

Péter Antal – Ádám Arany – Bence Bolgár – András Gézsi – Gergely Hajós
 – Gábor Hullám – Péter Marx – András Millinghoffer – László Poppe
 – Péter Sárközy

BIOINFORMATICS

The Bioinformatics book covers new topics in the rapidly expanding field of bioinformatics, from next-generation sequencing to drug discovery and metagenomics.

The first two chapters overviews genetic measurement methods. The next four chapters discuss topics related to the effect of genetic variants from protein modeling to gene regulatory networks. Standard statistical analysis in association studies are discussed in the next two chapters. The systems biology approach is illustrated by discussing a systems-based biomarker analysis method, the graph-based network science, the dynamical systems based approaches and a Bayesian causal inference method in subsequent chapters. The next chapter discusses text-mining methods in biomedicine, especially their application in interpretation and translation. The decision theoretic approach to study design, especially multi-stage, sequential study design is discussed in the next chapter, introducing the concepts of value of information and the expected value of an experiment. Next, the heterogeneity of biomedical big data sources is overviewed, together with data and knowledge fusion methods, and with the discussion of semantic publishing, which can lead to a new unification of biomedicine. Subsequently, bioinformatic workflow methods are summarized. At last, drug discovery methods are overviewed with an outlook for personalized medicine and the final chapter presents the main steps and workflows in metagenomics.

Keywords: genotyping, next-generation sequencing methods, protein modeling, gene regulatory networks, omic networks, study design, data and knowledge fusion, workflow systems, association study, biomarker analysis, medical decision support systems, semantic publishing, similarity based drug discovery, metagenomics.

Budapest University of Technology & Economics and Semmelweis University



Typotex Kiadó
 2014

COPYRIGHT: © 2014–2019, Péter Antal, Ádám Arany, Bence Bolgár, András Gézsi, Gergely Hajós, Gábor Hullám, Péter Marx, András Millinghoffer, László Poppe, Péter Sárközy, Budapest University of Technology and Economics, Semmelweis University

Creative Commons NonCommercial-NoDerivs 3.0 (CC BY-NC-ND 3.0)

“Terms of use of ©: This work can be reproduced, circulated, published and performed for non-commercial purposes without restriction by indicating the author’s name, but it cannot be modified.”

Scientific lecturers: Viktor Molnár, András Antos

ISBN 978 963 279 179 1

Prepared under the editorship of Typotex Kiadó
Responsible manager: Zsuzsa Votisky

Prepared within the framework of the project “Konzorcium a biotechnológia aktív tanulásáért” (“Consortium for the Active Studying of Biotechnology”) Grant No. TÁMOP-4.1.2/A/1-11/1-2011-0079.

Nemzeti Fejlesztési Ügynökség
www.ujszechenyiterv.gov.hu
06 40 638 638



A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Contents

1	DNA recombinant measurement technology, noise and error models	11
1.1	Historic overview	11
1.1.1	Clinical aspects of genome sequencing	12
1.1.2	Partial Genetic Association Studies	12
1.1.3	Genome Wide Association Studies	12
1.2	First generation automated Sanger sequencing	12
1.3	Next generation sequencing technologies	14
1.3.1	Pyrosequencing and pH based sequencing	15
1.3.2	Reversible terminator based sequencing	16
1.3.3	Nanopore based sequencing	16
1.4	Error characteristics of Next Generation Sequencing	16
1.4.1	Carry forward/incomplete extension	18
1.4.2	Homopolymer errors	18
1.5	Capture technologies	19
1.5.1	PCR capture	19
1.6	Emulsion PCR	21
1.7	Bridge amplification	24
1.8	Targeted resequencing	24
1.9	De-novo sequencing	25
1.10	Next generation sequencing workflows	25
1.10.1	Filtering	25
1.10.2	Mapping	25
1.10.3	Assembly	26
1.10.4	Variant calling	26
1.10.5	Paired end sequencing	26
1.11	Multiplexing samples	26
2	The post-processing, haplotype reconstruction, and imputation...	28
2.1	Genome	28
2.2	Genotype	28
2.2.1	Single nucleotide polymorphisms	29
2.2.2	Types of point mutation	30
2.3	Haplotypes and recombination	30

2.4	Linkage Disequilibrium	31
2.5	Haplotype reconstruction	32
2.6	Imputation	33
2.7	Genotyping platforms	33
2.7.1	Sample preparation	34
2.7.2	Regions of interest	34
2.7.3	Primer Design	35
2.7.4	PCR	35
2.7.5	Probe-tag based genotyping	37
2.7.6	Sanger sequencing	37
2.7.7	Real-time qualitative polymerase chain reaction	37
2.7.8	SNP arrays	38
2.8	Genotyping vs. gene expression	38
2.8.1	Call rate and accuracy	39
3	Comparative protein modeling and molecular docking	40
3.1	Introduction	40
3.1.1	The protein structure gap	40
3.1.2	Methods of protein modeling	41
3.2	Comparative protein modeling	43
3.2.1	Steps of homology modeling	43
3.2.2	Tools for homology modeling	48
3.3	Molecular docking	50
3.3.1	Protein-ligand interaction predictions	51
3.3.2	Protein-biomacromolecule interaction predictions	51
4	Methods of determining structure of proteins and protein structure databases	56
4.1	Introduction	56
4.1.1	Protein identification tools	56
4.1.2	Simple protein analyses	57
4.1.3	Levels and problems of protein structure predictions	57
4.2	Experimental methods to determine the secondary structure of proteins . .	58
4.2.1	Protein circular dichroism (CD)	59
4.2.2	Synchrotron radiation circular dichroism (SRCd)	60
4.3	Experimental methods to determining atomic structures of proteins	60
4.3.1	Protein X-ray crystallography	61
4.3.2	Protein NMR spectroscopy	63
4.3.3	Protein electron microscopy, electron diffraction and electron crystallography	66
4.3.4	Protein neutron crystallography	67

5	Quantitative models of the functional effects of genetic variants	70
5.1	Introduction	70
5.2	Variants	70
5.2.1	SNP, indel	71
5.2.2	Alternative splicing	71
5.3	Levels of regulation	72
5.4	Different regulatory elements	72
5.5	microRNA	72
5.5.1	miRNA development	73
5.5.2	miRNA regulatory methods	73
5.6	Transcription factors	73
5.7	Epigenetics	74
5.7.1	Methylation	74
5.7.2	Histone modifications	74
6	Mathematical models of gene regulatory networks	78
6.1	Introduction	78
6.2	Learning networks	78
6.2.1	Representation	78
6.2.2	Types of network learning algorithms	78
6.3	TF, miRNA, mRNA regulatory networks	81
7	Standard analysis of genetic association studies	84
7.1	Introduction	84
7.2	Genetic data transformation	85
7.2.1	Filtering	85
7.2.2	Standard test for Hardy–Weinberg equilibrium	85
7.3	Phenotype data transformation	86
7.3.1	Transformation	86
7.3.2	Discretization	87
7.4	Univariate analysis methods	87
7.4.1	Standard association tests	87
7.4.2	Cochran–Armitage test for trend	89
7.4.3	Odds ratios	90
7.4.4	Univariate Bayesian methods	92
7.5	Multivariate analysis methods	92
7.5.1	Logistic regression	92
7.5.2	Haplotype association	93
7.5.3	Analysis of statistical power	97
8	Analyzing gene expression studies	101
8.1	Introduction	101
8.2	Pre-procession	102

8.2.1	Background correction	102
8.2.2	Normalization	102
8.2.3	Summarization	103
8.2.4	Filtering	104
8.3	Data analysis	104
8.3.1	Clustering	104
8.3.2	Differential expression	108
8.3.3	Biological interpretation of results	110
9	Biomarker analysis	115
	Notation	115
9.1	Introduction	118
9.2	Background	118
9.3	Bayesian multilevel analysis of relevance	120
9.4	Multivariate scalability: k-MBS and k-MBG features	121
9.5	A knowledge-rich aggregation of input features	122
9.6	Interaction, redundancy based on posterior decomposition	122
9.7	Relevance for multiple targets	123
9.8	Conditional and contextual relevance	124
9.9	Posteriors for the predictive power of input features	125
9.10	Algorithmic aspects and applications	125
9.11	Summary	125
10	Network biology	129
10.1	Introduction	129
10.2	Biological networks	130
10.3	Basics of graph theory	131
10.4	Network analysis	132
10.4.1	Network topology	132
10.4.2	Network models and dynamics	133
10.4.3	Assortativity, degree distribution and scale-free networks	134
10.4.4	Tasks and challenges	135
10.5	An application to drug discovery	137
11	Dynamic modeling in cell biology	141
11.1	Biochemical concepts and their computational representations	141
11.2	Modeling with ordinary differential equations	144
11.3	Stochastic modeling	145
11.4	Hybrid methods	146
11.5	Reaction–diffusion systems	147
11.6	Model fitting	148
11.7	Whole-cell simulation	149
11.8	Overview	149

12 Causal inference in biomedicine	152
Notation	152
12.1 Introduction	155
12.2 Representing independence and causal relations by Bayesian networks . . .	155
12.3 Constraint based inference of causal relations and models	159
12.4 Learning complete causal domain models	160
12.5 Bayesian inference of causal features	161
12.5.1 Edges: direct pairwise dependencies	161
12.5.2 Pairwise causal relations	162
12.5.3 MBG subnetworks	162
12.5.4 Ordering of the variables	162
12.5.5 Effect modifiers	163
13 Text mining methods in bioinformatics	168
13.1 Introduction	168
13.2 Biomedical text mining	168
13.2.1 Constructing the corpus	169
13.2.2 Constructing the vocabulary	171
13.2.3 Text mining tasks	172
13.3 Basic techniques	173
13.3.1 Pattern matching	173
13.3.2 Document representation	173
13.3.3 Methods for named entity recognition	175
13.3.4 Methods for relation extraction	175
13.3.5 Lexicalized probabilistic context-free grammars	176
13.3.6 Difficulties in biomedical text mining	178
13.4 Text mining and knowledge management	179
14 Experimental design: from the basics to active learning extensions	182
14.1 Introduction	182
14.2 The elements of experimental design	182
14.2.1 Phases of biomedical DOE	183
14.2.2 Types of biological experiments	183
14.3 A decision theoretic approach to DoE	185
14.3.1 Expected value of an experiment	185
14.3.2 Adaptive designs and budgeted learning	186
14.3.3 A Bayesian treatment of sequential decision processes	188
14.4 Approaches to target variable selection	189
14.4.1 Gene Prioritization	189
14.4.2 Active learning	191
14.4.3 Other practical tasks relying on bioinformatics	191

15 Big data in biomedicine	194
15.1 Introduction	194
15.2 The first wave of biomedical big data	196
15.3 Post-genomic big data: the second wave	196
15.4 The common big data	197
15.5 The health-related common big data in biomedicine	199
15.6 Bioinformatic challenges of common big data	201
16 Analysis of heterogeneous biomedical data through information fusion	206
16.1 Introduction	206
16.2 Information fusion and data fusion	208
16.3 Types of data fusion	209
16.3.1 Early fusion	210
16.3.2 Intermediate fusion	211
16.3.3 Late fusion	211
16.4 Similarity-based data fusion	212
17 The Bayesian Encyclopedia	217
17.1 Introduction	217
17.2 The three worlds of data, knowledge and computation	220
17.3 From fragmentation problems to workflow for unification	221
17.4 Data repositories with semantic technologies	224
17.5 Semantic publishing for the literature world	224
17.6 Causal Bayesian network-based data analytic knowledge bases	227
17.7 Examples for links between worlds	228
17.8 Prospects for the Bayesian Encyclopedia	228
18 Bioinformatical workflow systems – case study	235
18.1 Overview of tasks	235
18.2 Data model and representation	236
18.3 Use cases and architecture	237
18.4 Implementation details of the server	238
18.5 Postprocessing steps	240
19 Computational aspects of pharmaceutical research	241
19.1 Overview of the process	241
19.2 Chemoinformatical background	242
19.3 Screening criteria	244
19.4 Method	247
19.5 Fragment-based design	249
19.6 Drug repositioning	250

20 Metagenomics	253
20.1 Introduction	253
20.2 Metagenome analysis	254
20.2.1 Community profiling	254
20.2.2 Functional metagenomics	255
20.3 Metagenomics step by step	255
20.3.1 Sampling	255
20.3.2 Sequencing	257
20.3.3 Assembly	257
20.3.4 Binning	258
20.3.5 Gene calling and functional inference	259