Contents

Preface

THIS BOOK IS WRITTEN MAINLY for biologists who want to understand the methods of sequence and structure analysis. I strongly believe that a person using a computer program should understand how it works. Accordingly, one of my main objectives is to help biologists appreciate the underlying algorithms used and assumptions made, as well as limitations of the methods used and strategies for their use. To this end, I have tried to avoid complex formulas and notations and to give instead simple numerical examples whenever possible. I hope that the book will also be of interest to computational biologists who want to learn a little more about the biological questions related to the field of bioinformatics. This book is intended to be a laboratory reference text, as well as a textbook for a course in bioinformatics, rather than a user guide for a specific set of sequence analysis programs.

Most of the chapters include a flowchart that is designed to propose an orderly use of the methods that are discussed in the chapter. There are very few examples of these types of charts and they are quite difficult to produce, requiring assumptions and over-simplifications that may not always be justified. I hope that these charts will be useful for the less experienced in this field, but I expect that the more-experienced practitioners in the field will have other, probably better, ways of achieving the same goal.

There are many references to Web sites and FTP locations where these methods may be applied or programs obtained. In some cases, as for the commonly used and important BLAST and CLUSTALW programs, I have provided a great deal of information about using the program and analyzing the results. However, there are many other important tools and approaches available for biological sequence and genome analysis and I have tried to cover as many of them as possible, given time and space limitations. I have not paid particular attention to simpler types of sequence analyses, e.g., searching for restriction sites, translating sequences, and compositional analysis. There are many commercial and noncommercial packages for performing these tasks, and commercial packages for genome analysis are now appearing.

In writing this book, my first, I found that the amount of information available in the published literature was far more than I could include. I have tried to be thorough and to cover the most significant problems in sequence and genome analysis, but there are also many excellent papers that have not been cited for reasons of time and space, and I apologize to colleagues whose valuable contributions are not mentioned. Because of the space limitations of a printed text, and the ever-changing nature of bioinformatics, material not included in the book, as well as links to all of the Web sites cited, examples, and problems, will appear on a special Web site for the book, which can be found at http://www.bioinformaticsonline.org.

One aspect of this discipline that has been quite remarkable to me is the willingness of most investigators, especially the pioneers in the field, to share their results with colleagues. I have had the privilege of personally knowing several of these early investigators, especially David Lipman, Hugo Martinez (with whom I spent a sabbatical year), and Temple Smith. The tremendous accomplishments of these people became even more meritorious because they freely shared the results of their efforts with colleagues. In doing so, they were very much responsible for the eventual success of the sequence analysis field in both the academic and commercial areas.

This large project has required much support and help. Part of this book was derived from class notes for a course in "Bioinformatics and Genome Analysis" at the University of Arizona in the 1999 and 2000 academic years. Many students made very useful suggestions and were helpful in finding errors; I want to particularly thank Bryan Zeitler for providing many corrections. Any remaining errors will be corrected on the book's Web site. I am grateful to Bill Pearson for information about the FASTA suite of programs, to Julie Thompson and John Kececioglu for comments on Chapter 4, to Steve Henikoff for reading Chapter 3, and to Michael Zuker for helpful comments on the writing of Chapter 5. Bill Montfort provided information about PDB files for Chapter 9, and Roger Miesfeld provided the example of complex gene regulation in Chapter 8. Jun Zhu was very kind in answering my questions about the Bayes block aligner for Chapter 3. My department has been most patient and supportive as I skipped meetings and seminars to complete or revise another chapter, over a period of three years. During this time, Rob Han and Juwon Kim provided the very large number of papers and book chapters that I needed on a regular basis with a very short turnaround time, allowing me more time to digest the information. My editor, Judy Cuddihy of Cold Spring Harbor Laboratory Press, guided me through the process of writing with great skill and was very patient as she tried to keep me to a reasonable writing schedule, providing needed encouragement for completing the project. Elisabeth Cuddihy checked most of the Web sites, carefully went through formulas and numerical examples, and helped to write parts of the glossary. I also thank Joan Ebert and Jan Argentine in the Development Department and Pat Barker and Denise Weiss in the Production Department at the Press.

Last, but not least, I thank my wife Jennifer Hall for her patience and understanding during the many times that book-writing took precedence over family matters.

David W. Mount