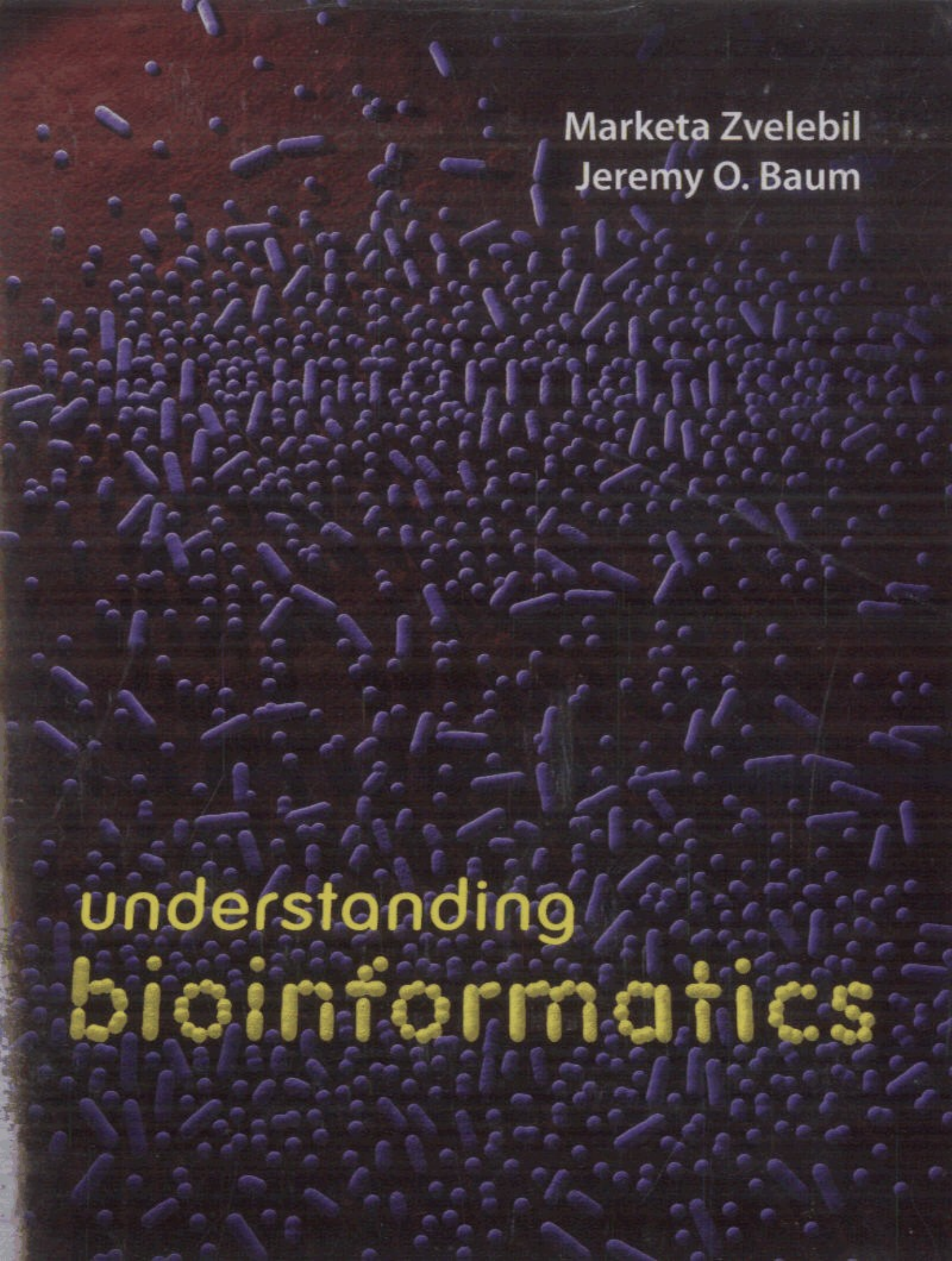
Marketa Zvelebil
Jeremy O. Baum
understanding
bioinformatics

CONTENTS

Preface v
A Note to the Reader vii
List of Reviewers xii
Contents in Brief xiii

Part 1 Background Basics

Chapter 1 The Nucleic Acid World

1.1 The Structure of DNA and RNA 5
DNA is a linear polymer of only four different bases 5
Two complementary DNA strands interact by base pairing to form a double helix 7
RNA molecules are mostly single stranded but can also have base-pair structures 9

1.2 DNA, RNA, and Protein: The Central Dogma 10
DNA is the information store, but RNA is the messenger 11
Messenger RNA is translated into protein according to the genetic code 12
Translation involves transfer RNAs and RNA-containing ribosomes 13

1.3 Gene Structure and Control 14
RNA polymerase binds to specific sequences that position it and identify where to begin transcription 15
The signals initiating transcription in eukaryotes are generally more complex than those in bacteria 17
Eukaryotic mRNA transcripts undergo several modifications prior to their use in translation 18
The control of translation 19

1.4 The Tree of Life and Evolution 20
A brief survey of the basic characteristics of the major forms of life 21
Nucleic acid sequences can change as a result of mutation 22

Summary 23
Further Reading 24

Chapter 2 Protein Structure

2.1 Primary and Secondary Structure 25
Protein structure can be considered on several different levels 26
Amino acids are the building blocks of proteins 27
The differing chemical and physical properties of amino acids are due to their side chains 28
Amino acids are covalently linked together in the protein chain by peptide bonds 29
Secondary structure of proteins is made up of α-helices and β-strands 33
Several different types of β-sheet are found in protein structures 35
Turns, hairpins and loops connect helices and strands 36

2.2 Implication for Bioinformatics 37
Certain amino acids prefer a particular structural unit 37
Evolution has aided sequence analysis 38
Visualization and computer manipulation of protein structures 38

2.3 Proteins Fold to Form Compact Structures 40
The tertiary structure of a protein is defined by the path of the polypeptide chain 41
The stable folded state of a protein represents a state of low energy 41
Many proteins are formed of multiple subunits 42

Summary 43
Further Reading 44

Chapter 3 Dealing with Databases

3.1 The Structure of Databases 46
Flat-file databases store data as text files 48
Relational databases are widely used for storing biological information 49
XML has the flexibility to define bespoke data classifications 50
Many other database structures are used for biological data 51
Databases can be accessed locally or online and often link to each other 52

3.2 Types of Database 52
There's more to databases than just data 53
Primary and derived data 53
How we define and connect things is very important: Ontologies 54

3.3 Looking for Databases 55
Sequence databases 55
Microarray databases 58

Protein interaction databases 58
Structural databases 59

3.4 Data Quality 61
Nonredundancy is especially important for some applications of sequence databases 62
Automated methods can be used to check for data consistency 63
Initial analysis and annotation is usually automated 64
Human intervention is often required to produce the highest quality annotation 65
The importance of updating databases and entry identifier and version numbers 65

Summary 66
Further Reading 67

Part 2 Sequence Alignments

APPLICATIONS CHAPTER

Chapter 4 Producing and Analyzing Sequence Alignments

4.1 Principles of Sequence Alignment 72
Alignment is the task of locating equivalent regions of two or more sequences to maximize their similarity 73
Alignment can reveal homology between sequences 74
It is easier to detect homology when comparing protein sequences than when comparing nucleic acid sequences 75

4.2 Scoring Alignments 76
The quality of an alignment is measured by giving it a quantitative score 76
The simplest way of quantifying similarity between two sequences is percentage identity 76
The dot-plot gives a visual assessment of similarity based on identity 77
Genuine matches do not have to be identical 79
There is a minimum percentage identity that can be accepted as significant 81
There are many different ways of scoring an alignment 81

4.3 Substitution Matrices 81
Substitution matrices are used to assign individual scores to aligned sequence positions 81
The PAM substitution matrices use substitution frequencies derived from sets of closely related protein sequences 82
The BLOSUM substitution matrices use mutation data from highly conserved local regions of sequence 84
The choice of substitution matrix depends on the problem to be solved 84

4.4 Inserting Gaps 85
Gaps inserted in a sequence to maximize similarity require a scoring penalty 85
Dynamic programming algorithms can determine the optimal introduction of gaps 86

4.5 Types of Alignment 87
Different kinds of alignments are useful in different circumstances 87
Multiple sequence alignments enable the simultaneous comparison of a set of similar sequences 90
Multiple alignments can be constructed by several different techniques 90
Multiple alignments can improve the accuracy of alignment for sequences of low similarity 91
ClustalW can make global multiple alignments of both DNA and protein sequences 92
Multiple alignments can be made by combining a series of local alignments 92
Alignment can be improved by incorporating additional information 93

4.6 Searching Databases 93
Fast yet accurate search algorithms have been developed 94
FASTA is a fast database-search method based on matching short identical segments 95
BLAST is based on finding very similar short segments 95
Different versions of BLAST and FASTA are used for different problems 95
PSI-BLAST enables profile-based database searches 96
SSEARCH is a rigorous alignment method 97

4.7 Searching with Nucleic Acid or Protein Sequences 97
DNA or RNA sequences can be used either directly or after translation 97
The quality of a database match has to be tested to ensure that it could not have arisen by chance 97
Choosing an appropriate *E*-value threshold helps to limit a database search 98
Low-complexity regions can complicate homology searches 100
Different databases can be used to solve particular problems 102

4.8 Protein Sequence Motifs or Patterns 103
Creation of pattern databases requires expert knowledge 104
The BLOCKS database contains automatically compiled short blocks of conserved multiply aligned protein sequences 105

4.9 Searching Using Motifs and Patterns 107
The PROSITE database can be searched for protein motifs and patterns 107

The pattern-based program PHI-BLAST searches for both homology and matching motifs 108
Patterns can be generated from multiple sequences using PRATT 108
The PRINTS database consists of fingerprints representing sets of conserved motifs that describe a protein family 109
The Pfam database defines profiles of protein families 109

4.10 Patterns and Protein Function 109
Searches can be made for particular functional sites in proteins 109
Sequence comparison is not the only way of analyzing protein sequences 110

Summary 111
Further Reading 112

THEORY CHAPTER
Chapter 5 Pairwise Sequence Alignment and Database Searching

5.1 Substitution Matrices and Scoring 117
Alignment scores attempt to measure the likelihood of a common evolutionary ancestor 117
The PAM (MDM) substitution scoring matrices were designed to trace the evolutionary origins of proteins 119
The BLOSUM matrices were designed to find conserved regions of proteins 122
Scoring matrices for nucleotide sequence alignment can be derived in similar ways 125
The substitution scoring matrix used must be appropriate to the specific alignment problem 126
Gaps are scored in a much more heuristic way than substitutions 126

5.2 Dynamic Programming Algorithms 127
Optimal global alignments are produced using efficient variations of the Needleman–Wunsch algorithm 129
Local and suboptimal alignments can be produced by making small modifications to the dynamic programming algorithm 135
Time can be saved with a loss of rigor by not calculating the whole matrix 139

5.3 Indexing Techniques and Algorithmic Approximations 141
Suffix trees locate the positions of repeats and unique sequences 141
Hashing is an indexing technique that lists the starting positions of all k-tuples 143
The FASTA algorithm uses hashing and chaining for fast database searching 144
The BLAST algorithm makes use of finite-state automata 147
Comparing a nucleotide sequence directly with a protein sequence requires special modifications to the BLAST and FASTA algorithms 150

5.4 Alignment Score Significance 153
The statistics of gapped local alignments can be approximated by the same theory 156

5.5 Aligning Complete Genome Sequences 156
Indexing and scanning whole genome sequences efficiently is crucial for the sequence alignment of higher organisms 157
The complex evolutionary relationships between the genomes of even closely related organisms require novel alignment algorithms 159

Summary 159
Further Reading 161

THEORY CHAPTER
Chapter 6 Patterns, Profiles, and Multiple Alignments

6.1 Profiles and Sequence Logos 167
Position-specific scoring matrices are an extension of substitution scoring matrices 168
Methods for overcoming a lack of data in deriving the values for a PSSM 171
PSI-BLAST is a sequence database searching program 176
Representing a profile as a logo 177

6.2 Profile Hidden Markov Models 179
The basic structure of HMMs used in sequence alignment to profiles 180
Estimating HMM parameters using aligned sequences 185
Scoring a sequence against a profile HMM: The most probable path and the sum over all paths 187
Estimating HMM parameters using unaligned sequences 190

6.3 Aligning Profiles 193
Comparing two PSSMs by alignment 193
Aligning profile HMMs 195

6.4 Multiple Sequence Alignments by Gradual Sequence Addition 196
The order in which sequences are added is chosen based on the estimated likelihood of incorporating errors in the alignment 198
Many different scoring schemes have been used in constructing multiple alignments 200

The multiple alignment is built using the guide tree and profile methods and may be further refined 204

6.5 Other Ways of Obtaining Multiple Alignments 207
The multiple sequence alignment program DIALIGN aligns ungapped blocks 207
The SAGA method of multiple alignment uses a genetic algorithm 209

6.6 Sequence Pattern Discovery 211
Discovering patterns in a multiple alignment: eMOTIF and AACC 213
Probabilistic searching for common patterns in sequences: Gibbs and MEME 215
Searching for more general sequence patterns 217

Summary 218
Further Reading 219

Part 3 Evolutionary Processes

APPLICATIONS CHAPTER
Chapter 7 Recovering Evolutionary History

7.1 The Structure and Interpretation of Phylogenetic Trees 225
Phylogenetic trees reconstruct evolutionary relationships 225
Tree topology can be described in several ways 230
Consensus and condensed trees report the results of comparing tree topologies 232

7.2 Molecular Evolution and its Consequences 235
Most related sequences have many positions that have mutated several times 236
The rate of accepted mutation is usually not the same for all types of base substitution 236
Different codon positions have different mutation rates 238
Only orthologous genes should be used to construct species phylogenetic trees 239
Major changes affecting large regions of the genome are surprisingly common 247

7.3 Phylogenetic Tree Reconstruction 248
Small ribosomal subunit rRNA sequences are well suited to reconstructing the evolution of species 249
The choice of the method for tree reconstruction depends to some extent on the size and quality of the dataset 249
A model of evolution must be chosen to use with the method 251
All phylogenetic analyses must start with an accurate multiple alignment 255
Phylogenetic analyses of a small dataset of 16S RNA sequence data 255
Building a gene tree for a family of enzymes can help to identify how enzymatic functions evolved 259

Summary 264
Further Reading 265

THEORY CHAPTER
Chapter 8 Building Phylogenetic Trees

8.1 Evolutionary Models and the Calculation of Evolutionary Distance 268
A simple but inaccurate measure of evolutionary distance is the p-distance 268
The Poisson distance correction takes account of multiple mutations at the same site 270
The Gamma distance correction takes account of mutation rate variation at different sequence positions 270
The Jukes–Cantor model reproduces some basic features of the evolution of nucleotide sequences 271
More complex models distinguish between the relative frequencies of different types of mutation 272
There is a nucleotide bias in DNA sequences 275
Models of protein-sequence evolution are closely related to the substitution matrices used for sequence alignment 276

8.2 Generating Single Phylogenetic Trees 276
Clustering methods produce a phylogenetic tree based on evolutionary distances 276
The UPGMA method assumes a constant molecular clock and produces an ultrametric tree 278
The Fitch–Margoliash method produces an unrooted additive tree 279
The neighbor-joining method is related to the concept of minimum evolution 282
Stepwise addition and star-decomposition methods are usually used to generate starting trees for further exploration, not the final tree 285

8.3 Generating Multiple Tree Topologies 286
The branch-and-bound method greatly improves the efficiency of exploring tree topology 288
Optimization of tree topology can be achieved by making a series of small changes to an existing tree 288
Finding the root gives a phylogenetic tree a direction in time 291

8.4 Evaluating Tree Topologies 293
Functions based on evolutionary distances can be used to evaluate trees 293
Unweighted parsimony methods look for the trees with the smallest number of mutations 297

Mutations can be weighted in different ways in the parsimony method 300
Trees can be evaluated using the maximum likelihood method 302
The quartet-puzzling method also involves maximum likelihood in the standard implementation 305
Bayesian methods can also be used to reconstruct phylogenetic trees 306

8.5 Assessing the Reliability of Tree Features and Comparing Trees 307
The long-branch attraction problem can arise even with perfect data and methodology 308
Tree topology can be tested by examining the interior branches 309
Tests have been proposed for comparing two or more alternative trees 310

Summary 311
Further Reading 312

Part 4 Genome Characteristics

APPLICATIONS CHAPTER
Chapter 9 Revealing Genome Features

9.1 Preliminary Examination of Genome Sequence 318
Whole genome sequences can be split up to simplify gene searches 319
Structural RNA genes and repeat sequences can be excluded from further analysis 319
Homology can be used to identify genes in both prokaryotic and eukaryotic genomes 322

9.2 Gene Prediction in Prokaryotic Genomes 322

9.3 Gene Prediction in Eukaryotic Genomes 323
Programs for predicting exons and introns use a variety of approaches 323
Gene predictions must preserve the correct reading frame 324
Some programs search for exons using only the query sequence and a model for exons 327
Some programs search for genes using only the query sequence and a gene model 332
Genes can be predicted using a gene model and sequence similarity 334
Genomes of related organisms can be used to improve gene prediction 336

9.4 Splice Site Detection 337
Splice sites can be detected independently by specialized programs 338

9.5 Prediction of Promoter Regions 338
Prokaryotic promoter regions contain relatively well-defined motifs 339
Eukaryotic promoter regions are typically more complex than prokaryotic promoters 340
A variety of promoter-prediction methods are available online 340
Promoter prediction results are not very clear-cut 341

9.6 Confirming Predictions 342
There are various methods for calculating the accuracy of gene-prediction programs 342
Translating predicted exons can confirm the correctness of the prediction 343
Constructing the protein and identifying homologs 343

9.7 Genome Annotation 346
Genome annotation is the final step in genome analysis 347
Gene ontology provides a standard vocabulary for gene annotation 348

9.8 Large Genome Comparisons 353

Summary 354
Further Reading 355

THEORY CHAPTER
Chapter 10 Gene Detection and Genome Annotation

10.1 Detection of Functional RNA Molecules Using Decision Trees 361
Detection of tRNA genes using the tRNAscan algorithm 361
Detection of tRNA genes in eukaryotic genomes 362

10.2 Features Useful for Gene Detection in Prokaryotes 364

10.3 Algorithms for Gene Detection in Prokaryotes 368
GeneMark uses inhomogeneous Markov chains and dicodon statistics 368
GLIMMER uses interpolated Markov models of coding potential 371
ORPHEUS uses homology, codon statistics, and ribosome-binding sites 372
GeneMark.hmm uses explicit state duration hidden Markov models 373
EcoParse is an HMM gene model 376

10.4 Features Used in Eukaryotic Gene Detection 377
Differences between prokaryotic and eukaryotic genes 377
Introns, exons, and splice sites 379
Promoter sequences and binding sites for transcription factors 381

10.5 Predicting Eukaryotic Gene Signals 381
Detection of core promoter binding signals is a key element of some eukaryotic gene-prediction methods 381
A set of models has been designed to locate the site of core promoter sequence signals 383
Predicting promoter regions from general sequence properties can reduce the numbers of false-positive results 387
Predicting eukaryotic transcription and translation start sites 389
Translation and transcription stop signals complete the gene definition 389

10.6 Predicting Exon/Intron Structure 389
Exons can be identified using general sequence properties 390
Splice-site prediction 392
Splice sites can be predicted by sequence patterns combined with base statistics 393
GenScan uses a combination of weight matrices and decision trees to locate splice sites 394
GeneSplicer predicts splice sites using first-order Markov chains 394
NetPlantGene uses neural networks with intron and exon predictions to predict splice sites 395
Other splicing features may yet be exploited for splice-site prediction 396
Specific methods exist to identify initial and terminal exons 396
Exons can be defined by searching databases for homologous regions 397

10.7 Complete Eukaryotic Gene Models 397

10.8 Beyond the Prediction of Individual Genes 399
Functional annotation 400
Comparison of related genomes can help resolve uncertain predictions 403
Evaluation and reevaluation of gene-detection methods 405

Summary 405
Further Reading 406

Part 5 Secondary Structures

APPLICATIONS CHAPTER

Chapter 11 Obtaining Secondary Structure from Sequence

11.1 Types of Prediction Methods 413
Statistical methods are based on rules that give the probability that a residue will form part of a particular secondary structure 414
Nearest-neighbor methods are statistical methods that incorporate additional information about protein structure 414
Machine-learning approaches to secondary structure prediction mainly make use of neural networks and HMM methods 415

11.2 Training and Test Databases 416
There are several ways to define protein secondary structures 417

11.3 Assessing the Accuracy of Prediction Programs 417
Q_3 measures the accuracy of individual residue assignments 417
Secondary structure predictions should not be expected to reach 100% residue accuracy 418
The Sov value measures the prediction accuracy for whole elements 419
CAFASP/CASP: Unbiased and readily available protein prediction assessments 419

11.4 Statistical and Knowledge-Based Methods 421
The GOR method uses an information theory approach 422
The program Zpred includes multiple alignment of homologous sequences and residue conservation information 425
There is an overall increase in prediction accuracy using multiple sequence information 426
The nearest-neighbor method: The use of multiple nonhomologous sequences 428
PREDATOR is a combined statistical and knowledge-based program that includes the nearest-neighbor approach 428

11.5 Neural Network Methods of Secondary Structure Prediction 430
Assessing the reliability of neural net predictions 432
Several examples of Web-based neural network secondary structure prediction programs 432
PROF: Protein forecasting 434
PSIPRED 434
Jnet: Using several alternative representations of the sequence alignment 434

11.6 Some Secondary Structures Require Specialized Prediction Methods 435
Transmembrane proteins 436
Quantifying the preference for a membrane environment 437

11.7 Prediction of Transmembrane Protein Structure 438
Multi-helix membrane proteins 439
A selection of prediction programs to predict transmembrane helices 441

Statistical methods 443
Knowledge-based prediction 443
Evolutionary information from protein families improves the prediction 444
Neural nets in transmembrane prediction 445
Predicting transmembrane helices with hidden Markov models 446
Comparing the results: What to choose 447
What happens if a non-transmembrane protein is submitted to transmembrane prediction programs 448
Prediction of transmembrane structure containing β-strands 448

11.8 Coiled-coil Structures 451
The COILS prediction program 452
PAIRCOIL and MULTICOIL are an extension of the COILS algorithm 453
Zipping the Leucine zipper: A specialized coiled coil 453

11.9 RNA Secondary Structure Prediction 455

Summary 458
Further Reading 459

THEORY CHAPTER
Chapter 12 Predicting Secondary Structures

12.1 Defining Secondary Structure and Prediction Accuracy 463
The definitions used for automatic protein secondary structure assignment do not give identical results 464
There are several different measures of the accuracy of secondary structure prediction 469

12.2 Secondary Structure Prediction Based on Residue Propensities 472
Each structural state has an amino acid preference which can be assigned as a residue propensity 473
The simplest prediction methods are based on the average residue propensity over a sequence window 476
Residue propensities are modulated by nearby sequence 479
Predictions can be significantly improved by including information from homologous sequences 484

12.3 The Nearest-Neighbor Methods are Based on Sequence Segment Similarity 485
Short segments of similar sequence are found to have similar structure 487
Several sequence similarity measures have been used to identify nearest-neighbor segments 488
A weighted average of the nearest-neighbor segment structures is used to make the prediction 490
A nearest-neighbor method has been developed to predict regions with a high potential to misfold 491

12.4 Neural Networks Have Been Employed Successfully for Secondary Structure Prediction 492
Layered feed-forward neural networks can transform a sequence into a structural prediction 494
Inclusion of information on homologous sequences improves neural network accuracy 502
More complex neural nets have been applied to predict secondary and other structural features 503

12.5 Hidden Markov Models Have Been Applied to Structure Prediction 504
HMM methods have been found especially effective for transmembrane proteins 506
Nonmembrane protein secondary structures can also be successfully predicted with HMMs 509

12.6 General Data Classification Techniques Can Predict Structural Features 510
Support vector machines have been successfully used for protein structure prediction 511
Discriminants, SOMs, and other methods have also been used 512

Summary 514
Further Reading 515

Part 6 Tertiary Structures

APPLICATIONS CHAPTER
Chapter 13 Modeling Protein Structure

13.1 Potential Energy Functions and Force Fields 524
The conformation of a protein can be visualized in terms of a potential energy surface 525
Conformational energies can be described by simple mathematical functions 525
Similar force fields can be used to represent conformational energies in the presence of averaged environments 526
Potential energy functions can be used to assess a modeled structure 527
Energy minimization can be used to refine a modeled structure and identify local energy minima 527
Molecular dynamics and simulated annealing are used to find global energy minima 528

13.2 Obtaining a Structure by Threading 529
The prediction of protein folds in the absence of known structural homologs 531
Libraries or databases of nonredundant protein folds are used in threading 531
Two distinct types of scoring schemes have been used in threading methods 531
Dynamic programming methods can identify optimal alignments of target sequences and structural folds 533

Several methods are available to assess the confidence to be put on the fold prediction 534
The C2-like domain from the Dictyostelia: A practical example of threading 535

13.3 Principles of Homology Modeling 537
Closely related target and template sequences give better models 539
Significant sequence identity depends on the length of the sequence 540
Homology modeling has been automated to deal with the numbers of sequences that can now be modeled 541
Model building is based on a number of assumptions 541

13.4 Steps in Homology Modeling 542
Structural homologs to the target protein are found in the PDB 543
Accurate alignment of target and template sequences is essential for successful modeling 543
The structurally conserved regions of a protein are modeled first 544
The modeled core is checked for misfits before proceeding to the next stage 545
Sequence realignment and remodeling may improve the structure 545
Insertions and deletions are usually modeled as loops 545
Nonidentical amino acid side chains are modeled mainly by using rotamer libraries 547
Energy minimization is used to relieve structural errors 548
Molecular dynamics can be used to explore possible conformations for mobile loops 548
Models need to be checked for accuracy 549
How far can homology models be trusted? 551

13.5 Automated Homology Modeling 552
The program MODELLER models by satisfying protein structure constraints 553
COMPOSER uses fragment-based modeling to automatically generate a model 553
Automated methods available on the Web for comparative modeling 554
Assessment of structure prediction 554

13.6 Homology Modeling of PI3 Kinase p110α 557
Swiss-Pdb Viewer can be used for manual or semi-manual modeling 557
Alignment, core modeling, and side-chain modeling are carried out all in one 558
The loops are modeled from a database of possible structures 559
Energy minimization and quality inspection can be carried out within Swiss-Pdb Viewer 559
MolIDE is a downloadable semi-automatic modeling package 560
Automated modeling on the Web illustrated with p110α kinase 561
Modeling a functionally related but sequentially dissimilar protein: mTOR 563
Generating a multidomain three-dimensional structure from sequence 564

Summary 564
Further Reading 565

APPLICATIONS CHAPTER
Chapter 14 Analyzing Structure-Function Relationships

14.1 Functional Conservation 568
Functional regions are usually structurally conserved 569
Similar biochemical function can be found in proteins with different folds 570
Fold libraries identify structurally similar proteins regardless of function 571

14.2 Structure Comparison Methods 574
Finding domains in proteins aids structure comparison 574
Structural comparisons can reveal conserved functional elements not discernible from a sequence comparison 576
The CE method builds up a structural alignment from pairs of aligned protein segments 576
The Vector Alignment Search Tool (VAST) aligns secondary structural elements 577
DALI identifies structure superposition without maintaining segment order 578
FATCAT introduces rotations between rigid segments 579

14.3 Finding Binding Sites 580
Highly conserved, strongly charged, or hydrophobic surface areas may indicate interaction sites 582
Searching for protein–protein interactions using surface properties 584
Surface calculations highlight clefts or holes in a protein that may serve as binding sites 585
Looking at residue conservation can identify binding sites 586

14.4 Docking Methods and Programs 587
Simple docking procedures can be used when the structure of a homologous protein bound to a ligand analog is known 588
Specialized docking programs will automatically dock a ligand to a structure 588

Scoring functions are used to identify the most likely docked ligand 590
The DOCK program is a semirigid-body method that analyzes shape and chemical complementarity of ligand and binding site 590
Fragment docking identifies potential substrates by predicting types of atoms and functional groups in the binding area 591
GOLD is a flexible docking program, which utilizes a genetic algorithm 591
The water molecules in binding sites should also be considered 592

Summary 593
Further Reading 594

Part 7 Cells and Organisms

Chapter 15 Proteome and Gene Expression Analysis

15.1 Analysis of Large-scale Gene Expression 601
The expression of large numbers of different genes can be measured simultaneously by DNA microarrays 602
Gene expression microarrays are mainly used to detect differences in gene expression in different conditions 602
Serial analysis of gene expression (SAGE) is also used to study global patterns of gene expression 604
Digital differential display uses bioinformatics and statistics to detect differential gene expression in different tissues 605
Facilitating the integration of data from different places and experiments 606
The simplest method of analyzing gene expression microarray data is hierarchical cluster analysis 606
Techniques based on self-organizing maps can be used for analyzing microarray data 608
Self-organizing tree algorithms (SOTAs) cluster from the top down by successive subdivision of clusters 610
Clustered gene expression data can be used as a tool for further research 610

15.2 Analysis of Large-scale Protein Expression 612
Two-dimensional gel electrophoresis is a method for separating the individual proteins in a cell 613
Measuring the expression levels shown in 2D gels 614
Differences in protein expression levels between different samples can be detected by 2D gels 615
Clustering methods are used to identify protein spots with similar expression patterns 615
Principal component analysis (PCA) is an alternative to clustering for analyzing microarray and 2D gel data 618
The changes in a set of protein spots can be tracked over a number of different samples 618
Databases and online tools are available to aid the interpretation of 2D gel data 620
Protein microarrays allow the simultaneous detection of the presence or activity of large numbers of different proteins 621
Mass spectrometry can be used to identify the proteins separated and purified by 2D gel electrophoresis or other means 621
Protein-identification programs for mass spectrometry are freely available on the Web 622
Mass spectrometry can be used to measure protein concentration 623

Summary 623
Further Reading 624

Chapter 16 Clustering Methods and Statistics

16.1 Expression Data Require Preparation Prior to Analysis 626
Data normalization is designed to remove systematic experimental errors 627
Expression levels are often analyzed as ratios and are usually transformed by taking logarithms 628
Sometimes further normalization is useful after the data transformation 630
Principal component analysis is a method for combining the properties of an object 631

16.2 Cluster Analysis Requires Distances to be Defined Between all Data Points 633
Euclidean distance is the measure used in everyday life 634
The Pearson correlation coefficient measures distance in terms of the shape of the expression response 635
The Mahalanobis distance takes account of the variation and correlation of expression responses 636

16.3 Clustering Methods Identify Similar and Distinct Expression Patterns 637
Hierarchical clustering produces a related set of alternative partitions of the data 639
k-means clustering groups data into several clusters but does not determine a relationship between clusters 641
Self-organizing maps (SOMs) use neural network methods to cluster data into a predetermined number of clusters 644
Evolutionary clustering algorithms use selection, recombination, and mutation to find the best possible solution to a problem 646

The self-organizing tree algorithm (SOTA) determines the number of clusters required 648
Biclustering identifies a subset of similar expression level patterns occurring in a subset of the samples 649
The validity of clusters is determined by independent methods 650

16.4 Statistical Analysis can Quantify the Significance of Observed Differential Expression 651
t-tests can be used to estimate the significance of the difference between two expression levels 654
Nonparametric tests are used to avoid making assumptions about the data sampling 656
Multiple testing of differential expression requires special techniques to control error rates 657

16.5 Gene and Protein Expression Data Can be Used to Classify Samples 659
Many alternative methods have been proposed that can classify samples 660
Support vector machines are another form of supervised learning algorithms that can produce classifiers 661

Summary 662
Further Reading 664

Chapter 17 Systems Biology

17.1 What is a System? 669
A system is more than the sum of its parts 669
A biological system is a living network 670
Databases are useful starting points in constructing a network 671
To construct a model more information is needed than a network 672
There are three possible approaches to constructing a model 674
Kinetic models are not the only way in systems biology 678

17.2 Structure of the Model 679
Control circuits are an essential part of any biological system 680
The interactions in networks can be represented as simple differential equations 680

17.3 Robustness of Biological Systems 683
Robustness is a distinct feature of complexity in biology 684
Modularity plays an important part in robustness 685
Redundancy in the system can provide robustness 686
Living systems can switch from one state to another by means of bistable switches 688

17.4 Storing and Running System Models 689
Specialized programs make simulating systems easier 691
Standardized system descriptions aid their storage and reuse 692

Summary 692
Further Reading 693

APPENDICES Background Theory

Appendix A: Probability, Information, and Bayesian Analysis

Probability Theory, Entropy, and Information 695
Mutually exclusive events 695
Occurrence of two events 696
Occurrence of two random variables 696
Bayesian Analysis 697
Bayes' theorem 697
Inference of parameter values 698

Further Reading 699

Appendix B: Molecular Energy Functions

Force Fields for Calculating Intra- and Intermolecular Interaction Energies 701
Bonding terms 702
Nonbonding terms 704
Potentials used in Threading 706
Potentials of mean force 706
Potential terms relating to solvent effects 707

Further Reading 708

Appendix C: Function Optimization

Full Search Methods 710
Dynamic programming and branch-and-bound 710
Local Optimization 710
The downhill simplex method 711
The steepest descent method 711
The conjugate gradient method 714
Methods using second derivatives 714
Thermodynamic Simulation and Global Optimization 715
Monte Carlo and genetic algorithms 716
Molecular dynamics 718
Simulated annealing 719
Summary 719

Further Reading 719

List of Symbols 721
Glossary 734
Index 751