

Contents

1	Basics for Bioinformatics	1
	Xuegong Zhang, Xueya Zhou, and Xiaowo Wang	
1.1	What Is Bioinformatics	1
1.2	Some Basic Biology	2
1.2.1	Scale and Time	3
1.2.2	Cells	3
1.2.3	DNA and Chromosome	5
1.2.4	The Central Dogma	6
1.2.5	Genes and the Genome	7
1.2.6	Measurements Along the Central Dogma	10
1.2.7	DNA Sequencing	10
1.2.8	Transcriptomics and DNA Microarrays	13
1.2.9	Proteomics and Mass Spectrometry	16
1.2.10	ChIP-Chip and ChIP-Seq	17
1.3	Example Topics of Bioinformatics	18
1.3.1	Examples of Algorithmic Topics	19
1.3.2	Examples of Statistical Topics	20
1.3.3	Machine Learning and Pattern Recognition Examples	21
1.3.4	Basic Principles of Genetics	21
	References	26
2	Basic Statistics for Bioinformatics	27
	Yuanlie Lin and Rui Jiang	
2.1	Introduction	27
2.2	Foundations of Statistics	27
2.2.1	Probabilities	27
2.2.2	Random Variables	30
2.2.3	Multiple Random Variables	32
2.2.4	Distributions	34

2.2.5	Random Sampling	37
2.2.6	Sufficient Statistics	39
2.3	Point Estimation	40
2.3.1	Method of Moments	41
2.3.2	Maximum Likelihood Estimators	41
2.3.3	Bayes Estimators	42
2.3.4	Mean Squared Error	44
2.4	Hypothesis Testing	44
2.4.1	Likelihood Ratio Tests	45
2.4.2	Error Probabilities and the Power Function	46
2.4.3	<i>p</i> -Values	48
2.4.4	Some Widely Used Tests	50
2.5	Interval Estimation	52
2.6	Analysis of Variance	54
2.6.1	One-Way Analysis of Variance	55
2.6.2	Two-Way Analysis of Variance	59
2.7	Regression Models	61
2.7.1	Simple Linear Regression	62
2.7.2	Logistic Regression	65
2.8	Statistical Computing Environments	66
2.8.1	Downloading and Installation	66
2.8.2	Storage, Input, and Output of Data	67
2.8.3	Distributions	67
2.8.4	Hypothesis Testing	68
2.8.5	ANOVA and Linear Model	68
	References	68
3	Topics in Computational Genomics	69
	Michael Q. Zhang and Andrew D. Smith	
3.1	Overview: Genome Informatics	69
3.2	Finding Protein-Coding Genes	71
3.2.1	How to Identify a Coding Exon?	72
3.2.2	How to Identify a Gene with Multiple Exons?	72
3.3	Identifying Promoters	73
3.4	Genomic Arrays and aCGH/CNP Analysis	75
3.5	Introduction on Computational Analysis of Transcriptional Genomics Data	76
3.6	Modeling Regulatory Elements	77
3.6.1	Word-Based Representations	77
3.6.2	The Matrix-Based Representation	78
3.6.3	Other Representations	79
3.7	Predicting Transcription Factor Binding Sites	79
3.7.1	The Multinomial Model for Describing Sequences	80
3.7.2	Scoring Matrices and Searching Sequences	81

3.7.3	Algorithmic Techniques for Identifying High-Scoring Sites	82
3.7.4	Measuring Statistical Significance of Matches	83
3.8	Modeling Motif Enrichment in Sequences	84
3.8.1	Motif Enrichment Based on Likelihood Models	84
3.8.2	Relative Enrichment Between Two Sequence Sets	86
3.9	Phylogenetic Conservation of Regulatory Elements	88
3.9.1	Three Strategies for Identifying Conserved Binding Sites	88
3.9.2	Considerations When Using Phylogenetic Footprinting	90
3.10	Motif Discovery	91
3.10.1	Word-Based and Enumerative Methods	92
3.10.2	General Statistical Algorithms Applied to Motif Discovery	93
3.10.3	Expectation Maximization	94
3.10.4	Gibbs Sampling	95
	References	96
4	Statistical Methods in Bioinformatics	101
	Jun S. Liu and Bo Jiang	
4.1	Introduction	101
4.2	Basics of Statistical Modeling and Bayesian Inference	102
4.2.1	Bayesian Method with Examples	102
4.2.2	Dynamic Programming and Hidden Markov Model	104
4.2.3	Metropolis-Hastings Algorithm and Gibbs Sampling ..	107
4.3	Gene Expression and Microarray Analysis	109
4.3.1	Low-Level Processing and Differential Expression Identification	110
4.3.2	Unsupervised Learning	113
4.3.3	Dimension Reduction Techniques	117
4.3.4	Supervised Learning	119
4.4	Sequence Alignment	126
4.4.1	Pair-Wise Sequence Analysis	126
4.4.2	Multiple Sequence Alignment	129
4.5	Sequence Pattern Discovery	133
4.5.1	Basic Models and Approaches	133
4.5.2	Gibbs Motif Sampler	136
4.5.3	Phylogenetic Footprinting Method and the Identification of <i>Cis</i> -Regulatory Modules	138
4.6	Combining Sequence and Expression Information for Analyzing Transcription Regulation	140
4.6.1	Motif Discovery in ChIP-Array Experiment	140
4.6.2	Regression Analysis of Transcription Regulation	141
4.6.3	Regulatory Role of Histone Modification	143

4.7	Protein Structure and Proteomics	144
4.7.1	Protein Structure Prediction	145
4.7.2	Protein Chip Data Analysis	146
	References	147
5	Algorithms in Computational Biology	151
	Tao Jiang and Jianxing Feng	
5.1	Introduction	151
5.2	Dynamic Programming and Sequence Alignment	153
5.2.1	The Paradigm of Dynamic Programming	153
5.2.2	Sequence Alignment	155
5.3	Greedy Algorithms for Genome Rearrangement	157
5.3.1	Genome Rearrangements	157
5.3.2	Breakpoint Graph, Greedy Algorithm and Approximation Algorithm	159
	References	161
6	Multivariate Statistical Methods in Bioinformatics Research	163
	Lingsong Zhang and Xihong Lin	
6.1	Introduction	163
6.2	Multivariate Normal Distribution	163
6.2.1	Definition and Notation	163
6.2.2	Properties of the Multivariate Normal Distribution	164
6.2.3	Bivariate Normal Distribution	165
6.2.4	Wishart Distribution	167
6.2.5	Sample Mean and Covariance	167
6.3	One-Sample and Two-Sample Multivariate Hypothesis Tests	168
6.3.1	One-Sample t Test for a Univariate Outcome	168
6.3.2	Hotelling's T^2 Test for the Multivariate Outcome	169
6.3.3	Properties of Hotelling's T^2 Test	170
6.3.4	Paired Multivariate Hotelling's T^2 Test	171
6.3.5	Examples	172
6.3.6	Two-Sample Hotelling's T^2 Test	174
6.4	Principal Component Analysis	178
6.4.1	Definition of Principal Components	178
6.4.2	Computing Principal Components	179
6.4.3	Variance Decomposition	179
6.4.4	PCA with a Correlation Matrix	180
6.4.5	Geometric Interpretation	181
6.4.6	Choosing the Number of Principal Components	183
6.4.7	Diabetes Microarray Data	184
6.5	Factor Analysis	187
6.5.1	Orthogonal Factor Model	187
6.5.2	Estimating the Parameters	188
6.5.3	An Example	190

6.6	Linear Discriminant Analysis	193
6.6.1	Two-Group Linear Discriminant Analysis	194
6.6.2	An Example	198
6.7	Classification Methods	200
6.7.1	Introduction of Classification Methods	200
6.7.2	k -Nearest Neighbor Method	202
6.7.3	Density-Based Classification Decision Rule	205
6.7.4	Quadratic Discriminant Analysis	208
6.7.5	Logistic Regression	212
6.7.6	Support Vector Machine	214
6.8	Variable Selection	219
6.8.1	Linear Regression Model	220
6.8.2	Motivation for Variable Selection	221
6.8.3	Traditional Variable Selection Methods	222
6.8.4	Regularization and Variable Selection	223
6.8.5	Summary	231
	References	231
7	Association Analysis for Human Diseases: Methods and Examples	233
	Jurg Ott and Qingrun Zhang	
7.1	Why Do We Need Statistics?	233
7.2	Basic Concepts in Population and Quantitative Genetics	234
7.3	Genetic Linkage Analysis	236
7.4	Genetic Case-Control Association Analysis	237
7.4.1	Basic Steps in an Association Study	238
7.4.2	Multiple Testing Corrections	239
7.4.3	Multi-locus Approaches	241
7.5	Discussion	241
	References	241
8	Data Mining and Knowledge Discovery Methods with Case Examples	243
	S. Bandyopadhyay and U. Maulik	
8.1	Introduction	243
8.2	Different Tasks in Data Mining	245
8.2.1	Classification	245
8.2.2	Clustering	248
8.2.3	Discovering Associations	252
8.2.4	Issues and Challenges in Data Mining	254
8.3	Some Common Tools and Techniques	256
8.3.1	Artificial Neural Networks	256
8.3.2	Fuzzy Sets and Fuzzy Logic	258
8.3.3	Genetic Algorithms	258

8.4	Case Examples	259
8.4.1	Pixel Classification	260
8.4.2	Clustering of Satellite Images	262
8.5	Discussion and Conclusions	267
	References	267
9	Applied Bioinformatics Tools	271
	Jingchu Luo	
9.1	Introduction	271
9.1.1	Welcome	271
9.1.2	About This Web Site	273
9.1.3	Outline	274
9.1.4	Lectures	275
9.1.5	Exercises	276
9.2	Entrez	277
9.2.1	PubMed Query	277
9.2.2	Entrez Query	278
9.2.3	My NCBI	278
9.3	ExPASy	278
9.3.1	Swiss-Prot Query	278
9.3.2	Explore the Swiss-Prot Entry HBA_HUMAN	279
9.3.3	Database Query with the EBI SRS	279
9.4	Sequence Alignment	280
9.4.1	Pairwise Sequence Alignment	280
9.4.2	Multiple Sequence Alignment	281
9.4.3	BLAST	281
9.5	DNA Sequence Analysis	282
9.5.1	Gene Structure Analysis and Prediction	282
9.5.2	Sequence Composition	283
9.5.3	Secondary Structure	283
9.6	Protein Sequence Analysis	283
9.6.1	Primary Structure	283
9.6.2	Secondary Structure	283
9.6.3	Transmembrane Helices	284
9.6.4	Helical Wheel	284
9.7	Motif Search	284
9.7.1	SMART Search	284
9.7.2	MEME Search	284
9.7.3	HMM Search	285
9.7.4	Sequence Logo	285
9.8	Phylogeny	285
9.8.1	Protein	285
9.8.2	DNA	286

9.9	Projects	286
9.9.1	Sequence, Structure, and Function Analysis of the Bar-Headed Goose Hemoglobin	286
9.9.2	Exercises	287
9.10	Literature	287
9.10.1	Courses and Tutorials	287
9.10.2	Scientific Stories	288
9.10.3	Free Journals and Books	288
9.11	Bioinformatics Databases	289
9.11.1	List of Databases	289
9.11.2	Database Query Systems	289
9.11.3	Genome Databases	290
9.11.4	Sequence Databases	291
9.11.5	Protein Domain, Family, and Function Databases	292
9.11.6	Structure Databases	293
9.12	Bioinformatics Tools	294
9.12.1	List of Bioinformatics Tools at International Bioinformatics Centers	295
9.12.2	Web-Based Bioinformatics Platforms	295
9.12.3	Bioinformatics Packages to be Downloaded and Installed Locally	295
9.13	Sequence Analysis	296
9.13.1	Dotplot	296
9.13.2	Pairwise Sequence Alignment	296
9.13.3	Multiple Sequence Alignment	296
9.13.4	Motif Finding	297
9.13.5	Gene Identification	297
9.13.6	Sequence Logo	297
9.13.7	RNA Secondary Structure Prediction	297
9.14	Database Search	298
9.14.1	BLAST Search	298
9.14.2	Other Database Search	298
9.15	Molecular Modeling	299
9.15.1	Visualization and Modeling Tools	299
9.15.2	Protein Modeling Web Servers	300
9.16	Phylogenetic Analysis and Tree Construction	300
9.16.1	List of Phylogeny Programs	300
9.16.2	Online Phylogeny Servers	300
9.16.3	Phylogeny Programs	301
9.16.4	Display of Phylogenetic Trees	301
	References	301

10	Foundations for the Study of Structure and Function of Proteins	303
	Zhirong Sun	
10.1	Introduction	303
10.1.1	Importance of Protein	303
10.1.2	Amino Acids, Peptides, and Proteins.....	304
10.1.3	Some Noticeable Problems	306
10.2	Basic Concept of Protein Structure	306
10.2.1	Different Levels of Protein Structures.....	306
10.2.2	Acting Force to Sustain and Stabilize the High-Dimensional Structure of Protein	308
10.3	Fundamental of Macromolecules Structures and Functions	310
10.3.1	Different Levels of Protein Structure	310
10.3.2	Primary Structure.....	311
10.3.3	Secondary Structure	312
10.3.4	Supersecondary Structure	314
10.3.5	Folds	319
10.3.6	Summary	321
10.4	Basis of Protein Structure and Function Prediction	322
10.4.1	Overview	322
10.4.2	The Significance of Protein Structure Prediction	322
10.4.3	The Field of Machine Learning.....	323
10.4.4	Homological Protein Structure Prediction Method	331
10.4.5	Ab Initio Prediction Method	334
	Reference	336
11	Computational Systems Biology Approaches for Deciphering Traditional Chinese Medicine	337
	Shao Li and Le Lu	
11.1	Introduction	337
11.2	Disease-Related Network.....	338
11.2.1	From a Gene List to Pathway and Network.....	338
11.2.2	Construction of Disease-Related Network	340
11.2.3	Biological Network Modularity and Phenotype Network.....	346
11.3	TCM ZHENG-Related Network	349
11.3.1	"ZHENG" in TCM	350
11.3.2	A CSB-Based Case Study for TCM ZHENG	352
11.4	Network-Based Study for TCM "Fu Fang"	358
11.4.1	Systems Biology in Drug Discovery	358
11.4.2	Network-Based Drug Design	359
11.4.3	Progresses in Herbal Medicine	360
11.4.4	TCM <i>Fu Fang</i> (Herbal Formula)	361
11.4.5	A Network-Based Case Study for TCM <i>Fu Fang</i>	361
	References.....	364

12 Advanced Topics in Bioinformatics and Computational Biology	369
Bailin Hao, Chunting Zhang, Yixue Li, Hao Li, Liping Wei, Minoru Kanehisa, Luhua Lai, Runsheng Chen, Nikolaus Rajewsky, Michael Q. Zhang, Jingdong Han, Rui Jiang, Xuegong Zhang, and Yanda Li	
12.1 Prokaryote Phylogeny Meets Taxonomy	369
12.2 Z-Curve Method and Its Applications in Analyzing Eukaryotic and Prokaryotic Genomes	372
12.3 Insights into the Coupling of Duplication Events and Macroevolution from an Age Profile of Transmembrane Gene Families	374
12.4 Evolution of Combinatorial Transcriptional Circuits in the Fungal Lineage	375
12.5 Can a Non-synonymous Single-Nucleotide Polymorphism (nsSNP) Affect Protein Function? Analysis from Sequence, Structure, and Enzymatic Assay	377
12.6 Bioinformatics Methods to Integrate Genomic and Chemical Information	379
12.7 From Structure-Based to System-Based Drug Design	381
12.8 Progress in the Study of Noncoding RNAs in <i>C. elegans</i>	383
12.9 Identifying MicroRNAs and Their Targets	385
12.10 Topics in Computational Epigenomics	387
12.11 Understanding Biological Functions Through Molecular Networks	389
12.12 Identification of Network Motifs in Random Networks	390
12.13 Examples of Pattern Recognition Applications in Bioinformatics	392
12.14 Considerations in Bioinformatics	394

X. Zhang (✉), X. Zhao, X. Wang

MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Tsinghua University, Department of Automation, Tsinghua University, Beijing 100084, China
e-mail: zhangxg@sem.tsinghua.edu.cn

X. Wang et al. (eds.), *Series of Bioinformatics: Lecture Notes of the Chinese Academy of Sciences on Bioinformatics in China*, DOI 10.1007/978-1-441-91591-4_1
© Tsinghua University Press, Beijing and Springer-Verlag Berlin Heidelberg 2011