

Contents

Preface	xiii
List of Contributors	xvii
PART I FUNDAMENTALS	1
Section 1 Knowledge-Driven Approaches	3
1 Knowledge-based bioinformatics	5
<i>Eric Karl Neumann</i>	
1.1 Introduction	5
1.2 Formal reasoning for bioinformatics	7
1.3 Knowledge representations	10
1.4 Collecting explicit knowledge	10
1.5 Representing common knowledge	11
1.6 Capturing novel knowledge	15
1.7 Knowledge discovery applications	15
1.8 Semantic harmonization: the power and limitation of ontologies	18
1.9 Text mining and extraction	19
1.10 Gene expression	20
1.11 Pathways and mechanistic knowledge	22
1.12 Genotypes and phenotypes	24
1.13 The Web's role in knowledge mining	25
1.14 New frontiers	26
1.14.1 Requirements for linked knowledge discovery	26
1.14.2 Information aggregation	26
1.14.3 The Linked Open Data initiative	28
1.14.4 Information articulation	28
1.14.5 Next-generation knowledge discovery	30
1.15 References	31
2 Knowledge-driven approaches to genome-scale analysis	33
<i>Hannah Tipney and Lawrence Hunter</i>	
2.1 Fundamentals	33
2.1.1 The genomic era and systems biology	33

2.1.2	The exponential growth of biomedical knowledge	34
2.1.3	The challenges of finding and interacting with biomedical knowledge	35
2.2	Challenges in knowledge-driven approaches	37
2.2.1	We need to read; development of automatic methods to extract data housed in the biomedical literature	37
2.2.2	Implicit and implied knowledge; the forgotten data source	41
2.2.3	Humans are visual beings: so should their knowledge be	42
2.3	Current knowledge-based bioinformatics tools	43
2.3.1	Enrichment tools	44
2.3.2	Integration and expansion: from gene lists to networks	46
2.3.3	Expanding the concept of an interaction	48
2.3.4	A systematic failure to support advanced scientific reasoning	50
2.4	3R systems: reading, reasoning and reporting the way towards biomedical discovery	50
2.4.1	3R knowledge networks populated by reading and reasoning	52
2.4.2	Implied association results in uncertainty	53
2.4.3	Reporting: using 3R knowledge networks to tell biological stories	54
2.5	The Hanalyzer: a proof of 3R concept	55
2.6	Acknowledgements	62
2.7	References	62
3	Technologies and best practices for building bio-ontologies	67
	<i>Mikel Egaña Aranguren, Robert Stevens, Erick Antezana, Jesualdo Tomás Fernández-Breis, Martin Kuiper, and Vladimir Mironov</i>	
3.1	Introduction	67
3.2	Knowledge representation languages and tools for building bio-ontologies	68
3.2.1	RDF (resource description framework)	71
3.2.2	OWL (Web ontology language)	72
3.2.3	OBO format	77
3.3	Best practices for building bio-ontologies	78
3.3.1	Define the scope of the bio-ontology	78
3.3.2	Identity of the represented entities	79
3.3.3	Commit to agreed ontological principles	79
3.3.4	Knowledge acquisition	80
3.3.5	Ontology Design Patterns (ODPs)	80
3.3.6	Ontology evaluation	81
3.3.7	Documentation	83
3.4	Conclusion	83
3.5	Acknowledgements	84
3.6	References	84

4	Design, implementation and updating of knowledge bases	87
	<i>Sarah Hunter, Rolf Apweiler, and Maria Jesus Martin</i>	
4.1	Introduction	87
4.2	Sources of data in bioinformatics knowledge bases	90
4.2.1	Data added by internal curators	90
4.2.2	Data submitted by external users and collaborators	90
4.2.3	Data added automatically	91
4.3	Design of knowledge bases	91
4.3.1	Understanding your end users and understanding their data	92
4.3.2	Interactions and interfaces: their impact on design	93
4.4	Implementation of knowledge bases	93
4.4.1	Choosing a database architecture	93
4.4.2	Good programming practices	96
4.4.3	Implementation of interfaces	97
4.5	Updating of knowledge bases	98
4.5.1	Manual curation and auto-annotation	98
4.5.2	Clever pipelines and data flows	101
4.5.3	Lessening data maintenance overheads	104
4.6	Conclusions	105
4.7	References	105
Section 2	Data-Analysis Approaches	107
5	Classical statistical learning in bioinformatics	109
	<i>Mark Reimers</i>	
5.1	Introduction	109
5.2	Significance testing	109
5.2.1	Multiple testing and false discovery rate	110
5.2.2	Correlated errors	111
5.3	Exploratory analysis	112
5.3.1	Clustering	112
5.3.2	Principal components	116
5.3.3	Multidimensional scaling (MDS)	117
5.4	Classification and prediction	119
5.4.1	Discriminant analysis	120
5.4.2	Modern procedures	120
5.5	References	122
6	Bayesian methods in genomics and proteomics studies	125
	<i>Ning Sun and Hongyu Zhao</i>	
6.1	Introduction	125
6.2	Bayes theorem and some simple applications	126
6.3	Inference of population structure from genetic marker data	129
6.4	Inference of protein binding motifs from sequence data	130

6.5	Inference of transcriptional regulatory networks from joint analysis of protein–DNA binding data and gene expression data	131
6.6	Inference of protein and domain interactions from yeast two-hybrid data	132
6.7	Conclusions	134
6.8	Acknowledgements	135
6.9	References	135
7	Automatic text analysis for bioinformatics knowledge discovery	137
	<i>Dietrich Rebholz-Schuhmann and Jung-jae Kim</i>	
7.1	Introduction	137
	7.1.1 Knowledge discovery through text mining	138
	7.1.2 Need for processing biomedical texts	139
	7.1.3 Developing text mining solutions	141
7.2	Information needs for biomedical text mining	142
	7.2.1 Efficient analysis of normalized information	142
	7.2.2 Interactive seeking of textual information	145
7.3	Principles of text mining	147
	7.3.1 Components	147
	7.3.2 Methods	150
7.4	Development issues	152
	7.4.1 Information needs	153
	7.4.2 Corpus construction	153
	7.4.3 Language analysis	154
	7.4.4 Integration framework	154
	7.4.5 Evaluation	155
7.5	Success stories	156
	7.5.1 Interactive literature analysis	156
	7.5.2 Integration into bioinformatics solutions	157
	7.5.3 Discovery of knowledge from the literature	158
7.6	Conclusion	159
7.7	References	160
PART II	APPLICATIONS	169
Section 3	Gene and Protein Information	171
8	Fundamentals of gene ontology functional annotation	173
	<i>Varsha K. Khodiyar, Emily C. Dimmer, Rachael P. Huntley, and Ruth C. Lovering</i>	
8.1	Introduction	173
	8.1.1 Data submission curation	174
	8.1.2 Value-added curation	174
8.2	Gene Ontology (GO)	175

8.2.1	Gene Ontology and the annotation of the human proteome	175
8.2.2	Gene Ontology Consortium data sets	176
8.2.3	GO annotation methods	176
8.2.4	Different approaches to manual annotation	183
8.2.5	Ontology development	183
8.3	Comparative genomics and electronic protein annotation	186
8.3.1	Manual methods of transferring functional annotation	186
8.3.2	Electronic methods of transferring functional annotation	187
8.3.3	Electronic annotation methods	188
8.4	Community annotation	189
8.4.1	Feedback forms	190
8.4.2	Wiki pages	190
8.4.3	Community annotation workshops	190
8.5	Limitations	191
8.5.1	GO cannot capture all relevant biological aspects	191
8.5.2	The ontology is always evolving	192
8.5.3	The volume of literature	192
8.5.4	Missing published data	192
8.5.5	Manual curation is expensive	192
8.6	Accessing GO annotations	193
8.6.1	Tools for browsing the GO	194
8.6.2	Functional classification	199
8.6.3	GO slims	202
8.6.4	GO displays in other databases	203
8.7	Conclusions	203
8.8	References	204
9	Methods for improving genome annotation	209
	<i>Jonathan Mudge and Jennifer Harrow</i>	
9.1	The basis of gene annotation	209
9.1.1	Introduction to gene annotation	209
9.1.2	Progression in <i>ab initio</i> gene prediction	211
9.1.3	Annotation based on transcribed evidence	211
9.1.4	A comparison of annotation processes	213
9.1.5	The CCDS project	214
9.1.6	Pseudogene annotation	215
9.1.7	The annotation of non-coding genes	218
9.2	The impact of next generation sequencing on genome annotation	220
9.2.1	The annotation of multispecies genomes	220
9.2.2	Community annotation	222
9.2.3	Alternative splicing and new transcriptomics data	223
9.2.4	The annotation of human genome variation	225

9.2.5	The annotation of polymorphic gene families	226
9.3	References	228
10	Sequences from prokaryotic, eukaryotic, and viral genomes available clustered according to phylotype on a Self-Organizing Map	233
	<i>Takashi Abe, Shigehiko Kanaya, and Toshimichi Ikemura</i>	
10.1	Introduction	233
10.2	Batch-learning SOM (BLSOM) adapted for genome informatics	235
10.3	Genome sequence analyses using BLSOM	237
10.3.1	BLSOMs for 13 eukaryotic genomes	237
10.3.2	Diagnostic oligonucleotides for phylotype-specific clustering	238
10.3.3	A large-scale BLSOM constructed with all sequences available from species-known genomes	240
10.3.4	Phylogenetic estimation for environmental DNA sequences and microbial community comparison using the BLSOM	242
10.3.5	Reassociation of environmental genomic fragments according to species	245
10.4	Conclusions and discussion	247
10.5	References	248
Section 4	Biomolecular Relationships and Meta-Relationships	251
11	Molecular network analysis and applications	253
	<i>Minlu Zhang, Jingyuan Deng, Chunsheng V. Fang, Xiao Zhang, and Long Jason Lu</i>	
11.1	Introduction	253
11.2	Topology analysis and applications	254
11.2.1	Global structure of molecular networks: scale-free, small-world, disassortative, and modular	254
11.2.2	Network statistics/measures	258
11.2.3	Applications of topology analysis	258
11.2.4	Challenges and future directions of topology analysis	262
11.3	Network motif analysis	263
11.3.1	Motif analysis: concept and method	263
11.3.2	Applications of motif analysis	263
11.3.3	Challenges and future directions of motif analysis	266
11.4	Network modular analysis and applications	267
11.4.1	Density-based clustering methods	268
11.4.2	Partition-based clustering methods	269

11.4.3	Centrality-based clustering methods	270
11.4.4	Hierarchical clustering methods	271
11.4.5	Applications of modular analysis	272
11.4.6	Challenges and future directions of modular analysis	273
11.5	Network comparison	274
11.5.1	Network comparison algorithms: from computer science to systems biology	274
11.5.2	Network comparison algorithms for molecular networks	275
11.5.3	Applications of molecular network comparison	277
11.5.4	Challenges and future directions of network comparison	278
11.6	Network analysis software and tools	279
11.7	Summary	279
11.8	Acknowledgement	282
11.9	References	282
12	Biological pathway analysis: an overview of Reactome and other integrative pathway knowledge bases	289
	<i>Robin A. Haw, Marc E. Gillespie, and Michael A. Caudy</i>	
12.1	Biological pathway analysis and pathway knowledge bases	289
12.2	Overview of high-throughput data capture technologies and data repositories	290
12.3	Brief review of selected pathway knowledge bases	293
12.3.1	Reactome	293
12.3.2	KEGG	296
12.3.3	WikiPathways	297
12.3.4	NCI-Pathway Interaction Database	297
12.3.5	NCBI-BioSystems	298
12.3.6	Science Signaling	299
12.3.7	PharmGKB	299
12.4	How does information get into pathway knowledge bases?	300
12.5	Introduction to data exchange languages	301
12.5.1	SBML	301
12.5.2	BioPAX	302
12.5.3	PSI MI	303
12.5.4	Comparison of data exchange formats for different pathway knowledge bases	303
12.6	Visualization tools	304
12.7	Use case: pathway analysis in Reactome using statistical analysis of high-throughput data sets	305
12.8	Discussion: challenges and future directions of pathway knowledge bases	310
12.9	References	311

13 Methods and challenges of identifying biomolecular relationships and networks associated with complex diseases/phenotypes, and their application to drug treatments	315
<i>Mie Rizig</i>	
13.1 Complex traits: clinical phenomenology and molecular background	315
13.2 Why it is challenging to infer relationships between genes and phenotypes in complex traits?	317
13.3 Bottom-up or top-down: which approach is more useful in delineating complex traits key drivers?	325
13.4 High-throughput technologies and their applications in complex traits genetics	327
13.5 Integrative systems biology: a comprehensive approach to mining high-throughput data	328
13.6 Methods applying systems biology approach in the identification of functional relationships from gene expression data	331
13.6.1 Methods using quantitative expression data to identify correlations in expression between genes (clustering)	334
13.6.2 Methods integrating functional genomics into cellular functional classes	338
13.6.3 Methods combining functional genomics results and existing biological information to construct novel biological networks	343
13.7 Advantages of networks exploration in molecular biology and drug discovery	353
13.8 Practical examples of applying systems biology approaches and network exploration in the identification of functional modules and disease-causing genes in complex phenotypes/diseases	354
13.9 Challenges and future directions	363
13.10 References	364
Trends and conclusion	367
Index	369