# Contents

Preface v

Acknowledgement vii

About the Author viii

1 Introduction to Modern Molecular Biology 1
- 1.1 Cells store large amounts of information in DNA 1
- 1.2 Cells process complex information 7
- 1.3 Cellular life is chemically complex and somewhat stochastic 12
- 1.4 Challenges in analyzing complex biodata 19
- References 19

2 Biodata Explosion 21
- 2.1 Primary sequence and structure data 22
  - 2.1.1 DNA sequence databases 22
  - 2.1.2 Protein sequence databases 27
  - 2.1.3 Molecular structure databases 28
- 2.2 Secondary annotation data 31
  - 2.2.1 Motif annotations 32
  - 2.2.2 Gene function annotations 35
  - 2.2.3 Genomic annotations 36
  - 2.2.4 Inter-species phylogeny and gene family annotations 36

2.3 Experimental and personalized data 38
2.3.1 DNA expression profiles 38
2.3.2 Proteomics data and degradomics 40
2.3.3 Protein expression profiles, 2D gel and protein interaction data 41
2.3.4 Metabolomics and metabolic pathway databases 42
2.3.5 Personalized data 44
2.4 Semantic and processed text data 48
2.4.1 Ontologies 49
2.4.2 Text-mined annotation data 51
2.5 Integrated and federated databases 52
References 55

3 Local Pattern Discovery and Comparing Genes and Proteins 60
3.1 DNA/RNA motif discovery 64
3.1.1 Single motif models: MEME, AlignAce etc. 64
3.1.2 Multiple motif models: LOGOS and MotifRegressor 70
3.1.3 Informative k-mers approach 73
3.2 Protein motif discovery 78
3.2.1 InterProScan and other traditional methods 79
3.2.2 Protein k-mer and other string based methods 82
3.3 Genetic algorithms, particle swarms and ant colonies 84
3.3.1 Genetic algorithms 84
3.3.2 Particle swarm optimization 86
3.3.3 Ant colony optimization 87
3.4 Sequence visualization 88
References 90

4 Global Pattern Discovery and Comparing Genomes 97
4.1 Alignment-based methods 98
4.1.1 Pairwise genome-wide search algorithms: LAGAN, AVID etc. 98

4.1.2 Multiple alignment methods: MLAGAN, MAVID, MULTIZ etc. 98
4.1.3 Dotplots 103
4.1.4 Visualization of genome comparisons 104
4.1.5 Global motif maps 105
4.2 Alignmentless methods 108
4.2.1 K-mer based methods 109
4.2.2 Average common substring and compressibility based methods 114
4.2.3 2D portraits of genomes 117
4.3 Genome scale non-sequence data analysis 125
4.3.1 DNA physical structure based methods 125
4.3.2 Secondary structure based comparisons 131
References 137

5 Molecule Structure Based Searching and Comparison 145
5.1 Molecule structures as graphs or strings 148
5.1.1 3D to 1D transformations 148
5.1.2 Graph matching methods 151
5.1.3 Graph visualization 155
5.1.4 Graph grammars 156
5.2 RNA structure comparison and prediction 157
5.3 Image comparison based methods 162
5.3.1 Gabor filter based methods 165
5.3.2 Image symmetry set based methods 166
5.3.3 Other graph topology based methods 168
References 169

6 Function Annotation and Ontology Based Searching and Classification 176
6.1 Annotation ontologies 176
6.2 Gene Ontology based mining 179
6.3 Sequence similarity based function prediction 182
6.4 Cellular location prediction 184
6.5 New integrative methods: Utilizing networks 186

6.6 Text mining bioliterature for automated annotation 192
6.6.1 Natural language processing (NLP) 193
6.6.2 Semantic profiling 197
6.6.3 Matrix factorization methods 199
References 205

7 New Methods for Genomics Data: SVM and Others 212
7.1 SVM kernels 212
7.2 SVM trees 219
7.3 Methods for microarray data 221
7.3.1 Gene selection algorithms 223
7.3.2 Gene selection by consistency methods 225
7.4 Genome as a time series and discrete wavelet transform 227
7.5 Parameterless clustering for gene expression 231
7.6 Transductive confidence machines, conformal predictors and ROC isometrics 232
7.7 Text compression methods for biodata analysis 236
References 238

8 Integration of Multimodal Data: Toward Systems Biology 245
8.1 Comparative genome annotation systems 246
8.2 Phylogenetics methods 249
8.3 Network inference from interaction and coexpression data 253
8.4 Bayesian inference, association rule mining and Petri nets 258
References 262

9 Future Challenges 266
9.1 Network analysis methods 266
9.2 Unsupervised and supervised clustering 269
9.3 Neural networks and evolutionary methods 270
9.4 Semantic web and ontologization of biology 273

9.5 Biological data fusion 277
9.6 Rise of the GPU machines 279
References 290

Index 297