Contents

Volume 1

Preface  XXV

List of Contributors  XXIX

Part 1  Introduction  1

1  Bioinformatics - From Genomes to Therapies  1
   Thomas Lengauer

Part 2  Sequencing Genomes  25

2  Bioinformatics Support for Genome-Sequencing Projects  25
   Knut Reinert and Daniel Huson

Part 3  Sequence Analysis  57

3  Sequence Alignment and Sequence Database Search  57
   Martin Vingron

4  Phylogeny Reconstruction  83
   Ingo Ebersberger, Arndt von Haeseler and Heiko A. Schmidt

5  Finding Protein-coding Genes  129
   David C. Kulp

6  Analyzing Regulatory Regions in Genomes  159
   Thomas Werner
### Contents

**Part 3 Finding Repeats in Genome Sequences**

7 Finding Repeats in Genome Sequences 197  
*Brian J. Haas and Steven L. Salzberg*

8 Analyzing Genome Rearrangements 235  
*Guillaume Bourque*

**Part 4 Molecular Structure Prediction** 261

9 Predicting Simplified Features of Protein Structure 261  
*Dariusz Przybylski and Burkhard Rost*

10 Homology Modeling in Biology and Medicine 297  
*Roland L. Dunbrack, Jr.*

11 Protein Fold Recognition Based on Distant Homologs 351  
*Ingolf Sommer*

12 De Novo Structure Prediction: Methods and Applications 389  
*Richard Bonneau*

13 Structural Genomics 419  
*Philip E. Bourne and Adam Godzik*

14 RNA Secondary Structures 439  
*Ivo L. Hofacker and Peter F. Stadler*

15 RNA Tertiary Structure Prediction 491  
*François Major and Philippe Thibault*

**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
Contents

17  Modeling Protein–Protein and Protein–DNA Docking  601
    Andreas Hildebrandt, Oliver Kohlbacher and Hans-Peter Lenhof
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
    Andreas Kämper, Didier Rognan and Thomas Lengauer
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
Chapter 19: Efficient Strategies for Lead Optimization by Simultaneously Addressing Affinity, Selectivity and Pharmacokinetic Parameters

Karl-Heinz Baringhaus and Hans Matter

Introduction 705

1 The Origin of Lead Structures 708

2 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 Global ADME Models for Intestinal Absorption and Protein Binding 724

5.1 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

Chapter 20: Modeling and Simulating Metabolic Networks

Stefan Schuster and David Fell

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
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
Udo Reichl and Yury Sidorenko
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
John Quackenbush
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
Classification of Patients 957
Claudio Lottaz, Dennis Kostka and Rainer Spang

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
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

30 Inferring Protein Function from Sequence 1087
Douglas Lee Brutlag
31 Analyzing Protein Interaction Networks  1121
Johannes Goll and Peter Uetz

32 Inferring Protein Function from Genomic Context  1179
Christian von Mering

33 Inferring Protein Function from Protein Structure  1211
Francisco S. Domingues and Thomas Lengauer

34 Mining Information on Protein Function from Text  1253
Martin Krallinger and Alfonso Valencia

35 Integrating Information for Protein Function Prediction  1297
William Stafford Noble and Asa Ben-Hur

36 The Molecular Basis of Predicting Druggability  1315
Bissan Al-Lazikani, Anna Gaulton, Gaia Paolini, Jerry Lanfear, John Overington and Andrew Hopkins

Part 9 Comparative Genomics and Evolution of Genomes  1335

37 Comparative Genomics  1335
Martin S. Taylor and Richard R. Copley

38 Association Studies of Complex Diseases  1375
Momiao Xiong and Li Jin

39 Pharmacogenetics/Pharmacogenomics  1427
Xing Jian Lou, Russ B. Altman and Teri E. Klein

40 Evolution of Drug Resistance in HIV  1457
Niko Beerenwinkel, Kirsten Roomp and Martin Däumer

41 Analyzing the Evolution of Infectious Bacteria  1497
Dawn Field, Edward J. Feil, Gareth Wilson and Paul Swift

Part 10 Basic Bioinformatics Technologies  1525

42 Integrating Biological Databases  1525
Zoé Lacroix, Bertram Ludäscher and Robert Stevens