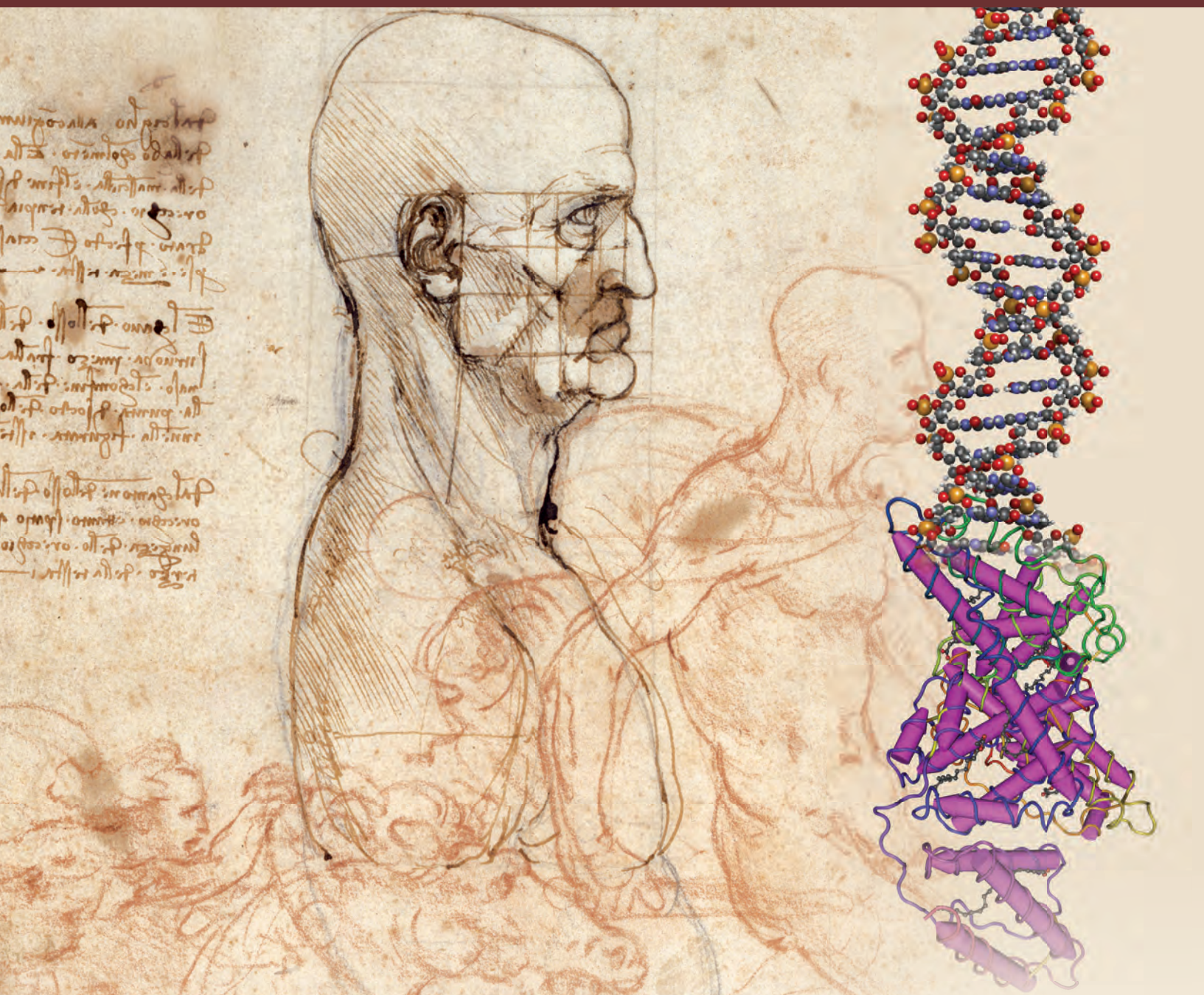
BIOINFORMATICS AND FUNCTIONAL GENOMICS

third edition



Jonathan Pevsner

WILEY Blackwell

BIOINFORMATICS AND FUNCTIONAL GENOMICS

Bioinformatics and Functional Genomics

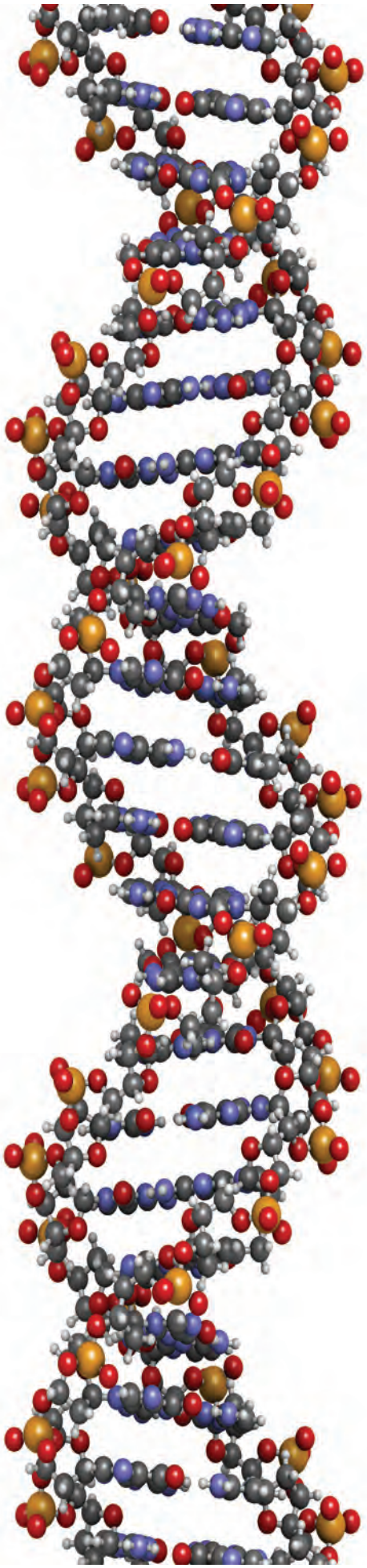
Third Edition

Jonathan Pevsner

Department of Neurology, Kennedy Krieger Institute, Baltimore, Maryland, USA

and

Department of Psychiatry and Behavioral Sciences, The Johns Hopkins School of Medicine, Baltimore, Maryland, USA



WILEY Blackwell

This edition first published 2015 © 2015 by John Wiley & Sons Inc

Registered office: John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial offices: 9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data
Pevsner, Jonathan, 1961- , author.
Bioinformatics and functional genomics / Jonathan Pevsner.—Third edition.
p. ; cm.
Includes bibliographical references and indexes.
ISBN 978-1-118-58178-0 (cloth)
I. Title.
[DNLM: 1. Computational Biology—methods. 2. Genomics. 3. Genetic Techniques. 4. Proteomics. QU 26.5]
QH441.2
572.8′6–dc23

2015014465

A catalogue record for this book is available from the British Library.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

The cover image is by Leonardo da Vinci, a study of a man in profile with studies of horse and riders (reproduced with kind permission of the Gallerie d'Accademia, Venice, Ms. 7r [236r], pen, black and red chalk). To the upper right a DNA molecule is shown (image courtesy of Wikimedia Commons) and a protein (human serum albumin, the most abundant protein in blood plasma, accession 1E7I, visualized with Cn3D software described in Chapter 13). Leonardo's text reads: "From the eyebrow to the junction of the lip with the chin, and the angle of the jaw and the upper angle where the ear joins the temple will be a perfect square. And each side by itself is half the head. The hollow of the cheek bone occurs half way between the tip of the nose and the top of the jaw bone, which is the lower angle of the setting on of the ear, in the frame here represented. From the angle of the eye-socket to the ear is as far as the length of the ear, or the third of the face." (Translation by Jean-Paul Richter, *The Notebooks of Leonardo da Vinci*, London, 1883.)

Set in Times LT Std 10.5/13 by Aptara, India

Printed in Singapore

1 2015

For three generations of family: to my parents
Aihud and Lucille; to my wife Barbara; to my daughters Kim,
Ava, and Lillian; and to my niece Madeline

Contents in Brief

Contents

Preface to the Third Edition

When the first edition of this textbook was published in 2003, the Human Genome Project had just been completed at a cost of nearly US$ 3 billion. When the second edition came into print in 2009, the first genome sequence of an individual (J. Craig Venter) had recently been published at an estimated cost of US$ 80 million.

Let me tell you a remarkable story. It is now 2015 and it costs just several thousand dollars to obtain the complete genome sequence of an individual. Sturge-Weber syndrome is a rare neurocutaneous disorder (affecting the brain and skin) that is sometimes debilitating: some patients must have a hemispherectomy (removal of half the brain) to alleviate the severe seizures. We obtained paired samples from just three individuals with Sturge-Weber syndrome: biopsies were from affected parts of the body (such as port-wine stains that occur on the face, neck, or shoulder) or from presumably unaffected regions. We purified DNA and sequenced these six whole genomes, compared the matched pairs, and identified a single base pair mutation in the *GNAQ* gene as responsible for Sturge-Weber syndrome. (The mutation is somatic, mosaic, and activating: somatic in that it occurs during development but is not transmitted from the parents; mosaic in that it affects just part of the body; and activating because *GNAQ* encodes a protein that in the mutated form turns on a signaling cascade.) We found that mutations in this gene also cause port-wine stain birthmarks (which affect 1 in 300 people or about 23 million people worldwide). Matt Shirley, then a graduate student in my lab, performed the bioinformatics analyses that led to this discovery. He analyzed about 700 billion bases of DNA. After finding the mutation he confirmed it by re-sequencing dozens of samples, typically at over 10,000-fold depth of coverage. We reported these findings in the *New England Journal of Medicine* in 2013.

This story illustrates several aspects of the fields of bioinformatics and genomics. First, we are in a time period when there is an explosive growth in the availability of DNA sequence. This is enabling us to address biological questions in unprecedented ways. Second, while it is inexpensive to acquire DNA sequences, it is essential to know how to analyze them. One goal of this book is to introduce sequence analysis. Third, bioinformatics serves biology: we can only interpret the significance of DNA sequence variation in the context of some biological process (such as a disease state). In the case of the *GNAQ* mutation, that gene encodes a protein (called Gαq) that we can study in tremendous detail using the tools of bioinformatics; we can evaluate its three-dimensional structure, the proteins and chemical messengers it interacts with, and the cellular pathways it participates in. Fourth, bioinformatics and genomics offer us hope. For Sturge-Weber syndrome patients and those with port-wine stain birthmarks, we are hopeful that a molecular understanding of these conditions will lead to treatments.

This book is written by a biologist who has used the tools of bioinformatics to help understand biomedical research questions. I introduce concepts in the context of biological problem-solving. Compared to earlier editions, this new text emphasizes command-line software on the Linux (or Mac) platform, complemented by web-based approaches.

In an era of "Big Data" there is a great divide between those whose intellectual core is centered in biomedical science and those whose focus involves computer science. I hope this book helps to bridge the divide between these two cultures.

Writing a book like this is a wonderful and constant learning experience. I thank past and present members of my lab who taught me including Shruthi Bandyadka (for advice on R), Christopher Bouton, Carlo Colantuoni, Donald Freed (for extensive advice on next-generation sequencing or NGS), Laurence Frelin, Mari Kondo, Sarah McClymont, Nathaniel Miller, Alicia Rizzo, Eli Roberson, Matt Shirley (who also provided extensive NGS advice), Eric Stevens, and Jamie Wangen. For advice on specific chapters, I thank: Ben Busby of the National Center for Biotechnology Information (NCBI) for advice regarding Chapters 1, 2, and 5 and detailed comments on Chapters 9 and 10; Eric Sayers and Jonathan Kans of NCBI for advice on EDirect in Chapter 2; Heiko Schmidt for advice on TREE-PUZZLE and MrBayes in Chapter 7; Joel Benington for detailed comments on Chapters 8 and 15–19 and helpful discussions about teaching; Harold Lehmann for guidance on various fields of informatics; and N. Varg for helpful comments on all chapters. I thank many colleagues who participated in teaching bioinformatics and genomics courses over the years. I've learned from all these teachers, including Dimitri Avramopoulos, Jef Boeke, Kyle Cunningham, Garry Cutting, George Dimopoulos, Egert Hoiczyk, Rafael Irizarry, Akhilesh Pandey, Sean Prigge, Ingo Ruczinski, Alan Scott, Alan F. Scott, Kirby D. Smith, David Sullivan, David Valle, and Sarah Wheelan. I am grateful to faculty members with whom I taught genomics workshops including Elana Fertig, Luigi Marchionni, John McGready, Loris Mulroni, Frederick Tan, and Sarah Wheelan. This book includes several thousand literature references, but I apologize to the many more colleagues whose work I did not cite. I also cite 900 websites and again apologize to the developers of the many I did not include.

I also acknowledge the support of Dr Gary W. Goldstein, President and CEO of the Kennedy Krieger Institute where I work. Kennedy Krieger Institute sees 22,000 patients a year, mostly children with neurodevelopmental disorders from common conditions (such as autism spectrum disorder and intellectual disability) to rare genetic diseases. I am motivated to try to apply the tools of bioinformatics and genomics to help these children. This perspective has guided my writing of this book, which emphasizes the relevance of all the topics in bioinformatics and genomics to human disease in general. We are hopeful that genomics will lead to an understanding of the molecular bases of so many devastating conditions, and this in turn may one day lead to better diagnosis, prevention, treatment, and perhaps even cures.

It is my pleasure to thank my editors at Wiley-Blackwell – Laura Bell, Celia Carden, Beth Dufour, Elaine Rowan, Fiona Seymour, Audrie Tan, and Rachel Wade – for generous support throughout this project. I appreciate all their dedication to the quality of the book.

On a personal note I thank my wife Barbara for her love and support throughout the very long process of writing this textbook. Finally, to my girls Ava and Lillian: I hope you'll always be inspired to be curious and full of wonder about the world around us.

About the Companion Website

This book is accompanied by a companion website:

www.wiley.com/go/pevsnerbioinformatics

Readers can visit this website for supplemental information, such as PowerPoint files of all the figures and tables from the book, solutions to the Self-Test Quizzes and Problems found at the end of each chapter.

The author also maintains a comprehensive website for the book:

www.bioinfbook.org

This site features lecture files (in PowerPoint and audiovisual format), over 900 Web Links and over 130 Web Documents that are referred to throughout the book as well as videocasts of how to perform many basic operations.