

FORMATICS

LEARNING APPROACH

ION

AND SØREN BRUNAK

FORMATICS

LEARNING APPROACH

ION

AND SØREN BRUNAK

FORMATICS

LEARNING APPROACH

ION

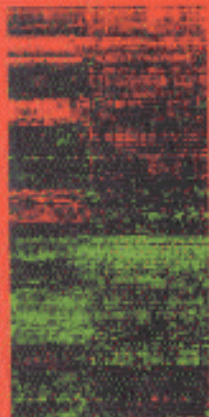
AND SØREN BRUNAK

FORMATICS

LEARNING APPROACH

ION

AND SØREN BRUNAK



Contents

Series Foreword	ix
Preface	xi
1 Introduction	1
1.1 Biological Data in Digital Symbol Sequences	1
1.2 Genomes—Diversity, Size, and Structure	7
1.3 Proteins and Proteomes	16
1.4 On the Information Content of Biological Sequences	24
1.5 Prediction of Molecular Function and Structure	43
2 Machine-Learning Foundations: The Probabilistic Framework	47
2.1 Introduction: Bayesian Modeling	47
2.2 The Cox Jaynes Axioms	50
2.3 Bayesian Inference and Induction	53
2.4 Model Structures: Graphical Models and Other Tricks	60
2.5 Summary	64
3 Probabilistic Modeling and Inference: Examples	67
3.1 The Simplest Sequence Models	67
3.2 Statistical Mechanics	73
4 Machine Learning Algorithms	81
4.1 Introduction	81
4.2 Dynamic Programming	82
4.3 Gradient Descent	83
4.4 EM/GEM Algorithms	84
4.5 Markov-Chain Monte-Carlo Methods	87
4.6 Simulated Annealing	91
4.7 Evolutionary and Genetic Algorithms	93
4.8 Learning Algorithms: Miscellaneous Aspects	94

5 Neural Networks: The Theory	99
5.1 Introduction	99
5.2 Universal Approximation Properties	104
5.3 Priors and Likelihoods	106
5.4 Learning Algorithms: Backpropagation	111
6 Neural Networks: Applications	113
6.1 Sequence Encoding and Output Interpretation	114
6.2 Sequence Correlations and Neural Networks	119
6.3 Prediction of Protein Secondary Structure	120
6.4 Prediction of Signal Peptides and Their Cleavage Sites	133
6.5 Applications for DNA and RNA Nucleotide Sequences	136
6.6 Prediction Performance Evaluation	153
6.7 Different Performance Measures	155
7 Hidden Markov Models: The Theory	165
7.1 Introduction	165
7.2 Prior Information and Initialization	170
7.3 Likelihood and Basic Algorithms	172
7.4 Learning Algorithms	177
7.5 Applications of HMMs: General Aspects	184
8 Hidden Markov Models: Applications	189
8.1 Protein Applications	189
8.2 DNA and RNA Applications	209
8.3 Advantages and Limitations of HMMs	222
9 Probabilistic Graphical Models in Bioinformatics	225
9.1 The Zoo of Graphical Models in Bioinformatics	225
9.2 Markov Models and DNA Symmetries	230
9.3 Markov Models and Gene Finders	234
9.4 Hybrid Models and Neural Network Parameterization of Graphical Models	239
9.5 The Single-Model Case	241
9.6 Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction	255
10 Probabilistic Models of Evolution: Phylogenetic Trees	265
10.1 Introduction to Probabilistic Models of Evolution	265
10.2 Substitution Probabilities and Evolutionary Rates	267
10.3 Rates of Evolution	269
10.4 Data Likelihood	270
10.5 Optimal Trees and Learning	273

10.6	Parsimony	273
10.7	Extensions	275

11 Stochastic Grammars and Linguistics 277

11.1	Introduction to Formal Grammars	277
11.2	Formal Grammars and the Chomsky Hierarchy	278
11.3	Applications of Grammars to Biological Sequences	284
11.4	Prior Information and Initialization	288
11.5	Likelihood	289
11.6	Learning Algorithms	290
11.7	Applications of SCFGs	292
11.8	Experiments	293
11.9	Future Directions	295

12 Microarrays and Gene Expression 299

12.1	Introduction to Microarray Data	299
12.2	Probabilistic Modeling of Array Data	301
12.3	Clustering	313
12.4	Gene Regulation	320

13 Internet Resources and Public Databases 323

13.1	A Rapidly Changing Set of Resources	323
13.2	Databases over Databases and Tools	324
13.3	Databases over Databases in Molecular Biology	325
13.4	Sequence and Structure Databases	327
13.5	Sequence Similarity Searches	333
13.6	Alignment	335
13.7	Selected Prediction Servers	336
13.8	Molecular Biology Software Links	341
13.9	Ph.D. Courses over the Internet	343
13.10	Bioinformatics Societies	344
13.11	HMM/NN simulator	344

A Statistics 347

A.1	Decision Theory and Loss Functions	347
A.2	Quadratic Loss Functions	348
A.3	The Bias/Variance Trade-off	349
A.4	Combining Estimators	350
A.5	Error Bars	351
A.6	Sufficient Statistics	352
A.7	Exponential Family	352
A.8	Additional Useful Distributions	353

A.9	Variational Methods	354
B	Information Theory, Entropy, and Relative Entropy	357
B.1	Entropy	357
B.2	Relative Entropy	359
B.3	Mutual Information	360
B.4	Jensen's Inequality	361
B.5	Maximum Entropy	361
B.6	Minimum Relative Entropy	362
C	Probabilistic Graphical Models	365
C.1	Notation and Preliminaries	365
C.2	The Undirected Case: Markov Random Fields	367
C.3	The Directed Case: Bayesian Networks	369
D	HMM Technicalities, Scaling, Periodic Architectures, State Functions, and Dirichlet Mixtures	375
D.1	Scaling	375
D.2	Periodic Architectures	377
D.3	State Functions: Bendability	380
D.4	Dirichlet Mixtures	382
E	Gaussian Processes, Kernel Methods, and Support Vector Machines	387
E.1	Gaussian Process Models	387
E.2	Kernel Methods and Support Vector Machines	389
E.3	Theorems for Gaussian Processes and SVMs	395
F	Symbols and Abbreviations	399
	References	409
	Index	447