Basic Overview of Chemoinformatics†

Thomas Engel*

Chemical Computing Group AG, Kaiser-Wilhelm-Ring 11, 50672 Cologne, Germany

There is no particular point in time that determines when chemoinformatics was founded or established. It slowly evolved from several, often quite humble beginnings. Scientists in various fields of chemistry struggled with the development of computer methods which allowed them to manage the enormous amount of chemical information and to find relationships between the structure and properties of a compound. During the 1960s some early developments appeared that led to a flurry of activities in the 1970s. This review provides a general overview of basic methods in the specific fields of chemoinformatics, from encoding chemical compounds, storing and searching data in databases, to generating and analyzing these data. In addition, the chief interconnecting points of chemoinformatics applications are highlighted including the contributions of Johann Gasteiger to this field.

1. THE SCOPE OF CHEMOINFORMATICS

It was realized quite some decades ago that the huge amount of information accumulated in chemistry can only be kept manageable and accessible to the scientific community in electronic form. However, the new discipline emerging from storing, manipulating, and processing chemical information was not given a proper name. Scientists working in this field would state that they were dealing with "chemical information". As this expression did not make a clear distinction between librarianships and the development of computer methods, other scientists said they were working in "computational chemistry" which better stressed their efforts in developing new techniques for processing chemical information. However, chemoinformatics can easily be confused with computational chemistry which primarily focuses on theoretical quantum mechanical calculations.

The term "chemoinformatics" took shape only a few years ago but rapidly gained widespread use. Here are some of the first definitions of this term:

"The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization." by F. K. Brown.[1]

"Chemoinformatics—A new name for an old problem" by M. Hann and R. Green,[2] and "Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information." by G. Paris (August 1999 Meeting of the American Chemical Society), quoted by W. Warr at http://www.warr.com/warrzone2000.html.

The long history of applying informatics methods to chemical problems and the contribution of scientists who started out decades ago to develop this interdisciplinary field inspired J. Gasteiger and T. Engel led to a much broader definition of chemoinformatics:[3] "The application of informatics methods to solve chemical problems".

There are still different opinions about the scope of chemoinformatics and even on the term itself. Two neologisms, "chemoinformatics" and "cheminformatics", presently occur with near-equal frequency as a search in the database of the Chemical Abstracts Service reveals. In the following the term "chemoinformatics" will be used without providing extensive justification but to show the linguistic relation to bioinformatics. Besides this linguistic relationship there are many others, and often there is in fact no clear separation between chemoinformatics and bioinformatics. Traditionally, chemoinformatics has mainly dealt with small molecules, whereas bioinformatics addresses genes, proteins, and other larger chemical compounds. The structure and function of proteins, the binding of a ligand to its binding site, the conversion of a substrate within its enzyme receptor, and the catalysis of a biochemical reaction by an enzyme are areas where chemoinformatics and bioinformatics complement each other to deepen our insight and knowledge for biomolecular processes.

The tight connection between both disciplines is evident especially in drug design. Genomics methods are developed to identify protein targets for novel drug candidates. Drugs, on the other hand, are usually rather small molecules, and chemoinformatics methods are used to find new lead structures and optimize them to drug candidates.

One might surmise from the definition of chemoinformatics, the application of computer science—or informatics— for solving chemical problems has expanded into various areas of chemistry. Thus, it is very challenging to give a review about this broad topic where so many different application fields are covered; however, it is inevitable that important work of some people in the field of chemoinformatics might be omitted here. Some of these fields have since evolved into individual areas with more explicit scope, e.g. molecular modeling,[4,5] quantitative structure/activity relation-

---

† Dedicated to Professor Johann Gasteiger.
* Corresponding author fax: +49 221 9776 129 33; e-mail: thengel@chemcomp.com.

ships (QSAR),[6-10] chemometrics,[11-13] and molecular and quantum mechanics.[14-17] Specific reviews, textbooks, or even journals have been published for all of these scientific areas.

Since the number of references for each topic would go far beyond this scope, only the most relevant or updated literature is provided. More detailed information and continuative literature can also be found in two comprehensive reference books: the "*Handbook of Chemoinformatics*"[18] which is a revised version of "*The Encyclopedia of Computational Chemistry*".[19]

2. DIFFERENT ASPECTS OF CHEMOINFORMATICS

Many of the chemoinformatics approaches were initiated in the 1960s and early 1970s and were implemented into software systems that are now widely used and being continuously refined. Some research groups from the early days are still actively pursuing further developments; many new groups have joined this community, enriching it with novel ideas and new software systems.

The following sections serve as an introduction to chemoinformatics and provide a general overview on basic methods and some specific fields—from encoding chemical compounds, storing and searching data in databases, to generating and analyzing data.
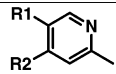
2.1. Representation of Chemical Compounds and Reactions. The input and output of chemical structures in computer programs and databases requires specific methods for their representation in computer-readable form. Structure representations also form the basis for methods used for (sub)structure searching in databases (see section 2.4).

Scientists may have diverse concepts and perceptions when talking about the structural information of chemical compounds. A compound may be characterized by its name, by its two-dimensional drawing, or by providing a three-dimensional molecular model. This hierarchy also reflects the various levels of refinement for representing chemical structures in electronic form—starting from linear notations to two-dimensional chemical graphs and 3D structure representations culminating in molecular surfaces.

Linear notations represent the structure of a chemical compound as a linear sequence of characters and numbers. The most popular linear notations are WLN (Wiswesser Line Notation), ROSDAL (Representation of Organic Structures Description Arranged Linearly), SMILES (Simplified Molecular Input Line Entry Specification), and SLN (SYBYL Line Notation). Although early linear notations were conceived before the advent of the computer, it was quickly realized that the compactness of such string notations is well suited to account for the expensive computational and storage resources. The heydays of line notations were in the 1960s.

Whereas the WLN—introduced in 1946[20-24]—and the ROSDAL syntax developed by S. Welford, J. Barnard, and M. F. Lynch in 1985 for the Beilstein Institute[25-29] has almost become obsolete, SMILES is still widely used nowadays. Compared to WLN and ROSDAL, SMILES is closely related to the native comprehension of organic chemists and is based on a few fundamental rules to convert a structure into a character string. SMILES was created in 1986 by David Weininger for chemical data processing[30-34] and has found widespread distribution as a universal chemical nomenclature for the representation and exchange of chemical structure

Table 1. Examples to Specific Structure Representations

Markush structure	
Fragment code	[OH]c1ccccc1
Fingerprint	0000100110100111
Hash code	5244987098423150

information, independent of software or hardware architecture. The SMILES notation has been subject to many extensions and enhancements since 1988 (e.g. SMARTS, SMIRKS, etc). The current definition and some examples of the SMILES notation can be found at http://www.daylight.com/dayhtml/smiles/index.html.

Other types of line notations presently used are, for instance, the Sybyl Line Notation (SLN) distributed by Tripos Inc.[35]

The most common international language for representing chemical structures is—by nature—a graphical one: the structure diagram. From a mathematical point of view, a structure diagram can be considered as a topological graph constructed with nodes or vertices (atoms) connected by edges or arcs (bonds). Consequently, many problems faced in processing structure information can be related to the mathematical field of graph theory.[36-42] Matrices for example, like adjacency and distance matrices, are commonly used in graph theory and may also be used to represent chemical structures. However, they represent only the connections within a structure and do not generally contain information about atom types and bond orders. Unlike adjacency and distance matrices, a bond-electron matrix describes the number of bonds and free valence electrons.[43,44] A variation derived from the matrix representation is the connection table (CT), first introduced in 1965.[45] Since the early 1980s, the connection table has been the predominant form of chemical structure representation in computer systems. A standard exchange format based on a connection table is the Molfile format developed by MDL Information Systems.[46,47] Graph-theoretical methods can be equally well applied to connection table representations of a molecule.

An important issue in processing structure information in database systems or in chemoinformatics applications is canonical numbering. Algorithms, like the one proposed by Morgan,[48] classify congeneric atoms of a compound and select invariant-labeled atoms. This provides the basis for a unique and unambiguous structure representation, which is also inherent to the perception of constitutional symmetry.[49] Additionally, various algorithms for ring perception have been developed over the years (see section 2.4).[51-54] These issues can also be solved with algorithms based on graph theory.

Further representations of chemical structures were developed for specific tasks (see Table 1): (i) Markush structures are mainly used as generic structure diagrams in patents for protecting a series of chemical compounds rather than single ones.[55] (ii) Fragment codes, e.g. SMARTS, specify substructure fragments as alphanumeric strings; they have always played an important role in chemical information systems.[56,57] (iii) Fingerprints or bit-wise representations of

BASIC OVERVIEW OF CHEMOINFORMATICS

J. Chem. Inf. Model., Vol. 46, No. 6, 2006 **2269**

fragments indicate specific substructure fragments in a compound with a binary bit (as present = 1 or absent = 0) at specific positions. These special characteristics of fingerprints can be used in similarity searches.[58-61] (iv) Hash coding is a very efficient method for creating a small digital "fingerprints", which are useful for registration purposes.[62-65]

So far, only representations of the constitution of chemical compounds have been discussed. The very fact that molecules are three-dimensional objects leads to a situation, in which different spatial atom arrangements exhibit the same unique chemical constitution, that is, stereoisomerism. The first chemoinformatics applications providing a module for the detection and analysis of stereochemical centers according to the CIP rules were the program systems LHASA[66] and CHIRON.[67] Excellent reviews about the representation and manipulation of stereochemistry are given by B. Rohde.[68,69]

A 3D structure representation can be automatically obtained from the connection table of a compound, which includes only the 2D constitution of the molecule.[70,71] The most well-known 3D structure generators are CORINA[72,73] and CONCORD.[74] Deriving the 3D molecular model raises the problem of conformational flexibility, the effect of molecule rotation around single bonds. Conformational flexibility is an important effect which allows, for instance, a drug molecule to adjust its shape to the structural requirements of a receptor protein.[71,75]

Thus two-dimensional structure diagrams as well as 3D molecular structures can form the basis for describing many chemical and physical properties of compounds. All models described so far, only represent the 3D skeleton of a molecule and not the actual appearance in space, in particular, the molecular surface. An important definition of a molecular surface was conceived by Richards.[76-79] The interpretation of molecular surfaces is particularly important wherever molecular interactions, reactions, and properties play a dominant role, as in drug design or docking experiments. Generally, molecular properties such as electrostatic potential, hydrophilicity/lipophilicity, and hydrogen bonding ability can be visualized by mapping them on surfaces.[80]

In the late 1960s, R. Langridge and co-workers developed methods to visualize 3D molecular models on the screens of cathode ray tubes. At the same time, G. Marshall started visualizing protein structures on graphics screens—the origin of molecular modeling.[4,5] Interactive molecular graphics in the computer and their corresponding graphical representation, such as the widely used ball-and-stick molecular models, were introduced in the 1960s by Levinthal.[81]

Since then, the development of computers and the challenging graphical visualization progressed immensely.

The next step is the dynamic aspect of chemistry—chemical reactions. Understanding chemical reactions is exceptionally important for the planning of many chemical experiments performed daily in the laboratory by chemists. In contrast to chemical structures, the chemoinformatics of chemical reactions is not so well developed—only a few scientific groups have dealt with this problem. There are various reasons for this situation: (i) the complexity of the problem, (ii) the need for much more chemical knowledge, and (iii) the lack of high quality electronic information on chemical reactions.

Nevertheless, important efforts have been made, and many of the results have been stored in reaction databases (see section 2.3).

With the enormous increase in the number of reaction databases—the larger ones contain millions of reactions—automatic knowledge acquisition from reaction databases has become of increasing interest. An important step of knowledge acquisition and to learn from individual reactions is classifying reaction instances to reaction types. The traditional and most simple classification scheme groups organic reactions into four categories: addition, elimination, rearrangement, and substitution. With the advent of computers the focus on reaction classification changed from describing full reaction classes in a general system to formulating algorithms for hierarchies to generate all examples of reactions from basic forms, which are themselves derived by computer.[82] A model-driven approach classifies reactions according to a predefined model or classification scheme.[83-85] On the other hand, data-driven approaches, e.g. the HORACE (Hierarchical Organization of Reactions through Attribute and Condition Eduction) system, rely on the computer to automatically analyze a set of reactions and dynamically generate classification results from given reaction data, without using predefined models.[86,87]

An excellent review about the process of knowledge acquisition and the classification of reaction instances into reaction types was described by L. Chen.[88]

To prepare chemical reactions for proper electronic processing, information should be given on the reaction center, on the bonds broken and made in a reaction, and how the valence electrons are rearranged in the course of a reaction.

Besides reaction classification, the chemical reactivity of reactions can also be quantified by different approaches, e.g. by the Frontier Molecular Orbital (FMO) theory[89] or by Linear Free Energy Relationships (LFER) which is based on the Hammett equation.[90-93]

2.2. The Data Types in Chemistry. Much of our knowledge of chemistry is gained from the analysis of experimental data. The acquisition and analysis of data are therefore important tasks that need much consideration.

Chemistry deals with quite a variety of data types which are very complex, probably the most complex among all scientific disciplines. Numerical and textual data may be electronically processed in a straightforward manner; the representation and processing of chemical structure data (2D/3D structures, bit vectors, hash codes, structural keys, fingerprints, etc.) as well as various molecular spectra data (NMR, IR, UV/VIS, MS, etc.) is more challenging.[94] But these data types are essential in chemical database design or data analysis. The various individual proprietary formats of software programs require only few standard data formats in order to facilitate the exchange of data, for instance spectral data.

The release of the (initially proprietary) MDL Molfile format to the scientific community in 1982 led to its acceptance as a general exchange format for chemical data sets. Several extensions have been made to the Molfile format leading to the SDfile, RGfile, Rxnfile, or RDfile, each one having special additional information on one or more molecules.[46,47] Besides the MDL Molfile format, other file formats are often used in chemistry. SMILES was already mentioned as linear notation. Another one, the PDB file format, is primarily used for storing 3D structure information

on biological macromolecules such as proteins and polynucleotides.[95,96] The CIF (crystallographic information file)[97,98] is another 3D structure information file format with more than three different file versions that are typically used in crystallography. In spectroscopy, JCAMP is applied as a spectroscopic exchange file format.[99] Here, two modifications can be distinguished: the JCAMP-DX and the JCAMP-CS formats. Whereas JCAMP-CS is an alternative to the Molfile and contains structure data, JCAMP-DX contains spectroscopic or chromatographic data. Finally, the CML (chemical markup language),[100−103] which is an extension of XML (extensible markup language), is an approach to unify all available chemical information for (Internet) publishing and computer processing. This standard data exchange format supports various chemical concepts, such as molecules, reactions, spectra, and other chemical data.

2.3. Chemical Databases and Data Sources. The multifaceted information about chemical compounds (more than 30 million compounds presently known in the CAS registry) and reactions, such as literature data, physicochemical properties, spectra, etc. can only be handled in a comprehensive manner by electronic methods. The amount of data and information constantly increases: each year more than 1 million new compounds and more than 700 000 publications that contribute in some way to chemical information.[104] Hence, chemistry was one of the first scientific disciplines using databases to store its treasure of information. Today, a competent overview on a topic can only be gained by querying databases and by data mining. It is only possible here to give a cursory overview of this dynamic topic and to refer to more detailed treatments.[105]

There is an amazing variety of databases in chemistry and its related sciences with useful or even indispensable content. The evaluation and selection of the appropriate information sources can be done by formal classification of databases according to their type of content into three broad categories: literature (textual), factual (alphanumeric), and structural (topological).

A strict separation into these categories and their subtypes is impossible, since many important chemistry databases cover several types of content. For instance, the Beilstein database contains structures, reactions, numerous physical properties of compounds, and related literature references.[106−107]

2.3.1. Literature Databases. Literature databases—divided into bibliographic and full-text databases—contain individual publications from the primary literature as objects, using character strings, e.g. alphanumeric characters, numbers, and special characters. Bibliographic and content descriptions (indexing) of publications such as author names, titles of articles, publication year, keywords, etc., as well as abstracts, can be retrieved from these databases in specific ways. A. Barth reviews in ref 108 how text may be indexed and retrieved in bibliographic databases such as the CA File from the Chemical Abstracts Service (CAS), Medline of the U.S. National Library of Medicine, Biosis (database version of the printed Biological Abstracts), or SciSearch. Most journals provide articles as full-text, like e-journals in ACS Journals (American Chemical Society),[109] Elsevier,[110] ScienceDirect,[111] etc.

Although general databases such as Chemical Abstracts[112] or Beilstein[106,108] do cover patents, there is a strong demand for specific bibliographic patent databases. One of the most important is the Derwent World Patent Index (WPI)[113] founded in 1963 as a printed information service. Another major source is INPADOC (International Patent Documentation Center)[114] from the European Patent Office (EPO).[115]

2.3.2. Factual Databases. In contrast to bibliographic databases which refer only to full publications in the primary literature, factual databases immediately provide the required textual or alphanumeric information: physical properties of chemical compounds, spectra, descriptions of research projects, legal information, etc. Although these databases often provide literature references to the origin of the data presented, the user does not necessarily need to go back to the primary literature as with bibliographic databases.

Factual databases can be divided into numerical databases, meta-databases, research project databases, and catalogs of chemical compounds.

Numerical databases primarily contain numeric data of chemical compounds such as physicochemical values and results of a series of measurements.

Typical numerical databases are Beilstein,[106,107] Gmelin,[116] SpecInfo,[117] KnowItAll,[118] DETHERM,[119] Cambridge Structural Database (CSD),[120−122] etc.

Besides the well-established databases Beilstein and Gmelin, there are many specialized numerical databases available, for example in the field of thermodynamics, safety, and toxicity (RTECS,[123,124] EINECS,[125−127] HSDB,[128,129] etc.), or biochemical sources such as the enzyme database BRENDA.[130,131]

Meta-databases provide information about the content of databases. Examples are the Gale Directory of Databases,[132,133] STNGuide,[134] or the Dialogue Dialindex.[135]

Research project databases include information on abstracts and reports categorized by research projects. Typical research project databases are UFORDAT (Environment Research in Progress) or Federal Research in Progress (FEDRIP).[136]

Catalog databases of chemical compounds including catalogs of many different chemical suppliers (including package sizes, purity information, prices, and ordering information) do not provide spectacular content but useful examples of factual databases. Some of the most important catalog databases are the Available Chemicals Directory[137] from MDL, CHEMCATS,[138] and ChemSources.[139]

2.3.3. Structure and Reaction Databases. Structure and reaction databases play a central role in chemistry since they contain information on chemical structures, both as individual compounds and as participants in reactions. The structure diagrams are not stored as graphics (i.e., pictures which are not structure-searchable) but represented e.g. as connection tables. This representation for two- and three-dimensional structures includes the topological arrangement of atoms and their connections as well as their stereochemistry.[3]

CAS Registry is the largest and most comprehensive structure database (2D and predicted 3D structures), containing millions of records of organic and inorganic compounds, peptide sequences, proteins, and nucleic acids.[111]

Another example of a structure database is the National Cancer Institute (NCI) database.[140]

Crystallographic structure databases, like the Inorganic Crystal Structure Database (ICSD),[141−143] Cambridge Structural Database (CSD),[120−122] and Protein Data Bank (PDB),[144,145]

BASIC OVERVIEW OF CHEMOINFORMATICS

J. Chem. Inf. Model., Vol. 46, No. 6, 2006 **2271**

provide "real" 3D structures from X-ray crystallographic structure analyses.

Patent databases including special structures are called Markush databases e.g. MARPAT and Merged Markush Service[146] by Questel-Orbit.[147]

In reaction databases, structures of reaction participants are stored in a similar manner. Additionally, the role of each compound in the reaction (starting material, product, solvent, reagent, catalyst), the reaction center information (formal, not mechanistic), and atoms added/eliminated as well as bonds formed/broken in the reaction are stored.

A large structure and reaction database is SPRESI (SPeicherung und REcherche Strukturchemischer Information; cp. ChemReact).[148]

Additional reaction databases e.g. CASREACT,[149] Chem-Inform, or Beilstein,[150] contain only reaction information.

G. Paris presented an interesting overview of chemical structure databases and retrieval strategies to access the required information.[151]

2.3.4. Molecular Biology Databases. Many molecular biology databases with sophisticated topics for different problems have been developed. Since 1996 the first issue of each journal volume of "*Nucleic Acid Research*" has been reserved for the presentation of molecular biology databases.[152] A comprehensive catalog on the Internet is DBCAT currently listing 511 databases.[153,154]

The largest bibliographic database in the biological sciences is BIOSIS (content of Biological Abstracts).[155] Besides such textual databases that provide bibliographic information and amino acid as well as nucleotide sequence databases have attained an even more important role in biochemistry. The three largest primary sequences databases are GenBank (USA),[156] EMBL (European Molecular Biology Laboratory),[157] and DDBJ (DNA Data Bank of Japan).[158]

Furthermore very multifaceted molecular biology databases have emerged (family, ligand databases, etc.), and the reader is referred to the review of von Homeyer[159] for further information.

In addition to most of the proprietary databases, the Internet presents a wide range of opportunities for searching chemical information in the life sciences: finding and sharing information, building databases, processing data, performing computations, etc. It is difficult to put these Internet sources into a rigid scheme. However, the sharing of information is probably the most important aspect of the Internet; the chemical search engine eMolecules[160] is a useful approach for implementation.

2.4. Search Methods. Storing and searching chemical structures and associated information in databases is probably the earliest foundation of what is now called chemoinformatics. The majority of actual databases in chemistry contain chemical structures in computer-readable form, either 2D or 3D, connection tables, or other representations. Therefore one of the primary methods for getting access to chemical information is searching for chemical structures or for a set of compounds that share a specific substructure or structural similarities.

The structure representation is a major key in finding the appropriate results in a data set. Therefore the representation should include as much functionality of the structure as possible, and it should also be unique and unambiguous. Various approaches have been devised for this purpose. They comprise the use of molecular formulas, molecular weights,

trade and/or trivial names, various canonical line notations, registry numbers, constitutional diagrams (2D representations), atom coordinates (2D or 3D representations), topological indices, hash codes, etc.

Rigorous methods of structure search (e.g. atom-by-atom) are computationally expensive; alternative approaches for molecular identification were introduced in the 1950s. In particular, structure perception methods are developed by consideration of items such as completeness, optimality, nonredundancy, and time and memory complexity.

2.4.1. Full Structure Search. The principle of a full structure search algorithm is to search for identical structures using molecular identifiers (e.g. topological index,[161,162] hashcode,[64] or linear notation). Unfortunately, most of these representations are nonunique, and the requirement of nonredundancy and optimum is not fulfilled. Therefore, a general structure search scenario applies molecular identification technology to narrow down or to filter the number of structure search hit candidates. Then, the actual hits are determined by an atom-by-atom match algorithm which is based exclusively on graph theory. This algorithm can also be used in substructure search.

Another structure search method used in the CAS registry system is based on a compact and easy calculated numerical representation of a chemical structure called the augmented connectivity molecular formula (ACMF).[163]

2.4.2. Substructure Search. Substructure searching is the process of identifying parts of a given structure that are equivalent to a specified query substructure. In graph-theoretical terms substructure searching verifies whether a query graph is isomorphic with a subgraph of another target graph by an atom-by-atom match algorithm, which represents a nonpolynomial (NP)-complete problem.[164] The atom-by-atom match algorithm has been utilized since 1957. A review on this algorithm was published by Barnard in ref 165. Representative algorithms are the Sussenguth algorithm (based on a partitioning procedure),[166] Figueras' algorithm (based on set reduction),[167] Ullmann's algorithm (based on backtracking[168,169] and relaxation refinements),[170] von Scholley's algorithm (based on relaxation),[171] and Xu's algorithms (based on backtracking and partial ordered sets).[172]

The methods available for 2D structure and substructure searching have been reviewed in ref 173.

Other important graph-based structure perceptions such as identification of equivalent atoms, determination of maximal common substructure, ring detection, the calculation of topological indices, etc. may be also comprised in substructure search algorithms.[174]

3D structure and substructure searching is another topic of high interest since the main purpose of 3D searching is to locate potential structure candidates that may fit into a given receptor in drug discovery. An excellent review about this topic has been written by O. F. Güner et al.[175]

2.4.3. Markush Search. Markush structures are used for the representation of families of structures in patents. Those generic structures pose specific retrieval problems to the search systems and are therefore complicated.

Two topological search systems for Markush structures in patents were developed and commercialized in the 1980s: Markush DARC (MDARC)[176] and MARPAT.[177,178] Both systems were built on the efforts of the Michael Lynch group at the University of Sheffield, which published several

papers on their research from about 1981 to 1995.[179] MDARC was released to the public in February, 1989, containing indexing from 1987 forward, and MARPAT was released to the public in June, 1990, with indexing from 1988 forward. Several reviews and papers are published and have compared the two systems.[55,180,181]

2.4.4. Similarity Search. The specification of a certain substructure is quite often insufficient for the definition of a family of related structures. Occasionally one wants to obtain a group of structures that have a number of structural features in common: structures that are similar to each other.[61,182,183]

A similarity search compares a set of characteristics describing the target structure with the corresponding sets of characteristics for each of the database structures. The measure of similarity between the target and each database structure is calculated based on the degree of resemblance of these two sets of characteristics. The database structures are usually sorted into order of decreasing similarity with the target.

The most important similarity measure has three main components: the structural representation that is used to characterize the molecules; the weighting scheme that is used to differentiate more important features from less important features; and the similarity coefficient that is used to quantify the degree of similarity between pairs of molecules. These coefficients may be based on fragments (e.g. the Tanimoto coefficient), on topological indices (and other calculated physicochemical properties, e.g. molecular connectivity, dipole moment, etc.), and on graphs (e.g. maximum common subgraphs).[184]

2.5. Calculation of Descriptors, Physical and Chemical Data. One major topic of chemoinformatics is how chemical structure information can be correlated with physical, chemical, or biological data and how such a model, once established, can be used for the prediction of new data. Thus, the appropriate description of chemical structures for the property to be modeled is an essential task. With a multitude of different types of descriptors being available (more than 1500), great care has to be devoted to the selection of the most suitable set of descriptors.

A comprehensive treatise on all available molecular descriptors was published by R. Todeschini and V. Consonni.[185]

Quantitative structure−property relationships (QSPR) or quantitative structure−activity relationships (QSAR) (see section 3) are usually established by inductive learning methods. An essential requirement for such learning methods to be invoked is that all structures of a data set are represented by the same, fixed number of descriptors. Thus, the representations of 2D structures (connection tables) or 3D structures (e.g. Cartesian coordinates) cannot be used as such since the number of descriptors would depend directly on the number of atoms in a molecule—the more atoms, the more descriptors.

Therefore, chemical structure information has somehow to be mathematically transformed to provide for any molecule descriptors, irrespective of the size of a molecule, and of the number of atoms in a molecule.

Various structure coding methods mainly developed in Gasteiger's research group take account of various physicochemical effects and can be scaled either to consider only the constitution of a molecule, also to reflect the 3D structure, or even include molecular surface properties.[186]

It should, however, not be forgotten that beyond inductive learning processes there is also sometimes a deductive access to physical or chemical data through direct computational methods.

Primary to this are quantum mechanical methods that allow the calculation of a variety of physical properties such as total energy, dipole moment, polarizability, ionization potential, etc., and 3D structure. Furthermore, a variety of electronic and energy values (such as coefficients and energies of frontier molecular orbitals) can be calculated that can be used as descriptors in correlating structure and properties.

Despite the enormous advances in hardware technology and software development, quantum mechanical calculations might still be too time-consuming to apply them to large molecules or large data sets. Simpler methods such as molecular mechanics or even empirical calculations might provide an alternative. Molecular mechanics shows the merits of this approach, particularly for proteins and other large biological systems.[187]

2.6. Methods for Data Analysis. In science, in general, and in chemistry and pharmaceutical research, in particular, huge amounts of data are produced. The intrinsic information in these data is often difficult to grasp. In many cases, not only the data themselves are interesting but also the intrinsic relationships among the data are of particular interest.

Chemical data contains information about various characteristics of chemical compounds, and a wide spectrum of methods is applied to extract the relevant information from the data sets. Data analysis, however, deals not only with the extraction of primary information from data but also with the generation of secondary information, for example the generation of descriptive models which can be used for prediction purposes. This task can be accomplished by machine learning methods.

However, the pieces are assembled to establish a relationship between chemical compounds and their properties: descriptors represent the compounds and the experimental data. The relationship has then to be established by some learning method, a statistical or pattern recognition method, or an artificial neural network. These methods are grouped here as "inductive learning methods" because they learn from sets of pairs of objects and their properties. Other buzzwords are also used in this context, with different authors having different definitions: the term "machine learning" [188] is often used among computer scientists; the expression "data mining" was introduced some years ago,[189] mainly in the context of processing data from large data sets or from different sources/databases.

Clearly, many of the topics touched here—particularly on chemometrics,[11−13] neural networks,[190] or evolutionary algorithms[191,192]—are also covered by entire books of their own.

A second major topic, expert systems contain a knowledge base which either has been built by interviewing experts or by using inductive learning methods.[193]

3. APPLICATIONS OF CHEMOINFORMATICS METHODS

Since the applications of chemoinformatics methods are many, any selection of examples must be arbitrary and fragmentary.
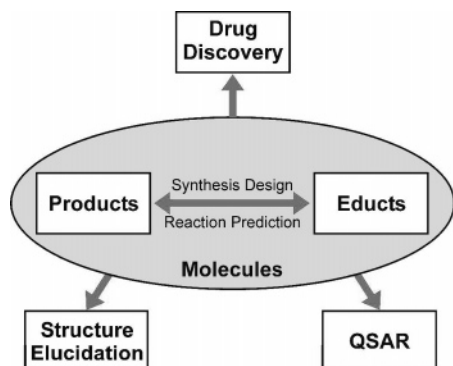
Figure 1. The interconnection of fundamental task in chemistry with chemoinformatics applications.

A major task of chemists is to make compounds with desired properties. Thus, the first fundamental task in chemistry is to make inferences on which structure might have the desired property (Figure 1). This is the domain of establishing structure−property or structure−activity relationships (SPR or SAR) or even finding such relationships in a quantitative manner (QSPR or QSAR).

QSAR originates in the work by Hammett and Taft in the 1950s that was dedicated to the separation and quantification of steric and electronic influences on chemical reactivity.[194,195] Building on this, Hansch started from 1964 to quantify the steric, electrostatic, and hydrophobic effects and their influences on a variety of properties, not the least on the biological activity of drugs. In 1964, the Free-Wilson analysis was introduced to relate biological activity to the presence or absence of certain substructures in a molecule.[196,197]

QSAR is closely associated with the calculation of molecular descriptors.[198]

As already mentioned there are many books[6−10] and journals[199,200] dealing with this topic.

If an appropriate property could be matched with structural features, suitable sets of compounds have to be synthesized. Thus, a strategy has to be established how to synthesize these compounds, which reaction or sequence of reactions to perform to make a structure from available starting materials, etc. (Figure 1). This is the domain of synthesis design and the planning of chemical reactions.[201−204] There, also the analysis of reaction information retrieved from reaction databases is very important. Indexing chemical reactions for database building was presented in 1967. Two years later, E. J. Corey and W. T. Wipke presented its first steps in the development of a system for computer-assisted synthesis design.[205]

Once a reaction has been performed, it has to be verified whether the reaction took the desired course and whether the desired structure was obtained (Figure 1). The factors influencing the course of chemical reactions are many, and the knowledge of chemical reactions is still too cursory. However, the structure of the reaction product has to be established. This is the domain of structure elucidation that largely builds on information obtained from various spectroscopic methods (infrared, NMR, mass spectra, etc.). The simultaneous use of different spectroscopic information at various stages of the structure elucidation process presented a challenge to chemoinformatics from the very beginning.

The development of automatic systems for structure elucidation can be traced back more than three decades.[206−209]

However, the access to spectral information in structure elucidation is an indispensable requirement, but the number of spectra stored in spectroscopic databases is minute in comparison to the number of known compounds. Thus one needs methods for the simulation of spectra in order to predict those spectra that are not contained in databases.[210−213]

Most applications in chemoinformatics cover the area of drug design. Central tasks herein are the establishment of a relationship between a chemical structure and its biological activity and the prediction of pharmacological properties (e.g. binding affinities, log P, pK_a, ADMET, etc.) [214,215] in addition to lead finding. Therefore, huge amounts of data are gathered in the drug development process, particularly so through the synthesis of combinatorial libraries and subsequent high-throughput screening.[216−218] QSAR methods are applied to those data sets to guide the further development of a new drug. Probably the most widely used QSAR method is CoMFA.[219−222] The pharmacophore concept shows how inferences on the 3D structural requirements for a biologically active compound can be drawn from a set of active ligands.[223−227] De novo design, on the other hand, needs the 3D structure of the binding pocket of the receptor and fills it with 3D structures of potential ligands.[228−231] Having both the 3D structure of various ligands and of the receptor, the question is then which one fits or docks best.[232−236]

All tasks are generally too complex to be solved from first principles. They are, therefore, tackled by making use of prior information and of information that has been condensed into knowledge. The amount of information that has to be processed is often quite large. All compounds have a series of physical, chemical, or biological properties which can be processed in many different ways (synthesized by a wide range of reactions, characterized by a host of spectra). This immense amount of information can only be processed by electronic means, by the power of the computer, and with applications in chemoinformatics.

4. SUMMARY AND OUTLOOK

The instruments and software methods available for chemistry research will continue to drown us in data. The incredible challenge of this is to manage these data to increase our chemical knowledge, to better understand, and ultimately to exploit the results of our experiments. With this knowledge, better experiments can be planned, and fewer yet more efficient experiments may be performed. The result is an improved insight into the relationships between chemical structures and their properties. In this context, chemoinformatics has matured into a scientific discipline that will, and in some cases has already, changed the way we perceive chemistry.

With better hardware and software more exact methods can be used for the representation of chemical structures and reactions. More and more quantum mechanical calculations and further applications can be utilized for chemoinformatics tasks.

Additionally the combination of chemoinformatics and bioinformatics allows studying the problems in genomics, proteomics, metabolomics, and drug design with a battery of specific methods.

Also teaching chemoinformatics will become of increasing importance. In fact, chemoinformatics curricula have already been integrated at specific universities (University of Sheffield, U.K.; University of Manchester Institute of Science and Technology, U.K.; Indiana University, U.S.A.; and University of Erlangen-Nuremberg, Germany).

REFERENCES AND NOTES

(1) Brown, F. K. Chemoinformatics: What is it and How does it Impact Drug Discovery. *Annu. Rep. Med. Chem.* **1998**, *33*, 375−384.

(2) Hann, M.; Green, R. Chemoinformatics−A new name for an old problem. *Curr. Opin. Chem. Biol.* **1999**, *3*, 379−383.

(3) Gasteiger J.; Engel T. *Chemoinformatics−A Textbook*; Wiley-VCH: Weinheim, 2003; 600 pp.

(4) Hoeltje, H.-D.; Sippl, W.; Rognan, D.; Folkers, G. *Molecular Modeling: Basic Principles and Applications*; John Wiley & Sons: Chichester, U.K., 2003; 270 pp.

(5) Leach, A. R. *Molecular Modelling: Principles and Applications*; Pearson Education Limited: 2001; 744 pp.

(6) Devillers, J. *Genetic Algorithms in Molecular Modeling. Principles of QSAR and Drug Design*; Academic Press: New York, 1996; 327 pp.

(7) Kubinyi, H. *3D Qsar in Drug Design: Volume 1: Theory Methods and Applications*; ESCOM, Science Publishers B.V.: Leiden, 1993; 788 pp.

(8) Ivanciuc, O. Topological Indices in QSAR and QSPR. In *SAR QSAR Environ. Res.* **2001**, *12* (1−2), 257 pp.

(9) Diudea, M. V. *QSPR/QSAR Studies by Molecular Descriptors*; Nova Sci. Publ: Huntington, NY, 2001; 438 pp.

(10) Devillers, J.; Balaban, A. T. *Topological indices and related descriptors in QSAR and QSPR*; Gordon & Breach: Amsterdam, The Netherlands, 1999; 811 pp.

(11) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. C. M.; De Jong, S.; Smeyers-Verbeke, J. *Handbook of chemometrics and qualimetrics: Part A*; Elsevier: Amsterdam, 1997; 884 pp.

(12) Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. C. M.; De Jong, S.; Smeyers-Verbeke, J. *Handbook of chemometrics and qualimetrics: Part B*; Elsevier: Amsterdam, 1998; 876 pp.

(13) Otto, M. Chemometrics. *Statistics and Computer Application in Analytical Chemistry*; WILEY-VCH.; Weinheim, 1998; 314 pp.

(14) Hameka, H. F. *Quantum Mechanics: A Conceptual Approach*; John Wiley & Sons: Chichester, U.K. 2004; 224 pp.

(15) *Quantum Mechanics*, 4th ed.; Rae, A. I. M., Ed.; Institute of Physics Publishing: Bristol, U.K., 2002; 301 pp.

(16) Atkins, P. W.; Friedman, R. S. *Molecular Quantum Mechanics*, 3rd ed.; Oxford University Press: Oxford, U.K., 2000; 568 pp.

(17) Townsend, J. S. *A Modern Approach to Quantum Mechanics*; University Science: Sausalito, CA, 2000; 497 pp.

(18) *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, 1870 pp.

(19) *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998.

(20) Wiswesser, W. J. *A Line-Formula Chemical Notation*; Y. Thomas Crowell Company: New York, 1954; 149 pp.

(21) Smith, E. G.; Baker P. A. *The Wiswesser Line-Formula Chemical Notation*; Chemical Information Management: Cherry Hill, NY, 1975.

(22) Wiswesser, W. J. Historic Development of Chemical Notations. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 258−263.

(23) Wiswesser, W. J. How the WLN began in 1949 and how it might be in 1999. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 88−93.

(24) Granito, C. E.; Rosenberg, M. D. Chemical Substructure Index (CSI). A new Research Tool. *J. Chem. Doc.* **1971**, *11*, 251−256.

(25) Barnard, J. M.; Jochum, C. J.; Welford, S. M. ROSDAL: A Universal Structure/Substructure Representation for PC-host Communication. In *Chemical Structure Information Systems: Interfaces Communication; and Standards*; Warr, W. A., Ed.; ACS Symposium Series 400; American Chemical Society: Washington, DC, 1989; pp 76−81.

(26) Rohbeck, H.-G. Representation of Structure Description Arranged Linearly. In *Software Development in Chemistry 5*; Gmehling, J., Ed.; Springer-Verlag: Berlin, Heidelberg, 1991; pp 49−58.

(27) Heller, S. R. *The Beilstein Online Database−Implementation, Content and Retrieval*; American Chemical Society: Washington, DC, 1990; 168 pp.

(28) Goebels, L.; Lawson, A. J.; Wisniewski, J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC−Compatible Names. 2. Nomenclature of Chains and Rings. *J. Chem. Inf. Comput. Sci.* **1991**, *31* (2), 216−225.

(29) Wisniewski, J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC−Compatible Names. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 324−332.

(30) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.

(31) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97−101.

(32) Weininger, D. SMILES. 3. DEPICT. Graphical Depiction of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (3), 237−243.

(33) Hinze, J.; Welz, U. Broad SMILES. In *Software Development in Chemistry 10*; Gasteiger, J., Ed.; Springer-Verlag: Berlin, Heidelberg, 1991; pp 59−65.

(34) Bone, R. G. A.; Firth, M. A.; Sykes, R. A. SMILES Extensions for Pattern Matching and Molecular Transformations: Applications in Chemoinformatics. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 846−860.

(35) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(36) Balaban, A. T. Chemical Graphs: Looking Back and Glimpsing Ahead. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (3), 339−50.

(37) Balaban, A. T. *Chemical Applications of Graph Theory*; Academic Press: London, 1976; 389 pp.

(38) Schultz, H. P. Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (3), 227−8.

(39) Ivanciuc, O. Graph Theory in Chemistry. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; pp 103−138.

(40) Beck, A.; Bleicher, M.; Crowe, D. *Excursion into Mathematics*; Worth Publishers: 1969; 499 pp.

(41) Trinajstic, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1991; 350 pp.

(42) Tarjan, R. E. Graph algorithms in chemical computation. In *Algorithms for chemical computations*; Christoffersen, R. E., Ed.; ACS Symposium Series No. 46; American Chemical Society: Washington, DC, 1977; pp 1−19.

(43) Spialter, L. Atom Connectivity Matrix (ACM) and its Characteristic Polynomial (ACMCP). A new Computer-Oriented Chemical Nomenclature. *J. Am. Chem. Soc.* **1963**, *85* (13), 2012−2013.

(44) Dugundji, J.; Ugi, I. Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19−64.

(45) Gluck, D. J. A chemical Structure Storage and Search System Developed at Du Pont. *J. Chem. Doc.* **1965**, *5*, 43−51.

(46) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(47) Ctfile Formats, MDL Information Systems Inc., CA, 1998. http://www.mdli.com/downloads.

(48) Morgan, H. L. Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(49) Ivanciuc, O. Canonical Numbering and Constitutional Symmetry. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; John Wiley-VCH: Weinheim, 2003; pp 139−160.

(50) Razinger, M.; Balasubramanian, K.; Munk, M. E. Graph automorphism perception algorithms in computer-enhanced structure elucidation. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 197−201.

(51) Downs, G. M. Ring Perception. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; pp 161−177.

(52) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Review of Ring Perception Algorithms for Chemical Graphs. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172−187.

(53) Balaban, A. T.; Filip, P.; Balaban, T. S. Computer Program for Finding all Possible Cycles in Graphs. *J. Comput. Chem.* **1985**, *6*, 316−329.

(54) Hanser, Th. Jauffret, Ph.; Kaufmann, G. A New Algorithm for Exhaustive Ring Perception in a Molecular Graph. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (6), 1146−1152.

(55) Barnard, J. M. A Comparison of Different Approaches to Markush Structure Handling. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 64−68.

(56) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular Diversity and Representativity in Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.

(57) Wild, D.; Blankley, C. J. Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155−162.

(58) Lewis, R. A.; Pickett, S. D.; Clark, D. E. Computer-Aided Molecular Diversity Analysis and Combinatorial Library Design. In *Reviews in*

BASIC OVERVIEW OF CHEMOINFORMATICS

J. Chem. Inf. Model., Vol. 46, No. 6, 2006 **2275**

Computational Chemistry; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 2000; Vol. 16, pp 8−51.

(59) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987; 254 pp.

(60) Rhodes, N.; Willett, P. Bit-String Methods for Selective Compound Acquisition. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 210−214.

(61) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983−996.

(62) Wipke, W. T.; Krishnan, S. K.; Ouchi, G. I. Hash Functions for Rapid Storage and Retrieval of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 32−37.

(63) Zupan, J. *Algorithms for Chemists*; John Wiley & Sons Ltd.: Chichester, 1989; 290 pp.

(64) Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. The Chemical Abstracts Service Chemical Registry System. II. Augmented connectivity molecular formula. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 94−97.

(65) Ihlenfeldt, W. D.; Gasteiger, J. Hash Codes for the Identification and Classification of Molecular Structure Elements. *J. Comput. Chem.* **1994**, *Vol 15* (8), 793−813.

(66) Mata, P.; Lobo, A. M.; Marshall, C.; Johnson, A. P. Implementation of the Cahn−Ingold−Prelog System for Stereochemical Perception in the LHASA Program. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 491−504.

(67) Hanessian, S.; Franco, J.; Gagnon, G.; Laramée, D.; Larouche, B. Computer-Assisted Analysis and Perception of Stereochemical Features in Organic Molecules Using the CHIRON Program. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 413−425.

(68) Rohde, B. Representation and Manipulation of Stereochemistry. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; John Wiley-VCH: Weinheim, 2003; pp 206−230.

(69) Barnard, J. M.; Cook, A. P. F.; Rohde, B. Storage and searching of stereochemistry in substructure search systems. *Chem. Inf. Syst.* **1990**, 29−41.

(70) Sadowski, J. 3D Structure Generation. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; John Wiley-VCH: Weinheim, 2003; pp 231−261.

(71) Sadowski, J.; Schwab, C. H.; Gasteiger, J. 3D structure generation and conformation searching. *Comput. Med. Chem. Drug Discovery* **2004**, 151−212.

(72) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *7*, 2567−2581.

(73) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(74) Pearlman, R. S. Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Auto. News* **1987**, *2*, 1−7.

(75) Schwab, C. H. Conformational Analysis and Searching. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; John Wiley-VCH: Weinheim, 2003; pp 262−301.

(76) Lee, B.; Richards, F. M. Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55* (3), 379−400.

(77) Connolly, M. L. Analytical Molecular Surface Calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548−558.

(78) Connolly, M. L. Solvent-Accessible Surfaces of Proteins and Nucleic Acids. *Science* **1983**, *221*, 709−713.

(79) Duncan, B. S.; Olson, A. J. Approximation and Visualization of Large-Scale Motion of Proteins Surfaces. *J. Mol. Graphics Modell.* **1995**, *13*, 250−257.

(80) Brickmann, J.; Exner, T.; Keil, M.; Marhöfer, R.; Moeckel, G. *The Encyclopedia of Computational Chemistry, Molecular Models: Visualization*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; pp 1679−1693.

(81) Levinthal, C. Molecular Model-Building by Computer. *Sci. Am.* **1966**, *214*, 42−52.

(82) Hendrickson, J. B.; Chen, L. *Reaction Classification in Encyclopedia of Computational Chemistry*; Schleyer, P. (v) R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; Vol. 4, pp 2381−2402. http://www.mrw.interscience.wiley.com/ecc.

(83) Hendrickson, J. B.; Sander, T. COGNOS: A Beilstein-Type System for Organizing Organic Reactions. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 251−260.

(84) Ugi, I.; Gillespie, P. D. Chemistry and Logical Structures. 4. Matter-Preserving Synthetic Pathways and Semiempirical Computer-Assisted Planning of Syntheses. *Angew. Chem., Int. Ed. Engl.* **1971**, *10*, 915−919.

(85) Dugundji, J.; Ugi, I. Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19−64.

(86) Gelernter, H.; Rose, J. R.; Chen, C. Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Machine Learning. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 492−504.

(87) Rose, J. R.; Gasteiger, J. HORACE: An Automatic System for the Hierarchical Classification of Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74−90.

(88) Chen, L. Reaction Classification and Knowledge Acquisition. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; John Wiley-VCH: Weinheim, 2003; pp 348−388.

(89) Fleming, I. *Frontier Orbitals and Organic Chemical Reactions*; John Wiley & Sons: New York, 1976; 249 pp.

(90) Hammett, L. P. *Physical Organic Chemistry*; McGraw-Hill: New York, 1970; 480 pp.

(91) Chapman, N. B.; Shorter, J. *Advances in Linear Free Energy Relationships*; Plenum Press: London, 1972; 486 pp.

(92) Chapman, N. B.; Shorter, J. *Correlation Analysis in Chemistry*; Plenum Press: London, 1978; 546 pp.

(93) Shorter, J. Linear Free Energy Relationships (LFER). In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; Vol. 2, pp 1487−1496. http://www.mrw.interscience.wiley.com/ecc.

(94) Tomczak, J. Data Types. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; John Wiley-VCH: Weinheim, 2003; pp 392−409.

(95) Bernstein, R. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(96) PDB Format Description. http://www.rcsb.org.

(97) Hall, S. R.; Allen, F. H.; Brown, I. D. The Crystallographic Information File (CIF): A new Standard Archive File for Crystallography. *Acta Crystallogr.* **1991**, *A47*, 655−685.

(98) Macromolecular Crystallographic Information File. http://www.iu-cr.org/iucr-top/cif/mmcif.

(99) Gasteiger, J.; Hendriks, B. M. P.; Hoever, P.; Jochum, C.; Somberg, H. JCAMP-CS: A Standard Exchange Format for Chemical Structure Information in Computer-Readable Form. *Appl. Spectrosc.* **1991**, *45*, 4−11.

(100) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 928.

(101) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the World-Wide Web. 2. Information Objects and the CMLDOM. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1113−1123.

(102) Gkoutos, G.; Murray-Rust, P.; Rzepa, H. S.; Wright, M. Chemical Markup, XML, and the World-Wide Web. 3. Toward a Signed Semantic Chemical Web of Trust. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1124−1130.

(103) Chemical Markup Language. http://www.xml-cml.org.

(104) http://www.cas.org/PRINTED/printca.html.

(105) Engel, T.; Zass, E. Chemical Information Systems and Databases. In *Comprehensive Medicinal Chemistry II*; Kubinyi, H., Ed.; Elsevier Ltd.: 2007; Vol. 3, in press.

(106) Lawson, A. J. The Beilstein Database. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2, pp 608−628.

(107) Heller, S. R. *The Beilstein System: Strategies for Effective Searching*; American Chemical Society: Washington, DC, 1998.

(108) Barth, A. Bibliographic Databases. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2, pp 507−522.

(109) http://pubs.acs.org/about.html.

(110) http://www.elsevier.com/wps/find/homepage.cws_home.

(111) http://www.sciencedirect.com/.

(112) Fisanick, W.; Shively, E. R. The CAS Information System: Applying Scientific Knowledge and Technology for Better Information. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2, pp 556−607.

(113) Thomson Derwent (producer of WPI). http://www.derwent.com/.

(114) Lingua, D. G. *World Pat. Inf.* **2005**, *27*, 105−111.

(115) EPO (European Patent Office). http://www.european-patent-office.org/index.en.php.

(116) Nebel, A.; Toelle, U.; Maass, R.; Olbrich, G.; Deplanque, R.; Lister, P. The Integrated Gmelin Information System. New Developments in Information Processing. *Anal. Chim. Acta* **1992**, *265*, 305−312.

(117) Barth, A. SpecInfo: An Integrated Spectroscopic Information System. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 52−58.

(118) http://www.biorad.com.

(119) http://www.dechema.de/Detherm.html/.

(120) http://www.ccdc.cam.ac.uk/products/csd/.

(121) Allen, F. H. The Cambridge structural database. *ACA Trans.* **2000**, Volume Date 1997, *32*, 1−5.

(122) Allen, F. H.; Lipscomb, K. J.; Battle, G. The Cambridge Structural Database (CSD) of Small Molecule Crystal Structures. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2, pp 645−666.

(123) http://www.cdc.gov/niosh/rtecs/default.html.

(124) Sweet, D. V.; Anderson, V. P.; Fang, J. C. F. An Overview of the Registry of Toxic Effects of Chemical Substances (RTECS): Critical Information on Chemical Hazards. *Chem. Health Saf.* **1999**, *6*, 12−16.

(125) http://ecb.jrc.it/; http://ecb.jrc.it/existing-chemicals/.

(126) Heidorn, C. J. A.; Rasmussen, K.; Hansen, B. G.; Norager, O.; Allanou, R.; Seynaeve, R.; Scheer, S.; Kappes, D.; Bernasconi, R. IUCLID: An Information Management Tool for Existing Chemicals and Biocides. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 779−786.

(127) Stephan, U.; Strobel, U. The IUCLID Database−Utilization and Limitations. *Nachr. Chem.* **2003**, *51*, 1052−1053.

(128) http://www.nlm.nih.gov/pubs/factsheets/hsdbfs.html.

(129) Fonger, G. C. Hazardous Substances Data Bank (HSDB) as a Source of Environmental Fate Information on Chemicals. *Toxicology* **1995**, *103*, 137−145.

(130) Wexler, P. The U.S. National Library of Medicine's Toxicology and Environmental Health Information Program. *Toxicology* **2004**, *198*, 161−168.

(131) http://www.brenda.uni-koeln.de/.

(132) Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D. BRENDA, The Enzyme Database: Updates and Major New Developments. *Nucleic Acids Res.* **2004**, *32*, D431−D433.

(133) http://www.galegroup.com/.

(134) http://www.stn-international.de/stndatabases/databases/stnguide.html.

(135) http://library.dialogue.com/bluesheets/html/bl0411.html.

(136) http://grc.ntis.gov/fedrip.htm. http://www.stn-international.de/stndatabases/databases/fedrip.html (STN). http://library.dialogue.com/bluesheets/html/bl0266.html (Dialogue).

(137) http://www.mdli.com/products/experiment/index.jsp (MDL). http://chemfinder.cambridgesoft.com/chemicals/chemacxpro.asp (CambridgeSoft). http://www.daylight.com/products/databases/ACD.html (Daylight).

(138) http://www.cas.org/CASFILES/chemcats.html. http://www.stn-international.de/stndatabases/databases/chemcats.html (STN CHEMCATS).

(139) http://www.chemsources.com/chemonline.html.

(140) Ihlenfeldt, W.-D.; Voigt, J. H.; Bienfait, B.; Oellien, F.; Nicklaus, M. C. Enhanced CACTVS Browser of the Open NCI Database. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 46−57.

(141) http://icsd.ill.fr/icsd/. http://www.stn-international.de/stndatabases/databases/icsd.html (STN ICSD).

(142) Hellenbrandt, M. The Inorganic Crystal Structure Database (ICSD)−Present and Future. *Crystallogr. Rev.* **2004**, *10*, 17−22.

(143) Bergerhoff, G. Inorganic Three-Dimensional Strcuture Databases. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III; Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; Vol. 2, pp 1325−1337.

(144) http://www.rcsb.org/pdb/.

(145) Berman, H. M.; Westbrook, J.; Zardecki, C.; Bourne, P. E. The Protein Data Bank. *Protein Struct.* **2003**, 389−405. Berman, H. M; et al. The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *D58*, 899−907.

(146) http://thomsonderwent.com/products/patentresearch/mergedmarkush/.

(147) http://www.questel.orbit.com/.

(148) (a) Schinzer, D. Three New Reaction Data Bases. *Nachr. Chem. Tech. Lab.* **1993**, *41*, 826−828. (b) Griepke, G. Chemical Databases from Springer-Verlag. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 154−155. (c) http://www.infochem.de.

(149) Blower, P. E., Jr.; Myatt, G. J.; Petras, M. W. Exploring Functional Group Transformations on CASREACT. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 54−58.

(150) Cooke, F.; Kopelev, N.; Schofield, H.; Boyce, G.; Dunne, S. Approaches to Understanding the Searching Behavior of CrossFire Users. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1016−1027.

(151) Paris, C. G. Databases of Chemical Structures. In *Handbook of Chemoinformatics;* Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2, pp 523−555.

(152) *Nucleic Acids Res.* **1996**, *24*. Baxevanis, A. D. The Molecular Biology Database Collection: 2003 Update. *Nucleic Acids Res.* **2003**, *31*, 1−12.

(153) Discala, C.; Benigni, X.; Barillot, E.; Vaysseix, G. DBcat: A Catalog of 500 Biological Databases. *Nucleic Acids Res.* **2000**, *28*, 8−9.

(154) http://www.infobiogen.fr/services/dbcat.

(155) www.biosis.org.

(156) http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.

(157) http://www.ebi.ac.uk/embl/.

(158) Miyazaki, S.; Sugawara, H.; Gojobori, T.; Tateno, Y. DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Res.* **2003**, *31*, 13−16.

(159) von Homeyer, A.; Reitz, M. Databases in Biochemistry and Molecular Biology. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2, pp 756−793.

(160) http://www.emolecules.com/.

(161) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley: New York, 1986; 280 pp.

(162) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: New York, 1999; 286 pp.

(163) Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. The Chemical Abstracts Service Chemical Registry System. II. Augmented connectivity molecular formula. *J. Chem. Inf. Comput. Sci.* **1979**, *19* (2), 94−8.

(164) Garey, M.; Johnson, D. *Computers and Intractability; A Guide to the Theory of NP-Completeness*; W. H. Freeman & Co.: 1979; 340 pp.

(165) Barnard, J. M. Substructure Searching methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532−538.

(166) Sussenguth, E. H. A Graph-Theoretic Algorithm for Matching Chemical Structures. *J. Chem. Doc.* **1965**, *5*, 36−43.

(167) Figueras, J. Substructure Search by Set Reduction. *J. Chem. Doc.* **1972**, *12*, 237−244.

(168) Ray, L. C.; Kirsch, R. A. Finding Chemical Records by Digital Computers. *Science* **1957**, *126*, 814−819.

(169) Bayada, D. M.; Simpson, R. W.; Johnson, A. P.; Laurenco, C.An algorithm for the multiple common subgraph problem. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (6), 680−5.

(170) Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. Assoc. Comput. Mach.* **1976**, *23*, 31−42.

(171) von Scholley, A. A Relaxation Algorithm for Generic Chemical Structure Screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 235−241.

(172) Xu, J. GMA: A Generic Match Algorithm for Structural Homomorphism, Isomorphism, Maximal Common Substructure Match and Its Applications. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 25−34.

(173) Xu, J. Two-dimensional Structure and Substructure Searching. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2, pp 868−884.

(174) Chen, L. Substructure and maximal common substructure searching. *Comput. Med. Chem. Drug Discovery* **2004**, 483−513.

(175) Güner, O. F.; Henry, D. 3D Structure Searching. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; Vol. 5, pp 2988−3003.

(176) Benichou, P.; Klimczak, C.; Borne, P. Handling Genericity in Chemical Structures Using the Markush Darc Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 43−53.

(177) Fisanick, W. The Chemical Abstract's Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145−154.

(178) Ebe, T.; Sanderson, K. A.; Wilson, P. S. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 2. The MARPAT File. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 31−36.

(179) Lynch, M. F.; Holliday, J. D. The Sheffield Generic Structures Project-a Retrospective Review. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 930−936.

(180) Berks, A. H.; Barnard, J. M.; O'Hara, M. P. Markush Structure Searching in Patents, In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; Vol. 3, pp 1552−1559.

(181) Schmuff, N. R. A Comparison of the MARPAT and Markush DARC Software. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 53−59.

(182) Willet, P. Similarity Searching in Chemical Structure Databases. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2, pp 904−915.

BASIC OVERVIEW OF CHEMOINFORMATICS

J. Chem. Inf. Model., Vol. 46, No. 6, 2006 **2277**

(183) Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. *J. Med. Chem.* **2005**, *48* (13), 4183−4199.

(184) Raymond, J. W.; Willet, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002**, *16* (7), 521−533.

(185) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000; 690 pp.

(186) Gasteiger, J. A Hierarchy of Structure Representations. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 3, pp 1034−1061.

(187) Lanig, H. Molecular Mechanics. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 3, pp 920−946.

(188) Mitchell, T. *Machine Learning*; McGraw-Hill: New York, 1997; 414 pp.

(189) Thuraisingham, B. *Data Mining: Technologies, Techniques, Tools, and Trends*; CRC Press: 1999; 270 pp.

(190) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999; 380 pp.

(191) Goldberg, D. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley Longman: Reading, 1989; 432 pp.

(192) von Homeyer, A. Evolutionary Algorithm and their Applications in Chemistry. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 3, pp 1239−1280.

(193) Jackson, P. *Introduction to Expert Systems*; Addison-Wesley Longman: Harlow, 1999; 542 pp.

(194) Hansch, C.; Fujita, T. ρ-σ-π Analysis; Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(195) Hansch, C.; Leo, A. Exploring QSAR. *Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(196) Free, S. M., Jr.; Wilson, J. W. A Mathematical contribution to Structure−Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(197) Kubinyi, H. QSAR: *Hansch Analysis and Related Approaches*; VCH: Weinheim, 1994; 240 pp.

(198) Jurs, P. C. Quantitative Structure−Property Relationships. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, pp 1314−1335.

(199) *QSAR & Combinatorial Science*; Wiley-VCH: Weinheim.

(200) Taylor & Francis: *SAR QSAR Environ. Res.*

(201) Corey, E. J.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969**, *166* (3902), 178−92.

(202) Gasteiger, J.; Ihlenfeldt, W. D.; Roese, P.; Wanke, R. Computer−assisted reaction prediction and synthesis design. *Anal. Chim. Acta* **1990**, *235* (1), 65−75.

(203) Ott, M. A. Cheminformatics and organic chemistry. Computer-assisted synthetic analysis. In *Cheminformatics Developments*; Noordik, J. H., Ed.; IOS Press: Amsterdam, 2004; pp 83−109.

(204) Pförtner, M.; Sitzmann, M. Computer-Assisted Synthesis Design by WODCA (CASD). In *Handbook of Chemoinformatics;* Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, pp 1457−1507.

(205) Corey, E. J.; Cramer, R. D., III; Howe, W. J. Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates. *J. Am. Chem. Soc.* **1972**, *94*, 440−459.

(206) Munk, M. E.; Sodano, C. S.; McLean, R. L.; Haskell, T. D. Actinobolin. I. Structure of Actinobolamine. *J. Am. Chem. Soc.* **1967**, *89*, 4158−4165.

(207) Shelly, C. A.; Munk, M. A.; Roman, R. V. A Unique Computer Representation for Molecular Structures. *Anal. Chim. Acta* **1978**, *103* (3), 245−251.

(208) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; Wiley: New York, 1986; 536 pp.

(209) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J.; *Application of Artificial Intelligence for Organic Chemistry*; McGraw-Hill: New York, 1980; 256 pp.

(210) Munk, M. E. Computer-Based Structure Determination: Then and Now. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 997−1009.

(211) Steinbeck, C. Recent developments in automated structure elucidation of natural products. *Nat. Prod. Rep.* **2004**, *21* (4), 512−518.

(212) Selzer, P. Correlations between Chemical Structures and Infrared Spectra. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 3, pp 1349−1367.

(213) Steinbeck, C. Correlations between Chemical Structures and NMR. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 3, pp 1368−1377.

(214) Oprea, T. I. Chemoinformatics in Drug Discovery. In *Methods Princ. Med. Chem.*; Wiley-VCH: Weinheim, 2005; Vol. 23, 493 pp.

(215) Wermuth, C. G. *The Practice of Medicinal Chemistry*; Academic Press: London, 1996; 968 pp.

(216) Wermuth, C. G. Possible alternatives to high−throughput screening. *Drug Discovery Dev.* **2006**, *1*, 213−232.

(217) Warr, W. A. High-Throughput Chemistry. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, pp 1604−1639.

(218) Farnum, M. A.; DesJarlais, R. L.; Agrafiotis, D. K. Molecular Diversity. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, pp 1640−1686.

(219) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Recent advances in comparative molecular field analysis (CoMFA). *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(220) Kubinyi, H., Folkers, G., Martin, Y. C., Eds. *3D QSAR in Drug Design. Ligand-Protein Interactions and Molecular Similarity*; Kluwer/ESCOM: Dordrecht, 1998. Also published as *Perspect. Drug Discovery Des. 1998*, *9*−*11*.

(221) Podlogar, B. L.; Ferguson, D. M. Qsar and CoMFA: a perspective on the practical application to drug discovery. *Drug Des. Discovery* **2000**, *17* (1), 4−12.

(222) Kubinyi, H. Comparative Molecular Field Analysis (CoMFA). In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, pp 1555−1575.

(223) Milne, G. W. A.; Nicklaus, M. C.; Wang, S. Pharmacophores in drug design and discovery. *SAR QSAR Environ. Res.* **1998**, *9*, 23−38.

(224) Langer, T.; Wolber, G. Pharmacophore definition and 3D searches. *Drug Discovery Today: Technol.* **2004**, *1* (3), 203−207.

(225) Guener, O. F. The impact of pharmacophore modeling in drug design. *Drugs* **2005**, *8* (7), 567−572.

(226) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. Novel approach to pharmacophore modeling and 3D database searching. *Chem. Biol. Drug Des.* **2006**, *67* (5), 370−372.

(227) Nicklaus, M. C. Pharmacophore and Drug Discovery. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, pp 1687−1711.

(228) Kubinyi, H. Structure-based drug design. *Chim. Oggi* **1998**, *16* (10), 17−22.

(229) Gillet, V. J. De novo molecular design. In *Methods and Principles in Medicinal Chemistry*; 2000; Vol. 8 (Evolutionary Algorithms in Molecular Design), pp 49−69.

(230) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4* (8), 649−663.

(231) Johnson, A. P. *De-novo* Design Systems. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, pp 1712−1731.

(232) Gohlke H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.

(233) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409−443.

(234) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 151−166.

(235) Rester, U.; Dock around the clock−current status of small molecule docking and scoring. *QSAR Comb. Sci.* **2006**, *25* (7), 605−615.

(236) Sotriffer, C.; Stahl, M.; Klebe, G. The Docking Problem. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, pp 1732−1768.

CI600234Z