

Contents

Preface	xi
Christine FROIDEVAUX, Marie-Laure MARTIN-MAGNIETTE and Guillem RIGAILL	
Part 1. Knowledge Integration	1
Chapter 1. Clinical Data Warehouses	3
Maxime WACK and Bastien RANCE	
1.1. Introduction to clinical information systems and biomedical warehousing: data warehouses for what purposes?	3
1.1.1. Warehouse history	4
1.1.2. Using data warehouses today	4
1.2. Challenge: widely scattered data	5
1.3. Data warehouses and clinical data	6
1.3.1. Warehouse structures	6
1.3.2. Warehouse construction and supply	11
1.3.3. Uses	11
1.4. Warehouses and omics data: challenges	15
1.4.1. Challenges of data volumetry and structuring omic data	16
1.4.2. Attempted solutions	17
1.5. Challenges and prospects	18
1.5.1. Toward general-purpose warehouses	18
1.5.2. Ethical dimension of the implementation and the use of warehouses	19
1.5.3. Origin and reproducibility	19
1.5.4. Data quality	20
1.5.5. Data warehousing federation and data sharing	21
1.6. References	21

Chapter 2. Semantic Web Methods for Data Integration in Life Sciences	25
Olivier DAMERON	
2.1. Data-related requirements in life sciences	26
2.1.1. Databases for the life sciences	26
2.1.2. Requirements	27
2.1.3. Common approaches: InterMine and BioMart	30
2.2. Semantic Web	31
2.2.1. Techniques	32
2.2.2. Implementation	42
2.3. Perspectives	43
2.3.1. Facilitating appropriation to users	43
2.3.2. Facilitating the appropriation by software programs: FAIR data	44
2.3.3. Federated queries	45
2.4. Conclusion	46
2.5. References	47
Chapter 3. Workflows for Bioinformatics Data Integration	53
Sarah COHEN-BOULAKIA and Frédéric LEMOINE	
3.1. Introduction	53
3.2. Bioinformatics data processing chains: difficulties	54
3.2.1. Designing a data processing chain	55
3.2.2. Analysis execution and reproducibility	56
3.2.3. Maintenance, sharing and reuse	58
3.3. Solutions provided by scientific workflow systems	59
3.3.1. Fundamentals of workflow systems	59
3.3.2. Workflow systems	64
3.4. Use case: RNA-seq data analysis	69
3.4.1. Study description	69
3.4.2. From data processing chain to workflows	72
3.4.3. Data processing chains implemented as workflows: conclusion . .	75
3.5. Challenges, open problems and research opportunities	77
3.5.1. Formalizing workflow development	77
3.5.2. Workflow testing	78
3.5.3. Discovering and sharing workflows	79
3.6. Conclusion	80
3.7. References	81

Part 2. Integration and Statistics	87
Chapter 4. Variable Selection in the General Linear Model: Application to Multiomic Approaches for the Study of Seed Quality	89
Céline LÉVY-LEDUC, Marie PERROT-DOCKÈS, Gwendal CUEFF and Loïc RAJJOU	
4.1. Introduction	90
4.2. Methodology	93
4.2.1. Estimation of the covariance matrix Σ_q	93
4.2.2. Estimation of \mathcal{B}	96
4.3. Numerical experiments	99
4.3.1. Statistical performance	99
4.3.2. Numerical performance	100
4.4. Application to the study of seed quality	103
4.4.1. Metabolomics data	104
4.4.2. Proteomics data	105
4.5. Conclusion	108
4.6. Appendices	108
4.6.1. Example of using the package <code>MultiVarSel</code> for metabolomic data analysis	108
4.6.2. Example of using the package <code>MultiVarSel</code> for proteomic data analysis	110
4.7. Acknowledgments	113
4.8. References	113
Chapter 5. Structured Compression of Genetic Information and Genome-Wide Association Study by Additive Models	117
Florent GUINOT, Marie SZAFRANSKI and Christophe AMBROISE	
5.1. Genome-wide association studies	118
5.1.1. Introduction to genetic mapping and linkage analysis	118
5.1.2. Principles of genome-wide association studies	119
5.1.3. Single nucleotide polymorphism	120
5.1.4. Disease penetrance and <i>odds ratio</i>	122
5.1.5. Single marker analysis	124
5.1.6. Multi-marker analysis	126
5.2. Structured compression and association study	132
5.2.1. Context	132
5.2.2. New structured compression approach	133
5.3. Application to ankylosing spondylitis (AS)	142
5.3.1. Data	142
5.3.2. Predictive power evaluation	143

5.3.3. Manhattan diagram	144
5.3.4. Estimation for the most significant SNP aggregates	144
5.4. Conclusion	146
5.5. References	146
Chapter 6. Kernels for Omics	151
Jérôme MARIETTE and Nathalie VIALANEIX	
6.1. Introduction	152
6.2. Relational data	153
6.2.1. Data described by the kernel	153
6.2.2. Data described by a general (dis)similarity measure	155
6.3. Exploratory analysis for relational data	158
6.3.1. Kernel clustering	158
6.3.2. Kernel principal component analysis	161
6.3.3. Kernel self-organizing maps	163
6.3.4. Limitations of relational methods	166
6.4. Combining relational data	168
6.4.1. Data integration in systems biology	168
6.4.2. Kernel approaches in data integration	169
6.4.3. A consensual kernel	172
6.4.4. A parsimonious kernel that preserves the topology of the initial data	173
6.4.5. A complete kernel preserving the topology of the initial data . . .	175
6.5. Application	176
6.5.1. Loading Tara Ocean data	176
6.5.2. Data integration by kernel approaches	177
6.5.3. Exploratory analysis: kernel PCA	179
6.6. Session information for the results of the example	186
6.7. References	188
Chapter 7. Multivariate Models for Data Integration and Biomarker Selection in 'Omics Data	195
Sébastien DÉJEAN and Kim-Anh LÊ CAO	
7.1. Introduction	195
7.2. Background	197
7.2.1. Mathematical notations	197
7.2.2. Terminology	198
7.2.3. Multivariate projection-based approaches	198
7.2.4. A criterion to maximize specific to each methodology	199
7.2.5. A linear combination of variables to reduce the dimension of the data	199

7.2.6. Identifying a subset of relevant molecular features	200
7.2.7. Summary	200
7.3. From the biological question to the statistical analysis	201
7.3.1. Exploration of one dataset: PCA	201
7.3.2. Classify samples: projection to latent structure discriminant analysis	206
7.3.3. Integration of two datasets: projection to latent structure and related methods	210
7.3.4. Integration of several datasets: multi-block approaches	215
7.4. Graphical outputs	220
7.4.1. Individual plots	220
7.4.2. Variable plots	221
7.5. Overall summary	222
7.6. Liver toxicity study	223
7.6.1. The datasets	223
7.6.2. Biological questions and statistical methods	223
7.6.3. Single dataset analysis	224
7.6.4. Integrative analysis	231
7.7. Conclusion	238
7.8. Acknowledgments	238
7.9. Appendix: reproducible R code	239
7.9.1. Toy examples	239
7.9.2. Liver toxicity	243
7.10. References	247
List of Authors	251
Index	255