

Contents

Preface	xi
List of Contributors	xiii
SECTION I INTRODUCTION – DATA DIVERSITY AND INTEGRATION	1
1 Integrative Data Analysis and Visualization: Introduction to Critical Problems, Goals and Challenges	3
<i>Francisco Azuaje and Joaquín Dopazo</i>	
1.1 Data Analysis and Visualization: An Integrative Approach	3
1.2 Critical Design and Implementation Factors	5
1.3 Overview of Contributions	8
References	9
2 Biological Databases: Infrastructure, Content and Integration	11
<i>Allyson L. Williams, Paul J. Kersey, Manuela Pruess and Rolf Apweiler</i>	
2.1 Introduction	11
2.2 Data Integration	12
2.3 Review of Molecular Biology Databases	17
2.4 Conclusion	23
References	26
3 Data and Predictive Model Integration: an Overview of Key Concepts, Problems and Solutions	29
<i>Francisco Azuaje, Joaquín Dopazo and Haiying Wang</i>	
3.1 Integrative Data Analysis and Visualization: Motivation and Approaches	29
3.2 Integrating Informational Views and Complexity for Understanding Function	31
3.3 Integrating Data Analysis Techniques for Supporting Functional Analysis	34
3.4 Final Remarks	36
References	38

SECTION II	INTEGRATIVE DATA MINING AND VISUALIZATION – EMPHASIS ON COMBINATION OF MULTIPLE DATA TYPES	41
4	Applications of Text Mining in Molecular Biology, from Name Recognition to Protein Interaction Maps	43
	<i>Martin Krallinger and Alfonso Valencia</i>	
4.1	Introduction	44
4.2	Introduction to Text Mining and NLP	45
4.3	Databases and Resources for Biomedical Text Mining	47
4.4	Text Mining and Protein–Protein Interactions	50
4.5	Other Text-Mining Applications in Genomics	55
4.6	The Future of NLP in Biomedicine	56
	Acknowledgements	56
	References	56
5	Protein Interaction Prediction by Integrating Genomic Features and Protein Interaction Network Analysis	61
	<i>Long J. Lu, Yu Xia, Haiyuan Yu, Alexander Rives, Haoxin Lu, Falk Schubert and Mark Gerstein</i>	
5.1	Introduction	62
5.2	Genomic Features in Protein Interaction Predictions	63
5.3	Machine Learning on Protein–Protein Interactions	67
5.4	The Missing Value Problem	73
5.5	Network Analysis of Protein Interactions	75
5.6	Discussion	79
	References	80
6	Integration of Genomic and Phenotypic Data	83
	<i>Amanda Clare</i>	
6.1	Phenotype	83
6.2	Forward Genetics and QTL Analysis	85
6.3	Reverse Genetics	87
6.4	Prediction of Phenotype from Other Sources of Data	88
6.5	Integrating Phenotype Data with Systems Biology	90
6.6	Integration of Phenotype Data in Databases	93
6.7	Conclusions	95
	References	95
7	Ontologies and Functional Genomics	99
	<i>Fátima Al-Shahrour and Joaquín Dopazo</i>	
7.1	Information Mining in Genome-Wide Functional Analysis	99
7.2	Sources of Information: Free Text Versus Curated Repositories	100
7.3	Bio-Ontologies and the Gene Ontology in Functional Genomics	101
7.4	Using GO to Translate the Results of Functional Genomic Experiments into Biological Knowledge	103

- 7.5 Statistical Approaches to Test Significant Biological Differences
- 7.6 Using FatiGO to Find Significant Functional Associations in Clusters of Genes
- 7.7 Other Tools
- 7.8 Examples of Functional Analysis of Clusters of Genes
- 7.9 Future Prospects
- References

8 The *C. elegans* Interactome: its Generation and Visualization

Alban Chesnau and Claude Sardet

- 8.1 Introduction
- 8.2 The ORFeome: the first step toward the interactome of *C. elegans*
- 8.3 Large-Scale High-Throughput Yeast Two-Hybrid Screens to Map the *C. elegans* Protein-Protein Interaction (Interactome) Network: Technical Aspects
- 8.4 Visualization and Topology of Protein-Protein Interaction Networks
- 8.5 Cross-Talk Between the *C. elegans* Interactome and other Large-Scale Genomics and Post-Genomics Data Sets
- 8.6 Conclusion: From Interactions to Therapies
- References

SECTION III INTEGRATIVE DATA MINING AND VISUALIZATION – EMPHASIS ON COMBINATION OF MULTIPLE PREDICTION MODELS AND METHODS

9 Integrated Approaches for Bioinformatic Data Analysis and Visualization – Challenges, Opportunities and New Solutions

Steve R. Pettifer, James R. Sinnott and Teresa K. Attwood

- 9.1 Introduction
- 9.2 Sequence Analysis Methods and Databases
- 9.3 A View Through a Portal
- 9.4 Problems with Monolithic Approaches: One Size Does Not Fit All
- 9.5 A Toolkit View
- 9.6 Challenges and Opportunities
- 9.7 Extending the Desktop Metaphor
- 9.8 Conclusions
- Acknowledgements
- References

10 Advances in Cluster Analysis of Microarray Data

Qizheng Sheng, Yves Moreau, Frank De Smet, Kathleen Marchal and Bart De Moor

- 10.1 Introduction
- 10.2 Some Preliminaries
- 10.3 Hierarchical Clustering
- 10.4 *k*-Means Clustering

10.5	Self-Organizing Maps	159
10.6	A Wish List for Clustering Algorithms	160
10.7	The Self-Organizing Tree Algorithm	161
10.8	Quality-Based Clustering Algorithms	162
10.9	Mixture Models	163
10.10	Biclustering Algorithms	166
10.11	Assessing Cluster Quality	168
10.12	Open Horizons	170
	References	171
11	Unsupervised Machine Learning to Support Functional Characterization of Genes: Emphasis on Cluster Description and Class Discovery	175
	<i>Olga G. Troyanskaya</i>	
11.1	Functional Genomics: Goals and Data Sources	175
11.2	Functional Annotation by Unsupervised Analysis of Gene Expression Microarray Data	177
11.3	Integration of Diverse Functional Data For Accurate Gene Function Prediction	179
11.4	MAGIC – General Probabilistic Integration of Diverse Genomic Data	180
11.5	Conclusion	188
	References	189
12	Supervised Methods with Genomic Data: a Review and Cautionary View	193
	<i>Ramón Díaz-Uriarte</i>	
12.1	Chapter Objectives	193
12.2	Class Prediction and Class Comparison	194
12.3	Class Comparison: Finding/Ranking Differentially Expressed Genes	194
12.4	Class Prediction and Prognostic Prediction	198
12.5	ROC Curves for Evaluating Predictors and Differential Expression	201
12.6	Caveats and Admonitions	203
12.7	Final Note: Source Code Should be Available	209
	Acknowledgements	210
	References	210
13	A Guide to the Literature on Inferring Genetic Networks by Probabilistic Graphical Models	215
	<i>Pedro Larrañaga, Iñaki Inza and Jose L. Flores</i>	
13.1	Introduction	215
13.2	Genetic Networks	216
13.3	Probabilistic Graphical Models	218
13.4	Inferring Genetic Networks by Means of Probabilistic Graphical Models	229
13.5	Conclusions	234
	Acknowledgements	235
	References	235

14 Integrative Models for the Prediction and Understanding of Protein Structure Patterns	239
<i>Inge Jonassen</i>	
14.1 Introduction	239
14.2 Structure Prediction	241
14.3 Classifications of Structures	244
14.4 Comparing Protein Structures	246
14.5 Methods for the Discovery of Structure Motifs	249
14.6 Discussion and Conclusions	252
References	254
Index	257