

Edited by Thomas Lengauer

 WILEY-VCH

Bioinformatics – From Genomes to Therapies

Volume 1:
Molecular Sequences and Structures



Contents

Volume 1

Preface XXV

List of Contributors XXIX

Part 1 Introduction 1

1 Bioinformatics - From Genomes to Therapies 1

Thomas Lengauer

1 Introduction 1

2 The Molecular Basis of Disease 1

3 The Molecular Approach to Curing Diseases 6

4 Finding Protein Targets 8

4.1 Genomics versus Proteomics 10

4.2 Extent of Information Available on the Genes/Proteins 11

5 Developing Drugs 12

6 Optimizing Therapies 14

7 Organization of the Book 15

References 23

Part 2 Sequencing Genomes 25

2 Bioinformatics Support for Genome-Sequencing Projects 25

Knut Reinert and Daniel Huson

1 Introduction 25

2 Assembly Strategies for Large Genomes 25

2.1 Introduction 25

2.2 Properties of the Data 29

2.2.1 Reads, Mate-pairs and Quality Values 29

2.2.2 Physical Maps 30

2.3	Assembly strategies	31
3	Algorithmic Problems and their Treatment	33
3.1	Overlap Comparison of all Reads	34
3.2	Contig Phase: Layout of Reads	37
3.3	Error Correction and Resolving Repeats	40
3.4	Layout of Contigs	42
3.5	Computation of the Consensus Sequences	45
4	Examples of Existing Assemblers	47
4.1	The Celera Assembler	47
4.2	The GigAssembler	48
4.3	The ARACHNE Assembler	48
4.4	The JAZZ Assembler	49
4.5	The RePS Sssembler	49
4.6	The Barnacle Assembler	49
4.7	The PCAP Assembler	50
4.8	The Phusion Assembler	50
4.9	The Atlas Assembler	51
4.10	Other Assemblers	52
5	Conclusion	52
	References	53
Part 3	Sequence Analysis	57
3	Sequence Alignment and Sequence Database Search	57
	<i>Martin Vingron</i>	
1	Introduction	57
2	Pairwise Sequence Comparison	58
2.1	Dot plots	58
2.2	Sequence Alignment	60
3	Database Searching I: Single-sequence Heuristic Algorithms	65
4	Alignment and Search Statistics	68
5	Multiple Sequence Alignment	71
6	Multiple Alignments, HMMs and Database Searching II	74
7	Protein Families and Protein Domains	78
8	Conclusions	79
	References	79
4	Phylogeny Reconstruction	83
	<i>Ingo Ebersberger, Arndt von Haeseler and Heiko A. Schmidt</i>	
1	Introduction	83
1.1	Reconstructing a Tree from its Leaves	84
1.2	Phylogenetic Relationships of Taxa and their Characters	85

1.2.1	The Problem of Character Inconsistencies	86
1.2.2	Finding the Appropriate Character Set	87
2	Modeling DNA Sequence Evolution	88
2.1	Nucleotide Substitution Models	90
2.2	Modeling Rate Heterogeneity	90
2.3	Codon Models	91
3	Tracing the Evolutionary Signal	92
3.1	The Parsimony Principle of Evolution	93
3.1.1	Generalized Parsimony	94
3.1.2	Multiple/Parallel Hits	95
3.2	Distance-based Methods	95
3.2.1	UPGMA	95
3.2.2	Neighbors-relation Methods	96
3.2.3	Neighbor-joining Method	97
3.2.4	Least-squares Methods	98
3.3	The Criterion of Likelihood	98
3.4	Calculating the Likelihood of a Tree	99
3.5	Bayesian Statistics in Phylogenetic Analysis	99
3.6	Rooting Trees/Molecular Clock	101
3.6.1	Outgroup Rooting	101
3.6.2	Midpoint Rooting and Molecular Clock	102
4	Finding the Optimal Tree	103
4.1	Exhaustive Search Methods	103
4.2	Heuristic Search Methods	104
4.2.1	Hill Climbing and the Problem of Local Optimization	105
4.2.2	Modeling Tree Quality	108
4.2.3	Heuristics for Large Datasets	108
5	The Advent of Phylogenomics	109
5.1	Multilocus Datasets	109
5.2	Combining Incomplete Multilocus Datasets: Supertrees and their Methods	112
5.2.1	Agreement Supertrees	112
5.2.2	Optimization Supertrees	114
5.2.3	The Supertrees/Consensus versus Total Evidence Debate	115
5.2.4	Medium-level Combination	115
6	Phylogenetic Network Methods	116
6.1	From Trees to Split Networks	116
6.1.1	Split Systems and their Visualization	116
6.1.2	Constructing Split Systems from Trees	118
6.1.3	Constructing Split Systems from Sequence Data	118
6.2	Reconstructing Reticulate Evolution and Further Analyses	119
	References	121

- 5 Finding Protein-coding Genes 129**
David C. Kulp
 - 1 Introduction 129
 - 2 Basic DNA Terminology 129
 - 3 Detecting Coding Sequences 131
 - 3.1 Reading Frames 132
 - 3.2 Coding Potential 132
 - 4 Gene Contents 135
 - 5 Gene Signals 137
 - 5.1 Splice Sites 137
 - 5.2 Translation Initiation 140
 - 5.3 Translation and Transcription Termination 140
 - 6 Integrating Gene Features 141
 - 6.1 Combining Local Features 141
 - 6.2 Dynamic Programming 142
 - 6.3 Gene Grammars 143
 - 7 Performance Comparisons 145
 - 8 Using Homology 147
 - 8.1 cDNA Clustering and Alignments 147
 - 8.2 Orthologous DNA 150
 - 8.3 Protein Homology 152
 - 8.4 Integrative Methods 153
 - 9 Pitfalls: Pseudogenes, Splice Variants and the Cruel Biological Reality 153
 - 10 Further Reading 154
 - References 155

- 6 Analyzing Regulatory Regions in Genomes 159**
Thomas Werner
 - 1 General Features of Regulatory Regions in Eukaryotic Genomes 159
 - 1.1 General Functions of Regulatory Regions 159
 - 1.2 Most Important Elements in Regulatory Regions 160
 - 1.3 TFBSs 160
 - 1.4 Sequence Features 161
 - 1.5 Structural Elements 161
 - 1.6 Organizational Principles of Regulatory Regions 162
 - 1.6.1 Overall Structure of Pol II Promoters 162
 - 1.6.2 TFBS in Promoters 162
 - 1.6.3 Module Properties of the Core Promoter 163
 - 1.7 Bioinformatics Models for the Analysis and Detection of Regulatory Regions 168

1.8	Statistical Models	168
1.8.1	Mixed Models	168
1.8.2	Organizational Models	169
2	Methods for Element Detection	169
2.1	Detection of TFBSs	169
2.2	Detection of Novel TFBS Motifs	171
2.3	Detection of Structural Elements	172
2.4	Assessment of Other Elements	172
3	Analysis of Regulatory Regions	173
3.1	Comparative Sequence Analysis	173
3.2	Training Set Selection	173
3.3	Statistical and Biological Significance	174
3.4	Context Dependency	174
4	Methods for Detection of Regulatory Regions	175
4.1	Scaffold/Matrix Attachment Regions (S/MARs)	176
4.2	Enhancers/Silencers	177
4.3	Promoters	177
4.4	Programs for Recognition of Regulatory Sequences	177
4.4.1	Programs Based on Statistical Models (General Promoter Prediction)	178
4.4.2	Programs Utilizing Mixed Models	179
4.4.3	Programs Based on Specific Promoter Recognition	179
4.4.4	Early Attempts at Promoter Prediction	181
5	Annotation of Large Genomic Sequences	182
5.1	Balance between Sensitivity and Specificity	182
5.2	Genes – Transcripts – Promoters	183
5.3	Sources for Finding Alternative Transcripts and Promoters	185
5.4	Comparative Genomics of Promoters	185
6	Genome-wide Analysis of Transcription Control	186
6.1	Context-specific Transcripts and Pathways	187
6.2	Consequences for Microarray Analysis	187
7	Conclusions	189
	References	190
7	Finding Repeats in Genome Sequences	197
	<i>Brian J. Haas and Steven L. Salzberg</i>	
1	Introduction	197
2	Algorithms and Tools for Mining Repeats	199
2.1	Finding Intra- and Inter-sequence Repeats as Pairwise Alignments	200
2.2	Miropeats (alias Printrepeats)	201
2.3	REPuter	202

2.4	RepeatFinder	206
2.5	RECON	207
2.6	PILER	209
2.7	RepeatScout	212
3	Tandem Repeats	215
3.1	TRF	216
3.2	STRING (Search for Tandem Repeats IN Genomes)	218
3.3	MREPS	219
4	Repeats and Genome Assembly Algorithms	220
4.1	Repeat Management in the Celera Assembler and other Assemblers	221
4.2	Repeat Identification by k -mer Counts	221
4.3	Repeat Identification by Depth of Coverage (Arrival Rates)	222
4.4	Repeat Identification by Conflicting Links	223
4.5	Repeat Placement: Rocks and Stones	223
4.6	Repeat Placement: Surrogates	223
4.7	Repeat Resolution in Euler	224
5	Untangling the Mosaic Nature of Repeats (The A-Bruijn Graph)	225
6	Repeat Annotation in Genomes	227
	References	230
8	Analyzing Genome Rearrangements	235
	<i>Guillaume Bourque</i>	
1	Introduction	235
2	Basic Concepts	236
2.1	Genome Representation	236
2.1.1	Circular, Linear and Multichromosomal Genomes	237
2.1.2	Unsigned Genomes	238
2.1.3	Unequal Gene Content	238
2.1.4	Homology Markers	238
2.2	Types of Genome Rearrangements	239
3	Distance between Two Genomes	240
3.1	Breakpoint Distance	240
3.2	Rearrangement Distance	241
3.2.1	HP Theory	242
3.3	Conservation Distance	244
3.3.1	Common Intervals	244
3.3.2	Conserved Intervals	245
4	Genome Rearrangement Phylogenies	245
4.1	Distance-based Methods	246
4.2	Maximum Parsimony Methods	247

4.3	Maximum Likelihood Methods	248
5	Recent Applications	249
5.1	Rearrangements in Large Genomes	249
5.2	Genomes Rearrangements and Cancer	252
6	Conclusion	253
6.1	Challenges	253
6.2	Promising New Approaches	255
	References	256
Part 4	Molecular Structure Prediction	261
9	Predicting Simplified Features of Protein Structure	261
	<i>Dariusz Przybylski and Burkhard Rost</i>	
1	Introduction	261
1.1	Protein Structures are Determined Much Slower than Sequences	261
1.2	Reliable and Comprehensive Computations of 3-D Structures are not yet Possible	261
1.3	Predictions of Simplified Aspects of 3-D Structure are often very Successful	262
2	Secondary Structure Prediction	262
2.1	Assignment of Secondary from 3-D Structure	262
2.1.1	Regular Secondary Structure Formation is Mostly a Local Process	262
2.1.2	Secondary Structures can be Somehow Flexible	263
2.1.3	Automatic Assignments of Secondary Structure	263
2.1.4	Reduction to Three Secondary Structure States	264
2.2	Measuring Performance	265
2.2.1	Performance has Many Aspects Relating to Many Different Measures	265
2.2.2	Per-residue Percentage Accuracy: Q_K	266
2.2.3	Per-residue Confusion between Regular Elements: <i>BAD</i>	266
2.2.4	Per-segment Prediction Accuracy: <i>SOV</i>	266
2.3	Comparing Different Methods	267
2.3.1	Generic Problems	267
2.3.2	Numbers can often not be Compared between Two Different Publications	267
2.3.3	Appropriate Comparisons of Methods Require Large, "Blind" Data Sets	268
2.4	History	269
2.4.1	First Generation: Single-residue Statistics	269
2.4.2	Second Generation: Segment Statistics	269

- 2.4.3 Third Generation: Evolutionary Information 269
- 2.4.4 Recent Improvements of Third-generation Methods 271
- 2.4.5 Meta-predictors Improve Somehow 272
- 2.5 State-of-the-art Performance 272
 - 2.5.1 Average Predictions Have Good Quality 272
 - 2.5.2 Prediction Accuracy Varies among Proteins 273
 - 2.5.3 Reliability of Prediction Correlates with Accuracy 273
 - 2.5.4 Understandable Why Certain Proteins Predicted Poorly? 274
- 2.6 Applications 274
 - 2.6.1 Better Database Searches 274
 - 2.6.2 One-dimensional Predictions Assist in the Prediction of Higher-dimensional Structure 275
 - 2.6.3 Predicted Secondary Structure Helps Annotating Function 275
 - 2.6.4 Secondary Structure-based Classifications in the Context of Genome Analysis 276
 - 2.6.5 Regions Likely to Undergo Structural Change Predicted Successfully 276
- 2.7 Things to Remember when using Predictions 277
 - 2.7.1 Special Classes of Proteins 277
 - 2.7.2 Better Alignments Yield Better Predictions 277
- 2.8 Resources 277
 - 2.8.1 Internet Services are Widely Available 277
 - 2.8.2 Interactive Services 277
 - 2.8.3 Servers 278
- 3 Transmembrane Regions 278
 - 3.1 Transmembrane Proteins are an Extremely Important Class of Proteins 278
 - 3.2 Prediction Methods 279
 - 3.3 Performance 279
 - 3.4 Servers 280
- 4 Solvent Accessibility 280
 - 4.1 Solvent Accessibility Somehow Distinguishes Structurally Important from Functionally Important 280
 - 4.2 Measuring Solvent Accessibility 280
 - 4.3 Best Methods Combine Evolutionary Information with Machine Learning 281
 - 4.4 Performance 282
 - 4.5 Servers 282
- 5 Inter-residue Contacts 282
 - 5.1 Two-dimensional Predictions may be a Step Toward 3-D Structures 282
 - 5.2 Measuring Performance 282

5.3	Prediction Methods	283
5.4	Performance and Applications	283
5.5	Servers	283
6	Flexible and Intrinsically Disordered Regions	284
6.1	Local Mobility, Rigidity and Disorder all are Features that Relate to Function	284
6.2	Measuring Flexibility and Disorder	284
6.3	Prediction Methods	284
6.4	Servers	285
7	Protein Domains	285
7.1	Independent Folding Units	285
7.2	Prediction Methods	285
7.3	Servers	286
	References	286
10	Homology Modeling in Biology and Medicine	297
	<i>Roland L. Dunbrack, Jr.</i>	
1	Introduction	297
1.1	The Concept of Homology Modeling	297
1.2	How do Homologous Protein Arise?	298
1.3	The Purposes of Homology Modeling	299
1.4	The Effect of the Genome Projects	301
2	Input Data	303
3	Methods	307
3.1	<i>Modeling at Different Levels of Complexity</i>	307
3.2	Side-chain Modeling	309
3.2.1	Input Information	309
3.2.2	Rotamers and Rotamer Libraries	311
3.2.3	Side-chain Prediction Methods	312
3.2.4	Available Programs for Side-chain Prediction	317
3.3	Loop Modeling	317
3.3.1	Input Information	317
3.3.2	<i>Loop Conformational Analysis</i>	318
3.3.3	Loop Prediction Methods	320
3.3.4	Available Programs	321
3.4	Methods for Complete Modeling	322
3.4.1	MODELLER	322
3.4.2	MolIDE: A Graphical User Interface for Modeling	323
3.4.3	RAMP and PROTINFO	323
3.4.4	SWISS-MODEL	323
4	Results	324
4.1	Range of Targets	324

4.2	Example: Protein Kinase STK11/LKB1	324
4.3	The Importance of Protein Interactions	331
5	Strengths and Limitations	334
6	Validation	335
6.1	The CASP Meeting	336
6.2	Protein Health	336
	References	337
11	Protein Fold Recognition Based on Distant Homologs	351
	<i>Ingolf Sommer</i>	
1	Introduction	351
2	Overview of Template-based Modeling	352
2.1	<i>Key Steps in Template-based Modeling</i>	<i>352</i>
2.1.1	Identifying Templates	352
2.1.2	Assessing Significance	353
2.1.3	Model Building	353
2.1.4	Evaluation	354
2.2	Template Databases	354
3	Sequence-based Methods for Identifying Templates	356
3.1	Sequence–Sequence Comparison Methods	356
3.2	Frequency Profile Methods	357
3.2.1	Definition of a Frequency Profile and PSSM	357
3.2.2	Generating Frequency Profiles	359
3.2.3	Scoring Frequency Profiles	360
3.2.4	Scoring Profiles Against Sequences	360
3.2.5	Scoring Profiles against Profiles	361
3.3	Hidden Markov Models (HMMs)	363
3.3.1	Definition	363
3.3.2	Profile HMM Technology	364
3.3.3	HMMs in Fold Recognition	365
3.3.4	HMM–HMM Comparisons	365
3.4	Support Vector Machines (SVMs)	365
3.4.1	Definition	365
3.4.2	Various Kernels	366
3.4.3	Experimental Assessment	366
4	Structure-based Methods for Identifying Templates	367
4.1	Boltzmann’s Principle and Knowledge-based Potentials	368
4.2	Threading Using Pair-interaction Potentials	369
4.3	Threading using Frozen Approximation Algorithms	371
5	Hybrid Methods and Recent Developments	372
5.1	Using Different Sources of Information	372

5.1.1	Incorporating Secondary Structure Prediction into Frequency Profiles and HMMs	372
5.1.2	Intrinsically Disordered Regions in Proteins	373
5.1.3	Incorporating 3-D Structure into Frequency Profiles	374
5.2	Combining Information	374
5.3	Meta-servers	375
6	Assessment of Models	376
6.1	Estimating Significance of Sequence Hits	376
6.2	Scoring 3-D Model Quality: Model Quality Assessment Programs (MQAPs)	377
6.3	Evaluation of Protein Structure Prediction: Critical Assessment of Techniques for Protein Structure Prediction	378
7	Programs and Web Resources	379
	References	380
12	<i>De Novo</i> Structure Prediction: Methods and Applications	389
	<i>Richard Bonneau</i>	
1	Introduction	389
1.1	Scope of this Review and Definition of <i>De Novo</i> Structure Prediction	389
1.2	The Role of Structure Prediction in Biology	390
1.3	<i>De novo</i> Structure Prediction in a Genome Annotation Context, Synergy with Other Methods	391
2	Core Features of Current Methods of <i>De Novo</i> Structure Prediction	393
2.1	Rosetta <i>De Novo</i>	393
2.2	Evaluation of Structure Predictions	396
2.3	Domain Prediction is Key	399
2.4	Local Structure Prediction and Reduced Complexity Models are Central to Current <i>De Novo</i> Methods	403
2.5	Clustering as a Heuristic Approach to Approximating Entropic Determinants of Protein Folding	405
2.6	Balancing Resolution with Sampling, Prospects for Improved Accuracy and Atomic Detail	406
3	Applying Structure Prediction: <i>De Novo</i> Structure Prediction in a Systems Biology Context	408
3.1	Structure Prediction as a Road to Function	408
3.2	Initial Application of <i>De Novo</i> Structure Prediction	409
3.3	Application on Genome-wide Scale and Examples of Data Integration	410

3.4	Scaling-up <i>De Novo</i> Structure Prediction: Rosetta on the World Community Grid 412
4	Future Directions 412
4.1	Structure Prediction and Systems Biology: Data Integration 412
4.2	Need for Improved Accuracy and Extending the Reach of <i>De Novo</i> Methods 413
	References 413
13	Structural Genomics 419 <i>Philip E. Bourne and Adam Godzik</i>
1	Overview 419
1.1	What is Structural Genomics? 419
1.2	What are the Motivators? 419
1.2.1	Fold Coverage as a Motivator 420
1.2.2	Structural Coverage of an Organism as a Motivator 424
1.2.3	Structure Coverage of Central Metabolism Pathways as a Motivator 424
1.2.4	Disease as a Motivator 425
1.3	How Does Structural Genomics Relate to Conventional Structural Biology? 425
2	Methodology 427
2.1	Target Selection 427
2.2	Crystallomics 428
2.3	Data Collection 429
2.4	Structure Solution 430
2.5	Structure Refinement 431
2.6	PDB Deposition 431
2.7	Functional Annotation 432
2.7.1	Biological Multimeric State 432
2.7.2	Active-site Determination 432
2.8	Publishing 433
3	Results – Number and Characteristics of Structures Determined 434
4	Discussion 435
4.1	Follow-up Studies 435
4.2	Examples of Functional Discoveries 436
5	The Future 436
	References 436
14	RNA Secondary Structures 439 <i>Ivo L. Hofacker and Peter F. Stadler</i>
1	Secondary Structure Graphs 439

1.1	Introduction	439
1.2	Secondary Structure Graphs	440
1.3	Mountain Plots and Dot Plots	443
1.4	Trees and Forests	443
1.5	Notes	444
2	Loop-based Energy Model	444
2.1	Loop Decomposition	444
2.2	Energy Parameters	445
2.3	Notes	447
3	The Problem of RNA Folding	447
3.1	Counting Structures and Maximizing Base Pairs	447
3.2	Backtracing	449
3.3	Energy Minimization in the Loop-based Energy Model	450
3.4	RNA Hybridization	453
3.5	Pseudoknotted Structures	454
3.6	Notes	454
4	Conserved Structures, Consensus Structures and RNA Gene Finding	456
4.1	The Phylogenetic Method	456
4.2	Conserved Structures	457
4.3	Consensus Structures	459
4.4	RNA Gene Finding	460
4.5	Notes	463
5	Grammars for RNA Structures	463
5.1	Context-free Grammars (CFGs) and RNA Secondary Structures	463
5.2	Cocke–Younger–Kasami (CYK) Algorithm	465
5.3	Inside and Outside Algorithms	465
5.4	Parameter Estimation	466
5.5	Algebraic Dynamic Programming	466
5.6	Notes	467
6	Comparison of Secondary Structures	468
6.1	String-based Alignments	469
6.2	Tree Editing	469
6.3	Tree Alignments	472
6.4	The Sankoff Algorithm and Variants	475
6.5	Multiple Alignments	475
6.6	Notes	476
7	Kinetic Folding	476
7.1	Folding Energy Landscapes	476
7.2	Kinetic Folding Algorithms	477
7.3	Approximate Folding Trajectories and Barrier Trees	478

7.4	RNA Switches	480
7.5	Notes	481
8	Concluding Remarks	481
	References	482
15	RNA Tertiary Structure Prediction	491
	<i>François Major and Philippe Thibault</i>	
1	Introduction	491
2	Annotation	493
2.1	Nucleotide Conformations	494
2.2	Nucleotide Interactions	501
2.2.1	Base Stacking	502
2.2.2	Base Pairing	505
2.2.3	Isosteric Base Pairs	508
3	Motif Discovery	508
3.1	RNA Motifs	509
3.1.1	Classical Examples	509
3.2	Catalytic Motifs	513
3.3	Transport and Localization	519
4	Modeling	521
4.1	The CSP	522
4.2	MC-Sym	524
4.2.1	Backbone Optimization	527
4.2.2	Probabilistic Backtracking	529
4.2.3	“Divide and Conquer”	529
4.3	MC-Sym at Work	530
4.3.1	Modeling a Yeast tRNA-Phe Stem-Loop	532
4.3.2	Modeling a Pseudoknot	533
4.3.3	Cycles of Interactions	535
5	Perspectives	535
	References	536

Volume 2

Part 5 Analysis of Molecular Interactions 541

16 Docking and Scoring for Structure-based Drug Design 541

Matthias Rarey, Jörg Degen and Ingo Reulecke

1 Introduction 541

1.1 A Taxonomy of Docking Problems 543

1.2 Application Scenarios in Structure-based Molecular Design 544

2 Scoring Protein–Ligand Complexes 546

2.1 Modeling Protein–Ligand Interactions 546

2.2 Scoring Functions based on Force Fields 548

2.3 Empirical Scoring 550

2.4 Knowledge-based Scoring 551

2.5	Evaluation	551
3	Methods for Protein–Ligand Docking	552
3.1	Rigid-body Docking Algorithms	552
3.1.1	Approaches based on Clique Search	553
3.1.2	Geometric Hashing	554
3.1.3	Pose Clustering	555
3.1.4	Fast Shape Comparison	557
3.2	Flexible Ligand-docking Algorithms	558
3.2.1	Conformation Ensembles	558
3.2.2	Flexible Docking based on Fragmentation	559
3.2.2.1	“Place & Join” Algorithms	559
3.2.2.2	Incremental Construction Algorithms	560
3.2.3	Genetic Algorithms and Evolutionary Programming	563
3.2.4	Distance Geometry	565
3.2.5	Random Search	565
3.3	Docking by Simulation	566
3.3.1	Simulated Annealing	566
3.3.2	MD Simulations	567
3.3.3	MC Algorithms	568
3.3.4	Hybrid Methods	570
4	Structure-based Virtual Screening	570
4.1	Considering Pharmacophoric Constraints	571
4.2	Docking of Combinatorial Libraries	571
4.3	Database Approaches	573
5	From Molecules to Fragment Spaces: Structure-based <i>De Novo</i> Design	574
5.1	Modeling Fragment Spaces	575
5.2	<i>De Novo</i> Design Algorithms	575
5.2.1	Rigid-body Algorithms	576
5.2.2	Simulation Methods	576
5.2.3	“Place & Join” Algorithms	577
5.2.4	Sequential Growth Algorithms	578
5.2.5	Genetic Algorithms and Evolutionary Programming	579
5.3	Synthetic Accessibility	580
5.3.1	Fragment Selection	580
5.3.2	Virtual Synthesis	581
5.3.3	Compound Analysis	581
6	Structure-based Drug Design at Work: Validation Studies and Applications	582
7	Concluding Remarks	583
	References	584

17	Modeling Protein–Protein and Protein–DNA Docking	601
	<i>Andreas Hildebrandt, Oliver Kohlbacher and Hans-Peter Lenhof</i>	
1	Introduction	601
2	Protein–Protein Interactions	603
2.1	Basic Concepts of Docking	603
2.2	Rigid Body Docking	606
2.2.1	Correlation Techniques	606
2.2.2	Graph-based Structure Generation Methods	610
2.2.3	Slice Decomposition and Polygon Descriptors	612
2.2.4	Critical Surface Points and Geometric Hashing	614
2.2.5	Other Approaches	615
2.3	Realizing Protein Flexibility	615
2.3.1	Side Chain Placement	617
2.3.1.1	Dead End Elimination	618
2.3.1.2	“Branch & Bound” and the A* Algorithm	619
2.3.1.3	Integer Linear Programming	621
2.3.2	Hinge-bending	624
2.3.3	Biased Probability Monte Carlo (BPMC) Conformational Search	626
2.4	Scoring Functions	627
2.4.1	Empirical Potentials	628
2.4.2	Knowledge-based Potentials	630
2.5	Data-driven Docking	632
2.5.1	Experimental Techniques	632
2.5.2	Algorithmic Approaches	633
2.6	Assessment of Docking Predictions	634
3	Protein–DNA Interactions	638
3.1	Peculiarities of Protein–DNA Binding	638
3.2	Algorithmic Techniques	639
3.2.1	Correlation Techniques	639
3.2.2	Monte Carlo Techniques	640
3.3	Scoring Functions	641
4	Conclusion	642
	References	644
18	Lead Identification by Virtual Screening	651
	<i>Andreas Kämper, Didier Rognan and Thomas Lengauer</i>	
1	Introduction	651
1.1	Screening Techniques	652
1.2	Drug Discovery Process	653
1.3	Compound Collections	654
2	Filtering and Preparation of Ligands	655

2.1	Library Preprocessing	656
2.2	Bioavailability	658
2.3	Drug-likeness	659
2.4	Molecular Diversity	660
3	Ligand-based VS	662
3.1	Descriptor-based Similarity Measures	664
3.2	Bit String Descriptors	665
3.3	Feature Trees	666
3.4	Molecular Superimposition Approaches	667
3.5	Pharmacophore Searches	669
3.6	QSARs	670
3.7	Other Techniques	672
4	Postprocessing of Hitlists	672
4.1	Data Mining	673
4.2	Analysis of the Protein–Ligand Interface	674
4.3	Consensus Techniques	675
4.4	Visualization	676
5	Critical Evaluation of Structure-based VS	677
5.1	Influence of Parameter Settings	677
5.1.1	Which Library?	677
5.1.2	Which Ligand Conformation(s)?	678
5.1.3	Which Protein Coordinates?	678
5.1.4	Which Docking Tool?	678
5.1.5	Which Scoring Function?	679
5.1.6	Which Postprocessing?	680
5.2	Recent Success Stories	681
5.2.1	Some Privileged Targets	681
5.2.2	First-in-class Compounds	684
5.2.3	Fragment Screening	685
5.2.4	Lead Optimization	686
5.2.5	Homology Models as VS Targets	686
5.3	Concluding Remarks	687
6	Critical Evaluation of Ligand-based VS	687
6.1	Influence of Parameter Settings	687
6.2	Recent Success Stories	688
6.3	Comparison of Structure- and Ligand-based Techniques	691
6.4	Concluding Remarks	692
	References	693

19	Efficient Strategies for Lead Optimization by Simultaneously Addressing Affinity, Selectivity and Pharmacokinetic Parameters	705
	<i>Karl-Heinz Baringhaus and Hans Matter</i>	
1	Introduction	705
2	The Origin of Lead Structures	708
3	Optimization for Affinity and Selectivity	711
3.1	Lead Optimization as a Challenge in Drug Discovery	711
3.2	Use and Limitation of Structure-based Design Approaches	712
3.3	Integration of Ligand- and Structure-based Design Concepts	713
3.4	The Selectivity Challenge from the Ligand's Perspective	716
3.5	Selectivity Approaches Considering Binding Site Topologies	717
4	Addressing Pharmacokinetic Problems	721
4.1	Prediction of Physicochemical Properties	721
4.2	Prediction of ADME Properties	722
4.3	Prediction of Toxicity	724
4.4	Physicochemical and ADMET Property-based Design	724
5	ADME/Antitarget Models for Lead Optimization	724
5.1	Global ADME Models for Intestinal Absorption and Protein Binding	724
5.2	Selected Examples to Address ADME/Toxicology Antitargets	728
6	Integrated Approach	732
6.1	Strategy and Risk Assessment	732
6.2	Integration	734
6.3	Literature and Aventis Examples on Aspects of Multidimensional Optimization	735
7	Conclusions	742
	References	743
Part 6	Molecular Networks	755
20	Modeling and Simulating Metabolic Networks	755
	<i>Stefan Schuster and David Fell</i>	
1	Introduction	755
2	Fundamentals	756
2.1	Motivation	756
2.2	Stoichiometry	757
2.3	Balance Equations	759
2.4	Enzyme Kinetics	760
3	Network Analysis	762
3.1	Conservation Relations	762
3.2	Stationary States and Stability Analysis	764

3.3	Constraints on Steady-state Fluxes	766
3.4	Defining Component Pathways of a Network	769
3.5	Examples of Elementary-modes Analysis	771
3.6	Extreme Pathways	777
3.7	Optimization of Molar Yields and Flux Balance Analysis (FBA)	778
3.8	Analyzing the Robustness of Metabolism	781
4	Dynamic Simulation	782
4.1	How is a Dynamic Model Constructed?	782
4.2	Metabolic Databases	788
4.3	Example: Red Blood Cell Metabolism	790
4.4	Oscillations	792
4.5	Whole-cell Modeling	794
5	Conclusions	797
	References	798
21	Inferring Gene Regulatory Networks	807
	<i>Michael Q. Zhang</i>	
1	Introduction	807
2	Gene Regulation at the Transcriptional Level	808
2.1	Finding TFBSs and Motifs	809
2.2	Identifying Target Genes	809
2.3	Discovering Novel Motifs and Target Genes	810
2.4	Inferring GRN Modules and Integrating Diverse Types of Data	812
3	Gene Regulation at the Level of RNA Processing	813
3.1	Identification of Splicing Enhancers and Silencers	814
3.2	Splicing Microarrays	814
4	Gene Regulation at the Translational Level	815
5	Gene Regulation by Small ncRNAs	816
6	GRNs in Development and Evolution	817
	References	819
22	Modeling Cell Signaling Networks	829
	<i>Anthony Hasseldine, Azi Lipshtat, Ravi Iyengar and Avi Ma'ayan</i>	
1	Introduction	829
1.1	Components and Cascades	829
1.2	From Pathways to Networks	832
1.2.1	Interactions between Signaling Pathways	832
1.2.2	Implications of Network Topology	834
1.2.3	Network Motifs	835
2	Types of Models and the Information they can Yield	839
2.1	Boolean Networks and Bayesian Networks Modeling Approaches	839

2.2	Quantitative Dynamics Modeling	841
2.2.1	Deterministic Models	843
2.2.2	Stochastic Models	846
2.2.3	Hybrid Models	849
3	Identifying Parameters/Data Sets for Modeling	850
3.1	Functionally Relevant Connections	850
3.2	Qualitative Relationships	850
3.3	Quantitative Specifications	851
4	Model Validation	853
4.1	Parameter Variation and Sensitivity Analysis	853
4.2	Constraints and Predictions	854
5	Perspective	855
	References	858
23	Dynamics of Virus–Host Cell Interaction	861
	<i>Udo Reichl and Yury Sidorenko</i>	
1	Introduction	861
2	Viral Infection of Cells	863
2.1	Viral Infection of Prokaryotic Cells	864
2.2	Viral Infection of Eukaryotic Cells	866
3	Mathematical Models of Virus Dynamics	868
3.1	Unstructured Models of Virus Dynamics	869
3.2	Structured Models of Virus Dynamics	871
4	Influenza Virus as an Example for Virus–Host Cell Interaction	872
4.1	The Influenza A Virus Life Cycle	873
4.2	Mathematical Model of the Influenza A Virus Life Cycle	877
4.3	Influenza A Virus Growth Dynamics	880
4.4	Discussion and Outlook	886
5	Conclusions	887
	References	892
Part 7	Analysis of Expression Data	899
24	DNA Microarray Technology and Applications – An Overview	899
	<i>John Quackenbush</i>	
1	Introduction to DNA Microarrays	899
2	Microarrays and Clinical Applications	899
3	Microarray Data Collection, Transformation and Representation	902
4	Identifying Patterns of Expression	905
5	Class Discovery	906
5.1	Hierarchical Clustering	906

5.2	<i>k</i> -means Clustering	907
5.3	Other Unsupervised Approaches	911
6	Classification	911
6.1	<i>k</i> NN Classification	912
7	Validation	914
8	Sample Selection and Classification	915
9	Limitations and Success of Classification	915
10	Data Reporting and Comparisons	916
11	Meta-analysis	919
12	The Path Forward	922
	References	923
25	Low-level Analysis of Microarray Experiments	929
	<i>Wolfgang Huber, Anja von Heydebreck and Martin Vingron</i>	
1	Introduction	929
1.1	Microarray technology	929
1.2	Prerequisites	930
1.3	Preprocessing	931
2	Visualization and Exploration of the Raw Data	932
2.1	Image Analysis	932
2.2	Dynamic Range and Spatial Effects	933
2.3	Scatterplot	934
2.4	Batch Effects	938
2.5	Along Chromosome Plots	941
2.6	Sensitivity and Specificity of Probes	942
3	Error Models	943
3.1	Motivation	943
3.1.1	Obtaining Optimal Estimates	943
3.1.2	Biological Inference	944
3.1.3	Quality Control	944
3.2	The Additive–Multiplicative Error Model	944
3.2.1	Induction from Data	944
3.2.2	A Theoretical Deduction	946
4	Normalization	947
5	Detection of Differentially Expressed Genes	949
5.1	Stepwise versus Integrated Approaches	949
5.2	Measures of Differential Expression: The Variance Bias Trade-off	950
5.3	Identifying Differentially Expressed Genes from Replicated Measurements	951
6	Software	953
	References	954

26	Classification of Patients	957
	<i>Claudio Lottaz, Dennis Kostka and Rainer Spang</i>	
1	Introduction	957
2	Molecular Diagnosis	958
2.1	Problem Statement	958
2.1.1	Notation	959
2.1.2	Loss and Risk	960
2.1.3	Bayes Classifier and Bayes Error	960
2.1.4	Minimal Empirical Risk and Maximum Likelihood	961
2.1.5	Regularized Risk and Priors	961
2.2	Supervised Classification	963
2.2.1	Discriminant Analysis and Feature Selection	964
2.2.2	Penalized Logistic Regression	965
2.2.3	Support Vector Classification	966
2.2.4	Bagging	967
2.2.5	Boosting	968
2.3	Gene Selection	968
2.3.1	Filter Approaches	969
2.3.2	Wrapper Approaches	969
2.4	Adaptive Model Selection and Validation	970
2.4.1	Adaptive Model Selection	970
2.4.1.1	Bias-variance Trade-off	970
2.4.1.2	Choosing a Trade-off via the Hold Out	971
2.4.1.3	Using Data More Efficiently via Cross-Validation	972
2.4.2	Validation of the Predictive Performance of a Molecular Signature	972
2.4.2.1	Estimating Error Rates	973
2.4.2.2	Selection Bias and Nested Loop Cross-validation	974
2.5	Discussion	975
3	Finding Molecular Disease Entities	975
3.1	Clustering	976
3.1.1	Clustering Algorithms	976
3.1.2	The Problem of Distances	977
3.2	Searching for Partitionings	978
3.2.1	Overlapping Partitionings	978
3.2.2	Search and Find	978
3.2.3	ISIS – Identifying Splits with Clear Separation	978
3.2.4	Overabundance of Differential Genes	980
3.2.5	Best-fitting Gaussian Model	980
3.3	Biclustering	980
3.4	Semisupervised Methods	981
3.4.1	Molecular Symptoms	981

3.4.2	Survival-driven Class-finding	981
3.4.3	Towards Survival Prediction	983
3.5	Validating Unsupervised Analysis	983
3.5.1	Statistical Significance	983
3.5.2	Stability	983
3.5.3	Detect Consensus by Subsampling	984
3.5.4	Adding Simulated Noise	984
3.5.5	Over-represented Pathways	984
4	Conclusions	985
	References	986
27	Classification of Genes	993
	<i>Jörg Rahnenführer and Thomas Lengauer</i>	
1	Introduction	993
2	Overview of Gene Classification Tasks	994
2.1	Grouping Genes without Additional Information	995
2.2	Functional Predictions	995
3	Grouping Genes on the Basis of Expression Data	996
3.1	Cluster Analysis	996
3.1.1	Similarity Measures	996
3.1.2	Hierarchical Clustering Algorithms	997
3.1.3	Partitioning Clustering Algorithms	999
3.1.4	Model-based Clustering	1001
3.1.5	Biclustering Algorithms	1002
3.2	Heuristic Gene Grouping of Expression Data	1003
3.2.1	CLICK Algorithm	1003
3.2.2	CAST	1004
3.2.3	Gene shaving	1004
4	Predicting Gene Function from Expression Data	1005
4.1	Classification methods	1006
4.1.1	Support Vector Machines (SVMs)	1006
4.1.2	Rule-based Models	1006
4.2	Supplementing Expression Data with Additional Biological Information	1007
4.2.1	Adding Sequence Data	1009
4.2.2	Adding Gene Ontology Data	1009
4.2.3	Integrating Pathway Information	1011
4.2.4	Combination of Multiple Data Types	1012
5	Evaluation	1014
5.1	Assessing the Biological Relevance of Gene Groups	1015
5.1.1	Validation of Clustering Results	1015
5.1.2	Estimating the Number of Clusters in a Data Set	1016

5.2 Assessing Function Prediction Accuracy 1017
6 Conclusions 1017
References 1018

28 Proteomics: Beyond cDNA 1023
*Patricia M. Palagi, Yannick Brunner, Jean-Charles Sanchez
and Ron D. Appel*

1 Introduction and Principles 1023
2 Proteomics Analytical Methods 1026
2.1 Electrophoresis Gels 1026
2.2 LC 1028
2.3 MS 1030
2.4 Protein Chips 1033
3 Computer Analysis of Proteomics Images 1034
3.1 Analysis of 2-DE Gels 1034
3.1.1 Data Analysis and Validation 1035
3.1.2 Annotation and Databases 1038
3.2 Analysis of LC-MS Images 1038
4 Identification and Characterization of Proteins after
Separation 1039
4.1 Identification with MS 1041
4.2 Characterization with MS 1046
5 Proteome Databases 1047
5.1 Protein Sequence Databases 1048
5.2 2-DE Gel Databases 1049
5.3 Mass Spectra Repositories 1051
5.4 PTM Databases 1051
5.5 General Considerations on Databases 1053
6 Conclusion 1053
References 1054

Volume 3

Part 8 Protein Function Prediction 1061

29 Ontologies for Molecular Biology 1061
Chris Wroe and Robert Stevens

1	Introduction	1061
2	Ontologies and their Components	1063
2.1	Ontology Representation	1065
3	Ontologies in the Real World	1067
3.1	Ontology Tools	1068
3.2	Bio-ontology Communities	1069
3.3	Incremental Development of Ontologies	1071
3.4	Ontology Features to Manage Database Content	1072
3.4.1	A Controlled Vocabulary with Human Readable Definitions	1072
3.4.1.1	Gene Ontology	1072
3.4.1.2	MGED Ontology	1073
3.4.2	A Structured Controlled Vocabulary	1075
3.4.3	A Subsumption Hierarchy	1075
3.4.4	Multiple Hierarchies	1076
3.4.5	Formal Definition of Concepts	1077
3.5	Ontology Features to Manage Data Schemata	1080
3.5.1	TAMBIS	1081
3.6	Ontologies for Prediction and Simulation	1082
3.6.1	EcoCyc	1082
3.7	The Physiome Project	1083
4	Summary	1083
	References	1085

30 Inferring Protein Function from Sequence 1087

Douglas Lee Brutlag

1	Introduction	1087
2	Sequence-based Motif Representations	1090
2.1	Consensus Sequences as Regular Expressions	1090
2.2	Accuracy and Precision of Motifs	1091
2.3	Position-specific Scoring Matrix (PSSM) Motifs	1094
2.4	Dirichlet-mixture Prior Probabilities and Pseudocounts	1094
2.5	Sensitivity and Specificity of PSSM Motifs	1096
2.6	HMMs	1098
2.7	Network Models	1099
2.8	Neural Networks	1101
3	Descriptions of Several Useful Motif Databases	1101
3.1	The Prosite Database	1101
3.2	The Blocks Databases	1104
3.3	The PRINTS Database	1105
3.4	The eBLOCKs Database	1106
3.5	The eMOTIF Database	1107
3.6	The eMATRIX Database	1108

3.7	HMM Databases	1109
3.8	The InterPro Database	1110
3.9	Supervised versus Unsupervised Learning of Motifs	1111
4	Summary and Conclusions	1112
	References	1113
31	Analyzing Protein Interaction Networks	1121
	<i>Johannes Goll and Peter Uetz</i>	
1	Introduction	1121
2	Experimental Methods and Interaction Data	1122
3	Validation of Experimental Protein–Protein Interaction Data	1125
3.1	Crystal Structures as Benchmarks	1126
3.2	Overlap with Protein Complex Data	1126
3.3	Correlation with Expression Data	1126
3.4	Functional Annotation	1127
3.5	Localization	1127
3.6	Paralogous Proteins and Evolutionary Rate	1127
3.7	Other Approaches	1128
3.8	Combined Approaches	1128
3.9	Comparison of Specific Data Sets	1129
3.9.1	Comparison of Tandem Affinity Purification (TAP) and High-throughput MS (HMS) complex purification data	1129
3.9.2	Comparison between Y2H and MS data sets	1131
3.9.3	Comparison of Spoke versus Matrix Models	1131
4	Predicting Protein–Protein Interactions	1133
4.1	Predictions Based on Genomic Context	1138
4.1.1	The Rosetta Stone Method	1138
4.1.2	Gene Neighborhood	1139
4.1.3	Phylogenetic Profiles	1139
4.1.4	Similarity of Phylogenetic Trees (SPT)	1140
4.1.5	<i>In Silico</i> Two-hybrid (I2H)	1140
4.2	Predictions Based on Known 3-D Structures	1140
4.3	Predicting Interaction Domains	1140
4.4	Predicting Homologous Interactions: Interologs	1141
4.5	Predictions based on Literature Mining	1143
4.6	Validation of Predicted Protein–Protein Interactions	1144
5	Representing Protein–Protein Interactions as Graphs	1145
5.1	Graph Terminology	1145
5.2	Network Models	1148
5.3	Random Networks	1149
5.4	Small-world Networks	1149
5.5	Scale-free Networks	1150

- 5.6 Connectivity Distributions of Protein–Protein Interaction Networks 1151
- 5.7 Error Tolerance and Attack Vulnerability 1151
- 5.8 Modules and Motifs in Networks 1152
- 5.9 Comparing Protein Interaction Networks: Pathblast 1152
- 6 Integrating Multiple Protein–Protein Interaction Evidence 1154
- 6.1 Protein Interactions and Gene Expression Data 1157
- 6.2 Integration for Predicting Protein Function 1157
- 7 Predicting Protein Functions from Protein Networks 1157
- 8 Evolution of Protein–Protein Interactions 1158
- 8.1 The Network Level 1159
- 8.1.1 The Rates of Interaction Loss and Gain 1159
- 8.2 Sequence and Interaction Divergence in Proteins 1160
- 8.2.1 Protein Evolution Rate and Protein–Protein Interactions 1161
- 8.2.2 Phylogenetic Relationships between Families of Interacting Proteins 1162
- 8.3 Structural Aspects of Conserved Interactions 1165
- 9 Databases and Other Information Sources 1165
- 10 Analysis and Visualization Tools 1166
- 11 Outlook/Perspectives 1166
- References 1171

32 Inferring Protein Function from Genomic Context 1179

Christian von Mering

- 1 Introduction 1179
- 1.1 Genomic Context – Genomes, Genes and Gene Arrangements 1179
- 1.2 Genome Comparisons Reveal Protein–Protein Associations 1180
- 1.3 Prerequisites for Genomic Context Analysis 1181
- 1.4 How Specific are the Inferred Functions? 1182
- 2 Gene Neighborhood 1183
- 2.1 Conserved Neighborhood versus Simple Synteny 1183
- 2.2 Operons and “Über-Operons” 1185
- 2.3 Divergently Transcribed Gene Pairs 1187
- 2.4 Gene Neighborhood in Eukaryotes 1189
- 3 Gene Fusion 1190
- 3.1 Gene Fusions and Gene Fissions 1190
- 3.2 Functional Implications 1191
- 3.3 Gene Fusions versus Domain Analysis 1193
- 4 Gene Co-occurrence 1194
- 4.1 Phylogenetic Profiles 1194
- 4.2 Discrete versus Continuous Profiles 1196
- 4.3 Profile Distance Measures 1197

4.4	Tree-based Methods	1198
4.5	Anti-correlated Profiles	1199
5	Outlook	1200
5.1	Methods based on Sequence Evolution	1200
5.2	Web-based Implementations of Genomic Context Tools	1202
5.3	Scoring and Integration	1204
5.4	Genome Sequencing Strategies: Impact on Genomic Context Analysis	1205
5.5	Environmental Context	1207
	References	1208
33	Inferring Protein Function from Protein Structure	1211
	<i>Francisco S. Domingues and Thomas Lengauer</i>	
1	Introduction	1211
1.1	Different Levels of Protein Function	1212
1.2	Structural Models	1212
1.3	Homology and Function	1213
1.4	Structure and Function	1214
1.5	Why Predict Function from Structure	1216
1.6	The Challenges of Automatic Prediction of Function from Structure	1217
1.7	Structure of the Chapter	1217
2	Localization of Functional Sites	1218
2.1	Supersites	1218
2.2	Electrostatics	1218
2.3	Surface Geometry	1218
2.4	Structure and Evolutionary Information	1219
2.4.1	Evolutionary Trace (ET)	1219
2.4.2	ConSurf	1220
2.4.3	Residue Conservation and Structural Information	1221
2.5	Network Centrality	1222
2.6	Combined Approaches	1223
2.6.1	Catalytic Sites in Enzymes	1223
2.6.2	Protein–protein Interactions	1223
3	Characterization of Molecular Function	1224
3.1	General Principles	1224
3.1.1	Homology versus Nonhomology	1224
3.1.2	Uncertainty and Flexibility in the Structural Models	1225
3.1.3	Functional Descriptors, Comparison and Scoring	1226
3.2	Descriptors based on Atom Coordinates	1227
3.2.1	ASSAM	1227
3.2.2	SPASM	1228

3.2.3	PINTS	1228
3.2.4	SuMo	1229
3.2.5	TESS and Jess	1230
3.3	Descriptors based on Chemical Environment and Surface	1232
3.3.1	FEATURE	1232
3.3.2	CavBase and SiteEngine	1233
3.3.3	eF-site	1234
3.3.4	pvSOAR	1234
3.3.5	Enzyme Classifier	1236
3.3.6	3D Shape Descriptors	1236
3.4	Databases of Functional Sites	1237
3.4.1	Relibase	1237
3.4.2	MSDsite	1238
3.4.3	CSA	1238
3.4.4	SURFACE	1238
3.4.5	Databases of Structural Motifs	1239
3.4.6	Protein–protein Binding Sites	1239
4	Integration Efforts	1239
5	Resources for Structural Characterization	1241
5.1	Available Tools and Databases	1241
5.2	Characterizing a Protein	1242
6	Current Applications	1243
7	Future Perspectives	1244
	References	1245
34	Mining Information on Protein Function from Text	1253
	<i>Martin Krallinger and Alfonso Valencia</i>	
1	Introduction	1253
2	Information Types of Protein Function Descriptions	1255
3	Literature Databases in Biomedicine	1256
4	NLP	1258
4.1	Grammatical Features	1258
4.2	Morphological Features	1259
4.3	Syntactic Features	1259
4.4	Semantic Features	1260
4.5	Contextual Features	1261
5	Main NLP Tasks	1261
5.1	IR	1261
5.2	IE	1265
5.3	QA	1266
5.4	NLG	1268
6	Difficulties when Processing Biological Texts	1268

7	Strategies of Extracting Functional Information from Text	1271
7.1	NER and Protein Tagging	1271
7.2	Associating Proteins with Biological Features from Databases and Ontologies	1274
7.3	Mining Interactions and Relations from Text	1278
7.4	Discovering Information Associated with Groups of Proteins	1281
7.5	Other Applications	1282
8	Evaluation of Text Mining Strategies	1283
9	Resources for Text Mining	1285
9.1	Literature Databases	1286
9.2	Annotated Text Corpora	1286
9.3	Generic NLP Tools	1286
9.4	Dictionaries and Ontologies	1288
9.5	Biomedical Domain NLP Systems	1289
10	Concluding Remarks	1289
	References	1291
35	Integrating Information for Protein Function Prediction	1297
	<i>William Stafford Noble and Asa Ben-Hur</i>	
1	Introduction	1297
2	Vector-space Integration	1298
3	Classifier Integration	1301
4	Kernel Methods	1302
5	Learning Functional Relationships	1304
6	Learning Function from Networks of Pairwise Relationships	1307
7	Discussion	1311
	References	1311
36	The Molecular Basis of Predicting Druggability	1315
	<i>Bissan Al-Lazikani, Anna Gaulton, Gaia Paolini, Jerry Lanfear, John Overington and Andrew Hopkins</i>	
1	Introduction	1315
2	Chemical Properties of Drugs, Leads and Tools	1316
3	Molecular Recognition is the Basis for Druggability	1316
4	Estimating the Size of the Druggable Genome	1319
4.1	Initial Estimates	1320
4.2	Hopkins and Groom's Method	1320
4.3	Orth and Coworkers Update 2004	1321
4.4	Russ and Lampel's Update 2005	1321
5	Homology-based Analysis of Drug Targets	1322
6	Feature-based Druggability Prediction	1327

7	Structure-based Druggability Analysis of Protein Data Base (PDB) Structures	1327
8	How Many Drug Targets are Accessible to Protein Therapeutics?	1329
9	Conclusions	1331
	References	1333

Part 9 Comparative Genomics and Evolution of Genomes 1335

37	Comparative Genomics	1335
	<i>Martin S. Taylor and Richard R. Copley</i>	
1	Introduction	1335
2	The Genomic Landscape	1336
3	Concepts	1339
4	Practicalities	1343
4.1	Available Genomic Sequences	1343
4.2	Defining and Obtaining Genomic Sequences	1345
5	Technology	1347
5.1	Alignments	1347
5.1.1	Local Genomic Alignments	1349
5.1.2	Global Genomic Alignments	1350
5.1.3	Multiple Sequence Alignments	1351
5.1.4	Assessing the Quality of Genomic Alignment Tools	1353
5.1.5	Using Whole-genome Alignments	1354
5.2	Visualizing Genomic Alignments	1355
5.3	Detecting Selection	1357
6	Applications	1361
6.1	How Much of the Human Genome is Constrained?	1362
6.2	Ultra-conserved Regions	1363
6.3	Specific Locus Studies	1364
7	Challenges and Future Directions	1367
8	Conclusion	1368
	References	1368
38	Association Studies of Complex Diseases	1375
	<i>Momiao Xiong and Li Jin</i>	
1	Introduction	1375
2	Linkage Disequilibrium (LD), Haplotype and Association Studies	1378
2.1	Concepts of LD	1378
2.2	Measures of LD	1379
2.2.1	LD Coefficient D	1379

2.2.2	Normalized Measure of LD D'	1379
2.2.3	Correlation Coefficient r	1380
2.2.4	Composite Measure of LD	1380
2.2.5	The Relationship between the Measure of LD and Physical Distance	1380
2.3	SNPs and Haplotype Blocks in the Human Genome	1381
2.3.1	SNPs	1381
2.3.2	Tagging SNPs	1381
2.3.3	Haplotype Block Model	1381
2.3.4	Definitions of Haplotype Block	1383
2.3.4.1	Definition of Haplotype Blocks based on Pairwise LD	1384
2.3.4.2	Definition of Haplotype Blocks based on Haplotype Diversity	1384
2.3.4.3	Definition of Haplotype Blocks based on both Pairwise LD and Haplotype Diversity	1384
2.3.5	Haplotype Reconstruction	1385
2.3.5.1	Clark's Algorithm	1385
2.3.5.2	Expectation Maximization (EM) Algorithm	1386
2.3.5.3	Bayesian and Coalescence-based Methods	1386
2.3.6	Measure of Haplotype Block LD	1387
3	A General Framework for Population-based Association Studies	1387
3.1	Motivation	1387
3.2	The Traditional χ^2 Test Statistic	1389
3.3	Test Statistics	1391
3.4	Null Distribution of the Nonlinear Statistics	1392
3.5	Power of the Nonlinear Test Statistics and the Standard χ^2 Test Statistic	1393
4	Similarity-based Statistics for Association Studies	1400
4.1	Similarity Measures	1400
4.1.1	Matching Measure	1402
4.1.2	Counting Measure	1403
4.1.3	Length Measure	1403
4.2	Test Statistics	1403
5	Generalized T^2 Test Statistic	1404
5.1	Test Statistic	1405
5.2	Nonlinear T^2 test	1406
6	Family-based Association Studies	1406
6.1	TDT at a Single Locus with Two Alleles	1407
6.2	TDT at a Single Locus with Multiple Alleles or at Multiple Loci with Phase-known Haplotypes	1407
6.3	Sib-TDT	1409
6.3.1	Comparison of Genotype Frequencies	1409

6.3.2	Comparison of Allele Frequencies	1410
7	Nonlinear Transmission/Disequilibrium Test	1410
7.1	General Procedures for the Construction of the Nonlinear TDT	1412
7.1.1	A Single Locus with Two Alleles	1412
7.1.2	A Single Locus with Multiple Alleles or Multiple Loci with Phase-known Haplotypes	1413
7.2	Power of the $N \setminus$ nonlinear TDT	1414
7.3	Real Examples	1415
8	Perspective of Genome-wide Association Studies	1416
	References	1417
39	Pharmacogenetics/Pharmacogenomics	1427
	<i>Xing Jian Lou, Russ B. Altman and Teri E. Klein</i>	
1	Introduction	1427
2	An Overview of Pharmacogenetics and Pharmacogenomics	1427
2.1	Background of Pharmacogenetics and Pharmacogenomics	1428
2.2	Influence of Pharmacogenetics and Pharmacogenomics on Drug Development and Therapy	1429
3	Biomedical Informatics Resources Relevant to Pharmacogenomics	1430
4	Building the PharmGKB	1433
4.1	Establishing a Repository of Pharmacogenetics and Pharmacogenomics Information	1435
4.1.1	The Data Model	1435
4.1.2	Primary Data	1436
4.1.3	Data from Literature	1437
4.1.4	Linking to other Data Resources	1438
4.2	Turning Data into Knowledge	1439
4.2.1	Categorizing Data	1440
4.2.1.1	Genotype	1440
4.2.1.2	Clinical Outcome	1441
4.2.1.3	Pharmacodynamics and Drug Responses	1441
4.2.1.4	Pharmacokinetics	1441
4.2.1.5	Molecular and Cellular Functional Assays	1442
4.2.2	Establishing Genotype–Phenotype Correlation	1442
4.2.3	Using Pathways to Summarize Current Pharmacogenetics and Pharmacogenomics Knowledge	1443
4.3	Providing Easy Access of Knowledge for the Research Community	1443
4.3.1	Querying System	1445
4.3.2	Visualization and Browsing	1445

4.3.3	Privacy Protection	1447
4.3.4	Data Exchange Strategy	1449
5	Analytic Tools for Pharmacogenomics	1449
6	Future Perspectives on Informatics for Pharmacogenetics/Pharmacogenomics	1451
	References	1452
40	Evolution of Drug Resistance in HIV	1457
	<i>Niko Beerenwinkel, Kirsten Roomp and Martin Däumer</i>	
1	Introduction	1457
2	Biomedical Background	1458
2.1	Biology of HIV	1458
2.1.1	Epidemiology of HIV/AIDS	1458
2.1.2	Structure, Genome and Replication Cycle	1459
2.1.3	Basic Immunology and Course of Infection	1461
2.2	Antiretroviral Therapy	1462
2.2.1	Antiretroviral Drugs	1462
2.2.2	Drug Resistance	1464
2.3	Resistance Testing	1464
2.3.1	Genotypic Resistance Testing	1465
2.3.2	Phenotypic Resistance Testing	1465
3	Prediction of Phenotypic Resistance from Genotypes	1466
3.1	Drug Resistance Data	1466
3.2	Methods of Phenotype Prediction	1467
3.3	Comparisons	1468
4	Development of Resistance-associated Mutations	1470
4.1	Viral Evolution	1470
4.2	Learning Mutational Pathways	1472
4.3	Genetic Barrier	1473
4.4	Transitions between Sequence Clusters	1475
5	Selecting Optimal Combination Therapies	1476
5.1	Clinical Databases	1477
5.2	Simple Scoring Functions	1477
5.3	Look-ahead Techniques	1478
5.4	Rules-based Approaches	1479
6	Host Genetic Profiles and Viral Evolution	1480
6.1	Immunobiological Background	1480
6.1.1	HLA Genes	1480
6.1.2	Chemokine Receptors	1482
6.2	Epitope Prediction	1483
6.2.1	Problem Definition	1483
6.2.2	Methods	1484

6.3	Analysis of Escape Mutations	1485
7	Conclusions	1488
8	Web resources	1488
8.1	Los Alamos HIV Databases (http://www.hiv.lanl.gov)	1488
8.2	Stanford HIV Drug Resistance Database (http://hivdb.stanford.edu)	1488
8.3	Geno2pheno (http://www.geno2pheno.org)	1489
8.4	IMGT/HLA Databases (http://www.ebi.ac.uk/imgt/hla)	1489
	References	1489
41	Analyzing the Evolution of Infectious Bacteria	1497
	<i>Dawn Field, Edward J. Feil, Gareth Wilson and Paul Swift</i>	
1	Introduction	1497
1.1	Introduction to Molecular Evolutionary Theory	1498
1.2	The Quantity and Quality of Data Available	1501
1.3	A Practical Overview of Online Resources	1502
2	Identification and Study of Determinants of Virulence and Pathogenicity	1504
2.1	Homology-based Detection	1506
2.2	Pattern-based Detection	1506
2.3	Comparative Genomic Methods of Detection	1507
2.4	Taxonomically Restricted Genes (TRGs) and Orphans	1508
3	Putting Isolates of Infectious Bacteria into a Phylogenetic Framework	1509
4	Mixing of Genetic Material among Bacteria	1512
4.1	The Importance of Phage and Plasmids	1513
5	Coevolution of Infectious Bacteria with Their Hosts	1516
5.1	Reconstructing Metabolic Pathways	1516
5.2	The Genetic Arms Race between Pathogen and Host	1517
6	Conclusions	1518
	References	1520
Part 10	Basic Bioinformatics Technologies	1525
42	Integrating Biological Databases	1525
	<i>Zoé Lacroix, Bertram Ludäscher and Robert Stevens</i>	
1	Biological Resources	1525
2	Data Modeling	1527
2.1	Conceptual Model	1528
2.1.1	ER	1528
2.1.2	Unified Modeling Language	1530
2.2	“Flat” Data Models	1532

2.3	Tree-structured Representations	1533
2.4	Graph Representations	1534
2.5	Multi-dimensional Data Model	1536
3	Data Integration	1537
3.1	Scientific View of Data	1537
3.2	Data Warehouse	1540
3.3	Link-driven Federations	1541
3.4	Mediations	1541
4	Integrating Applications and Data	1542
4.1	Middleware	1543
4.2	CORBA	1544
4.3	Web Services	1545
4.4	P2P	1546
4.5	Grid	1547
5	Semantic Integration	1547
5.1	Identifying Objects	1549
5.2	Representing Metadata	1550
5.3	Ontologies and Data Integration	1552
5.3.1	Example	1553
5.3.2	From Information to Reasoning	1554
5.3.3	Biological Ontologies	1555
5.3.4	Ontologies and Data Integration	1556
5.4	Semantic Web	1557
6	Scientific Workflows	1558
6.1	Example: Promoter Identification Workflow (PIW)	1559
6.2	Scientific Workflow Requirements and Desiderata	1561
6.3	Semantic Extensions and Scientific Workflow Design	1565
7	Conclusion	1567
	References	1567
43	Visualization of Biological Data	1573
	<i>Harry Hochheiser, Kevin W. Eliceiri and Ilya G. Goldberg</i>	
1	Introduction	1573
2	Microscopy Image Visualization	1574
2.1	Fluorescence Microscopy Techniques Applicable to HCS Screening	1574
2.1.1	Spectral Imaging	1575
2.1.2	Lifetime Imaging	1575
2.1.3	Fluorescence Resonant Energy Transfer (FRET)	1576
2.1.4	Optical Sectioning	1577
2.1.5	MP Imaging	1578
2.1.6	Second Harmonic Imaging	1579

2.2	Functional Genomics	1580
2.2.1	RNAi	1580
2.2.2	Chemical Compound Libraries	1581
2.3	Tools for Scientist-driven Analysis Development and Deployment	1582
2.3.1	ImageJ	1582
2.3.2	VisBio	1583
3	Biological Information Visualization	1585
3.1	Genome and Sequence Data	1586
3.2	Gene Expression Data	1594
3.3	Proteomics	1601
3.4	Interaction Networks and Pathways	1601
3.5	Phylogenies and Taxonomies	1605
3.6	Phenotypes and Lineages	1607
3.7	Visualization of the Scientific Process	1608
4	Image Informatics	1608
4.1	Data and Information Management	1611
4.2	Image Analysis	1611
4.3	Analysis Workflows	1613
4.4	Provenance	1613
4.5	Federation	1614
4.6	Visualization and User Tools	1614
5	Conclusion: Research Questions and Challenges	1614
	References	1616
44	Using Distributed Data and Tools in Bioinformatics Applications	1627
	<i>Robert Stevens, Phillip Lord and Duncan Hull</i>	
1	Introduction to Distributed Resources	1627
2	Heterogeneity in Bioinformatics Resources	1629
3	Type Systems in Bioinformatics	1631
4	Plumbing Bioinformatics Resources	1634
4.1	CORBA	1635
4.2	XML in Bioinformatics	1638
4.3	Web Services	1640
5	Case Studies in Distributed Bioinformatics	1642
5.1	ISYS	1642
5.2	BioMOBY	1643
5.2.1	MOBY-S	1643
5.2.2	S-MOBY	1644
5.3	The Grid Future – the <i>my</i> Grid Project	1644
5.4	The <i>my</i> Grid Project	1645

6	Discussion	1647
	References	1649

Part 11 Outlook 1651

45	Future Trends	1651
	<i>Thomas Lengauer</i>	
1	Introduction	1651
2	Building Blocks – Post-translational Modification of Proteins	1653
3	Regulation – Synthesis and Degradation Pipeline of RNA and Proteins	1655
4	Regulation – RNAi	1656
5	Regulation – Tiling Arrays, ChIP-on-chip and array-CGH	1657
6	Regulation – Epigenetics	1659
7	Protein Function – Alternative Splicing	1663
8	Interaction Networks – Immunoinformatics	1665
9	Cell Engineering – Synthetic Biology	1670
9.1	Genetic Engineering	1671
9.2	Protein Engineering	1672
9.3	Genetic Networks	1672
10	Imaging	1673
10.1	Obtaining Pictures of Cellular Structures	1673
10.2	Movies of Cellular Processes	1675
10.3	Organism Development	1676
11	Modeling Organs	1676
12	Outlook	1677
	References	1678

Index 1687

Name Index 1727