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