Preface

With recent advances in biotechnology spurred by the Human Genome Project, tremendous amounts of sequence, gene, protein, and pathway data have been accumulating at an exponential rate. Ontologies are emerging as an increasingly critical framework for coping with the onslaught of information encountered in genomics, transcriptomics, and proteomics. This onslaught involves not only an increase in sheer volume but also increases in both complexity and diversity. An ontology is a precise formulation of the concepts that form the basis for communication within a specific field. Because of this it is expected that the use of ontology and ontology languages will rise substantially in the postgenomic era. This book introduces the basic concepts and applications of ontologies and ontology languages in bioinformatics.

Distilling biological knowledge is primarily focused on unveiling the fundamental hidden structure as well as the grammatical and semantic rules behind the inherently related genomic, transcriptomic, and proteomic data within the boundary of a biological organism. Sharing vocabulary constitutes only the first step toward information retrieval and knowledge discovery. Once data have been represented in terms of an ontology, it is often necessary to transform the data into other representations which can serve very different purposes. Such transformations are crucial for conducting logical and critical analyses of existing facts and models, as well as deriving biologically sensible and testable hypotheses. This is especially important for bioinformatics because of the high degree of heterogeneity of both the format and the data models of the myriads of existing genomic and transcriptomic databases. This book presents not only how ontologies can be constructed but also how they can be used in reasoning, querying, and combining information. This includes transforming data to serve diverse purposes as well as combining information from diverse sources.

Our purpose in writing this book is to provide an introductory, yet in-depth analysis of ontologies and ontology languages to bioinformaticists, computer scientists, and other biomedical researchers who have intensive interests in exploring the meaning of the gigantic amounts of data generated by high-throughput technologies. Thus, this book serves as a guidebook for how one could approach questions like ontology development, inference, and reasoning in bioinformatics using contemporary information technologies and tools.

One of the most common ways that people cope with complexity is to classify into categories and then organize the categories hierarchically. This is a powerful technique, and modern ontologies make considerable use of it. Accordingly, classification into hierarchies is the starting point of the book.

The main division of the book is in three parts. We think of the parts as answering three questions: What ontologies are, How ontologies are used, and What ontologies could be. The actual titles are less colorful, but more informative. Since the audience of the book consists of scientists, the last part focuses on how ontologies could be used to represent techniques for reasoning with uncertainty.

The first part introduces the notion of an ontology, starting from hierarchically organized ontologies to the more general network organizations. It ends with a survey of the best-known ontologies in biology and medicine.

The second part shows how to use and construct ontologies. Ontologies have many uses. One might build an ontology just to have a better understanding of the concepts in a field. However, most uses are related in some way to the problem of coping with the large amount of information being generated by modern bioinformatics technologies. Such uses can be classified into three main categories: querying, viewing, and transforming. The first of these can be done using either imprecise natural language queries or precise queries using a formal query language. The second is actually a special case of the third, and this is explained in the first chapter in the subpart devoted to transformations. The other two chapters on transformations show two different approaches to transformations. The last part covers how to create an ontology.

The first two parts of the book consider only one style of reasoning: deductive or Boolean logic. The third part of the book considers the process of thinking in which a conclusion is made based on observation, also known as inductive reasoning. The goal of this part is to achieve a synthesis that supports both inductive and deductive reasoning. It begins by contrasting inductive and deductive reasoning. Then it covers Bayesian networks, a

popular formalism that shows great promise as a means of expressing uncertainty. One important activity of science is the process of combining multiple independent observations of phenomena. The third chapter in this part gives a brief introduction to this very large subject. The final chapter of the part and the book is the most speculative. It proposes that the World Wide Web can be extended to support reasoning with uncertainty, as expressed using Bayesian networks. The result is an inductive reasoning web which we have named the Bayesian Web.

The authors would like to thank the many friends and colleagues who contributed their time and expertise. We especially appreciate John Bottoms who read the manuscript more than once and contributed many insightful suggestions. We wish to thank JoAnn Manson, Simin Liu, and the Division of Preventive Medicine, Brigham and Women's Hospital, for their help and encouragement. We also appreciate the contributions by our many colleagues at Northeastern University, Versatile Information Systems, and Composable Logic, including Mitch Kokar, Jerzy Letkowski, and Jeff Smith. We thank Xiaobin Wang at Children's Memorial Hospital in Chicago for sharing with us the microarray data on preterm delivery. KB would like to acknowledge his debt to his mentors, the late Gian-Carlo Rota and Mark Kac. Robert Prior and Katherine Almeida deserve special praise for their patience in what turned out to be a rather larger project than we originally anticipated. Finally, we wish to thank our families for their love, support and encouragement to complete this work.

Throughout the book there are many references to web resources. These references are Uniform Resource Identifiers (URIs). A Uniform Resource Locator (URL) is a special case of a URI that specifies the location of a web resource. A URL is used by a web browser to find and download a resource, such as a webpage. A URI is a unique identifier of a web resource and need not correspond to a downloadable resource, although they often do. Some web resources have a URL that is not the same as its URI. This is becoming an increasingly common practice for ontologies and schemas. The "typewriter" font was used in this book for URIs. Most URLs begin with `http://`. This initial part of the URL specifies the protocol for obtaining the resource. When the protocol is omitted, one obtains the Uniform Resource Name (URN). Most web browsers are capable of finding a resource even when the protocol has not been specified. In this book we will usually use the URN rather than the URL to save space. For typographical purposes, some URIs (and other constructs) in this book have been split so as to fit in the space available.

The URI for this book is `ontobio.org`, and this URI is also the URL of the website for the book. Because URIs are constantly changing, the website for the book has updated information about the URIs that appear in the book as well as new ones that may be of interest to readers. The book website also has additional exercises and solutions to them.