Preface

We have been very pleased, beyond our expectations, with the reception of the first edition of this book. Bioinformatics, however, continues to evolve very rapidly, hence the need for a new edition. In the past three years, full-genome sequencing has blossomed with the completion of the sequence of the fly and the first draft of the Human Genome Project. In addition, several other high-throughput/combinatorial technologies, such as DNA microarrays and mass spectrometry, have considerably progressed. Altogether, these high-throughput technologies are capable of rapidly producing terabytes of data that are too overwhelming for conventional biological approaches. As a result, the need for computer/statistical/machine learning techniques is today *stronger* rather than weaker.

Bioinformatics in the Post-genome Era

In all areas of biological and medical research, the role of the computer has been dramatically enhanced in the last five to ten year period. While the first wave of computational analysis did focus on sequence analysis, where many highly important unsolved problems still remain, the current and future needs will in particular concern sophisticated *integration* of extremely diverse sets of data. These novel types of data originate from a variety of experimental techniques of which many are capable of data production at the levels of entire cells, organs, organisms, or even populations.

The main driving force behind the changes has been the advent of new, efficient experimental techniques, primarily DNA sequencing, that have led to an exponential growth of linear descriptions of protein, DNA and RNA molecules. Other new data producing techniques work as massively parallel versions of traditional experimental methodologies. Genome-wide gene expression measurements using DNA microrarrays is, in essence, a realization of tens of thousands of Northern blots. As a result, computational support in experiment design, processing of results and interpretation of results has become essential.

These developments have greatly widened the scope of bioinformatics.

As genome and other sequencing projects continue to advance unabated, the emphasis progressively switches from the accumulation of data to its interpretation. Our ability in the future to make new biological discoveries will depend strongly on our ability to combine and correlate diverse data sets along multiple dimensions and scales, rather than a continued effort focused in traditional areas. Sequence data will have to be integrated with structure and function data, with gene expression data, with pathways data, with phenotypic and clinical data, and so forth. Basic research within bioinformatics will have to deal with these issues of *system* and *integrative* biology, in the situation where the amount of data is growing exponentially.

The large amounts of data create a critical need for theoretical, algorithmic, and software advances in storing, retrieving, networking, processing, analyzing, navigating, and visualizing biological information. In turn, biological systems have inspired computer science advances with new concepts, including genetic algorithms, artificial neural networks, computer viruses and synthetic immune systems, DNA computing, artificial life, and hybrid VLSI-DNA gene chips. This cross-fertilization has enriched both fields and will continue to do so in the coming decades. In fact, all the boundaries between carbon-based and silicon-based information processing systems, whether conceptual or material, have begun to shrink [29].

Computational tools for classifying sequences, detecting weak similarities, separating protein coding regions from non-coding regions in DNA sequences, predicting molecular structure, post-translational modification and function, and reconstructing the underlying evolutionary history have become an essential component of the research process. This is essential to our understanding of life and evolution, as well as to the discovery of new drugs and therapies. Bioinformatics has emerged as a strategic discipline at the frontier between biology and computer science, impacting medicine, biotechnology, and society in many ways.

Large databases of biological information create both challenging data-mining problems and opportunities, each requiring new ideas. In this regard, conventional computer science algorithms have been useful, but are increasingly unable to address many of the most interesting sequence analysis problems. This is due to the inherent complexity of biological systems, brought about by evolutionary tinkering, and to our lack of a comprehensive theory of life's organization at the molecular level. Machine-learning approaches (e.g. neural networks, hidden Markov models, vector support machines, belief networks), on the other hand, are ideally suited for domains characterized by the presence of large amounts of data, "noisy" patterns, and the absence of general theories. The fundamental idea behind these approaches is to *learn the theory automatically from the data*, through a process of inference, model

fitting, or learning from examples. Thus they form a viable complementary approach to conventional methods. The aim of this book is to present a broad overview of bioinformatics from a *machine-learning perspective*.

Machine-learning methods are computationally intensive and benefit greatly from progress in computer speed. It is remarkable that both computer speed and sequence volume have been growing at roughly the same rate since the late 1980s, doubling every 16 months or so. More recently, with the completion of the first draft of the Human Genome Project and the advent of high-throughput technologies such as DNA microarrays, biological data has been growing even faster, doubling about every 6 to 8 months, and further increasing the pressure towards bioinformatics. To the novice, machine-learning methods may appear as a bag of unrelated techniques—but they are not. On the theoretical side, a unifying framework for all machine-learning methods also has emerged since the late 1980s. This is the Bayesian probabilistic framework for modeling and inference. In our minds, in fact, there is little difference between machine learning and Bayesian modeling and inference, except for the emphasis on computers and number crunching implicit in the first term. It is the confluence of all three factors—data, computers, and theoretical probabilistic framework—that is fueling the machine-learning expansion, in bioinformatics and elsewhere. And it is fair to say that bioinformatics and machine learning methods have started to have a significant impact in biology and medicine.

Even for those who are not very sensitive to mathematical rigor, modeling biological data probabilistically makes eminent sense. One reason is that biological measurements are often inherently "noisy", as is the case today of DNA microarray or mass spectrometer data. Sequence data, on the other hand, is becoming noise free due to its discrete nature and the cost-effectiveness of repeated sequencing. Thus measurement noise cannot be the sole reason for modeling biological data probabilistically. The real need for modeling biological data probabilistically comes from the complexity and variability of biological systems brought about by eons of evolutionary tinkering in complex environments. As a result, biological systems have inherently a very high dimensionality. Even in microarray experiments where expression levels of thousands of genes are measured simultaneously, only a small subset of the relevant variables is being observed. The majority of the variables remain “hidden” and must be factored out through probabilistic modeling. Going directly to a systematic probabilistic framework may contribute to the acceleration of the discovery process by avoiding some of the pitfalls observed in the history of sequence analysis, where it took several decades for probabilistic models to emerge as the proper framework.

An often-met criticism of machine-learning techniques is that they are “black box” approaches: one cannot always pin down exactly how a complex

neural network, or hidden Markov model, reaches a particular answer. We have tried to address such legitimate concerns both within the general probabilistic framework and from a practical standpoint. It is important to realize, however, that many other techniques in contemporary molecular biology are used on a purely empirical basis. The polymerase chain reaction, for example, for all its usefulness and sensitivity, is still somewhat of a black box technique. Many of its adjustable parameters are chosen on a trial-and-error basis. The movement and mobility of sequences through matrices in gels is another area where the pragmatic success and usefulness are attracting more attention than the lack of detailed understanding of the underlying physical phenomena. Also, the molecular basis for the pharmacological effect of most drugs remains largely unknown. Ultimately the proof is in the pudding. We have striven to show that machine-learning methods yield good puddings and are being elegant at the same time.

Audience and Prerequisites

The book is aimed at both students and more advanced researchers, with diverse backgrounds. We have tried to provide a succinct description of the main biological concepts and problems for the readers with a stronger background in mathematics, statistics, and computer science. Likewise, the book is tailored to the biologists and biochemists who will often know more about the biological problems than the text explains, but need some help to understand the new data-driven algorithms, in the context of biological data. It should in principle provide enough insights while remaining sufficiently simple for the reader to be able to implement the algorithms described, or adapt them to a particular problem. The book, however, does not cover the informatics needed for the management of large databases and sequencing projects, or the processing of raw fluorescence data. The technical prerequisites for the book are basic calculus, algebra, and discrete probability theory, at the level of an undergraduate course. Any prior knowledge of DNA, RNA, and proteins is of course helpful, but not required.

Content and General Outline of the Book

We have tried to write a comprehensive but reasonably concise introductory book that is self-contained. The book includes definitions of main concepts and proofs of main theorems, at least in sketched form. Additional technical details can be found in the appendices and the references. A significant portion of the book is built on material taken from articles we have written over

the years, as well as from tutorials given at several conferences, including the ISMB (Intelligent Systems for Molecular Biology) conferences, courses given at the Technical University of Denmark and UC Irvine, and workshops organized during the NIPS (Neural Information Processing Systems) conference. In particular, the general Bayesian probabilistic framework that is at the core of the book has been presented in several ISMB tutorials starting in 1994.

The main focus of the book is on methods, not on the history of a rapidly evolving field. While we have tried to quote the relevant literature in detail, we have concentrated our main effort on presenting a number of techniques, and perhaps a general way of thinking that we hope will prove useful. We have tried to illustrate each method with a number of results, often but not always drawn from our own practice.

Chapter 1 provides an introduction to sequence data in the context of molecular biology, and to sequence analysis. It contains in particular an overview of genomes and proteomes, the DNA and protein "universes" created by evolution that are becoming available in the public databases. It presents an overview of genomes and their sizes, and other comparative material that, if not original, is hard to find in other textbooks.

Chapter 2 is the most important theoretical chapter, since it lays the foundations for all machine-learning techniques, and shows explicitly how one must reason in the presence of uncertainty. It describes a general way of thinking about sequence problems: the Bayesian statistical framework for inference and induction. The main conclusion derived from this framework is that the proper language for machine learning, and for addressing all modeling problems, is the language of probability theory. All models *must* be probabilistic. And probability theory is all one needs for a scientific discourse on models and on their relationship to the data. This uniqueness is reflected in the title of the book. The chapter briefly covers classical topics such as priors, likelihood, Bayes theorem, parameter estimation, and model comparison. In the Bayesian framework, one is mostly interested in probability distributions over high-dimensional spaces associated, for example, with data, hidden variables, and model parameters. In order to handle or approximate such probability distributions, it is useful to exploit independence assumptions as much as possible, in order to achieve simpler factorizations. This is at the root of the notion of graphical models, where variable dependencies are associated with graph connectivity. Useful tractable models are associated with relatively sparse graphs. Graphical models and a few other techniques for handling high-dimensional distributions are briefly introduced in Chapter 2 and further elaborated in Appendix C. The inevitable use of probability theory and (sparse) graphical models are really the two central ideas behind all the methods.

Chapter 3 is a warm-up chapter, to illustrate the general Bayesian probabilistic framework. It develops a few classical examples in some detail which

are used in the following chapters. It can be skipped by anyone familiar with such examples, or during a first quick reading of the book. All the examples are based on the idea of generating sequences by tossings one or several dices. While such a dice model is extremely simplistic, it is fair to say that a substantial portion of this book, Chapters 7–12, can be viewed as various generalizations of the dice model. Statistical mechanics is also presented as an elegant application of the dice model within the Bayesian framework. In addition, statistical mechanics offers many insights into different areas of machine learning. It is used in particular in Chapter 4 in connection with a number of algorithms, such as Monte Carlo and EM (expectation maximization) algorithms.

Chapter 4 contains a brief treatment of many of the basic algorithms required for Bayesian inference, machine learning, and sequence applications, in order to compute expectations and optimize cost functions. These include various forms of dynamic programming, gradient-descent and EM algorithms, as well as a number of stochastic algorithms, such as Markov chain Monte Carlo (MCMC) algorithms. Well-known examples of MCMC algorithms are described, such as Gibbs sampling, the Metropolis algorithm, and simulated annealing. This chapter can be skipped in a first reading, especially if the reader has a good acquaintance with algorithms and/or is not interested in implementing such algorithms.

Chapters 5–9 and Chapter 12 form the core of the book. Chapter 5 provides an introduction to the theory of neural networks. It contains definitions of the basic concepts, a short derivation of the "backpropagation" learning algorithm, as well as a simple proof of the fact that neural networks are universal approximators. More important, perhaps, it describes how neural networks, which are often introduced without any reference to probability theory, are in fact best viewed within the general probabilistic framework of Chapter 2. This in turn yields useful insights on the design of neural architectures and the choice of cost functions for learning.

Chapter 6 contains a selected list of applications of neural network techniques to sequence analysis problems. We do not attempt to cover the hundreds of applications produced so far, but have selected seminal examples where advances in the methodology have provided significant improvements over other approaches. We especially treat the issue of optimizing training procedures in the sequence context, and how to combine networks to form more complex and powerful algorithms. The applications treated in detail include protein secondary structure, signal peptides, intron splice sites, and gene-finding.

Chapters 7 and 8, on hidden Markov models, mirror Chapters 5 and 6. Chapter 7 contains a fairly detailed introduction to hidden Markov models (HMMs), and the corresponding dynamic programming algorithms (forward,

backward, and Viterbi algorithms) as well as learning algorithms (EM, gradient-descent, etc.). Hidden Markov models of biological sequences can be viewed as generalized dice models with insertions and deletions.

Chapter 8 contains a selected list of applications of hidden Markov models to both protein and DNA/RNA problems. It demonstrates, first, how HMMs can be used, among other things, to model protein families, derive large multiple alignments, classify sequences, and search large databases of complete or fragment sequences. In the case of DNA, we show how HMMs can be used in gene-finding (promoters, exons, introns) and gene-parsing tasks.

HMMs can be very effective, but they have their limitations. Chapters 9–11 can be viewed as extensions of HMMs in different directions. Chapter 9 uses the theory of probabilistic graphical models systematically both as a unifying concept and to derive new classes of models, such as hybrid models that combine HMMs with artificial neural networks, or bidirectional Markov models that exploit the spatial rather than temporal nature of biological sequences. The chapter includes applications to gene-finding, analysis of DNA symmetries, and prediction of protein secondary structure.

Chapter 10 presents phylogenetic trees and, consistent with the framework of Chapter 2, the inevitable underlying probabilistic models of evolution. The models discussed in this chapter and throughout the book can be viewed as generalizations of the simple dice models of Chapter 3. In particular, we show how tree reconstruction methods that are often presented in a nonprobabilistic context (i.e., parsimony methods) are in fact a special case of the general framework as soon as the underlying probabilistic model they approximate is made explicit.

Chapter 11 covers formal grammars and the Chomsky hierarchy. Stochastic grammars provide a new class of models for biological sequences, which generalize both HMMs and the simple dice model. Stochastic regular grammars are in fact equivalent to HMMs. Stochastic context-free grammars are more powerful and roughly correspond to dice that can produce pairs of letters rather than single letters. Applications of stochastic grammars, especially to RNA modeling, are briefly reviewed.

Chapter 12 focuses primarily on the analysis of DNA microarray gene expression data, once again by generalizing the die model. We show how the Bayesian probabilistic framework can be applied systematically to array data. In particular, we treat the problems of establishing whether a gene behaves differently in a treatment versus control situation and of gene clustering. Analysis of regulatory regions and inference of gene regulatory networks are discussed briefly.

Chapter 13 contains an overview of current database resources and other information that is publicly available over the Internet, together with a list of useful directions to interesting WWW sites and pointers. Because these

resources are changing rapidly, we focus on general sites where information is likely to be updated regularly. However, the chapter contains also a pointer to a page that contains regularly-updated links to all the other sites.

The book contains in appendix form a few technical sections that are important for reference and for a thorough understanding of the material. Appendix A covers statistical notions such as errors bars, sufficient statistics, and the exponential family of distributions. Appendix B focuses on information theory and the fundamental notions of entropy, mutual information, and relative entropy. Appendix C provides a brief overview of graphical models, independence, and Markov properties, in both the undirected case (random Markov fields) and the directed case (Bayesian networks). Appendix D covers technical issues related to hidden Markov models, such as scaling, loop architectures, and bendability. Finally, appendix E briefly reviews two related classes of machine learning models of growing importance, Gaussian processes and support vector machines. A number of exercises are also scattered throughout the book: from simple proofs left to the reader to suggestions for possible extensions.

For ease of exposition, standard assumptions of positivity or differentiability are sometimes used implicitly, but should be clear from the context.

What Is New and What Is Omitted

On several occasions, we present new unpublished material or old material but from a somewhat new perspective. Examples include the discussion around MaxEnt and the derivation of the Boltzmann–Gibbs distribution in Chapter 3, the application of HMMs to fragments, to promoters, to hydropathy profiles, and to bendability profiles in Chapter 8, the analysis of parsimony methods in probabilistic terms, the higher-order evolutionary models in Chapter 10, and the Bayesian analysis of gene differences in microarray data. The presentation we give of the EM algorithm in terms of free energy is not widely known and, to the best of our knowledge, was first described by Neal and Hinton in an unpublished technical report.

In this second edition we have benefited from and incorporated the feedback received from many colleagues, students, and readers. In addition to revisions and updates scattered throughout the book to reflect the fast pace of discovery set up by complete genome sequencing and other high-throughput technologies, we have included a few more substantial changes.

These include:

- New section on the human genome sequence in Chapter 1.
- New sections on protein function and alternative splicing in Chapter 1.

- New neural network applications in Chapter 6.
- A completely revised Chapter 9, which now focuses systematically on graphical models and their applications to bioinformatics. In particular, this chapter contains entirely new section about gene finding, and the use of recurrent neural networks for the prediction of protein secondary structure.
- A new chapter (Chapter 12) on DNA microarray data and gene expression.
- A new appendix (Appendix E) on support vector machines and Gaussian processes.

The book material and treatment reflect our personal biases. Many relevant topics had to be omitted in order to stay within reasonable size limits. At the theoretical level, we would have liked to be able to go more into higher levels of Bayesian inference and Bayesian networks. Most of the book in fact could have been written using Bayesian networks only, providing an even more unified treatment, at some additional abstraction cost. At the biological level, our treatment of phylogenetic trees, for example, could easily be expanded and the same can be said of the section on DNA microarrays and clustering (Chapter 12). In any case, we have tried to provide ample references where complementary information can be found.

Vocabulary and Notation

Terms such as "bioinformatics," "computational biology," "computational molecular biology," and "biomolecular informatics" are used to denote the field of interest of this book. We have chosen to be flexible and use all those terms essentially in an interchangeable way, although one should not forget that the first two terms are extremely broad and could encompass entire areas not directly related to this book, such as the application of computers to model the immune system, or the brain. More recently, the term "computational molecular biology" has also been used in a completely different sense, similar to "DNA computing," to describe attempts to build computing devices out of biomolecules rather than silicon. The adjective "artificial" is also implied whenever we use the term "neural network" throughout the book. We deal with artificial neural networks from an algorithmic-pattern-recognition point of view only.

And finally, a few words on notation. Most of the symbols used are listed at the end of the book. In general, we do not systematically distinguish between scalars, vectors, and matrices. A symbol such as "D" represents the data, regardless of the amount or complexity. Whenever necessary, vectors should be

regarded as column vectors. Boldface letters are usually reserved for probabilistic concepts, such as probability (**P**), expectation (**E**), and variance (**Var**). If X is a random variable, we write $\mathbf{P}(x)$ for $\mathbf{P}(X = x)$, or sometimes just $\mathbf{P}(X)$ if no confusion is possible. Actual distributions are denoted by P, Q, R, and so on.

We deal mostly with discrete probabilities, although it should be clear how to extend the ideas to the continuous case whenever necessary. Calligraphic style is reserved for particular functions, such as the energy ($\mathcal{E}$) and the entropy ($\mathcal{H}$). Finally, we must often deal with quantities characterized by many indices. A connection weight in a neural network may depend on the units, i and j, it connects; its layer, l; the time, t, during the iteration of a learning algorithm; and so on. Within a given context, only the most relevant indices are indicated. On rare occasions, and only when confusion is extremely unlikely, the same symbol is used with two different meanings (for instance, D denotes also the set of delete states of an HMM).

Acknowledgments

Over the years, this book has been supported by the Danish National Research Foundation and the National Institutes of Health. SmithKline Beecham Inc. sponsored some of the work on fragments at Net-ID. Part of the book was written while PB was in the Division of Biology, California Institute of Technology. We also acknowledge support from Sun Microsystems and the Institute for Genomics and Bioinformatics at UCI.

We would like to thank all the people who have provided feedback on early versions of the manuscript, especially Jan Gorodkin, Henrik Nielsen, Anders Gorm Pedersen, Chris Workman, Lars Juhl Jensen, Jakob Hull Kristensen, and David Ussery. Yves Chauvin and Van Mittal-Henkle at Net-ID, and all the members of the Center for Biological Sequence Analysis, have been instrumental to this work over the years in many ways.

We would like also to thank Chris Bishop, Richard Durbin, and David Haussler for inviting us to the Isaac Newton Institute in Cambridge, where the first edition of this book was finished, as well as the Institute itself for its great environment and hospitality. Special thanks to Geeske de Witte, Johanne Keiding, Kristoffer Rapacki, Hans Henrik Stærfeldt and Peter Busk Laursen for superb help in turning the manuscript into a book.

For the second edition, we would like to acknowledge new colleagues and students at UCI including Pierre-François Baisnée, Lee Bardwell, Thomas Briese, Steven Hampson, G. Wesley Hatfield, Dennis Kibler, Brandon Gaut, Richard Lathrop, Ian Lipkin, Anthony Long, Larry Marsh, Calvin McLaughlin, James Nowick, Michael Pazzani, Gianluca Pollastri, Suzanne Sandmeyer, and

Padhraic Smyth. Outside of UCI, we would like to acknowledge Russ Altman, Mark Borodovsky, Mario Blaum, Doug Brutlag, Chris Burge, Rita Casadio, Piero Fariselli, Paolo Frasconi, Larry Hunter, Emeran Mayer, Ron Meir, Burkhard Rost, Pierre Rouze, Giovanni Soda, Gary Stormo, and Gill Williamson.

We also thank the series editor Thomas Dietterich and the staff at MIT Press, especially Deborah Cantor-Adams, Ann Rae Jonas, Yasuyo Iguchi, Ori Kometani, Katherine Innis, Robert Prior, and the late Harry Stanton, who was instrumental in starting this project. Finally, we wish to acknowledge the support of all our friends and families.