
Preface

We are drowning in information,
but starved of knowledge.

– John Naisbitt, *Megatrends*

The turn of the millennium has been described as the dawn of a new scientific revolution, which will have as great an impact on society as the industrial and computer revolutions before. This revolution was heralded by a large-scale DNA sequencing effort in July 1995, when the entire 1.8 million base pairs of the genome of the bacterium *Haemophilus influenzae* was published – the first of a free-living organism. Since then, the amount of DNA sequence data in publicly accessible data bases has been growing exponentially, including a working draft of the complete 3.3 billion base-pair DNA sequence of the entire human genome, as pre-released by an international consortium of 16 institutes on June 26, 2000.

Besides genomic sequences, new experimental technologies in molecular biology, like microarrays, have resulted in a rich abundance of further data, related to the transcriptome, the spliceosome, the proteome, and the metabolome. This explosion of the “omes” has led to a paradigm shift in molecular biology. While pre-genomic biology followed a hypothesis-driven reductionist approach, applying mainly qualitative methods to small, isolated systems, modern post-genomic molecular biology takes a holistic, systems-based approach, which is data-driven and increasingly relies on quantitative methods. Consequently, in the last decade, the new scientific discipline of *bioinformatics* has emerged in an attempt to interpret the increasing amount of molecular biological data. The problems faced are essentially statistical, due to the inherent complexity and stochasticity of biological systems, the random processes intrinsic to evolution, and the unavoidable error-proneness and variability of measurements in large-scale experimental procedures.

Since we lack a comprehensive theory of life's organization at the molecular level, our task is to learn the theory by induction, that is, to extract patterns from large amounts of noisy data through a process of statistical inference based on model fitting and learning from examples.

Medical informatics is the study, development, and implementation of algorithms and systems to improve communication, understanding, and management of medical knowledge and data. It is a multi-disciplinary science at the junction of medicine, mathematics, logic, and information technology, which exists to improve the quality of health care.

In the 1970s, only a few computer-based systems were integrated with hospital information. Today, computerized medical-record systems are the norm within the developed countries. These systems enable fast retrieval of patient data; however, for many years, there has been interest in providing additional decision support through the introduction of knowledge-based systems and statistical systems.

A problem with most of the early clinically-oriented knowledge-based systems was the adoption of ad hoc rules of inference, such as the use of certainty factors by MYCIN. Another problem was the so-called knowledge-acquisition bottleneck, which referred to the time-consuming process of eliciting knowledge from domain experts. The renaissance in neural computation in the 1980s provided a purely data-based approach to probabilistic decision support, which circumvented the need for knowledge acquisition and augmented the repertoire of traditional statistical techniques for creating probabilistic models.

The 1990s saw the maturity of Bayesian networks. These networks provide a sound probabilistic framework for the development of medical decision-support systems from knowledge, from data, or from a combination of the two; consequently, they have become the focal point for many research groups concerned with medical informatics.

As far as the methodology is concerned, the focus in this book is on probabilistic graphical models and Bayesian networks. Many of the earlier methods of data analysis, both in bioinformatics and in medical informatics, were quite ad hoc. In recent years, however, substantial progress has been made in our understanding of and experience with probabilistic modelling. Inference, decision making, and hypothesis testing can all be achieved if we have access to conditional probabilities. In real-world scenarios, however, it may not be clear what the conditional relationships are between variables that are connected in some way. Bayesian networks are a mixture of graph theory and probability theory and offer an elegant formalism in which problems can be portrayed and conditional relationships evaluated. Graph theory provides a framework to represent complex structures of highly-interacting sets of variables. Probability theory provides a method to infer these structures from observations or measurements in the presence of noise and uncertainty. This method allows a system of interacting quantities to be visualized as being composed of sim-

pler subsystems, which improves model transparency and facilitates system interpretation and comprehension.

Many problems in computational molecular biology, bioinformatics, and medical informatics can be treated as particular instances of the general problem of learning Bayesian networks from data, including such diverse problems as DNA sequence alignment, phylogenetic analysis, reverse engineering of genetic networks, respiration analysis, Brain-Computer Interfacing and human sleep-stage classification as well as drug discovery.

Organization of This Book

The first part of this book provides a brief yet self-contained introduction to the methodology of Bayesian networks. The following parts demonstrate how these methods are applied in bioinformatics and medical informatics.

This book is by no means comprehensive. All three fields – the methodology of probabilistic modeling, bioinformatics, and medical informatics – are evolving very quickly. The text should therefore be seen as an introduction, offering both elementary tutorials as well as more advanced applications and case studies.

The first part introduces the methodology of statistical inference and probabilistic modelling. Chapter 1 compares the two principle paradigms of statistical inference: the frequentist versus the Bayesian approach. Chapter 2 provides a brief introduction to learning Bayesian networks from data. Chapter 3 interprets the methodology of feed-forward neural networks in a probabilistic framework.

The second part describes how probabilistic modelling is applied to bioinformatics. Chapter 4 provides a self-contained introduction to molecular phylogenetic analysis, based on DNA sequence alignments, and it discusses the advantages of a probabilistic approach over earlier algorithmic methods. Chapter 5 describes how the probabilistic phylogenetic methods of Chapter 4 can be applied to detect interspecific recombination between bacteria and viruses from DNA sequence alignments. Chapter 6 generalizes and extends the standard phylogenetic methods for DNA so as to apply them to RNA sequence alignments. Chapter 7 introduces the reader to microarrays and gene expression data and provides an overview of standard statistical pre-processing procedures for image processing and data normalization. Chapters 8 and 9 address the challenging task of reverse-engineering genetic networks from microarray gene expression data using dynamical Bayesian networks and state-space models.

The third part provides examples of how probabilistic models are applied in medical informatics.

Chapter 10 illustrates the wide range of techniques that can be used to develop probabilistic models for medical informatics, which include logistic regression, neural networks, Bayesian networks, and class-probability trees.

The examples are supported with relevant theory, and the chapter emphasizes the Bayesian approach to probabilistic modeling.

Chapter 11 discusses Bayesian models of groups of individuals who may have taken several drug doses at various times throughout the course of a clinical trial. The Bayesian approach helps the derivation of predictive distributions that contribute to the optimization of treatments for different target populations.

Variable selection is a common problem in regression, including neural-network development. Chapter 12 demonstrates how Automatic Relevance Determination, a Bayesian technique, successfully dealt with this problem for the diagnosis of heart arrhythmia and the prognosis of lupus.

The development of a classifier is usually preceded by some form of data preprocessing. In the Bayesian framework, the preprocessing stage and the classifier-development stage are handled separately; however, Chapter 13 introduces an approach that combines the two in a Bayesian setting. The approach is applied to the classification of electroencephalogram data.

There is growing interest in the application of the variational method to model development, and Chapter 14 discusses the application of this emerging technique to the development of hidden Markov models for biosignal analysis.

Chapter 15 describes the Treat decision-support system for the selection of appropriate antibiotic therapy, a common problem in clinical microbiology. Bayesian networks proved to be particularly effective at modelling this problem task.

The medical-informatics part of the book ends with Chapter 16, a description of several software packages for model development. The chapter includes example codes to illustrate how some of these packages can be used.

Finally, an appendix explains the conventions and notation used throughout the book.

Intended Audience

The book has been written for researchers and students in statistics, machine learning, and the biological sciences. While the chapters in Parts II and III describe applications at the level of current cutting-edge research, the chapters in Part I provide a more general introduction to the methodology for the benefit of students and researchers from the biological sciences.

Chapters 1, 2, 4, 5, and 8 are based on a series of lectures given at the Statistics Department of Dortmund University (Germany) between 2001 and 2003, at Indiana University School of Medicine (USA) in July 2002, and at the “International School on Computational Biology”, in Le Havre (France) in October 2002.

Website

The website

<http://robots.ox.ac.uk/~parg/pmbmi.html>

complements this book. The site contains links to relevant software, data, discussion groups, and other useful sites. It also contains colored versions of some of the figures within this book.

Acknowledgments

This book was put together with the generous support of many people.

Stephen Roberts would like to thank Peter Sykacek, Iead Rezek and Richard Everson for their help towards this book. Particular thanks, with much love, go to Clare Waterstone.

Richard Dybowski expresses his thanks to his parents, Victoria and Henry, for their unfailing support of his endeavors, and to Wray Buntine, Paulo Lisboa, Ian Nabney, and Peter Weller for critical feedback on Chapters 3, 10, and 16.

Dirk Husmeier is most grateful to David Allcroft, Lynn Broadfoot, Thorsten Forster, Vivek Gowri-Shankar, Isabelle Grimmenstein, Marco Grzegorzczak, Anja von Heydebreck, Florian Markowetz, Jochen Maydt, Magnus Rattray, Jill Sales, Philip Smith, Wolfgang Urfer, and Joanna Wood for critical feedback on and proofreading of Chapters 1, 2, 4, 5, and 8. He would also like to express his gratitude to his parents, Gerhild and Dieter; if it had not been for their support in earlier years, this book would never have been written. His special thanks, with love, go to Ulli for her support and tolerance of the extra workload involved with the preparation of this book.

Edinburgh, London, Oxford
UK
July 2003

Dirk Husmeier
Richard Dybowski
Stephen Roberts

Probabilistic Modeling in Bioinformatics and Medical
Informatics

Husmeier, D.; Dybowski, R.; Roberts, S. (Eds.)

2005, XX, 508 p., Hardcover

ISBN: 978-1-85233-778-0