Preface to the Second Edition

The Neurobehavioral Unit of the Kennedy Krieger Institute has 16 hospital beds. Most of the patients are children who have been diagnosed with autism, and most engage in self-injurious behavior. They engage in self-biting, self-hitting, head-banging, and other destructive behaviors. In most cases, we do not understand the genetic contributions to such behaviors, limiting the available strategies for treatment. In my research, I am motivated to understand molecular changes that underlie childhood brain diseases. The field of bioinformatics provides tools we can use to understand disease processes through the analysis of molecular sequence data. More broadly, bioinformatics facilitates our understanding of the basic aspects of biology including development, metabolism, adaptation to the environment, genetics (e.g., the basis of individual differences), and evolution.

Since the publication of the first edition of this textbook in 2003, the fields of bioinformatics and genomics have grown explosively. In the preface to the first edition (2003) I noted that tens of billions of base pairs (gigabases) of DNA had been deposited in GenBank. Now in 2009 we are reaching tens of trillions (terabases) of DNA, presenting us with unprecedented challenges in how to store, analyze, and interpret sequence data. In this second edition I have made numerous changes to the content and organization of the book. All of the chapters are rewritten, and about 90% of the figures and tables are updated. There are two new chapters, one on functional genomics and one on the eukaryotic chromosome. I now focus on the globins as examples throughout the book. Globins have a special place in the history of biology, as they were among the first proteins to be identified (in the 1830s) and sequenced (in the 1950s and 1960s). The first protein to have its structure solved by X-ray crystallography was myoglobin (Chapter 11); molecular phylogeny was applied to the globins in the 1960s (Chapter 7); and the globin gene loci were among the first to be sequenced (in the 1980s; see Chapter 16).

The fields of bioinformatics and genomics are far too broad to be understood by one person. Thus many textbooks are written by multiple authors, each of whom brings a deeper knowledge of the subject matter. I hope that this book at least offers the benefit of a single author's vision of how to present the material. This is essentially two textbooks: one on bioinformatics (parts I and II) and one on genomics (part III). I feel that presenting bioinformatics on its own would be incomplete without further applying those approaches to sequence analysis of genomes across the tree of life. Similarly I feel that it is not possible to approach genomics without first treating the bioinformatics tools that are essential engines of that field.

As with the previous edition a companion website is available which provides up-to-date web links referred to in the book and PowerPoint slides arranged by

chapter (www.bioinfbook.org). A resource site for instructors is also available giving detailed solutions to problems (www.wiley.com/go/pevsnerbioinformatics).

In preparing each edition of this book I read many papers and reviewed several thousand websites. I sincerely apologize to those authors, researchers and others whose work I did not cite. It is a great pleasure to acknowledge my colleagues who have helped in the preparation of this book. Some read chapters including Jef Boeke (Chapter 12), Rafael Irizarry (Chapter 9), Stuart Ray (Chapter 7), Ingo Ruczinski (Chapter 11), and Sarah Wheelan (Chapters 3 and 5–7). I thank many students and faculty at Johns Hopkins and elsewhere who have provided critical feedback, including those who have lectured in bioinformatics and genomics courses (Judith Bender, Jef Boeke, Egbert Hoiczyk, Ingo Ruczinski, Alan Scott, David Sullivan, David Valle, and Sarah Wheelan). Many others engaged in helpful discussions including Charles D. Cohen, Bob Cole, Donald Coppock, Laurence Frelin, Hugh Gelch, Gary W. Goldstein, Marjan Gucek, Ada Hamosh, Nathaniel Miller, Akhilesh Pandey, Elisha Roberson, Kirby D. Smith, Jason Ting, and N. Varg. I thank my wife Barbara for her support and love as I prepared this book.

Preface to the First Edition

Origins of this Book

This book emerged from lecture notes I prepared several years ago for an introductory bioinformatics and genomics course at the Johns Hopkins School of Medicine. The first class consisted of about 70 graduate students and several hundred auditors, including postdoctoral fellows, technicians, undergraduates, and faculty. Those who attended the course came from a broad variety of fields—students of genetics, neuroscience, immunology or cell biology, clinicians interested in particular diseases, statisticians and computer scientists, virologists and microbiologists. They had a common interest in wanting to understand how they could apply the tools of computer science to solve biological problems. This is the domain of bioinformatics, which I define most simply as the interface of computer science and molecular biology. This emerging field relies on the use of computer algorithms and computer databases to study proteins, genes, and genomes. Functional genomics is the study of gene function using genome-wide experimental and computational approaches.

Comparison

At its essence, the field of bioinformatics is about comparisons. In the first third of the book we learn how to extract DNA or protein sequences from the databases, and then to compare them to each other in a pairwise fashion or by searching an entire database. For the student who has a gene of particular interest, a natural question is to ask "what other genes (or proteins) are related to mine?"

In the middle third of the book, we move from DNA to RNA (gene expression) and to proteins. We again are engaged in a series of comparisons. We compare gene expression in two cell lines with or without drug treatment, or a wildtype mouse heart versus a knockout mouse heart, or a frog at different stages of development. These comparisons extend to the world of proteins, where we apply the tools of proteomics to complex biological samples under assorted physiological conditions. The alignment of multiple, related DNA or protein sequences is another form of comparison. These relationships can be visualized in a phylogenetic tree.

The last third of the book spans the tree of life, and this provides another level of comparison. Which forms of human immunodeficiency virus threaten us, and how can we compare the various HIV subtypes to learn how we might develop a vaccine? How are a mosquito and a fruitfly related? What genes do vertebrates such as fish and humans share in common, and which genes are unique to various phylogenetic lineages?

I believe that these various kinds of comparisons are what distinguish the newly emerging fields of bioinformatics and genomics from traditional biology. Biology has always concerned comparisons; in this book I quote 19th century biologists such as Richard Owen, Ernst Haeckel, and Charles Darwin who engaged in comparative studies at the organismal level. The problems we are trying to solve have not changed substantially. We still seek a more complete understanding of the unifying concepts of biology, such as the organization of life from its constituent parts (e.g., genes and proteins), the behavior of complex biological systems, and the continuity of life through evolution. What *has* changed is how we pursue this more complete understanding. This book describes databases filled with raw information on genes and gene products and the tools that are useful to analyze these data.

The Challenge of Human Disease

My training is as a molecular biologist and neuroscientist. My laboratory studies the molecular basis of childhood brain disorders such as Down syndrome, autism, and lead poisoning. We are located at the Kennedy Krieger Institute, a hospital for children for developmental disorders. (You can learn more about this Institute at http://www.kennedykrieger.org.) Each year over 10,000 patients visit the Institute. The hospital includes clinics for children with a variety of conditions including language disorders, eating disorders, autism, mental retardation, spina bifida, and traumatic brain injury. Some have very common disorders, such as Down syndrome (affecting about 1:700 live births) and mental retardation. Others have rare disorders, such as Rett syndrome or adrenoleukodystrophy.

We are at a time when the number of base pairs of DNA deposited in the world's public repositories has reached tens of billions, as described in Chapter 2. We have obtained the first sequence of the human genome, and since 1995 hundreds of genomes have been sequenced. Throughout the book, you can follow the progress of science as we learn how to sequence DNA, and study its RNA and protein products. At times the pace of progress seems dazzling.

Yet at the same time we understand so little about human disease. For thousands of diseases, a defect in a single gene causes a pathological effect. Even as we discover the genes that are defective in diseases such as cystic fibrosis, muscular dystrophy, adrenoleukodystrophy, and Rett syndrome, the path to finding an effective treatment or cure is obscure. But single gene disorders are not nearly as common as complex diseases such as autism, depression, and mental retardation that are likely due to mutations in multiple genes. And all genetic disease is not nearly as common as infectious disease. We know little about why one strain of virus infects only humans, while another closely related species infects only chimpanzees. We do not understand why one bacterial strain may be pathogenic, while another is harmless. We have not learned how to develop an effective vaccine against any eukaryotic pathogen, from protozoa (such as *Plasmodium falciparum* that causes malaria) to parasitic nematodes.

The prospects for making progress in these areas are very encouraging specifically because of the recent development of new bioinformatics tools. We are only now beginning to position ourselves to understand the genetic basis of both disease-causing agents and the hosts that are susceptible. Our hope is that the information so rapidly accumulating in new bioinformatics databases can be translated through research into insights into human disease and biology in general.

Note to Readers

This book describes over 1,000 websites related to bioinformatics and functional genomics. All of these sites evolve over time (and some become extinct). In an effort to keep the web links up-to-date, a companion website (http://www.bioinfbook.org) maintains essentially all of the website links, organized by chapter of the book. We try our best to maintain this site over time. We use a program to automatically scan all the links each month, and then we update them as necessary.

An additional site is available to instructors, including detailed solutions to problems (see http://www.wiley.com).

Acknowledgments

Writing this book has been a wonderful learning experience. It is a pleasure to thank the many people who have contributed. In particular, the intellectual environment at the Kennedy Krieger Institute and the Johns Hopkins School of Medicine has been extraordinarily rich. These chapters were developed from lectures in an introductory bioinformatics course. The Johns Hopkins faculty who lectured during its first three years were Jef Boeke (yeast functional genomics), Aravinda Chakravarti (human disease), Neil Clarke (protein structure), Kyle Cunningham (yeast), Garry Cutting (human disease), Rachel Green (RNA), Stuart Ray (molecular phylogeny), and Roger Reeves (the human genome). I have benefited greatly from their insights into these areas.

I gratefully acknowledge the many reviewers of this book, including a group of anonymous reviewers who offered extremely constructive and detailed suggestions. Those who read the book include Russ Altman, Christopher Aston, David P. Leader, and Harold Lehmann (various chapters), Conover Talbot (Chapters 2 and 18), Edie Sears (Chapter 3), Tom Downey (Chapter 7), Jef Boeke (Chapter 8 and various other chapters), Michelle Nihei and Daniel Yuan (Chapter 8), Mario Amzel and Ingo Ruczinski (Chapter 9), Stuart Ray (Chapter 11), Marie Hardwick (Chapter 13), Yukari Manabe (Chapter 14), Kyle Cunningham and Forrest Spencer (Chapter 15), and Roger Reeves (Chapter 16). Kirby D. Smith read Chapter 18 and provided insights into most of the other chapters as well. Each of these colleagues offered a great deal of time and effort to help improve the content, and each served as a mentor. Of the many students who read the chapters I mention Rong Mao, Ok-Hee Jeon, and Vinoy Prasad. I particularly thank Mayra Garcia and Larry Frelin who provided invaluable assistance throughout the writing process. I am grateful to my editor at John Wiley & Sons, Luna Han, for her encouragement.

I also acknowledge Gary W. Goldstein, President of the Kennedy Krieger Institute, and Solomon H. Snyder, my chairman in the Department of Neuroscience at Johns Hopkins. Both provided encouragement, and allowed me the opportunity to write this book while maintaining an academic laboratory.

On a personal note, I thank my family for all their love and support, as well as N. Varg, Kimberly Reed, and Charles Cohen. Most of all, I thank my fiancée Barbara Reed for her patience, faith, and love.