

PREFACE

Over the past decade, computer scientists have increasingly been enlisted as “bioinformaticians” to assist molecular biologists in their research. This book is conceived as a practical introduction to bioinformatics for these computer scientists. While it is not possible to be exhaustive in coverage, the chapters are in-depth discussions by expert bioinformaticians on both general techniques and specific approaches to a range of selected bioinformatics problems. Let us provide a brief overview of these chapters here.

A practical bioinformatician must learn to speak the language of biology. We thus begin with Chapter 1 by Ng and Chapter 2 by Schönbach that form an overview of modern molecular biology and of the planning and execution of bioinformatics experiments. At the same time, a practical bioinformatician must also be conversant in a broad spectrum of topics in computer science—data mining, machine learning, mathematical modeling, sequence alignment, data integration, database management, workflow development, and so on. These diversified topics are surveyed in three separate chapters, *viz.* Chapter 3 by Li *et al.* which provides an in-depth review of data mining techniques that are amongst the key computing technologies for the analysis of biological data; Chapter 10 by Brown *et al.* which discusses the advances through the past thirty years in both global and local alignment, and present methods for general purpose homology that are widely adopted; and Chapter 17 by Wong which reviews some of the requirements and technologies relevant to data integration and warehousing.

DNA sequences contain a number of functional and regulatory sites, such as the site where the transcription of a gene begins, the site where the translation of a gene begins, the site where an exon of a gene ends and an intron begins, and so on. The next several chapters of the book deal with the computational recognition of these sites, *viz.* Chapter 4 by Li *et al.* which is an in-depth survey spanning two decades of research of methods for computational recognition of translation initiation sites from mRNA, cDNA, and DNA sequences; and Chapters 5–7 by Bajić *et al.* which discuss general frameworks, conceptual issues, and performance tuning related to the use of statistical and neural network modeling

for computational recognition of promoters and regulatory sites. The recognition of translation initiation sites is among the simplest problems in the recognition of functional sites from DNA sequences. On the other hand, among the toughest problems in the recognition of functional sites from DNA sequences is the determination of locations of promoters and related regulatory elements and functional sites. Thus we hope these four chapters together can bring out clearly the whole range of approaches to this group of problems.

We next move to analysis of RNA sequences. We consider the problem of predicting the secondary structure of RNAs, which is relevant to applications such as function classification, evolution study, and pseudogene detection. Chapter 8 by Sung provides a detailed review of computational methods for predicting secondary structure from RNA sequences. Unlike the prediction methods introduced in earlier chapters for recognition of functional sites from DNA sequences, which are mostly data mining and machine learning methods, the methods described in this chapter come from the realm of mathematical modeling.

After looking at RNA sequences, we move on to protein sequences, and look at aspects relating to protein function prediction. For example, each compartment in a cell has a unique set of functions, and it is thus reasonable to assume that the compartment or membrane in which a protein resides is a determinant of its function. So we have Chapter 9 by Horton *et al.* to discuss various aspects of protein subcellular localization in the context of bioinformatics and review the twenty years of progress in predicting protein subcellular localization. As another example, the homology relationship between a protein and another protein is also suggestive of the function of that protein. So we have Chapter 12 by Kaplan *et al.* to describe two bioinformatics tools, ProtoNet and PANDORA. ProtoNet uses an approach of protein sequence hierarchical clustering to detect remote protein relatives. PANDORA uses a graph-based method to interpret complex protein groups through their annotations. As a third example, motifs are commonly used to classify protein sequences and to provide functional clues on binding sites, catalytic sites, and active sites, or structure/functions relations. So we have Chapter 16 by Schönbach and Matsuda to present a case study of a workflow for mining new motifs from the FANTOM1 mouse cDNA clone collection by a linkage-clustering method, with an all-to-all sequence comparison, followed by visual inspection, sequence, topological, and literature analysis of the motif candidates.

Next we sample the fascinating topic of phylogenetics—the study of the origin, development, and death of a taxon—based on sequence and other information. Chapter 11 by Meng is an introduction to phylogenetics using a case study on Saururaceae. In contrast to earlier chapters, which emphasize the computational aspect, this chapter is written from the perspective of a plant molecular biologist,

and emphasizes instead the care that must be exercised in the use of computational tools and the analysis that must be performed on the results produced by computational tools.

The genomics and proteomics efforts have helped identify many new genes and proteins in living organisms. However, simply knowing the existence of genes and proteins does not tell us much about the biological processes in which they participate. Many major biological processes are controlled by protein interaction networks and gene regulation networks. Thus we have Chapter 13 by Tan and Ng to give an overview of the various current methods for discovering protein-protein interactions experimentally and computationally.

The development of microarray technology in the last decade has made possible the simultaneous monitoring of the expression of thousands of genes. This development offers great opportunities in advancing the diagnosis of diseases, the treatment of diseases, and the understanding of gene functions. Chapter 14 by Li and Wong is an in-depth survey of several approaches to some of the gene expression analysis challenges that accompany these opportunities. On the other hand, Chapter 15 by Lin *et al.* presents a method for selecting probes in the design of a microarray to profile genome-wide gene expression of a given genome.

Biological data is being created at ever-increasing rates as different high-throughput technologies are implemented for a wide variety of discovery platforms. It is crucial for researchers to be able to not only access this information but also to integrate it well and synthesize new holistic ideas about various topics. So it is appropriate that we devote the remaining chapters of this book to the issues of integrating databases, cleansing databases, and large-scale experimental and computational analysis workflows as follows. Chapter 18 by Kolatkar and Lin demonstrates the construction of a purpose-built integrated database PPDB using the powerful general data integration engine Kleisli. Chapter 19 by Wu and Barker presents the classification-driven rule-based approach in PIR database to the functional annotation of proteins. Wu and Barker also provide two case studies: the first looks at error propagation to secondary databases using the example of IMP Dehydrogenase; the second looks at transitive identification error using the example of His-I bifunctional proteins. Finally, Chapter 20 by Scheetz and Casavant describes the sophisticated informatics tools and workflow underlying the large-scale effort in EST-based gene discovery in Rat, Human, Mouse, and other species being conducted at the University of Iowa.

Limsoon Wong
11 December 2003

CONTENTS

| | |
|--|-----------|
| Dedication | v |
| Preface | vii |
| Chapter 1. Molecular Biology for the Practical Bioinformatician | 1 |
| 1 Introduction | 1 |
| 2 Our Molecular Selves | 3 |
| 2.1 Cells, DNAs, and Chromosomes | 4 |
| 2.2 Genes and Genetic Variations | 7 |
| 2.2.1 Mutations and Genetic Diseases | 9 |
| 3 Our Biological Machineries | 10 |
| 3.1 From Gene to Protein: The Central Dogma | 12 |
| 3.2 The Genetic Code | 14 |
| 3.3 Gene Regulation and Expression | 15 |
| 4 Tools of the Trade | 18 |
| 4.1 Basic Operations | 18 |
| 4.1.1 Cutting DNA | 18 |
| 4.1.2 Copying DNA | 19 |
| 4.1.3 Separating DNA | 22 |
| 4.1.4 Matching DNA | 23 |
| 4.2 Putting It Together | 24 |
| 4.2.1 Genome Sequencing | 24 |
| 4.2.2 Gene Expression Profiling | 26 |
| 5 Conclusion | 28 |
| Chapter 2. Strategy and Planning of Bioinformatics Experiments | 31 |
| 1 Multi-Dimensional Bioinformatics | 31 |
| 2 Reasoning and Strategy | 33 |
| 3 Planning | 34 |

| | |
|--|-----------|
| Chapter 3. Data Mining Techniques for the Practical Bioinformatician . | 35 |
| 1 Feature Selection Methods | 36 |
| 1.1 Curse of Dimensionality | 36 |
| 1.2 Signal-to-Noise Measure | 38 |
| 1.3 T-Test Statistical Measure | 40 |
| 1.4 Fisher Criterion Score | 40 |
| 1.5 Entropy Measure | 41 |
| 1.6 χ^2 Measure | 43 |
| 1.7 Information Gain | 44 |
| 1.8 Information Gain Ratio | 44 |
| 1.9 Wilcoxon Rank Sum Test | 45 |
| 1.10 Correlation-Based Feature Selection | 46 |
| 1.11 Principal Component Analysis | 47 |
| 1.12 Use of Feature Selection in Bioinformatics | 49 |
| 2 Classification Methods | 50 |
| 2.1 Decision Trees | 50 |
| 2.2 Bayesian Methods | 51 |
| 2.3 Hidden Markov Models | 52 |
| 2.4 Artificial Neural Networks | 52 |
| 2.5 Support Vector Machines | 55 |
| 2.6 Prediction by Collective Likelihood of Emerging Patterns | 57 |
| 3 Association Rules Mining Algorithms | 59 |
| 3.1 Association Rules | 59 |
| 3.2 The Apriori Algorithm | 60 |
| 3.3 The Max-Miner Algorithm | 61 |
| 3.4 Use of Association Rules in Bioinformatics | 62 |
| 4 Clustering Methods | 64 |
| 4.1 Factors Affecting Clustering | 64 |
| 4.2 Partitioning Methods | 66 |
| 4.3 Hierarchical Methods | 66 |
| 4.4 Use of Clustering in Bioinformatics | 68 |
| 5 Remarks | 69 |
| Chapter 4. Techniques for Recognition of Translation Initiation Sites . . | 71 |
| 1 Translation Initiation Sites | 72 |
| 2 Data Set and Evaluation | 73 |
| 3 Recognition by Perceptrons | 74 |
| 4 Recognition by Artificial Neural Networks | 75 |
| 5 Recognition by Engineering of Support Vector Machine Kernels | 76 |

| | | |
|-----|--|----|
| 6 | Recognition by Feature Generation, Feature Selection, and Feature Integration | 77 |
| 6.1 | Feature Generation | 78 |
| 6.2 | Feature Selection | 80 |
| 6.3 | Feature Integration for Decision Making | 82 |
| 7 | Improved Recognition by Feature Generation, Feature Selection, and Feature Integration | 82 |
| 8 | Recognition by Linear Discriminant Function | 84 |
| 9 | Recognition by Ribosome Scanning | 86 |
| 10 | Remarks | 88 |

| | | |
|--|--|-----------|
| Chapter 5. How Neural Networks Find Promoters Using Recognition of Micro-Structural Promoter Components | | 91 |
| 1 | Motivation and Background | 92 |
| 1.1 | Problem Framework | 93 |
| 1.2 | Promoter Recognition | 94 |
| 1.3 | ANN-Based Promoter Recognition | 95 |
| 2 | Characteristic Motifs of Eukaryotic Promoters | 96 |
| 3 | Motif-Based Search for Promoters | 97 |
| 3.1 | Evaluation Study by Fickett and Hatzigeorgiou | 100 |
| 3.2 | Enhancers May Contribute to False Recognition | 100 |
| 4 | ANNs and Promoter Components | 101 |
| 4.1 | Description of Promoter Recognition Problem | 101 |
| 4.2 | Representation of Nucleotides for Network Processing | 103 |
| 5 | Structural Decomposition | 104 |
| 5.1 | Parallel Composition of Feature Detectors | 104 |
| 5.2 | First- and Second-Level ANNs | 109 |
| 5.3 | Cascade Composition of Feature Detectors | 112 |
| 5.4 | Structures Based on Multilayer Perceptrons | 112 |
| 6 | Time-Delay Neural Networks | 114 |
| 6.1 | Multistate TDNN | 118 |
| 6.2 | Pruning ANN Connections | 118 |
| 7 | Comments on Performance of ANN-Based Programs for Eukaryotic Promoter Prediction | 119 |

| | | |
|---|--|------------|
| Chapter 6. Neural-Statistical Model of TATA-Box Motifs in Eukaryotes | | 123 |
| 1 | Promoter Recognition via Recognition of TATA-Box | 124 |
| 2 | Position Weight Matrix and Statistical Analysis of the TATA-Box and Its Neighborhood | 126 |

| | | |
|---|--|-----|
| 2.1 | TATA Motifs as One of the Targets in the Search for Eukaryotic Promoters | 126 |
| 2.2 | Data Sources and Data Sets | 127 |
| 2.3 | Recognition Quality | 128 |
| 2.4 | Statistical Analysis of TATA Motifs | 128 |
| 2.4.1 | PWM | 129 |
| 2.4.2 | Numerical Characterization of Segments Around TATA-Box | 130 |
| 2.4.3 | Position Analysis of TATA Motifs | 131 |
| 2.4.4 | Characteristics of Segments S_1 , S_2 , and S_3 | 132 |
| 2.5 | Concluding Remarks | 134 |
| 3 | LVQ ANN for TATA-Box Recognition | 135 |
| 3.1 | Data Preprocessing: Phase 1 | 135 |
| 3.2 | Data Preprocessing: Phase 2 — Principal Component Analysis | 138 |
| 3.3 | Learning Vector Quantization ANN | 140 |
| 3.4 | The Structure of an LVQ Classifier | 141 |
| 3.5 | LVQ ANN Training | 142 |
| 3.6 | Initial Values for Network Parameters | 143 |
| 3.6.1 | Genetic Algorithm | 144 |
| 3.6.2 | Searching for Good Initial Weights | 144 |
| 4 | Final Model of TATA-Box Motif | 146 |
| 4.1 | Structure of Final System for TATA-Box Recognition | 146 |
| 4.2 | Training of the ANN Part of the Model | 146 |
| 4.3 | Performance of Complete System | 149 |
| 4.3.1 | Comparison of MLVQ and System with Single LVQ ANN | 151 |
| 4.4 | The Final Test | 152 |
| 5 | Summary | 155 |
| Chapter 7. Tuning the Dragon Promoter Finder System for Human Promoter Recognition 157 | | |
| 1 | Promoter Recognition | 158 |
| 2 | Dragon Promoter Finder | 159 |
| 3 | Model | 161 |
| 4 | Tuning Data | 161 |
| 4.1 | Promoter Data | 161 |
| 4.2 | Non-Promoter Data | 162 |

| | | |
|--|--|------------|
| 5 | Tuning Process | 162 |
| 6 | Discussions and Conclusions | 164 |
| Chapter 8. RNA Secondary Structure Prediction | | 167 |
| 1 | Introduction to RNA Secondary Structures | 168 |
| 2 | RNA Secondary Structure Determination Experiments | 171 |
| 3 | RNA Structure Prediction Based on Sequence | 172 |
| 4 | Structure Prediction in the Absence of Pseudoknot | 173 |
| 4.1 | Loop Energy | 174 |
| 4.2 | First RNA Secondary Structure Prediction Algorithm | 175 |
| 4.2.1 | $W(j)$ | 175 |
| 4.2.2 | $V(i, j)$ | 176 |
| 4.2.3 | $VBI(i, j)$ | 176 |
| 4.2.4 | $VM(i, j)$ | 176 |
| 4.2.5 | Time Analysis | 177 |
| 4.3 | Speeding up Multi-Loops | 178 |
| 4.3.1 | Assumption on Free Energy of Multi-Loop | 178 |
| 4.3.2 | Modified Algorithm for Speeding Up Multi-Loop Computation | 178 |
| 4.3.3 | Time Complexity | 179 |
| 4.4 | Speeding Up Internal Loops | 179 |
| 4.4.1 | Assumption on Free Energy for Internal Loop | 179 |
| 4.4.2 | Detailed Description | 180 |
| 4.4.3 | Time Analysis | 181 |
| 5 | Structure Prediction in the Presence of Pseudoknots | 181 |
| 6 | Akutsu's Algorithm | 182 |
| 6.1 | Definition of Simple Pseudoknot | 182 |
| 6.2 | RNA Secondary Structure Prediction with Simple Pseudoknots | 183 |
| 6.2.1 | $V_L(i, j, k), V_R(i, j, k), V_M(i, j, k)$ | 185 |
| 6.2.2 | To Compute Basis | 186 |
| 6.2.3 | Time Analysis | 186 |
| 7 | Approximation Algorithm for Predicting Secondary Structure with General Pseudoknots | 187 |
| Chapter 9. Protein Localization Prediction | | 193 |
| 1 | Motivation | 194 |
| 2 | Biology of Localization | 196 |
| 3 | Experimental Techniques for Determining Localization Sites | 198 |
| 3.1 | Traditional Methods | 198 |

| | | |
|--|---|------------|
| 3.1.1 | Immunofluorescence Microscopy | 198 |
| 3.1.2 | Gradient Centrifugation | 199 |
| 3.2 | Large-Scale Experiments | 199 |
| 3.2.1 | Immunofluorescent Microscopy | 199 |
| 3.2.2 | Green Fluorescent Protein | 200 |
| 3.2.3 | Comments on Large-Scale Experiments | 200 |
| 4 | Issues and Complications | 201 |
| 4.1 | How Many Sites are There and What are They? | 201 |
| 4.2 | Is One Site Per Protein an Adequate Model? | 202 |
| 4.3 | How Good are the Predictions? | 203 |
| 4.3.1 | Which Method Should I Use? | 204 |
| 4.4 | A Caveat Regarding Estimated Prediction Accuracies | 204 |
| 4.5 | Correlation and Causality | 206 |
| 5 | Localization and Machine Learning | 209 |
| 5.1 | Standard Classifiers Using (Generalized) Amino Acid Content | 210 |
| 5.2 | Localization Process Modeling Approach | 212 |
| 5.3 | Sequence Based Machine Learning Approaches with Architectures Designed to Reflect Localization Signals | 212 |
| 5.4 | Nearest Neighbor Algorithms | 213 |
| 5.5 | Feature Discovery | 213 |
| 5.6 | Extraction of Localization Information from the Literature and Experimental Data | 214 |
| 6 | Conclusion | 215 |
| Chapter 10. Homology Search Methods | | 217 |
| 1 | Overview | 218 |
| 2 | Edit Distance and Alignments | 219 |
| 2.1 | Edit Distance | 219 |
| 2.2 | Optimal Alignments | 220 |
| 2.3 | More Complicated Objectives | 221 |
| 2.3.1 | Score Matrices | 221 |
| 2.3.2 | Gap Penalties | 222 |
| 3 | Sequence Alignment: Dynamic Programming | 222 |
| 3.1 | Dynamic Programming Algorithm for Sequence Alignment | 222 |
| 3.1.1 | Reducing Memory Needs | 224 |
| 3.2 | Local Alignment | 225 |
| 3.3 | Sequence Alignment with Gap Open Penalty | 227 |
| 4 | Probabilistic Approaches to Sequence Alignment | 228 |
| 4.1 | Scoring Matrices | 228 |

Contents

xvii

| | | |
|---|---|------------|
| 4.1.1 | PAM Matrices | 229 |
| 4.1.2 | BLOSUM Matrices | 230 |
| 4.1.3 | Weaknesses of this Approach | 231 |
| 4.2 | Probabilistic Alignment Significance | 231 |
| 5 | Second Generation Homology Search: Heuristics | 232 |
| 5.1 | FASTA and BLAST | 232 |
| 5.2 | Large-Scale Global Alignment | 233 |
| 6 | Next-Generation Homology Search Software | 234 |
| 6.1 | Improved Alignment Seeds | 234 |
| 6.2 | Optimized Spaced Seeds and Why They Are Better | 236 |
| 6.3 | Computing Optimal Spaced Seeds | 238 |
| 6.4 | Computing More Realistic Spaced Seeds | 239 |
| 6.4.1 | Optimal Seeds for Coding Regions | 239 |
| 6.4.2 | Optimal Seeds for Variable-Length Regions | 240 |
| 6.5 | Approaching Smith-Waterman Sensitivity Using Multiple Seed Models | 240 |
| 6.6 | Complexity of Computing Spaced Seeds | 241 |
| 7 | Experiments | 242 |
| Chapter 11. Analysis of Phylogeny: A Case Study on Saururaceae | | 245 |
| 1 | The What, Why, and How of Phylogeny | 246 |
| 1.1 | What is Phylogeny? | 246 |
| 1.2 | Why Study Phylogeny? | 246 |
| 1.3 | How to Study Phylogeny | 247 |
| 2 | Case Study on Phylogeny of Saururaceae | 248 |
| 3 | Materials and Methods | 249 |
| 3.1 | Plant Materials | 249 |
| 3.2 | DNA Extraction, PCR, and Sequencing | 249 |
| 3.3 | Alignment of Sequences | 250 |
| 3.4 | Parsimony Analysis of Separate DNA Sequences | 250 |
| 3.5 | Parsimony Analysis of Combined DNA Sequences | 250 |
| 3.6 | Parsimony Analysis of Morphological Data | 250 |
| 3.7 | Analysis of Each Morphological Characters | 251 |
| 4 | Results | 251 |
| 4.1 | Phylogeny of Saururaceae from 18S Nuclear Genes | 251 |
| 4.2 | Phylogeny of Saururaceae from <i>trnL-F</i> Chloroplast DNA Sequences | 253 |
| 4.3 | Phylogeny of Saururaceae from <i>matR</i> Mitochondrial Genes | 253 |
| 4.4 | Phylogeny of Saururaceae from Combined DNA Sequences | 254 |
| 4.5 | Phylogeny of Saururaceae from Morphological Data | 254 |

| | | |
|--|--|------------|
| 5 | Discussion | 256 |
| 5.1 | Phylogeny of Saururaceae | 256 |
| 5.2 | The Differences Among Topologies from 18S, <i>trnL-F</i> , and <i>matR</i> | 258 |
| 5.3 | Analysis of Important Morphological Characters | 259 |
| 6 | Suggestions | 260 |
| 6.1 | Sampling | 260 |
| 6.2 | Selecting Out-Group | 261 |
| 6.3 | Gaining Sequences | 261 |
| 6.4 | Aligning | 261 |
| 6.5 | Analyzing | 262 |
| 6.6 | Dealing with Morphological Data | 263 |
| 6.7 | Comparing Phylogenies Separately from Molecular Data and Morphological Data | 263 |
| 6.8 | Doing Experiments | 264 |
| Chapter 12. Functional Annotation and Protein Families: From Theory to Practice | | 270 |
| 1 | Introduction | 270 |
| 2 | ProtoNet — Tracing Protein Families | 272 |
| 2.1 | The Concept | 272 |
| 2.2 | The Method and Principle | 273 |
| 2.3 | In Practice | 274 |
| 3 | ProtoNet-Based Tools for Structural Genomics | 282 |
| 4 | PANDORA — Integration of Annotations | 282 |
| 4.1 | The Concept | 282 |
| 4.2 | The Method and Principle | 283 |
| 4.3 | In Practice | 286 |
| 5 | PANDORA-Based Tools for Functional Genomics | 290 |
| Chapter 13. Discovering Protein-Protein Interactions | | 293 |
| 1 | Introduction | 294 |
| 2 | Experimental Detection of Protein Interactions | 294 |
| 2.1 | Traditional Experimental Methods | 295 |
| 2.1.1 | Co-Immunoprecipitation | 295 |
| 2.1.2 | Synthetic Lethal Screening | 296 |
| 2.2 | High Throughput Experimental Methods | 296 |
| 2.2.1 | Yeast Two-Hybrid | 297 |
| 2.2.2 | Phage Display | 299 |
| 2.2.3 | Affinity Purification and Mass Spectrometry | 301 |

| | | |
|-------|---|------------|
| 2.2.4 | Protein Microarrays | 303 |
| 3 | Computational Prediction Protein Interaction | 304 |
| 3.1 | Structure-Based Predictions | 305 |
| 3.1.1 | Structural Homology | 306 |
| 3.2 | Sequence-Based Predictions | 307 |
| 3.2.1 | Interacting Orthologs | 307 |
| 3.2.2 | Interacting Domain Pairs | 308 |
| 3.3 | Genome-Based Predictions | 308 |
| 3.3.1 | Gene Locality Context: Gene Neighborhood | 309 |
| 3.3.2 | Gene Locality Context: Gene Fusion | 310 |
| 3.3.3 | Phylogenetic Context: Phylogenetic Profiles | 312 |
| 3.3.4 | Phylogenetic Context: Phylogenetic Tree Similarity | 313 |
| 3.3.5 | Gene Expression: Correlated mRNA Expression | 316 |
| 4 | Conclusion | 317 |
| | Chapter 14. Techniques for Analysis of Gene Expression | 319 |
| 1 | Microarray and Gene Expression | 320 |
| 2 | Diagnosis by Gene Expression | 322 |
| 2.1 | The Two-Step Approach | 323 |
| 2.2 | Shrunk Centroid Approach | 325 |
| 3 | Co-Regulation of Gene Expression | 328 |
| 3.1 | Hierarchical Clustering Approach | 328 |
| 3.2 | Fuzzy K-Means Approach | 330 |
| 4 | Inference of Gene Networks | 332 |
| 4.1 | Classification-Based Approach | 333 |
| 4.2 | Association Rules Approach | 335 |
| 4.3 | Interaction Generality Approach | 336 |
| 5 | Derivation of Treatment Plan | 339 |
| | Chapter 15. Genome-Wide cDNA Oligo Probe Design and its | |
| | Applications in <i>Schizosaccharomyces Pombe</i> | 347 |
| 1 | Biological Background | 348 |
| 2 | Problem Formulation | 349 |
| 3 | Algorithm Overview | 350 |
| 4 | Implementation Details | 351 |
| 4.1 | Data Schema | 351 |
| 4.2 | Data Objects Creation | 352 |
| 4.3 | Criteria for Probe Production | 354 |
| 4.4 | Probe Production | 355 |

| | | |
|--|--|------------|
| 4.5 | Optimal Probe Set | 356 |
| 5 | Results and Discussions | 357 |
| Chapter 16. Mining New Motifs from cDNA Sequence Data | | 359 |
| 1 | What is a Motif? | 360 |
| 2 | Motif Discovery | 362 |
| 3 | Preparations for Computational Motif Detection | 363 |
| 4 | Extraction of Homologous Sequences and Clustering | 363 |
| 5 | Detection of Homologous Regions with Maximum-Density Subgraphs | 364 |
| 6 | Visualization of Graph-Based Clustering | 364 |
| 7 | Filtering Out Known Motifs | 365 |
| 8 | Extension of New Motif Candidates | 365 |
| 9 | Motif Exploration and Extended Sequence Analysis | 365 |
| 10 | Biological Interpretation of Motifs | 367 |
| 11 | How Many New Motifs are Waiting to be Discovered? | 369 |
| Chapter 17. Technologies for Biological Data Integration | | 375 |
| 1 | Requirements of Integration Systems for Biological Data | 375 |
| 2 | Some Data Integration Solutions | 380 |
| 2.1 | EnsEMBL | 380 |
| 2.2 | GenoMax | 381 |
| 2.3 | SRS | 382 |
| 2.4 | DiscoveryLink | 383 |
| 2.5 | OPM | 385 |
| 2.6 | Kleisli | 386 |
| 2.7 | XML | 387 |
| 3 | Highlight of Selected Features | 388 |
| 3.1 | Data Model and Data Exchange Format | 388 |
| 3.2 | Query Capability | 391 |
| 3.3 | Warehousing Capability | 394 |
| 3.4 | Application Programming Interfaces | 396 |
| 4 | Concluding Remarks | 398 |
| Chapter 18. Construction of Biological Databases: A Case Study on the Protein Phosphatase DataBase (PPDB) | | 401 |
| 1 | Biological Background | 402 |
| 2 | Overview of Architecture and Workflow of PPDB | 403 |
| 3 | Data Integration Tool and Object Representation | 404 |
| 4 | Protein and DNA Sequences | 406 |
| 5 | Structure | 406 |

| | | |
|----|---|-----|
| 6 | Biological Function and Related Information | 407 |
| 7 | Other Data Objects | 408 |
| 8 | Classification Tree | 409 |
| 9 | Data Integration and Classification | 410 |
| 10 | Data Collection | 410 |
| 11 | Phosphatase Classification Strategy | 410 |
| 12 | Automatic Classification and Validation | 412 |
| 13 | Data Integration and Updates | 412 |
| 14 | Data Publication and Internet Access | 413 |
| 15 | Future Development | 414 |

Chapter 19. A Family Classification Approach to Functional

| | | |
|---|---|------------|
| | Annotation of Proteins | 417 |
| 1 | Classification-Driven and Rule-Based Annotation with Evidence Attribution | 418 |
| | 1.1 Protein Family Classification | 418 |
| | 1.2 Rule-Based Annotation and Evidence Attribution | 419 |
| 2 | Case Studies | 421 |
| | 2.1 IMP Dehydrogenase: Error Propagation to Secondary Databases . | 421 |
| | 2.2 His-I Bifunctional Proteins: Transitive Identification Catastrophe . | 424 |
| 3 | Analysis of the Common Identification Errors | 424 |
| 4 | Integrated Knowledge Base System to Facilitate Functional Annotation | 427 |
| | 4.1 PIR-NREF Non-Redundant Reference Database | 427 |
| | 4.2 iProClass Integrated Protein Classification Database | 429 |
| | 4.3 Analytical Tools and Graphical Interfaces | 430 |
| 5 | Functional Associations Beyond Sequence Homology | 433 |

Chapter 20. Informatics for Efficient EST-Based Gene Discovery in

| | | |
|---|---|------------|
| | Normalized and Subtracted cDNA Libraries | 435 |
| 1 | EST-Based Gene Discovery | 436 |
| 2 | Gene Discovery Pipeline Overview | 437 |
| 3 | Pipeline Component Details | 439 |
| | 3.1 Raw Data Gathering and Archival | 441 |
| | 3.2 Sequence Assessment and Sequence Editing | 442 |
| | 3.3 Annotation | 449 |
| | 3.4 Novelty Assessment | 449 |

| | |
|--|-----|
| 3.5 Submission to Local and Public Databases | 452 |
| 4 Discussion | 454 |
| References | 457 |
| List of Contributors | 511 |