BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 5. 376–377
Advance Access publication July 18, 2007
doi:10.1093/bib/bbm022

Book Review

Bayesian Inference for Gene Expression and Proteomics.

Edited by Kim-Anh Do, Peter Müller and Marina Vannucci
Cambridge University Press,
Cambridge; 2006;
ISBN: 978-0521860925;
Hardback; 456pp.

A text that has a systematic account of Bayesian analysis in computational biology has been needed for a long time. This book is a timely publication entirely devoted to cutting-edge Bayesian methods in genomics and proteomics research and many of its contributors are leading authorities in the field. It is thus an indispensable reference for researchers who are interested in applying Bayesian techniques in their own biological research. Moreover, the book calls for more methodological and theoretical research to surmount challenging issues that still remain.

The book starts with a brief review of three key high-throughput biotechnologies, microarray, serial analysis of gene expression (SAGE) and mass spectrometry experiments. It then continues to discuss their data structures as well as their statistical analysis strategy. The rest of the book is divided into three sections corresponding to these three high throughput technologies, each of which focuses on Bayesian analysis.

Chapters 2 through 11, the first section of the book, concentrate on Bayesian inference for microarray data. Among them, Chapter 4 focuses on low level analysis, including the estimation of gene expression index from image data. Other chapters in this section are primarily concerned with high level analysis, including identifying differentially expressed genes, clustering, marker identification as well as joint modeling of gene expression and disease status. The second and third sections of the book fill a critical gap in existing bioinformatics books by providing in depth coverage of rigorous analysis on mass spectrometry data and DNA motif identification. To that end, Chapters 12 through 15 focus on analysis of protein spectrometry and SAGE data, where mixture models and mixed models are the basic ingredients. The third section begins with Chapter 16, which gives a wonderful review of methods for the identification of transcription factor binding motifs. The elucidation of regulatory networks is the focus of Chapters 17 through 21. It is well known that Bayesian methods can effectively extract information, and this was nicely illustrated by the authors by showing the power of Bayesian methods to integrate multiple types of data, including gene expression data, protein–DNA binding data, and Gene Ontology annotations. Finally, a chapter on sample size calculation in microarray experiments concludes the book.

Even though the book was contributed by many authors, it is well structured and the reader can easily follow the presentation. I found the discussion in many chapters very useful. For example, Chapter 16 provides the current research trends in motif discovery and sheds a fresh light on future research in this area.

Since the book consists of many novel applications of Bayesian methods in genomics and proteomics, it may help guide a new wave of research in the field. On the other hand, the effectiveness of some methods needs further investigation. For example, estimation of the gene expression index in Chapter 3, has not been compared to similar methods, such as the Li–Wong model of the Wong group and robust multi-array average (RMA) method of the Speed group.

One cautionary note for practitioners who are eager to adapt these methods in their own research, is that many Bayesian models presented in this book are hierarchical Bayesian models. In a typical Bayesian paradigm, one needs to specify some prior (probability distribution) for each parameter. Then combining with data, one can get the posterior distribution through the model using Bayes theorem. Another layer of complexity of a hierarchical Bayesian model is that one need to specify

probability distributions (hierarchical prior) for priors of parameters, which could be very tricky. Moreover, the posterior distribution of the parameters obtained in hierarchical Bayesian model, in many circumstances, can not be expressed in a closed analytic form. Thus, statistical inference heavily relies on iterative sampling techniques, such as Markov chain Monte Carlo (MCMC) methods. However, it is well known that one can blindly design a MCMC method to sample the posterior even if the posterior does not exist. Therefore, a good practice is to check the existence of the posterior after model setup. A necessary condition of existence is that posterior distribution needs to approach to zero if parameters are approaching infinity.

One related aspect of Bayesian analysis is the issues concerning the convergence of the iterative samples, which do not get too much attention in this book. The authors assume the readers already have hands-on experience in Bayesian analysis. For a novice, a standard textbook on Bayesian methods is essential before attempting the methods presented in this book.

In conclusion, this book provides a wealth of material, deep insight, and up-to-date coverage of Bayesian methods in bioinformatics. I use it as a reference and parts of it will be used in graduate classes that I teach. I have no doubt that it will be an invaluable source for all researchers in the field for years to come.

Ping Ma
Department of Statistics and
Institute for Genomic Biology
University of Illinois at Urbana-Champaign, USA