Handbook of Chemoinformatics

From Data to Knowledge in 4 Volumes

Volume 2
Contents

Volume 1

Foreword xlvii
Preface xlix
Contributors liii

1 Introduction 1

1.1 The Scope of Chemoinformatics 3
Johann Gasteiger
Textbook–Handbook 4
References 5

1.2 A History of Chemoinformatics 6
Peter Willett

2.1 Introduction 6
2.2 The Early Days: Representations and Structure Searching 7
2.3 The Early Days: Substructure Searching 8
2.4 Similarity Searching, Clustering, and Diversity Analysis 9
2.5 Searching Databases of Patents 11
2.6 Searching Databases of Reactions and Synthesis Design 12
2.7 Computer-aided Structure Elucidation 13
2.8 3D Substructure Searching 13
2.9 QSAR and Docking 14
2.10 Conclusions 16
References 16

II Representation of Chemical Compounds 21

Introduction 23

II.1 Representation of Molecular Structures – Overview 27
John M. Barnard

1.1 Introduction 27
1.2 History 28
1.3 Structure Diagrams 28
1.4 Chemical Nomenclature 30
1.5 Line Notations 30
1.6 Connection Tables 32
1.6.1 Connection Tables and Topological Graphs 33
1.6.2 Internal and External Connection Tables 34
1.6.3 Standard Connection Tables 34
1.7 Fragment Codes and Fingerprints 35
1.8 Special Problems of Structure Representation 37
1.8.1 Alternating Bonds and Aromaticity 37
1.8.2 Tautomerism 39
1.8.3 Inorganic and Coordination Compounds 40
1.8.4 Macromolecules and Polymers 41
1.9 Markush Structures 42
1.10 The Interconversion of Structure Representations 46
1.11 Conclusions and Outlook 47
References 48

II.2 Chemical Nomenclature and Structure Representation: Algorithmic Generation and Conversion 51

Janusz Leon Wisniewski
2.1 Introduction 51
2.2 Computer Representation of a Structural Diagram – Connection Tables 55
2.3 Naming Organic Compounds – Chemical Nomenclature Systems 56
2.3.1 Systematic Chemical Nomenclature 57
2.4 Organic Chemical Nomenclature – Algorithmic Approach 57
2.4.1 Conversion of Chemical Names into Structures 58
2.4.2 Conversion of Structural Diagrams into Chemical Names 61
2.5 Solutions for Algorithmic Name Generation – Case Study AutoNom 63
2.5.1 Ring System Perception 65
2.5.2 Recognition of Functional Groups 65
2.5.3 Ring System Identification 67
2.5.4 Parent Structure Selection 69
2.5.5 Creation of the Name Tree 70
2.5.6 Name Assembly 71
2.5.7 Performance and Limitations of the AutoNom Algorithm 73
2.6 IUPAC Standard Representation – Electronic Signature of a Structure 76
2.7 Outlook 77
References 77

II.3 SMILES – A Language for Molecules and Reactions 80

David Weininger
3.1 Introduction 80
4.5 Molecular Graph Matrices 113
4.5.1 The Adjacency Matrix 114
4.5.2 The Burden Matrix 115
4.5.3 The Laplacian Matrix 116
4.5.4 The Distance Matrix 117
4.5.5 The Reciprocal Distance Matrix 118
4.5.6 The Detour Matrix 119
4.5.7 The Detour-Distance Matrix 119
4.5.8 The Distance-Detour Quotient Matrix 119
4.5.9 The Distance-Valence Matrix 119
4.5.10 The Resistance Distance Matrix 120
4.5.11 The Electrical Conductance Matrix 122
4.5.12 The Distance-Path Matrix 123
4.5.13 The Reciprocal Distance-Path Matrix 123
4.5.14 The Distance Complement Matrix 123
4.5.15 The Reciprocal Distance Complement Matrix 124
4.5.16 The Complementary Distance Matrix 124
4.5.17 The Reciprocal Complementary Distance Matrix 125
4.5.18 The Reverse Wiener Matrix 125
4.5.19 The Reciprocal Reverse Wiener Matrix 125
4.5.20 The Szeged Matrix 126
4.6 Molecular Graph Polynomials 128
4.6.1 The Characteristic Polynomial 128
4.6.2 The Acyclic (Matching) Polynomial 130
4.6.3 The Characteristic Polynomial of a Molecular Matrix 130
4.7 Enumeration of Kekulé Structures 130
4.8 Molecular Graphs and Hückel Molecular Orbital Theory 132
4.9 The Topological Resonance Energy 133
4.10 Isomer Enumeration 134
4.10.1 Pólya's Theorem 135
4.11 Conclusions 136
References 137

II.5 Processing Constitutional Information 139

5.1 Canonical Numbering and Constitutional Symmetry 139
Ovidiu Ivanciuc
5.1.1 Introduction 139
5.1.2 Graph Labeling 140
5.1.3 Constitutional Symmetry of Graphs 142
5.1.4 Canonical Coding of Graphs 146
5.1.5 The MORGAN Algorithm 149
5.1.6 The Augmented Connectivity Molecular Formula 151
5.1.7 Modifications of the Extended Connectivity Algorithm 154
5.1.8 Other Symmetry Perception Algorithms 155
5.2  Ring Perception  161
   Geoffrey M. Downs
   5.2.1  Introduction  161
   5.2.2  Terminology and Definitions  162
   5.2.3  Perception Methods  165
   5.2.3.1  Graph Theoretic Methods  165
   5.2.3.2  Linear Algebraic Methods  166
   5.2.3.3  Pre-processing Methods  166
   5.2.3.3.1  Pruning and Processing Components  166
   5.2.3.3.2  Graph Reduction  166
   5.2.4  Ring Sets and Algorithms  167
   5.2.4.1  All Cycles/Simple Cycles  168
   5.2.4.2  Beta-ring  169
   5.2.4.3  ESER (Essential Set of Essential Rings)  169
   5.2.4.4  Essential Cycles  170
   5.2.4.5  Extended Set of Smallest Rings (ESSR)  171
   5.2.4.6  K-rings (Relevant Cycles)  172
   5.2.4.7  SER (Set of Elementary Rings)  172
   5.2.4.8  SSCE (Set of Smallest Cycles at Edges)  173
   5.2.4.9  SSSR (Smallest Set of Smallest Rings)  174
   5.2.5  Conclusions  176
       Acknowledgment  176
       References  177

5.3  Topological Structure Generators  178
    Ivan P. Banov
   5.3.1  Introduction  178
   5.3.2  Structure Generation Fundamentals  179
   5.3.3  The Isomorphism Problem  182
   5.3.4  Structure Generation Approaches  183
   5.3.5  Outlook  193
       References  194

5.4  Combinatorics of Organic Molecular Structures  195
    David Weininger
   5.4.1  Introduction  195
   5.4.2  Conventions  196
   5.4.3  Small n-Hexane Derivatives  197
   5.4.3.1  Problem Definition  197
   5.4.3.2  Enumeration Strategy  197
   5.4.3.3  Enumerating Small Substituent Graphs  197
   5.4.3.4  Valence and Stereochemistry  198
   5.4.3.5  Enumerating Small Substituents  198
II.6 Representation and Manipulation of Stereochemistry 206

Bernhard Rohde

6.1 Introduction 206
6.2 Connection Tables 206
6.3 3D Representations 208
6.4 Connection Table Representation of Stereochemistry 208
6.4.1 Cahn–Ingold–Prelog Descriptors 209
6.4.2 Priority Coding of Coordination Compounds 210
6.4.3 Relative Stereo Descriptors 210
6.4.4 Stereoparent Descriptors 211
6.4.5 Local Environment Descriptors 211
6.4.5.1 Stereoparities 211
6.4.5.2 Stereovertex Lists 213
6.4.5.3 Relational Systems 213
6.5 Graphical Representations 214
6.6 External Formats 217
6.6.1 MOLFile 217
6.6.2 SMD and MIF Format 217
6.6.3 JCAMP-CS 218
6.6.4 SMILES 219
6.6.5 Other Important Formats 219
6.7 Uses of Stereochemistry in Computer Programs 220
6.7.1 Conversion of Depictions to Connection Table Representations 220
6.7.2 Unique Naming 220
6.7.2.1 Constitutional Canonicalization 220
6.7.2.2 Stereochemical Canonicalization 222
6.7.2.3 Partial Stereochemistry 223
6.7.2.4 Compact Coding 223
6.7.3 Substructure Search 224
6.7.4 Other Uses of Stereochemical Connection Tables 226
6.7.4.1 Computer-assisted Synthesis Planning and Reaction Simulation 226
6.7.4.2 CIP Descriptor Determination 226
6.7.4.3 2D/3D Conversions 227
6.7.4.4 Depictions from Connection Tables 227
6.7.4.5 Applications in Spectroscopy 227
6.8 Conclusion and Outlook 228
References 229
II.7  Representation of 3D Structures  231

7.1  3D Structure Generation  231

  Jens Sadowski

  7.1.1 Introduction  231
  7.1.2 Problem Description  232
  7.1.2.1 Computational Requirements  232
  7.1.2.2 General Problems  233
  7.1.2.3 Classification of Specific Concepts  235
  7.1.2.3.1 Manual Methods  235
  7.1.2.3.2 Automatic Methods  236
  7.1.3 3D Structure Generation: Methods and Programs  237
  7.1.3.1 Early Precursors  237
  7.1.3.1.1 Conformational Analysis for Six-Membered Rings in the LIASA Program  237
  7.1.3.1.2 The Script Program  239
  7.1.3.1.3 SCA: Systematic Conformational Analysis for Cyclic Systems  240
  7.1.3.2 Rule- and Data-Based Methods  241
  7.1.3.2.1 Wizard and Cobra  241
  7.1.3.2.2 CONCORD  243
  7.1.3.2.3 CORINA  244
  7.1.3.3 Fragment-based Methods  250
  7.1.3.3.1 AIMB  250
  7.1.3.3.2 Chem-X  253
  7.1.4 3D Structure Generation: Evaluation of Available Programs  254
  7.1.4.1 Evaluation Procedure  254
  7.1.4.2 Programs  255
  7.1.4.3 Results and Discussion  255
  7.1.4.4 Comparison of CONCORD and CORINA using 25 017 X-ray Structures  256
  7.1.5 Conclusions  259
  References  259

7.2  Conformational Analysis and Searching  262

  Christof H. Schwab

  7.2.1 Introduction  262
  7.2.2 Problem Description  263
  7.2.2.1 Computational Requirements  263
  7.2.2.2 General Problems  265
  7.2.3 Conformational Analysis and Searching: Methods and Programs  266
  7.2.3.1 Systematic Searches  267
  7.2.3.1.1 Acyclic Portions  268
  7.2.3.1.2 Ring Systems  269
  7.2.3.1.3 ConFirm (Catalyst)  271
  7.2.3.2 Rule- and Data-Based Approaches  272
Contents

7.2.3.2.1 COBRA 272
7.2.3.2.2 MIMUMBA 273
7.2.3.2.3 CORINA and ROTATE 276
7.2.3.3 Random Methods 280
7.2.3.3.1 Step 1 – Selection of a Starting Geometry 280
7.2.3.3.2 Step 2 – Changing of Coordinates 280
7.2.3.3.3 Step 3 – Stop Criterion 281
7.2.3.4 Genetic Algorithms 281
7.2.3.5 Distance Geometry 285
7.2.3.6 Simulation Methods 288
7.2.3.6.1 Molecular Dynamics and Monte Carlo Simulations 288
7.2.3.6.2 Simulated Annealing 289
7.2.3.7 Peptide Structures 291
7.2.4 Conformational Analysis and Searching: Evaluation of Available Programs 295
7.2.4.1 Evaluation Procedure 295
7.2.4.2 Programs 296
7.2.4.3 Results and Discussion 297
7.2.5 Conclusions 298
References 299

II.8 Molecular Shape Analysis 302
Jarosław Polański
8.1 Introduction 302
8.2 Molecular Shape and Atomic Configuration 302
8.3 Molecular Graphs for the Description of Molecular Shape 303
8.4 Molecular-shape Descriptors from 3D Structures 304
8.5 Molecular-shape Descriptors from Molecular Configuration 304
8.6 Molecular Surfaces 306
8.6.1 Neural Networks for Comparison of Molecular Surfaces 307
8.7 Receptor Site Models 309
8.8 Molecular Volumes 311
8.8.1 Molecular Similarity Indices and Molecular Shape 312
8.8.2 Molecular Shape Analysis 312
8.8.3 Shape Analyses with Fuzzy Volume Representations 313
8.9 Quantum Chemical Molecular Shape Representations 314
8.10 Protein Shapes 316
8.11 Conclusion and Outlook 317
References 317

II.9 Computer Visualization of Molecular Models – Tools for Man–Machine Communication in Molecular Science 320
Matthias Keil, Thorsten Borosch, Thomas E. Exner, and Jürgen Brickmann
9.1 Introduction 320
9.2 A Guideline to the Development of Molecular Graphics
   Representations 322
9.3 Visualization of the Molecular Recognition Problem 323
9.4 Visual Representation of Molecular Models 324
9.4.1 Standard Models 324
9.4.2 Simplified Molecular Representations and Special Features 325
9.5 Molecular Surfaces 326
9.5.1 Isovalue “Surfaces” 327
9.5.2 Connolly Surfaces 328
9.6 Visualization of Molecular Properties 331
9.6.1 Mapping of Information on Molecular Surfaces 331
9.6.2 Isosurfaces for Molecular Properties 333
9.6.3 Volumetric Properties 334
9.6.4 Representations of Vector Fields 336
9.6.5 Combination of Different Graphical Representations 339
9.7 Conclusions and Outlook 339
References 341

III Representation of Chemical Reactions 345

Introduction 347

III.1 Reaction Classification and Knowledge Acquisition 348
   Lingran Chen
1.1 Introduction 348
1.2 Model-driven Methods of Reaction Classification 349
1.2.1 The Theilheimer Reaction Classification Method 350
1.2.2 The Balaban Reaction Classification Method 350
1.2.3 Hendrickson's Extension of Balaban Reaction Classification Method 352
1.2.4 The Arens Reaction Classification Method 353
1.2.5 The Vladutz Reaction Classification Method 354
1.2.6 The Zefirov Reaction Classification Method 355
1.2.7 The Fujita Reaction Classification Method 358
1.2.8 The Unification of Reaction Descriptions 359
1.2.9 The Ugi Reaction Classification Method 360
1.2.10 The Hendrickson Reaction Classification Method 362
1.3 Data-driven Methods of Reaction Classification 366
1.3.1 The Wilcox Reaction Classification Method 366
1.3.2 The Blurock Reaction Classification Method 367
1.3.3 The Gelernter Reaction Classification Method 367
1.3.4 The InfoChem Reaction Classification Method 368
1.3.5 The HORACE Reaction Classification Method 370
1.3.5.1 Physicochemical Features 371
1.3.5.2 Functional Groups (Topological Features) 371
Contents

1.3.6  The Chen–Gasteiger Classification Method  374
1.3.6.1  Kohonen Neural Networks  374
1.3.6.2  Description of Reactions by Use of Physicochemical Variables  375
1.3.6.3  Automatic Detection of Clusters in the Kohonen Map  375
1.3.6.4  Assigning Empty Neurons  377
1.3.6.5  The Merit of Two-dimensional Classification Map  379
1.3.6.6  Comparison with HORACE  383
1.3.6.7  Reaction Prediction  384
1.4  Summary and Outlook  385
References  386

IV  The Data  389

Introduction  391

IV.1  Data Types  392

Jaroslaw Tomczak
1.1  Introduction  392
1.2  Numeric Data  394
1.3  Molecular Structures  395
1.4  Bit Vectors  397
1.4.1  Hash Codes  397
1.4.2  Structural Keys  400
1.4.3  Fingerprints  403
1.5  Molecular Spectra  404
1.6  Conclusions  407
References  407

2.1  Quality Control and Data Analysis  410

David E. Booth, Thomas L. Isenhour, John K. Mahaney Jr, Michael Suh, and Christine Wright
2.1.1  Introduction  410
2.1.2  Process Control  411
2.1.2.1  Incoming Inspection  411
2.1.2.2  Process Control Background  411
2.1.2.2.1  Pareto Charts  411
2.1.2.2.2  Cause and Effect Diagram  412
2.1.2.3  Control Charts  412
2.1.2.3.1  Control Charts for Variable Data in Statistical Process Control (SPC)  412
2.1.2.3.2  Control Charts for Attributes  413
2.1.2.3.3  Multivariate Control Charts  413
2.1.3  Design of Experiments  414
2.1.3.1  Background  414
2.1.3.2  Importance of DOE in Statistical Process Control  414
2.1.3.3 DOE Techniques 415
2.1.3.4 Summary 415
2.1.4 Taguchi Methods 416
2.1.4.1 Introduction 416
2.1.4.2 Taguchi's Definition of Quality and the Loss Function 416
2.1.4.3 Taguchi's Off-line Quality Control 416
2.1.4.4 Taguchi's Off-line Methods and Statistical Design of Experiments 417
2.1.4.5 Taguchi's On-line Process Control 417
2.1.5 Illustrations of Chemical Statistical Process Control: Nuclear Material Safeguards and Open-path Fourier Transform Infrared Spectroscopy 417
2.1.5.1 Introduction 417
2.1.5.2 Background 418
2.1.5.2.1 A Modern Approach to SPC when Standard Methods are not Appropriate 418
2.1.5.3 Applications to Nuclear Material Safeguards and Other Prior Research 418
2.1.5.3.1 The Material Balance Equation 418
2.1.5.3.2 Material Balances as a Time Series 419
2.1.5.4 Three New Algorithms for the Detection of Nuclear Material Losses and the General SPC Problem 419
2.1.5.4.1 Smoothing and Data Bounding 419
2.1.5.4.2 Joint Estimation Method 419
2.1.5.4.3 Neural Networks 420
2.1.5.5 Open-path Fourier Transform Infrared Spectroscopy in SPC 420
2.1.6 Conclusions 421
References 421

2.2 Experimental Design 423

Mario Marsili

2.2.1 Introduction 423
2.2.1.1 Information 423
2.2.2 Creating Design Matrices 424
2.2.2.1 Orthogonality 424
2.2.2.2 Full Factorial Designs 425
2.2.2.3 Main Effects 427
2.2.2.4 Precision 427
2.2.3 Interactions 428
2.2.4 Fractional Factorial Designs 430
2.2.4.1 Confounding 431
2.2.4.2 Saturated Designs 433
2.2.4.3 An Example 434
2.2.5 Non-linear Designs 436
2.2.5.1 Composite Designs 436
2.2.5.2 Box–Behnken Design  437
2.2.5.3 Equiradiial Designs  438
2.2.5.4 Multivariate Multilevel Designs  438
2.2.5.5 Non-orthogonal Designs  439
2.2.5.6 Design of Organic Analog Compounds  440
2.2.6 Quality Design  440
2.2.6.1 Noise Factors  441
2.2.7 Optimizing Quality by Matrix Design  444
2.2.8 Conclusions  444
References  444

IV.3 Standard Exchange Formats for Spectral Data  446

Antony N. Davies
3.1 Introduction  446
3.2 JCAMP-DX Protocols  447
3.2.1 Background  447
3.2.2 The JCAMP-DX File Format  448
3.2.3 JCAMP-DX for Infrared Spectroscopy  449
3.2.3.1 History  449
3.2.3.2 Required Labeled Data Records for Infrared  450
3.2.4 JCAMP-DX for Nuclear Magnetic Resonance Spectroscopy  451
3.2.4.1 History  451
3.2.4.2 The NMR Specific Keywords  451
3.2.4.3 Recent Advances: JCAMP-DX v5.01  452
3.2.5 JCAMP-DX for Mass Spectrometry  452
3.2.5.1 History  452
3.2.5.2 Mass Spectrometry Specific Keywords  453
3.2.6 JCAMP-CS for Chemical Structures  453
3.2.6.1 History  453
3.2.6.2 Chemical Structure-specific Keywords  454
3.2.7 JCAMP-DX for Ion Mobility Spectrometry  454
3.2.7.1 History  454
3.2.7.2 Ion Mobility Spectrometry-specific Keywords  455
3.2.8 Guidelines for the Representation of Pulse Sequences for Solution-state NMR  456
3.2.8.1 History  456
3.2.8.2 NMR Pulse Sequences  456
3.2.9 JCAMP-DX version 6.0 Generic  456
3.2.9.1 History  456
3.2.9.2 The NTUPLE format  457
3.2.10 Future Work (IUPAC)  457
3.3 Analytical Instrument Association Standards  458
3.3.1 Background  458
3.3.2 The Unidata netCDF Standard  458
3.3.3 The AIA Information Categories  459
3.3.4 The AIA Chromatography 460
3.3.5 The AIA Data Interchange Specification for Mass Spectrometry 461
3.3.6 The AIA Data Interchange Specification for Infrared Spectroscopy 461
3.3.7 The AIA and the ASTM 462
3.4 XML in Chemistry and Spectroscopy 462
3.4.1 IUPAC and the XML in Chemistry Initiative 462
3.4.2 ASTM E13 XML Sub-committee Meeting 464
3.5 Conclusions 464
References 465

IV.4 XML and Its Application in Chemistry 466
Peter Murray-Rust and Henry S. Rzepa
4.1 Scope of this Section 466
4.2 The Need for an Extensible Markup Language 466
4.2.1 The Web 466
4.2.2 Software Quality 467
4.2.3 Representation of Knowledge 467
4.3 What is XML? 468
4.4 Information Objects ("Datums") in Chemistry 469
4.4.1 Namespaces 470
4.5 Software Implementations of XML 471
4.6 Creation of XML 471
4.7 Validation of XML 472
4.8 Processing XML 473
4.9 Scientific Technical Medical Markup Language (STMML) 475
4.10 Scientific Dictionaries in XML 476
4.11 Chemical Markup Language (CML) 478
4.12 Computational Chemical Markup Language CMLComp 480
4.12.1 Architecture of CMLComp 481
4.12.1.1 Phase 1. Interfacing with Computational Codes 481
4.12.1.2 Phase 2. Program Flow 482
4.12.2 Information Flow 483
4.12.3 Job Control and Workflow 483
4.12.4 XML Software 485
4.13 Conclusions 485
References 486
Appendix 486
Volume 2

V Databases/Data Sources 491

Introduction 493

V.1 Overview of Databases/Data Sources 496
Gary D. Wiggins
1.1 Introduction 496
1.2 Commercial Database Vendors and Databases 497
1.2.1 Common Features of Vendor Systems 497
1.2.1.1 Front-End Search Software 497
1.2.1.2 Database Search Costs 497
1.2.1.3 Data Analysis Tools 498
1.2.2 STN International and CAS Databases 498
1.2.3 SciFinder 499
1.2.4 Other Vendors and Databases 499
1.2.4.1 Major Chemical Database Vendors 500
1.2.4.2 Hybrid Publishers/Vendors 501
1.2.4.3 Beilstein and Gmelin 502
1.2.4.4 Knovel Databases 502
1.2.4.5 Handbooks, Encyclopedias, Physical Property Data Compilations 503
1.2.4.6 Cambridge Structural Database 503
1.2.5 Electronic Journals 503
1.2.6 Free Internet Sources 504
1.2.7 The Future 505
References 505

V.2 Bibliographic Databases 507
Andreas Barth
2.1 Introduction 507
2.2 Abstracting and Indexing in Bibliographic Databases 508
2.2.1 Metadata, Data Structures and Representation 508
2.2.2 Subject Indexing 508
2.2.3 Retrieval Functions 510
2.2.4 Thesaurus Function 512
2.3 Important Bibliographic Databases in Chemistry 512
2.3.1 The Chemical Abstracts Plus File (CAplus) 512
2.3.2 Databases with Relevance to Chemistry 513
2.3.3 SCISEARCH (Science Citation Index) 515
2.4 Search Strategies and Examples 515
2.4.1 Concepts for Searching 515
2.4.2 Example: Simple Search in CAplus 516
2.4.3 Example: Search in More than one Database 518
2.5 Analysis and Postprocessing 518
4.5 Conclusion 604
References 606

V.5 The Beilstein Database 608
Alexander J. Lawson
5.1 Historical Background 608
5.2 The Beilstein Information System 612
5.3 The CrossFire Revolution 615
5.3.1 Opening up Interaction with End Users, Directly 615
5.3.2 The Search and Retrieval Performance of CrossFire 616
5.3.2.1 Structure Searching 616
5.3.2.2 Text and Numerical Searching 617
5.3.3 Point and Click Access to Relevance in Context 618
5.3.4 Abstracts, Reactions, and Ecopharm Extensions to the Beilstein File 619
5.3.5 Incorporation in a Fully Integrated Environment 622
5.4 The Future 626
References and Notes 627

V.6 Databases in Inorganic Chemistry 629
Jürgen Vogt, Natalja Vogt, and Axel Schunk
6.1 Introduction 629
6.2 Features for Retrieving Inorganic and Organometallic Compounds 630
6.3 The Databases of Chemical Abstracts 631
6.4 INSPEC Database 631
6.5 Gmelin Database 632
6.5.1 Literature Coverage 632
6.5.2 Stereochemistry and 3D Structures 633
6.5.3 Ligand Search System 634
6.5.4 Reaction Retrieval 634
6.5.5 Abstracts 635
6.5.6 Multifile Search 636
6.5.7 Information about Catalysts 636
6.6 Crystallographic Databases 636
6.7 Database for Gas-phase Compounds 637
6.8 Landolt–Börnstein 640
6.9 Resume 642
References 642

V.7 The Cambridge Structural Database (CSD) of Small Molecule Crystal Structures 645
Frank H. Allen, Karen J. Lipscomb, and Gary Battle
7.1 Introduction 645
7.2 The Cambridge Structural Database 646
Contents

7.2.1 Information Content of the CSD 646
7.2.2 Data Acquisition 648
7.2.3 Data Processing, Validation and Annotation 648
7.2.4 Statistics 649
7.3 The CSD Software System 650
7.3.1 Searching the CSD Using ConQuest 650
7.3.2 Visualizing Crystal Structures: Mercury 651
7.3.3 Analysis and Display of Geometrical Data: Vista 654
7.4 Research Applications of the CSD 655
7.4.1 Overview and Leading References 655
7.4.2 The DBUse Bibliography of CSD Applications 656
7.5 CSD Applications Examples 656
7.5.1 Molecular Dimensions 656
7.5.2 Conformational Analysis 657
7.5.3 Intermolecular Interactions: Weak Hydrogen-bonds 658
7.6 IsoStar – A Knowledge Base of Intermolecular Interactions 660
7.7 CSD System Releases and Data Availability 663
7.8 CCDC Applications Software 663
7.8.1 Protein–Ligand Interactions 663
7.8.2 Crystal Structures from Powder Diffraction Data 664
7.9 Conclusions 665
References 665

V.8 Databases of Chemical Reactions 667
Engelbert Zass

8.1 Introduction 667
8.2 Sources of Reaction Information 668
8.2.1 Interfaces to Reaction Databases 669
8.2.2 General Organic Reaction Databases 670
8.2.2.1 Theilheimer and Journal of Synthetic Methods 670
8.2.2.2 ChemInform RX 671
8.2.2.3 Current Chemical Reactions (CCR) and Related ISI Databases 671
8.2.2.4 Methods in Organic Synthesis (MOS) 673
8.2.2.5 Reference Library of Synthetic Methodology (RefLib) 673
8.2.2.6 ChemReact 673
8.2.2.7 Science of Synthesis 674
8.2.2.8 CASREACT 675
8.2.2.9 CrossFire Beilstein 676
8.2.2.10 Summary 678
8.2.3 Special Organic Reaction Databases 680
8.2.3.1 Organic Syntheses 680
8.2.3.2 Comprehensive Heterocyclic Chemistry (CHC) 680
8.2.3.3 Encyclopedia of Reagents in Organic Synthesis (EROS) 680
8.2.3.4 Protecting Groups 681
8.2.3.5 Biotransformations and Metabolic Reactions 683
10.3.1.5 Catalogs of Chemical Substances (CA), Material Safety Data Sheets (MSDS) 730
10.3.2 Text-based Databases 732
10.3.2.1 Bibliographic Databases (BI) 732
10.3.2.2 Full-text Databases (FT) 734
10.3.3 Integrated Databases (INDB) 735
10.3.3.1 Structural Databases (ST) 735
10.3.3.2 Reaction Databases (RE) 735
10.4 DAIN – Metadatabase of Internet Resources for Environmental Chemicals 736
10.4.1 General Description of DAIN 736
10.4.2 Content-related Evaluation of DAIN 738
10.5 Conclusions and Outlook 739
References 741

V.11 Patent Databases 743
Jürgen Vogt
11.1 Introduction 743
11.2 Patents as Important Source of Novel Chemical Information 743
11.3 Databases of the Chemical Abstracts Service 744
11.4 World Patents Index 746
11.5 Beilstein and Gmelin Database 748
11.6 Full Text Databases 750
11.7 Other Patent Databases 751
11.8 Summary 754
References 755

V.12 Databases in Biochemistry and Molecular Biology 756
Alexander von Homeyer and Martin Reitz
12.1 Introduction 756
12.2 Classification 757
12.3 Data in Biochemistry and Molecular Biology 758
12.4 Growth of Data 759
12.5 Annotation and Documentation 760
12.6 Entry Coding 760
12.7 Redundancy 762
12.8 Formats 763
12.9 Search Options and Algorithms 764
12.10 Sequence Databases 768
12.10.1 Nucleotide Sequence Databases 768
12.10.2 Protein Sequence Databases 769
12.11 Motif, Domain and Family Databases 771
12.12 Macromolecular 3D Structure Databases 775
12.13 Molecular Interaction Databases 778
## Contents

V.14  
**Laboratory Information Management Systems (LIMS)**  
Markus Hemme  
14.1  
Introduction  844  
14.2  
LIMS and Regulatory Compliance  846  
14.2.1  
Good Automated Laboratory Practice (GALP)  846  
14.3  
FDA 21 CFR Part 11-compliant Data Management  848  
14.4  
LIMS Characteristics  849  
14.5  
Why Use a LIMS?  849  
14.6  
The Basic LIMS  850  
14.6.1  
A Functional Model  852  
14.6.1.1  
Sample Tracking  852  
14.6.1.2  
Sample Analysis  852  
14.6.1.3  
Information Structure  853  
14.7  
The Modern LIMS  853  
14.7.1  
The Planning System  853  
14.7.1.1  
Basic Data  854  
14.7.1.2  
Product Standards  854  
14.7.2  
Controlling System  855  
14.7.3  
Laboratory Processing  856  
14.7.4  
The Assurance System  858  
14.7.5  
Automatic Test Programs (ATP)  858  
14.7.5.6  
Offline Client  859  
14.8  
Additional LIMS Modules  859  
14.8.1  
Stability Management  859  
14.8.2  
Complaints Management  861  
14.8.3  
Reference Substance Module  861  
14.8.4  
Recipe Administration  862  
14.9  
LIMS and Knowledge Management in the Analytical Laboratory  863  
Further Reading  864  

VI  
**Searching Chemical Structures**  865  

**Introduction**  867  

VI.1  
**Two-dimensional Structure and Substructure Searching**  868  
Jun Xu  
1.1  
Introduction  868  
1.2  
2D Structure Representation  870  
1.2.1  
Connection Table  870  
1.2.2  
Linear Notation  871  
1.2.3  
Structure Representation Data Exchange Formats  871  
1.3  
Structure Searching  872  
1.3.1  
Molecular Indices  872  
1.3.2  
Canonic Linear Notation  873  
1.3.3  
Structure Search  875
1.4 Substructure Searching  875
1.4.1 Structure Normalization  875
1.4.2 Substructure Search Query  877
1.4.3 Atom-by-atom Match  877
1.4.4 Backtracking Algorithm (Using the Example from Figure 1-9)  878
1.4.4.1 Step 1  878
1.4.4.2 Step 2  878
1.4.5 Techniques to Enhance Atom-by-atom Match Performance  879
1.5 Substructure Mapping and Other Structure Perceptions  880
1.5.1 Finding the Smallest Set of Smallest Rings (SSSR)  880
1.5.2 Determination of Topological RS Chirality  881
1.6 Summary and Conclusions  883
References  884

VI.2 Current State of the Art of Markush Topological Search Systems  885
Andrew H. Berks
2.1 Introduction  885
2.2 Markush DARC  886
2.3 MARPAT  888
2.3.1 Database Content Issues  889
2.3.1.1 Old Content Issues  889
2.3.1.2 Current Content Issues  889
2.3.2 System Issues  891
2.3.2.1 Translation Capability  891
2.3.2.2 Other Search Engine Issues  892
2.3.2.3 Client Software  896
2.3.2.4 Display Issues  897
2.3.2.5 Costs  901
2.4 Summary and Future Directions  902
References  903

VI.3 Similarity Searching in Chemical Structure Databases  904
Peter Willett
3.1 Introduction  904
3.2 Similarity Searching in 2D Databases  906
3.2.1 Fragment Substructures  906
3.2.2 Topological Indices  907
3.2.3 Graph-based Approaches  907
3.2.4 Weighting Schemes and Similarity Coefficients  908
3.3 Similarity Searching in 3D Databases  909
3.3.1 Fragment Substructures  910
3.3.2 Alignment-based Approaches  911
3.4 Conclusions  911
References  913
Volume 3

VII Calculation of Physical and Chemical Data 917

Introduction 919

VII.1 Molecular Mechanics 920
Harald Lang

1.1 Introduction 920
1.2 The Common Functional Form of Many Force Fields 921
1.2.1 Can Molecular Mechanics Give Correct Results? 921
1.2.2 The Mathematics on One Page 921
1.3 Answers to Some Frequently Asked Questions 922
1.3.1 Can Force Fields Describe Reactive Chemical Systems? 923
1.3.2 How Can We Simulate the Effect of Liquid Water? 924
1.3.3 How Do Force Field Energies Correspond to Thermodynamics? 925
1.4 Available Force Fields – An Incomplete Overview 926
1.4.1 Following Force Field Development May not be that Easy 926
1.4.2 Force Fields for Mainly Small Molecules 927
1.4.2.1 MM2/MM3/MM4 (http://europa.chem.uga.edu/allinger/mn2mm3.html) 927
1.4.2.2 Tinker (http://dasher.wustl.edu/tinker) 927
1.4.2.3 UFF (http://franklin.chem.colostate.edu/mmac/uff.html) 928
1.4.2.4 Mornec (http://www.uni-heidelberg.de/institute/fak12/AC/comba/) 928
1.4.2.5 Cosmos (http://www.cosmos-software.de/) 929
1.4.3 Force Fields for Biological Molecules 929
1.4.3.1 AMBER (http://www.amber.ucsf.edu/amber/) 930
1.4.3.2 CHARMM (http://yuri.harvard.edu) 931
1.4.3.3 Gromos (http://www.igc.ethz.ch/gromos/) 933
1.4.3.4 OPLS (http://zarbi.chem.yale.edu) 934
1.4.3.5 ECEPP (http://www.tc.cornell.edu/reports/NIH/resource/CompBiologyTools/eceppak/) 935
1.4.3.6 CVFF/CFF (http://struktrum.kemi.dtu.dk/cff/cffhome.html) 936
1.4.3.7 MMFF (http://www.schrodinger.com) 937
1.5 Applications of Force Field Techniques 937
1.5.1 For What Else Can We Use Molecular Mechanics Energies? 937
1.5.2 Mixed QM/MM Methods 938
1.5.3 Examples of State-of-the-Art Simulations 939
1.6 Conclusion and Outlook 940
References 941

VII.2 Quantum Mechanics 947
Tim Clark

2.1 Introduction 947
2.2 Hückel Molecular Orbital Theory 948
2.3 Semiempirical MO Theory 953
2.4 Ab initio Molecular Orbital Theory 957
2.5 Density Functional Theory 964
2.6 Properties from Quantum Mechanical Calculations 966
2.6.1 Net Atomic Charges 966
2.6.2 Dipole and Higher Multipole Moments 968
2.6.3 Polarizabilities 968
2.6.4 Orbital Energies 969
2.6.5 Surface Descriptors 969
2.6.6 Local Ionization Potential 969
2.7 Quantum Mechanical Techniques for Very Large Molecules 970
2.7.1 Linear Scaling Methods 970
2.7.2 Hybrid QM/MM Calculations 972
2.8 The Future of Quantum Mechanical Methods in Chemoinformatics 973
References 973

VIII Descriptors for Chemical Compounds 977

Introduction 979

VIII.1 Topological Indices 981

Ovidiu Ivanciuc

1.1 Introduction 981
1.2 Topological Indices 982
1.2.1 The Wiener Index and Related Topological Indices 982
1.2.2 The Szeged Index 983
1.2.3 The Connectivity Indices 984
1.2.4 The Electrotopological State 985
1.2.5 The Hosoya Index 985
1.2.6 The Balaban Index J 986
1.2.7 The Information-Theory Indices $I_{D}^{W}$, $I_{D}^{W}$, $I_{D}^{W}$, and $I_{D}^{W}$ 986
1.2.8 The Information on Distances Indices U, V, X, and Y 987
1.2.9 Triplet Topological Indices T.P.R 988
1.3 Computing Topological Indices with Graph Operators 989
1.3.1 The Vertex Sum Operator 989
1.3.2 The Connectivity Chi Operator 990
1.3.3 The Wiener Operator 990
1.3.4 The Hyper-Wiener Operator 990
1.3.5 The Szeged Operator 991
1.3.6 The Characteristic Polynomial Operator 992
1.3.7 The Matrix Spectrum Operators 992
1.3.8 The Spectral Moment Operator 993
1.3.9 The Hosoya Operator 994
Contents

1.3.10 The Ivanciuc–Balaban Operator 994
1.3.11 The Information-theory Operators U(M), V(M), X(M), and Y(M) 995
1.3.12 Information Theory Indices 996
1.3.13 Information Theory Operators Derived from the Equality of Matrix Elements 997
1.3.14 Information Theory Operators Derived from the Size of Matrix Elements 998
1.3.15 Information Theory Operators Derived from the Equality of Vertex Invariants 999
1.3.16 Information Theory Operators Derived from the Size of Vertex Invariants 1000
1.4 Topological Indices for Combinatorial Chemistry 1000
1.5 Conclusions 1001
References 1002

VIII.2 Descriptors from Molecular Geometry 1004
Robert Todeschini and Viviana Consonni
2.1 Introduction 1004
2.2 Geometrical Descriptors for Molecular Size and Shape 1005
2.3 WHIM Descriptors 1012
2.4 GETAWAY Descriptors 1017
2.5 3D Autocorrelation Descriptors 1024
2.6 3D-MoRSE and RDF Descriptors 1025
2.7 EVA and EEVA Descriptors 1029
2.8 Conclusions 1031
References 1031

VIII.3 A Hierarchy of Structure Representations 1034
Johann Gasteiger
3.1 Introduction 1034
3.2 Constitution, 3D Structure, Molecular Surfaces 1037
3.2.1 Constitution 1037
3.2.2 3D Structures and Conformational Flexibility 1037
3.2.3 Molecular Surfaces 1038
3.3 Physicochemical Effects 1038
3.3.1 Charge Distribution 1039
3.3.2 Polarizability Effect 1040
3.3.3 Additional Physicochemical Effects 1042
3.3.4 PETRA 1044
3.4 Uniform Structure Representations 1044
3.4.1 Inductive Learning Methods 1044
3.4.2 Coding the Constitution 1045
3.4.3 Coding the 3D Structure 1049
3.4.3.1 3D Autocorrelation 1049
3.4.3.2 3D-MoRSE Code 1050
3.4.3.3 Radial Distribution Function Code 1051
3.4.3.4 Descriptors of Molecular Chirality 1052
3.4.4 Coding of Molecular Surface Properties 1052
3.4.4.1 Autocorrelation of Molecular Surface Properties 1052
3.4.4.2 Kohonen Maps of Molecular Surface Properties 1056
3.5 Summary 1058
References 1059

VIII.4 Representation of Molecular Chirality 1062
João Aires-de-Sousa
4.1 Introduction 1062
4.2 Detection and Specification of Chirality 1064
4.3 Quantitative Descriptions of Chirality 1066
4.3.1 Continuous Chirality Measure (CCM) 1067
4.3.2 Dissymmetry Function 1068
4.3.3 Chirality Coefficients 1069
4.3.4 Chirality Functions 1069
4.3.5 Chirality Measure Within a Data Set 1070
4.3.6 Chirality-modified Topological Indices 1071
4.3.7 Chirality Codes 1072
4.3.8 Chirality-aware Molecular Fingerprints 1075
4.4 Conclusion 1076
References 1077

IX Methods for Data Analysis 1079

Introduction 1081

IX.1 Inductive Learning Methods 1082

1.1 Machine Learning Techniques in Chemistry 1082
John R. Rose
1.1.1 Introduction 1082
1.1.2 Supervised Learning 1083
1.1.2.1 Decision Trees 1084
1.1.2.2 Bayesian Belief Networks 1085
1.1.2.3 Version Space 1088
1.1.2.4 Explanation-Based Learning 1089
1.1.2.5 General Neural Networks 1090
1.1.2.6 Support Vector Learning 1090
1.1.2.7 Genetic Algorithms 1093
1.1.3 Unsupervised Learning 1093
1.1.3.1 Conceptual Clustering 1094
1.1.3.2 Neural Networks 1095
1.1.4 Conclusions 1095
## Contents

1.1.5 Related Topics 1095  
References 1095  

### 1.2 Multivariate Data Analysis in Chemistry 1098  
*Kurt Varmuza*

1.2.1 Introduction 1098  
1.2.1.1 Basic Ideas 1098  
1.2.1.2 Literature 1099  
1.2.1.3 Matrix Algebra 1100  
1.2.2 Multivariate Data 1103  
1.2.2.1 Basic Definitions 1103  
1.2.2.2 Preprocessing 1103  
1.2.2.3 Variances and Covariances 1104  
1.2.2.4 Distances and Similarities 1105  
1.2.3 Linear Latent Variables 1106  
1.2.3.1 Concept (Features, Loadings, Scores) 1106  
1.2.3.2 Projection onto an Axis and a Plane (Scatter Plots) 1107  
1.2.4 Principal Component Analysis 1109  
1.2.4.1 Concept 1109  
1.2.4.2 Mathematics 1111  
1.2.4.3 Applications and Summary 1113  
1.2.5 Multivariate Exploratory Data Analysis 1114  
1.2.5.1 Methods 1114  
1.2.5.2 Example 1116  
1.2.6 Multivariate Calibration 1117  
1.2.6.1 Methods 1117  
1.2.6.2 Example 1125  
1.2.7 Multivariate Classification 1127  
1.2.7.1 Methods 1127  
1.2.7.2 Example 1129  
1.2.8 Summary and Outlook 1131  
References 1131  

### 1.3 Partial Least Squares (PLS) in Cheminformatics 1134  
*Lennart Eriksson, Henrik Antti, Elaine Holmes, Erik Johansson, Torbjörn Lundstedt, John Shockcor, and Svante Wold*

1.3.1 Introduction 1134  
1.3.1.1 Notation 1135  
1.3.2 PLS and the Underlying Scientific Model 1135  
1.3.2.1 The Data – X and Y 1136  
1.3.2.2 Transformation, Scaling, Centering, and Normalization 1136  
1.3.2.3 The PLS Model 1138  
1.3.2.4 Interpretation of the PLS Model 1140  
1.3.2.5 Geometric Interpretation 1140  
1.3.2.6 Incomplete X and Y Matrices (Missing Data) 1141
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3.2.7 One Y at a Time, or All in a Single Model?</td>
<td>1142</td>
</tr>
<tr>
<td>1.3.2.8 The Number of PLS Components, A</td>
<td>1142</td>
</tr>
<tr>
<td>1.3.2.9 Model Validation</td>
<td>1143</td>
</tr>
<tr>
<td>1.3.2.10 PLS Algorithms</td>
<td>1144</td>
</tr>
<tr>
<td>1.3.2.11 Standard Errors and Confidence Intervals</td>
<td>1145</td>
</tr>
<tr>
<td>1.3.3 Assumptions Underlying PLS</td>
<td>1145</td>
</tr>
<tr>
<td>1.3.3.1 Latent Variables</td>
<td>1145</td>
</tr>
<tr>
<td>1.3.3.2 Alternative Derivation</td>
<td>1146</td>
</tr>
<tr>
<td>1.3.3.3 Homogeneity</td>
<td>1147</td>
</tr>
<tr>
<td>1.3.4 Illustrations</td>
<td>1147</td>
</tr>
<tr>
<td>1.3.4.1 Example 1, A Modeling and Design Study of Hexapeptides</td>
<td>1147</td>
</tr>
<tr>
<td>1.3.4.2 Results of Example 1 (Hexapep)</td>
<td>1148</td>
</tr>
<tr>
<td>1.3.4.2.1 Basic Principles of Peptide QSAR</td>
<td>1148</td>
</tr>
<tr>
<td>1.3.4.2.2 Initial PLS Modeling</td>
<td>1149</td>
</tr>
<tr>
<td>1.3.4.2.3 Refined PLS Modeling</td>
<td>1152</td>
</tr>
<tr>
<td>1.3.4.2.4 Molecular Design of New Hexapeptides</td>
<td>1153</td>
</tr>
<tr>
<td>1.3.4.2.5 Virtual Screening of New Hexapeptides</td>
<td>1154</td>
</tr>
<tr>
<td>1.3.4.2.6 Discussion of Example</td>
<td>1154</td>
</tr>
<tr>
<td>1.3.4.3 Example 2, a Metabonomic Investigation of Phospholipidosis in</td>
<td>1155</td>
</tr>
<tr>
<td>Rat</td>
<td></td>
</tr>
<tr>
<td>1.3.4.4 Results of Example 2 (Metabonomics)</td>
<td>1156</td>
</tr>
<tr>
<td>1.3.4.4.1 An Overview PCA Model</td>
<td>1156</td>
</tr>
<tr>
<td>1.3.4.4.2 PLS Discriminant Analysis (PLS-DA)</td>
<td>1159</td>
</tr>
<tr>
<td>1.3.4.4.3 PLS-DA of Groups &quot;s&quot;, &quot;sa&quot;, and &quot;sc&quot;</td>
<td>1160</td>
</tr>
<tr>
<td>1.3.4.4.4 Discussion of Second Example</td>
<td>1161</td>
</tr>
<tr>
<td>1.3.5 Conclusions and Discussion</td>
<td>1164</td>
</tr>
<tr>
<td>References</td>
<td>1165</td>
</tr>
</tbody>
</table>

1.4 **Neural Networks** 1167

*Jure Zupan*

1.4.1 History 1167
1.4.2 Introduction 1168
1.4.3 Basic Concepts of ANN 1170
1.4.3.1 Artificial Neuron 1170
1.4.3.2 Networking of Neurons 1173
1.4.4 Learning by ANN 1173
1.4.4.1 Error-backpropagation Learning 1173
1.4.4.2 Kohonen Learning 1176
1.4.4.3 Counterpropagation Learning 1179
1.4.4.4 When to Stop Learning of ANN 1181
1.4.4.5 Overtraining 1182
1.4.4.6 Radial-basis Functions 1183
1.4.4.7 Other ANN 1186
1.4.5 Standard Applications 1186
1.4.5.1 Classification 1186
1.5 Fuzzy Set Theory and Fuzzy Logic and Its Application to Molecular Recognition 1216
Thomas E. Exner, Matthias Keil, and Jürgen Brickmann
1.5.1 Introduction 1216
1.5.2 Basic Principles of Fuzzy Set Theory and Fuzzy Logic 1217
1.5.2.1 Fuzzy Sets 1217
1.5.2.2 Decision-making in Fuzzy Environments 1218
1.5.2.3 Linguistic Variables 1219
1.5.2.4 LR Representation of Fuzzy Numbers and Fuzzy Intervals 1220
1.5.2.5 Fuzzy Relations 1222
1.5.2.6 Fuzzy Logic and Approximate Reasoning 1222
1.5.3 Pattern Recognition of the Human Brain 1223
1.5.4 Molecular Descriptors Using Molecular Surfaces 1226
1.5.5 Segmentation of Molecular Surfaces 1228
1.5.5.1 Non-overlapping Patches 1228
1.5.5.2 Overlapping Patches 1229
1.5.6 Identification of Binding Sites by Surface Patch Comparison 1231
1.5.7 Results and Discussion 1231
1.5.8 Conclusion 1234
References 1236

1.6 Evolutionary Algorithms and their Applications in Chemistry 1239
Alexander von Honeyer
1.6.1 Introduction 1239
1.6.2 Biological Motivation 1239
1.6.3 Classification and Comparison 1241
1.6.3.1 General Classification 1241
1.6.3.2 Comparison of Methods 1242
1.6.3.2.1 Encoding the Problem 1243
1.6.3.2.2 Selection 1244
1.6.3.2.3 Crossover/Recombination 1245
1.6.3.2.4 Mutation 1247
1.6.4 Problems Suitable for Evolutionary Algorithms 1249
1.6.5 Hybrid EA Implementations 1249
1.6.6 New Techniques 1250
1.6.6.1 Problem-specific Operators 1250
IX.2 Expert Systems 1281

Markus Hemmer

2.1 Introduction 1281
2.2 What Are Expert Systems? 1281
2.3 Some Definitions 1282
2.4 Development of Expert Systems 1283
2.4.1 Rule-based Programming 1285
2.4.2 Forward and Backward Chaining 1285
2.4.3 The Inference Engine 1286
2.5 The Fuzzy World 1286
2.5.1 Certainty Factors 1287
2.5.2 Bayesian Networks 1287
2.5.3 Hidden Markov Models 1288
2.5.4 Dempster–Shafer Theory of Evidence 1288
2.5.5 Fuzzy Logic 1289
2.6 Knowledge Acquisition 1290
2.7 Expert Systems in Spectroscopy 1290
2.8 Will Expert Systems Replace Experts? 1292
References 1293
Contents

X Applications 1295

Introduction 1297

X.1 Prediction of Physical and Chemical Properties 1300

1.1 Octanol/Water Partition Coefficients 1300

Raimund Mannel

1.1.1 Impact of Lipophilicity in Bio-Studies 1300
1.1.2 Measurement of Octanol/Water Partition Coefficients 1301
1.1.3 Calculation of Octanol/Water Partition Coefficients 1301
1.1.3.1 Substructure Approaches: Fragmental Methods 1302
1.1.3.2 Substructure Approaches: Atom Contribution Methods 1306
1.1.3.3 Whole-Molecule Approaches: MLP-Based Methods 1308
1.1.3.4 Whole-Molecule Approaches: Topology-Based Methods 1310
1.1.3.5 Whole-Molecule Approaches: Methods Based on Molecular Properties 1310
1.1.4 Summary 1311
References 1312

1.2 Quantitative Structure–Property Relationships 1314

Peter C. Jurs

1.2.1 Introduction 1314
1.2.2 Molecular Representations 1315
1.2.2.1 Topological Descriptors 1316
1.2.2.2 Electronic Descriptors 1320
1.2.2.3 Geometrical Descriptors 1321
1.2.3 Feature Selection 1324
1.2.4 Model Building 1325
1.2.5 Examples of QSPR Studies 1327
1.2.5.1 Boiling Points of Heteroatom-Containing Organic Compounds 1328
1.2.5.2 QSPR for Surface Tension of Solvents 1329
1.2.5.3 Aqueous Solubility Prediction Study 1331
1.2.6 Summary and Conclusions 1331
References 1332

1.3 Web-Based Calculation of Molecular Properties 1336

Peter Ertl and Paul Selzer

1.3.1 Introduction 1336
1.3.2 Tools for Web-Based Calculation of Molecular Properties in the Chemical Industry 1337
1.3.3 Novartis Web-Based Chemoinformatics and Molecular Processing System 1338
1.3.3.1 Property Calculation – In Silico Profiling 1339
1.3.3.2 Visualization of Molecules and 3D Surface Properties 1339
1.3.3.3 Calculation of Molecular Polar Surface Area 1341
1.3.3.4 Identification of Bioisosteric Substituents 1342
1.3.4 Free Web-Based Calculation of Molecular Properties on the Internet 1343
1.3.5 Conclusions 1347
References 1348

X.2 Structure–Spectra Correlations 1349

2.1 Correlations between Chemical Structure and Infrared Spectra 1349
Paul Seltzer
2.1.1 Introduction 1349
2.1.2 Correlation Techniques 1350
2.1.2.1 Rule-Based Approach 1350
2.1.2.2 Artificial Neural Networks (ANN) 1350
2.1.3 Digital Encoding of Infrared Spectra 1353
2.1.4 The Coding of the Chemical Structure 1354
2.1.4.1 Substructure-Based Approaches 1355
2.1.4.2 Connectivity-Based Approaches 1356
2.1.4.3 3D Structure-Based Approach 1357
2.1.4.4 Spectroscopic Data as a Structure Descriptor 1357
2.1.5 Correlating Structural and Spectroscopic Information 1357
2.1.5.1 Comparison of Data 1359
2.1.5.2 Spectrum Prediction 1360
2.1.6 Conclusions and Outlook 1364
References 1365

2.2 Correlations between Chemical Structures and NMR Data 1368
Christoph Steinbeck
2.2.1 Introduction 1368
2.2.2 NMR Databases 1369
2.2.3 Spectrum Estimation 1370
2.2.3.1 Increment Rules 1370
2.2.3.2 HOSE Codes 1371
2.2.4 Neural Networks for Shift Prediction 1372
2.2.5 Structure Validation by Quantum Mechanical Chemical Shift Calculation 1375
2.2.6 Stereochemical Aspects 1376
2.2.7 Conclusion 1377
References 1377

2.3 Computer-Assisted Structure Elucidation 1378
Christoph Steinbeck
2.3.1 Introduction 1378
2.3.2 A Prototypic CASE Process 1380
2.3.3 Spectroscopic Databases 1382
2.3.4 Constraint Generation from Spectroscopic and Other Sources 1385
2.3.4.1 General Remarks 1385
2.3.4.2 Constraints from Spectroscopic Methods 1387
2.3.4.3 Nonspectroscopic Constraints 1391
2.3.5 Structure Generators 1394
2.3.5.1 Introduction 1394
2.3.5.2 Deterministic Generators 1395
2.3.5.3 Stochastic Generators 1397
2.3.6 Examples 1399
2.3.6.1 Structure Elucidation of Monochaetin 1399
2.3.6.2 A Simulated Annealing-Based SENECA Calculation 1401
2.3.7 Conclusion 1401
References 1405
X.3 Chemical Reactions and Synthesis Design 1407

3.1 Analysis of Reaction Information 1407
   Guenter Grethe
   3.1.1 Introduction 1407
   3.1.2 Analysis and Processing of Reaction Data Information 1408
   3.1.3 Topology-Based Reaction Classification for Reaction Data Analysis 1411
   3.1.4 Analysis of Reaction Information with Kohonen Neural Networks 1421
   3.1.5 Summary 1426
   References 1427

3.2 Computer-Assisted Synthesis Design (CASP) 1428
   René Barone and Michel Chanon
   3.2.1 Introduction 1428
   3.2.2 General Approach 1430
   3.2.3 Description of Molecules 1431
   3.2.3.1 By a Connectivity Table 1431
   3.2.3.2 By Matrices 1432
   3.2.3.3 By Numerical Linear Notation 1433
   3.2.4 Description of Reactions 1434
   3.2.4.1 The Transform Approach 1434
   3.2.4.2 The BE-Matrices Approach 1437
   3.2.4.3 The Numerical Approach 1439
   3.2.5 Pruning the Synthesis Tree 1440
   3.2.5.1 Strategies Based upon a Given Transform 1441
   3.2.5.2 Starting Material-Oriented Strategy 1444
   3.2.5.3 Topological Strategies 1444
   3.2.5.4 Tactical Combinations of Transforms Strategy 1445
   3.2.5.5 Other Strategies in other Programs 1446
   3.2.6 Other Programs 1448
   3.2.6.1 Synthesis Programs 1448
   3.2.6.2 Reaction Generator 1449
   3.2.7 Conclusion 1450
   Acronyms for the Different Programs 1451
   Internet Sites 1452
   References 1452

3.3 Computer-Assisted Synthesis Design by WODCA (CASP) 1457
   M. Pförtner and M. Sitzmann
   3.3.1 Motivation 1457
   3.3.2 Introduction 1458
   3.3.3 Problem Domain 1458
3.3.3.1 Reaction Planning 1458
3.3.3.2 Reaction Prediction 1459
3.3.3.3 Synthesis Design 1459
3.3.4 Overview 1459
3.3.5 Methods and Tools 1460
3.3.5.1 Searching for Available Starting Materials: Similarity Searches 1460
3.3.5.2 Generating Synthesis Precursors: Strategic Bonds 1463
3.3.5.3 Synthesis Tree 1478
3.3.5.4 Functional Group Interconversion and Protecting Groups 1480
3.3.5.5 Synthesis Design of Combinatorial Libraries: Substructure Searches 1480
3.3.5.6 Finding Known Reactions: Reaction Substructure Searches 1483
3.3.5.7 Extraction of Knowledge from Reaction Databases 1486
3.3.6 Examples of Synthesis Studies 1488
3.3.6.1 Combinatorial Library of 4-Thiazolidinones 1488
3.3.6.2 Synthesis Design Study of Trioxifene 1498
3.3.7 Conclusion 1505
References 1507

X.4 Drug Design 1508

4.1 Chemoinformatics and the Quest for Leads in Drug Discovery 1509
  Tudor I. Oprea
  4.1.1 Chemoinformatics in the Context of Modern Pharmaceutical Research 1509
    4.1.1.1 Drug Discovery and Society 1509
    4.1.1.2 Chemoinformatics in Drug Discovery 1510
    4.1.2 The Quest for High-Activity Compounds 1512
    4.1.2.1 The Drug Discovery Paradigm 1512
    4.1.2.2 Is there a Chemical Space for High-Activity Molecules? 1514
    4.1.3 The Quest for High-Quality Leads 1518
    4.1.4 Virtual Screening in Lead Discovery 1525
    4.1.5 Conclusions 1526
    References 1527

4.2 QSAR in Drug Design 1532
  Hugo Kubinyi
  4.2.1 Introduction 1532
  4.2.2 History of QSAR 1533
  4.2.3 QSAR Methodology 1534
    4.2.3.1 A Practical Example 1534
    4.2.3.2 Nonlinear QSAR Models 1536
    4.2.3.3 Dissociation and Ionization of Acids and Bases 1537
    4.2.3.4 Scope and Limitations of Hansch and Free–Wilson Analysis 1538
    4.2.4 Biological and Physicochemical Parameters 1539
4.2.4.1 Biological Activity Values and the Additivity of Group Contributions 1539
4.2.4.2 Lipophilicity Parameters 1539
4.2.4.3 Electronic, Polar, and Other Parameters 1540
4.2.5 QSAR Applications in Drug Design 1541
4.2.5.1 Transport and Distribution of Drugs in Biological Systems 1541
4.2.5.2 Enzyme Inhibition 1543
4.2.5.3 Other Biological Activities 1545
4.2.5.4 Activity–Activity Relationships 1547
4.2.6 QSAR Model Selection and Validation 1548
4.2.6.1 Variable Selection Methods 1548
4.2.6.2 QSAR Model Validation 1549
4.2.7 2D and 3D Similarity QSAR Analyses 1550
4.2.8 Summary and Conclusions 1552
Abbreviations 1552
Notes 1552
References 1553

4.3 Comparative Molecular Field Analysis (CoMFA) 1555
Hugo Kubinyi
4.3.1 Introduction 1555
4.3.2 CoMFA Methodology 1556
4.3.2.1 History of CoMFA 1556
4.3.2.2 Steps in a Comparative Molecular Field Analysis 1556
4.3.2.3 A CoMFA Application 1557
4.3.2.4 CoMFA Applications in Drug Design 1558
4.3.3 Problems in CoMFA Applications 1559
4.3.3.1 Series Design: Training and Test Set Selection 1560
4.3.3.2 Pharmacophore Hypotheses and Alignment 1560
4.3.3.3 Box, Grid Size, and Calculation of 3D Molecular Fields 1563
4.3.3.4 PLS Analysis 1566
4.3.3.5 Model Validation 1567
4.3.4 Summary and Conclusions 1570
Abbreviations 1571
Notes 1571
References 1572

4.4 3D- and nD-QSAR Methods 1576
Emilio Xavier Esposito, Anton J. Hopfinger, and Jeffrey D. Madura
4.4.1 Introduction 1576
4.4.2 3D-QSAR Methodology 1577
4.4.2.1 Comparative Molecular Similarity Index Analysis (CoMSIA) 1578
4.4.2.2 Encoding the 3D Molecular Structure by Mathematical Transformation 1579
4.4.2.3 Comparative Molecular Moment Analysis (CoMMA) 1582
4.4.2.4 Self-Organizing Molecular Field Analysis (SOMFA) 1584
4.4.2.5 Probabilistic Receptor Potentials (PRP) 1585
4.4.3 nD-QSAR Methodology 1586
4.4.3.1 4D-QSAR 1586
4.4.3.2 EigenValue Descriptor (EVA) 1589
4.4.4 Virtual High-Throughput Screening (VHTS) 1591
4.4.4.1 Binary-QSAR 1591
4.4.4.2 Virtual 3D-Pharmacophore Screening using 4D-QSAR 1592
4.4.5 Receptor-Dependent (RD) QSAR 1593
4.4.5.1 5D-QSAR 1594
4.4.5.2 Free Energy Force Field (FEFF) 3D-QSAR 1595
4.4.5.3 Membrane-Interaction (M1) QSAR 1596
4.4.6 Conclusion 1599
Abbreviations 1600
References 1602

4.5 High-Throughput Chemistry 1604
Wendy A. Warr
4.5.1 Introduction 1604
4.5.2 History 1606
4.5.3 Mix and Split Synthesis 1608
4.5.3.1 Deconvolution 1608
4.5.3.2 Encoded Libraries 1611
4.5.4 Solid-Phase Synthesis 1614
4.5.4.1 Reviews of Chemistry 1615
4.5.4.2 Natural Product Chemistry 1615
4.5.4.3 Resins 1616
4.5.4.4 Linkers 1617
4.5.4.5 Microarrays 1619
4.5.5 Solution-Phase Synthesis 1619
4.5.5.1 Polymer-Assisted Solution-Phase Synthesis 1620
4.5.5.2 Fluorous Mixture Synthesis 1621
4.5.6 Combinatorial Biosynthesis 1622
4.5.7 Dynamic Combinatorial Chemistry 1623
4.5.8 Laboratory Instrumentation 1623
4.5.8.1 Microwave Synthesis 1624
4.5.8.2 Purification and Analysis 1625
4.5.9 Information Sources 1626
4.5.10 Data Management 1627
4.5.10.1 Library Registration 1628
4.5.10.2 Research Asset Management and Inventory Control 1630
4.5.11 Library Design 1631
4.5.12 Virtual High-Throughput Screening 1633
4.5.13 Conclusion 1634
References 1635
4.6 Molecular Diversity 1640

Michael A. Farnum, Renée L. Desjarlais, and Dimitris K. Agrafiotis

4.6.1 Introduction 1640
4.6.2 Molecular Representation 1641
4.6.3 Diversity Metrics 1642
4.6.3.1 Distance-Based Diversity Metrics 1642
4.6.3.2 Cell-Based Diversity Metrics 1650
4.6.3.3 Variance-Based Diversity Metrics 1651
4.6.3.4 Distribution-Based Distance Metrics 1653
4.6.4 Selection Algorithms 1653
4.6.4.1 Clustering 1653
4.6.4.2 Maximin 1655
4.6.4.3 Stepwise Elimination 1655
4.6.4.4 Cluster Sampling 1655
4.6.4.5 Experimental Design 1656
4.6.4.6 Partitioning Techniques 1658
4.6.4.7 Stochastic Techniques 1659
4.6.4.8 Boolean Logic 1660
4.6.4.9 Conformational Sampling 1661
4.6.4.10 Vector Analysis 1662
4.6.4.11 Cosine Coefficient 1662
4.6.4.12 Multi-objective Selection 1664
4.6.5 Dimensionality Reduction 1664
4.6.5.1 The Curse of Dimensionality 1664
4.6.5.2 Dimensionality Reduction 1665
4.6.6 Visualization 1668
4.6.6.1 Histograms and Multivariate Scatter Diagrams 1669
4.6.6.2 Chernoff Faces, Andrews' Curves, Star Diagrams, and Parallel Coordinate Plots 1669
4.6.6.3 Flower Plots 1670
4.6.6.4 Pharmacophore Plots 1670
4.6.6.5 Self-Organizing Maps 1672
4.6.6.6 Multidimensional Scaling 1675
4.6.6.7 Nonlinear Mapping 1676
4.6.6.8 Stochastic Proximity Embedding 1678
4.6.7 Application: Examples 1680
4.6.8 Conclusions 1683
Abbreviations 1683
References 1683

4.7 Pharmacophore and Drug Discovery 1687

Marc C. Nicklaus

4.7.1 Introduction 1687
4.7.2 What's Necessary for Drug Design? 1690
4.7.2.1 Pharmacophore Generation 1691
4.7.2.2 Database Building and Conformer Generation 1693
4.7.2.3 Query Generation and Submission 1695
4.7.2.4 Searches in the Database 1696
4.7.2.5 Presentation, Evaluation, and Further Processing of Hit Sets 1697
4.7.3 Software 1698
4.7.3.1 Catalyst/HipHop, HypoGen 1698
4.7.3.2 UNITY/RECEPTOR, DISCOtech, GASP; CoMFA, FlexS 1699
4.7.3.3 Other Software 1701
4.7.4 Example 1702
4.7.5 Limitations of the Pharmacophore Concept 1705
4.7.6 Current Trends and Future Perspectives 1707
References 1708

4.8 De-Novo Design Systems 1712
A. Peter Johnson
4.8.1 Introduction 1712
4.8.1.1 Structure-Based Drug Design 1713
4.8.1.2 De-Novo Design 1713
4.8.2 Generating the Constraints Model 1714
4.8.2.1 Known Receptor Structure 1714
4.8.2.2 Pharmacophores 1716
4.8.3 Finding Structures 1717
4.8.3.1 De-Novo Structure Generation 1717
4.8.4 Sorting and Selecting 1722
4.8.5 Synthetic Accessibility 1724
4.8.6 Experimental Validation 1726
4.8.7 Conclusion 1729
References 1730

4.9 The Docking Problem 1732
Christoph Sotriffer, Martin Stahl, and Gerhard Klebe
4.9.1 Introduction: The Basic Aspects of Docking 1732
4.9.2 Concepts to Address the Docking Problem 1735
4.9.2.1 Representation of the Macromolecular Receptor 1735
4.9.2.2 Ligand Handling 1737
4.9.2.3 Strategies for Searching the Configuration and Conformation Space 1740
4.9.2.4 Special Aspects of the Docking Problem 1747
4.9.3 The Scoring Aspect of the Docking Problem 1756
4.9.3.1 Classes of Scoring Functions 1756
4.9.3.2 Critical Aspects of Current Scoring Functions 1760
4.9.4 Outlook 1762
References 1763
4.10 From Structural Genomics to Drug Design: Knowledge Discovery in Crystallographic Databases to Assist Lead Discovery and Optimization 1769
  G. Klebe
  4.10.1 Introduction 1769
  4.10.2 Structure-Based Design Needs Structural Data 1770
  4.10.3 Databases in Structural Biology 1770
  4.10.4 The Database System Relibase 1772
  4.10.5 Protein-Based Features as Guidelines for the Design of New Ligands 1774
    4.10.5.1 Versatile Recognition Properties of Binding Site-Exposed Protein Functional Groups 1774
    4.10.5.2 Distribution of Preferred Protein-to-Ligand Contacts to Map “Hot Spots” of Binding in Active Sites 1775
    4.10.6 Protein Flexibility – Implications for the Selection of an Appropriate Ligand Docking Strategy 1777
    4.10.6.1 Docking to a Protein with a Rigid Binding Site 1778
    4.10.6.2 Docking to a Protein Showing Pronounced Induced-Fit Adaptations 1779
    4.10.7 Water Molecules – the Often Neglected Binding Partner in Protein–Ligand Complexes 1779
    4.10.7.1 Surface-Exposed and Deeply Buried Water Molecules in the Protein–Ligand Interface 1780
    4.10.7.2 Waters as Versatile Partners to Fill Empty Space in a Binding Pocket 1782
    4.10.8 Detection of Protein with Similar Function: Comparative Analysis of Ligand Binding Pockets 1782
    4.10.9 Secondary Structural Elements in Proteins Provide Preferred Ligand Binding Site Motifs 1783
    4.10.10 Summary and Outlook 1786
      Acknowledgements 1787
      References 1787

X.5 Chemoinformatics/Bioinformatics 1789

5.1 Prediction of Protein Structure Through Evolution 1789
  Burkhard Rost, Jinfeng Liu, Dariusz Przybylski, Rajesh Nair, Kazimierz O. Wrzeszczynski, Henry Bigelow, and Yanay Ofran
  5.1.1 Introduction 1789
  5.1.2 Protein Structure Prediction for Comparative Genomics 1792
  5.1.2.1 Synopsis 1792
  5.1.2.2 Bridging the Gap Between Sequence and Structure for Proteomes 1792
5.1.3 Sequence Alignments 1794
5.1.3.1 Synopsis 1794
5.1.3.2 Alignment Methods in Practice 1794
5.1.4 Prediction in One Dimension 1798
5.1.4.1 Synopsis 1798
5.1.4.2 Secondary Structure 1798
5.1.4.3 Solvent Accessibility 1799
5.1.4.4 Transmembrane Helices 1800
5.1.4.5 Transmembrane Strands 1802
5.1.5 Prediction in Two Dimensions 1803
5.1.5.1 Synopsis 1803
5.1.5.2 Inter-residue Contacts 1803
5.1.5.3 Inter-strand Contacts 1804
5.1.6 Prediction in Three Dimensions 1804
5.1.6.1 Synopsis 1804
5.1.6.2 Known Folds: Comparative Modeling 1805
5.1.6.3 Known Folds: Remote Homology Modeling (Threading) 1806
5.1.6.4 Unknown Folds: Ab-Initio Prediction of Structure? 1807
5.1.7 Conclusions 1808
      Abbreviations 1809
      References 1809

5.2 Sequence and Genome Bioinformatics 1812
      H. Werner Mewes

5.2.1 Introduction 1812
5.2.2 Sequence Analysis and Genome Annotation 1815
5.2.2.1 Detecting the Time Course of Evolution Encrypted into Protein
      Sequences 1815
5.2.2.2 Interpretation of Individual Sequences 1816
5.2.2.3 Homology Based Approaches to Gene Function 1816
5.2.2.4 Non-homologous Methods for Functional Assignment 1818
5.2.2.5 Large-Scale Genome Analysis 1819
5.2.2.6 Analysis of Coding Regions 1820
5.2.2.7 Genome Redundancy 1821
5.2.2.8 Analysis of Non-coding Regions 1821
5.2.2.9 Protein–Protein Interactions 1822
5.2.3 Functional Classification and Data Resources 1823
5.2.3.1 Data Resources 1823
5.2.3.2 Primary Databases 1823
5.2.3.3 Secondary Databases 1823
5.2.4 Processing Genome Information 1824
5.2.4.1 Automatic Annotation of Genomes 1824
5.2.4.2 Comparative Genomics 1825
5.2.4.3 Data Structures and Integration 1826
5.2.4.4 Data Mining 1827
5.2.5 Postgenomics: Genome-Based High-Throughput Functional Analysis 1833

5.2.5.1 Expression Analysis 1834
5.2.5.2 Proteomics 1835
5.2.5.3 Metabolomics 1836

5.2.6 Outlook 1837
Glossary 1838
References 1839

XI Future Directions 1845

J. Gasteiger

Index 1849