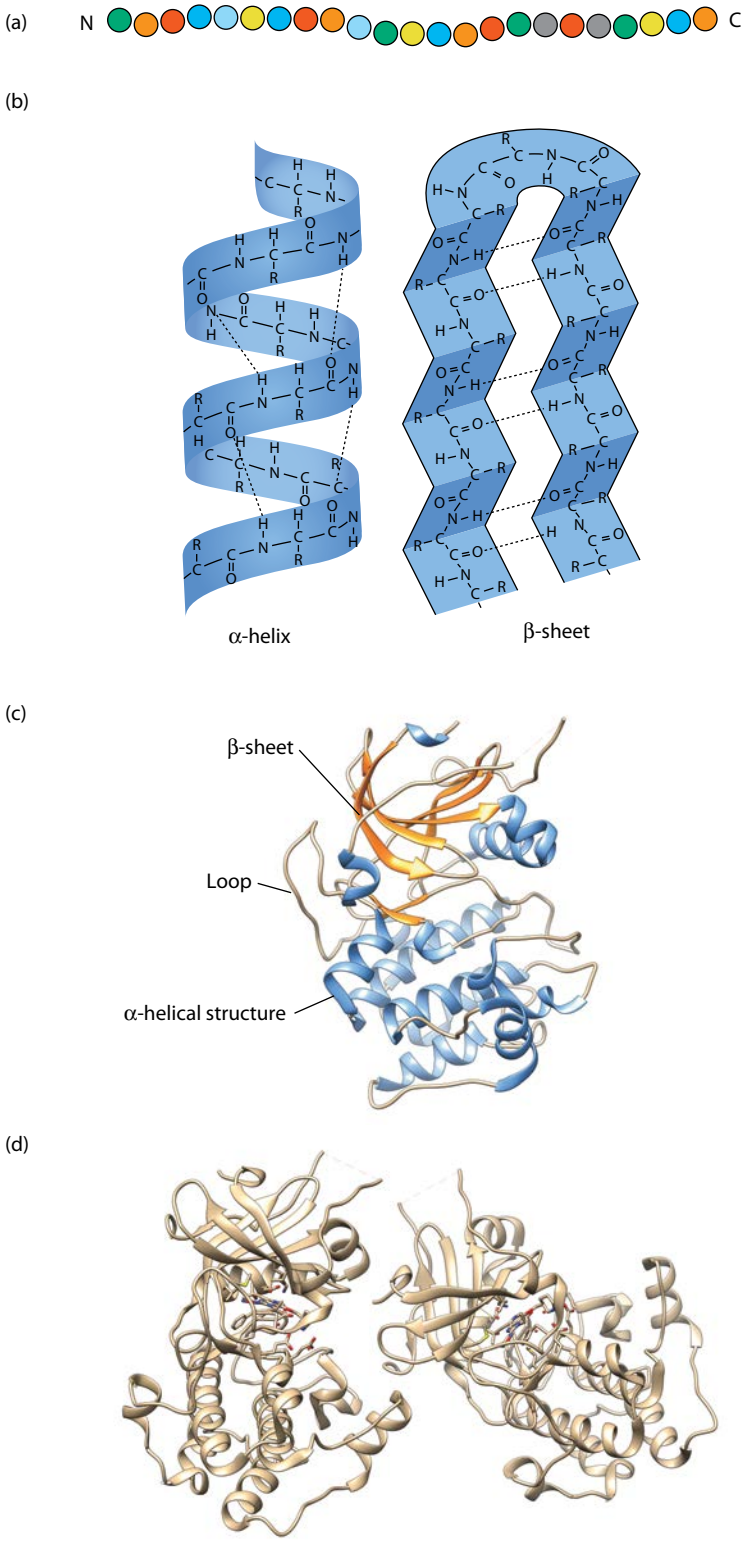

Figure 2.10 **The four levels of protein structure.** (a) Primary structure is the linear sequence of amino acids along the polypeptide chain. (b) Protein secondary structure elements. An α-helix is a common helical structure in proteins in which every backbone N-H group donates a hydrogen bond to the backbone C = O group of the amino acid located three or four residues earlier along the protein sequence. Another common secondary structure element in proteins is the β-sheet . It consists of strands connected laterally by backbone hydrogen bonds. (c) The tertiary structure of a protein is its global fold, including α-helices (blue), β-sheets (orange), and loop regions without ordered structure. The protein shown is a subunit of the enzyme Akt2 with both α-helical and β-sheet structures. (d) The quaternary structure of Akt2 is shown—a dimer of two identical polypeptide chains.

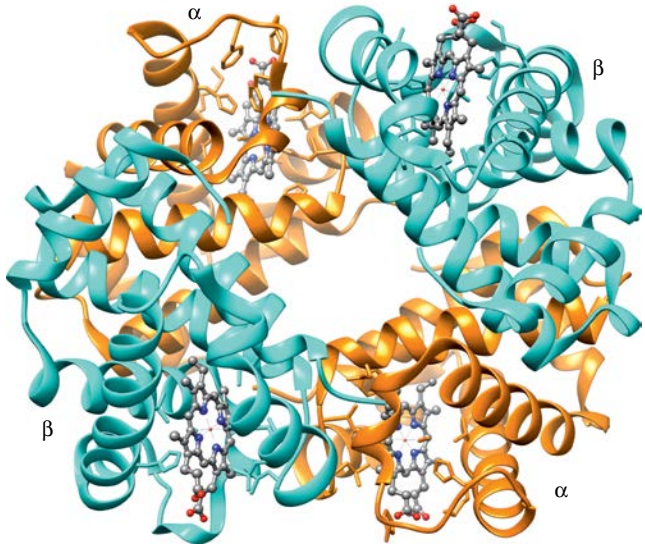

Figure 2.11 Three-dimensional structure of hemoglobin. Hemoglobin is built from two α-chains (orange) and two β-chains (cyan). Each of the four subunits contains heme groups (shown in ball-and-stick representation). (Adapted from Paoli M et al. 1996. *J Mol Biol* 256:775. PDB ID: 1GZX.)

The replacement in sickle cell anemia of glutamic acid—a negatively charged amino acid (see Figure 2.6 and **Figure 2.12a**)—with valine has a significant effect on the properties of hemoglobin. Valine is a hydrophobic amino acid and interacts with a pair of hydrophobic amino acids—phenylalanine in position 85 and leucine in position 88 of the protein chain—that are located in the β-chain of a neighboring globin molecule (**Figure 2.12b**). As a result, an aggregation process is initiated where globin molecules form fibers. These fibers extend through the red blood cells such that the cells are distorted. Because of this alteration in shape and because sickle cells tend to stick to the wall of blood vessels, normal blood flow is prevented. Pain may arise when blood flow is restricted. Sickled cells have a shorter life span, giving rise to a lowered amount of red blood cells and poor oxygen transport (anemia).

A MOLECULAR SEQUENCE: PROTEINS CAN BE DEPICTED AS A STRING OF AMINO ACID SYMBOLS

Whereas Ingram focused on the amino acid sequence of selected peptides, the complete amino acid sequences of the human α- and β-globins were elucidated in the early 1960s. The sequence of the β-chain is shown in **Figure 2.13**, where the amino acids are shown with a background color based on their physical and chemical properties. A polypeptide chain built from amino acids has a distinct polarity, where one end of the polypeptide has a free amino-terminal group (the N-terminus) and the other end has a free carboxyl terminal group (C-terminus). Amino acid sequences are, as a rule, displayed with the N-terminus to the left and the C-terminus to the right. The amino acids are typically abbreviated with one letter code (see Figure 2.13) or three letter codes. The reader is referred to Figure A.1 for details of these codes.

Amino acid sequences of proteins may be compared using a computational procedure of **alignment**. Such an alignment is shown in **Figure 2.14**, where α- and β-chains of five different vertebrates are compared. From this alignment, we observe that all the globin sequences are similar, reflecting the fact that they have been well conserved during evolution. We also identify features that distinguish the α- and β-chains,

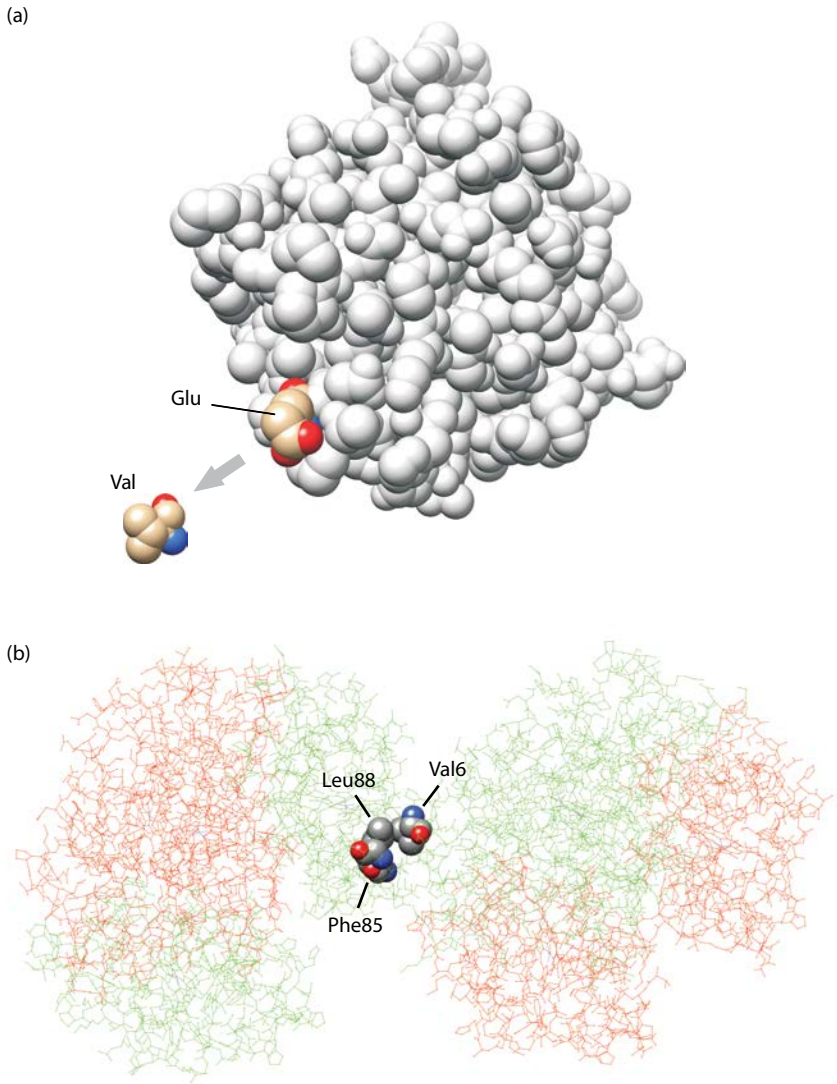

Figure 2.12 Sickle cell hemoglobin molecules interact with each other in an abnormal manner. (a) Space-filling model of normal β-chain of hemoglobin showing surface location of glutamic acid (colored with side-chain carboxyl oxygen atoms in red) in position 6 of the polypeptide chain. For comparison is shown the structure of valine, the amino acid replacing glutamic acid in sickle cell anemia. (b) Two molecules of sickle cell hemoglobin showing interaction between hydrophobic amino acids in neighboring subunits. The β-chains are in green and α-chains in red (wireframe representation). Highlighted in space-fill representation is an interaction between valine 6 in one polypeptide with a pair of two other hydrophobic amino acids—phenylalanine 85 and leucine 88—in a neighboring polypeptide. This interaction initiates aggregation of globin molecules into harmful fibers. (Adapted from Tame J, Vallone B. 2000. *Acta Crystallogr* 56:805–811. PBD ID: 1A3N. Harrington DJ et al. 1997. *J Mol Biol* 272:398–407. PDB ID: 2HBS.)

Figure 2.13 Amino acid sequence of the β-chain of hemoglobin. Amino acids are colored according to their physical and chemical properties. For instance, the basic (positively charged) amino acids arginine (R) and lysine (K) have a red background. Cleavage sites for trypsin as used in the experiment of Figure 2.9 are indicated with arrows.

such as an insertion of five amino acids in the β-chains, as compared to the α-chains. However, there is a lot of additional information that we may extract from amino acid sequences and alignments as further discussed in Chapter 12.

Amino acids in proteins are our first instance of a **molecular sequence**. There will be more of these as we enter into the world of nucleic acids—DNA and RNA—in the next chapter. We then address the important problem of how, at a molecular level, glutamic acid was changed into valine in sickle cell anemia.

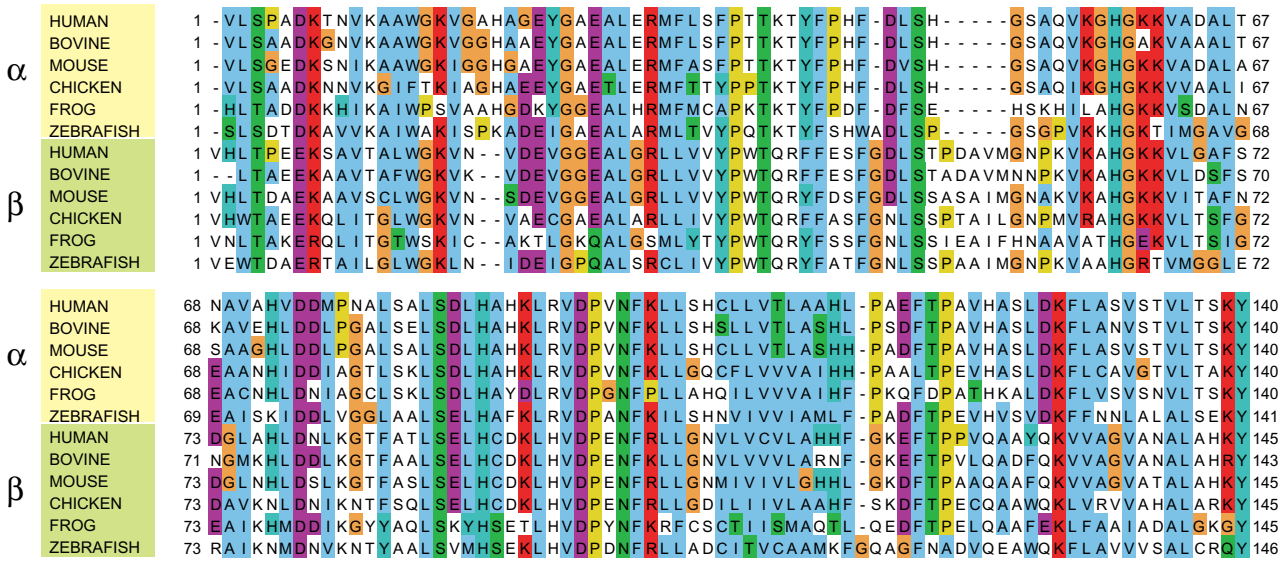

Figure 2.14 Alignment of vertebrate hemoglobin amino acid sequences. Coloring of amino acids is as in Figure 2.13.

SUMMARY

- Sickle cell anemia is a recessive inherited disorder. For the disease to be expressed, individuals must carry two copies of the sickle cell allele. Individuals with only one copy have the sickle cell trait with less severe symptoms.
- Protein structure may be described at the four levels: primary, secondary, tertiary, and quaternary.
- Hemoglobin is built from two chains of α-globin and two chains of β-globin.
- In sickle cell anemia, the amino acid glutamic acid of β-globin is replaced by the hydrophobic amino acid valine, resulting in an aggregation process.
- Sickle cell anemia was the first disease to be characterized where a genetic change is associated with a well-defined change in a protein molecule.
- The primary structure of a protein—the amino acid sequence—may be depicted as a string of amino acid symbols. Amino acid sequences may be compared using a computational procedure of alignment.

QUESTIONS

1. What is characteristic of recessive and dominant inheritance, respectively?
2. What is the difference between *sickle cell trait* and *sickle cell anemia*?
3. Explain the concepts *genotype* and *phenotype*.
4. Explain what is meant when we say that an individual is *homozygous* for a specific genetic variant (allele).
5. Use a Punnett square to show the offspring when one parent is homozygous for the sickle cell allele and one parent has two normal globin alleles.
6. Why is the sickle cell gene variant (allele) fairly common in specific populations?
7. Draw the chemical structure of an *amino acid*. Also draw a dipeptide (two amino acids joined with a *peptide bond*).
8. What are the acidic and basic amino acids, respectively?
9. Give examples of hydrophobic amino acids.

10. What were the experiments carried out in the 1940s and 1950s to elucidate the molecular basis of sickle cell anemia?
11. Explain what is meant by *primary*, *secondary*, *tertiary*, and *quaternary* structures of proteins.
12. What are the methods used to infer the three-dimensional structure of proteins and other large molecules?
13. In sickle cell anemia, one amino acid is replaced with another in the β-globin subunit of hemoglobin. What are the consequences in terms of amino acid interactions and protein structure? What are the physiological consequences?
14. What is meant by a *molecular sequence*?

FURTHER READING

General on sickle cell anemia and its history

Bjorklund R. 2011. *Sickle cell anemia*. Marshall Cavendish Benchmark, New York.

Orkin SH, Higgs DR. 2010. Medicine. Sickle cell disease at 100 years. *Science* 329(5989):291–292.

Serjeant GR. 2001. The emerging understanding of sickle cell disease. *Br J Haematol* 112(1):3–18.

Serjeant GR. 2010. One hundred years of sickle cell disease. *Br J Haematol* 151(5):425–429.

Early work on human hemoglobin and myoglobin

Baglioni C. 1961. An improved method for the fingerprinting of human hemoglobin. *Biochim Biophys Acta* 48:392–396.

Braunitzer G, Hilschmann N, Rudloff V et al. 1961. The hæmoglobin particles. *Nature* 190:480–482.

Braunitzer G, Gehring-Mueller R, Hilschmann N et al. 1961. The structure of normal adult human hemoglobins. *Hoppe Seylers Z Physiol Chem* 325:283–286.

Ingram VM. 1959. Abnormal human haemoglobins. III. The chemical difference between normal and sickle cell haemoglobins. *Biochim Biophys Acta* 36:402–411.

Kendrew JC, Bodo G, Dintzis HM et al. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181(4610):662–666.

Perutz MF, Rossmann MG, Cullis AF et al. 1960. Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by x-ray analysis. *Nature* 185(4711):416–422.

Sickle cell anemia: Early genetic studies

Neel JV. 1949. The inheritance of sickle cell anemia. *Science* 110(2846):64–66.

Sickle cell anemia: Early paper by Pauling

Pauling L, Itano HA, Singer SJ, Wells IC. 1949. Sickle cell anemia: A molecular disease. *Science* 110(2865):543–548.