# CONTENTS

Extended contents ix  
Preface xv  
Acknowledgments xxi  
Editors and contributors xxiv  
A computational micro primer xxvi

## PART I Genomes 1

1 Identifying the genetic basis of disease 3  
   *Vineet Bafna*

2 Pattern identification in a haplotype block 23  
   *Kun-Mao Chao*

3 Genome reconstruction: a puzzle with a billion pieces 36  
   *Phillip E. C. Compeau and Pavel A. Pevzner*

4 Dynamic programming: one algorithmic key for many biological locks 66  
   *Mikhail Gelfand*

5 Measuring evidence: who’s your daddy? 93  
   *Christopher Lee*

## PART II Gene Transcription and Regulation 109

6 How do replication and transcription change genomes? 111  
   *Andrey Grigoriev*

7 Modeling regulatory motifs 126  
   *Sridhar Hannenhalli*

8 How does the influenza virus jump from animals to humans? 148  
   *Haixu Tang*
PART III  Evolution  165
  9  Genome rearrangements  167
    Steffen Heber and Brian E. Howard
  10  Comparison of phylogenetic trees and search for a central trend in the “Forest of Life”  189
    Eugene V. Koonin, Pere Puigbò, and Yuri I. Wolf
  11  Reconstructing the history of large-scale genomic changes: biological questions and computational challenges  201
    Jian Ma

PART IV  Phylogeny  225
  12  Figs, wasps, gophers, and lice: a computational exploration of coevolution  227
    Ran Libeskind-Hadas
  13  Big cat phylogenies, consensus trees, and computational thinking  248
    Seung-Jin Sul and Tiffani L. Williams
  14  Phylogenetic estimation: optimization problems, heuristics, and performance analysis  267
    Tandy Warnow

PART V  Regulatory Networks  289
  15  Biological networks uncover evolution, disease, and gene functions  291
    Nataša Pržulj
  16  Regulatory network inference  315
    Russell Schwartz

Glossary  344
Index  350
EXTENDED CONTENTS

Preface xv
Acknowledgments xxi
Editors and contributors xxiv
A computational micro primer xxvi

PART I Genomes 1

1 Identifying the genetic basis of disease 3
   Vineet Bafna

1 Background 3
2 Genetic variation: mutation, recombination, and coalescence 6
3 Statistical tests 9
3.1 LD and statistical tests of association 12
4 Extensions 12
4.1 Continuous phenotypes 12
4.2 Genotypes and extensions 14
4.3 Linkage versus association 15
5 Confound it 16
5.1 Sampling issues: power, etc. 16
5.2 Population substructure 17
5.3 Epistasis 18
5.4 Rare variants 19
   Discussion 20
   Questions 20
   Further Reading 21
Extended contents

2 Pattern identification in a haplotype block 23
   Kun-Mao Chao
   1 Introduction 23
   2 The tag SNP selection problem 25
   3 A reduction to the set-covering problem 26
   4 A reduction to the integer-programming problem 30
       Discussion 33
       Questions 33
       Bibliographic notes and further reading 34

3 Genome reconstruction: a puzzle with a billion pieces 36
   Phillip E. C. Compeau and Pavel A. Pevzner
   1 Introduction to DNA sequencing 36
      1.1 DNA sequencing and the overlap puzzle 36
      1.2 Complications of fragment assembly 38
   2 The mathematics of DNA sequencing 40
      2.1 Historical motivation 40
      2.2 Graphs 43
      2.3 Eulerian and Hamiltonian cycles 43
      2.4 Euler's Theorem 44
      2.5 Euler's Theorem for directed graphs 45
      2.6 Tractable vs. intractable problems 48
   3 From Euler and Hamilton to genome assembly 49
      3.1 Genome assembly as a Hamiltonian cycle problem 49
      3.2 Fragment assembly as an Eulerian cycle problem 50
      3.3 De Bruijn graphs 52
      3.4 Read multiplicities and further complications 54
   4 A short history of read generation 55
      4.1 The tale of three biologists: DNA chips 55
      4.2 Recent revolution in DNA sequencing 58
   5 Proof of Euler's Theorem 58
       Discussion 63
       Notes 63
       Questions 64

4 Dynamic programming: one algorithmic key for many biological locks 66
   Mikhail Gelfand
   1 Introduction 66
   2 Graphs 69
   3 Dynamic programming 70
   4 Alignment 77
   5 Gene recognition 81
6 Dynamic programming in a general situation. Physics of polymers 83
Answers to quiz 86
History, sources, and further reading 91

5 Measuring evidence: who's your daddy? 93
Christopher Lee
1 Welcome to the Maury Povich Show! 93
1.1 What makes you you 94
1.2 SNPs, forensics, Jacques, and you 96
2 Inference 97
2.1 The foundation: thinking about probability “conditionally” 97
2.2 Bayes’ Law 100
2.3 Estimating disease risk 100
2.4 A recipe for inference 102
3 Paternity inference 103
Questions 108

PART II Gene Transcription and Regulation 109
6 How do replication and transcription change genomes? 111
Andrey Grigoriev
1 Introduction 111
2 Cumulative skew diagrams 112
3 Different properties of two DNA strands 116
4 Replication, transcription, and genome rearrangements 120
Discussion 124
Questions 125

7 Modeling regulatory motifs 126
Sridhar Hannenhalli
1 Introduction 126
2 Experimental determination of binding sites 129
3 Consensus 130
4 Position Weight Matrices 132
5 Higher-order PWM 134
6 Maximum dependence decomposition 135
7 Modeling and detecting arbitrary dependencies 138
8 Searching for novel binding sites 139
8.1 A PWM-based search for binding sites 140
8.2 A graph-based approach to binding site prediction 140
9 Additional hallmarks of functional TF binding sites 141
9.1 Evolutionary conservation 142
9.2 Modular interactions between TFs 142
8 How does the influenza virus jump from animals to humans? 148

Haixu Tang

Introduction 148

1 Host switch of influenza: molecular mechanisms 151
   2.1 Diversity of glycan structures 152
   2.2 Molecular basis of the host specificity of influenza viruses 155
   2.3 Profiling of hemagglutinin–glycan interaction by using glycan arrays 156

3 The glycan motif finding problem 157
   Discussion 161
   Questions 161
Further Reading 163

PART III Evolution 165

9 Genome rearrangements 167

Steffen Heber and Brian E. Howard

1 Review of basic biology 167
2 Distance metrics and the genome rearrangement problem 171
3 Unsigned reversals 175
4 Signed reversals 178
5 DCJ operations and algorithms for multiple chromosomes 180
   Discussion 186
   Questions 187

10 Comparison of phylogenetic trees and search for a central trend in the
   “Forest of Life” 189

Eugene V. Koonin, Pere Puigbó, and Yuri I. Wolf

1 The crisis of the Tree of Life in the age of genomics 189
2 The bioinformatic pipeline for analysis of the Forest of Life 193
3 Trends in the Forest of Life 195
   3.1 The NUTs contain a consistent phylogenetic signal, with independent HGT events 195
   3.2 The NUTs versus the FOL 198
   Discussion: the Tree of Life concept is changing, but is not dead 199
   Questions 200

11 Reconstructing the history of large-scale genomic changes: biological
   questions and computational challenges 201

Jian Ma

1 Comparative genomics and ancestral genome reconstruction 202
   1.1 The Human Genome Project 202
1.2 Comparative genomics 202
1.3 Genome reconstruction provides an additional dimension for comparative genomics 205
1.4 Base-level ancestral reconstruction 206
2 Cross-species large-scale genomic changes 207
2.1 Genome rearrangements 207
2.2 Synteny blocks 209
2.3 Duplications and other structural changes 211
3 Reconstructing evolutionary history 211
3.1 Ancestral karyotype reconstruction 211
3.2 Rearrangement-based ancestral reconstruction 212
3.3 Adjacency-based ancestral reconstruction 213
3.4 Challenges and future directions 217
4 Chromosomal aberrations in human disease genomes 219
Discussion 221
Questions 221

PART IV Phylogeny 225
12 Figs, wasps, gophers, and lice: a computational exploration of coevolution 227
Ran Libeskind-Hadas
1 Introduction 228
2 The cophylogeny problem 229
3 Finding minimum cost reconstructions 233
4 Genetic algorithms 235
5 How Jane works 237
6 See Jane run 241
Discussion 245
Questions 245

13 Big cat phylogenies, consensus trees, and computational thinking 248
Seung-Jin Sul and Tiffani L. Williams
1 Introduction 249
2 Evolutionary trees and the big cats 250
2.1 Evolutionary hypotheses for the pantherine lineage 251
2.2 Methodology for reconstructing pantherine phylogenetic trees 252
2.3 Implications of consensus trees on the phylogeny of the big cats 254
3 Consensus trees and bipartitions 254
3.1 Phylogenetic trees and their bipartitions 255
3.2 Representing bipartitions as bitstrings 256
4 Constructing consensus trees 256
4.1 Step 1: collecting bipartitions from a set of trees 256
4.2 Step 2: selecting consensus bipartitions 258
4.3 Step 3: constructing consensus trees from consensus bipartitions 261
Discussion 264
Questions 264
Extended contents

14 Phylogenetic estimation: optimization problems, heuristics, and performance analysis 267
Tandy Warnow

1 Introduction 268
2 Computational problems 269
2.1 The 2-colorability problem 271
2.2 Maximum independent set 274
3 NP-hardness, and lessons learned 275
4 Phylogeny estimation 277
4.1 Maximum parsimony 277
Discussion and recommended reading 286
Questions 286

PART V Regulatory Networks 289
15 Biological networks uncover evolution, disease, and gene functions 291
Nataša Pržulj

1 Interaction network data sets 293
2 Network comparisons 295
3 Network models 300
4 Using network topology to discover biological function 303
5 Network alignment 306
Discussion 312
Questions 312

16 Regulatory network inference 315
Russell Schwartz

1 Introduction 315
1.1 The biology of transcriptional regulation 317
2 Developing a formal model for regulatory network inference 320
2.1 Abstracting the problem statement 320
2.2 An intuition for network inference 322
2.3 Formalizing the intuition for an inference objective function 323
2.4 Generalizing to arbitrary numbers of genes 332
3 Finding the best model 333
4 Extending the model with prior knowledge 335
5 Regulatory network inference in practice 337
5.1 Real-valued data 338
5.2 Combining data sources 339
Discussion and further directions 341
Questions 342

Glossary 344
Index 350