Contents

Foreword *Steven L. Salzberg* *i*
Preface *iii*
1. Introduction **1**
1.1 The Central Dogma of Molecular Biology 1
1.2 Evolution 12
1.3 Genome Sequencing and Assembly 15
1.4 Genomic Annotation 19
1.5 The Problem of Computational Gene Prediction 25
Exercises 26
2. Mathematical Preliminaries **29**
2.1 Numbers and Functions 29
2.2 Logic and Boolean Algebra 34
2.3 Sets 36
2.4 Algorithms and Pseudocode 37
2.5 Optimization 40
2.6 Probability 42
2.7 Some Important Distributions 51
2.8 Parameter Estimation 57
2.9 Statistical Hypothesis Testing 58
2.10 Information 62
2.11 Computational Complexity 66
2.12 Dynamic Programming 68
2.13 Searching and Sorting 71
2.14 Graphs 72

2.15 Languages and Parsing 79
Exercises 84

3. Overview of Gene Prediction 87

3.1 Genes, Exons, and Coding Segments 87
3.2 Orientation 91
3.3 Phase and Frame 93
3.4 Gene Finding as Parsing 97
3.5 Common Assumptions in Gene Prediction 103
3.5.1 No overlapping genes 103
3.5.2 No nested genes 104
3.5.3 No partial genes 104
3.5.4 No non-canonical signal consensuses 104
3.5.5 No frame shifts or sequencing errors 105
3.5.6 Optimal parse only 105
3.5.7 Constraints on feature lengths 105
3.5.8 No split start codons 106
3.5.9 No split stop codons 106
3.5.10 No alternative splicing 106
3.5.11 No selenocysteine codons 106
3.5.12 No ambiguity codes 107
3.5.13 One haplotype only 107
Exercises 108

4. Gene Finder Evaluation 109

4.1 Testing Protocols 109
4.2 Evaluation Metrics 118
Exercises 124

5. A Toy Exon Finder 125

5.1 The Toy Genome and its Toy Genes 125
5.2 Random Exon Prediction as a Baseline 128
5.3 Predicting Exons Based on {G,C} Bias 132
5.4 Predicting Exons Based on Codon Bias 134
5.5 Predicting Exons Based on Codon Bias and WMM Score 136
5.6 Summary 140
Exercises 141

6. Hidden Markov Models 143

6.1 Introduction to HMM's 143
6.1.1 An Illustrative Example 145
6.1.2 Representing HMM's 146
6.2 Decoding and Similar Problems 147
6.2.1 Finding the Most Probable Path 147
6.2.2 Computing the Probability of a Sequence 151

6.3 Training with Labeled Sequences 153
6.4 Example: Building an HMM for Gene Finding 155
6.5 Case Study: VEIL and UNVEIL 165
6.6 Using Ambiguous Models 167
6.6.1 Viterbi Training 168
6.6.2 Merging Submodels 169
6.6.3 Baum-Welch Training 171
6.6.3.1 Naive Baum-Welch Algorithm 171
6.6.3.2 Baum-Welch with Scaling 174
6.7 Higher-order HMM's 178
6.7.1 Labeled Sequence Training for Higher-order HMM's 179
6.7.2 Decoding with Higher-order HMM's 180
6.8 Variable-order HMM's 180
6.8.1 Back-off Models 181
6.8.2 Example: Incorporating Variable-order Emissions 182
6.8.3 Interpolated Markov Models 182
6.9 Discriminative Training of HMM's 184
6.10 Posterior Decoding of HMM's 187
Exercises 189

7. Signal and Content Sensors 195

7.1 Overview of Feature Sensing 195
7.2 Content Sensors 196
7.2.1 Markov Chains 196
7.2.2 Markov Chain Implementation 200
7.2.3 Improved Markov Chain Implementation 200
7.2.4 Three-periodic Markov Chains 202
7.2.5 Interpolated Markov Chains 203
7.2.6 Nonstationary Markov Chains 204
7.3 Signal Sensors 205
7.3.1 Weight Matrices 207
7.3.2 Weight Array Matrices 209
7.3.3 Windowed Weight Array Matrices 211
7.3.4 Local Optimility Criterion 211
7.3.5 Coding-Noncoding Boundaries 213
7.3.6 Case Study: GeneSplicer 214
7.3.7 Maximal Dependence Decomposition 214
7.3.8 Interpolated Context Models 218
7.3.9 Case Study: Signal Sensing in GENSCAN 221
7.4 Other Methods of Feature Sensing 222
7.5 Case Study: Bacterial Gene Finding 223
Exercises 225

8. Generalized Hidden Markov Models **227**
8.1 Generalization and its Advantages 227
8.2 Typical Model Topologies 232
8.2.1 One Exon Model or Four? 234
8.2.2 One Strand or Two? 235
8.3 Decoding with a GHMM 236
8.3.1 PSA Decoding 242
8.3.2 DSP Decoding 252
8.3.3 Equivalence of DSP and PSA 257
8.3.4 A DSP Example 260
8.3.5 Shortcomings of DSP and PSA 262
8.4 Higher-fidelity Modeling 263
8.4.1 Modeling Isochores 263
8.4.2 Explicit Modeling of Noncoding Lengths 265
8.5 Prediction with an ORF Graph 268
8.5.1 Building the Graph 268
8.5.2 Decoding with a Graph 269
8.5.3 Extracting Suboptimal Parses 270
8.5.4 Posterior Decoding for GHMM's 271
8.5.5 The ORF Graph as a Data Interchange Format 273
8.6 Training a GHMM 276
8.6.1 Maximum Likelihood Training for GHMM's 277
8.6.2 Discriminative Training for GHMM's 277
8.7 Example: GHMM Versus HMM 281
Exercises 281
9. Comparative Gene Finding **285**
9.1 Informant Techniques 287
9.1.1 Case Study: TWINSCAN 287
9.1.2 Case Study: GenomeScan 289
9.1.3 Case Study: SGP-2 290
9.1.4 Case Study: HMMgene 291
9.1.5 Case Study: GENIE 292
9.2 Combiners 294
9.2.1 Case Study: JIGSAW 294
9.2.2 Case Study: GAZE 296
9.3 Alignment-based Prediction 296
9.3.1 Case Study: ROSETTA 297
9.3.2 Case Study: SGP-1 297
9.3.3 Case Study: CEM 298
9.4 Pair HMM's 300
9.4.1 Case Study: Doublescan 305
9.5 Generalized Pair HMM's 307
9.5.1 Case Study: TWAIN 308

9.6 Phylogenomic Gene Finding 319
9.6.1 Phylogenetic HMM's 320
9.6.2 Decoding with a PhyloHMM 326
9.6.3 Evolution Models 328
9.6.4 Parameterization of Rate Matrices 332
9.6.5 Estimation of Evolutionary Parameters 335
9.6.6 Modeling Higher-order Dependencies 339
9.6.7 Enhancing Discriminative Power 341
9.6.8 Selection of Informants 341
9.7 Auto-annotation Pipelines 342
9.8 Looking Toward the Future 343
Exercises 344

10. Machine Learning Methods 347

10.1 Overview of Automatic Classification 347
10.2 K-Nearest Neighbors 351
10.3 Naive Bayes Models 352
10.4 Bayesian Networks 353
10.5 Neural Networks 355
10.5.1 Case Study: GRAIL 359
10.6 Decision Trees 360
10.6.1 Case Study: GlimmerM 363
10.7 Linear Discriminant Analysis 363
10.8 Quadratic Discriminant Analysis 366
10.9 Multivariate Regression 366
10.10 Logistic Regression 367
10.11 Regularized Logistic Regression 369
10.12 Genetic Programming 370
10.13 Simulated Annealing 373
10.14 Support Vector Machines 374
10.15 Hill Climbing with the GSL 376
10.16 Feature Selection and Dimensionality Reduction 377
10.17 Applications 378
Exercises 380

11. Tips and Tricks 383

11.1 Boosting 383
11.2 Bootstrapping 384
11.3 Modeling Additional Gene Features 386
11.4 Masking Repeats 391
Exercises 392

12. Advanced Topics 395

12.1 Alternative Splicing 395

12.2 Prediction of Noncoding Genes 399
12.3 Promoter Prediction 406
12.4 Generative Versus Discriminative Modeling 408
12.5 Parallelization and Grid Computing 411
Exercises 413

Appendix A – Online Resources 415

A.1 Official Book Website 415
A.2 Open Source Gene Finders 415
A.3 Gene-finding Web Sites 416
A.4 Gene-finding Bibliographies 416

References 417

Index 437