

Preface

Bioinformatics has been one of the most important multidisciplinary subjects in the last century. Initially, the major task of bioinformatics research was to handle large genomic data for knowledge extraction and for making predictions. More recently, the practices of bioinformatics have extended from genomics to proteomics, metabolomics, and most importantly systems biology. In addition to most traditional bioinformatics exercises which focus on large database management and sequence homology alignment for molecular structure prediction and function annotation, modelling biological data using statistical/machine learning has been an important trend. This part of the exercise has gained great attention because it can help carry out efficient, effective, and accurate knowledge extraction and prediction model construction. However, the application of machine learning approaches in bioinformatics researches and practices has a series of challenges compared with other applications. The challenges include data size, data quality, and the imbalance between different data resources. These challenges are particularly obvious in systems biology research. For instance, genomics data size has a scale of around 25K, but proteomics data size can reach up to a scale of millions. Currently, it is hard to use modern computers to handle such large scale data in one machine learning model. Furthermore, due to experimental variation, tissue corruption, and equipment resolution, most metabolite data suffer a problem of data quality. This casts a challenge in machine learning model construction in terms of data noise and missing data. In using next generation sequencing equipment such as Illumina, we are faced with mega-byte of fragments of sequences. The challenge is how to assembly

these fragments accurately without any reference sequences. An urgent requirement in systems biology proposes to use different sources of data for analysing systems behaviour. This then casts a challenge about how to efficiently incorporate these data with different resolutions, with different data format, with different data quality, and with different data dimensionalities in one machine learning model. This book therefore tries to discuss some of these challenges.

This book is written based on my teaching and research notes in bioinformatics in the past ten years. I thank Prof Jason Wang and the publisher for inviting me to write this book. The book is written mainly for postgraduates and researchers at the start of their bioinformatics research and practice. The pre-requisite to using this book is some basic linear algebra and statistics knowledge. The book can be used for both advanced undergraduate and postgraduate teaching reference. Readers are encouraged to be familiar with basic R programming before using this book as most case studies presented in the book are implemented in R.

The book is composed of three parts. The first part covers several unsupervised learning approaches which can be used in bioinformatics. For instance, multidimensional scaling is commonly used in bioinformatics for biological data visualisation. Various cluster analysis approaches as well as self-organising map have been used for biological pattern recognition. After data partitioning, molecules can then be clustered leading to prototype pattern discovery and new hypothesis generation.

The second part mainly discusses supervised learning approaches. In many bioinformatics projects, a typical question is how to accurately predict unknowns based on experimental data. For instance, how can we identify the most important genes for most efficient and accurate disease diagnosis? Additionally, given a huge number of molecular sequence data in which most functions are still unknown, how can we make prediction models based on limited information of known functions in sequence data? This part therefore introduces several commonly used supervised learning algorithms as well as their applications to bioinformatics.

The third part of this book introduces the concepts relevant to computational systems biology which is now the most important research targets in bioinformatics. Computational systems biology research mainly focuses on large biological systems aiming to reveal the complex interplay between molecules and molecular entities. Gene network, systems dynamics and pathway recognition have been of much interest in recent years. The third part then demonstrates how machine learning algorithms can be used for these issues.

As mentioned above, this book is based on the revision of my teaching and research notes. It is therefore important to name several research collaborators. My key research collaborators include T Charlie Hodgman, Andrew Dalby, Murray Grant, Richard Titball, Nick Smirnoff, and Tom Richards. The students who have contributed to the improvement of my teaching of bioinformatics in University of Exeter are Rebecca Hamer, Jon Dry, Emily Berry, Dave Trudgun, Hanieh Yaghootkar and Susie Clark. I am very grateful to Susie Clark for proof-reading the book.

Finally, I would like to thank my parents, wife and daughter for their great support. During the writing of this book, I regret not being able to spend more time with them. I hope the publication of this book will make up for the sacrifice.

Zheng Rong Yang

29 November 2009

Exeter, England, UK

1 Introduction to Unsupervised Learning

2 Probabilistic Density Estimation Approaches

2.1 Histogram approach

2.2 Parametric approach

2.3 Non-parametric approach

2.3.1 K-nearest neighbour approach

2.3.2 Kernel approach

Summary

3 Dimensional Reduction

3.1 General