AN INTRODUCTION TO CHEMOINFORMATICS

by

ANDREW R. LEACH

GlaxoSmithKline Research and Development,
Stevenage, U.K.

and

VALERIE J. GILLET

Department of Information Studies,
University of Sheffield, U.K.

KLUWER ACADEMIC PUBLISHERS

DORDRECHT / BOSTON / LONDON

CONTENTS

PREFACE xi

ACKNOWLEDGEMENTS xv

CHAPTER 1. REPRESENTATION AND MANIPULATION OF 2D MOLECULAR STRUCTURES 1

1. Introduction 1
2. Computer Representations of Chemical Structures 1
   2.1 Graph Theoretic Representations of Chemical Structures 2
   2.2 Connection Tables and Linear Notations 4
   2.3 Canonical Representations of Molecular Structures 6
3. Structure Searching 8
4. Substructure Searching 9
   4.1 Screening Methods 11
   4.2 Algorithms for Subgraph Isomorphism 13
   4.3 Practical Aspects of Structure Searching 17
5. Reaction Databases 19
6. The Representation of Patents and Patent Databases 21
7. Relational Database Systems 24
8. Summary 26

CHAPTER 2. REPRESENTATION AND MANIPULATION OF 3D MOLECULAR STRUCTURES 27

1. Introduction 27
2. Experimental 3D databases 28
3. 3D Pharmacophores 31
4. Implementation of 3D database Searching 33

5. Theoretical 3D Databases 34
5.1 Structure-Generation Programs 34
5.2 Conformational Search and Analysis 35
5.3 Systematic Conformational Search 35
5.4 Random Conformational Search 38
5.5 Other Approaches to Conformational Search 39
5.6 The Generation of Distance Keys for Flexible Molecules 39
6. Methods to Derive 3D Pharmacophores 40
6.1 Pharmacophore Mapping using Constrained Systematic Search 41
6.2 Pharmacophore Mapping using Clique Detection 43
6.3 The Maximum Likelihood Method for Pharmacophore Mapping 45
6.4 Pharmacophore Mapping using a Genetic Algorithm 47
6.5 Practical Aspects of Pharmacophore Mapping 49
7. Applications of 3D Pharmacophore Mapping and 3D Database Searching 51
8. Summary 52

CHAPTER 3. MOLECULAR DESCRIPTORS 53
1. Introduction 53
2. Descriptors Calculated from the 2D Structure 54
2.1 Simple Counts 54
2.2 Physicochemical Properties 54
2.3 Molar Refractivity 57
2.4 Topological Indices 57
2.5 Kappa Shape Indices 59
2.6 Electrotopological State Indices 62
2.7 2D Fingerprints 63
2.8 Atom-Pairs and Topological Torsions 63
2.9 BCUT Descriptors 64
3. Descriptors Based on 3D Representations 65
3.1 3D Fragment Screens 65
3.2 Pharmacophore Keys 66
3.3 Other 3D Descriptors 67
4. Data Verification and Manipulation 68
4.1 Data Spread and Distribution 68
4.2 Scaling 69
4.3 Correlations 69
4.4 Reducing the Dimensionality of a Data Set: Principal Components Analysis 71
5. Summary 75

CHAPTER 4. COMPUTATIONAL MODELS 77
1. Introduction 77
2. Historical Overview 77
3. Deriving a QSAR Equation: Simple and Multiple Linear Regression 79
3.1 The Squared Correlation Coefficient, R^2 81
3.2 Cross-Validation 82
3.3 Other Measures of a Regression Equation 83
4. Designing a QSAR "Experiment" 85
4.1 Selecting the Descriptors to Include 86
4.2 Experimental Design 86
4.3 Indicator Variables 88
4.4 Free-Wilson Analysis 89
4.5 Non-Linear Terms in QSAR Equations 89
4.6 Interpretation and Application of a QSAR Equation 90
5. Principal Components Regression 91
6. Partial Least Squares 92
7. Molecular Field Analysis and Partial Least Squares 97
8. Summary 101

CHAPTER 5. SIMILARITY METHODS 103
1. Introduction 103
2. Similarity Based on 2D Fingerprints 105
3. Similarity Coefficients 106
3.1 Properties of Similarity and Distance Coefficients 108
4. Other 2D Descriptor Methods 110
4.1 Maximum Common Subgraph Similarity 110
4.2 Reduced Graph Similarity 111
5. 3D Similarity 112
5.1 Alignment-Independent Methods 113
5.2 Alignment Methods 115
5.3 Field-Based Alignment Methods 115
5.4 Gnomonic Projection Methods 118
5.5 Finding the Optimal Alignment 119
5.6 Comparison and Evaluation of Similarity Methods 119
6. Summary 122

CHAPTER 6. SELECTING DIVERSE SETS OF COMPOUNDS 123
1. Introduction 123
2. Cluster Analysis 124
2.1 Hierarchical Clustering 126
2.2 Selecting the Appropriate Number of Clusters 128
2.3 Non-Hierarchical Clustering 130

2.4 Efficiency and Effectiveness of Clustering Methods 131
3. Dissimilarity-Based selection methods 133
3.1 Efficiency and Effectiveness of DBCS Methods 135
4. Cell-Based Methods 137
4.1 Partitioning Using Pharmacophore Keys 141
5. Optimisation Methods 142
6. Comparison and Evaluation of Selection Methods 144
7. Summary 145

CHAPTER 7. ANALYSIS OF HIGH-THROUGHPUT SCREENING DATA 147
1. Introduction 147
2. Data Visualisation 150
2.1 Non-Linear Mapping 151
3. Data Mining Methods 153
3.1 Substructural Analysis 154
3.2 Discriminant Analysis 155
3.3 Neural Networks 157
3.4 Decision Trees 160
4. Summary 163

CHAPTER 8. VIRTUAL SCREENING 165
1. Introduction 165
2. "Drug-Likeness" and Compound Filters 167
3. Structure-Based Virtual Screening 170
3.1 Protein-Ligand Docking 171
3.2 Scoring Functions for Protein-Ligand Docking 174
3.3 Practical Aspects of Structure-Based Virtual Screening 179
4. The Prediction of ADMET Properties 181
4.1 Hydrogen Bonding Descriptors 182
4.2 Polar Surface Area 183
4.3 Descriptors Based on 3D Fields 185
4.4 Toxicity Prediction 186
5. Summary 188

CHAPTER 9. COMBINATORIAL CHEMISTRY AND LIBRARY DESIGN 189
1. Introduction 189
2. Diverse and Focussed Libraries 191
3. Library Enumeration 193
4. Combinatorial Library Design Strategies 195
4.1 Monomer-Based Selection 195
4.2 Product-Based Selection 196

5. Approaches to Product-Based Library Design 197
6. Multiobjective Library Design 200
6.1 Multiobjective Library Design using a MOGA 201
7. Practical Examples of Library Design 203
7.1 Structure-Based Library Design 203
7.2 Library Design in Lead Optimisation 204
8. Summary 207

APPENDIX 1. MATRICES, EIGENVECTORS AND EIGENVALUES. 209

APPENDIX 2. CONFORMATION, ENERGY CALCULATIONS AND ENERGY SURFACES. 213

FURTHER READING 219

REFERENCES 225

INDEX 251