

PREFACE

At the crossroads between statistics and machine learning, probabilistic graphical models provide a powerful formal framework to model complex data. Examples of probabilistic graphical models are Bayesian networks and Markov random fields, which represent two of the most popular classes of such models. With the rapid advancements of high-throughput technologies and the ever decreasing costs of these next-generation technologies, a fast-growing volume of biological data of various types—the so-called omics—is in need of accurate and efficient modeling methods, prior to further downstream analysis. As probabilistic graphical models are able to deal with high-dimensional data and non-linear dependences, it is foreseeable that such models will have a prominent role to play in advances in genome-wide analyses.

Currently, few people are specialists in the design of cutting-edge methods using probabilistic graphical models for genetics, genomics, and postgenomics. This seriously hinders the diffusion of such methods. The prime aim of this book is therefore to bring the concepts underlying these advanced models within understanding of a broader audience of scientists, engineers, and graduate students.

If they are not specialists of probabilistic graphical models, bioinformaticians, statisticians, biostatisticians, and experts in statistical genetics with an intuition that their solution to a problem should involve such models are compelled to glean incomplete information from publications. We are not even talking of surveys whose consultation will never allow launching out into the design of advanced methods. Some academic courses may well be delivered here and there, that dwell on cutting-edge approaches using probabilistic graphical models for the targeted topics; neither are such courses widely available for the potentially interested audience, nor do they cover a sufficiently illustrative set of models and applications.

The target readers of this book include researchers and engineers as well as graduate students starting a master's or a PhD thesis. Besides, if there is one area where transdisciplinarity is the daily lot, it is the advanced analysis of genome-wide data. Constructive cooperation with a domain specialist requires the ability to hold a productive dialogue, which therefore demands a deep understanding of the models as well as a solid background regarding these models. Often, scientists from different fields such as genetics, statistics, or computer science do not use the same scientific language, and this might lead to confusion and misunderstanding. Bridging the gap between different scientific worlds thus helps scientists to better communicate, and from a

higher perspective, contributes to the emergence of new fields of research. Currently, the only solution for such people to gain a deep understanding is finding spare time to gather information to learn from it. The book intends to spare such readers this task.

Hopefully, this book will be of equal interest, if still not higher, for the graduate students supervised by members of the aforementioned audience. Depending on their academic institution, students taught computational methods for genetics, genomics, or postgenomics rarely have access to a course presenting the advanced use of probabilistic graphical models in such fields. One reason for this lies in the fact that these models and their potentialities have only rather recently created renewed interest in genetics in the broad sense. Another reason might be the lack of experts possessing this two-fold skill in these students' institutions. Besides, a few hours taught on the subject are not sufficient to provide both enough material and hindsight on the topic. This book attempts to fill this gap.

This book is also designed to help experts in machine learning grasp the interest in designing advanced methods based on probabilistic graphical models in transdisciplinary collaborations.

This book arises out of a six-year collaboration between its scientific editors. Our various interests in computer science, machine learning, applied mathematics, Bayesian statistics, applications in genetics, genomics, and postgenomics have found in probabilistic graphical models a breeding ground for both our own investigations and the preparation and direction of this book. Besides, coming from different backgrounds, we found a common ground in demanding the highest self-containedness in the contributions of the invited authors. In addition to the intrinsic richness of these contributions, our guiding thread was then providing added value through accessibility for non-specialists of probabilistic graphical models, with no concession on the informativeness of the book's contents.

We have been fortunate to obtain the widest consent regarding invited authors' participation in our project. We subsequently enjoyed a fruitful period of dense exchanges with these authors, who accepted this extra workload.

The book is divided into a general introduction, a tutorial on probabilistic graphical networks, and six main sections devoted to specific application fields in genetics (in the broad sense). The introductory chapter aims at providing a minimal background for readers that are not familiar with biology or need information about the high-throughput biological data addressed by the models described in the book. Moreover, such terms and expressions as genetics, genomics, postgenomics, systems biology, and integrative biology are clarified. Indeed, a leitmotif of the book is the integration of heterogeneous sources of omics data, to boost downstream biological applications. Finally, this introduction provides the motivation for using probabilistic graphical models to handle high-throughput biological data and provides a brief evocation of the use of probabilistic graphical networks in the six applications highlighted by the book: gene network inference, causality discovery, association genetics, epigenetics, detection of copy number variations, and prediction of outcomes from high-dimensional genomic data.

The essentials for understanding probabilistic graphical models are offered in a tutorial at the beginning of the book. This tutorial was carefully designed to be accessible to the largest audience. Since the concepts and techniques presented in this tutorial may require broader and non-trivial knowledge, accessibility and self-containedness were again the targeted objectives.

Together with a thorough review chapter focusing on selected domains in genetics, fourteen chapters illustrate the design of advanced approaches, for the six abovementioned applications. This book offers a lot of new insights that could only be gleaned from the literature available through excruciating labor. The chapters are self-contained, and they can be read independently of each other.

C. S. and R. M.