

Contents

Preface	xi
1 Basic Molecular Biology for Statistical Genetics and Genomics	1
1.1 Mendelian genetics	1
1.2 Cell biology	2
1.3 Genes and chromosomes	3
1.4 DNA	5
1.5 RNA	6
1.6 Proteins	7
1.6.1 Protein pathways and interactions	9
1.7 Some basic laboratory techniques	11
1.8 Bibliographic notes and further reading	13
1.9 Exercises	13
2 Basics of Likelihood Based Statistics	15
2.1 Conditional probability and Bayes theorem	15
2.2 Likelihood based inference	16
2.2.1 The Poisson process as a model for chromosomal breaks	17
2.2.2 Markov chains	18
2.2.3 Poisson process continued	19
2.3 Maximum likelihood estimates	21
2.3.1 The EM algorithm	26
2.4 Likelihood ratio tests	28
2.4.1 Maximized likelihood ratio tests	28

2.5	Empirical Bayes analysis	29
2.6	Markov chain Monte Carlo sampling	30
2.7	Bibliographic notes and further reading	33
2.8	Exercises	33
3	Markers and Physical Mapping	37
3.1	Introduction	37
3.2	Types of markers	39
3.2.1	Restriction fragment length polymorphisms (RFLPs)	40
3.2.2	Simple sequence length polymorphisms (SSLPs)	40
3.2.3	Single nucleotide polymorphisms (SNPs)	40
3.3	Physical mapping of genomes	41
3.3.1	Restriction mapping	41
3.3.2	Fluorescent in situ hybridization (FISH) mapping	45
3.3.3	Sequence tagged site (STS) mapping	46
3.4	Radiation hybrid mapping	46
3.4.1	Experimental technique	46
3.4.2	Data from a radiation hybrid panel	46
3.4.3	Minimum number of obligate breaks	47
	Consistency of the order	47
3.4.4	Maximum likelihood and Bayesian methods	48
3.5	Exercises	50
4	Basic Linkage Analysis	53
4.1	Production of gametes and data for genetic mapping	53
4.2	Some ideas from population genetics	54
4.3	The idea of linkage analysis	55
4.4	Quality of genetic markers	61
4.4.1	Heterozygosity	61
4.4.2	Polymorphism information content	62
4.5	Two point parametric linkage analysis	62

CONTENTS	v
4.5.1 LOD scores	63
4.5.2 A Bayesian approach to linkage analysis	63
4.6 Multipoint parametric linkage analysis	64
4.6.1 Quantifying linkage	65
4.6.2 An example of multipoint computations	66
4.7 Computation of pedigree likelihoods	67
4.7.1 The Elston Stewart algorithm	68
4.7.2 The Lander Green algorithm	68
4.7.3 MCMC based approaches	69
4.7.4 Sparse binary tree based approaches	70
4.8 Exercises	70
5 Extensions of the Basic Model for Parametric Linkage	73
5.1 Introduction	73
5.2 Penetrance	74
5.3 Phenocopies	75
5.4 Heterogeneity in the recombination fraction	75
5.4.1 Heterogeneity tests	76
5.5 Relating genetic maps to physical maps	77
5.6 Multilocus models	80
5.7 Exercises	81
6 Nonparametric Linkage and Association Analysis	83
6.1 Introduction	83
6.2 Sib-pair method	83
6.3 Identity by descent	84
6.4 Affected sib-pair (ASP) methods	84
6.4.1 Tests for linkage with ASPs	85
6.5 QTL mapping in human populations	86
6.5.1 Haseman Elston regression	87
6.5.2 Variance components models	88

Coancestry	89
6.5.3 Estimating IBD sharing in a chromosomal region	90
6.6 A case study: dealing with heterogeneity in QTL mapping	92
6.7 Linkage disequilibrium	98
6.8 Association analysis	100
6.8.1 Use of family based controls	100
Haplotype relative risk	101
Haplotype-based haplotype relative risk	102
The transmission disequilibrium test	103
6.8.2 Correcting for stratification using unrelated individuals	104
6.8.3 The HAPMAP project	106
6.9 Exercises	106
7 Sequence Alignment	109
7.1 Sequence alignment	109
7.2 Dot plots	110
7.3 Finding the most likely alignment	111
7.4 Dynamic programming	114
7.5 Using dynamic programming to find the alignment	115
7.5.1 Some variations	119
7.6 Global versus local alignments	119
7.7 Exercises	120
8 Significance of Alignments and Alignment in Practice	123
8.1 Statistical significance of sequence similarity	123
8.2 Distributions of maxima of sets of iid random variables	124
8.2.1 Application to sequence alignment	127
8.3 Rapid methods of sequence alignment	128
8.3.1 FASTA	130
8.3.2 BLAST	130
8.4 Internet resources for computational biology	132
8.5 Exercises	133

9 Hidden Markov Models	135
9.1 Statistical inference for discrete parameter finite state space Markov chains	135
9.2 Hidden Markov models	136
9.2.1 A simple binomial example	136
9.3 Estimation for hidden Markov models	137
9.3.1 The forward recursion	137
The forward recursion for the binomial example	138
9.3.2 The backward recursion	138
The backward recursion for the binomial example	139
9.3.3 The posterior mode of the state sequence	140
9.4 Parameter estimation	141
Parameter estimation for the binomial example	142
9.5 Integration over the model parameters	143
9.5.1 Simulating from the posterior of ϕ	145
9.5.2 Using the Gibbs sampler to obtain simulations from the joint posterior	145
9.6 Exercises	146
10 Feature Recognition in Biopolymers	147
10.1 Gene transcription	149
10.2 Detection of transcription factor binding sites	150
10.2.1 Consensus sequence methods	150
10.2.2 Position specific scoring matrices	151
10.2.3 Hidden Markov models for feature recognition	153
A hidden Markov model for intervals of the genome	153
A HMM for base-pair searches	154
10.3 Computational gene recognition	154
10.3.1 Use of weight matrices	156
10.3.2 Classification based approaches	156
10.3.3 Hidden Markov model based approaches	157
10.3.4 Feature recognition via database sequence comparison	159
10.3.5 The use of orthologous sequences	159
10.4 Exercises	160

11 Multiple Alignment and Sequence Feature Discovery	161
11.1 Introduction	161
11.2 Dynamic programming	162
11.3 Progressive alignment methods	163
11.4 Hidden Markov models	165
11.4.1 Extensions	167
11.5 Block motif methods	168
11.5.1 Extensions	172
11.5.2 The propagation model	173
11.6 Enumeration based methods	174
11.7 A case study: detection of conserved elements in mRNA	175
11.8 Exercises	177
12 Statistical Genomics	179
12.1 Functional genomics	179
12.2 The technology	180
12.3 Spotted cDNA arrays	181
12.4 Oligonucleotide arrays	181
12.4.1 The MAS 5.0 algorithm for signal value computation	182
12.4.2 Model based expression index	184
12.4.3 Robust multi-array average	185
12.5 Normalization	187
12.5.1 Global (or linear) normalization	188
12.5.2 Spatially varying normalization	189
12.5.3 Loess normalization	189
12.5.4 Quantile normalization	190
12.5.5 Invariant set normalization	190
12.6 Exercises	190

CONTENTS	ix
13 Detecting Differential Expression	193
13.1 Introduction	193
13.2 Multiple testing and the false discovery rate	194
13.3 Significance analysis for microarrays	199
13.3.1 Gene level summaries	199
13.3.2 Nonparametric inference	200
13.3.3 The role of the data reduction	202
13.3.4 Local false discovery rate	203
13.4 Model based empirical Bayes approach	203
13.5 A case study: normalization and differential detection	207
13.6 Exercises	211
14 Cluster Analysis in Genomics	213
14.1 Introduction	213
14.1.1 Dissimilarity measures	215
14.1.2 Data standardization	215
14.1.3 Filtering genes	215
14.2 Some approaches to cluster analysis	216
14.2.1 Hierarchical cluster analysis	216
14.2.2 K -means cluster analysis and variants	219
14.2.3 Model based clustering	220
14.3 Determining the number of clusters	223
14.4 Biclustering	226
14.5 Exercises	228
15 Classification in Genomics	231
15.1 Introduction	231
15.2 Cross-validation	233
15.3 Methods for classification	234
15.3.1 Discriminate analysis	234
15.3.2 Regression based approaches	237

15.3.3 Regression trees	238
15.3.4 Weighted voting	239
15.3.5 Nearest neighbor classifiers	240
15.3.6 Support vector machines	240
15.4 Aggregating classifiers	244
15.4.1 Bagging	244
15.4.2 Boosting	245
15.4.3 Random forests	246
15.5 Evaluating performance of a classifier	246
15.6 Exercises	247
References	249
Index	261