

Contents

1	Introduction	1
1.1	Some Basic Genetics	1
1.2	The Central Dogma	4
1.3	The Structure of a Gene	6
1.4	How Many Genes Do We Have?	8
1.5	Problems of Gene Definitions	11
1.6	The Gene Finding Problem	13
1.7	Comparative Gene Finding	15
1.8	History of Algorithm Development	16
1.9	To Build a Gene Finder	22
	References	23
2	Single Species Gene Finding	29
2.1	Hidden Markov Models (HMMs)	29
2.1.1	Markov Chains	30
2.1.2	Hidden Markov Models	41
2.1.3	Dynamic Programming	44
2.1.4	The Forward Algorithm	47
2.1.5	The Backward Algorithm	48
2.1.6	The Viterbi Algorithm	49
2.1.7	EasyGene: A Prokaryotic Gene Finder	52
2.2	Generalized Hidden Markov Models (GHMMs)	55
2.2.1	Preliminaries	55
2.2.2	The Forward and Backward Algorithms	57
2.2.3	The Viterbi Algorithm	59
2.2.4	Genscan: A GHMM-Based Gene Finder	60
2.3	Interpolated Markov Models (IMMs)	70
2.3.1	Preliminaries	71
2.3.2	Linear and Rational Interpolation	71
2.3.3	GLIMMER: A Microbial Gene Finder	73

2.4	Neural Networks	76
2.4.1	Biological Neurons	76
2.4.2	Artificial Neurons and the Perceptron	77
2.4.3	Multilayer Neural Networks	80
2.4.4	GRAIL: A Neural Network-Based Gene Finder	81
2.5	Decision Trees	84
2.5.1	Classification	84
2.5.2	Decision Tree Learning	86
2.5.3	MORGAN: A Decision Tree-Based Gene Finder	89
2.6	Conditional Random Fields	91
2.6.1	Preliminaries	91
2.6.2	Generative Versus Discriminative Models	92
2.6.3	Graphical Models and Markov Random Fields	94
2.6.4	Conditional Random Fields (CRFs)	98
2.6.5	Conrad: CRF-Based Gene Prediction	100
	References	103
3	Sequence Alignment	107
3.1	Pairwise Sequence Alignment	107
3.1.1	Dot Plot Matrix	109
3.1.2	Nucleotide Substitution Models	110
3.1.3	Amino Acid Substitution Models	116
3.1.4	Gap Models	125
3.1.5	The Needleman–Wunsch Algorithm	126
3.1.6	The Smith–Waterman Algorithm	130
3.1.7	Pair Hidden Markov Models (PHMMs)	132
3.1.8	Database Similarity Searches	136
3.1.9	The Significance of Alignment Scores	141
3.2	Multiple Sequence Alignment	143
3.2.1	Scoring Schemes	144
3.2.2	Phylogenetic Trees	147
3.2.3	Dynamic Programming	149
3.2.4	Progressive Alignments	152
3.2.5	Iterative Methods	155
3.2.6	Hidden Markov Models	158
3.2.7	Genetic Algorithms	160
3.2.8	Simulated Annealing	163
3.2.9	Alignment Profiles	166
	References	171
4	Comparative Gene Finding	175
4.1	Similarity-Based Gene Finding	175
4.1.1	GenomeScan: GHMM-Based Gene Finding Using Homology	176

4.1.2	Twinscan: GHMM-Based Gene Finding Using Informant Sequences	178
4.2	Heuristic Cross-Species Gene Finding	180
4.2.1	ROSETTA: A Heuristic Cross-Species Gene Finder	180
4.3	Pair Hidden Markov Models (PHMMs)	182
4.3.1	DoubleScan: A PHMM-Based Comparative Gene Finder	182
4.4	Generalized Pair Hidden Markov Models (GPHMMs)	185
4.4.1	Preliminaries	185
4.4.2	SLAM: A GPHMM-Based Comparative Gene Finder	188
4.5	Gene Mapping	192
4.5.1	Projector: A Gene Mapping Tool	193
4.5.2	GeneMapper—Reference-Based Annotation	194
4.6	Multiple Sequence Gene Finding	195
4.6.1	N-SCAN: A Multiple Informant-Based Gene Finder	196
	References.	198
5	Gene Structure Submodels	201
5.1	The State Space	201
5.1.1	The Exon States	203
5.1.2	Splice Sites	204
5.1.3	Introns and Intergenic Regions	205
5.1.4	Untranslated Regions (UTRs)	206
5.1.5	Promoters and PolyA-Signals	207
5.2	State Length Distributions	208
5.2.1	Geometric and Negative Binomial Lengths.	209
5.2.2	Empirical Length Distributions	211
5.2.3	Acyclic Discrete Phase-Type Distributions	213
5.3	Sequence Content Sensors	217
5.3.1	GC-Content Binning	217
5.3.2	Start Codon Recognition	218
5.3.3	Codon and Amino Acid Usage	219
5.3.4	K-Tuple Frequency Analysis	220
5.3.5	Markov Chain Content Sensors	222
5.3.6	Interpolated Markov Models.	224
5.4	Splice Site Detection	225
5.4.1	Weight Matrices and Weight Array Models	225
5.4.2	Variable-Length Markov Models (VLMMs)	228
5.4.3	Maximal Dependence Decomposition (MDD).	230
5.4.4	Neural Networks.	236
5.4.5	Linear Discriminant Analysis	238

5.4.6	Maximum Entropy	243
5.4.7	Bayesian Networks	249
5.4.8	Support Vector Machines	255
	References	264
6	Parameter Training	269
6.1	Introduction	269
6.2	Pseudocounts	270
6.3	Maximum Likelihood Estimation	273
6.4	The Expectation–Maximization (EM) Algorithm	279
6.5	The Baum–Welch Algorithm	286
6.6	Gradient Ascent/Descent	290
6.7	The Backpropagation Algorithm	293
6.8	Discriminative Training	299
6.9	Gibbs Sampling	303
6.10	Simulated Annealing	305
	References	308
7	Implementation of a Comparative Gene Finder	311
7.1	Program Structure	311
7.1.1	Command Line Arguments	312
7.1.2	Parameter Files	314
7.1.3	Candidate Exon Boundaries	316
7.1.4	Output Files	317
7.2	The GPHMM Model	318
7.2.1	Modeling Intron and Intergenic Pairs	319
7.2.2	Modeling Exon Pairs	320
7.2.3	Approximate Alignment	321
7.3	Accuracy Assessment	322
7.4	Possible Model Extensions	323
	References	324
8	Annotation Pipelines for Next-Generation Sequencing Projects	325
8.1	Introduction	325
8.2	History of DNA Sequencing	326
8.2.1	The Origin of Bioinformatics	331
8.3	Next-Generation Sequencing (NGS)	333
8.3.1	NGS Technologies	334
8.3.2	Genome Sequence Assembly	336
8.3.3	NGS Applications	342
8.4	NGS Genome Sequencing Annotation Pipelines	349
8.4.1	Assembly Quality	350
8.4.2	Repeat Masking	350
8.4.3	Gene Annotation	352

8.4.4	De Novo Annotation Assessment	355
8.4.5	MAKER: An Annotation Pipeline for Next-Generation Sequencing Projects	357
	References	359
	Index	369