

Preface

This book is intended for students in Statistics, Biostatistics, Computer Science and related fields in applied Mathematics who seek to gain a basic knowledge of some of the problems arising in the analysis of genetics and genomics. As such, it is assumed the reader is familiar with calculus and linear algebra and has been exposed to a course that covers the basics of probability and statistics assuming knowledge of calculus and linear algebra. The unifying theme of all the topics revolves around the questions: “What does this segment of DNA do?” and “Which segment of DNA does this?” The selection of topics is largely based on a course I took while a student at Columbia that was offered by Shaw-Hwa Lo and Herman Chernoff (although they obviously can’t be blamed for any of the shortcomings one may notice here). The material of that course was the basis for a course I started in conjunction with Wei Pan and Hegang Chen. Wei and I continued to teach this course and refined the subject matter as we progressed. To all of these people, I am deeply indebted for their help. The largest change that Wei, Hegang and I made to the material that Drs. Lo and Chernoff included in their class is the inclusion of a substantial amount of material on microarrays (which were just emerging when I took Lo and Chernoff’s course).

This book is not intended to be the authoritative treatment of the topics it covers, as such a book would be enormous and unsuitable as a text. Rather, the goal here is to introduce the student to a diverse set of problems and a number of approaches that have been used to address these problems. In the process, it is hoped that the reader will learn enough Biology, Statistics, Computer Science and Bioinformatics so that when confronted with a new problem at the interface of these areas, the reader will have some ideas regarding what will likely constitute a reasonable approach. This seems preferable to covering this material in an encyclopedic fashion since technology is constantly changing. Indeed, even over the course of time in which this book was drafted, technological developments have altered our view regarding optimal approaches to certain problems. For example, the role of linkage studies relative to association studies for disease gene mapping has changed substantially over the last several years (and as I write this, some are questioning the role of association analysis relative to sequencing entire genomes of carefully selected subjects). Nonetheless, topics were selected that will likely have value as technology changes. For example, classification and cluster analysis will still be important tools when we master quantification of proteins or metabolites. Other topics were selected because

they illustrate the use of methods that are widespread among researchers who analyze genomic data, such as hidden Markov models and the extreme value distribution.

In addition to those already mentioned, I would like to thank all the students who attended this course and provided valuable feedback (even if some of this feedback just consisted of an odd facial expression betraying confusion, likely due to my imperfect presentation). In addition John Connett (the Division head while I drafted this book) deserves thanks for understanding that writing a book is a huge undertaking that necessarily detracts from time I could be doing other things. Thanks also to Julie Frommes and my family. Finally thanks to Rob Calver and all the people at Chapman & Hall/CRC Press for their assistance and patience.

Cavan Reilly
Minneapolis, MN
February, 2009