## BIOINFORMATICS

THE MACHINE LEARNING APPROACH
SECOND EDITION

PIERRE BALDI AND SØREN BRUNAK



RITTCCTARRA

5630042467497 krosbyrscror

EBBARBETRADE ARBABARTETET MARKARAN

> GCA A370 GABAY

TTCTCTGAGS TSAISTTTSA KGCCATGAGT

GAGARTTCATG GAAAGAGACG BUTGARDOGGC

TGBAGTGATBAG AGCTTGGTGCTG GGAACCGAGGG MDGAACCAGCT

STETTOADBEAG AGAATAOGGTET GAGGEAGGGAA ATAAGGETGATT ACACTRACTYTE

CCAACTTTCCAC STOCKAGTALCCA NAGGACCAGAGG

ATECCAPETAT EN TETATRICEA ATGARAGACTO

CTCCCTTCCCI ACCARTCARR BTAGGTGGCC

TAST

1101014400 CTO1412516A CCTTTBCTTB

STABABISTAC TCDITTICAAT ATCTCTOCITI TYSSTITTATI

TIGTOTOTATOT TTGTCCAAGACO CDAGAAGCTOGG ACCCTCCACCOT

STANCSCASTO TTANCACRASC GASCYCASAGA GOOMSGACTCA

ITIAATOISIT MAAASOICIOI SOCYSCISSO ASATOSATGA

> TTAGG AAGET ATETÉ

ACAGAS

OCCCCAAAA
TAAGCTAAAT
AAAGCTAAAT
CTGGGGGGA
CTTAAGGGA
TGGCGCGGG

AGATAAATBC ECTSATTUCC TOOTTIGGAS CTAATCAGAA AAAATACAA AATGCSACCA

AATOCAACCA! CITTOTETTTI COTTITTITI CACEBACTCI TITOTCACCCE CICEBACTACA! GTOCBATCCC!

CACTOCANA CCSTTTONA CCTSCCTCA CCCSACTAA CACTACACO ACCCACCAT COCCACCAT CCCTACTT

TITOTATTI
TAGABACOS
TOACOSTOT
CAGOSTAGT
ATUTCOTGA
LINEN GIGATO

## **Contents**

	Series For	eword	ix
	Preface		xi
1	Introducti	on	1
	1.1	Biological Data in Digital Symbol Sequences	1 7
	1.2	Genomes—Diversity, Size, and Structure	
	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-L	earning 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	Probabilis	tic Modeling and Inference: Examples	67
	3.1	The Simplest Sequence Models	67
	3.2	Statistical Mechanics	73
4	Machine 1	earning 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 Net	works: 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 Net	works: 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 Ma	rkov Models: The Theory	165
	<b>7.</b> 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 Ma	rkov Models: Applications	189
8	Hidden Ma 8.1	rkov Models: Applications Protein Applications	1 <b>89</b> 189
8			
8	8.1	Protein Applications	189
	8.1 8.2 8.3	Protein Applications DNA and RNA Applications	189 209
	8.1 8.2 8.3	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs	189 209 222 <b>225</b> 225
	8.1 8.2 8.3 Probabilist	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs ic Graphical Models in Bioinformatics	189 209 222 <b>225</b>
	8.1 8.2 8.3 <b>Probabilist</b> 9.1	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics	189 209 222 <b>225</b> 225
	8.1 8.2 8.3 <b>Probabilist</b> 9.1 9.2	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries	189 209 222 225 225 230
	8.1 8.2 8.3 <b>Probabilist</b> 9.1 9.2 9.3	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries Markov Models and Gene Finders	189 209 222 225 225 230
	8.1 8.2 8.3 <b>Probabilist</b> 9.1 9.2 9.3	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries Markov Models and Gene Finders Hybrid Models and Neural Network Parameterization of	189 209 222 <b>225</b> 225 230 234
	8.1 8.2 8.3 <b>Probabilist</b> 9.1 9.2 9.3 9.4	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  cic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries Markov Models and Gene Finders Hybrid Models and Neural Network Parameterization of Graphical Models	189 209 222 225 225 230 234 239
	8.1 8.2 8.3 <b>Probabilist</b> 9.1 9.2 9.3 9.4	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries Markov Models and Gene Finders Hybrid Models and Neural Network Parameterization of Graphical Models The Single-Model Case	189 209 222 225 225 230 234 239
9	8.1 8.2 8.3 <b>Probabilist</b> 9.1 9.2 9.3 9.4 9.5 9.6	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries Markov Models and Gene Finders Hybrid Models and Neural Network Parameterization of Graphical Models The Single-Model Case Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction  tic Models of Evolution: Phylogenetic Trees	189 209 222 225 225 230 234 239 241
9	8.1 8.2 8.3 <b>Probabilist</b> 9.1 9.2 9.3 9.4 9.5 9.6	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries Markov Models and Gene Finders Hybrid Models and Neural Network Parameterization of Graphical Models The Single-Model Case Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction	189 209 222 225 230 234 239 241 255
9	8.1 8.2 8.3 Probabilist 9.1 9.2 9.3 9.4 9.5 9.6	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries Markov Models and Gene Finders Hybrid Models and Neural Network Parameterization of Graphical Models The Single-Model Case Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction  tic Models of Evolution: Phylogenetic Trees	189 209 222 225 230 234 239 241 255 265
9	8.1 8.2 8.3 Probabilist 9.1 9.2 9.3 9.4 9.5 9.6 Probabilist 10.1	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries Markov Models and Gene Finders Hybrid Models and Neural Network Parameterization of Graphical Models The Single-Model Case Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction  ic Models of Evolution: Phylogenetic Trees Introduction to Probabilistic Models of Evolution	189 209 222 225 230 234 239 241 255 265 267 269
9	8.1 8.2 8.3 Probabilist 9.1 9.2 9.3 9.4 9.5 9.6 Probabilist 10.1 10.2	Protein Applications DNA and RNA Applications Advantages and Limitations of HMMs  ic Graphical Models in Bioinformatics The Zoo of Graphical Models in Bioinformatics Markov Models and DNA Symmetries Markov Models and Gene Finders Hybrid Models and Neural Network Parameterization of Graphical Models The Single-Model Case Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction  tic Models of Evolution: Phylogenetic Trees Introduction to Probabilistic Models of Evolution Substitution Probabilities and Evolutionary Rates	189 209 222 225 230 234 239 241 255 265 265

Contents	vii

	10.6	Parsimony	2/3
	10.7	Extensions	275
11 St	ochastic (	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 Mi	icroarray	s 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 In	ternet Re	esources 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		341
	13.9	Ph.D. Courses over the Internet	343
	13.10		344
	13.11	HMM/NN simulator	344
A St	atistics		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 Q** 

R.1

**R.2** 

**R.3** 

**B.4** 

R 5

**R** 6

C.1

References

index

359

360

361

361

362

365

365

367

369

375

375

377

380

382

387

Contente

Information Theory. Entropy, and Relative Entropy Relative Entropy Mutual Information

Jensen's Inequality

Maximum Entropy Minimum Relative Entropy

C Probabilistic Graphical Models Notation and Preliminaries The Undirected Case: Markov Random Fields

Variational Methods

Entropy

C.3 The Directed Case: Bavesian Networks D HMM Technicalities, Scaling, Periodic Architectures,

State Functions. and Dirichlet Mixtures D1Scaling D.2

Periodic Architectures D.3 State Functions: Bendability Dirichlet Mixtures

D4 E Gaussian Processes, Kernel Methods, and Support

Vector Machines F.1 Gaussian Process Models E.2

F.3

Symbols and Abbreviations

Theorems for Gaussian Processes and SVMs

Kernel Methods and Support Vector Machines

387

389 395

399

409

447