Contents

Preface V

A Personal Foreword VII
H. van de Waterbeemd

1 Introduction 1
H. van de Waterbeemd

1.1 Quantitative Molecular Design 2
1.2 Chemometrics 3
1.3 The Hansch Approach 5
1.4 Modern Chemometric Approaches in Molecular Design 7
1.5 Software 8
1.5.1 General Statistical Packages 8
1.5.2 Specialized Software for SPC Studies 9
References 10

2 Molecular Concepts 15

2.1 Representations of Molecules 15
P. C. Jurs, S. L. Dixon and L. M. Egolf

2.1.1 Introduction 16
2.1.2 Substituent Constants 17
2.1.2.1 Electronic Substituent Constants 17
2.1.2.2 The Hydrophobic Substituent Constant, π 21
2.1.2.3 Partition Coefficient – Log P 21
2.1.2.4 Steric Substituent Constants 23
2.1.3 Whole Molecule Representations 25
2.1.3.1 Topological Descriptions 25
2.1.3.2 Electronic Whole Molecule Descriptors 31
2.1.3.3 Geometric Descriptors 33
References 35

2.2 Atom-Level Descriptors for QSAR Analyzes 39
L. B. Kier

2.2.1 Introduction 39
2.2.2 An Atom-Level Description of Structure 40

2.2.2.1 The Field 40
2.2.2.2 The Intrinsic State of an Atom 41
2.2.2.3 The Field Effect on Each Atom 43
2.2.3 Strategies for Use of E-State Indices 44
2.2.4 Examples of E-State QSAR 44
2.2.4.1 MAO Inhibition with Hydrazides 44
2.2.4.2 Adenosine A_1 Inhibitors 45
2.2.4.3 Anesthetic Concentration of Haloalkanes 45
2.2.4.4 Odor Sensitivity of Pyrazines 46
2.2.5 Conclusions 46
References 47

3 Experimental Design in Synthesis Planning and Structure-Property Correlations 49

3.1 Experimental Design 49
V. Austel

3.1.1 The Importance of Experimental Design in Medicinal Chemistry 49
3.1.2 Strategy in Experimental Design 50
3.1.3 Selected Methods for Experimental Design 51
3.1.3.1 Methods for the Direct Optimization of Lead Compounds 51
3.1.3.2 Methods for the Systematic Investigation of Parameter Space 53
3.1.3.3 Choice of Molecular Descriptors 61
3.1.4 Summary and Conclusion 62
References 62

3.2 Applications of Statistical Experimental Design and PLS Modeling in QSAR 63
M. Sjöström and L. Eriksson

3.2.1 Introduction 63
3.2.2 A Strategy for QSAR Development in Drug Design 64
3.2.2.1 Formulation of Classes of Similar Compounds (Step 1) 64
3.2.2.2 Structural Description and Definition of Design Variables (Step 2) 65
3.2.2.3 Selection of the Training Set of Compounds (Step 3) 66
3.2.2.4 Biological Testing (Step 4) 67
3.2.2.5 QSAR Development (Step 5) 67
3.2.2.6 Validation and Predictions for Non-Tested Compounds (Step 6) 68
3.2.3 Examples of Design and PLS Modeling 68
3.2.3.1 Bradykinin Potentiating Pentapeptides 69
3.2.3.2 Dipeptides (Inhibiting the Angiotensin Converting Enzyme) 73
3.2.3.3 Dipeptides (Bitter Tasting) 76
3.2.3.4 Mimetics 78
3.2.3.5 Haloalkanes 80

3.2.3.6 Dibenzofurans 82
3.2.3.7 Monosubstituted Benzenes 84
3.2.3.8 Corrosive Carboxylic Acids 86
3.2.4 Discussion and Conclusions 88
Software Used 89
Acknowledgements 89
References 90

3.3 Total Response Surface Optimization 91
L. B. Hall

3.3.1 Background 91
3.3.2 Representation of a Response Surface 92
3.3.3 Structure Descriptors from Chemical Graph Theory 94
3.3.4 Examples 95
3.3.4.1 Neurotoxicity of Fluorophosphorous Compounds 95
3.3.4.2 Bioconcentration of Chlorinated Phenyls and Biphenyls 99
3.3.5 Conclusions 101
References 102

3.4 Disjoint Principle Properties of Organic Substituents 103
H. van de Waterbeemd, G. Constantino, S. Clementi, G. Cruciano and R. Valigi

3.4.1 The Design of Molecular Diversity 104
3.4.1.1 Combinatorial Chemistry 104
3.4.1.2 Statistical Experimental Design 104
3.4.2 Substituent Descriptors 104
3.4.2.1 Principal Properties (PPs) 105
3.4.2.2 Disjoint Principal Properties (DPPs) 107
3.4.2.3 Selection of Representative Substituents 107
3.4.3 An Example of DPPs in Design and Analysis 109
3.4.4 Conclusions 110
References 111

4 Multivariate Data Analysis of Chemical and Biological Data 113

4.1 Principal Component and Factor Analysis 113
R. Franke and A. Gruska

4.1.1 Introduction 115
4.1.2 Basic Principles 116
4.1.2.1 Principal Component Analysis 116
4.1.2.2 Factor Analysis 122
4.1.3 Applications of Principal Component and Factor Analysis in Medicinal Chemistry 124

4.1.3.1 Data from Parallel Biological Tests 125
4.1.3.2 Pharmacokinetic Data and Time Series 136
4.1.3.3 Analysis of QSAR Descriptors 141
References 158

4.2 Graphical Analysis as an Aid in Medicinal Chemistry 165
J. Devillers and D. Chessel

4.2.1 Introduction 165
4.2.2 Graphical Displays 166
4.2.2.1 Overall Strategy 166
4.2.2.2 Techniques in Graphical Design 166
4.2.2.3 Visual Perception 167
4.2.2.4 Software Availability 168
4.2.3 The Key Role of Graphics in Co-Structure Analysis 168
4.2.3.1 Background 168
4.2.3.2 Example: Structure-Reactivity Relationships for Unsaturated Dialdehydes 169
4.2.4 Conclusion 174
References 176

4.3 SIMCA Pattern Recognition and Classification 179
W. J. Dunn and S. Wold

4.3.1 Introduction 180
4.3.2 SIMCA Pattern Recognition 181
4.3.3 Steps in Pattern Recognition Study 183
4.3.4 Establishing the Training Sets 183
4.3.4.1 Consideration of Conformation and Alignment of Flexible Compounds 183
4.3.4.2 Novel Descriptors for Peptide QSAR 184
4.3.5 Symmetric and Asymmetric Data Structures 187
4.3.6 Variable Selection 188
4.3.7 Determining the Model Complexity 188
4.3.8 Developing, Optimizing and Validating Classification Rules 189
4.3.9 Discussion 192
References 193

4.4 PLS for Multivariate Linear Modeling 195
S. Wold

4.4.1 Introduction 196
4.4.2 Objectives and Data Homogeneity 197
4.4.3 The QSAR Model 200
4.4.3.1 The Conceptual Model 200
4.4.3.2 Transformation, Scaling and Centering 201
4.4.3.3 The PLS Model 202

4.4.3.4 Interpretation of the PLS Model 203
4.4.3.5 The Number of PLS Components, A 205
4.4.3.6 Model Validation 206
4.4.3.7 PLS Algorithms 206
4.4.4 The First PLS Analysis of the AA Data 206
4.4.4.1 Score Plots 207
4.4.4.2 PLS Weights w and c 207
4.4.4.3 A Comparison of PLS with Multiple Linear Regression (MLR) 209
4.4.4.4 Conclusion of the First Analysis 210
4.4.4.5 Conclusion of the Second Analysis 211
4.4.5 Selection of Important Variables 213
4.4.6 Residuals 214
4.4.7 Conclusions 215
4.4.8 Summary: How to Develop and Interpret a PLS Model 216
4.4.9 Conclusions and Discussion 217
Acknowledgements 218
References 218

4.5 Spectral Mapping of Drug Test Specificities 219
P. Lewi

4.5.1 Activity, Potency, Sensibility and Specificity 220
4.5.2 Historical Background 222
4.5.3 Spectral Map Analysis 233
4.5.4 Spectral Map of Neuroleptics 241
4.5.5 Mathematical Description of SMA 245
4.5.6 Discussion 249
Acknowledgement 252
References 252

4.6 Display of Multivariate Data Using Non-Linear Mapping 255
J. Devillers

4.6.1 Linear and Non-Linear Methods in SAR and QSAR Studies 255
4.6.2 Non-Linear Mapping Algorithms 256
4.6.3 Interpretation of Non-Linear Maps 257
4.6.4 Drawbacks and Limitations 257
4.6.5 An Illustrative Example in Medicinal Chemistry 259
4.6.6 Software Availability 261
4.6.7 Concluding Remarks 262
References 262

4.7 The Use of Canonical Correlation Analysis 265
M. G. Ford and D. W. Salt

4.7.1 Introduction 266
4.7.2 Formulation of the Problem 266

4.7.3 Features of Canonical Correlation Analysis (CCA) 268
4.7.3.1 Procedure for CCA 268
4.7.3.2 Canonical Weights and Canonical Loadings 268
4.7.3.3 Proportion of Explained Variance 269
4.7.3.4 Redundancy Coefficient 269
4.7.3.5 Hypothesis Testing 270
4.7.4 The Application of CCA to QSAR Problems 271
4.7.4.1 Pharmacokinetics of Pyrethroid Insecticides in Insects 271
4.7.4.2 The Relationship between the Physico-Chemical Properties of Pyrethroids and Pharmacokinetics, Pharmacodynamics and Toxicity 273
4.7.4.3 Variable Deletion Procedures and CCA 275
4.7.4.4 Mapping the Toxicological and Physico-Chemical Hyperspaces 278
4.7.5 Useful Features of CCA for the Design of Biologically Active Compounds 280
4.7.6 Predicting Biological Activity 281
4.7.7 The Advantages and Disadvantages of Using CCA in QSAR Studies 281
References 282

4.8 Discriminant Analysis for Activity Prediction 283
H. van de Waterbeemd

4.8.1 Theoretical Background 283
4.8.2 Descriptor Selection 286
4.8.3 Chance Correlations with Discriminant Analysis 286
4.8.4 Validation 287
4.8.5 Examples 287
4.8.5.1 Mode of Action of Pyrimidine Folic Acid Antagonists 287
4.8.5.2 Antiulcer Benzoguanamines 289
4.8.5.3 Calmodulin Inhibitors 289
4.8.5.4 MAO Inhibitors 290
4.8.6 Conclusions 291
References 292

4.9 Cluster Significance Analysis 295
J. W. McFarland and D. J. Gans

4.9.1 Introduction 295
4.9.2 Theoretical Background to CSA 296
4.9.2.1 Parameter Focusing 296
4.9.2.2 A Graphical Explanation 297
4.9.2.3 Calculations 297
4.9.2.4 Choosing Among Sets of Parameters (Sequential CSA) 298
4.9.3 Practical Considerations 299
4.9.3.1 Software Availability 299
4.9.3.2 Dividing the Dependent Variables into Two Categories 300

4.9.4 Examples 301
4.9.4.1 CSA1: The Hypothetical Example 301
4.9.4.2 CSA2: Inhibition of Tetracycline Efflux Antiport Protein 302
4.9.4.3 Sequential CSA: Sulfamate Sweetening Agents 305
4.9.4.4 Literature Examples 306
References 307

5 Statistical Validation of QSAR Results 309

5.1 Validation Tools 309
S. Wold and L. Eriksson

5.1.1 Introduction 309
5.1.2 Examples 310
5.1.3 Four Tools for Model Validation 311
5.1.3.1 Tool 1: Randomization of the Responses into an Array of Reordered Variables 311
5.1.3.2 Tool 2: Cross-Validation 312
5.1.3.3 Tool 3: Splitting of Parent Data Set into Training and Validation Sets 313
5.1.3.4 Tool 4: External Validation Using a Designed Validation Set 313
5.1.4 Results 313
5.1.4.1 The Epoxide Example 314
5.1.4.2 The Haloalkane Example 316
5.1.5 Concluding Remarks 317
References 318

5.2 How To Choose The Proper Statistical Method 319
S. Clementi and S. Wold

5.2.1 Introduction 320
5.2.2 Problem Formulation 321
5.2.2.1 Parameter Selection 321
5.2.2.2 Regression Methods 322
5.2.2.3 Model Evaluation 324
5.2.3 The SIMCA Philosophy 325
5.2.3.1 Factorial and D-Optimal Designs in PPs 326
5.2.3.2 Validation and Variable Selection 328
5.2.4 Other PLS Codes 331
5.2.5 3D QSAR 332
5.2.6 Conclusions 334
Acknowledgements 336
References 336

Index 339