

CHEMOINFORMATICS: THEORY, PRACTICE, & PRODUCTS

CHEMOINFORMATICS: THEORY, PRACTICE, & PRODUCTS

B. A. BUNIN

Collaborative Drug Discovery, San Mateo, CA, U.S.A.

B. SIESEL

Merrill Lynch & Co., San Francisco, CA, U.S.A.

G. A. MORALES

Telik Inc., Palo Alto, CA, U.S.A.

J. BAJORATH

Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany



Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-5000-3 (HB)

ISBN-13 987-1-4020-5000-8 (HB)

ISBN-10 1-4020-5001-1 (e-book)

ISBN-13 987-1-4020-5001-5 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

All Rights Reserved

© 2007 Springer

No part of this work may be reproduced, stored in a retrieval system, or transmitted
in any form or by any means, electronic, mechanical, photocopying, microfilming, recording
or otherwise, without written permission from the Publisher, with the exception
of any material supplied specifically for the purpose of being entered
and executed on a computer system,
for exclusive use by the purchaser of the work.

TABLE OF CONTENTS

| | |
|---|----------|
| Foreword | ix |
| 1. Chemoinformatics Theory | 1 |
| 1.1 Chemoinformatics – What is it? | 1 |
| 1.2 Chemo- versus Bio-informatics | 2 |
| 1.3 Scientific Origins | 4 |
| 1.4 Fundamental Concepts | 4 |
| 1.4.1 Molecular descriptors and chemical spaces | 4 |
| 1.4.2 Chemical spaces and molecular similarity | 7 |
| 1.4.3 Molecular similarity, dissimilarity, and diversity | 8 |
| 1.4.4 Modification and simplification of chemical spaces | 9 |
| 1.5 Compound Classification and Selection | 11 |
| 1.5.1 Cluster analysis | 12 |
| 1.5.2 Partitioning | 13 |
| 1.5.3 Support vector machines | 16 |
| 1.6 Similarity Searching | 17 |
| 1.6.1 Structural queries and graphs | 17 |
| 1.6.2 Pharmacophores | 18 |
| 1.6.3 Fingerprints | 21 |
| 1.7 Machine Learning Methods | 23 |
| 1.7.1 Genetic algorithms | 23 |
| 1.7.2 Neural networks | 24 |
| 1.8 Library Design | 26 |
| 1.8.1 Diverse libraries | 27 |
| 1.8.2 Diversity estimation | 28 |
| 1.8.3 Multi-objective design | 29 |
| 1.8.4 Focused libraries | 29 |
| 1.9 Quantitative Structure-Activity Relationship Analysis | 31 |
| 1.9.1 Model building | 31 |
| 1.9.2 Model evaluation | 32 |
| 1.9.3 3D-QSAR | 33 |
| 1.9.4 4D-QSAR | 34 |
| 1.9.5 Probabilistic methods | 35 |
| 1.10 Virtual Screening and Compound Filtering | 35 |
| 1.10.1 Biologically active compounds | 35 |

| | | |
|-----------|--|-----------|
| 1.10.2 | Virtual and high-throughput screening | 36 |
| 1.10.3 | Filter functions | 38 |
| 1.11 | From Theory to Practice | 40 |
| 1.11.1 | Database design | 40 |
| 1.11.2 | Compound selection for medicinal chemistry | 42 |
| 1.11.3 | Computational hit identification | 45 |
| | References | 47 |
| 2. | Practice and Products | 51 |
| 2.1 | Accelrys | 51 |
| 2.2 | ACD Labs | 59 |
| 2.3 | Barnard Chemical Information Ltd | 67 |
| 2.4 | BioByte | 69 |
| 2.5 | CambridgeSoft | 73 |
| 2.6 | CAS/Scifinder | 80 |
| 2.7 | ChemAxon | 87 |
| 2.8 | Chemical Computing Group | 98 |
| 2.9 | ChemInnovation Software | 103 |
| 2.10 | ChemNavigator | 109 |
| 2.11 | Chimera-Dock-Zinc from UCSF | 112 |
| 2.12 | Collaborative Drug Discovery (CDD, Inc.) | 115 |
| 2.13 | Daylight | 123 |
| 2.14 | Eidogen-Sertanty (previously Libraria) | 127 |
| 2.15 | Fujitsu Biosciences Group (previously Cache) | 137 |
| 2.16 | Genego | 140 |
| 2.17 | GVK-Bio | 144 |
| 2.18 | Hypercube | 148 |
| 2.19 | IDBS | 152 |
| 2.20 | Infochem | 156 |
| 2.21 | Jubilant Biosys | 164 |
| 2.22 | Leadscope | 169 |
| 2.23 | MDL | 171 |
| 2.24 | Milano Chemometrics and QSAR Research Group | 180 |
| 2.25 | Molecular Discovery | 184 |
| 2.26 | Molecular Networks | 187 |
| 2.27 | Open Eye Scientific Software | 194 |
| 2.28 | Planaria-Software | 202 |
| 2.29 | PubChem | 203 |
| 2.30 | PyMol | 208 |
| 2.31 | RasMol and Protein Explorer | 211 |
| 2.32 | Schrödinger, LLC | 215 |
| 2.33 | Scinova Technologies | 223 |
| 2.34 | Scitegic | 226 |

| | | |
|------|-----------------------|-----|
| 2.35 | Simulation Plus, Inc. | 229 |
| 2.36 | Spotfire | 236 |
| 2.37 | Summit PK | 239 |
| 2.38 | Symyx | 243 |
| 2.39 | TimTec | 254 |
| 2.40 | Tripes | 259 |

SUBJECT APPENDICES

Drug Discovery Informatics Registration Systems and Underlying Toolkits (Appendices 1 and 2)

| | | |
|------------|--|-----|
| Appendix 1 | Drug, Molecular Registration Systems, and Chemistry Data Cartridges | 271 |
| Appendix 2 | Chemoinformatics Toolkits to Develop Applications | 272 |

Content Databases (Appendices 3–7)

| | | |
|------------|-----------------------------------|-----|
| Appendix 3 | Compound Availability Databases | 273 |
| Appendix 4 | SAR Database | 273 |
| Appendix 5 | Chemical Reaction Databases | 274 |
| Appendix 6 | Patent Databases | 275 |
| Appendix 7 | Other Compound and Drug Databases | 275 |

Drug, Molecule, and Protein Visualization (Appendices 8–10)

| | | |
|-------------|---|-----|
| Appendix 8 | Chemical Drawing, Structure Viewing and Modeling Packages | 276 |
| Appendix 9 | Data Analysis and Mining Tools | 276 |
| Appendix 10 | Small Molecule – Protein Visualization Tools | 277 |

Modeling and Algorithms (Appendices 11–17)

| | | |
|-------------|--|-----|
| Appendix 11 | Molecular Descriptors | 278 |
| Appendix 12 | Clogp, Tpsa, and Lipinski Property Calculation Systems | 279 |
| Appendix 13 | Qsar/Pharmacophore Programs | 279 |
| Appendix 14 | Docking and Crystallographic Software | 280 |
| Appendix 15 | Quantum Mechanics Calculations | 280 |
| Appendix 16 | PK/ADME/Tox Databases and Predictors | 280 |
| Appendix 17 | Multi-parameter Drug Development/Identification Software | 281 |

| | |
|-------|-----|
| Index | 283 |
|-------|-----|

FOREWORD

Chemoinformatics: Theory, Practice & Products covers the theory, commercially available packages and applications of Chemoinformatics. Chemoinformatics is broadly defined as the use of information technology to assist in the acquisition, analysis and management of data and information relating to chemical compounds and their properties. This includes molecular modeling, reactions, spectra and structure-activity relationships associated with chemicals. Computational scientists, chemists, and biologists all rely on the rapidly evolving field of Chemoinformatics. *Chemoinformatics: Theory, Practice & Products* is an essential handbook for determining the right Chemoinformatics method or technology to use. There has been an explosion of new Chemoinformatics tools and techniques. Each technique has its own utility, scope, and limitations, as well as meeting resistance to use by experimentalists. The purpose of *Chemoinformatics: Theory, Practice & Products* is to provide computational scientists, medicinal chemists and biologists with unique practical information and the underlying theories relating to modern Chemoinformatics and related drug discovery informatics technologies.

The book also provides a summary of currently available, state-of-the-art, commercial Chemoinformatics products, with a specific focus on databases, toolkits, and modeling technologies designed for drug discovery. It will be broadly useful as a reference text for experimentalists wishing to rapidly navigate the expanding field, as well as the more expert computational scientists wishing to stay up to date.

It is primarily intended for applied researchers from the chemical and pharmaceutical industry, academic investigators, and graduate students.

The purpose of “Chemoinformatics: Theory, Practice, & Products” is to provide scientists with practical information and a fundamental understanding of the latest chemoinformatics technologies applied to drug discovery and other applications. Given an ever-expanding list of drug discovery informatics tools available to the modern researcher, understanding the underlying theories, organizing and summarizing the tools for best practices should be broadly useful. It is intended to be a regularly used text.

Chemoinformatics is broadly defined as information associated with molecules: both theoretical and experimental. This ranges from molecular modeling to reactions to spectra to structure-activity relationships associated with molecules. Chemoinformatics has the potential to revolutionize synthesis, drug discovery, or any science where one wants to optimize molecular properties. Computational scientists, chemists, and biologists all rely on the rapidly evolving field of chemoinformatics. The terms chemoinformatics and cheminformatics are often used interchangeably. As of July, 2006, the term “Cheminformatics” is leading “Chemoinformatics” ~306,000 to ~164,000 in a Google

search (thanks to Phil McHale for his original suggestion). Despite this difference in Google popularity, we use chemoinformatics throughout this book because cheminformatics is frequently mis-interpreted as an abbreviation of the expression “Chemical informatics”. As we will discuss, “Chemical informatics” has originally been used in a different context (and it is also not a very meaningful term).

Chemoinformatics, which can be viewed as either a subset or superset of Drug Discovery Informatics, has emerged as an interdisciplinary field of science of importance to chemists and biologists as well as computational scientists. Computational scientists use chemoinformatics tools to design and refine better models. Medicinal chemists use chemoinformatics tools to design and synthesize better compounds. Biologists use chemoinformatics to prioritize compounds for screening and assays for development. The drug discovery process is often analogized to the tale of the three blind men and the elephant where each “sees” a different beast by grabbing the tail, trunk, or side. The appropriate development of new and use of existing chemoinformatics tools is often directly a function of a specific problem ... and problem solver. Thus having a centrally-compiled resource describing relevant chemoinformatics tools allows researcher to find the appropriately shaped “hammer” for their “nail.”

“Chemoinformatics: Theory, Practice, & Products” provides the basic toolkits. It is a handbook that one can consult to determine the chemoinformatics method or technology of choice to use. The book covers the theory behind the methodologies *as well* as the practical information on commercially available products. The goal is to provide the perspective of computational chemists in a format accessible to experimentalists, too. Thus, there are sections on the underlying theory as well as sections over-viewing the modern commercially available software and applications to provide the information of interest to computational scientists as well as to the broader audience of experimentalists.

There has been an explosion of new chemoinformatics tools and techniques. Each technique has some utility, scope, and limitations, as well as resistance to use by experimentalists. There is no compilation describing all the modern tools that are available. This guide will allow both experts and non-experts to know how and when to best use these technologies.

“Chemoinformatics: Theory, Practice, & Products” is intended for chemists, biologists, and computational scientists. It is basically for anyone interested in cheminformatics for either synthesis or drug discovery. This includes the individuals at the companies mentioned in the book who work in the field of cheminformatics (MDL, Accelrys, Tripos, CambridgeSoft, etc.) as well as the computational chemistry or drug design departments at biotechnology and pharmaceutical companies engaged in small molecule drug discovery and those using cheminformatics for materials discovery too.

The book can be useful as a reference book for the experienced cheminformatics expert or as a text to introduce the new student to the field. The information from the leading commercial suppliers is covered and organized into tables to help a wider range of scientists benefit from the revolution in informatics technologies in their

day-to-day work. It is a reference of what is known as well as a guidebook to define what is possible with modern chemoinformatics technologies.

A quick disclaimer. Although a range of areas were covered including over a hundred product and methods, it is not possible to include everything under the sun. A more specialized book could be written entirely about any one of the seventeen subject appendices. Obviously tradeoffs had to be made between scope and depth of coverage. Furthermore, although it is inevitable that products and technologies will evolve over time, many of the most useful products are now mainstays of the modern chemoinformatics arsenal such as CAS-Scifinder, Beilstein, ChemDraw, Marvin, smiles strings, and Lipinski calculations – just to name a few. In addition to these well known products, there are often alternative products available with different specifications which are also described herein. Thus even as new trends emerge, the general state of modern chemoinformatics (and drug discovery informatics) is fundamentally represented. It is interesting to see the range of products that have historically been available as well as the evolution of new product areas such as gene-family wide SAR databases, data-pipelining, and metabolism predictors, just to name a few.

Perhaps most notable of the new initiatives is the publicly funded PubChem effort. A road map of existing products is useful both to differentiate new products and to prioritize the most important areas to focus future innovation. Understanding the landscape of existing products should be particularly useful to the buyers and sellers of chemoinformatics and drug discovery informatics technologies. Where might the field go in the future? With the emergence of open source software products in the broader software marketplace (for products like Linux, Apache, and MySQL), the integration of community-based tools with commercial tools has been a recently increasing phenomena. Similarly, the increasing number of openly available databases and tools emerging from the publicly funded initiatives such as the human genome project provide a fertile frontier for future innovation that combines the best of community and commercial chemoinformatics tools in new ways.

1. CHEMOINFORMATICS THEORY

The theoretical part of this book is intended to provide a general introduction into this still young and rapidly evolving scientific discipline. In addition, it is meant to provide a basis for researchers interested in applying products and tools that are detailed in the later sections. Therefore, it is attempted to outline some of the most relevant scientific concepts on which current chemoinformatics tools are based and provide some guidance as to which methodologies can be applied in a meaningful way to tackle specific problems. As such the theoretical sections are first and foremost written for practitioners with various scientific backgrounds and also students trying to access chemoinformatics tools. Therefore, the description of mathematical formalisms will be limited to the extent required to achieve a general understanding. In addition, rather than trying to provide an extensive bibliography covering this field, it is attempted to limit citations to key publications and contributions that are accessible to a readership with diverse scientific backgrounds.

As a still evolving discipline, chemoinformatics is an equally interesting playground for method development, chemical and drug discovery applications, and interdisciplinary research. This makes this field a rather exciting area to work in and it is hoped that the information provided herein might encourage many scientific minds to actively contribute to its further development.

1.1 CHEMOINFORMATICS – WHAT IS IT?

The term chemoinformatics (which is synonymously used with cheminformatics) was introduced in the literature by Brown in 1998 and defined as the combination of “all the information resources that a scientist needs to optimize the properties of a ligand to become a drug” (Brown 1998). Following this definition, both decision support by computer and drug discovery relevance are crucial aspects. On the other hand, the term chemical informatics was already used much earlier and generally understood as the application of information technology to chemistry, thus lacking a specific drug discovery focus. In addition, the chemometrics field focuses on the application of statistical methods to chemical data in order to derive predictive models or descriptors. Although these definitions and areas of research still co-exist, it appears to be increasingly difficult to distinguish between them, in particular, as far as method development is concerned. Therefore, it has recently been suggested to more broadly define chemoinformatics and include the types of

TABLE 1.1. The spectrum of chemoinformatics

| |
|---|
| Chemical data collection, analysis, and management |
| Data representation and communication |
| Database design and organization |
| Chemical structure and property prediction (including drug-likeness) |
| Molecular similarity and diversity analysis |
| Compound or library design and optimization |
| Database mining |
| Compound classification and selection |
| Qualitative and quantitative structure-activity or – property relationships |
| Information theory applied to chemical problems |
| Statistical models and descriptors in chemistry |
| Prediction of <i>in vivo</i> compound characteristics |

computational methodologies and infrastructures in the chemoinformatics spectrum that are shown in Table 1.1 (Bajorath, 2004).

This extended definition does no longer imply that chemoinformatics is necessarily linked to drug discovery and takes into account that this field is still evolving. Moreover, approaches that are long established as disciplines in their own right are also part of the chemoinformatics spectrum. This is well in accord with other views that chemoinformatics might largely be a new rationalization of tasks in chemical research that have already existed for considerable time (Hann and Green 1999). In fact, chemoinformatics research and development should be capable of adopting established scientific concepts and putting them into a novel context. Given the above topics, good examples for this might include, among others, the use of quantitative structure-activity relationship (QSAR) models for computational screening of large compound databases or the use of fragments of active compounds (so-called substructures) as a starting point for the design of targeted combinatorial libraries. In its extended definition, chemoinformatics includes all concepts and methods designed to interface theoretical and experimental programs involving small molecules. This is a crucial aspect because there is little doubt that the evolution of chemoinformatics as an independent discipline will much depend on its ability to demonstrate a measurable impact on experimental chemistry programs, regardless of whether these are in pharmaceutical research or elsewhere.

1.2 CHEMO- VERSUS BIO-INFORMATICS

There is little doubt that data explosion in chemistry and biology has been the major driver for the development of chemoinformatics and bioinformatics as disciplines. In the 1990s the advent of high-throughput technologies in biology (DNA sequencing) and chemistry (combinatorial synthesis) had caused much of the need for efficient computational infrastructures for data processing, management, and mining. In biology raw DNA sequences were the primary data source, whereas in chemistry rapidly

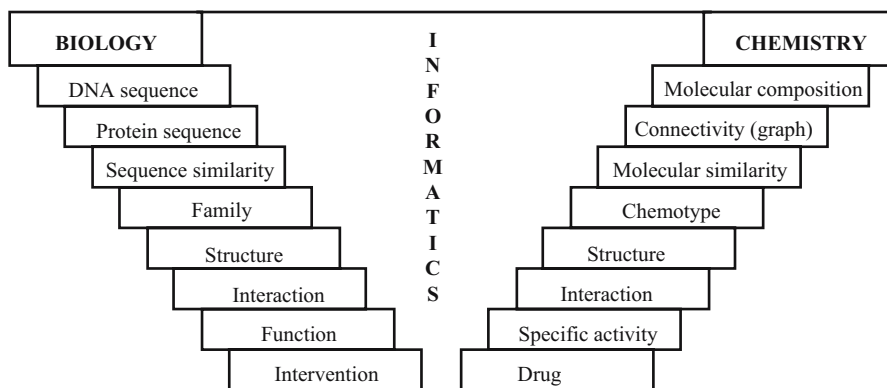


Figure 1.1. Hierarchy of bio- and chemoinformatics research

growing compound databases produced by combinatorial synthesis techniques provided previously unobserved amounts of primary data (structures) and secondary data (screening results). Over a relatively short period of time, however, both bio- and chemo-informatics have developed well beyond data processing and management and have become research-intense disciplines.

How distinct are bio- and chemoinformatics as disciplines? Figure 1.1 summarizes the topics at different stages of typical bio- or chemoinformatic analysis.

Clearly, proceeding from molecular composition to two- and three-dimensional structure and function or activity presents a number of similar challenges, regardless of whether the starting point is a DNA sequence or the chemical element distribution of a molecule. From an algorithmic point of view, many tasks for this type of analysis in biology and chemistry are often much more similar than one might think, considering the diversity of biological and chemical applications. Thus, many algorithms and computational techniques used in chemoinformatics, as will be described herein, are also used for many applications in bioinformatics. For example, cluster algorithms are not only applied to classify compound databases but also to analyze expression data sets. Similarly, statistical algorithms are used to correlate compound structures with specific activities and also to correlate expression patterns and experimental conditions in microarray analysis. Thus, “similar algorithms – diverse applications” is a general theme in applied informatics research. Such insights are also consistent with recent trends in the life science area where bio- and chemoinformatics are beginning to merge. This is particularly relevant for drug discovery where chemical and biological information needs to be integrated as much as possible to be ultimately successful and where the boundaries between different disciplines have become rather fluid. For example, it could hardly be decided whether the development of relational databases that link compound structure with assay data, biological targets, and pharmacological information would be a bio- or chemoinformatics

project. Thus, informatics research and development in the life sciences is expected to become much more global in the future.

1.3 SCIENTIFIC ORIGINS

Given the above outline of the chemoinformatics field, one should also review the scientific roots that have laid the foundation for the development of chemoinformatics as a research discipline, beyond data management. In the 1960s, efforts begun to correlate compound structures and activities in quantitative terms by modeling linear relationships with the aid of molecular descriptors (Hansch and Fujita 1964; Free and Wilson 1964). These studies provided the basis of quantitative structure-activity relationship (QSAR) analysis, which was ultimately extended to multi-dimensional QSAR in 1980 (Hopfinger 1980). Also in the 1960s, chemical structures were first stored as computer files in searchable form by Chemical Abstract Services, thus providing a basis for structure retrieval and searching (Willett 1987). During the 1970s, methods for two-dimensional substructure (Cramer *et al.* 1974) and three-dimensional pharmacophore searching (Gund 1977) were developed, which made it possible to search compound databases for desired structural motifs or active molecules. In the 1980s, clustering methods were adapted for chemical applications, became very popular for the classification of molecular data sets, and were applied to explore similarities from various points of view (Willett 1988). The concept of molecular similarity itself became a major research topic in the late 1980s (Johnson and Maggiora 1990). Molecular similarity analysis extended conventional QSAR approaches where the influence of small compound modifications on activity is studied. Thus, relationships between molecular structure (and properties) and biological activity were beginning to be explored from a more global point of view. During the 1990s, the concepts of molecular diversity and dissimilarity complemented similarity analysis and algorithms were developed for the design of chemically diverse compound libraries (Martin *et al.* 1995) and selection of diverse compounds from databases (Lajiness 1997; for a compendium of interesting personal accounts of the early days of molecular similarity and diversity research, see Martin 2001). Although many other efforts have – without doubt – significantly contributed to and helped to shape chemoinformatics, as we understand it today, it is evident that two major themes have largely dominated the development of this discipline: chemical data organization and mining and, in addition, the exploration of structure-activity relationships (from many different points of view).

1.4 FUNDAMENTAL CONCEPTS

1.4.1 Molecular descriptors and chemical spaces. The majority of chemoinformatics methods depend on the generation of chemical reference spaces into which molecular data sets are projected and where analysis or design is carried out. The definition of chemical spaces critically depends on the use of computational descriptors of molecular structure, physical or chemical properties, or pharmacophores. Essentially, any comparison of molecular characteristics that goes beyond simple structural comparison requires the calculation of property values and the application

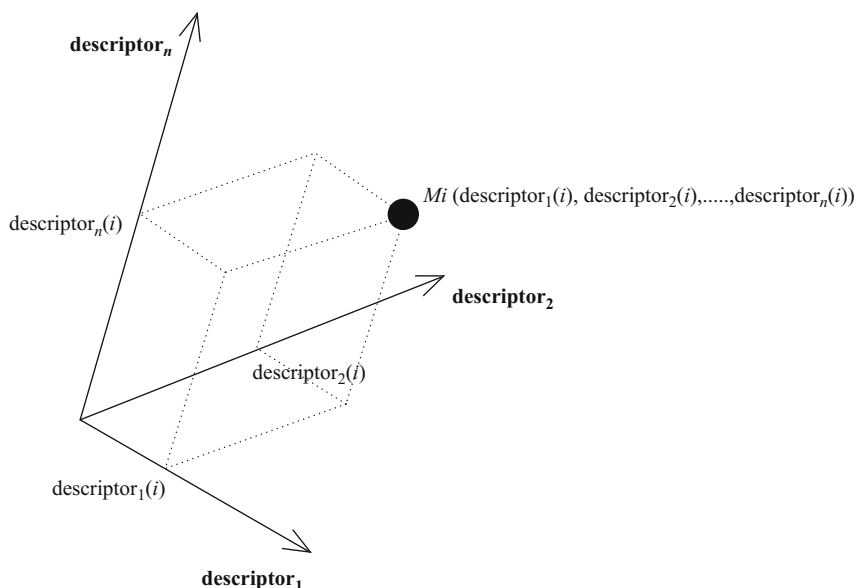


Figure 1.2. N -dimensional chemical space with a molecule M at position i

of mathematical models. In chemical space design, each chosen descriptor adds a dimension to the reference space, as illustrated in Figure 1.2.

For each molecule, calculation of n descriptor values produces an N -dimensional coordinate vector in descriptor space that determines its position:

$$\text{Molecule } i: M(i) = \sum_1^j d_j(i)$$

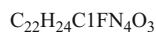
Hundreds if not thousands of molecular descriptors have been designed for chemical applications (for an encyclopedic descriptor compendium, see [Todeschini and Consonni 2000]) that can be divided into different types or classes. Some examples are given in Table 1.2. Descriptors are frequently divided into 1D, 2D, or 3D descriptors, dependent on the dimensionality of the molecular representation from which they can be calculated, as illustrated in Figure 1.3.

The design and complexity of different types of descriptors often varies dramatically. Among very simple descriptors are, for example, 2D structural fragments that have, however, high predictive value in many applications because they implicitly account for diverse molecular properties (such as complexity, polarity, hydrophic character etc.). Topological indices, for example, go beyond simple structural fragment description and introduce a next level of abstraction. To give an example, for a molecule containing

TABLE 1.2. Different types of molecular descriptors

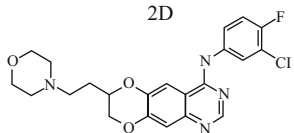
| Descriptor category | Examples |
|------------------------------------|---|
| Physical properties | Molecular weight logP(o/w) |
| Atom and bond counts | Number of nitrogen atoms Number of aromatic atoms Number of rotatable bonds |
| Pharmacophore features | Number of hydrogen bond acceptors Sum of van der Waal surface areas of basic atoms |
| Charge descriptors | Total positive partial charge Dipole moment from partial charges |
| Connectivity and shape descriptors | Kier and Hall molecular shape indices |
| Surface area and volume | Solvent-accessible surface area |

1D



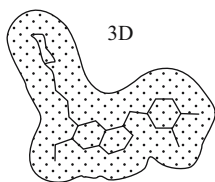
Number of carbon atoms

2D



Number of rotatable bonds
log P(o/W)
Molecular connectivity index

3D



Solvent-accessible surface area
Van der Waals volume

Figure 1.3. Examples of descriptors classified according to dimensionality
(adapted from Bajorath 2002)

n atoms and m bonds, so-called first and second order kappa shape indices (Kier 1997) are calculated as follows:

$$\kappa_1 = \frac{n(n-1)^2}{m^2}$$

$$\kappa_2 = \frac{(n-1)(n-2)^2}{m^2},$$

These indices are also 2D descriptors because they require a molecular drawing (or graph) in order to determine the number of bonds.

Other descriptor designs can become increasingly complex. Contributions of different types of descriptors can also be combined into composite formulations, for example, descriptors combining molecular surface and charge information such as charged partial surface area (CPSA) descriptors (Stanton and Jurs 1990). As an example, two of these descriptors, $PNSA_1$ and $PPSA_1$, capture the sum of the solvent-accessible surface area (SAA) of all negatively and positively charged atoms in a molecule, respectively:

$$PNSA_1 = \sum_n SAA^-$$

$$PPSA_1 = \sum_n SAA^+$$

Thus, calculation of these descriptor values for a molecule involves the separate calculation of atomic charges and SAA s.

1.4.2 Chemical spaces and molecular similarity. There are no generally preferred descriptor spaces for chemoinformatics applications and it is usually required to generate reference spaces for specific applications on a case-by-case basis, either intuitively, based on experience, or by applying machine learning techniques to automate and optimize descriptor selection for a given problem. However, descriptors are ultimately selected for chemical space design, n descriptors always produce an n -dimensional reference space, as discussed above, into which compound sets can be mapped. In meaningful chemical space representations, similar compounds should map to similar regions, in other words, their intermolecular distance should be small. This represents a basic interpretation of the similarity concept. Table 1.3 lists examples of conventional distance functions that are used for these calculations.

Here n_i and n_j are the number of descriptor values for molecules i and j , respectively, and n_{ij} is the number of common values. D_{ij} is the distance between molecules i and j , D the average distance, and n the total number of molecules.

It should be noted that the general understanding of molecular similarity goes beyond simple structural similarity and extends to biological activity, in accord with the so-called Similar Property Principle (Johnson and Maggiora 1990) postulating that molecules having similar structures and properties should also exhibit similar activity

TABLE 1.3. Distance functions

| | |
|--------------------|--|
| Hamming distance | $HD = \sum_{i=1}^n x_i \oplus y_i,$ |
| Euclidean distance | $ED = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$ |
| Average distance | $AD = \frac{\sum_{i=1}^n \sum_{j=1}^n D_{ij}}{n(n-1)}$ |

In the formula of the Hamming distance, \oplus means “exclusive disjunction” and detects non-identical values. In the formula of the average distance, D_{ij} is the distance between molecules i, j and n the total number of molecules.

TABLE 1.4. Similarity coefficients

| | |
|----------------------|------------------------------------|
| Tanimoto coefficient | $Tc = n_{ij}/(n_i + n_j - n_{ij})$ |
| Dice coefficient | $Dc = 2n_{ij}/(n_i + n_j)$ |
| Cosine coefficient | $Cc = n_{ij}/(n_i n_j)^{1/2}$ |

(which is often – but not always – true). Thus, molecules that are located closely together in chemical reference space are often considered to be functionally related, which is one of the hallmarks of molecular similarity analysis.

If descriptor combinations are expressed as bit strings (often called fingerprints, as described in more detail later on), each test molecule is assigned a characteristic bit pattern, and pair-wise molecular similarity can be assessed by quantifying the overlap of bit strings using various similarity metrics (coefficients). Examples are shown in Table 1.4.

In these formulations, n_i and n_j are the number of bits set on for molecules i and j , respectively, and n_{ij} is the number of bits in common to both molecules.

The values of these similarity coefficients range from zero (i.e., no overlap; no similarity) to one (i.e., complete overlap; identical or very similar molecules). In chemoinformatics, the most widely used metric is the Tanimoto coefficient.

1.4.3 Molecular similarity, dissimilarity, and diversity. How are similarity and diversity related to each other? As discussed, similar molecules can be identified by application of distance functions and analysis of nearest neighbors in chemical space. Diversity analysis, on the other hand, attempts to either select different compounds from a given population or, alternatively, evenly populate a given chemical space with candidate molecules. This can also be accomplished using distance functions by only selecting compounds that are at least a pre-defined minimum distance away from others or – in diversity design – by trying to maximize average inter-compound distances.

An alternative approach to diversity selection and design is to divide the descriptor axes into evenly spaced value intervals, a process called “binning”, which produces n -dimensional subsections of chemical space (also called “cells”, as discussed in a later section). Then it can be monitored how these cells are populated with compounds that are projected into chemical space. In diversity selection, one would attempt, for example, to select a representative compound from each populated cell; in diversity design, one would try to populate cells as evenly as possible with computed molecules. As will be discussed in the next section, such segment- or cell-based design strategies can, in practical terms, only be applied to low-dimensional descriptor spaces; otherwise, the vast majority of cells would remain empty, thereby preventing a meaningful analysis.

Molecular diversity is a global concept, which is applicable to the analysis of large compound distributions, but not to the study of pair-wise molecular relationships. This

is in contrast to molecular similarity analysis, which explores pair-wise relationships, the exploration of which is more local in nature. For example, one tries to find compounds similar to a given reference molecule or study the compound population within a limited region of chemical space. From this point of view, the inverse of molecular similarity is not diversity, but rather “dissimilarity”, which is local in nature (addressing the question which molecule in a collection is most dissimilar from a given compound or set of compounds). Like similarity, dissimilarity calculations can focus on the exploration of pair-wise compound relationships (e.g., distances in chemical space). When similarity metrics are applied, the dissimilarity d between two molecules i and j is thus defined as, for example:

$$d_{ij} = 1 - Cc(i, j) \quad \text{or} \quad 1 - Tc(i, j)$$

Dissimilarity analysis plays a major role in compound selection. Typical tasks include the selection of a maximally dissimilar subset of compounds from a large set or the identification of compounds that are dissimilar to an existing collection. Such issues have played a major role in compound acquisition in the pharmaceutical industry. A typical task would be to select a subset of k maximally dissimilar compounds from a data set containing n molecules. This represents a non-trivial challenge because of the combinatorial problem involved in exploring all possible subsets. Therefore, other dissimilarity-based selection algorithms have been developed (Lajiness 1997). The basic idea of such approaches is to initially select a seed compound (either randomly or, better, based on dissimilarity to others), then calculate dissimilarity between the seed compound and all others and select the most dissimilar one. In the next step, the database compound most dissimilar to these two compounds is selected and added to the subset, and the process is repeated until a subset of desired size is obtained.

1.4.4 Modification and simplification of chemical spaces. High-dimensional chemistry spaces might often be too complex for carrying out meaningful and interpretable analyses. One reason for this is that major areas or subsections of high-dimensional chemical space might not be populated with compounds and thus remain “empty”. Another reason is that correlation effects between selected descriptors dramatically distort the reference space, which often (but not always) complicates the analysis of compound distributions. Therefore, it is generally attempted to either design low-dimensional reference spaces, simplify high-dimensional spaces, or reduce their dimensionality. Descriptor correlation is a very common effect. For example, the number of carbon atoms in a molecule (a very simple 1D descriptor) correlates with molecular weight, hydrophobicity etc. In fact, it is rather difficult to find a set of completely uncorrelated descriptors. Compound analysis or design in low-dimensional spaces has the added bonus that it is often possible to further reduce the dimensionality to three without too much loss of information so that one can visualize the results. Visualization of chemical space representations, even if only approximate, is in general of high value, as it permits a more intuitive analysis of molecule distributions and makes it possible to complement computations with chemical knowledge and experience. There are several different ways to simplify chemical spaces or produce low-dimensional representations, as discussed in the following.

Regardless of space dimensionality, it is generally important to scale selected descriptors because their value ranges may substantially differ for a given data set. Descriptors with large value ranges will dominate those having smaller ones and distort the analysis (i.e., a very “long” coordinate axis in chemical space might render “short” axes nearly “invisible”). Therefore, auto-scaling or variance scaling with mean centering is typically applied:

$$d'_i = \frac{(d_i - d_{av})}{\sigma}$$

Here d_i is the descriptor value of molecule i , d_{av} the average (or mean) value of the entire data set, the σ standard deviation, and d'_i the scaled value of descriptor d for molecule i . This procedure ensures that all chosen descriptors have similar value ranges (i.e., that descriptor axes have comparable length) and thus prevents space distortions.

The most common way to generate low-dimensional reference spaces is dimension reduction of original descriptor spaces, as illustrated in Figure 1.4. This process attempts to define a low-dimensional representation that captures data variability to the same or a similar extent as the original descriptor space.

Dimension reduction relies on the assumption that high-dimensional descriptor spaces have at least some intrinsic redundancy, which is in most cases true as a consequence of descriptor correlation effects. There are two major categories of methods to facilitate dimension reduction, for which different algorithms are available. One class of methods attempts to identify those descriptors that are most important for representing the original data set to use them and the relationships they form between objects for lower-dimensional representations. An example for this approach is multi-dimensional scaling (Agrafiotis *et al.* 2001). The other type of methods attempts to generate new descriptors for low-dimensional spaces by combining important contributions from the original ones. A representative method is principal component analysis that processes descriptor variance and co-variance matrices of compound sets and

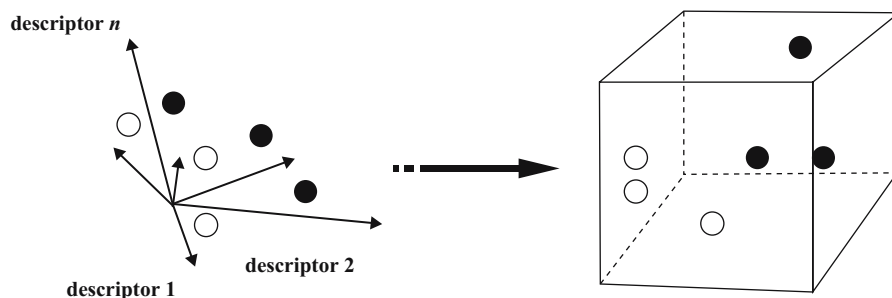


Figure 1.4. Dimension reduction. The figure illustrates the transformation of an n -dimensional descriptor space into an orthogonal three-dimensional space formed by three non-correlated descriptors either selected from the original ones or derived from them as new composite descriptors.

Test compounds are shown as white or black dots

ultimately calculates novel composite descriptors as linear combinations of the original ones (Xue and Bajorath 2000):

$$\begin{aligned}PC1 &= c_{1,1}d1 + c_{1,2}d2 + c_{1,3}d3 + \cdots + c_{1,n}dn + \text{const}_1 \\PC2 &= c_{2,1}d1 + c_{2,2}d2 + c_{2,3}d3 + \cdots + c_{2,n}dn + \text{const}_2 \\&\vdots \\PCn &= c_{n,1}d1 + c_{n,2}d2 + c_{n,3}d3 + \cdots + c_{n,n}dn + \text{const}_n\end{aligned}$$

Here the coefficients c reflect the importance of each descriptor (within each component) to capture data variance. Principal component analysis removes descriptor correlation effects and the resulting components account for data variance in descending order (i.e., the first accounts for more than the second, the second more than the third, and so on). A possible result would be that for an original 20-descriptor space, the first five or six principal components account for greater than 95% of the variance within the molecular data set, thus permitting the generation of an orthogonal reference space reduced to five or six dimensions.

An alternative to dimension reduction is the use of composite and uncorrelated descriptors that are suitable for the design of information-rich yet low-dimensional chemical spaces. An elegant example is presented by the popular BCUT (Burden-CAS-University of Texas) descriptors (Pearlman and Smith 1998). BCUTs are a set of uncorrelated descriptors that combine information about molecular connectivity, inter-molecular distances, and other molecular properties. BCUT spaces used for many applications are typically only six-dimensional and can frequently be further reduced to 3D representations for visualization purposes by identifying those BCUT axes around which most compounds map.

Simplification of n -dimensional descriptor spaces is another alternative to dimension reduction. This can be accomplished, for example, by analysis of descriptor value distributions in large databases. In statistics, the median of a value distribution is defined as the value that separates it into two equal halves (above and below the median). Thus, a descriptor with continuous database value range can be transformed into a binary scheme where a test compound with a descriptor value above or equal to the median is assigned a transformed descriptor value of "1" and a compound with a descriptor value below the median a transformed value of "0" (Godden *et al.* 2004). Binary descriptor transformation retains the dimensionality of the original space but greatly simplifies it because the length of each descriptor axis is either zero or unity. Thus, this simplified descriptor space is also scaled. As further discussed in the following, both low-dimensional and binary-transformed descriptor spaces have been proven very useful for partitioning analyses and compound classification or design.

1.5 COMPOUND CLASSIFICATION AND SELECTION

The classification of molecular data sets according to pre-defined criteria is one of the central themes in chemoinformatics. Methods designed to classify molecules are

applied in database organization and mining and also provide a basis for selection of compounds according to diversity, property, or biological activity criteria.

1.5.1 Cluster analysis. As mentioned before, clustering has been one of the roots of the chemoinformatics field and continues to be widely applied. Clustering methods are often divided into non-hierarchical and hierarchical techniques and hierarchical methods are further divided into divisive or agglomerative clustering. Hierarchical-divisive methods start from a large cluster containing all compounds (“top-down”), whereas hierarchical-agglomerative techniques begin from singletons (“bottom-up”). Hierarchical clustering builds relationships between clusters in subsequent steps, which means that the composition of each cluster depends on the one from which it was derived. Non-hierarchical clustering methods organize compounds into an initially defined number of independent clusters, which is often accomplished by calculating nearest neighbor distributions in chemical space. Molecules can be expressed as descriptor vectors and for each cluster, a center vector can be calculated (e.g., as an average position) that distinguishes different clusters from each other. New molecules are assigned to clusters based on their distances from different cluster centers in descriptor space. Figure 1.5 illustrates these different clustering approaches.

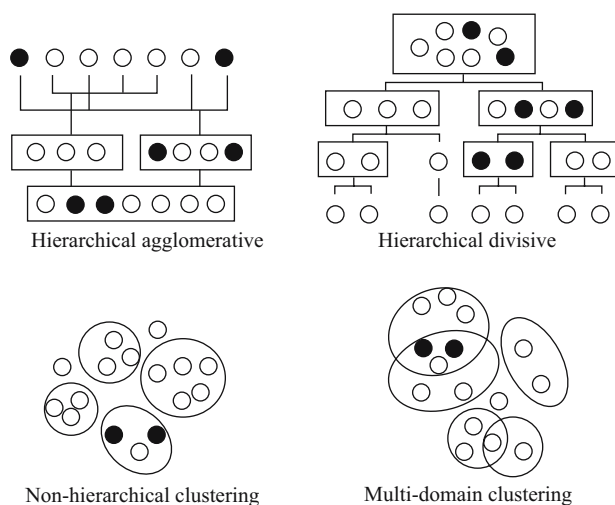


Figure 1.5. Clustering approaches described in the text
(adapted from Kitchen *et al.* 2004)

Fuzzy clustering methods that have recently become popular are distinct from traditional clustering techniques in that molecules are permitted to belong to multiple clusters or have fractional membership in all clusters. A potential advantage of such classification schemes is that more than one similarity relationship can be established by cluster analysis.

Hierarchical clustering also depends on the chosen “linkage scheme” that determines the way inter-cluster distances are calculated. For example, based on “single linkage”, inter-cluster distance is defined as the minimum distance between members of two clusters. By contrast, “complete linkage” calculates the maximum distance between members in two different clusters. Furthermore, for all clustering methods, clustering levels and cluster occupancy present additional variables. For example, too many clustering steps will result in sparsely populated clusters and too few in densely populated ones, both of which will distort molecular similarity relationships derived from clustering (which means that molecules within the same cluster should be “similar”). Therefore, level selection algorithms are typically applied in order to determine calculation parameters that balance clustering levels and cluster occupancy.

Among non-hierarchical methods, Jarvis-Patrick clustering (Jarvis and Patrick 1973) has been popular early on in chemical database analysis. It is a nearest neighbor method: two molecules are included in the same cluster if they share a pre-defined minimum number of nearest neighbors. However, the method has been found to produce rather unevenly sized clusters, often too large or too small. Another popular non-hierarchical method is k -means clustering where k clusters are randomly seeded, cluster averages or means are calculated in descriptor space, and molecules are re-assigned to other clusters if their position is closer to those means than to the one of their initial cluster. This clustering technique is fast but depends on the initial random seeding of clusters with test compounds and the choice of k . Over the years, agglomerative-hierarchical methods, in particular, Ward's clustering (Ward 1963), have become more popular in chemistry because this approach has been shown to produce more balanced cluster levels and distributions than non-hierarchical methods, resulting in more reliable classification of similar molecules.

As discussed, clustering algorithms generally involve distance comparisons between compounds or between compounds and cluster centers in chemical space, which renders calculations increasingly demanding as the compound databases grow in size. Given currently available computational power, classifications methods that involve exhaustive pair-wise compound or distance comparisons can be applied to thousands of compounds but become prohibitive when databases further increase in size by orders of magnitude.

1.5.2 Partitioning. In contrast to clustering techniques, partitioning algorithms do not rely on pair-wise molecular and distance comparison and can therefore be applied to very large compound source databases. Rather than comparing molecular positions,

partitioning methods establish a coordinate or reference system in chemical space that ultimately defines the position of each compound based on its calculated descriptor coordinates. Compounds that populate the same partitions or sub-sections of chemical space are considered similar. Partitioning in low-dimensional descriptors spaces, generated either by use of BCUT descriptors or dimension reduction techniques, has become a very popular approach. Cells are generated by dividing orthogonal (uncorrelated) descriptor axes into regularly spaced intervals or bins, as illustrated in Figure 1.6.

Regardless of whether clustering or partitioning algorithms are applied, compound classification calculations are often carried out to provide a basis for compound selection from large data sets. Major strategies include diversity- or activity-oriented selection, as illustrated in Figure 1.7. Diversity-based selection aims at generating a small representative subset of a compound collection. In this case, it

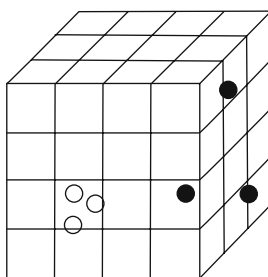


Figure 1.6. Axes of a low-dimensional orthogonal chemical space are binned in order to produce cells for partitioning. White dots represent similar compounds

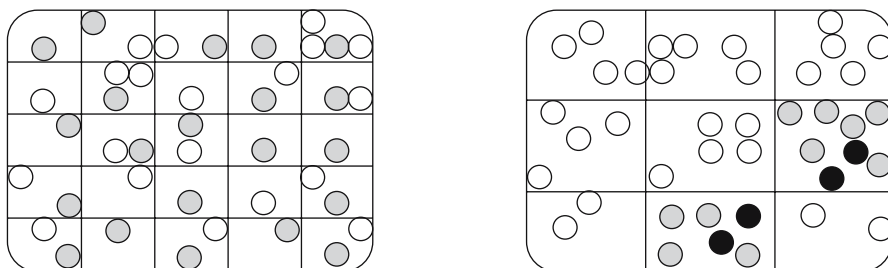


Figure 1.7. Diversity-based (left) and activity-based (right) compound selection from partitions. White dots represent database compounds, gray dots selected database compounds, and black dots known active compounds

is attempted to generate evenly populated partitions or clusters from which representative compounds are selected in order to mirror the overall diversity distribution in chemical space. By contrast, in activity-based selection, known active compounds are added to the source database prior to clustering or partitioning. Database compounds mapping close to known actives are then selected as candidates for testing to identify new hits.

Cell-based partitioning can not only be used for compound selection but also to aid in combinatorial diversity design. In this case, a chemical descriptor space is defined and “empty” partitions are generated by binning. Test compounds are then enumerated on the computer based on reaction schemes and selected to evenly populate these partitions.

In addition to cell-based partitioning, statistical partitioning methods are widely used for compound classification. One of the most popular approaches is recursive partitioning (Rusinko *et al.* 1999), a decision tree method, as illustrated in Figure 1.8. Recursive partitioning divides data sets along decision trees formed by sequences of molecular descriptors. At each node of the tree, a descriptor-based decision is made and the molecular data set is subdivided. For example, a chosen descriptor could simply detect the presence or absence of a structural fragment in a molecule. Alternatively, the

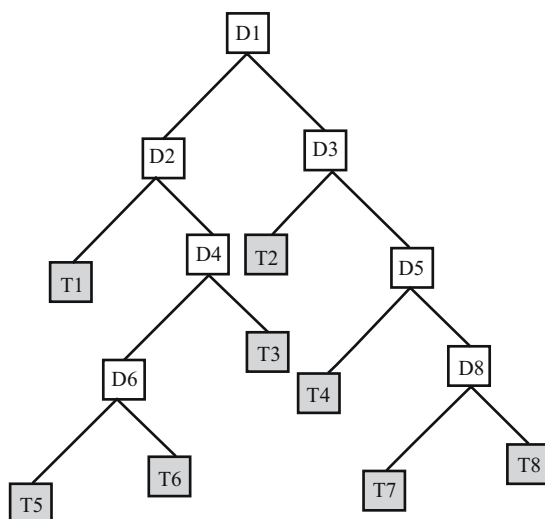


Figure 1.8. Decision tree. Shown is a rudimentary tree structure (D, descriptors; T, terminal nodes) for recursive partitioning. Terminal nodes are shaded gray.

compounds could be divided according to molecular weight (e.g., equal or greater than 400 or less). Very many different descriptors can be utilized and decision tree methods are computationally very efficient and applicable to very large data sets. Typically, a learning set would consist of active and inactive compounds and one would search for descriptor combinations that enrich active compounds in certain terminal nodes. The so derived descriptor pathways can then be used to search compound databases for other active compounds. Therefore, statistical partitioning methods such as recursive partitioning are also very attractive tools for the analysis of HTS data sets and the extraction of descriptor-activity relationships from them that can serve as predictive models of specific biological activities.

1.5.3 Support vector machines In addition to more “traditional” classification methods like clustering or partitioning, other computational approaches have recently also become popular in chemoinformatics and support vector machines (SVMs) (Warmuth *et al.* 2003) are discussed here as an example. Typically, SVMs are applied as classifiers for binary property predictions, for example, to distinguish active from inactive compounds. Initially, a set of descriptors is selected and training set molecules are represented as vectors based on their calculated descriptor values. Then linear combinations of training set vectors are calculated to construct a hyperplane in descriptor space that best separates active and inactive compounds, as illustrated in Figure 1.9.

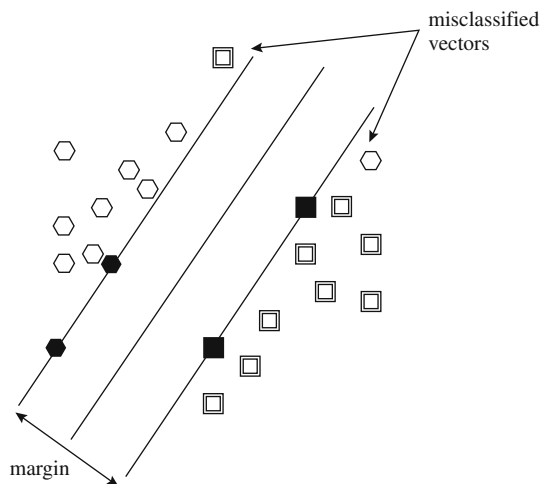


Figure 1.9. SVM-based hyperplane. Two classes of molecules are separated in descriptor space by a hyperplane ($H(x) = 0$) with margins ($H(x) = \pm 1$). Support vectors are shown as filled objects and used to construct the hyperplane and define its margins

A polynomial function is applied to return the inner products of descriptor vectors and the separating hyperplane is defined as

$$H(x) = 0 = \langle w, x \rangle + b$$

Here b is the distance between the hyperplane and the origin and w the distance between vector x and the hyperplane. The margin of the hyperplane is defined as its minimum distance to any training set vector. A small number of vectors constituting the margin are called support vectors and are sufficient to construct a hyperplane that separates the remaining data points into two subsets. SVM calculations and hyperplane construction can initially be carried out using small training sets and additional data can be added in a step-wise manner to further refine the prediction scheme.

1.6 SIMILARITY SEARCHING

Searching for compounds in databases that are similar to query molecules is one of the most widely applied molecular similarity-based approaches. Commonly used similarity search tools have different levels of complexity, as discussed in the following.

1.6.1 Structural queries and graphs. A simple but very popular form of similarity searching is the detection of structural fragments or substructures that are shared by query and database compounds. Figure 1.10 illustrates the idea of substructure searching.

In medicinal chemistry, substructure searches are often carried out to find analogs of active compounds in databases. Contemporary substructure search methods are mostly based on dictionaries or look-up tables of molecular fragments or fragment-type descriptors. Substructure search queries based on dictionaries of predefined molecular fragments can be transformed into an easily machine-readable format such as the Simplified Molecular Input Line Entry specification widely known as SMILES code (Weininger 1988). As illustrated in Figure 1.11, SMILES encodes 2D representations of molecules as linear strings of special alpha-numeric characters for atoms, their chemical character, bonding patterns, branch points, stereo centers etc. These query strings can be easily compared to those of database compounds.

Substructure searching has also been coupled with statistical analysis to identify molecular fragments that are associated with biological activity of test compounds (Roberts *et al.* 2000). For example, substructures can be taken from series of active compounds and their frequency of occurrence in compound databases can be determined. Then the statistical significance of this frequency of occurrence in active molecules and database compounds can be compared, which might indicate whether or not a specific structural motif could be responsible for a given biological activity.

In molecular graphs, atoms are represented as nodes and bonds as edges. Conventional 2D representations of compound structures are typical graphs, but graph representations of molecular connectivity can also be much simpler than that (as described below). Common substructures can also be determined by systematic mapping of corresponding node positions in graphs, which is called the analysis of “subgraph isomorphism”. However, computationally this is a much more expensive

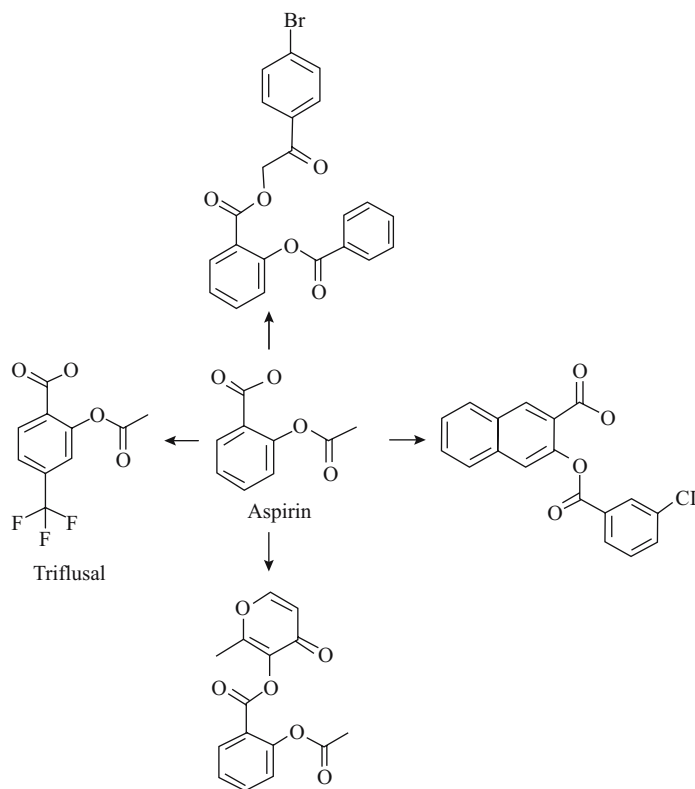


Figure 1.10. Example of compounds containing Aspirin as a substructure that can be used as a query for database searching

procedure than dictionary-based (or “grammatical”) approaches to substructure analysis. Therefore, similarity searching using molecular graphs has generally been computationally prohibitive, until recently when reduced graphs were developed for these purposes (Gillett *et al.* 2003). In reduced graphs, nodes do not represent atoms but features such as functionally important groups or whole ring systems, which reduces the level of detail of the representations of intra-molecular connectivity, as illustrated in Figure 1.12. Thus, such simplified graph representations become more suitable for node matching procedures and similarity searching.

1.6.2 Pharmacophores. Going beyond 2D substructures, pharmacophores are defined as spatial arrangements of atoms or groups that are responsible for biological activity. Such geometric arrangements of important moieties or groups are often used as 3D queries for databases searching. Pharmacophores are most often derived from computed conformations or conformational ensembles of active compounds and less so from


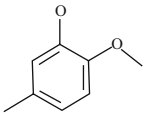
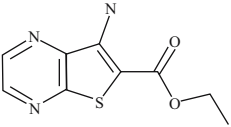
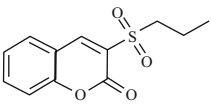
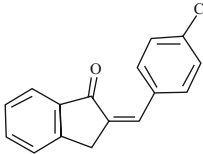
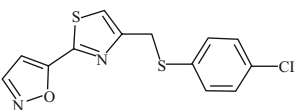
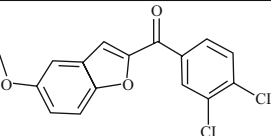
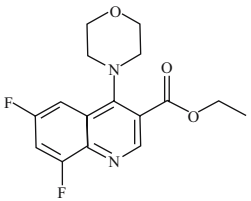
| Structures | Strings |
|---|---|
|  | <chem>c1ccccc1</chem> |
|  | <chem>Oc1cc(C)ccc1OC</chem> |
|  | <chem>s1c2[nH0]cc[nH0]c2c(N)c1C(=O)OCC</chem> |
|  | <chem>[S+2]([O-])([O-])(CCC)C1=Cc2ccccc2OC1=O</chem> |
|  | <chem>Clc1ccc(cc1)C=C1Cc2ccccc2C1=O</chem> |
|  | <chem>Clc1ccc(SCc2[nH0]c(sc2)c2o[nH0]cc2)cc1</chem> |
|  | <chem>Clc1ccc(cc1C1)C(=O)c1oc2ccc(OC)cc2c1</chem> |
|  | <chem>Fc1cc(F)c2[nH0]cc(c(N3CCOCC3)c2c1)C(=O)OCC</chem> |

Figure 1.11. Examples of structures and corresponding SMILES strings

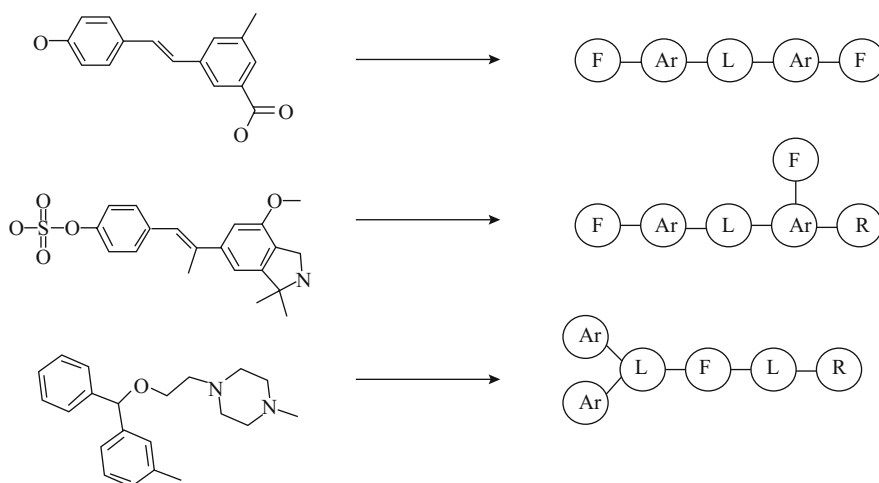


Figure 1.12. Examples of reduced graphs. Nodes corresponding to aromatic rings (Ar), aliphatic rings (R), functional groups (F) and linking groups (L) are shown (adapted from Gillet *et al.* 2003)

experimentally determined structures of ligands. Consequently, the majority of pharmacophore models used to identify similar compounds represent hypotheses of 3D features crucial for biological activity.

For database searching, pharmacophores are best defined by all possible distances between chosen groups or features (pharmacophore points). Therefore, as illustrated in Figure 1.13, they are best represented as a molecular graph (similar to reduced graphs). In this case, different from conventional graphs, however, nodes correspond to points (or centroids) and edges to inter-point distances, rather than bonds.

Graphs of query and test molecules can be compared by graph matching (subgraph detection) algorithms or systematic comparison of inter-feature distances. Two molecules are considered similar if their pharmacophores match for at least one predicted conformation. In order to explore conformational space and generate conformational ensembles, multiple compound conformations are typically generated by systematic conformational search (in increments) around rotatable bonds.

Three-point pharmacophores have traditionally been used for many applications but have recently been more and more replaced by four-point pharmacophores (Mason *et al.* 1999), which increases the complexity of the search but also the resolution of the pharmacophore analysis. This is the case because the additional point increases the total number of inter-point distances from three for a three-point pharmacophore to six for a four-point pharmacophore. Pharmacophore searching is further refined by assigning alternative features to each point (e.g., hydrogen bond acceptors, donors, or charged groups) and ranges to inter-point distances (rather than an exact distance). For example, five different features (e.g., atom types or groups) may be permitted for each point

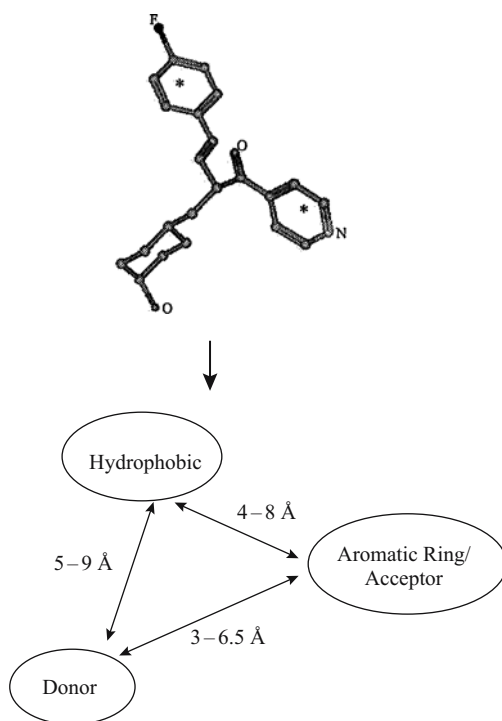


Figure 1.13. Example of a possible 3-point pharmacophore representation

and ten inter-point distance intervals for each range, which makes it possible to capture thousands (if not millions) of similar yet distinct conformation-dependent pharmacophore arrangements.

1.6.3 Fingerprints. Molecular fingerprints are widely used similarity search tools. They consist of various descriptors that are encoded as bit strings. As illustrated in Figure 1.14, bit strings of query and database compounds are calculated and quantitatively compared using similarity metrics such as the Tanimoto coefficient (Table 1.2). Fingerprint overlap between test compounds is regarded as a measure of molecular similarity. Thus, if the chosen coefficient reaches a pre-defined threshold value, compared molecules are considered to be similar.

In many fingerprint designs, each bit position accounts for a specific feature (for example, a structural fragment) and the bit is set on, if this feature is present in the molecule. Furthermore, value ranges of other molecular descriptors (e.g., molecular weight or the number of hydrogen bond acceptors) can also be incrementally encoded as bit strings. So designed fingerprints may consist of hundreds to thousands of bit positions. It is important to note that string representations of molecular