Preface

Computational biology is an unusual science, not because it lies at the intersection of two distinctly different disciplines, but because its development is strongly dependent on simultaneous technical advances in both areas—information technology and molecular biology. Moreover, molecular biology at its core has always been an information science built on an ever-growing list of code words, sequences, patterns, three-dimensional structures, and the general flow of information from gene to protein to complete organism.

Today we take it for granted that biological information is carried in gene sequences, which are ultimately transcribed into messages that form the template for protein synthesis. However, 50 years ago, when James Watson and Francis Crick published their landmark paper describing the three-dimensional structure of DNA, none of these concepts were obvious. The DNA code words had yet to be deciphered, there was no hint of a messenger molecule, and the mechanisms that underlie protein synthesis were still a mystery. During the ensuing years, some of the most ingenious scientific experiments of our time were used to solve these problems. The system that emerged is remarkably straightforward at the level of an overview, yet enormously complex and flexible. One might expect such a system to embody many levels of subtlety because it must be capable of storing enough information, in a very small space, to code for a living creature.

Recent advances in information technology coupled with new chemical and physical techniques have allowed researchers to capitalize on this understanding of molecular

biology. The result has been a new generation of tools for studying gene and protein sequences, macromolecular structures, and the interactions that comprise metabolic systems. These tools, both computational and physical, are the subject of this book.

APPROACH AND ORGANIZATION

Throughout the book I attempt to present an equal mix of biology and computer science. The goal has been to create a framework for thinking about molecular biology as an information science. In this sense the current work is unique. Its design parallels the flow of biological information from gene to message to protein and finally to metabolic system. I've also chosen to extend this theme by including a section on the emerging field of information-based medicine. Because it builds on topics presented throughout the book, such as database infrastructure design and transcriptional profiling, this chapter is presented near the end. However, some have commented that they skipped to this section first, and I have tried to make the information as accessible as possible to the first-time reader. Moreover, the book is intended as an in-depth introduction to many areas of computational biology, and you are encouraged to pursue specific subject areas in greater depth through the scientific literature.

I tried to focus on a number of topics that have enormous biological impact but have been almost universally overlooked elsewhere. Most significant among these is the discussion about minor messages—those expressed in vanishingly small quantities within a cell. It has recently come to light that the majority of intracellular mRNA species are present in very small copy counts and that, despite their scarcity, these messages can have important implications for health and disease. That said, the strengths and weaknesses of various transcriptional profiling techniques emerges as an important topic because some are specifically designed to provide accurate copy counts at very low levels. As a result, I have included a lengthy section on technologies for transcriptional profiling with the goal of presenting a balanced view of the strengths and weaknesses of each. I have also included a section on neural networks and pattern-discovery algorithms as an adjunct to the more common statistical techniques used to analyze such data.

AUDIENCE

When developing the content, I attempted to create a reference with broad appeal. In that regard, the discussions of gene and protein structure, transcription, and translation are perfectly reasonable standalone tutorials on those topics. The introductory sections on gene and protein sequence databases, basic bioinformatic tools, and transcription profiling share the same design characteristic; they should be accessible to anyone with a basic knowledge of biological processes. Generally speaking, the book was designed for undergraduate and graduate students entering the field of bioinformatics or professionals with a need to understand bioinformatics and a desire to frame that understanding in a biological context. Conversely, the chapter on information-based medicine is designed to be helpful to doctors and other clinical professionals who are in the process of developing next-generation computer infrastructure for delivering medical care. This portion of the book is also likely to appeal to pharmaceutical and insurance company executives whose organizations are core components of the healthcare delivery system. I hope that the content will prove helpful to those who already have strong backgrounds in biology and computation and are seeking a reference with a broad base of coverage across the field. For those readers, I have included information about many new developments, such as high-throughput whole genome sequencing, systems biology, and presymptomatic disease prediction. These topics are relatively new and usually covered only in the scientific literature.