

Brief Contents

Preface xi

Acknowledgment xiv

Chapter 1 Road Map 1

Chapter 2 Biological Basics 22

Chapter 3 Wet and Dry Lab Techniques 84

Chapter 4 Fragment Assembly 128

Chapter 5 Sequence Alignment 158

Chapter 6 Simulating and Modeling Evolution 219

Chapter 7 Gene Finding 276

Chapter 8 Gene Expression 329

Chapter 9 Projects 383

Index 445

Contents

Preface xi

Acknowledgment xiv

Chapter 1 Road Map 1

- 1.1 What Is Bioinformatics? 1
- 1.2 A Bioinformatics Team 6
- 1.3 What Defines Life? 8
- 1.4 A Systems Approach to Biology 10
- 1.5 Bioinformatics in Action 16
 - 1.5.1 Deciphering a Killer: HIV and Bioinformatics 17
- 1.6 The Road Ahead 18
- Summary 20
- Key Terms 20
- Bibliography 21

Chapter 2 Biological Basics 22

- 2.1 The Blind Engineer 22
 - 2.1.1 The Case of the Peppered Moth 23
 - 2.1.2 How Evolution Works 24
 - 2.1.3 Evolution's Palette 25
- 2.2 Compute Machine par Excellence 26
 - 2.2.1 Cellular Organization and Complexity 26
 - 2.2.2 Chemistry and Life 28
 - 2.2.3 A Parts List for Life 33
- 2.3 The Languages of the Cell 36
 - 2.3.1 Operating Systems for Cells 37
 - 2.3.2 Deciphering the Language of Cells 39
 - 2.3.3 Compiling DNAStrings Programs 46
 - 2.3.4 Executing Code from DNAStrings 49
- 2.4 Further Nuances in DNAStrings 57
- 2.5 Proteins: Cellular Machines 61
 - 2.5.1 Proteins as Molecules 62
 - 2.5.2 Proteins as Engineered Machines 67

2.6 Data Maintenance and Integrity Tasks 70

- 2.6.1 Backing up DNA Data 70
- 2.6.2 The Challenges of Data Management 74

Key Terms 82

Bibliography 83

Chapter 3 Wet and Dry Lab Techniques 84

- 3.1 Hybridization: Putting Base Pairs to Work 85
- 3.2 Making Copies of Nucleotide Sequences 87
- 3.3 An Explosion of Copies 91
- 3.4 Sequencing DNA Strings 94
- 3.5 The Human Genome Project: Computing to the Rescue 99
 - 3.5.1 Mission Impossible: Sequencing the Human Genome 100
- 3.6 Human Genome Sequencing Strategies 105
- 3.7 From Structure to Function 107
- 3.8 Profiling the Transcriptome 111
- 3.9 A Few Proteomics Techniques 115
- 3.10 Putting It All Together 118
- 3.11 A Few Selected Dry Lab Techniques 120
 - 3.11.1 Algorithms 120
 - 3.11.2 Analysis 122
- Key Terms 125
- Bibliography 126

Chapter 4 Fragment Assembly 128

- 4.1 The Nature of the Problem 128
 - 4.1.1 Two Analogies 128
 - 4.1.2 The Need for Multiple Coverage 130
- 4.2 Putting the Pieces Together 132
 - 4.2.1 Location, Location, Location 132

- 6.5 Discovering Evolutionary Relationships 249
 - 6.5.1 Parsimony 252
 - 6.5.2 Other Ways to Build Trees 263
 - 6.5.3 Maximum Likelihood 270

Key Terms 274

Bibliography 275

Chapter 7 Gene Finding 276

- 7.1 A Modern Cryptographic Puzzle 276
 - 7.1.1 Detecting Encryption 278
 - 7.1.2 Encoding Information 279
- 7.2 Cracking the Genome: A First Pass 281
 - 7.2.1 A Worked Example: HIV Integration Sites 282
 - 7.2.2 Regulating Genes: Transcription Factor-Binding Sites 293
- 7.3 A Biological Decoder Ring 295
 - 7.3.1 A First Try at Decryption: ORF Finding 297
 - 7.3.2 Accounting for Discontinuous Coding Regions 301
- 7.4 Finding Genes Through Mathematics 306
 - 7.4.1 Linguistic Complexity 307
 - 7.4.2 Looks Like a ... 307
 - 7.4.3 Markov Models 308
 - 7.4.4 Genes as Markov Processes 314
- 7.5 Gene Finding by Learning: Letting a Computer Do It 317
- 7.6 Exercises for Chapter 7 319
 - Key Terms 327
 - Bibliography 327

Chapter 8 Gene Expression 329

- 8.1 Introduction 329
- 8.2 Genes in Context 329
- 8.3 Genotype to Phenotype 330
- 8.4 The Expected (by now) Complications of Biology 331
- 8.5 A Flood of Data 334
- 8.6 Noisy Data 337
 - 8.6.1 Turning down the Noise 338
- 8.7 The Many Modes of Gene Expression Data 339

8.8 A Worked Example: Gene Expression in HIV-Infected Cells 342

- 8.8.1 Data Preprocessing 342

8.9 Programs to Work with Genes and Expression Vectors 350

8.10 Mining the Gene Expression Data 352

- 8.10.1 A Worked Example: Looking for Differentially Expressed Genes 353
- 8.10.2 Testing Biological Hypotheses with Statistical Hypotheses 354

8.11 A Worked Example: Forming New Hypotheses 356

- 8.11.1 Organizing the Data 356
- 8.11.2 Clustering 358
- 8.11.3 Classification 370
- 8.11.4 Using Visualization Techniques to Aid Interpretation 371
- 8.11.5 Advanced Classification Algorithms 373

8.12 Data Management 374

- 8.12.1 Controlled Vocabularies and Standardization of Microarray Data 375

8.13 Exercises for Chapter 8 378

Key Terms 380

Bibliography 381

Chapter 9 Projects 383

- 9.1 Visualization and Exploration of Complex Datasets 383
 - 9.1.1 Sequencing Gel Visualization 384
 - 9.1.2 Microarray Data Visualization 389
 - 9.1.3 Data Visualization Tools 390
 - 9.1.4 Over to You 392
 - 9.1.5 Resources for Visualization 393
- 9.2 RNA Structure and Function Prediction 394
 - 9.2.1 Solving Structures for Functional RNAs: Early Successes 395
 - 9.2.2 Structural RNAs and Gene Regulation 396
 - 9.2.3 RNA Structures in Machines: Solving Complex Structures 401
 - 9.2.4 Over to You 405

9.2.5	Resources for Structure Prediction	406	9.4.3	Over to You	423
9.3	Rational Drug Design Through Protein Structure and Function Prediction	408	9.4.4	Resources for Information-Based Medicine	424
9.3.1	A Pharmaceutical Fairy Tale	408	9.5	Systems Biology	424
9.3.2	Drug Development: One in a Million Chances	409	9.5.1	Introduction	424
9.3.3	Structure-Based Drug Design	410	9.5.2	Inputs	428
9.3.4	A Pharmaceutical Cautionary Tale	414	9.5.3	Outputs	429
9.3.5	Over to You	415	9.5.4	Modern Approach to Systems Biology	429
9.3.6	Resources for Rational Drug Design	416	9.5.5	Feedback, Equilibrium, and Attractors	430
9.4	Information-Based Medicine	416	9.5.6	What Kind of Model?	435
9.4.1	Identifying Simple Disease Genes	417	9.5.7	Over to You	436
9.4.2	The Challenge of Mapping Complex Diseases	420	9.5.8	Resources for Systems Biology	439
			Key Terms	440	
			Bibliography	441	
			Index	445	