

Contents

Preface	xi
About the Author	xiii
I Background	1
1 Probabilistic Informatics	3
1.1 What Is Informatics?	4
1.2 Bioinformatics	6
1.3 Probabilistic Informatics	7
1.4 Outline of This Book	8
2 Probability Basics	11
2.1 Probability Basics	11
2.1.1 Probability Spaces	12
2.1.2 Conditional Probability and Independence	14
2.1.3 Bayes' Theorem	17
2.2 Random Variables	18
2.2.1 Probability Distributions of Random Variables	18
2.2.2 Independence of Random Variables	23
2.3 The Meaning of Probability	27
2.3.1 Relative Frequency Approach to Probability	27
2.3.2 Subjective Approach to Probability	30
2.4 Random Variables in Applications	32
3 Statistics Basics	41
3.1 Basic Concepts	42
3.1.1 Expected Value	42
3.1.2 Variance and Covariance	43
3.1.3 Maximum Likelihood Estimation	49
3.2 Markov Chain Monte Carlo	51
3.2.1 Markov Chains	51
3.2.2 MCMC	54
3.3 The Normal Distribution	59

4	Genetics Basics	63
4.1	Organisms and Cells	63
4.1.1	Basic Definitions	63
4.1.2	Cell Division in Eukaryotes	65
4.1.3	Reproduction	67
4.2	Genes	68
4.2.1	High-Level Gene Behavior	68
4.2.2	Low-Level Gene Behavior	71
4.3	Mutations	74
4.3.1	Substitution Mutations	75
4.3.2	Insertions and Deletions	75
4.3.3	Mutation Rates	78
II	Bayesian Networks	83
5	Foundations of Bayesian Networks	85
5.1	What Is a Bayesian Network?	86
5.2	Properties of Bayesian Networks	88
5.2.1	Definition of a Bayesian Network	89
5.2.2	Representation of a Bayesian Network	91
5.3	Causal Networks as Bayesian Networks	95
5.3.1	Causality	95
5.3.2	Causality and the Markov Condition	100
5.3.3	The Markov Condition without Causality	104
5.4	Inference in Bayesian Networks	104
5.4.1	Examples of Inference	106
5.4.2	Inference Algorithms and Packages	108
5.4.3	Inference Using Netica	110
5.5	Networks with Continuous Variables	112
5.5.1	Gaussian Bayesian Networks	112
5.5.2	Hybrid Networks	115
5.6	How Do We Obtain the Probabilities?	117
5.6.1	The Noisy OR-Gate Model	117
5.6.2	Methods for Discretizing Continuous Variables	124
6	Further Properties of Bayesian Networks	135
6.1	Entailed Conditional Independencies	136
6.1.1	Examples of Entailed Conditional Independencies	136
6.1.2	d-Separation	139
6.2	Faithfulness	143
6.2.1	Unfaithful Probability Distributions	144
6.2.2	The Faithfulness Condition	146
6.3	Markov Equivalence	147
6.4	Markov Blankets and Boundaries	150

7	Learning Bayesian Network Parameters	157
7.1	Learning a Single Parameter	158
7.1.1	Binomial Random Variables	158
7.1.2	Multinomial Random Variables	162
7.1.3	Guidelines for Articulating Prior Belief	163
7.2	Learning Parameters in a Bayesian Network	165
7.2.1	Procedure for Learning Parameters	165
7.2.2	Equivalent Sample Size	167
8	Learning Bayesian Network Structure	177
8.1	Model Selection	178
8.2	Score-Based Structure Learning	179
8.2.1	The Bayesian Score	179
8.2.2	The BIC Score	188
8.2.3	Consistent Scoring Criteria	190
8.2.4	How Many DAGs Must We Score?	191
8.2.5	Using the Learned Network to Do Inference	191
8.3	Constraint-Based Structure Learning	192
8.3.1	Learning a DAG Faithful to P	192
8.3.2	Learning a DAG in Which P Is Embedded Faithfully	198
8.4	Causal Learning	199
8.4.1	Causal Faithfulness Assumption	199
8.4.2	Causal Embedded Faithfulness Assumption	203
8.5	Model Averaging	206
8.6	Approximate Structure Learning	209
8.6.1	Approximate Model Selection	209
8.6.2	Approximate Model Averaging	214
8.7	Software Packages for Learning	217

III Bioinformatics Applications 223

9	Nonmolecular Evolutionary Genetics	225
9.1	No Mutations, Selection, or Genetic Drift	226
9.1.1	Haploid Population	226
9.1.2	Diploid Populations	227
9.2	Natural Selection	230
9.2.1	Natural Selection in a Haploid Population	230
9.2.2	Natural Selection in a Diploid Population	232
9.3	Genetic Drift	248
9.3.1	A Coin-Tossing Experiment	250
9.3.2	Application to the Wright-Fisher Model	255
9.3.3	Difficulties with the Wright-Fisher Model	256
9.4	Natural Selection and Genetic Drift	257
9.5	Rate of Substitution	259

10 Molecular Evolutionary Genetics	263
10.1 Models of Nucleotide Substitution	264
10.1.1 The One-Parameter Jukes-Cantor Model	264
10.1.2 Equilibrium Frequencies	269
10.1.3 Other Models	270
10.1.4 Time Reversibility	271
10.2 Evolutionary Distance	272
10.2.1 Assuming Uniform Rates	273
10.2.2 Assuming Nonuniform Rates	278
10.3 Sequence Alignment	283
11 Molecular Phylogenetics	295
11.1 Phylogenetic Trees	296
11.1.1 Rooted Trees	297
11.1.2 Unrooted Trees	297
11.2 Distance Matrix Learning Methods	299
11.2.1 UPGMA	300
11.2.2 Neighbors-Relation Method	302
11.3 Maximum Likelihood Method	304
11.3.1 A Phylogenetic Tree as a Bayesian Network	304
11.3.2 Scoring a Tree Assuming Uniform Rates	306
11.3.3 Scoring a Tree Assuming Nonuniform Rates	317
11.3.4 Scoring a Tree Assuming Context-Dependent Rates	318
11.3.5 Scoring a Tree Using Other Substitution Models	318
11.3.6 Searching over the Space of Trees	319
11.3.7 Discussion	326
11.4 Distance Matrix Methods Using ML	326
11.4.1 Rates Not varying Among Sites	327
11.4.2 Rates Varying among Sites	329
12 Analyzing Gene Expression Data	333
12.1 DNA Microarrays	334
12.2 A Bootstrap Approach	335
12.2.1 Analysis of Path Features	337
12.2.2 Analysis of Markov Features	337
12.3 Model Averaging Approaches	339
12.3.1 The Probability of a Feature Given an Ordering	340
12.3.2 MCMC over Orderings	344
12.3.3 Experimental Results	345
12.3.4 Using a Different Prior Probability Distribution	346
12.4 Module Network Approach	348
12.4.1 Module Networks	348
12.4.2 The Bayesian Score for Module Networks	352
12.4.3 A Search Algorithm for Module Networks	354
12.4.4 Using Class Probability Trees in Module Networks	356
12.4.5 Experimental Results	361

13 Genetic Linkage Analysis	367
13.1 Introduction to Genetic Linkage Analysis	368
13.1.1 Analyzing the Fruit Fly <i>Drosophila</i>	368
13.1.2 Creating a Genetic Map	372
13.2 Genetic Linkage Analysis in Humans	374
13.2.1 Polymorphic DNA Sequences as Genetic Markers	375
13.2.2 Pedigrees	377
13.2.3 Examples of Pedigrees	377
13.2.4 Pedigrees and Genetic Linkage Analysis	379
13.3 A Bayesian Network Model	380

Bibliography	387
---------------------	------------

Index	401
--------------	------------

This is not a book on biology, but rather one on the application of probability and statistics, in particular Bayesian networks, to genetics, as there is nothing to discern chemical structure of behavior. For example, a DNA sequence is simply treated as a string composed of the characters A, G, C, and T. I do, however, present some material on high-level features of biological processes so as to provide a context for the applications.

As in most books I write, I have taken the following measures to make the content accessible. First, initially I do not introduce a concept or prove a theorem concerning the given material. Rather, I discuss a specific, concrete case and then mention that a result can be generalized. Second, I provide simple figures and examples. A picture may be worth a 1000 words, but an example is worth 10,000 words.

The book has three parts. Part I provides background material on probability, statistics, and genetics which the student or reader may already know. Many readers should be able to skip this part. Part II discusses Bayesian networks, inference in Bayesian networks, and learning Bayesian networks from data. Part III covers applications to bioinformatics. Specifically, Chapter 9 concerns non-inferential evolutionary genetics, Chapter 10 presents molecular evolutionary genetics, Chapter 11 shows an application of Bayesian networks to molecular phylogenetics, Chapter 12 introduces analyzing gene expression data, and Chapter 13 concerns gene linkage analysis.

For the purpose of classification, it seemed best to group the material into these three parts. However, I do not cover the material in this order in the