

Chemoinformatics

Basic Concepts and Methods

Edited by Thomas Engel and Johann Gasteiger

WILEY-VCH

Contents

Foreword *xxi*

List of Contributors *xxv*

- 1 Introduction** *1*
Thomas Engel and Johann Gasteiger
- 1.1 The Rationale for the Books *1*
- 1.2 The Objectives of Chemoinformatics *2*
- 1.3 Learning in Chemoinformatics *4*
- 1.4 Outline of the Book *5*
- 1.5 The Scope of the Book *7*
- 1.6 Teaching Chemoinformatics *8*
- References *8*
- 2 Principles of Molecular Representations** *9*
Thomas Engel
- 2.1 Introduction *9*
- 2.2 Chemical Nomenclature *11*
- 2.2.1 Non-systematic Nomenclature (Trivial Names) *11*
- 2.2.2 Systematic Nomenclature of Chemical Compounds *12*
- 2.2.3 Drawbacks of Chemical Nomenclature for Data Processing *12*
- 2.3 Chemical Notations *12*
- 2.3.1 Empirical Formulas of Inorganic and Organic Compounds *12*
- 2.3.2 Line Notations *14*
- 2.4 Mathematical Notations *14*
- 2.4.1 Introduction into Graph Theory *15*
- 2.4.2 Matrix Representations *18*
- 2.4.2.1 Adjacency Matrix *18*
- 2.4.2.2 Incidence Matrix *19*
- 2.4.2.3 Distance Matrix *20*
- 2.4.2.4 Bond Matrix *21*
- 2.4.2.5 Bond–Electron Matrix *21*
- 2.4.2.6 Summary on Matrix Representations *23*
- 2.4.3 Connection Table *23*
- 2.5 Specific Types of Chemical Structures *25*

2.5.1	General Concepts of Isomerism	25
2.5.2	Tautomerism	26
2.5.3	Markush Structures	27
2.5.4	Beyond a Connection Table Representation	28
2.5.4.1	Representation of Molecular Structures by Electron Systems	28
2.6	Spatial Representation of Structures	31
2.6.1	Representation of Configurational Isomers	32
2.6.2	Chirality	33
2.6.3	3D Coordinate Systems	36
2.7	Molecular Surfaces	37
	Selected Reading	38
	References	39
3	Computer Processing of Chemical Structure Information	43
	<i>Thomas Engel</i>	
3.1	Introduction	43
3.2	Standard File Formats for Chemical Structure Information	44
3.2.1	SMILES	44
3.2.1.1	Stereochemistry in SMILES	47
3.2.1.2	Summary on SMILES	47
3.2.2	SMARTS	47
3.2.3	SYBYL Line Notation	48
3.2.4	The International Chemical Identifier (InChI) and InChIKey	48
3.2.5	XYZ Format	50
3.2.6	Z-Matrix	51
3.2.7	The Molfile Format Family	52
3.2.7.1	Structure of a Molfile	53
3.2.7.2	Stereochemistry in the Molfile	57
3.2.7.3	Structure of an SDfile	57
3.2.8	The PDB File Format	58
3.2.8.1	Introduction/History	58
3.2.8.2	General Description	58
3.2.8.3	Analysis of a Sample PDB File	60
3.2.9	Metadata Formats	65
3.2.9.1	STAR-Based File Formats and Dictionaries	65
3.2.9.2	CIF File Format	66
3.2.9.3	mmCIF File Format	67
3.2.9.4	CML	68
3.2.9.5	CSRML	68
3.2.10	Libraries for Handling Information in Structure File Formats	69
3.3	Input and Output of Chemical Structures	70
3.3.1	Molecule Editors	72
3.3.2	Molecule Viewers	73
3.4	Processing Constitutional Information	73
3.4.1	Structure Isomers and Isomorphism	73
3.4.2	Tautomerism	74

3.4.3	Unambiguous and Biunique Representation by Canonicalization	76
3.4.3.1	The Morgan Algorithm	77
3.4.4	Ring Perception	79
3.4.4.1	Introduction	79
3.4.4.2	Graph Terminology	80
3.4.4.3	Ring Perception Strategies	81
3.5	Processing 3D Structure Information	86
3.5.1	Detection and Specification of Chirality	86
3.5.1.1	Detection of Chirality	87
3.5.1.2	Specification of Chirality	87
3.5.2	Automatic Generation of 3D Structures	90
3.5.3	Automatic Generation of Ensemble of Conformations	94
3.6	Visualization of Molecular Models	100
3.6.1	Introduction	100
3.6.2	Models of the 3D Structure	101
3.6.2.1	Wire Frame and Capped Sticks Model	101
3.6.2.2	Ball-and-Stick Model	101
3.6.2.3	Space-Filling Model	102
3.6.2.4	Crystallographic Models	102
3.6.3	Models of Biological Macromolecules	102
3.6.4	Virtual Reality	103
3.6.5	3D Printing	103
3.7	Calculation of Molecular Surfaces	103
3.7.1	Van der Waals Surface	104
3.7.2	Connolly Surface	104
3.7.3	Solvent-Accessible Surface	105
3.7.4	Enzyme Cavity Surface (Union Surface)	106
3.7.5	Isovalue-Based Electron Density Surface	106
3.7.6	Experimentally Determined Surfaces	106
3.7.7	Visualization of Molecular Surface Properties	107
3.7.8	Property-based Isosurfaces	107
3.7.8.1	Electrostatic Potentials	108
3.7.8.2	Hydrogen Bonding Potential	108
3.7.8.3	Polarizability and Hydrophobicity Potential	108
3.7.8.4	Spin Density	108
3.7.8.5	Vector Fields	108
3.7.8.6	Volumetric Properties	108
3.8	Chemoinformatic Toolkits and Workflow Environments	109
	Selected Reading	111
	References	111
4	Representation of Chemical Reactions	121
	<i>Oliver Sacher and Johann Gasteiger</i>	
4.1	Introduction	121
4.2	Reaction Equation	122
4.3	Reaction Types	123

4.4	Reaction Center and Reaction Mechanisms	125
4.5	Chemical Reactivity	126
4.5.1	Physicochemical Effects	126
4.5.1.1	Charge Distribution	126
4.5.1.2	Inductive Effect	127
4.5.1.3	Resonance Effect	127
4.5.1.4	Polarizability Effect	128
4.5.1.5	Steric Effect	128
4.5.1.6	Stereoelectronic Effects	128
4.5.2	Simple Methods for Quantifying Chemical Reactivity	128
4.5.2.1	Frontier Molecular Orbital Theory	128
4.5.2.2	Linear Free Energy Relationships	130
4.6	Learning from Reaction Information	132
4.7	Building of Reaction Databases	133
4.7.1	Contents	133
4.7.2	Reaction Data Exchange Formats	134
4.7.2.1	RXN/RDF format by MDL/Symyx	134
4.7.2.2	Reaction SMILES/SMIRKS by Daylight Chemical Information Systems	134
4.7.2.3	Chemical Markup Language	135
4.7.2.4	International Chemical Identifier for Reactions (RinChi)	135
4.7.3	Input and Output of Reactions	135
4.8	Reaction Center Perception	138
4.9	Reaction Classification	139
4.9.1	Model-Driven Approaches	139
4.9.1.1	Ugi's Scheme and Some Follow-Ups	140
4.9.1.2	InfoChem's Reaction Classification	143
4.9.2	Data-Driven Approaches	145
4.9.2.1	HORACE	145
4.9.2.2	Reaction Landscapes	146
4.10	Stereochemistry of Reactions	148
4.11	Reaction Networks	149
	Selected Reading	151
	References	152
5	The Data	155
5.1	Introduction	155
5.2	Data Types	156
5.2.1	Numerical Data	157
5.2.2	Molecular Structures	159
5.2.3	Bit Vectors	160
5.2.3.1	Hash Codes	160
5.2.3.2	Structural Keys	162
5.2.3.3	Fingerprints	163
5.2.4	Chemical Reactions	164
5.2.5	Molecular Spectra	165
5.3	Storage and Manipulation of Data	169
5.3.1	Experimental Data	169

5.3.1.1	Types of Data on Properties	170
5.3.1.2	Accuracy of the Data	170
5.3.2	Data Storage and Exchange	171
5.3.2.1	DAT File	171
5.3.2.2	JCAMP-DX	171
5.3.2.3	Predictive Model Markup Language (PMML)	172
5.3.3	Real-World Data	173
5.3.3.1	Data Complexity	173
5.3.3.2	Outliers and Redundant Objects	174
5.3.4	Data Transformation	175
5.3.4.1	Fast Fourier Transformation	175
5.3.4.2	Wavelet Transformation	175
5.3.5	Preparation of Datasets for Building of Models and Validations of Their Quality	176
5.4	Conclusions	177
	Selected Reading	178
	References	179
6	Databases and Data Sources in Chemistry	185
	<i>Engelbert Zass and Thomas Engel</i>	
6.1	Introduction	185
6.2	Chemical Literature and Databases	186
6.2.1	Classification of Chemical Literature	186
6.2.2	The Origin of Chemical Databases	187
6.2.3	Evolution of Database Systems and User Interfaces	187
6.3	Major Chemical Database Systems	188
6.3.1	SciFinder	188
6.3.2	Reaxys	189
6.3.3	SciFinder <i>versus</i> Reaxys	190
6.4	Compound Databases	191
6.4.1	2D Structures	191
6.4.1.1	Searching Organic Compounds	192
6.4.1.2	Searching Inorganic and Coordination Compounds	194
6.4.2	Sequences of Biopolymers	195
6.4.3	3D Structures	198
6.4.4	Catalog Databases	200
6.5	Databases with Properties of Compounds	200
6.5.1	Physical Properties	201
6.5.2	Thermodynamic and Thermochemical Data	202
6.5.3	Spectra	204
6.5.3.1	Spectroscopic Databases	205
6.5.3.2	Compound Databases with Spectroscopic Information	205
6.5.4	Biological, Environmental, and Safety Information Sources	206
6.5.4.1	Biological Information	207
6.5.4.2	Pharmaceutical and Medical Information	208
6.5.4.3	Toxicity, Environmental, and Safety Information	209
6.6	Reaction Databases	210

6.6.1	Comprehensive Reaction Databases	210
6.6.2	Synthetic Methodology Databases	212
6.7	Bibliographic and Citation Databases	212
6.7.1	Bibliographic Databases	213
6.7.1.1	Special Bibliographic Databases	213
6.7.1.2	Patent Bibliographic Databases	214
6.7.1.3	Searching Bibliographic Databases	216
6.7.1.4	Linking to Full Text	216
6.7.2	Citation Databases	217
6.7.2.1	General Citation Databases	218
6.7.2.2	Patent Citation Databases	219
6.8	Full-Text Databases	219
6.8.1	Electronic Journals	219
6.8.2	Patents	220
6.8.3	Lexika and Encyclopedias	221
6.9	Architecture of a Structure-Searchable Database	222
	Selected Reading	224
	References	224
7	Searching Chemical Structures	231
	<i>Nikolay Kochev, Valentin Monev, and Ivan Bangov</i>	
7.1	Introduction	231
7.2	Full Structure Search	232
7.3	Substructure Search	235
7.3.1	Basic Concepts	235
7.3.2	Backtracking Algorithm	236
7.3.3	Optimization of the Backtracking Algorithm	238
7.3.4	Screening	239
7.3.5	Superstructure Searching	241
7.3.6	Automorphism Searching	241
7.3.7	Maximum Common Substructure Searching	242
7.3.8	Specific Line Notations for Substructure Searching	243
7.3.9	Chemotypes for Database Searching	244
7.4	Similarity Search	245
7.4.1	Similarity Basics	245
7.4.2	Similarity Measures	247
7.4.3	Descriptor Selection and Coding	249
7.4.4	Similarity Measures Based on Maximum Common Substructure	250
7.5	Three-Dimensional Structure Search Methods	250
7.5.1	Pharmacophore Searching	251
7.5.2	3D Similarity Searching	252
7.6	Sequence Searching in Protein and Nucleic Acid Databases	254
7.6.1	Sequence Similarity Definition	255
7.6.2	Dynamic Programming Algorithm	256
7.6.3	Fast Sequence Searching in Large Databases	258
7.7	Summary	259

Selected Reading 261

References 262

8 Computational Chemistry 267

8.1 Empirical Approaches to the Calculation of Properties 269

Johann Gasteiger

8.1.1 Introduction 269

8.1.2 Additivity of Atomic Contributions 269

8.1.3 Attenuation Models 271

8.1.3.1 Calculation of Charge Distribution 271

8.1.3.2 Polarizability Effect 275

Selected Reading 277

References 277

8.2 Molecular Mechanics 279

Harald Lanig

8.2.1 Introduction 279

8.2.2 No Force Field Calculation without Atom Types 280

8.2.3 The Functional Form of Common Force Fields 281

8.2.3.1 Bond Stretching 282

8.2.3.2 Angle Bending 283

8.2.3.3 Torsional Terms 284

8.2.3.4 Out-of-Plane Bending 285

8.2.3.5 Electrostatic Interactions 286

8.2.3.6 Van der Waals Interactions 287

8.2.3.7 Cross Terms 289

8.2.3.8 Advanced Interatomic Potentials and Future Development 290

8.2.4 Available Force Fields 291

8.2.4.1 Force Fields for Small Molecules 292

8.2.4.2 Force Fields for Biomolecules 293

Selected Readings 296

References 296

8.3 Molecular Dynamics 301

Harald Lanig

8.3.1 Introduction 301

8.3.2 The Continuous Movement of Molecules 302

8.3.3 Methods 302

8.3.3.1 Algorithms 303

8.3.3.2 Ways for Speeding up the Calculations 304

8.3.3.3 Solvent Effects 305

8.3.3.4 Periodic Boundary Conditions 308

8.3.4 Constant Energy, Temperature, or Pressure? 308

8.3.5	Long-Range Forces	310
8.3.6	Application of Molecular Dynamics Techniques	311
8.3.7	Future Perspectives	315
	Selected Readings	317
	References	317
8.4	Quantum Mechanics	320
	<i>Tim Clark</i>	
8.4.1	Hückel Molecular Orbital Theory	320
8.4.2	Semiempirical MO Theory	324
8.4.3	<i>Ab Initio</i> Molecular Orbital Theory	327
8.4.4	Density Functional Theory	332
8.4.5	Properties from Quantum Mechanical Calculations	334
8.4.5.1	Net Atomic Charges	334
8.4.5.2	Dipole and Higher Multipole Moments	335
8.4.5.3	Polarizabilities	335
8.4.5.4	Orbital Energies	336
8.4.5.5	Surface Descriptors	336
8.4.5.6	Local Ionization Potential	336
8.4.6	Quantum Mechanical Techniques for Very Large Molecules	337
8.4.6.1	Linear Scaling Methods	337
8.4.6.2	Hybrid QM/MM Calculations	338
8.4.7	The Future of Quantum Mechanical Methods in Chemoinformatics	338
	Selected Reading	340
	References	341
9	Modeling and Prediction of Properties (QSPR/QSAR)	345
	<i>Johann Gasteiger</i>	
10	Calculation of Structure Descriptors	349
	<i>Lothar Terfloth and Johann Gasteiger</i>	
10.1	Introduction	349
10.1.1	QSPR/QSAR Modeling	349
10.1.2	Overview	349
10.1.3	Classification of Compounds and Similarity Searching	350
10.1.4	Definition of the Terms “Structure Descriptor” and “Molecular Descriptor”	351
10.1.5	Classification of Structure Descriptors	351
10.1.6	Structure Descriptors with a Fixed Length	351
10.2	Structure Descriptors for Classification and Similarity Searching	352
10.2.1	2D Structure Descriptors (Topological Descriptors)	352
10.2.1.1	Structural Keys	352
10.2.1.2	Fingerprints	353
10.2.1.3	Distance and Similarity Measures	354

10.2.1.4	Chemotypes: Data Mining for Compounds with Structural Features	356
10.2.1.5	Multilevel Neighborhoods of Atoms	358
10.2.1.6	Descriptors from Shannon Entropy Calculations	359
10.2.1.7	Chemically Advanced Template Search (CATS2D) Descriptors	360
10.2.1.8	Descriptors from Chemical Bond Information	360
10.2.2	3D Descriptors	361
10.2.2.1	Geometric Atom Pair Descriptors	361
10.2.2.2	CATS3D and CHARGE3D	361
10.2.2.3	Pharmacophores	362
10.2.3	Field-Based Molecular Similarity	362
10.2.3.1	Electron Density	362
10.2.3.2	General Field-Based Similarity Indices	363
10.3	Structure Descriptors for Quantitative Modeling	363
10.3.1	0-D Molecular Descriptors	363
10.3.2	1D Molecular Descriptors	363
10.3.3	2D Molecular Descriptors (Topological Descriptors)	365
10.3.3.1	Single-Valued Descriptors	365
10.3.3.2	Topological Descriptors as Vectors	366
10.3.4	3D Descriptors	369
10.3.4.1	3D Structure Generation	369
10.3.4.2	3D Autocorrelation Vector	370
10.3.4.3	3D Molecule Representation of Structures Based on Electron Diffraction Code (3D MoRSE Code)	370
10.3.4.4	Radial Distribution Function Code	371
10.3.4.5	Other 3D Descriptors	375
10.3.5	Chirality Descriptors	375
10.3.5.1	Chirality Codes	376
10.3.5.2	Conformation-Independent Chirality Code (CICC)	376
10.3.5.3	Conformation-Dependent Chirality Code (CDCC)	377
10.3.5.4	Descriptors of Molecular Shape and Molecular Surfaces	377
10.3.5.5	Global Shape Descriptors	378
10.3.5.6	Autocorrelation of Molecular Surface Properties	378
10.3.5.7	2D Maps of Molecular Surfaces	379
10.3.5.8	Charged Partial Surface Area	382
10.3.6	Field-Based Methods	383
10.3.6.1	Comparative Molecular Field Analysis (CoMFA)	383
10.3.6.2	Comparative Molecular Similarity Analysis (CoMSIA)	384
10.3.6.3	3D Molecular Interaction Fields	384
10.3.7	Descriptors for an Ensemble of Conformations (4D Descriptors)	384
10.3.7.1	4D-QSAR	384
10.3.8	Quantum Chemical Descriptors	385
10.4	Descriptors That Are Not Calculated from the Chemical Structure	385
10.5	Summary and Outlook	387

	Selected Reading	390
	References	390
11	Data Analysis and Data Handling (QSPR/QSAR)	397
11.1	Methods for Multivariate Data Analysis	399
	<i>Kurt Varmuza</i>	
11.1.1	Introduction into Multivariate Data Analysis	399
11.1.1.1	Aims	399
11.1.1.2	Notation and Symbols	400
11.1.2	Basics of Statistical Data Evaluation	401
11.1.2.1	Data Distribution, Central Value, and Spread	401
11.1.2.2	Correlation	404
11.1.2.3	Discrimination	405
11.1.3	Multivariate Data	406
11.1.3.1	Overview	406
11.1.3.2	Preprocessing	407
11.1.3.3	Distances and Similarities	408
11.1.3.4	Linear Latent Variables	410
11.1.4	Evaluation of Empirical Models	412
11.1.4.1	Overview	412
11.1.4.2	Optimum Model Complexity	412
11.1.4.3	Performance Criteria for Calibration Models	413
11.1.4.4	Performance Criteria for Classification Models	414
11.1.4.5	Cross-Validation	415
11.1.4.6	Bootstrap	416
11.1.5	Exploration: Analyzing the Independent Variables	417
11.1.5.1	Overview	417
11.1.5.2	Principal Component Analysis (PCA)	417
11.1.5.3	Nonlinear Mapping	419
11.1.5.4	Cluster Analysis	419
11.1.5.5	Example: Exploratory Data Analysis of Mass Spectra from Meteorite Samples	421
11.1.6	Calibration: Building a Quantitative Model	423
11.1.6.1	Overview	423
11.1.6.2	Ordinary Least Squares (OLS) Regression	424
11.1.6.3	Principal Component Regression (PCR)	424
11.1.6.4	Partial Least Squares (PLS) Regression	425
11.1.6.5	Variable Selection	426
11.1.6.6	Example: Prediction of Gas Chromatographic Retention Indices for Polycyclic Aromatic Hydrocarbons	427
11.1.7	Classification: Discriminating Samples	428
11.1.7.1	Overview	428
11.1.7.2	Linear Discriminant Analysis (LDA)	430
11.1.7.3	Discriminant Partial Least Squares (D-PLS) Analysis	430

- 11.1.7.4 *k*-Nearest Neighbor (KNN) Classification 430
- 11.1.7.5 Support Vector Machine (SVM) 431
- 11.1.7.6 Classification Trees (CART) 432
- 11.1.7.7 Example: Classification of Meteorite Samples Using Mass Spectral Data 432
 - Acknowledgements 434
 - Selected Reading 435
 - References 435

11.2 Artificial Neural Networks (ANNs) 438

Jure Zupan

- 11.2.1 How to Learn a New Method? 438
- 11.2.2 Multivariate Representation of Data 439
- 11.2.3 Overview of Artificial Neural Networks (ANNs) 442
- 11.2.4 Error Back-Propagation ANNs 443
- 11.2.5 Kohonen and Counter-Propagation ANN 445
- 11.2.6 Training of the ANN: Adapting the Weights 448
- 11.2.7 Controlling Model Complexity and Optimizing Predictivity 450
- 11.2.8 Few General Remarks about ANNs 450
 - Selected Reading 451
 - References 451

11.3 Deep and Shallow Neural Networks 453

David A. Winkler

- 11.3.1 Drug Design in the Era of Big Data and Artificial Intelligence (AI) 453
- 11.3.2 Deep Learning 454
- 11.3.3 Controlling Model Complexity and Optimizing Predictivity Using Regularization 455
- 11.3.4 Universal Approximation Theorem 458
- 11.3.5 Do QSAR Models Generated by Neural Networks Meet the Requirements of the Universal Approximation Theorem? 458
- 11.3.6 Comparison of the Performance of Deep and Shallow Regularized Neural Networks on Drug Datasets 459
- 11.3.7 A Few General Remarks about Neural Networks for Drug Discovery 460
 - Selected Reading 462
 - References 462

12 QSAR/QSPR Revisited 465

Alexander Golbraikh and Alexander Tropsha

- 12.1 Best Practices of QSAR Modeling 466
 - 12.1.1 Introduction 466
 - 12.1.2 Key Concepts 467
 - 12.1.3 Predictive QSAR Modeling Workflow 468

- 12.1.4 Dataset Curation 469
- 12.1.5 Modelability Studies 470
- 12.1.6 Development of QSAR Models: Internal and External Validation 471
- 12.1.7 Prediction Accuracy Criteria for QSAR Models for a Continuous Response Variable 472
- 12.1.8 Prediction Accuracy Criteria for Category QSAR Models 473
- 12.1.9 Time-Split Validation 475
- 12.1.10 Validation by Y-Randomization 475
- 12.1.11 Applicability Domain of QSAR Models 475
 - 12.1.11.1 Leverage AD for Regression QSAR Models 476
 - 12.1.11.2 Residual Standard Deviation (RSD) as AD 476
 - 12.1.11.3 Other widely Used ADs 476
- 12.1.12 Ensemble Modeling 478
- 12.1.13 Model Interpretation: Structural Alerts 478
- 12.1.14 Virtual Screening 479
- 12.1.15 Conclusions 480
- 12.2 The Data Science of QSAR Modeling 480
 - 12.2.1 Introduction 480
 - 12.2.2 Data Curation: Trust but Verify! 482
 - 12.2.3 Models as Decision Support Tools 487
 - 12.2.4 Conclusions 487
 - Selected Reading 489
 - References 489
- 13 Bioinformatics 497**
Heinrich Sticht
 - 13.1 Introduction 497
 - 13.2 Sequence Databases 499
 - 13.2.1 GenBank 499
 - 13.2.2 UniProt 501
 - 13.3 Searching Sequence Databases 502
 - 13.3.1 Tools for Sequence Database Searches 503
 - 13.3.2 Scoring Matrices 503
 - 13.3.3 Interpretation of the Results of a Database Search 507
 - 13.4 Characterization of Protein Families 509
 - 13.4.1 Multiple Sequence Alignment 509
 - 13.4.2 Sequence Signatures 512
 - 13.5 Homology Modeling 515
 - Selected Reading 520
 - References 520
- 14 Future Directions 525**
Johann Gasteiger
 - 14.1 Access to Chemical Information 525

14.2	Representation of Chemical Compounds	527
14.3	Representation of Chemical Reactions	527
14.4	Learning from Chemical Information	528
14.5	Training in Chemoinformatics	529

Answers Section 531

Index 555