A BIOINFORMATICS GUIDE FOR MOLECULAR BIOLOGISTS

ALSO FROM COLD SPRING HARBOR LABORATORY PRESS

At the Bench: A Laboratory Navigator, Updated Edition
At the Helm: Leading Your Laboratory, Second Edition
Experimental Design for Biologists, Second Edition
Introduction to Protein–DNA Interactions: Structure, Thermodynamics, and Bioinformatics
Lab Math: A Handbook of Measurements, Calculations, and Other Quantitative Skills for Use at the Bench, Second Edition
Lab Ref: A Handbook of Recipes, Reagents, and Other Reference Tools for Use at the Bench, Volumes 1 and 2
Next-Generation DNA Sequencing Informatics
Statistics at the Bench: A Step-by-Step Handbook for Biologists

A BIOINFORMATICS GUIDE FOR MOLECULAR BIOLOGISTS

SARAH J. AERNI
MARINA SIROTA
Biomedical Informatics Program
Stanford University School of Medicine

CSH PRESS
COLD SPRING HARBOR LABORATORY PRESS
Cold Spring Harbor, New York • www.cshlpress.org

A BIOINFORMATICS GUIDE FOR MOLECULAR BIOLOGISTS

Publisher	John Inglis
Acquisition Editors	Ann Boyle, Kaaren Janssen
Director of Editorial Development	Jan Argentine
Developmental Editors	Kaaren Janssen, Gabrielle Leblanc, Lori Martin, Christin Munkittrick, Kimberly Prohaska
Project Managers	Maryliz Dickerson, Inez Sialiano
Production Editor	Rena Springer
Production Manager	Denise Weiss
Cover Designer	Michael Albano

Front cover: Art courtesy of Ryan Katkov.

Library of Congress Cataloging-in-Publication Data

A bioinformatics guide for molecular biologists / [edited by] Sarah Aerni and Marina Sirota, Biomedical Informatics Program, Stanford University School of Medicine.
pages cm
ISBN 978-1-936113-22-4 (hardcover : alk. paper)
1. Molecular biology--Data processing. 2. Molecular biology--Computer simulation.
3. Bioinformatics. I. Aerni, Sarah, 1983-II. Sirota, Marina.

QH506.B534 2013
572.80285- -dc23
2013017412

10 9 8 7 6 5 4 3 2 1

All World Wide Web addresses are accurate to the best of our knowledge at the time of printing.

For a complete catalog of all Cold Spring Harbor Laboratory Press publications, visit our website at www.cshlpress.org.

Contents

Section 1: Introduction and Overview

Section 2: Fundamental Concepts

Section 3: Techniques for Analyzing Your Data

Section 4: Augmenting Your Data

Preface

Due to a staggering increase in the abundance and availability of molecular data during recent years, bioinformatics has enabled scientists to ask questions previously impossible to answer. The purpose of this book is to arm molecular biologists with the knowledge to enable their use of bioinformatics tools. The idea of the book was born when John Inglis and Kaaren Janssen visited Stanford in 2008 to discuss the future of the *Bioinformatics* textbook long used by students. From this meeting came the idea of a new project, to be led and executed by graduate students and postdocs, to create a bioinformatics "handbook," making computational and statistical techniques more accessible for molecular biologists. The sometimes oblique (to molecular biologists) language of bioinformatics would also be made more accessible by the introduction of significant terminology (within the text, these boldfaced terms are now defined in the context in which they appear). Such a book would drive the adoption of new techniques and technologies by providing a fresh perspective from the graduate students and create a natural source of new content with the new students entering the program, who would use this as an opportunity to "pass the torch" from one group of students to the next.

Without the important contributions and support of many individuals, this book would not have been possible. First, we thank all of the authors for their hard work. We are also grateful for the support of the Biomedical Informatics Program at Stanford. In particular, we thank the members of the executive committee of the Program, including Russ Altman, Atul Butte, Teri Klein, Larry Fagan, Mark Musen, Amar Das, David Paik, Dan Rubin, Mary Jeanne Oliva, and Darlene Vian, who were critical in enabling us to act as student representatives. Without their support and encouragement, this book would not have been possible. The authors and editors have chosen to donate all of the royalties from this publication to the Stanford Biomedical Informatics Program in support of future student endeavors. In addition, we thank all of the staff at Cold Spring Harbor Laboratory Press, who facilitated the creation of this book from its inception to publication, in particular, Kaaren Janssen, who initially presented the idea for the book to us while visiting Stanford University. Her ever-positive attitude and enthusiasm made this an amazing experience. We are grateful to many more at the Press, including John Inglis, Rena Springer, Maryliz Dickerson, and Inez Sialiano, who were also instrumental in making this book possible. We also sincerely thank Greg Cooper and Sarah Elgin who provided thoughtful and critical reviews of many of chapters, as well as others who provided valuable comments on individual chapters, including Russ Altman, Steve Briggs, Betty Cheng, Mario Roederer, Gary Stormo, Jason Swedlow, and Michael Weiner. In addition, we are individually indebted to past and present mentors and teachers who, often unknowingly, inspired each of us to pursue biomedical informatics as a field of study. Lastly, and most importantly, we are eternally grateful to our families and friends, who provided us with limitless encouragement and support for our endeavors, including this one.

Marina Sirota and Sarah J. Aerni
June 5, 2014