PREFACE

Bioinformatics Computing is a practical guide to computing in the burgeoning field of bioinformatics—the study of how information is represented and transmitted in biological systems, starting at the molecular level. This book, which is intended for molecular biologists at all levels of training and practice, assumes the reader is computer literate with modest computer skills, but has little or no formal computer science training. For example, the reader may be familiar with downloading bioinformatics data from the Web, using spreadsheets and other popular office automation tools, and/or working with commercial database and statistical analysis programs. It is helpful, but not necessary, for the reader to have some programming experience in BASIC, HTML, or C++.

In bioinformatics, as in many new fields, researchers and entrepreneurs at the fringes—where technologies from different fields interact—are making the greatest strides. For example, techniques developed by computer scientists enabled researchers at Celera Genomics, the Human Genome Project consortium, and other laboratories around the world to sequence the nearly 3 billion base pairs of the roughly 40,000 genes of the human genome. This feat would have been virtually impossible without computational methods.

No book on biotechnology would be complete without acknowledging the vast potential of the field to change life as we know it. Looking beyond the computational hurdles addressed by this text, there are broader issues and implications of biotechnology related to ethics, morality, religion, privacy, and economics. The high-stakes economic game of biotechnology pits proponents of custom medicines, genetically modified foods, cross-species cloning for spe-

cies conservation, and creating organs for transplant against those who question the bioethics of stem cell research, the wisdom of creating frankenfoods that could somehow upset the ecology of the planet, and the morality of creating clones of farm animals or pets, such as Dolly and CC, respectively.

Even the major advocates of biotechnology are caught up in bitter patent wars, with the realization that whoever has control of the key patents in the field will enjoy a stream of revenues that will likely dwarf those of software giants such as Microsoft. Rights to genetic codes have the potential to impede R&D at one extreme, and reduce commercial funding for research at the other. The resolution of these and related issues will result in public policies and international laws that will either limit or protect the rights of researchers to work in the field.

Proponents of biotechnology contend that we are on the verge of controlling the coding of living things, and concomitant breakthroughs in biomedical engineering, therapeutics, and drug development. This view is more credible especially when combined with parallel advances in nanoscience, nanoengineering, and computing. Researchers take the view that in the near future, cloning will be necessary for sustaining crops, livestock, and animal research. As the earth's population continues to explode, genetically modified fruits will offer extended shelf life, tolerate herbicides, grow faster and in harsher climates, and provide significant sources of vitamins, protein, and other nutrients. Fruits and vegetables will be engineered to create drugs to control human disease, just as bacteria have been harnessed to mass-produce insulin for diabetics. In addition, chemical and drug testing simulations will streamline pharmaceutical development and predict subpopulation response to designer drugs, dramatically changing the practice of medicine.

Few would argue that the biotechnology area presents not only scientific, but cultural and economic challenges as well. The first wave of biotechnology, which focused on medicine, was relatively well received by the public—perhaps because of the obvious benefits of the technology, as well as the lack of general knowledge of government-sponsored research in biological weapons. Instead, media stressed the benefits of genetic engineering, reporting that millions of patients with diabetes have ready access to affordable insulin.

The second wave of biotech, which focused on crops, had a much more difficult time gaining acceptance, in part because some consumers feared that engineered organisms have the potential to disrupt the ecosystem. As a result, the first genetically engineered whole food ever brought to market, the short-lived Flavr Savr™ Tomato, was an economic failure when it was introduced in the spring of 1994—only four years after the first federally approved

gene therapy on a patient. However, Calgene's entry into the market paved the way for a new industry that today holds nearly 2,000 patents on engineered foods, from virus-resistant papayas and bug-free corn, to caffeine-free coffee beans.

Today, nearly a century after the first gene map of an organism was published, we're in the third wave of biotechnology. The focus this time is on manufacturing military armaments made of transgenic spider webs, plastics from corn, and stain-removing bacilli. Because biotechnology manufacturing is still in its infancy and holds promise to avoid the pollution caused by traditional smokestack factories, it remains relatively unnoticed by opponents of genetic engineering.

The biotechnology arena is characterized by complexity, uncertainty, and unprecedented scale. As a result, researchers in the field have developed innovative computational solutions heretofore unknown or unappreciated by the general computer science community. However, in many areas of molecular biology R&D, investigators have reinvented techniques and rediscovered principles long known to scientists in computer science, medical informatics, physics, and other disciplines.

What's more, although many of the computational techniques developed by researchers in bioinformatics have been beneficial to scientists and entrepreneurs in other fields, most of these redundant discoveries represent a detour from addressing the main molecular biology challenges. For example, advances in machine-learning techniques have been redundantly developed by the microarray community, mostly independent of the traditional machine-learning research community. Valuable time has been wasted in the duplication of effort in both disciplines. The goal of this text is to provide readers with a roadmap to the diverse field of bioinformatics computing while offering enough in-depth information to serve as a valuable reference for readers already active in the bioinformatics field. The aim is to identify and describe specific information technologies in enough detail to allow readers to reason from first principles when they critically evaluate a glossy print advertisement, banner ad, or publication describing an innovative application of computer technology to molecular biology.

To appreciate the advantage of a molecular biologist studying computational methods at more than a superficial level, consider the many parallels faced by students of molecular biology and students of computer science. Most students of molecular biology are introduced to the concept of genetics through Mendel's work manipulating the seven traits of pea plants. There they learn Mendel's laws of inheritance. For example, the Law of Segregation of

Alleles states that the alleles in the parents separate and recombine in the offspring. The Law of Independent Assortment states that the alleles of different characteristics pass to the offspring independently.

Students who delve into genetics learn the limitations of Mendel's methods and assumptions—for example, that the Law of Independent Assortment applies only to pairs of alleles found on different chromosomes. More advanced students also learn that Mendel was lucky enough to pick a plant with a relatively simple genetic structure. When he extended his research to mice and other plants, his methods failed. These students also learn that Mendel's results are probably too perfect, suggesting that either his record-keeping practices were flawed or that he blinked at data that didn't fit his theories.

Just as students of genetics learn that Mendel's experiment with peas isn't adequate to fully describe the genetic structures of more complex organisms, students of computer science learn the exceptions and limitations of the strategies and tactics at their disposal. For example, computer science students are often introduced to algorithms by considering such basic operations as sorting lists of data.

To computer users who are unfamiliar with underlying computer science, sorting is simply the process of rearranging an unordered sequence of records into either ascending or descending order according to one or more keys—such as the name of a protein. However, computer scientists and others have developed dozens of searching algorithms, each with countless variations to suit specific needs. Because sorting is a fundamental operation used in everything from searching the Web to analyzing and matching patterns of base pairs, it warrants more than a superficial understanding for a biotechnology researcher engaged in operations that involve sorting.

Consider that two of the most popular sorting algorithms used in computer science, quicksort and bubblesort, can be characterized by a variety of factors, from stability and running time to memory requirements, and how performance is influenced by the way in which memory is accessed by the host computer's central processing unit. That is, just as Mendel's experiments and laws have exceptions and operating assumptions, a sorting algorithm can't simply be taken at face value.

For example, the running time of quicksort on large data sets is superior to that of many other stable sorting algorithms, such as bubblesort. Sorting long lists of a half-million elements or more with a program that implements the bubblesort algorithm might take an hour or more, compared to a half-second for a program that follows the quicksort algorithm. Although the performance of quicksort is nearly identical to that of bubblesort on a few hundred

or thousand data elements, the performance of bubblesort degrades rapidly with increasing data size. When the size of the data approaches the number of base pairs in the human genome, a sort that takes 5 or 10 seconds using quicksort might require half a day or more on a typical desktop PC.

Even with its superb performance, quicksort has many limitations that may favor bubblesort or another sorting algorithm, depending on the nature of the data, the limitations of the hardware, and the expertise of the programmer. For example, one virtue of the bubblesort algorithm is simplicity. It can usually be implemented by a programmer in any number of programming languages, even one who is a relative novice. In operation, successive sweeps are made through the records to be sorted and the largest record is moved closer to the top, rising like a bubble.

In contrast, the relatively complex quicksort algorithm divides records into two partitions around a pivot record, and all records that are less than the pivot go into one partition and all records that are greater go into the other. The process continues recursively in each of the two partitions until the entire list of records is sorted. While quicksort performs much better than bubblesort on long lists of data, it generally requires significantly more memory space than the bubblesort. With very large files, the space requirements may exceed the amount of free RAM available on the researcher's PC. The bubblesort versus quicksort dilemma exemplifies the common tradeoff in computer science of space for speed.

Although the reader may never write a sorting program, knowing when to apply one algorithm over another is useful in deciding which shareware or commercial software package to use or in directing a programmer to develop a custom system. A parallel in molecular biology would be to know when to describe an organism using classical Mendelian genetics, and when other mechanisms apply.

Given the multidisciplinary characteristic of bioinformatics, there is a need in the molecular biology community for reference texts that illustrate the computer science advances that have been made in the past several decades. The most relevant areas—the ones that have direct bearing on their research—are in computer visualization, very large database designs, machine learning and other forms of advanced pattern-matching, statistical methods, and distributed-computing techniques. This book, which is intended to bring molecular biologists up to speed in computational techniques that apply directly to their work, is a direct response to this need.

ORGANIZATION OF THIS BOOK

This book is organized into modular, stand-alone topics related to bioinformatics computing according to the following chapters:

- *Chapter 1: THE CENTRAL DOGMA*
  This chapter provides an overview of bioinformatics, using the Central Dogma as the organizing theme. It explores the relationship of molecular biology and bioinformatics to computer science, and how the purview of computational bioinformatics necessarily extends from the molecular to the clinical medicine level.
- *Chapter 2: DATABASES*
  Bioinformatics is characterized by an abundance of data stored in very large databases. The practical computer technologies related to very large databases are discussed, with an emphasis on object-oriented database methods, given that traditional relational database technology may be ill-suited for some bioinformatics needs. Data warehousing, data dictionaries, database design, and knowledge management techniques related to bioinformatics are also discussed in detail.
- *Chapter 3: NETWORKS*
  This chapter explores the information technology infrastructure of bioinformatics, including the Internet, World Wide Web, intranets, wireless systems, and other network technologies that apply directly to sharing, manipulating, and archiving sequence data and other bioinformatics information. This chapter reviews Web-based resources for researchers, such as GenBank and other systems maintained by NCBI, NIH, and other government agencies. The Great Global Grid and its potential for transforming the field of bioinformatics is also discussed.
- *Chapter 4: SEARCH ENGINES*
  The exponentially increasing amounts of data accessible in digital form over the Internet, from gene sequences to published references to the experimental methods used to determine specific sequences, is only accessible through advanced search engine technologies. This chapter details search engine operations related to the major online bioinformatics resources.
- *Chapter 5: DATA VISUALIZATION*
  Exploring the possible configurations of folded proteins has proven to be virtually impossible by simply studying linear sequences of bases. However, sophisticated 3D visualization

techniques allow researchers to use their visual and spatial reasoning abilities to understand the probable workings of proteins and other structures. This chapter explores data visualization techniques that apply to bioinformatics, from methods of generating 2D and 3D renderings of protein structures to graphing the results of the statistical analysis of protein structures.

- *Chapter 6: STATISTICS*
  The randomness inherent in any sampling process—such as measuring the mRNA levels of thousands of genes simultaneously with microarray techniques, or assessing the similarity between protein sequences—necessarily involves probability and statistical methods. This chapter provides an in-depth discussion of the statistical techniques applicable to molecular biology, addressing topics such as statistical analysis of structural features, gene prediction, how to extract maximal value from small sample sets, and quantifying uncertainty in sequencing results.
- *Chapter 7: DATA MINING*
  Given an ever-increasing store of sequence and protein data from several ongoing genome projects, data mining the sequences has become a field of research in its own right. Many bioinformatics scientists conduct important research from their PCs, without ever entering a wet lab or seeing a sequencing machine. The aim of this chapter is to explore data-mining techniques, using technologies, such as the Perl language, that are uniquely suited to searching through data strings. Other issues covered include taxonomies, profiling sequences, and the variety of tools available to researchers involved in mining the data in GenBank and other very large bioinformatics databases.
- *Chapter 8: PATTERN MATCHING*
  Expert systems and classical pattern matching or AI techniques—from reasoning under uncertainty and machine learning to image and pattern recognition—have direct, practical applicability to molecular biology research. This chapter covers a variety of pattern-matching approaches, using molecular biology as a working context. For example, microarray research lends itself to machine learning, in that it is humanly impossible to follow thousands of parallel reactions unaided, and several gene-prediction applications are based on neural network pattern-matching engines. The strengths and weakness of various pattern-matching approaches in bioinformatics are discussed.

- *Chapter 9: MODELING AND SIMULATION*
  This chapter covers a variety of simulation techniques, in the context of computer modeling events from drug-protein interactions and probable protein folding configurations to the analysis of potential biological pathways. The application of event-driven, time-driven, and hybrid simulation techniques are discussed, as well as linking computer simulations with visualization techniques.
- *Chapter 10: COLLABORATION*
  Bioinformatics is characterized by a high degree of cooperation among the researchers who contribute their part to the whole knowledge base of genomics and proteomics. As such, this chapter explores the details of collaboration with enabling technologies that facilitate multimedia communications, real-time videoconferencing, and Web-based application sharing of molecular biology information and knowledge.

HOW TO USE THIS BOOK

For readers new to bioinformatics, the best way to tackle the subject is to simply read each chapter in order; however, because each chapter is written as a stand-alone module, readers interested in, for example, data-mining techniques, can go directly to Chapter 7, "Data Mining." Where appropriate, "On the Horizon" sidebars provide glimpses of techniques and technologies that hold promise but have either not been fully developed or have yet to be embraced by the bioinformatics community. In addition, readers who want to delve deeper into bioinformatics are encouraged to refer to the list of publications and Web sites listed in the Bibliography.

THE LARGER CONTEXT

Bioinformatics may not be able solve the numerous social, ethical, and legal issues in the field of biotechnology, but it can address many of the scientific and economic issues. For example, there are technical hurdles to be overcome before advances such as custom pharmaceuticals and cures for genetic diseases can be affordable and commonplace. Many of these advances will require new technologies in molecular biology, and virtually all of these advances will be enabled by computational methods. For example, most molecular biologists concede that sequencing the human genome was a rela-

tively trivial task compared to the challenges of understanding the human proteome. The typical cell produces hundreds of thousands of proteins, many of which are unknown and of unknown function. What's more, these proteins fold into shapes that are a function of the linear sequence of amino acids they contain, the temperature, as well as the presence of fats, sugars, and water in the microenvironment.

As the history of the PC and the Internet has demonstrated, the rate of change in technological innovation is accelerating, and the practical applications of computing to unravel the proteome and other bioinformatics challenges are growing exponentially. In this regard, bioinformatics should be considered an empowering technology with which researchers in biotechnology can take a proactive role in defining and shaping the future of their field—and the world.