CHAPTER NINE

Chemical Information Computing Systems in Drug Discovery

DOUGLAS R. HENRY
MDL Information Systems, Inc.
San Leandro, California

Contents

1 Introduction, 358
1.1 Motivation for Chemical Information Management, 358
1.2 Literature, References, Societies, and Research Groups, 359
1.3 Brief History of Chemical Data Management, 360
1.3.1 Pre-1980—Flat File Storage of Chemical Structures, 360
1.3.2 The 1980s––Flat Database Storage, 362
1.3.3 The 1990s—Relational Data Storage, 363
1.3.4 The 2000s, 363
2 Chemical Representation, 363
2.1 Types of Chemical Entities, 363
2.1.1 Sequences, 363
2.1.2 2D Structures, 364
2.1.3 Reactions, 366
2.1.4 3D Models, 366
2.1.5 Mixtures, 367
2.1.6 Generic Structures, 368
2.1.7 Substances, 368
2.1.8 Search Queries, 368
2.2 Types of Chemical Representation, 368
2.2.1 Linear Notation, 368
2.2.2 Tabular Storage, 369
2.2.3 Graphical Representation, 371
2.2.4 Markup Languages, 371
2.3 Chemical Structure File Conversion, 372
2.4 Representing Nonstructural Chemical Data, 373
3 Storing and Searching Chemical Structures and Reactions, 373
3.1 Storing Chemical Information in Databases, 373

Burger's Medicinal Chemistry and Drug Discovery
Sixth Edition, Volume 1: Drug Discovery
Edited by Donald J. Abraham
ISBN 0-471-27090-3 © 2003 John Wiley & Sons, Inc.

3.2 Registering Chemical Information, **377**
3.2.1 Extract the Data, **377**
3.2.2 Cleaning and Transforming the Data, **378**
3.2.3 Loading the Data, **378**
3.3 Searching Chemical Structures and Reactions, **379**
3.3.1 Exact Match Searching, **379**
3.3.2 Substructure Searching, **381**
3.3.3 Similarity Searching, **382**
3.3.4 Reaction Searching, **383**
3.3.5 Searching Other Data, **383**
3.4 Chemical Information Management Systems and Databases, **384**
3.5 Commercial Database Systems for Drug-Sized Molecules, **384**
3.6 Sequence and 3D Structure Databases, **387**
3.7 In-House Proprietary and Academic Database Systems, **387**
4 Chemical Property Estimation Systems, **388**
4.1 Topological Descriptors, **388**
4.2 Physicochemical Descriptors, **389**
4.3 Absorption, Distribution, Metabolism, and Excretion Properties, **389**
4.4 Property Calculations Online, **390**
5 Data Warehouses and Data Marts, **390**
5.1 Data Warehouses of Chemical Information, **390**
5.2 Data Marts of Chemical Information, **391**
6 Future Prospects, **393**
7 Glossary of Terms, **397**
8 Acknowledgments, **412**

1 INTRODUCTION

The term drug discovery once encompassed only those activities that were traditionally practiced by synthetic chemists—the design, synthesis, analysis, and testing of new chemical entities. Until the 1980s, most drug discovery was conducted in a serial fashion. Thus, a chemist working on a given project would design a series of structures, then synthesize them one after another, in milligram quantities (large by today's standards), and finally send batches of the compounds for analysis and assay. Based on the assay results, the chemist would design new or modified sets of structures and repeat the cycle until a marketable entity was obtained. This serial, iterative procedure was adequate in a time when a few major drug companies were doing drug design, and the number of therapeutic targets was relatively small. One consequence of this approach was a much higher number of "me-too" drugs on the market than we see today, likely because of the intensive time and resource that was devoted to each new chemical entity.

1.1 Motivation for Chemical Information Management

The serial approach to drug discovery is very costly in time and resource. Figure 9.1 shows an idealized view of the drug discovery "funnel" in which a (very productive) hypothetical chemist could produce 10–20 structures a week. In the 1970s, it was estimated that 1 in 7000 compounds synthesized and tested would eventually reach the market. That number has risen over the years to about 1 in 10,000—a figure that holds true to this day, despite advances in combinatorial and high-throughput chemistry, molecular modeling, structure-based drug design, diversity analysis, and quantitative structure-activity relationships (QSAR). In real dollars, it almost certainly costs as much or more to bring a new drug to market than it did in the 1970s. A commonly quoted figure of $500 million per marketable drug has been questioned by a Ralph Nader watchdog group, but the figure is certainly in the hundreds of millions of dollars (1).To balance the many computational advances that have been made in the past 30 years are factors of increased competition in the field, many more therapeutic targets, increased regulation, and very importantly, the flood of information flowing from high-throughput methods. The advent of high-throughput combinatorial chemistry has increased the number of structures a chemist can generate by 100- to 1000-fold, with a corresponding increase in the amount of data that must be gathered, stored, and processed.

To deal with the flood of information—chemical, biological, and clinical—it became essential over the years to develop chemical information computing systems (i.e., chemical and reaction database systems) from which the chemist and biologist could obtain up-to-

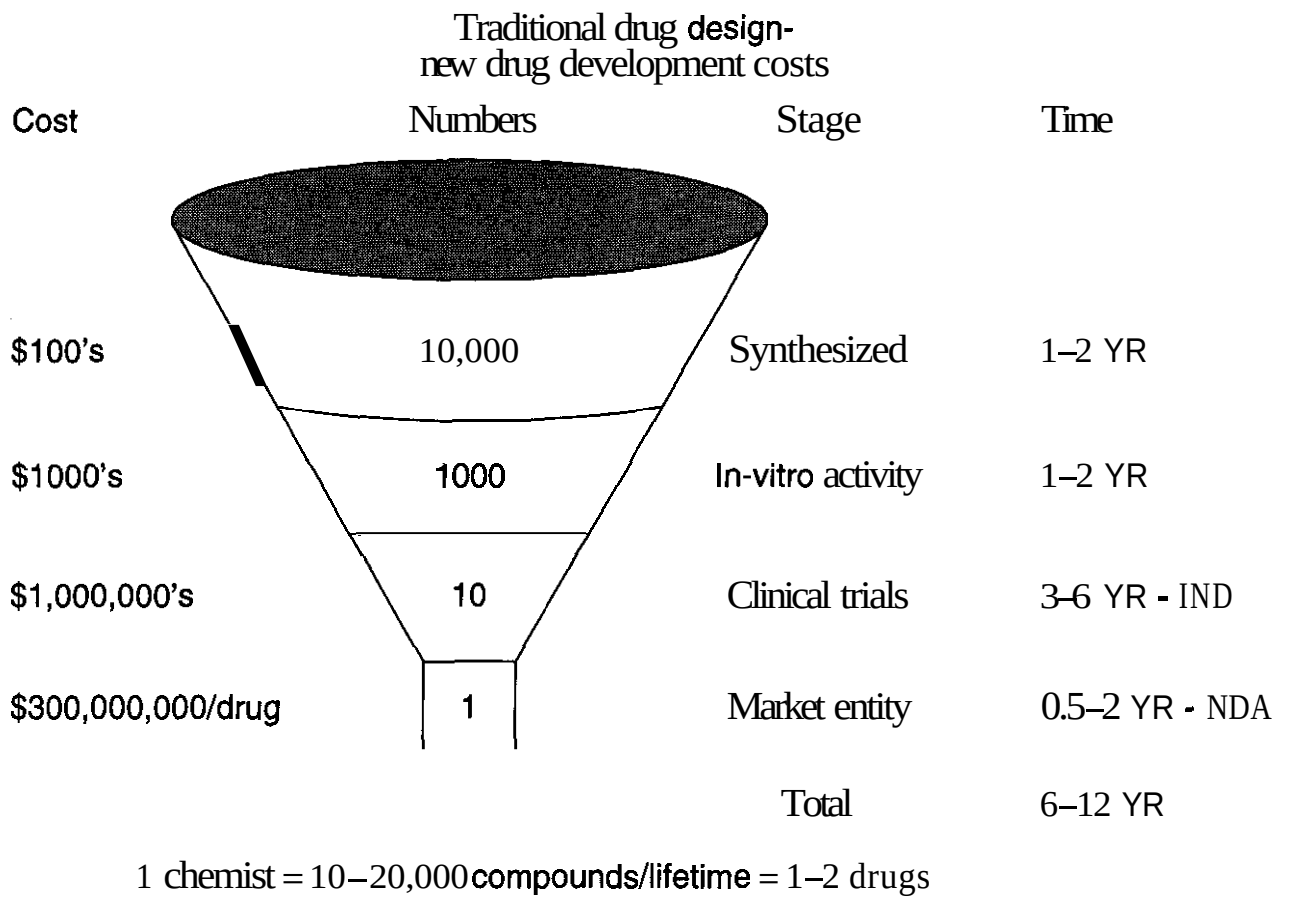


Figure 9.1. Traditional "serial" drug design costs. The drug discovery "funnel" typically shows about a 10-fold reduction at each stage in the process. **A** chemist who could produce 10-20 structures per week would be lucky to discover a single marketable drug in a 20- to 30-year career.

date information about commercially available and in-house structures, reactions, and data. This chapter briefly describes the history of these systems, the current state of chemical information management as it applies to drug discovery, and a look at future developments in the field. The coverage is primarily aimed at corporate applications of chemical information management, as practiced in the pharmaceutical industry. The expanding use of microcomputers running Microsoft Windows or Linux operating systems means that many of the programs and database systems now used in industry can also be installed and applied in academic settings. Much of the innovation in chemical information management comes from academia, whereas most of the application has been seen in industry. This review is limited to the management and storage of chemical structure information in databases. Other chapters deal with the generation of this information (molecular modeling, property calculation) and with the use of the information in drug discovery (library design, docking and structure-based drug design, and QSAR). By analogy with another rapidly expanding field, bioinformatics, the term chemoinformatics (or chemoinformatics) has recently become common to describe the acquisition, management, and use of chemical information.

1.2 Literature, References, Societies, and Research Groups

The literature of chemistry is vast, and chemical information management occupies a small corner of this domain. The chemical information literature overlaps that of computer science, database management, molecular modeling, QSAR, and even mathematics. The primary journals that publish chemical information articles are the American Chemical Society's Journal of Chemical Information and Computer Sciences, Kluwer's Journal of *Computer-Aided Molecular Design*, and Elsevier's Journal of Molecular Graphics and Modeling. Less frequently, chemical information articles appear in Wiley's Journal of Computational Chemistry, Quantitative Structure-Activity Relationships, and Journal of *Chemometrics*, the ACS Journal of Medicinal Chemistry and the *Journal* of Organic Chemistry, and Elsevier's Analytica *Chimica Acta,* Computers and Chemistry, and Chemometrics and Intelligent

Laboratory Systems. Other journals with articles on chemical information include the University of Bayreuth's Communications in Mathematical Chemistry (MATCH), Elsevier's Drug Discovery Today, ACS's Modern Drug Discovery, and a handful of newer periodicals (2).

The history of chemical information management has recently been catalogued online by the Chemical Heritage Foundation (3). The American Chemical Society has a Division of Chemical Information (CINF), and divisional symposia are held at national meetings of the ACS, often in conjunction with other divisions including Medicinal Chemistry, Computers in Chemistry, and Pesticide Chemistry. The Skolnik Award is given annually by the ACS Division of Chemical Information for "achievement in the areas of computerized information systems, chemical information, chemical indexing and notation systems, nomenclature, structure-activity relationships, and numerical data analysis and correlation." Herman Skolnik, who died in 1994, was the first recipient. He founded the Journal of Chemical Documentation, which became the Journal of Chemical Information and Computer Sciences, and he made many contributions to the field (4). Besides the ACS, other national and international meetings on chemical information include the Noordwijkerhout Conference on Chemical Structures (5), the Quantitative Structure-Activity Relationship Gordon Conference (6), and the International Conference on Chemical Information (7).

Except for journal articles and some conference proceedings, very recent general books on chemical information management are rather few in number. This is caused in part by the rapid changes in a field so closely tied to computer hardware and software development. Another reason for the paucity of texts is that most chemical information management systems are commercially developed and marketed, not widely used by universities, and in many cases, they use trademarked or even patented technology. Some texts of note in the last decade include several by Collier (8), Martin and Willett (9), and Warr and Suhr (10), one by Wiggins and Emry (11), and ones by Maizell (12) and Ash et al. (13), and a book on chemical searching by Ridley (14).

Most chemical information research and development is conducted by commercial software vendors and in-house at pharmaceutical firms. A small number of academic research groups study chemical information. The Computational Information Systems group at the University of Sheffield, under Peter Willett, has been very active in studying database searching (15). The Computer-Chemie-Centrum at the University of Erlangen under Johann Gasteiger focuses on organic structure representation and reaction classification (16). Numerous other academic groups are active in QSAR and modeling research, described in other chapters in this series.

In addition to the academic groups already mentioned, a number of online resources deal with chemical information management. Examples include the comprehensive CHEMINFO site at Indiana University (17), Cambridge Health Institute's Cheminformatics Glossary (18), the Chemical Structure Association (19), the Computational Chemistry List (CCL) (20), the Molecular Graphics and Modeling Society (21), the Open Molecule Foundation (22), the QSAR and Modeling Society (23), the Royal Society of Chemistry Chemical Information Group (24), and the UK QSAR and Cheminformatics Group (25).

1.3 Brief History of Chemical Data Management

The history of chemical information management parallels the history of computers. It can be roughly viewed in terms of decades of development (Fig. 9.2).

1.3.1 Pre-1980—Flat File Storage of Chemical Structures. Computers consisted of mainframe machines (e.g., IBM 3090) and small minicomputers (Digital, Prime). Users connected through low speed serial connections, using "dumb" terminals (no graphics capability) or monochrome vector graphics terminals such as Tektronix and Imlac. Chemical structures were mainly stored as either (1) individual structure files, indexed by name, and handled one or a few structures at a time or (2) in a flat-file database accessed by record number (26). A typical corporate database contained up to a few tens of thousands of structures.

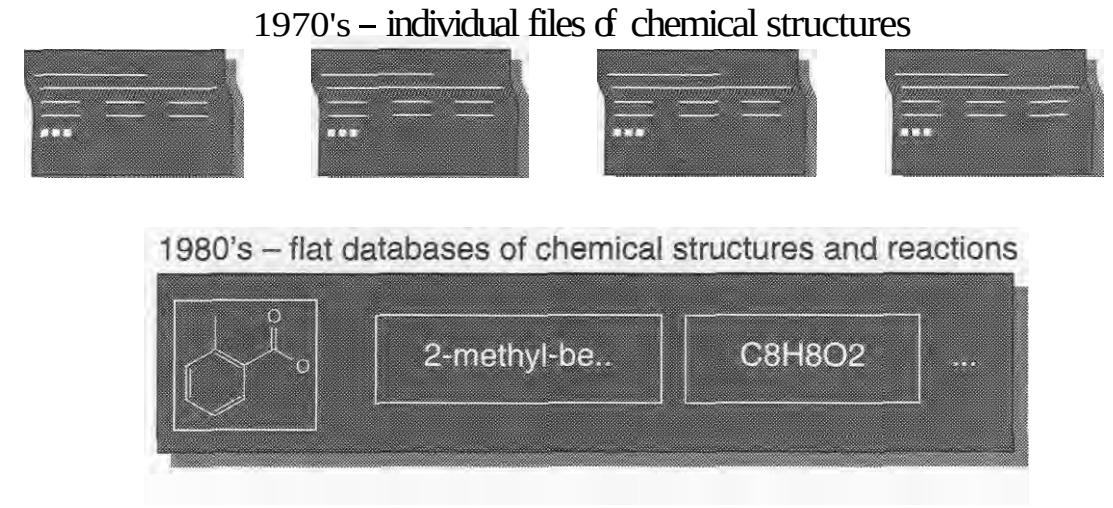


Figure 9.2. Evolution of chemical information storage. The storage of chemical information has typically lagged the development of database management systems, but it is catching up. In the 1970s, structures were stored in individual molecule files or large concatenated files. In the 1980s, proprietary databases of structures and reactions appeared, in which a single record contained all the information for a given structure. In the 1990s, this information was distributed into tables in a relational database. In the 2000s, we see the application of the concepts of data warehousing and data marts that consolidate information from a variety of sources for transactional and/or analytical purposes

In-house chemical information management systems began to emerge at some of the larger chemical and pharmaceutical firms. These included CONTRAST and SOCRATES at Pfizer, SYNLIB at SmithKline, COUSIN at Upjohn, MSDRL/CSIS at Merck, and CROSSBOW at ICI (27). The Chemical Abstracts database was made available online in 1967 (28). In 1980 this became CAS ONLINE. A comprehensive study of the user acceptance of CAS ONLINE was published in 1988 (29). The first commercial chemical structure database systems appeared in the late 1970s. These offered an in-house solution using a mainframe chemical structure management system with a graphical interface, which could be accessed by interactive graphics terminals. A standard program in widespread use was the MACCS

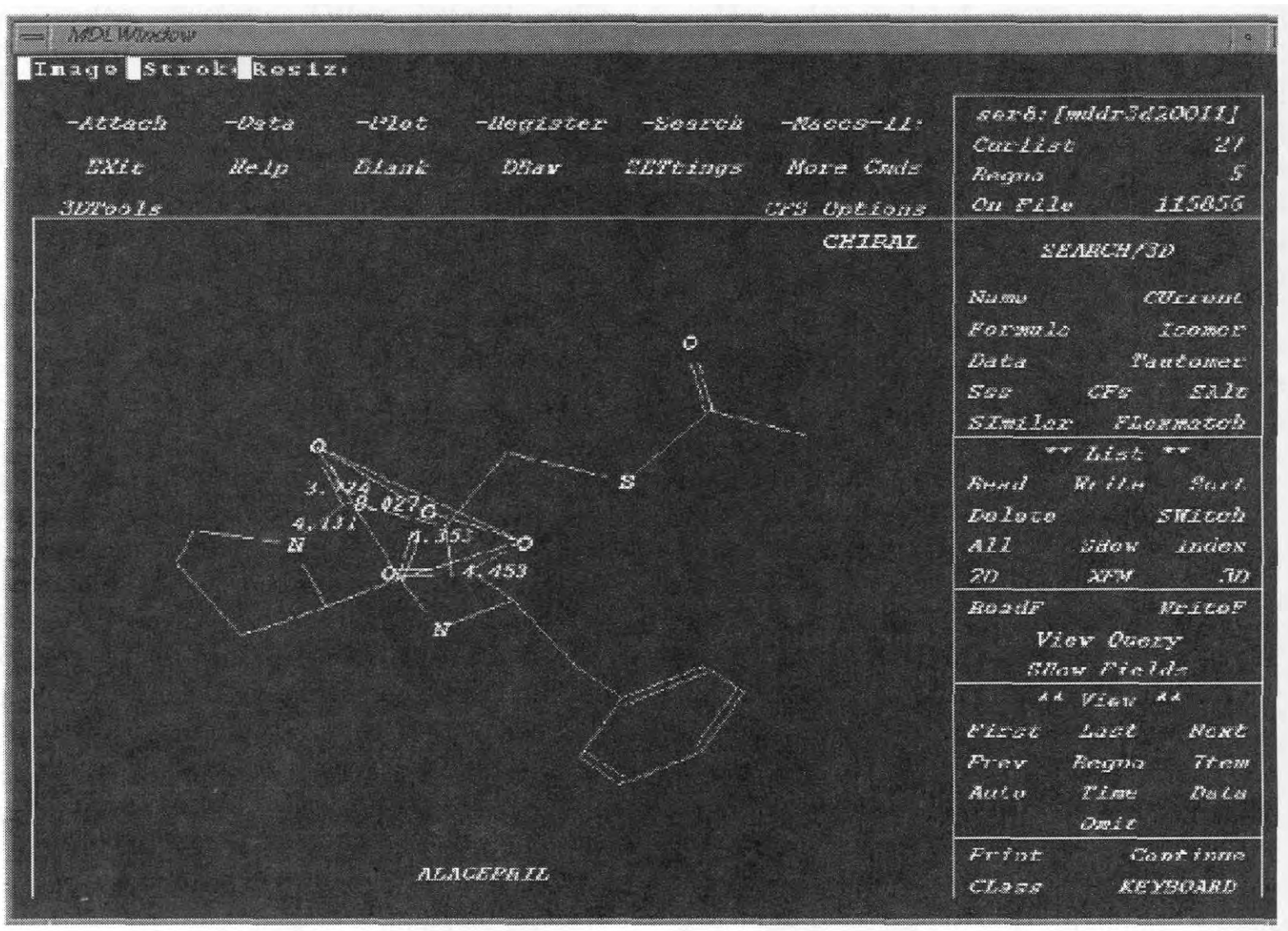


Figure 9.3. MACCS—the Molecular ACCess System—an early structure indexing system. This program originally used fixed menus for searching, registration, and reporting. Later versions allowed users to customize the menus. The figure shows the result of a 3D pharmacophore search for ACE inhibitors. Out of a database of 115,000 structures, 21 fit the 2D and 3D requirements of the search query. The user could typically browse the "hits" from the search, save the list of structures to a list file, and output the structures to a structure-data file (SDFile). The MACCS database was a proprietary flat database system in which data of a given type, say, formula, was stored in a given file, indexed by the compound ID number.

program (Fig. 9.3). Structures could be drawn, registered, searched, and output to files. The systems were only slightly customizable, and the graphics terminals, which used vector displays, were large and expensive.

1.3.2 The 1980s—Flat Database Storage. This was the era of minicomputers (Prime, Vax) and a period of immense growth for chemical information, molecular modeling, and QSAR. In industry, chemical structure databases consisted mainly of custom-designed "flat" databases (where each record in a given table refers to a given structure in the database—much like in a spreadsheet). Client-server architectures appeared, and personal computers replaced graphics terminals and workstations. Highly successful PC-based "personal" chemical information systems appeared, which included chemical structure drawing and text processing programs (e.g., ChemDraw, ChemText) and personal chemical databases (e.g., ChemBase) (30). Customizable mainframe systems appeared (31), as did reaction indexing and searching systems (32). Additional commercial chemical information vendors appeared including Daylight Chemical Information Systems, Chemical Design Ltd., DARC-Questel, and Cambridge Scientific Corporation. The Beilstein System came online in 1988 (33). In-house and commercial database sizes were typically 100-200K structures in size. The rapid and accurate conversion of two-dimensional (2D) structures to

three-dimensional (3D) models became possible using the program CONCORD, introduced by Pearlman in 1987 (34). This enabled the introduction of 3D structural databases with the ability to generate, store, and search 3D molecular models on a large scale. These 3D database systems included ALADDIN by Daylight Chemical Information Systems, UNITY3D by Tripos, CHEMDBS3D by Chemical Design Ltd., and MACCS3D by MDL (35).

1.3.3 The 1990s—Relational Data Storage. This period saw the decline of single-computer mainframe chemical management programs and the rise of server-based systems and distributed computing. By far, the most significant influences on chemical information management were the Internet, the introduction of relational database technology, and the shift to high-throughput combinatorial chemistry. In a relational database, information that formerly was kept in a single large table is stored in numerous smaller tables, indexed by "keys." This is a much more flexible architecture, and combining different fields from several tables into a "view" of the data gives the user the impression of a single large table, as before. At the end of the decade, chemical and pharmaceutical firms could obtain chemical structure, reaction, and 3D model databases from a variety of vendors. These databases were even somewhat integrated with molecular modeling, quantum mechanics, and docking programs, and to literature, spectra, and biological databases. The largest database of known chemical structures, the Chemical Abstracts Registry, grew to about 20 million structures, whereas a typical corporate inventory increased to between 100,000 and 1,000,000 structures. A database of billions of virtual chemical structures was constructed and made available for drug-design purposes by Tripos, Inc. (36).

1.3.4 The 2000s. Like the customization and distributed computing of the 1980s that followed the introduction of mini-mainframe systems, the 2000s are witnessing the customization and further distribution of relational and integrated database systems. Chemical structure-specific and reaction-specific search types can be integrated into relational databases, to take maximal advantage of the scale and performance of these systems. We see the increasing use of web-based clients, also known as "thin" clients, because they need little software other than a web browser. Former single databases are turning into distributed and replicated database systems, and we see increasing use of data marts and data warehouses, more fully integrated structure, reaction, data, and citation searching, and increasingly "intelligent" database systems.

2 CHEMICAL REPRESENTATION

Chemical structures and reactions can be represented in many ways. At the most fundamental level, the parameters of the time-dependent Schrödinger equation—the atomic and molecular orbitals—do a more or less complete job of characterizing a chemical compound. Storing and representing structures as mathematical wave functions is obviously not suitable for thousands or millions of structures; nor is such a representation useful for drug discovery, except perhaps to a molecular modeler. Synthetic chemists still function in a mostly 2D chemical structure space. Intuition, training, and experience allow a chemist to extrapolate from a flat representation with a few stereochemical hints—dashed and wedged bonds or Z/E double bonds—to a higher-dimensional mental representation of a structure. Chemical representation systems are a compromise of several factors, including the needs of the chemist, the storage and performance characteristics of the chemical database system, and the ultimate 3D reality of chemical structures.

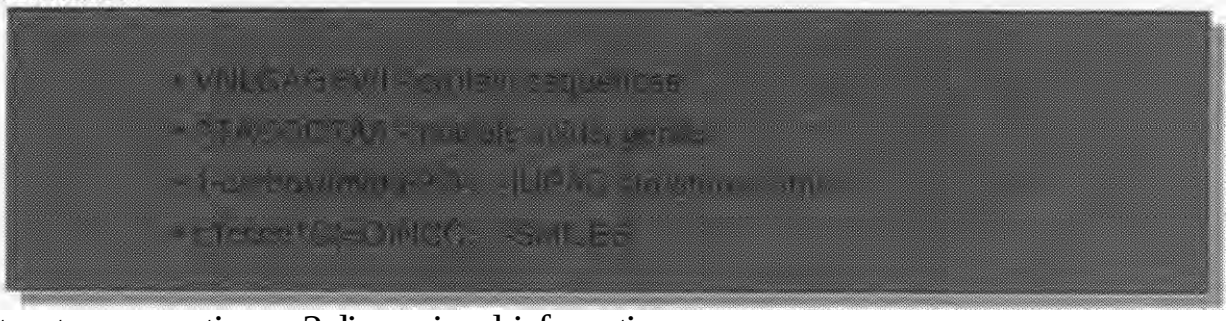
2.1 Types of Chemical Entities

There are several ways to look at chemical representation. One approach is to classify according to the type of chemical data that is stored. The most basic types of chemical structure data are shown in Fig. 9.4, including the following.

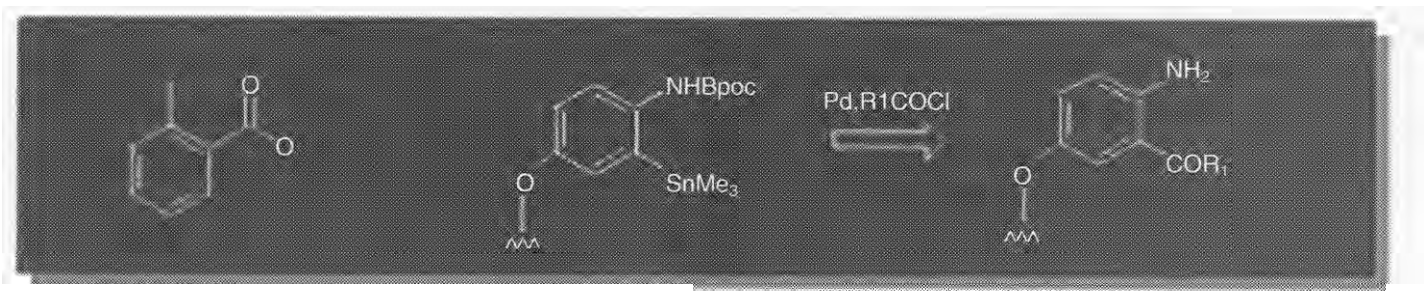
2.1.1 Sequences. For linear chemical systems, such as DNA, RNA, and proteins, the sequence of subunits (nucleotide bases or amino acids) provides most of the information

Types of chemical data

• Sequences, names, linear notations - 1-dimensional information



• Structures, reactions - 2-dimensional information



• 3D models - 3-dimensional information

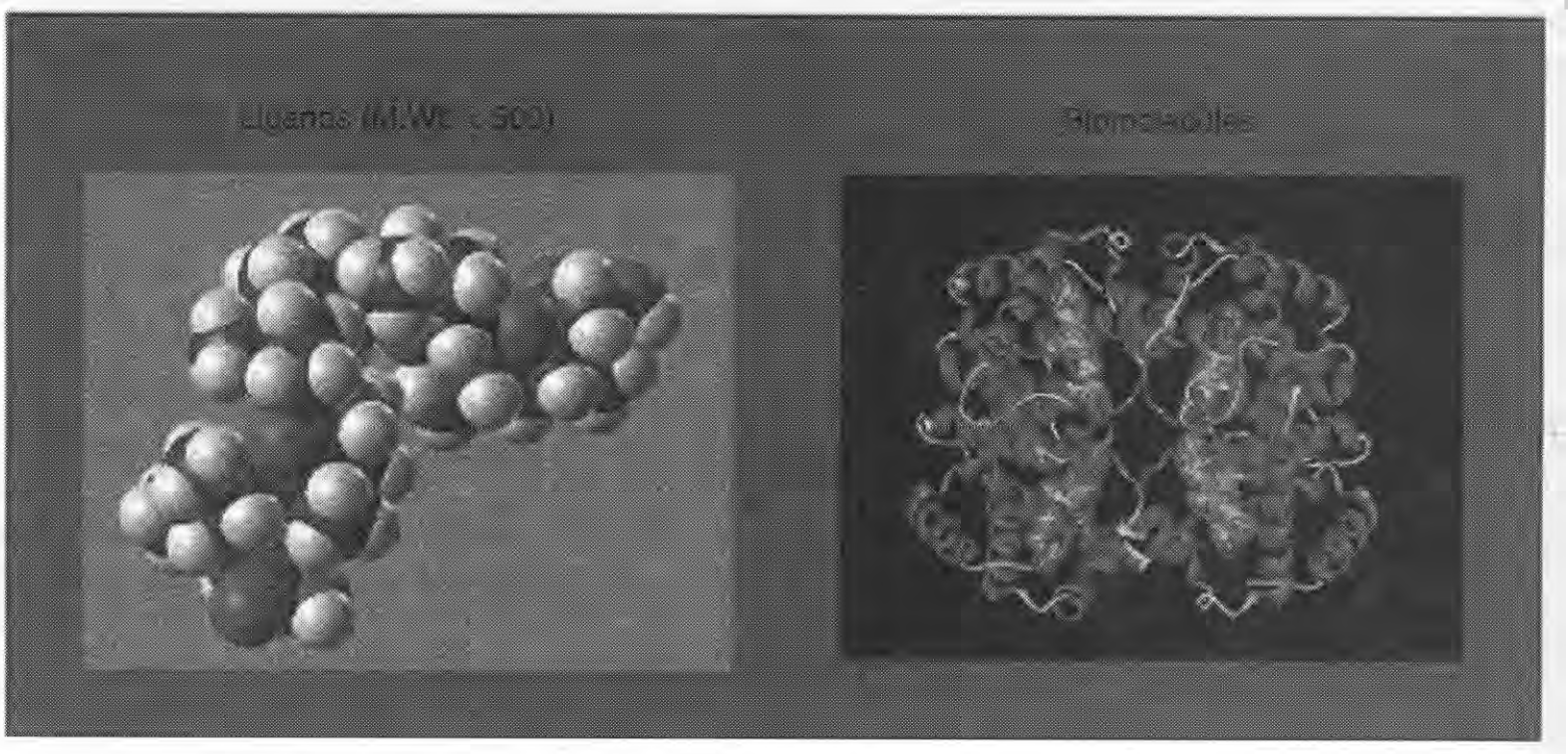


Figure 9.4. Basic types of 2D chemical structure data. The amount of information and the complexity of searching increases with the dimensionality of the data.

about the structure. The deciphering of the human genome and the exploding interest in bioinformatics as a means of identifying new drug targets means there will be an increasing growth in the use of sequence data. The use of a sequence representation depends on a natural "vocabulary" of fixed building blocks. This vocabulary consists of nucleotides in the case of nucleic acids and consists of the amino acids in the case of proteins. If any of the building blocks are unique, or even if the bonding attachment between building blocks differs, a simple sequence notation is not possible or it becomes more complex.

2.1.2 2D Structures. When the building blocks are unique or when dealing with the large variety of ordinary chemical structures, a 2D representation is used. In mathematical terms, this is a "graph" of the structure, which consists of a set of "nodes" (atoms) connected by "edges" (bonds). The important atom infor-

mation includes atom type (symbol or atomic number), its 2D coordinates, formal charge, valence state, atom stereochemistry, and isotope information. Note that atom stereochemistry can be local (i.e., relative) or it may follow Cahn-Ingold-Prelog (CIP) conventions. Local atom stereochemistry gives the clockwise or counter-clockwise direction of the attachment of neighboring atoms when viewed from some reference attached atom—often a hydrogen atom or the lowest atomic numbered atom (37). The order of atoms in the rotation usually depends on atomic number. CIP stereochemistry is the familiar "*R,S*" nomenclature that relates the stereochemistry of the given atom to the entire structure (38). CIP stereochemistry requires analyzing the entire structure to determine the stereochemistry values. It can occasionally be ambiguous, and if any part of the structure changes, the CIP stereochemistry on distant atoms in the structure may switch. For these reasons, it is common in chemical databases to store local atom stereochemistry, but to perceive CIP stereochemistry "on the fly."

A particular problem with relative stereochemistry is that a given combination of "up" and "down" bonds on a structure implies a mixture of at least two stereoisomers. If all the centers are specified, the structure represents at least the two enantiomers. If some of the stereo centers are not designated, the number of isomers the structure represents is 2^n, where n is the number of undesignated centers. Some database vendors (e.g., MDL) allow a "chiral" designation on the molecule, which indicates that the structure represents only a single stereoisomer, but does not specify which one. One approach to dealing with these problems, which is being adopted in MDL programs, is to allow three kinds of stereo designation at a given tetrahedral center:

1. Absolute—an atom is given a known absolute stereochemistry. If all the stereo centers are so designated, this represents a single stereoisomer of the structure, as drawn.
2. Relative as a single stereoisomer—an up or down bond represents the current relative configuration, with respect to some collection of other chiral centers in the structure. The structure represents a single stereoisomer among the possible ones. More than one collection of stereo centers may be present in the structure.
3. Relative as a mixture of stereoisomers—an up or down bond represents the current relative configuration, with respect to some collection of other chiral centers in the structure. Now, however, the structure represents a mixture of the possible stereoisomers, considering combinations of the stereo collections that are present.

Examples of these alternatives are shown in Fig. 9.5, which shows the present and the newer stereochemistry options, using a steroid structure as an example.

The bond information usually includes the bonding atoms, the bond type, and bond stereochemistry. Bond types include the common single, double, triple, and aromatic types. They may also include types that are unique to the type of structure, including dative, ionic, hydrogen bonds, etc. The bond stereochemistry for double bonds is usually Z (Zusammen-together), E (Entgegen-opposite), or either (indicating an unknown stereochemistry). For single bonds attached to a chiral or prochiral center it is typically "up" (wedge or thick bond), "down" (dashed or dotted bond), or "either" (often a wiggly bond). Some systems allow the representation of extended stereochemistry, as with the terminal groups of allene systems, which can show a type of tetrahedral stereochemistry if you collapse the allene system to a point. The bonding information—which atoms are attached to which other atoms and the bond types—is collected in the "connection table" of the structure. Table 9.1 shows a simple atom connection table for camphor. The diagonal elements of the table describe the type of atom at a given position in the structure. The off-diagonal elements describe the bonding of that atom with other atoms in the structure. Some information about a structure can be derived implicitly from the connection table. This includes the rings that are present, and the hydrogen atoms that could be attached. When a structure can be represented by more than one isomer, it is common to either (*1*) store multiple

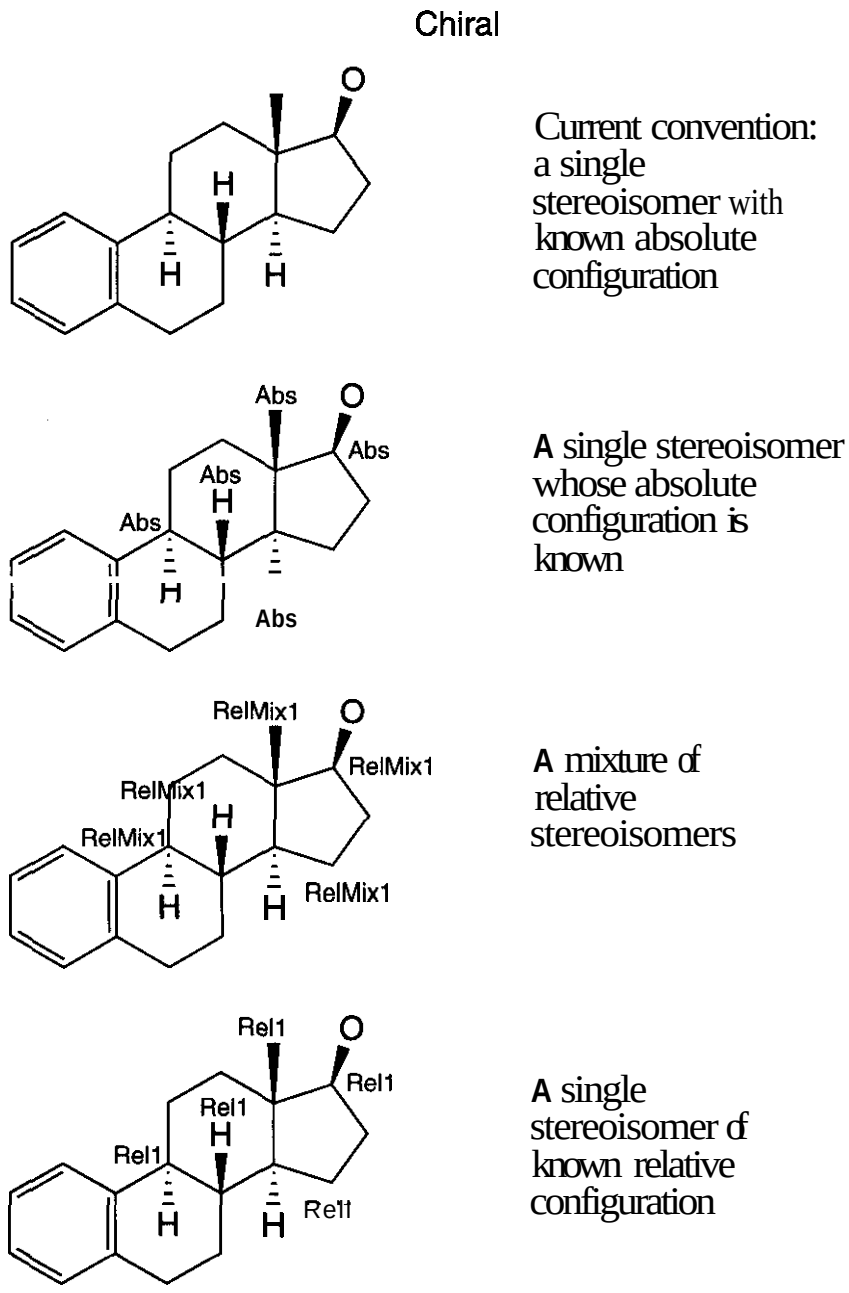


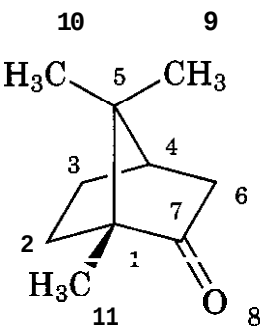
Figure 9.5. Defining absolute, relative collection, and relative single configuration stereochemistry. The older convention depends on a "chiral" flag on the molecule to specify whether a given structure represents one or several stereoisomers. In the newer convention, collections of stereo centers can be defined, and they can be designated absolute, relative-part-of-a-mixture, or relative-single-configuration.

isomers in the database, or (2)run a structure search using a search query that will hit the desired isomers. This is true for stereoisomers, enantiomers, and tautomers. Because the connection table is often symmetric, it is possible to store only, say, the upper diagonal part of it in the database.

2.1.3 **Reactions.** Chemical reactions extend the structure representation by adding information about what role the structure plays in the reaction (reactant, catalyst, solvent, product, etc.). Reaction representation may also include information about what bonds are made or broken during the reaction, and which atoms are involved in reacting centers. It is also common to use a hierarchical organization for reaction information (reaction > variation > reactants, catalysts, solvents, products, etc.).

2.1.4 **3D Models.** These extend the structure representation by adding one or more sets of 3D atomic coordinates for the various conformations that the molecule can adopt. 3D model representation may also include additional atom or bond information such as partial charge or partial bond order. It is common to generate approximate 3D models from 2D structures using fast abbreviated molecular modeling and fragment joining methods such as CONCORD, CORINA, and CONVERTER (39). These programs combine molecular mechanics with rules and heuristics to generate reasonable 3D structures in a fraction of the time required by molecular mechanics or quantum mechanics modeling. Typically, hundreds of structures can be processed per second. Although the resulting models are not the lowest energy models possible, they are quite suitable for 3D pharmacophore searching,

Table 9.1 Connection Table for D-Camphor



Atom	1	2	3	4	5	6	7	8	9	10	11
1	C	1			1		1				1
2	1	C	1								
3		1	C	1							
4			1	C	1	1					
5	1			1	C				1	1	
6				1		C	1				
7	1					1	C	2			
8							2	O			
9					1				C		
10					1					C	
11	1										C

and they serve as a good starting point for further optimization.

Recently, with the use of combinatorial and high-throughput chemistry, more general types of structure representation, so-called chemical libraries, have become common (Fig. 9.6). These are typically used to represent mixtures and generic structures.

2.1.5 **Mixtures.** Mixtures are useful to represent isomers, formulations, and the products of reactions. Their representation usually

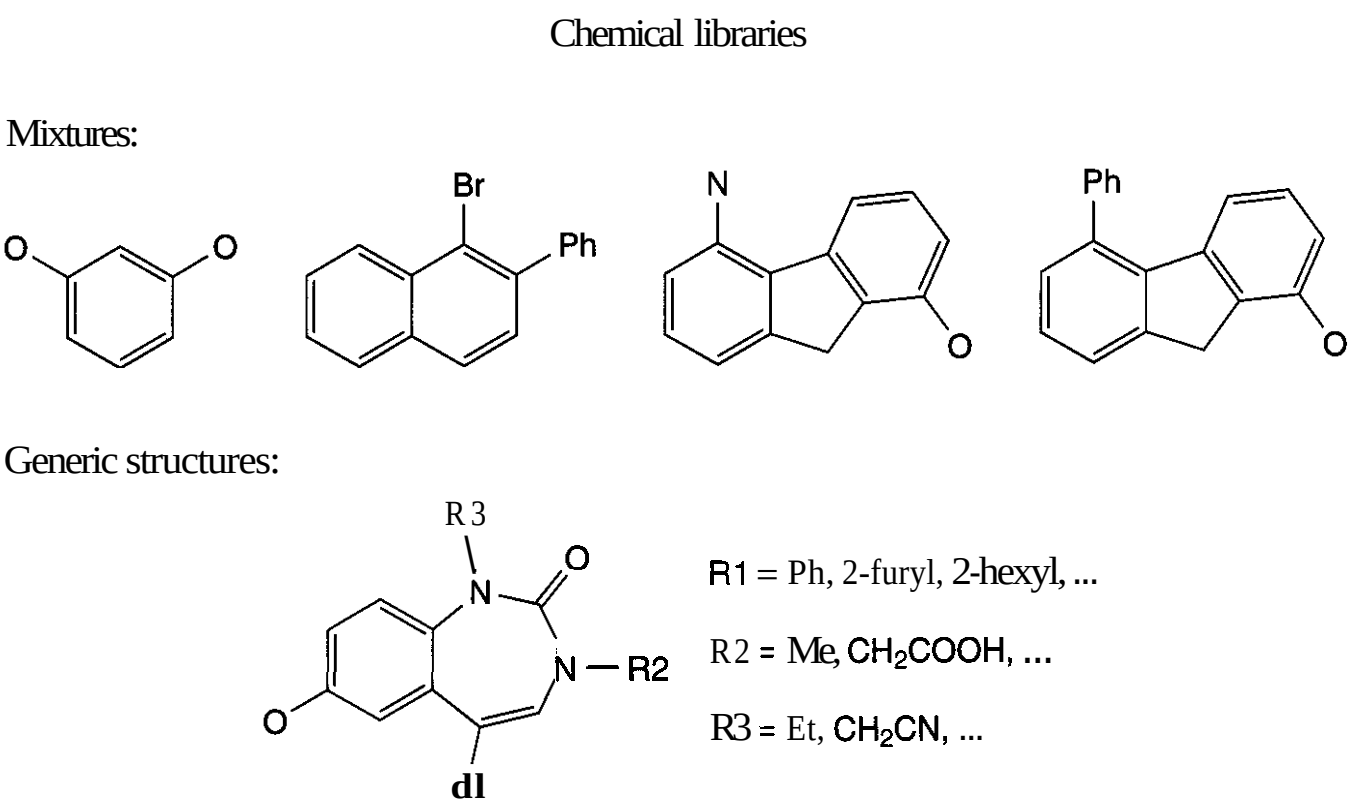


Figure 9.6. Chemical structure data for high-throughput chemistry. The generic structure representation is often referred to as a Markush structure.

requires adding data to specify percent or amount content in the mixture for each component.

2.1.6 Generic Structures. Generic or Markush structures are commonly used to represent structures for patent purposes. Since the introduction of combinatorial chemistry, generic structures and generic reactions have become a standard means of representing potentially huge numbers of specific compounds in a highly compact representation that is familiar to the chemist (40). The central structure of a generic, which is common to all the structures it represents, is commonly called the "root" or "parent." The variable parts of the structure (R_1, R_2, etc.) are referred to as the "Rgroups." The exact substituents that make up the various Rgroups (e.g., —Cl, —Br, —OH) are referred to as the "members" of the Rgroup. Finally, a specific combination of root and Rgroup members—which constitutes a single, real structure—is referred to as a specific or "enumerated" structure. Some chemical computations, like property and similarity calculations, can be performed on the generic structure without enumerating all the specific structures (41).

The remaining types of chemical data that need representation include substances and search queries.

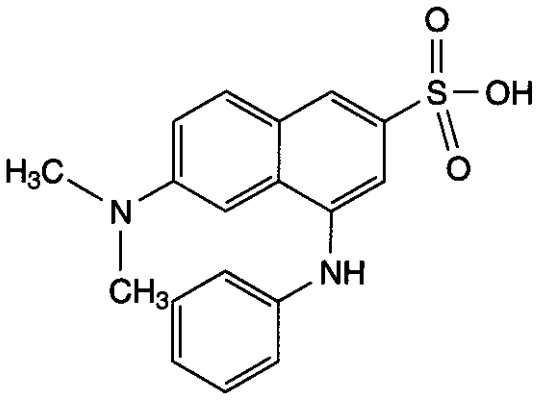
2.1.7 Substances. Less common in drug discovery, but very useful for material science and polymer chemistry, is the ability to store "substances." These include unspecified or uncertain chemical structures, polymers, and other chemical entities that cannot be classed with the other chemical representations (42). Polymers pose particular problems, as discussed in the article by Schultz and Wilks (43).

2.1.8 Search Queries. For all types of chemical representation, there are *query* representations that can be applied to a database to return a list of structures which match or "fit" the query, or that the query "hits" in the database. The same chemical drawing programs that are used to input structures can commonly be used to input chemical structure queries. These drawing programs currently include several programs in the commercial and public domains (44). A comparison of popular drawing programs has recently appeared on the Internet (45). Query structures often contain generalized atom types, bond types, and ring types. They may specify the required presence or absence of certain atom types or functional groups. In the case of 3D models, queries can be devised to represent pharmacophores for certain types of therapeutic activity (46). An important distinction must often be made between the query representation of a pharmacophore used for 3D searching and the conceptual pharmacophore used for drug development.

2.2 Types of Chemical Representation

A second way of looking at chemical representation is to consider the manner in which the chemical structure data is organized and exchanged, either in some file format or in a database. The most common ways of representing structures and reactions include the following.

2.2.1 Linear Notation. One of the earliest forms of chemical structure representation is Wiswesser line notation (WLN), developed in 1946. This notation used short letter codes to represent functional groups in molecules (47). An alternative early notation is the Beilstein ROSDAL string (48). These two formats are not used much today, having been replaced by the Daylight SMILES notation (49) and its extensions (50). Figure 9.7 shows a drug-like molecule along with WLN, SMILES, and ROSDAL notation. Also shown is a simple chemical reaction represented in SMILES. Note that SMILES and other linear notation schemes do not include 2D coordinates for display of the structure. These are either stored separately or generated on the fly (51). The SMILES notation has become especially popular for property estimation programs, because atom coordinates are not usually needed for connection table-based calculations. It is a very convenient method for web-based input of structures for property calculations (52). Note that the order of atoms in most linear notations is arbitrary, depending on where in the molecule the notation generator (program or chemist) starts. For this reason, some linear notations have a canonical (or "uniquified") form that



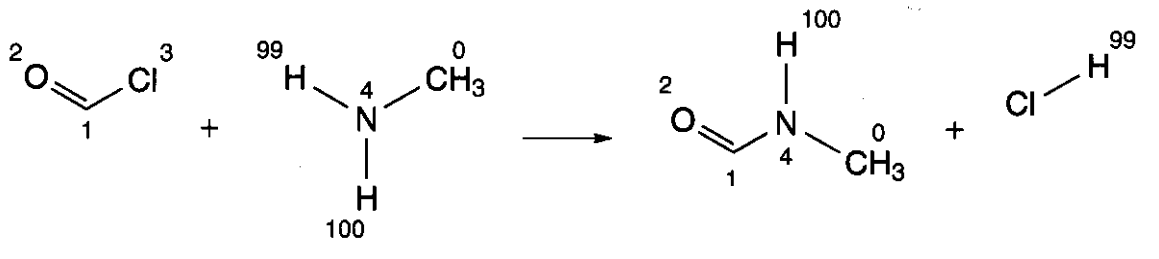
WLN: L66J BMR& DSWQ IN1&1

ROSDAL: 1 = -5- = 10 = 5, 10-1, 1-11N-12- = 17 = 12, 3-18S-19O, 18-200, 18 = 210, 8-22N-23, 22-24

SMILES: OS(=O)(=O)c1cc2ccc(cc2c(c1)Nc1ccccc1)N(C)C

SLN: OS(=O)(=O)C[1]:C:C[2]:C:C:C(:C:C(:@2):C(:C(:@1))NC[3]:C:C:C:C:C(:@3)) N(C)C

CHIME: 3aQf713AsUwQDjIMwyMWrSA7AOxHeqiAAPWRmMrZSZIJjTrAEfcsXH1JTUf...



SMILES:

[C:1](=[O:2])[Cl:3].[H:99][N:4]([H:100])[C:0]>> [C:1](=[O:2])[N:4]([H:100])[C:0].[Cl:3][H:99]

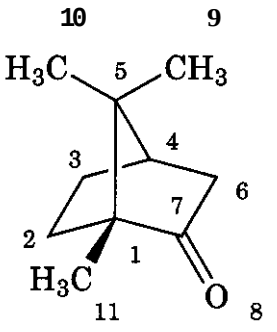
Figure 9.7. Various linear notation schemes for chemical representation. Some contain only atom types and connectivity (WLN, ROSDAL, SMILES, SLN) and are chemist-readable. Others are compressed versions of molecule file formats (CHIME) and are meant for computer interpretation.

places the atoms in a topological order, usually reflecting their degree of branching, the types of neighboring atoms and bonds, etc. This canonical ordering of the atoms reduces any user-input ordering to the same string. It can then be used for exact-match lookup of the structure, regardless of how it was drawn or typed. The SMILES notation has also been extended to include reactions as shown in Fig. 9.7 (53). Occasionally, other linear notations are described **(54).**

2.2.2 Tabular Storage. To preserve more specific information about atoms and bonds, such as coordinates, stereochemistry, charge, and isotope number, it is necessary to store molecule information in a tabular format. Each row of the table typically contains all the information about a single atom or bond. In some formats, the atom and bond information is combined on a single line. Table **9.2** shows three common file formats for a simple structure. In the MDL molfile format, the atom and bond information is separated into separate blocks. In the Hyperchem HIN file format, the bond information is mixed with the atom information, resulting in fewer records in the file. In the PDB format, the atoms can be assigned to residues. Descriptions of various formats can be found in the reference manuals for chemical management and molecular modeling programs or in the literature **(55).** The systems that manage reactions typically have their own file formats as well.

Both linear and tabular formats are capable of being transmitted over a network between computers. This allows passing structure information from a server to a workstation for display purposes. It is common to compress and/or encrypt the chemical structure information before it is transmitted, and then have the workstation or display program uncompress or decrypt the resulting structures. This is done for performance and

Table 9.2 Tabular Molecule File Formats



MDL Molfile format:

```
D-Camphor
  -ISIS-  03130218162D
2D molfile
 11 12  0  0  0  0  0  0  0  0999 V2000
   -2.0625   -1.1833    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.5583   -0.4167    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4208    0.1500    0.0000 C   0  0  3  0  0  0  0  0  0  0  0  0
   -0.8292   -0.7542    0.0000 C   0  0  2  0  0  0  0  0  0  0  0  0
   -0.7042    1.1667    0.0000 C   0  0  3  0  0  0  0  0  0  0  0  0
    0.9917   -0.4333    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.4667   -1.2167    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.5237    1.7332    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.9208    1.6686    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.9875   -2.2803    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.6004   -2.0787    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  1  0  0  0  0
  1  4  1  0  0  0  0
  4  5  1  0  0  0  0
  5  3  1  0  0  0  0
  3  6  1  0  0  0  0
  6  7  1  0  0  0  0
  7  4  1  0  0  0  0
  5  8  1  0  0  0  0
  5  9  1  0  0  0  0
  7  1  0  2  0  0  0  0
  4  1  1  1  1  0  0  0
M  END
```

Hyperchem HIN file format:

```
moll C:\TEMP\DCAMPHOR.HIN
atom     1  -  C  **  -  0  0.0000  0.8445  0.0000  2  2  s  4  s
atom     2  -  C  **  -  0  0.3881  1.4347  0.0000  2  1  s  3  s
atom     3  -  C  **  -  0  1.2639  1.8710  0.0000  3  2  s  5  s  6  s
atom     4  -  C  **  -  0  0.9494  1.1749  0.0000  4  1  s  5  s  7  s  11  s
atom     5  -  C  **  -  0  1.0457  2.6537  0.0000  4  4  s  3  s  8  s  9  s
atom     6  -  C  **  -  0  2.3513  1.4219  0.0000  2  3  s  7  s
atom     7  -  C  **  -  0  1.9471  0.8189  0.0000  3  6  s  4  s  10  d
atom     8  -  C  **  -  0  1.9910  3.0899  0.0000  1  5  s
atom     9  -  C  **  -  0  0.1090  3.0401  0.0000  1  5  s
atom    10  -  O  **  -  0  2.3481  0.0000  0.0000  1  7  d
atom    11  -  C  **  -  0  0.3557  0.1552  1.0000  1  4  s
endmol   1
```

Table 9.2 (Continued)

Protein Data Bank format:

```
HEADER    PROTEIN
COMPND    c:\temp\dcamphor.pdb
AUTHOR    GENERATED BY BABEL 1.6
ATOM      1  C   UNK     1      -2.063  -1.183   0.000  1.00  0.00
ATOM      2  C   UNK     1      -1.558  -0.417   0.000  1.00  0.00
ATOM      3  C   UNK     1      -0.421   0.150   0.000  1.00  0.00
ATOM      4  C   UNK     1      -0.829  -0.754   0.000  1.00  0.00
ATOM      5  C   UNK     1      -0.704   1.167   0.000  1.00  0.00
ATOM      6  C   UNK     1       0.992  -0.433   0.000  1.00  0.00
ATOM      7  C   UNK     1       0.467  -1.217   0.000  1.00  0.00
ATOM      8  C   UNK     1       0.524   1.733   0.000  1.00  0.00
ATOM      9  C   UNK     1      -1.921   1.669   0.000  1.00  0.00
ATOM     10  O   UNK     1       0.988  -2.280   0.000  1.00  0.00
ATOM     11  C   UNK     1      -1.600  -2.079   0.000  1.00  0.00
CONECT    1    2    4
CONECT    2    1    3
CONECT    3    2    5    6
CONECT    4    1    5    7   11
CONECT    5    4    3    8    9
CONECT    6    3    7
CONECT    7    6    4   10
CONECT    8    5
CONECT    9    5
CONECT   10    7
CONECT   11    4
MASTER        0    0    0    0    0    0    0    0   11    0   11    0
END
```

for security. In MDL systems, the Chime linear format is used to transmit structures and reactions, whereas Daylight systems simply use the SMILES representation and depict the structure on the fly (Fig. 9.7).

2.2.3 Graphical Representation. Occasionally it is desirable to store chemical structures as "pictures", usually for document purposes. For example, some chemical drawing packages and many molecular modeling packages can store structures as the following:

- WordPerfect or Microsoft Word document (.doc files)
- Extended postscript (.eps files)
- Windows metafile (.wmf files)
- A proprietary sketch (MDL .skc files)
- A variety of compressed graphics formats including JPEG (.jpg files), bitmap (.bmp files), GIF (.gif files), and TIFF (.tif files)

Often, the graphical format allows the connection table to be stored and transferred transparently with the image—through the computer's clipboard, for instance. This allows the receiving program to "interpret" the image as a chemical structure and manipulate it accordingly.

2.2.4 Markup Languages. The Internet has spawned a host of new "languages" that facilitate the exchange of information. The most common of these are HTML (hypertext markup language) and XML (extensible markup language). A variation of XML that is designed for chemical information exchange is the Chemical Markup Language CML (56). Although it is not widely used as of this writing, it bears watching as more web-based chemical information platforms become available. Problems with markup languages are that they are verbose compared with structure

Table 9.3 Chemical Markup Representation of Acetic Acid

```
(molecule convention="MDLMol" id="acetate" title="ACETATE")
(date day="23" month="11" year="1995" /)
⟨atomArray⟩
(atom id="a1")
(string builtin="elementType")C⟨/string⟩
(float builtin="x2")0.27⟨/float⟩
(float builtin="y2")0.1217⟨/float⟩
⟨/atom⟩
(atom id="a2")
(string builtin="elementType")C⟨/string⟩
(float builtin="x2")−1.27⟨/float⟩
(float builtin="y2")0.1246⟨/float⟩
⟨/atom⟩
(atom id="a3")
(string builtin="elementType")O⟨/string⟩
(float builtin="x2")1.0623⟨/float⟩
(float builtin="y2")−1.2937⟨/float⟩
⟨/atom⟩
(atom id="a4")
(string builtin="elementType")O⟨/string⟩
(float builtin="x2")1.1008⟨/float⟩
(float builtin="y2")1.4332⟨/float⟩
⟨/atom⟩
⟨/atomArray⟩
⟨bondArray⟩
(bond id="b1")
(string builtin="atomRef")a1⟨/string⟩
(string builtin="atomRef")a2⟨/string⟩
(string builtin="order")1⟨/string⟩
⟨/bond⟩
(bond id="b2")
(string builtin="atomRef")a1⟨/string⟩
(string builtin="atomRef")a3⟨/string⟩
(string builtin="order")1⟨/string⟩
(/bond)
(bond id="b3")
(string builtin="atomRef")a1⟨/string⟩
(string builtin="atomRef")a4⟨/string⟩
(string builtin="order")2⟨/string⟩
(/bond)
⟨/bondArray⟩
(Imolecule)
```

files, and they are difficult for chemists to read (although they are not usually meant for chemist interpretation). This is evident in Table **9.3,** which shows the CML for acetic acid. By comparison, the SMILES for acetic acid is simply "CC(=O)O".

2.3 Chemical Structure File Conversion

Many chemical information management systems, especially modeling programs, permit a chemist to import and export structures using a variety of file formats. Commercial programs designed specifically for file conversion are available (57). **A** widely used public domain program, Babel, is available in source code and in a Windows version (58). It is being extended by the "OpenBabel" programming project (59). It is possible, with a fair amount of accuracy, to convert a chemical structure from a connection table format to an acceptable

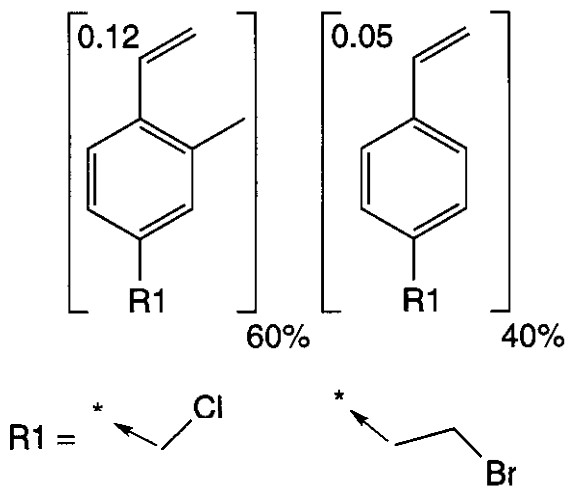


Figure 9.8. Structure representation with additional information, including atom (partial charge) and fragment (percent composition) data and Markush structure features.

IUPAC name (60). The reverse conversion of names to 2D structures is also possible (61).

2.4 Representing Nonstructural Chemical Data

Nonstructural chemical data includes any textual, numeric, or binary data that is not directly a part of the chemical structure. It includes the following:

- Whole-molecule data such as physicochemical properties, spectral data, literature citations, availability, biological or therapeutic activity, etc. In either the molecule file or in a database, data are typically maintained in fields that are separate from the structure field, but linked by some identifier.
- Atom, bond, or fragment-based data such as partial charge, component fraction, secondary structure, various fragment-based physicochemical and QSAR properties, etc. Because these data are linked to particular atoms, bonds, fragments, or components in the structure, they are typically stored along with some indication of the substructural features with which they are associated.

Figure 9.8 shows a structure that contains atom (partial charge) and fragment (component percent) data of various types, as well as Markush structure features (R_1). Table 9.4 shows the corresponding structure-data file (MDL sdfile). Each Rgroup member appears in its own "submolfile" in this file representation. So-called Sgroup data appears as part of the molfile, along with the name of the data field, its value, location on the display, and the atoms and bonds that bear the data.

3 STORING AND SEARCHING CHEMICAL STRUCTURES AND REACTIONS

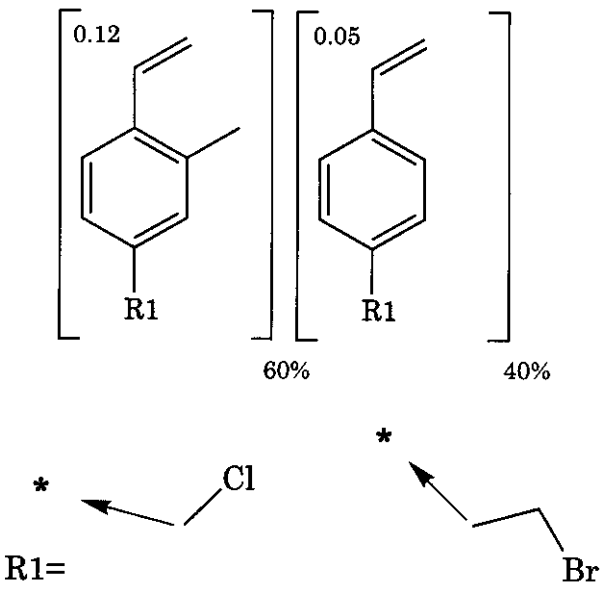
In the simplest sense, searching chemical information consists of (1) finding structures or reactions that meet the chemist's search criteria and/or (2) finding data that meets the search criteria. Data searching (numbers and text) is a well-established informatics activity, supported by spreadsheets, word processors, and relational database systems. Chemical structures and reactions are a unique form of data. Searching for full or partial matches to structures, models, and reactions requires highly specialized databases and search techniques.

3.1 Storing Chemical Information in Databases

When they were first developed, chemical structure databases consisted of record-oriented flat files, much like a spreadsheet whose columns have each been cut out and placed in a separate file. This organization has limitations in searching, access, and efficiency of, storage. Also, it is not the most appropriate form for storing generic structures and reactions, which are more hierarchical in nature. As a result, since the 1990s, chemical information has become increasingly stored in commercial relational database systems, chiefly Oracle and Microsoft Access. Relational storage has the added advantage of combining chemical structure storage with biological data and inventory data (location, cost, units on hand, etc.) that are often stored in the corporation's relational databases. One of the first reports of the relational storage of structures was by Hagadone and Lajiness, who modified the Upjohn COUSIN system (62).

An example of a current commercial relational chemical database is seen in Fig. 9.9, which shows the organization of a basic ISIS chemical database. Each labeled item in the figure is a table in an Oracle relational database. The tables in the database consist of the following.

Table 9.4 Example Molfile Showing: Markush Features, Atom, and Fragment Data




```
$MDL REV 1 29AUG01 17:47
$MOL
$HDR
Figure 8 molfile
  MACCS-II08290117472D              1            0.00487            0.00000            0

$END HDR
$CTAB
  1  9  1  9  0  0  0  0                      25  V2000
    -5.3301    1.0237    0.0000 C   0  0  0  0  0  0
    -5.3323   -0.5232    0.0000 C   0  0  0  0  0  0
    -3.9956   -1.2952    0.0000 C   0  0  0  0  0  0
    -2.6560   -0.5224    0.0000 C   0  0  0  0  0  0
    -2.6615    1.0307    0.0000 C   0  0  0  0  0  0
    -3.9990    1.7956    0.0000 C   0  0  0  0  0  0
    -3.9881    3.2993    0.0000 C   0  0  0  0  0  0
    -3.9960   -2.8378    0.0000 R#  0  0  0  0  0  0
    -2.6701    4.1134    0.0000 C   0  0  0  0  0  0
    -1.3283    1.8070    0.0000 C   0  0  0  0  0  0
     0.8559    0.9069    0.0000 C   0  0  0  0  0  0
     0.8538   -0.6400    0.0000 C   0  0  0  0  0  0
     2.1904   -1.4121    0.0000 C   0  0  0  0  0  0
     3.5299   -0.6393    0.0000 C   0  0  0  0  0  0
     3.5247    0.9138    0.0000 C   0  0  0  0  0  0
     2.1870    1.6787    0.0000 C   0  0  0  0  0  0
     2.1823    3.2214    0.0000 C   0  0  0  0  0  0
     3.5161    3.9966    0.0000 C   0  0  0  0  0  0
     2.1900   -2.9547    0.0000 R#  0  0  0  0  0  0
  8  3  1  0  0  0
  2  3  1  0  0  0
  7  9  2  0  0  0
  5  1  0  1  0  0  0
  3  4  2  0  0  0
  1  1  1  2  2  0  0  0
  4  5  1  0  0  0
  1  2  1  3  1  0  0  0
  1  3  1  4  2  0  0  0
  5  6  2  0  0  0
  1  4  1  5  1  0  0  0
  6  1  1  0  0  0
  1  5  1  6  2  0  0  0
  1  6  1  1  1  0  0  0
  1  2  2  0  0  0
  1  6  1  7  1  0  0  0
  6  7  1  0  0  0
  1  7  1  8  2  0  0  0
  1  3  1  9  1  0  0  0
```

Table 9.4 ***(Continued)***

```
M  RGP  2  8  1  19  1
M  STY  6  1 GEN  2 GEN  3 DAT  4 DAT  5 DAT  6 DAT
M  SLB  6  1  1  2  2  3  3  4  4  5  5  6  6
M  SAL  1  9  11  12  13  14  15  16  17  18  19
M  SDI  1  4  0.0515  -3.7303  0.0515  4.7961
M  SDI  1  4  4.8009  4.7961  4.8009  -3.7303
M  SAL  2  10  1  2  3  4  5  6  7  8  9  10
M  SDI  2  4  -6.1377  -3.6181  -6.1377  4.9083
M  SDI  2  4  -0.5469  4.9083  -0.5469  -3.6181
M  SAL  3  1  7
M  SDT  3 PCHARGE  F  MQ
M  SDD  3  -5.4645  3.7303  DA  ALL  1  5
M  SED  3  0.12
M  SDT  4  PCT  F  MQ
M  SDD  4  -0.9021  -4.9083  DA  ALL  1  5
M  SED  4  60%
M  SDT  5  PCT  F  MQ
M  SDD  5  5.2122  -4.7961  DA  ALL  1  5
M  SED  5  40%
M  SAL  6  1  17
M  SDT  6 PCHARGE  F  MQ
M  SDD  6  0.5002  3.9173  DA  ALL  1  5
M  SED  6  0.05
M  SPL  2  4  2  5  1
M  END
$END CTAB
$RGP
  1
$CTAB
  2  1  0  0  0  0  2 V2000
    -3.7453  0.0472  0.0000  C  0  0  0  0  0  0
    -2.5668  1.0385  0.0000  Cl  0  0  0  0  0  0
  1  2  1  0  0  0
M  APO  1  1  1
M  END
$END CTAB
$CTAB
  3  2  0  0  0  0  2 V2000
    -5.0122  0.6013  0.0000  C  0  0  0  0  0  0
    -3.5311  1.0229  0.0000  C  0  0  0  0  0  0
    -2.6113  -0.2666  0.0000  Br  0  0  0  0  0  0
  1  2  1  0  0  0
  3  2  1  0  0  0
M  APO  1  1  1
M  END
$END CTAB
$END RGP
$END MOL
```

- The master "data dictionary" table, which describes all the objects in the database, as well as some parameters that are specific to the database (exact match criteria, version of the database, etc.). This is sometimes referred to as "metadata" or "data about data."
- A handful of tables that contain database parameters. These include substructure search key definitions, the periodic table used with the database, and a list of salt moieties that can be considered during searches.

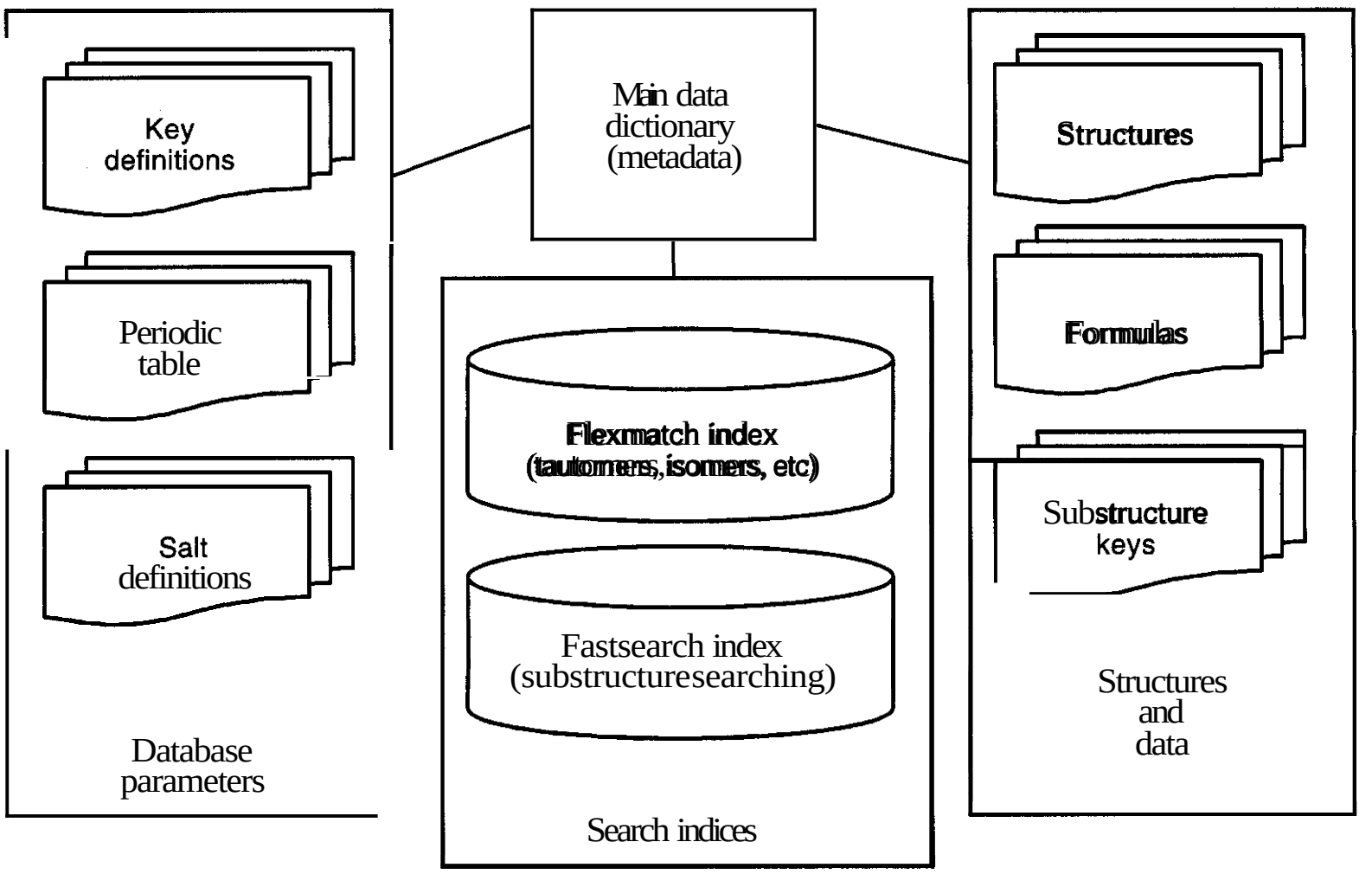


Figure 9.9. ISIS, a relational chemical structure database.

Structure and data storage is shown on the right. A structure table contains the structures, their internal identifiers, and their external identifiers, if any. The structures are stored in a compact binary representation that includes the connection table, the coordinates, the ring information, and any stereochemical, valence, isomer, isotope, or bond information. Certain types of structure-specific information such as polymer or component designations are stored here, whereas other types of structure-specific information (atom- or bond-specific data, and more verbose text data) are stored in their own tables, referenced by the internal identifier, and the atom or bond numbers to which the data correspond. A formula table contains the molecular formula and various atom and atom-type indexes to enhance formula searching and sorting.

- A table of substructure keys containing a binary or text string of the substructure fingerprint that was identified in the given structure at registration time. These keys represent the presence of either simple functional groups (e.g., phenyl ring, carbonyl), or more complex atom/bond combinations (e.g., carbonyl separated from a secondary amine by three bonds). In ISIS, a set of 166 searchable keys can be explicitly used as filters for structure searching. A larger set of 960 keys is used for similarity calculations. For 3D models, it is common to generate 3D pharmacophore keys, which encode all the possible 2- and 3-point pharmacophores represented in the structure, sometimes considering multiple conformations.
- A third kind of information includes indexes to enable structure and substructure searching. A "flexmatch" table contains a numerical hash (see Glossary) of certain features of the molecule, including stereochemistry, charge, and isotopes. This table can be used to retrieve a set of candidate structures quickly for exact match verification (63). It can also be used for "fuzzy" exact-match searches to retrieve tautomers and isomers of the input structure.
- Another index table contains a "fastsearch" index. This contains a single balanced tree (see Glossary) of all the substructural frag-

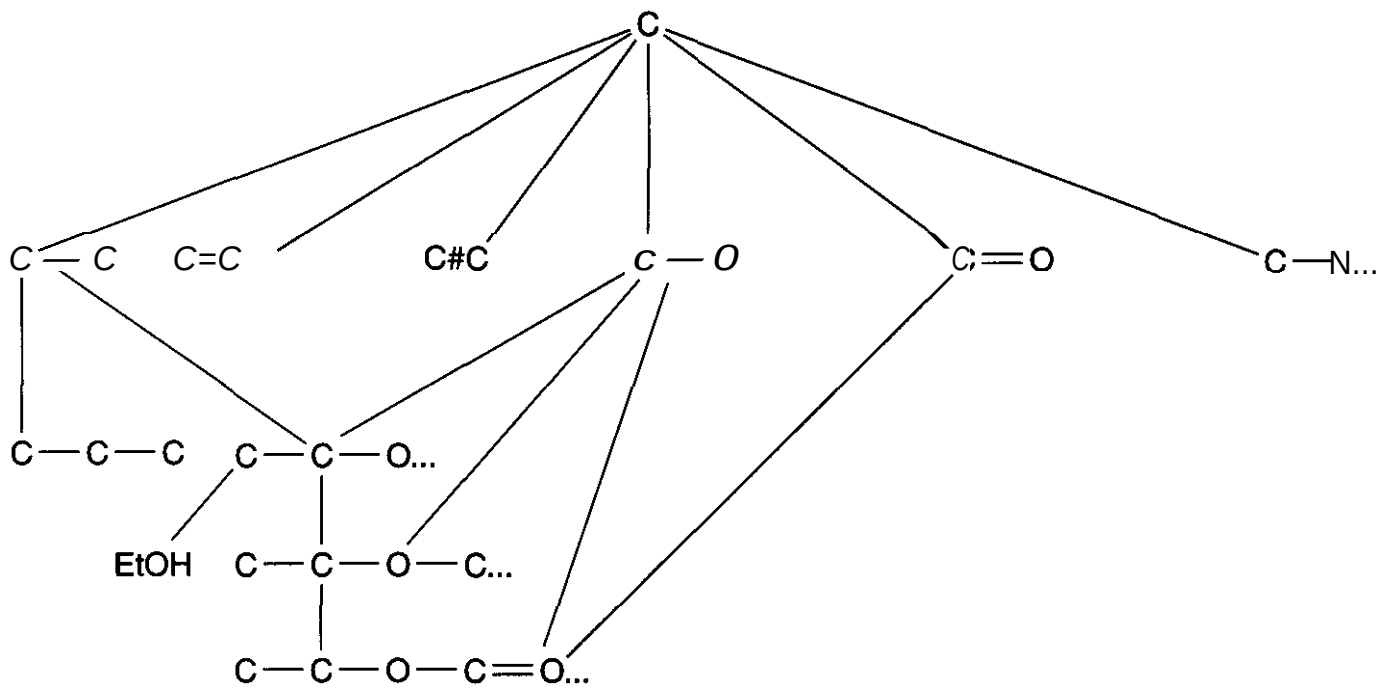


Figure **9.10.** Simplified ISIS Fastsearch index—ethanol is a leaf node that can be reached from several substructure nodes.

ments found in structures in the database, up to a fixed pathlength. These are stored in a highly compressed binary format (Fig. 9.10). Similar approaches have appeared in the literature (64). Leaf nodes in the tree contain identifiers of specific structures in the database (simplified in Fig. 9.10). An exact match or substructure search consists of traversing the tree to find structures in the database that have substructural fragments in common with the query structure. Because the fastsearch index is large—often as large as the rest of the structure database, updating it for the addition or removal of structures is time consuming.

This relational chemical database format is extended in ISIS to include 3D models, generic structures, and most recently, reactions. In these cases, additional "trees" in the database hierarchy connect 2D structures with 3D models, connect root structures with corresponding Rgroup members, or connect molecules with reactions.

Other relational *structure/reaction* database systems are available commercially. These include the Thor system from Daylight (65), Accord and RS^3 Discovery from Accelrys (66), and Unity from Tripos (67). Personal database systems that can be implemented on a desktop computer include ISIS/Base (68), Accord for Access (66), and Team Works from Afferent (69).

3.2 Registering Chemical Information

Chemical structure registration is an important activity that is necessary for drug discovery. The structures that have been developed by a pharmaceutical company constitute the "crown jewels" of chemical information, and they must be properly and securely archived. The registration process usually involves the process of extracting, cleaning, transforming, and loading the data—sometimes termed ECTL.

3.2.1 Extract the Data. First, the *structures/reactions* and corresponding data are extracted, collected, and validated. Increasingly, this is managed automatically, using output from the high-throughput chemistry process. Laboratory information management systems (LIMS) that are "structure smart" can manage chemical structure information starting from the design of a reaction, through the synthesis of the compounds, the chemical analysis of the structures, the in vitro biological assay, and finally the storage in the chemical database. Certain steps, such as drawing the initial structures/reactions, still remain an activity for the chemist, although many chemical information systems can take a generic structure, enumerate the many specific combinations, and layout the structures automatically (for example, the Monomer Toolkit by Day-

light, the Central Library program by MDL, and CombiLibMaker by Tripos).

3.2.2 Cleaning and Transforming the Data. Next, the structures/reactions are passed through a filtering program that searches for structure anomalies and corrects the chemical representation. In this step, chemical "business rules" are applied to the structures to insure that representations that can be drawn in different ways, such as nitro groups and tautomers, are represented by a single convention. Specialized chemical manipulation languages such as and Genie Control Language by Daylight, Cheshire by MDL, and Sybyl Programming Language by Tripos are used to implement this step. These languages are versatile and easily programmed, and they can be applied to other steps in the drug discovery process, such as searching, property calculation, and structure manipulation in general.

3.2.3 Loading the Data. Finally the structureslreactions are handed to a chemical registration system. The chemical registration system will typically "perceive" the structures—identify atoms, bonds, rings, stereochemistry, valence states, isotope values, and other chemical information as needed. In the case of reactions, it notes which structures are reactants, which are products, and which are agents or catalysts. Because there can be many valid ways of drawing a structure depending on which atom you start with, a structure may be given a canonical renumbering of the atoms using a variant of the Morgan algorithm (70). In the case of a linear representation like SMILES, this canonicalization yields a unique string for the structure, which can be generated from any valid SMILES string derived from the structure. In the case of a structure stored in a connection table, the Morgan algorithm results in the atoms being reordered in the connection table to generate a tree, branching outwards from the most highly connected atom in the structure. Because of the efficiency of indexing in modern relational chemical databases, Morgan renumbering is not used as much today as in the past.

The registration system then computes indexes. These include structure-searching indexes, substructure or similarity keys, molecular formula, molecular weight, and other structure-based properties. Substructure keys or "fingerprints" are particularly important. They consist of a number of binary descriptors for the presence of certain functional groups or more generalized atom/bond combinations. These keys can be used to filter structures before searching. They are also used for similarity calculations. Originally, substructure search keys were always used to filter structures before performing a substructure search of the database. If a query structure contains, say, a carbonyl group, then only carbonyl-containing structures should be examined during the substructure search. A key representing the carbonyl group can be used to filter structures that contain the group (the key turned on, or set to 1) from those lacking it (key set to 0). Tree-based substructure searching does not require prior filtering, so today, substructure keys are primarily used for similarity calculations between molecules. If the key values of two structures are compared, the more keys they have in common, the higher their similarity value will be. When registering reactions, the reactants and products may undergo automated or semi-automated perception of reacting bonds and atom centers (71). Generic structures may be analyzed and "clipped" or reverse-transformed to generate root and member structures, which may be stored separately (72).

Before finally storing the structure in the database, the registration program may search the database for some level of match to the input structure or reaction, and skip the registration if it is a duplicate. This is sometimes termed "deduplication" through "exact match" searching. There is usually some redundancy in chemical databases, and to save search time and disk space, most companies do not store duplicate structures or reactions, but rather store pointers to them.

The final step, after registering the structure or reaction, is to assign it a unique registry identifier, which is typically used throughout the company to identify the given structure/reaction and any chemical, biological, or inventory data that is associated with it. Some identifiers, like the Chemical Abstracts Service CAS number and the Beilstein BRN,

have wide application, and these may be used in addition to, or instead of, a corporate-assigned external registry identifier.

3.3 Searching Chemical Structures and Reactions

The type of chemical structure and reaction searching that a chemist does usually depends on the current stage of a project. For example, if the chemist is starting a new therapeutic project, a *therapeutic activity* search might be conducted, using a database such as the Derwent World Drug Index, the MDL Drug Data Report, or the MDL Comprehensive Medicinal Chemistry database. Retrieving many search hits, the chemist might organize them by *sorting* on name, molecular weight, ring system, or some topological basis. If the resulting list is too large, the chemist might perform a *cluster analysis* of the structures to see what general classes of compounds have been synthesized in the past. After sampling from the various clusters, and identifying a handful of interesting structures, the chemist might perform a *substructure* search to find structures that contain the features that are felt to be important to activity (i.e., the pharmacophore). If that search returns too many hits, the search query can be refined by making it more specific. If the search returns too few hits, the search query can be relaxed, or a *similarity* search can be used to find structures in the topological neighborhood of the query structure. Eventually, a number of structures will be obtained as candidates for synthesis and/or testing.

The next step is to design a set of reactions to synthesize the compounds. One or more reaction databases can be searched to find whether any reactions give the desired structures as products or give structures that are similar to the desired ones. The chemist may also use reaction similarity searching (73) and searching across reaction schemes (e.g., if A + B $\rightarrow$ C + D and C + E $\rightarrow$ F + G; a reaction scheme search will find the query A $\rightarrow$ F) (74). Once a reaction is found, the chemist needs to decide what reagents to use in the synthesis and where to obtain them. The selection of reagents will usually be based on a combination of physicochemical property considerations (i.e., QSAR and diversity), tempered by the chemist's experience and preferences, and balanced by synthetic feasibility and economics. The reagents may be located in-house, or they may require ordering from a chemical supplier.

A completely separate approach to reaction discovery is the reaction planning approach implemented in such programs as Logic and Heuristics Applied to Synthetic Analysis (LHASA) (75). This program works by searching a chemical knowledge base that contains information on approximately 2300 retro-reactions or transforms. The chemist draws a target molecule and indicates a strategy for the reverse-synthetic analysis. The program then searches the transform knowledge base for transforms that satisfy the strategy the chemist selected. The program decides which transforms are suitable for the particular target structure and displays the resulting precursors to the chemist. The chemist can then select a precursor for further analysis and choose another strategy option, on which the program returns a second level of precursors in the same way. Processing continues in this manner until the chemist is satisfied that one or more of the precursors correspond to a reasonable starting point for a synthesis. Retrosynthetic methods have not become as widely used in industry as reaction searching, partly because the certainty of the reactions is not guaranteed. Also, searching existing reaction databases generally yields the desired reaction or something close to it. Indeed, a major problem with search results from reaction databases is often an overabundance of hits, which typically need further organization and filtering to be useful. One approach to organizing the results of reaction searching is to apply some clustering or classification to the reactions (76).

To support the workflow just described, a number of structure and reaction search types have come into use (Fig. 9.11). These are briefly described as follows.

3.3.1 Exact Match Searching. Here, the chemist has a particular structure (or reaction) that he wishes to find in the database. The structure/reaction is drawn using a drawing program and then passed to a search program. The program submits the query to the

Searching chemical information

Type of search	Strings	Structures	Reactions	3D Models	Chemical libraries
Exact match	X	X	X	X	X
Substructure	X	X	X	X	X
Pharmacophore		X		X	
Similarity	X	X	X	X	X
Isomer/tautomer/salt		X			X

Figure 9.11. Search types depend on the nature of the chemical information.

search routine that typically generates index values from the query that are of the same type as those generated for structures/reactions when stored in the database. The index values are then used as filters to retrieve a set of candidate structures/reactions. In ISIS, these filters include the formula, the molecular weight, and the flexmatch index, a numeric hash code based on the presence of isomers, tautomers, isotopes, salts, charges, and stereochemistry (see Glossary). The resulting filtered structures have the minimum set of requirements to fit the search query, but typically only a fraction of these structures will fit the query exactly. Once this set of candidate structures is obtained, the query structure is *mapped* to the candidate structure using a process known as atom-atom mapping, which is known in topology as the "graph isomorphism" problem. This mapping is time-consuming, so the prior filtering step should be as efficient as possible. Each structure that maps exactly to the query is placed in the result set or "hit list." To accommodate various chemists' needs, exact match searching can usually be "relaxed" to permit the finding of isomers, tautomers, salts, charged or uncharged species, etc. In the case of reactions, variations of the reaction can be retrieved-by relaxing the constraints on the reaction conditions, solvent, and catalyst (Fig. **9.12).**

In a Daylight Thor database, where the

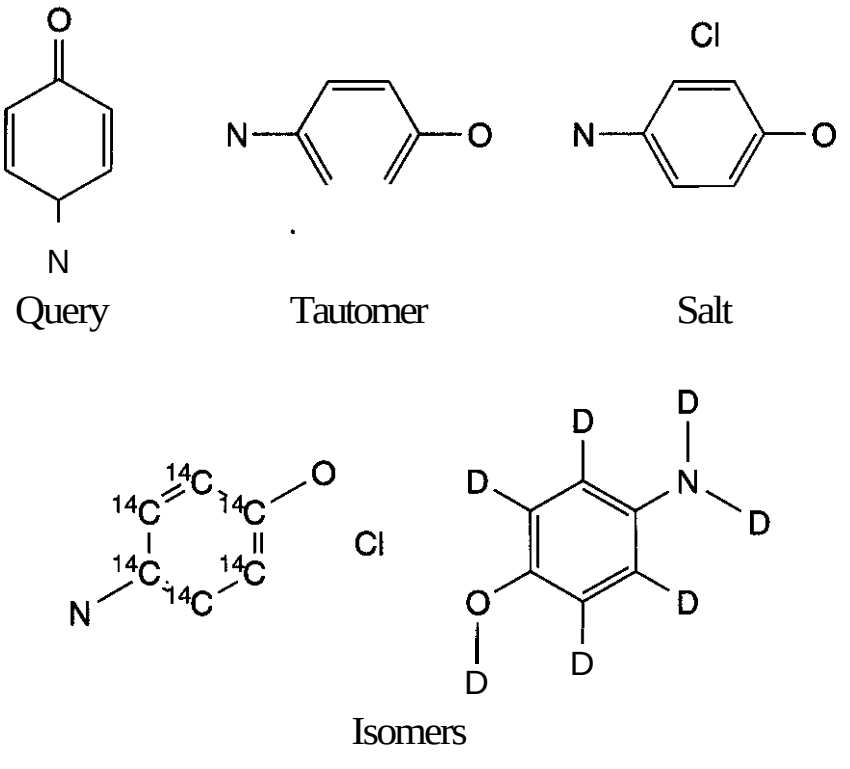


Figure 9.12. Different degrees of exactness can be defined by allowing tautomers, salts, and isomers successively in the search.

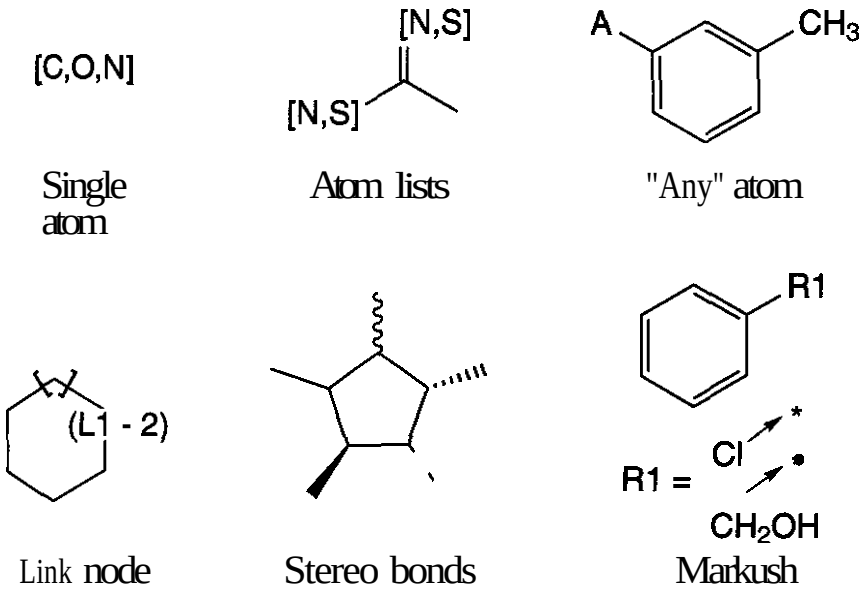


Figure 9.13. Example 2D substructure search queries with various atom and bond query features. The more features that are present, the more flexible the search becomes, but the search may also require more time to complete. There is a trade-off between putting the flexibility into the database (i.e., storing and indexing multiple forms of a structure) and putting the flexibility into the search query and the search software.

structure is stored as unique SMILES, the canonical query SMILES can be compared lexically with strings in the database using fast string comparison and indexing techniques, to find exact match structures and reactions. Because a structure in a Thor database consists of a meaningful, canonical sequence of characters, the computational efficiencies of string searching and comparison can be applied when searching the database. This is in contrast to the highly specialized search techniques used in other structure database formats.

3.3.2 Substructure Searching. A substructure search is performed when a chemist has in mind a pharmacophore consisting of a set of functional groups or a substructure which he knows must be present in the structures to be retrieved. Only part of the molecule is drawn, along with ***query features*** that generalize atoms, bonds, and rings in the structure. Figure 9.13 shows some typical substructure query features. The features include the following:

- Single atom—specifies a periodic table atom that must be present or a more generalized atom (hetero, metal, etc.) or "superatom" (condensed functional group, such as Ph, Et, Ala, etc.)
- Atom list—a list of atoms, any one of which may be present
- "Any" atom—which simply means some atom must be attached at the given position. As with structures, the hydrogen atoms in substructures are implicit, unless the user requests a particular hydrogen count or range at a given position
- Link node—which specifies a range of allowed atom or functional group links between atoms
- Stereo bond—including Z/E/either or up/down/either
- Markush feature—used for patent representation, for representation of generic structures for combinatorial chemistry, or to limit the substituents that can be present at a given position. Note that some systems allow logical operations on Markush features (if —OH at R_1, then no —Cl at R_2).

A specialized case of substructure searching is 3D pharmacophore searching, in which a substructure search is combined with the measurement/generation of 3D features, to identify models that could fit a 3D pharmacophore. Figure 9.14 shows an example of a 3D substructure search query that includes various 3D features or constraints. A given conformation of a molecule that is stored in the database may not exactly match a given query, but it could be modified by rotation about single bonds to fit the query. For this reason, conformationally flexible 3D searching is a feature of most 3D database systems (77). When searching conformers, the conformational flexibility can be incorporated into (1) the query, by tethering flexible groups to fixed anchor points in the structure, (2) the database, by storing multiple low energy conformations for each structure, or (3) the search process, by

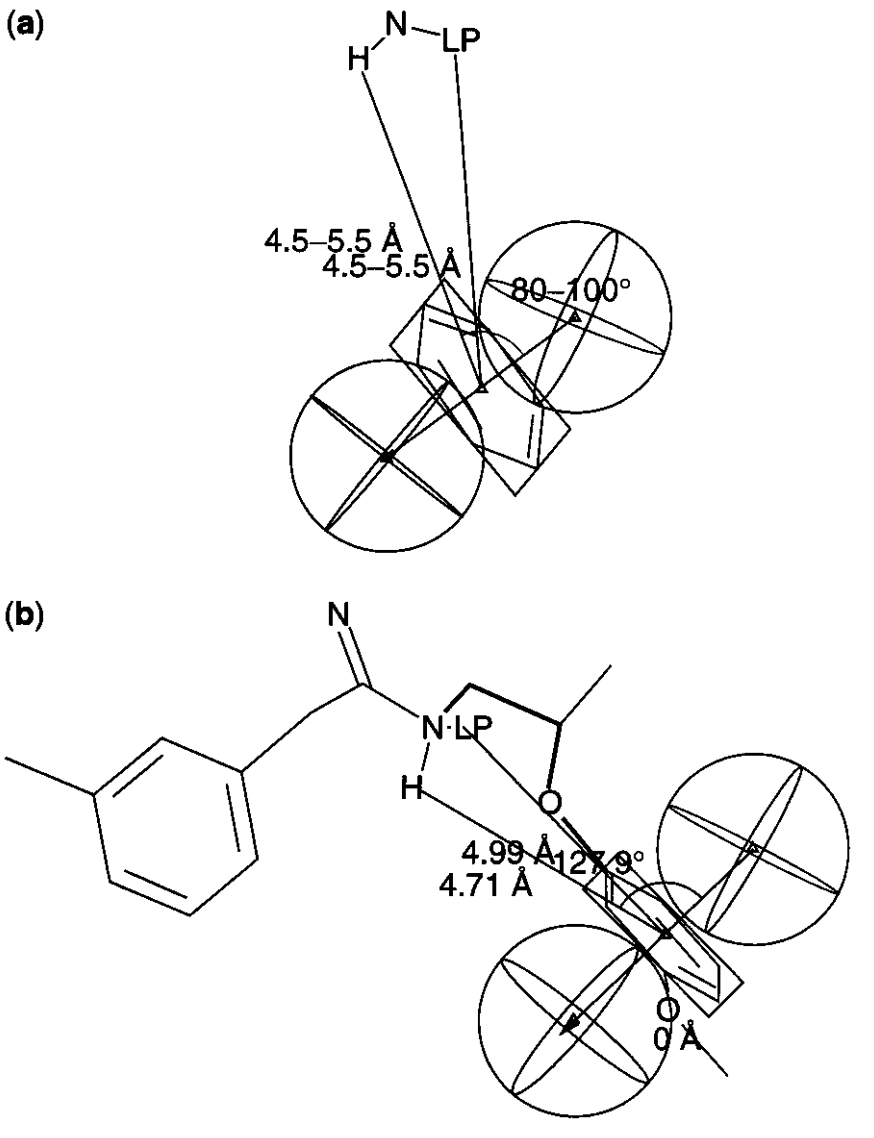


Figure 9.14. (a) Example 3D pharmacophore search query, showing substructure, distance, angle, and exclusion sphere constraints. (b) Example result of a conformationally flexible 3D search using this query. The molecule was "flexed" in 3D by rotating about the highlighted single bonds to fit the query. The atoms, bonds, and 3D features that match the query are colored. One problem with conformationally flexible 3D searching is that unwanted hits can be conformed to fit the query.

incorporating a rapid conformational analysis into the 3D search algorithm. The last is the most common approach and is a part of database systems from Tripos, Accelrys, and MDL.

Many different approaches to substructure searching have been devised (78). In ISIS, the fastsearch index file is used to retrieve candidate structures. If needed, the query is then mapped onto these structures using a "backtracking" approach. This involves successively matching atoms and bonds in the structure to those in the query in a stepwise manner. When a match fails at any given step, the program backtracks to the last successful step and selects an alternative atom or bond. Once all the atoms and bonds have been matched, the structure is considered a hit. An issue of the ***Journal of Chemical Information and Computer Sciences*** has been devoted to substructure search methods (79). Hicks and Jochum reported a comparison of several substructure search algorithms in 1990 (80). These authors found the Beilstein-Softron S4 search system to be superior in search speed at that time.

3.3.3 Similarity Searching. The most generalized type of structure/reaction searching is searching for "similar" structures or reactiong in the database. Chemical similarity has been a highly debated topic for some time, mostly from the standpoint of what constitutes good descriptors to use in the similarity calculations (81). Nevertheless, there are some general approaches that are widely used, not because of their theoretical soundness, but simply because they work for the chemist. For 2D structures, the most useful and efficient similarity approach is key-based similarity. This involves computing the overlap between a query structure and a candidate structure using substructure or fragment keys. ISIS uses the 960 keys that are generated when the structure is registered. The overlap is typically computed using the Tanimoto metric, which was first used in 2D structure similarity by Willett et al. (82). Depending on the nature and number of the keys, it may be desirable to weight the Tanimoto calculation inversely according to the prevalence of the key in the database. Thus, a cyclopropyl key, which may

not be highly prevalent in the database, and would be "swamped" by other, less relevant keys in an unweighted similarity calculation, may have more influence in a weighted calculation. This weighted calculation is used as the default in ISIS chemical databases. It is possible for an ISIS database administrator to regenerate the keys using custom values of the weights to enhance differences in the similarity calculations and select, say, more "drug-like" molecules in the search. In the reaction domain, similarity can be defined in terms of the structures, the reactions, or a combination of the two (83). Other similarity search systems have been described in the literature, including the one used by CAS (84). It is also possible to use 3D pharmacophore keys to compute similarity, although these have typically not performed as well as 2D keys. It is possible that conformational flexibility so vastly expands the "chemical space" of the molecules that a limited number of keys is simply inadequate for 3D similarity calculation. When attempting to predict the type of therapeutic activity a compound has, Briem and Lessel concluded that 2D and 3D keys have complementary information (85).

3.3.4 Reaction Searching. Reaction searching, sometimes called reaction indexing, has been available for over 20 years. Originally developed as online searching systems, the introduction of in-house systems like REACCS allowed pharmaceutical companies to augment published reaction sources with their own reactions and data (86). As with molecules, reaction storage has moved from proprietary database foundations to storage and access in relational systems. Reaction searching encompasses many of the same types of searches used for molecules. A reaction typically consists of three types of structures: reactants, products, and catalysts or agents, along with textual information about yield, conditions, etc. Reactant and product structures undergo structural changes in the reaction, whereas agents do not. The atom and bond changes that occur in a reaction are isolated in one or more reacting centers of the reactants and products. The atom changes consist of changes in atom valence, charge, number of attached hydrogens, number of bonds attached, retention or inversion of stereochemistry, etc. Bond changes include making and breaking of bonds, and changes in bond order and stereochemistry. When searching reactions, the chemist can search for exact, isomer, or substructure matches in the reactants, in the products, or both. The structure searching can be accompanied by a search of the reaction text information for yield and conditions. Several commercial reaction indexing systems are available from molecule database vendors, and online searching is even possible (87).

In most reactions, the majority of the atoms and bonds are not involved in the reaction, and they remain unchanged between reactants and products. To avoid examining these unchanging atoms and bonds, most reaction indexing systems allow the user to mark, in the reactants and products, those atoms and bonds that are involved in the reaction. These are termed ***reacting center*** atoms and bonds, and when they are present, they enable much faster reaction searching and they reduce the number of false hits obtained. A simple example is seen in Fig. 9.15. Some systems have semiautomatic perception of reacting centers, which must usually be augmented or checked by a chemist, especially with complex transformations.

As with molecules, it is also possible to do reaction similarity searching. Given a reaction with reactants, products, and agents, one can typically run molecule similarity searches for the reactants, the products, or both. This will retrieve reactions that have similar structures involved in them. This does not guarantee that the molecules undergo the same or even similar transformations. It is possible in some systems to also include the similarity of the transformation as part of the overall similarity search. This is usually carried out using special keys that have been generated for a fixed number of possible transformations. As with molecules, the more keys a query and a reaction have in common, the higher will be the similarity.

3.3.5 Searching Other Data. Data other than structures and reactions must also be searched in the drug discovery process. Various systems exist for indexing and searching literature and journal contents (88), patents

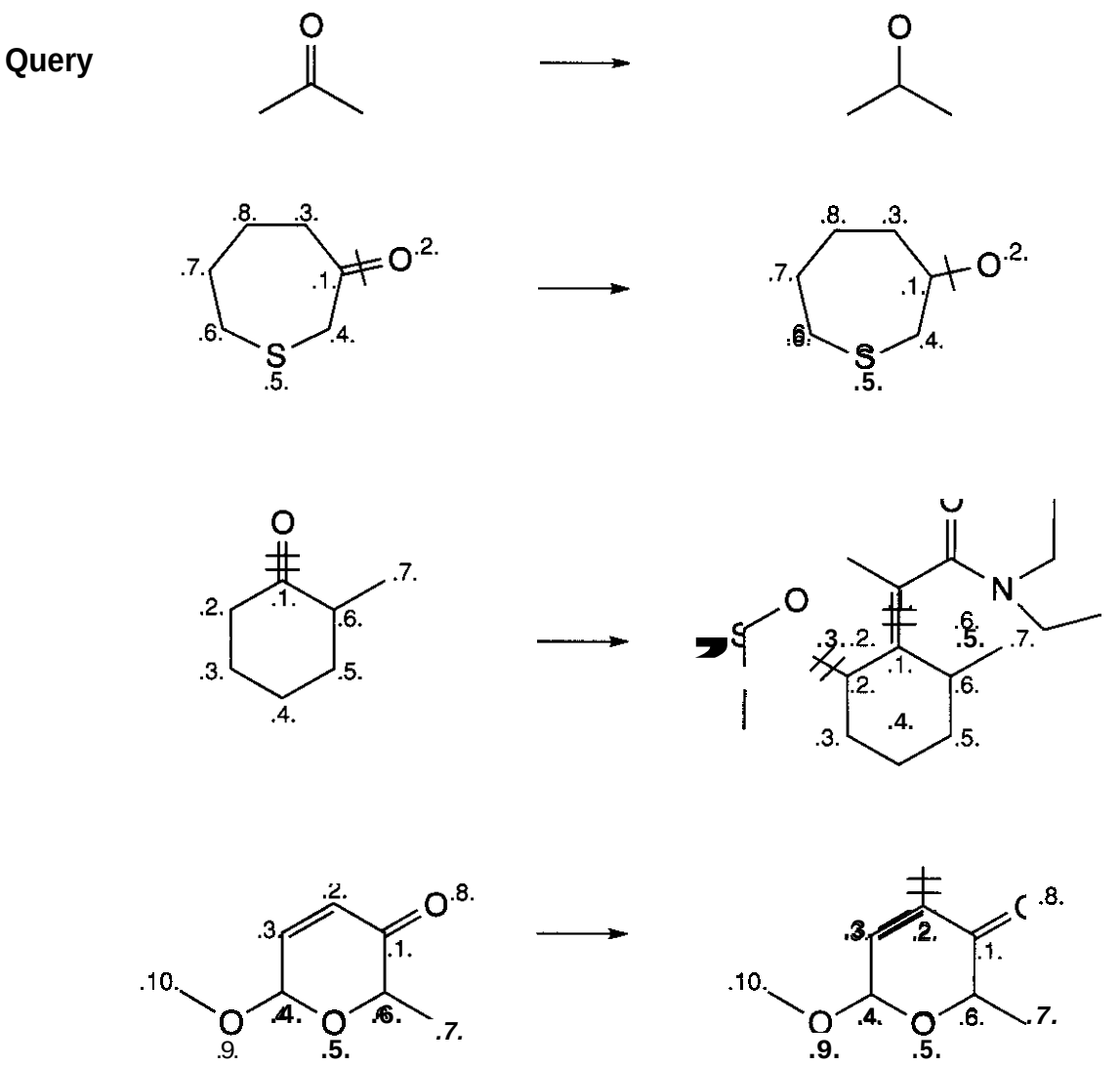


Figure 9.15. Reaction substructure search query and some example hits. If no reacting center or mapping information is used, all three hits are found. If reading bond information is used, hit c is excluded. If both reacting atom and reacting bond information is included, then false hits b and c are excluded.

(89), material safety data sheets (90), and chemical suppliers (91). Some useful tools include the Accord ChemExplorer program, which allows searching word processor documents and files for particular chemical structures, and the CambridgeSoft ChemFinder for Word (92).

3.4 Chemical Information Management Systems and Databases

A number of software and database vendors provide programs and database systems to implement representation, registration, and searching of chemical information in a corporate environment. Some of these vendors have smaller personal chemical database systems that support registration and searching on a personal computer. A handful of academic and public domain systems are also available. Finally, an increasing number of chemical information systems are being made available on the Internet. Some representative systems that are being sold or have been discussed recently in the literature are discussed below.

3.5 Commercial Database Systems for Drug-Sized Molecules

Accelrys. A subsidiary of Pharmacopeia, Inc, Accelrys was originally a provider of molecular modeling software. They recently acquired several companies that provide offerings in the chemical information and bioinformatics areas. The company provides unique databases including several for reactions.

- BioCatalysis—biomolecules as catalysts
- BioSter—pairs of biologically similar structures for bioisosterism applications
- Biotransformations—developed in conjunction with the Royal Society of Chemistry

- Failed Reactions—those that did not proceed as expected
- Metabolism—developed in conjunction with the Royal Society of Chemistry
- Methods in Organic Synthesis--33,000 reactions, Protecting Groups—functional group protection with region/stereoselectivity
- Solid Phase Synthesis—with emphasis on small-molecule and combinatorial chemistry

The chemical information programs provided by Accelrys include several database systems.

- Accord for Excel and Access—relational chemical storage for Microsoft programs
- Accord for Oracle—a chemical data cartridge (see Glossary)
- Accord Database Explorer—to access Accelrys reaction databases
- RS^3 Discovery System—with programs for chemical structure, data management, high-throughput screening, and inventory

Accelrys also provides programs for descriptor calculation, QSAR, and data mining (93).

The Beilstein Database. The Beilstein Database, with over 8 million structures, is the oldest in existence, based on the Beilstein Handbook of Organic Chemistry, and contains data that extend back to 1771. The database is produced by the independent Beilstein Institute (94). Access to the database is either through Beilstein Online, available through STN and Dialog, or through the Web using Crossfire Beilstein, which is marketed by MDL GmbH—formerly Beilstein Inc. (95). Data that are stored include the structure, Beilstein and CAS Registry Numbers, names, formula, preparations, reactions, natural product isolations, and chemical derivatives. Physical properties, if available, are also stored, including optical data, mechanical properties, multicomponent system data, spectral and thermodynamic properties, as well as biological function, ecological data, toxicity, and common uses. Citation data, including author, journal, CODENs, and patent information, are also stored. The data are organized into substance, reaction, and citation contexts, and a user can easily switch from one context to the other. An ACS symposium volume devoted to the Beilstein database has been published (96).

Chemical Abstracts Service. As a division of the American Chemical Society, CAS develops and manages the world's largest databases of chemical structures and reactions.

- CAS Registry-—35 million structures—19.5 million distinct structures—13 million biosequences
- CASREACT—4 million reactions
- CHEMCATS—2.5 million commercially available chemicals
- MARPAT—500,000 searchable Markush structures

The CAS databases are maintained online, with searching allowed on a subscription basis. SciFinder is a client/server application to search CAS databases by author, keyword, exact, and substructure. It includes a "keep me posted" update feature, reaction information back to 1974, nucleotide and protein sequence searching, browsing of 1600 journals, and integration of structure, data, and citation information. STN International is a collection of 200 databases covering chemistry, life sciences, engineering, patents, etc. STN Express provides wizard-assisted searching, and STN on the Web serves as a web client for STN. The ChemPort program provides web access to journals (97).

Daylight Chemical Information Systems, Inc. This company provides numerous third-party databases in the Thor format. These include the following:

- Databases of organic structures: Available Chemicals Directory—250,000 structures, Asinex catalog—115,000 structures, Maybridge catalog—62,000 structures, InfoChem SPRESI'95—2.5 million structures
- Drug and biological databases: BioScreen NP and SC—about 52,000 structures including natural products, Pomona College Medchem—36,000 structures with measured Log*P*, National Cancer Institute—

120,000 structures with cancer/HIV screening data, Derwent World Drug Index WDI—60,000 drugs.

- Toxicity: Aquire—5300 EPA structures with aquatic toxicity, TSCA—100,000 EPA substances
- Reactions: InfoChem ChemReact/ChemSynth—390,000 reactions with 470,000 structures, InfoChem SpresiReact—2.5 million reactions and 1.8 million structures

Software and applications from Daylight include the following (98):

- Numerous toolkits: SMILES, Depict, SMARTS, Fingerprint, Monomer, Thor, Merlin, X-Widgets, Program objects, Remote Access, and Reaction Toolkits (see Glossary)
- Daylight chemistry cartridge for Oracle: DayCart (see Glossary)
- Thor database manager—to build and manage thesaurus-oriented databases
- Merlin searching of structures and data
- Clustering package, with Jarvis-Patrick type cluster analysis
- Rubicon, a program for building 3D models using a distance geometry approach
- PCModels for Log*P* and other physical property calculations
- CombiChem Package to manage high-throughput synthesis
- Reaction Package
- DayCGI—a web development toolkit
- A set of Java tools for chemical information management

Derwent Information. A division of Thomson Scientific, Inc., Derwent is the leading supplier of value-added patent information. The Derwent databases, which are maintained online, include the following:

- Derwent World Patents Index—references to patents, including chemical structure and use patents
- Patents Citations Index—bibliographic and citation data, the Innovations index combined entries from WPI and PCI
- Derwent Selection database—customized subsets of the WPI

The databases are available through several hosting services, including STN, Dialog, and Questel Orbit. User guides for the PCI chemical indexing are available online at Derwent (99). Chemical patents can also be searched using the Merged Markush Service, MicroPatent, and for Japanese patents, the Japanese Patent and Trademark Documents (ISTA) among others (100).

The Gmelin Database. The most comprehensive database of structures, properties, and citations in inorganic and organometallic chemistry is the Gmelin database, based on the Gmelin Handbook of Inorganic and Organometallic Chemistry dating back to 1772. This database includes 1.4 million compounds including coordination compounds, alloys, solid solutions, glasses and ceramics, polymers, and minerals. As such, it is less valuable to drug discovery. The current Gmelin database is owned by the Gesellschaft Deutscher Chemiker and is licensed to MDL GmbH.

MDL Information Systems, Inc. Owned by Elsevier Science Publishing, MDL is a longtime provider of in-house databases and software. Databases include the following:

- Available Chemicals Directory ACD—300,000 structures—reagents and general chemicals, with supplier information
- Bioactivity databases—AIDS database—43,000 structures and data from the National Cancer Institute, Comprehensive Medicinal Chemistry (CMC)—7500 common drug structures, MDL Drug Data Report (MDDR)—120,000 patented drug structures
- Reactions—ChemInform—850,000 reactions and 1.2 million structures, Theilheimer/Chiras/Metalysis—171,000 reactions and 223,000 structures
- Metabolism—Metabolite—53,000 transformations—34,000 structures
- Toxicity—EPA RTECS-based—150,000 structures
- Material safety—OHS Material Safety Data Sheets

Software from MDL includes the ISIS scientific information system (ISIS/Draw, ISIS/Base, and ISIS/Direct), Cheshire for chemical structure manipulation, and Chime and Chemscape for Web access. Combinatorial and high-throughput chemistry programs include Afferent, Central Library, Project Library, Reagent Selector, and Elan. Biological data management programs include Apex and Assay Explorer; literature access through LitLink; reaction access through Reaction Browser/Web; and finally, molecular modeling through Sculpt (101).

Tripos, Inc. Originally the major provider of molecular modeling software, Tripos now offers chemical information content in the form of databases and the tools to manage them. These include the following:

- Several Chapman and Hall databases including ones for organic structures (180,000 structures), inorganic and organometallic structures (40,000 structures), natural products (105,000 structures), and pharmacological agents (22,000 structures)
- The National Cancer Institute structures in a Tripos-compatible format
- The Derwent World Drug Index (60,000 structures)

Chemical information software offered by Tripos now also extends beyond just molecular modeling. Their programs include the following:

- The Unity 3D database system, which features rapid flexible 3D pharmacophore searching
- Concord and Stereoplex—for generating 3D models of database structures including multiple stereochemical isomers
- ChemEnlighten for chemical data mining
- The AUSPYX structure data cartridge for Oracle
- A suite of programs for combinatorial chemistry—Legion to build and store virtual libraries, CombiLibMaker to enumerate structures, Selector to define diversity measures to select diverse subsets of structures, and DiverseSolutions to apply chemical diversity techniques to chemical populations to characterize and populate chemical space (102)

3.6 Sequence and 3D Structure Databases

Sequence databases of biological macromolecules are useful when defining new therapeutic targets. Databases for DNA, RNA, and proteins are available from such sources as the National Center for Biotechnology Information (NCBI) (103) and the European Bioinformatics Institute (104). Numerous online programs and tools are available to researchers to search and align sequences, generate phylogenetic analyses (chemical evolutionary trees), map genes, and predict secondary structure (105). The Protein Data Bank stores the largest collection of crystallographic, NMR, and molecular-modeling derived protein and nucleic acid 3D models (106). The Cambridge Crystallographic Data Center is the primary source for crystal structure data on small molecules, with more than 250,000 entries. The Cambridge Database can be searched using the programs ConQuest for searching, Mercury for structure visualization, and Vista for numerical display and statistical analysis (107).

3.7 In-House Proprietary and Academic Database Systems

Larger chemical and pharmaceutical firms have, over the years, developed in-house systems with capabilities that are specific to the chemist's needs. Today, the costs of developing from scratch and maintaining an in-house system are prohibitive, especially because commercial chemical information systems are highly efficient and customizable. Personal chemical information software is still being developed and reported in the literature. Examples include a relational database patterned after the Upjohn Cousin system (108), and CheD, which is a SQL-based system with a Web client (109).

Commercial personal database systems are available from several vendors, as described above. These products extend the productivity of an individual chemist or a small workgroup, but are not designed for corporate or enterprise applications. Other personal chemical

database programs that are available include ChemFinder from CambridgeSoft, ChemFolder from ACDLabs, ChemWindow from Softshell, and Aura-Mol from Cybula (110).

4 CHEMICAL PROPERTY ESTIMATION SYSTEMS

The design and screening of drug candidates is increasingly being conducted in *silico.* This is made possible by improvements in programs for property calculation and estimation. Here, the term property calculation refers to the generation of some topological (depending only on the 2D structure), topographical (depending on the 3D conformation), or physicochemical property of a molecule—directly from the structure. The term property estimation refers to the generation of some property as a function of other properties—either through a regression equation, a formula, neural network calculation, or some other indirect means.

The distinction between calculation and estimation is important because some properties, like molecular weight, polar surface area, molecular connectivity values, counts of chemical functional groups, partial charges, and other quantum mechanical descriptors, can be calculated precisely and de *novo* from the structure alone. Most of these properties have some fixed definition or algorithm that enables their calculation to be performed unambiguously, with little or no error. What error is present is usually systematic or deterministic. A second class of properties, including Log*P* and other additive-constitutive properties, may be calculated by fragment additivity with various correction terms. These properties differ from de novo properties because they are approximations to the true (sometimes measured) values. Often, there are multiple approaches to their calculation. The errors in the calculation of these properties are statistical or stochastic. A third class of properties includes those that can only be estimated from other properties, using a regression analysis, neural network, or other linear or nonlinear function of variables. The errors in these properties can be complex and difficult to determine. For all these reasons, it is important to carefully consider the use of any given property for drug discovery purposes. Too often, properties are calculated simply because they are available, then used in a QSAR analysis, and possibly applied to future predictions—all without proper consideration of their precision, accuracy, and relevance to the chemical problem.

Given this caveat, it must be noted that there are a multitude of programs available for the calculation of properties of structures. Some programs compute only a single property, like Log*P*. Others calculate a series of values in a given genre of property, like molecular connectivity (111) or BCUT descriptors (112). Still others compute a vast range of properties that include topological, topographical, and physicochemical descriptors alike. It is beyond the scope of this chapter to detail all the programs and vendors that provide property calculation and estimation software. Many of the calculations are provided as part of molecular modeling and QSAR program systems. Some programs and vendors whose products are solely for property calculation are described below.

4.1 Topological Descriptors

Descriptors based on the 2D structure or simply on the connectivity matrix of a structure have long been used for chemical similarity and for property correlations. Because they often lack any relationship to mechanism, these descriptors are best used within a congeneric series or at least a set of similar structures. They may be empirically useful for cluster analysis and chemical library design, because they are effective at representing structure differences and similarities. A few programs and providers of topological descriptors include the following:

- Barnard Chemical Information—provides chemical Fingerprint Generation Pack—to compute fragment-based fingerprints for cluster and diversity analysis (113)
- DRAGON—implementation of about 1400 descriptors of Todeschini and Consonni (114) including constitutional, topological, autocorrelation, geometrical and functional

groups, and including simple molar refractivity, polar surface area, and Moriguchi

- Molconnz-EduSoft LC—provides MOLCONNZ molecular connectivity and electrotopological state descriptors of Kier and Hall (115)

4.2 Physicochemical Descriptors

As a complement to topological descriptors, physicochemical descriptors often have a strong relationship to mechanism, and are widely used in lead optimization and QSAR. The classic triad—steric, electronic, and lipophilic descriptors—are considered the foundation of QSAR, and adequate coverage of the space of these factors is still a major goal in drug discovery. The most common physicochemical descriptor is Log*P*, the 1-octanol/water partition coefficient. Because it is so important, a number of programs and vendors provide Log*P* calculations based on a variety of methods. Many of these programs also compute other physicochemical properties, such as p*K*a and solubility.

- BioByte, Inc.—developers of CLOGP, premier Log*P*, and molar refractivity calculator (116)
- a Syracuse Research Corporation—provide KOWWIN and 11 other structure-based property calculations (117)
- CompuDrug Ltd.—the PALLAS System—including programs for for pK_a, log*P*, log*D* predictions, metabolism and toxicity, and high pressure liquid chromatography (HPLC) development (118)
- a ACDLabs—physicochemical laboratory program calculates p*K*a, Log*P*, log*D*, aqueous solubility, boiling point and vapor pressure, Hammett electronic constants, and a variety of liquid properties (119)
- XLOGP—The Peking University Log*P* calculator—a similar version for proteins is available as PLOGP (120)
- EduSoft LC—provider of Hint!-Lo@—to accompany the HINT! Hydropathic interaction modeling program (121)
- SciVision—provider of software for chemical property calculation, and to estimate QSAR, toxicology, oncology, and other biological properties (122)
- Sirius-Analytical—provider of instruments for Log*P* and p*K*a determination, and the Absolv program to predict physicochemical properties (123)

Most of the commercial molecular modeling systems also provide some property calculations, which range from simply calculating the polar surface area of a structure to a full range of topological and physicochemical descriptors. These may be based on fragment additivity, like most of the programs mentioned above, or they may involve correlations with quantum mechanical or even molecular dynamics-based calculations.

4.3 Absorption, Distribution, Metabolism, and Excretion Properties

Perhaps the most critical aspect of drug development—the behavior of the drug in *vivo*—is also one of the least predictable. Each year, many drug candidates reach the very expensive stage of clinical trials, only to be discontinued because of problems with absorption, distribution, metabolism, or excretion (ADME). Toxicity is often added to this acronym (ADMET), because we increasingly find critical differences in the way children respond versus adults, males versus females, etc. Among the hopes that accompany the deciphering of the human genome, is that drug selection can someday be tailored to an individual's genotype, to lessen the possibility of untoward drug response. For the present, drug designers are focusing increased attention on the prediction of ADME properties, pharmacokinetics, and in vivo behavior. Compared with topological and physicochemical predictions, ADME calculations are still rather crude and approximate. They are usually based on correlations with other properties. And, if the method for obtaining the correlation is a neural network, the predictions may be superior to simpler regression-based approaches, but the interpretability of the model is missing. Some of the programs that are used to predict ADME descriptors include the following:

- LION Bioscience—provider of iDEA, a modular ADME predictive system. The absorption module predicts Caco-2 cell permeation, and performs dose-response modeling of the oral absorption. The metabolism module predicts first-pass effects and models metabolic parameters. Future modules are planned for distribution and elimination (124).
- PASS—prediction of biological activity spectra—compares a test structure with those in a database of about 45,000 structures with known activity/toxicity, using topological descriptors and probability calculations (125).

4.4 Property Calculations Online

Many of the providers of software and databases of chemical properties also provide online calculation services. These include Daylight Chemical Information (126), ACD labs (127), and Syracuse Research (128). In addition, the following sites provide online calculations of a variety of properties:

- Molinspiration—calculates Log*P*, polar surface area, Lipinski Rule-of-5, and a drug-likeness index (129)
- Alogp—VCCLab online Log*P* calculation (130)
- PETRA—The University of Erlangen property calculation routines (131)

a USEPA Suite—including implementations of the Syracuse Research software (132)

5 DATA WAREHOUSES AND DATA MARTS

Even relational databases have their limitations when dealing with huge amounts of data and high user traffic. The burden of continual data updating and registration—activities known as OnLine Transaction Processing (OLTP), can considerably slow down searching and report generation activity—known as OnLine Analytical Processing (OLAP). For this reason, it is becoming common in the database field to build special large databases designed primarily for searching purposes—so-called data warehouses (133). These warehouses have pre-computed indexes and tables that facilitate repeated searching. A special database architecture, known as the star schema, facilitates OLAP activity. In this design, one or more large fact tables contain records of frequently searched data for each object (e.g., structure or reaction) in the database. The fact table is joined to smaller dimension tables that contain the relational information. The schema is known as a star schema because the architecture resembles a many-pointed star, with the fact table at the center, and dimension tables at the ends of the arms. The design of the fact and dimension tables in the warehouse should reflect the searching habits of the users to get the best performance. Probably the first mention of data warehousing in the pharmaceutical area was that of Axel and Song in 1997 (134).

5.1 Data Warehouses of Chemical Information

A data warehouse is designed to consolidate structures and data from many diverse sources, including relational databases, flat databases, and structure and data files. It is considered to be multidimensional. A true chemical data warehouse might contain sequences, 2D structures, 3D models, Markush structures, and reactions—all in the same database. No such commercial database currently exists, but databases presently being developed at MDL and other vendors are examples of chemical data warehouses of structures and their reactions. The MDL data warehouse framework is termed the concordance. The fact table for the concordance is the source table, which brings together structure and reaction identifiers from all the various data sources and links them to the unique structures in the warehouse (Fig. 9.16). Using the concordance, a substructure search can retrieve a set of unique, unduplicated structures, along with pointers to all the relevant identifiers and reactions in the various data sources. Similar pointers exist to the original citations and stored data.

Physicochemical properties that are based solely on the structure are stored in the data warehouse, but properties that are data-source dependent, such as citation or biological activity, are only referenced. A typical use of a chemical warehouse is to search for a set of

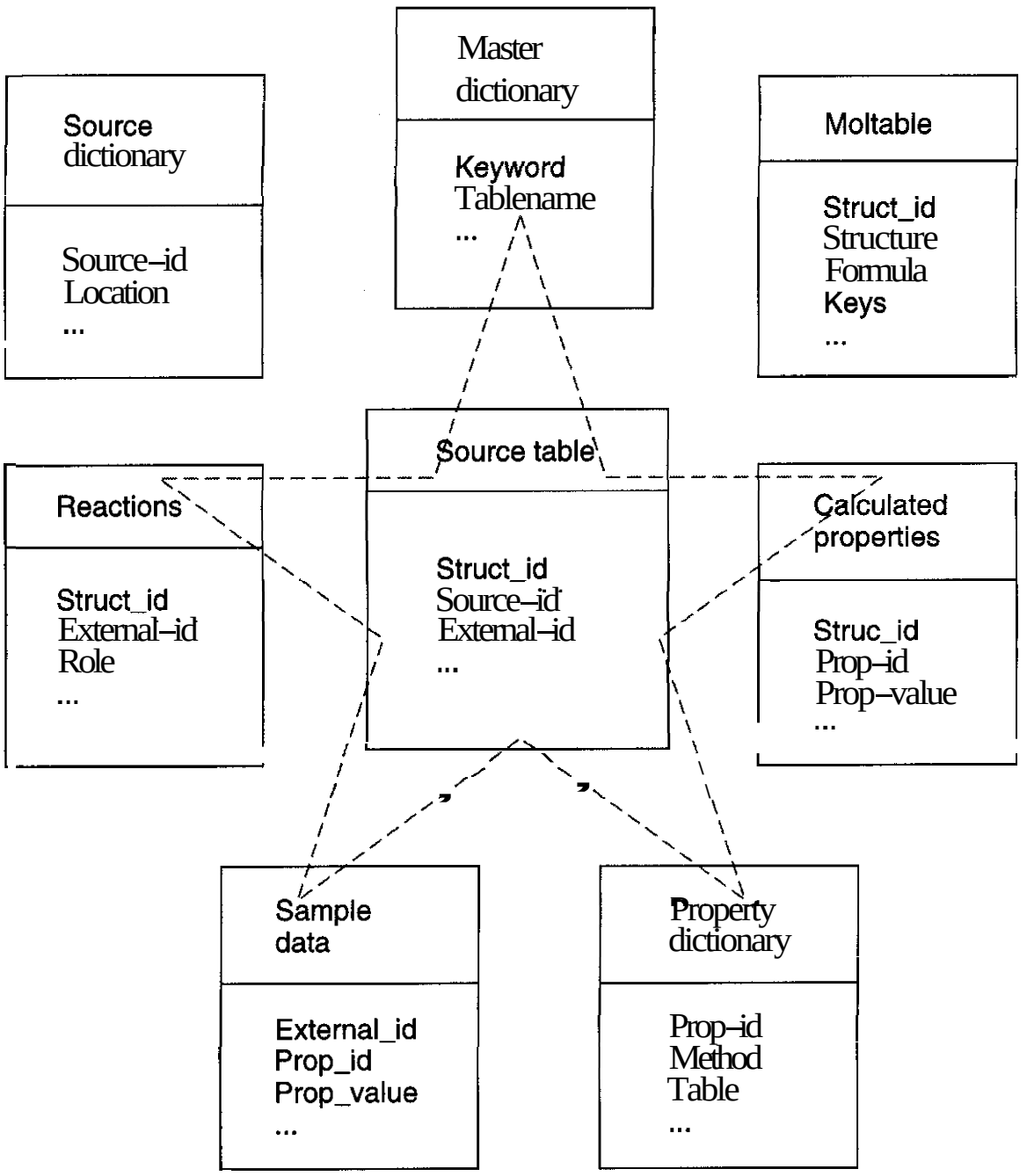


Figure 9.16. Star schema design of a chemical data warehouse. The central source table allows access to the External–ID of every molecule, arranged by source database. These External–ID values can be used to build multidimensional views of the data. For example, to see all the reactions with products that can be found in source database ACD, one would combine data from the source dictionary table (Source_ID for database ACD), the reactions table (Struct_ID, and Role), and moltable (Struct_ID) table, using identifiers (External–ID) from the central source table.

structures that satisfy a search query, then drill-down using a web browser to access the original data sources. In the case of reactions, the user might retrieve and browse a list of reactions that contain the structures that were found in the search. In addition to drill-down, a "hop-into" facility allows passing a set of structures into a search program or web browser that is native to the source database being accessed.

5.2 Data Marts of Chemical Information

For certain purposes, like reagent selection, a data warehouse is too large and comprehensive, and a much smaller database is sufficient. Such a data *mart* has the same architecture as a data warehouse, but it has only a single dimension of structural data—for example, synthetic reagents. The MDL Reagent Selector program is one example of a data mart of reagent structures, with information on their price and availability from various suppliers (Fig. 9.17). It has a fact table that links structures to their identifiers in the various source databases. It stores properties that can be used to filter reagents, such as the molecular weight and Log*P*, and it has pointers to supplier information stored in the MDL Chemical

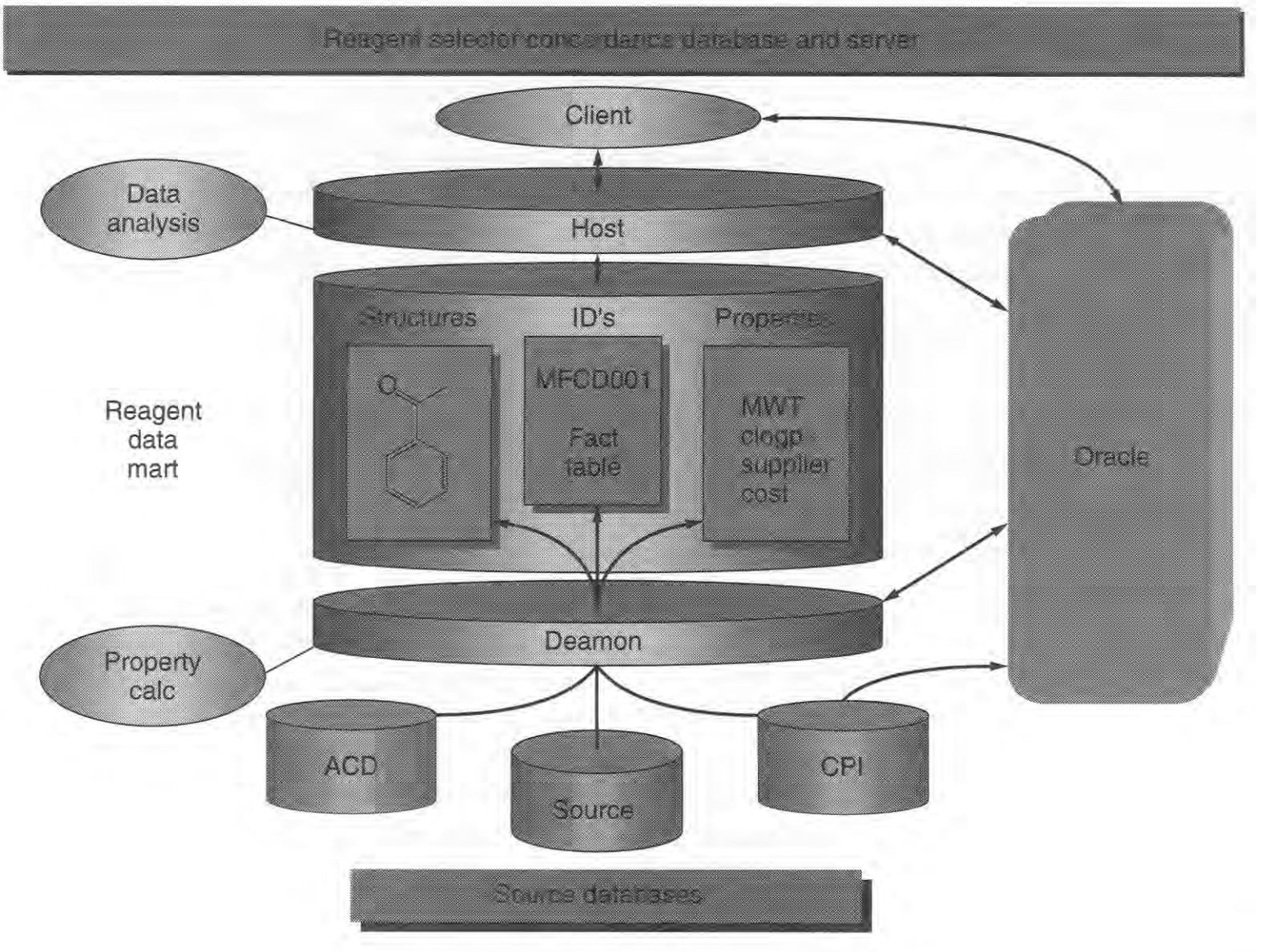


Figure 9.17. Reagent Selector—an example of a chemical data mart. Various components of the system are shown, including the data sources, the daemon program that automatically updates the mart, the concordance database, and the client/server architecture, which is implemented in a three-tier system.

Products Index (CPI) database. To aid in reducing the size of a hit list, a Reagent Selector user can filter reagents and sort on properties, availability, presence or absence of functional groups, etc. (Fig. 9.18). Further list reduction can be achieved by clustering the structures by means of a cluster analysis using substructure keys as descriptors.

An important feature of Reagent Selector is the daemon program, which runs in the background. This agent-like program "awakens" on a fixed schedule and checks the various source databases for new or deleted structures or for changes in the structures and data. If any changes or additions are found, the daemon updates the data mart accordingly, so users will see the latest information when they run searches. Another aspect of chemical warehouses and data marts concerns their physical architecture. It is increasingly common to see so-called "multitier" architectures in which the client program (the "application tier") may be a very "thin" Web client that communicates to a more extensive "middle tier" of programs that serve the immediate needs of the client (see Glossary). Requests for searching and registration, which demand database server resources, are passed from the middle tier to a "database tier" that corresponds mostly to the server part of former client-server architectures. There are many advantages to this arrangement. The programs can be distributed onto different computers to optimize performance of the system. The middle tier can be modified independently to accommodate changes in the client and server. From a development point of view, the various tiers in the architecture can be developed and maintained on their own schedule, with minimal dependence on other components.

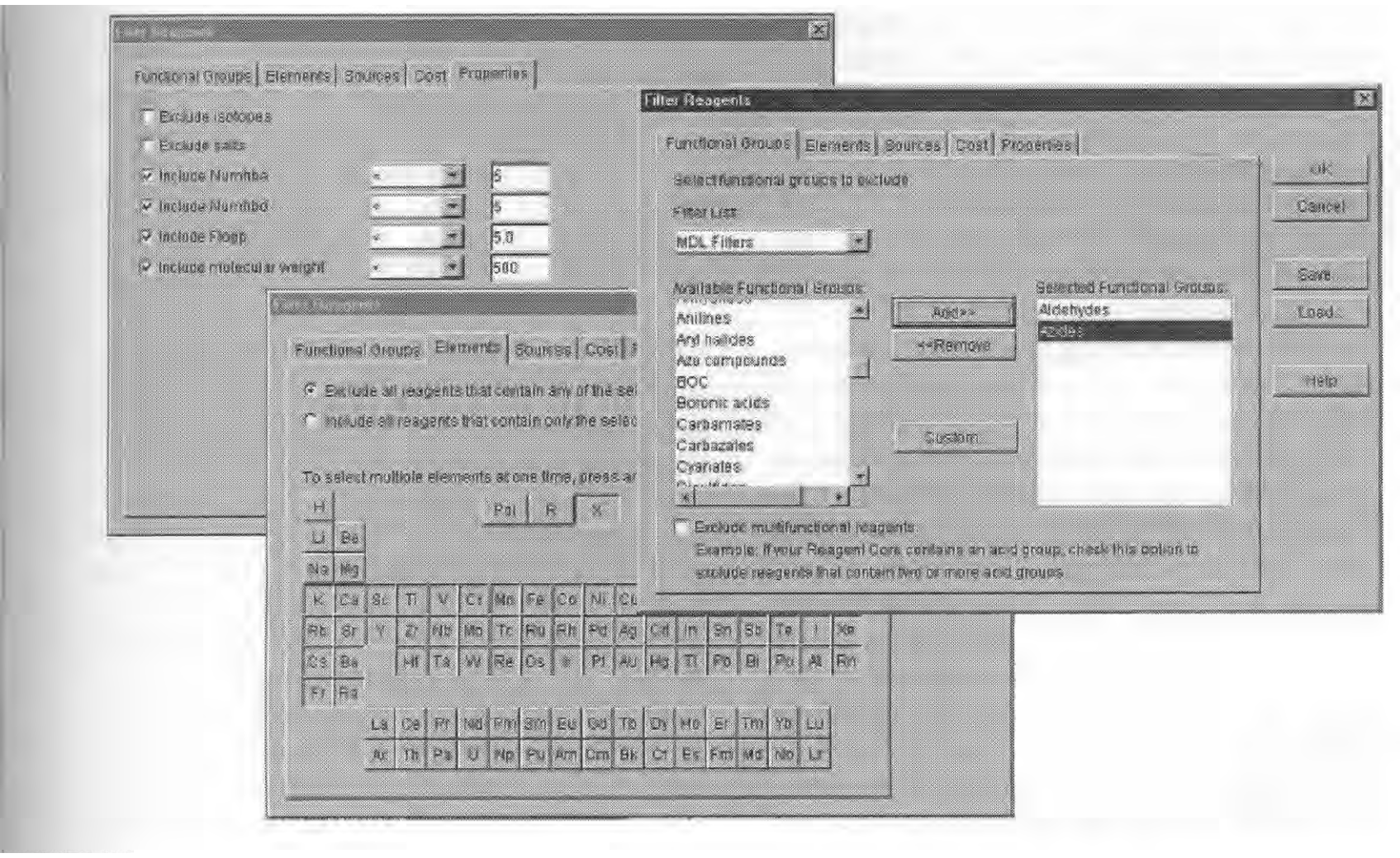


Figure **9.18.** Filtering structures as part of the reagent selection process. The filter criteria include criteria for structure complexity, log*P*, Hdonor/acceptor, molecular weight, formula, and substructures.

6 FUTURE PROSPECTS

It is always difficult to predict the direction of advances in information management. Much of the improvement in chemical structure management and searching has been because of advances in hardware and computer systems. Moore's Law, which states that computer power roughly doubles every 18 months, has held since the 1970s, but threatens to break soon (135). As computer manufacturers hasten to avert the leveling off of computer performance gains, new technologies will surely affect the way chemical information is stored and searched. Based on current trends, it is possible to make some short-term predictions about directions in this field.

- Information: Integration of chemical structure data with other types of data. There is a welcome tendency to treat structures, models, and reactions like other relational data. The biggest advantage of this approach is being able to run integrated searches using structures and data together in search queries. An interesting approach pioneered by the Merck group involves generating fingerprints for key words in documents, then searching for combined structure/document similarity (136). This approach and similar ones will be simplified by increasing integration with relational database systems, as described below.
- Knowledge Discovery in Databases: In the past, dating back to the DENDRAL project (137), attempts to apply artificial intelligence and machine learning to problems in chemistry and drug discovery have gained only moderate acceptance. One problem with expert system approaches has been the small base of information to access. In some cases, this has been a single expert chemist or a handful of example structures. This situation is changing as data accumulates in databases. As Fig. 9.19 shows, *data* can be organized, indexed, and stored in databases to produce *information*. This information, depending on how relevant, unique, and complete it is, can be analyzed and modeled

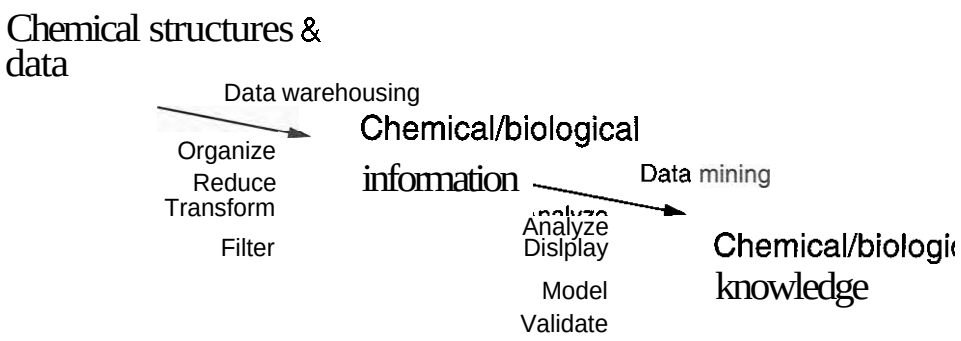


Figure 9.19. Turning chemical data into knowledge. Data becomes organized and indexed to produce information. Mining and analyzing information yields knowledge.

to generate knowledge that might not otherwise have been evident. Once this knowledge is materialized, it can be managed, shared, and deployed for future applications. This process is termed Knowledge Discovery in Databases (KDD), and it is becoming more widely practiced (138).

Data mining is the mechanism by which knowledge is derived from databases. It is generally defined as the extraction of predictive models and associations from large volumes of data using statistical and pattern recognition techniques, usually for some competitive advantage. Data mining is already well established in the marketing, sales, and telecommunications fields (139). Data mining is being used increasingly by scientists, especially in genomics and proteomics. Example applications include the clustering of DNA array data and using database information for protein secondary structure prediction (140). Examples are starting to appear in the field of drug design. Depending on the stage of a drug discovery project, one can mine chemical structure data for diversity, similarity, or specificity, as shown in Fig. 9.20. This figure shows that the lead discovery, refinement, and optimization phases of drug discovery proceed through mini-cycles, each with their own data

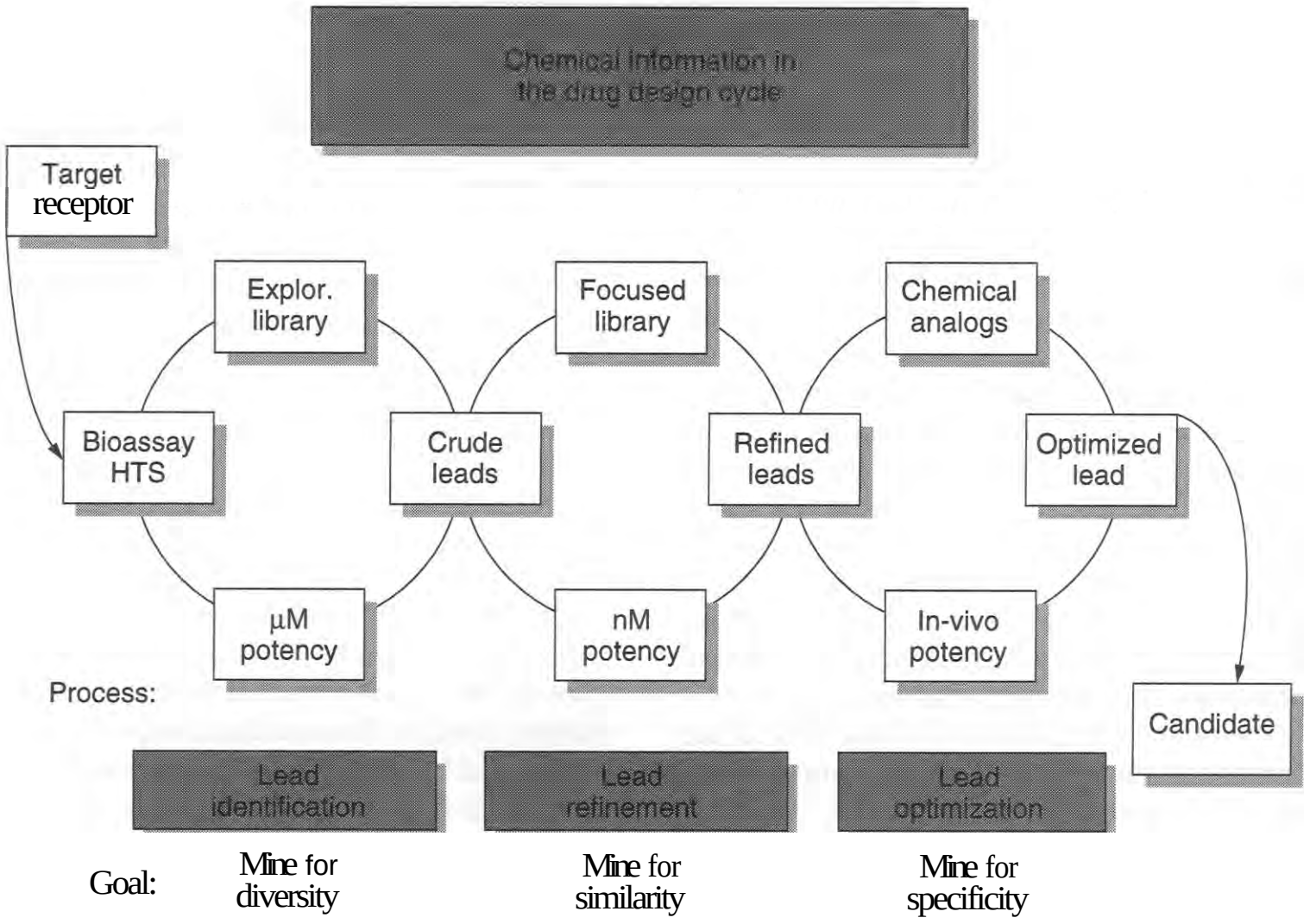


Figure 9.20. Mining chemical information in the drug discovery cycle. Each mini-cycle proceeds until a sufficient number of suitable compounds become available.

mining requirements. So far, data mining has mostly been applied to library design, QSAR, and ADME prediction (141). Whether the techniques become more widely used will depend on the accuracy of predictions made using them, on the availability of convenient software, and most of all, on clean and relevant data.

- Software: Integration with relational systems. A consequence of treating structures as relational data is a tighter integration of once-specialized structure management software techniques with relational database systems. In Oracle, so-called "data cartridges" are being increasingly used to allow a chemist to treat structures like other relational data in a search. Structures, models, and reactions can all be input, registered, and searched using standard SQL to which special operators have been added. SQL stands for Structured Query Language—the standard language for querying relational systems (see Glossary). For example, in the Daylight relational data cartridge, substructure and similarity searches in a reaction database can be conducted directly in SQL as follows:

  Substructure—to find reactions containing benzoic acid as a product:
  SELECT * FROM RXN WHERE CONTAINS (SMILES, '>>O=C(O)c1ccccc1') = 1;
  This statement translates to "Select everything from the table named RXN, where the SMILES field contains the substructure string for benzoic acid as a product". The "=1" clause is an artifact of the data cartridge implementation; it does not necessarily mean that only a single occurrence of the benzoic acid substructure should be found.

  Similarity—to find how many reactions have a solvent that is 80% or more similar to acetic acid:
  SELECT COUNT (*) FROM MEDIUM WHERE SIMILAR (SMILES, 'OC(=O)C', 0.8) = 1;
  This translates to "Tell me the number of rows in the table MEDIUM where the value of the SMILES column in that table has an 80% similarity to acetic acid." Again, the "=1" parameter is an artifact.

This approach greatly simplifies the development of applications. Also, the searches can take advantage of optimization that is built into the relational database system. Fig. 9.21 shows a Web browser that uses the MDL reaction data cartridge to perform structure and reaction searches. The use of a direct searching approach with an object-relational database for combined retrieval of chemical and biological information was reported by Cargill and MacCuish (142). In the field of data mining, the generation, storage, and deployment of predictive models is fully integrated into SQL Server 2000 and Oracle 9i, and this trend will soon extend to other relational database systems (143).

Another advance in chemical information software that promises to have considerable impact on drug discovery is "meta-layer" searching, as described by Hoctor (144). In this approach, queries entered by the chemist are submitted first to a middle-tier search engine, the meta-layer, which automatically and transparently generalizes and transforms the query into several queries. These are then submitted to various databases to retrieve "more of the same" kinds of information. The results are automatically formatted and presented to the chemist in the context of a Web browser (Fig. 9.22). Thus, a name search might get converted automatically to a structure, for substructure or reaction searching, a literature citation search, or a patent search, etc. The linking of searches across indirectly related literature can also be used to generate new knowledge (145).

- Hardware and Operating Systems: The value of parallel and distributed processing was reported early in the development of structure search systems (146). Since then, some commercial products have adopted parallel processing. These mostly involve CPU-intensive searching like conformationally flexible 3D searching and docking. With the exception of such tasks, the speed of most chemical information searching is determined by data input and output (i.e, "I/O

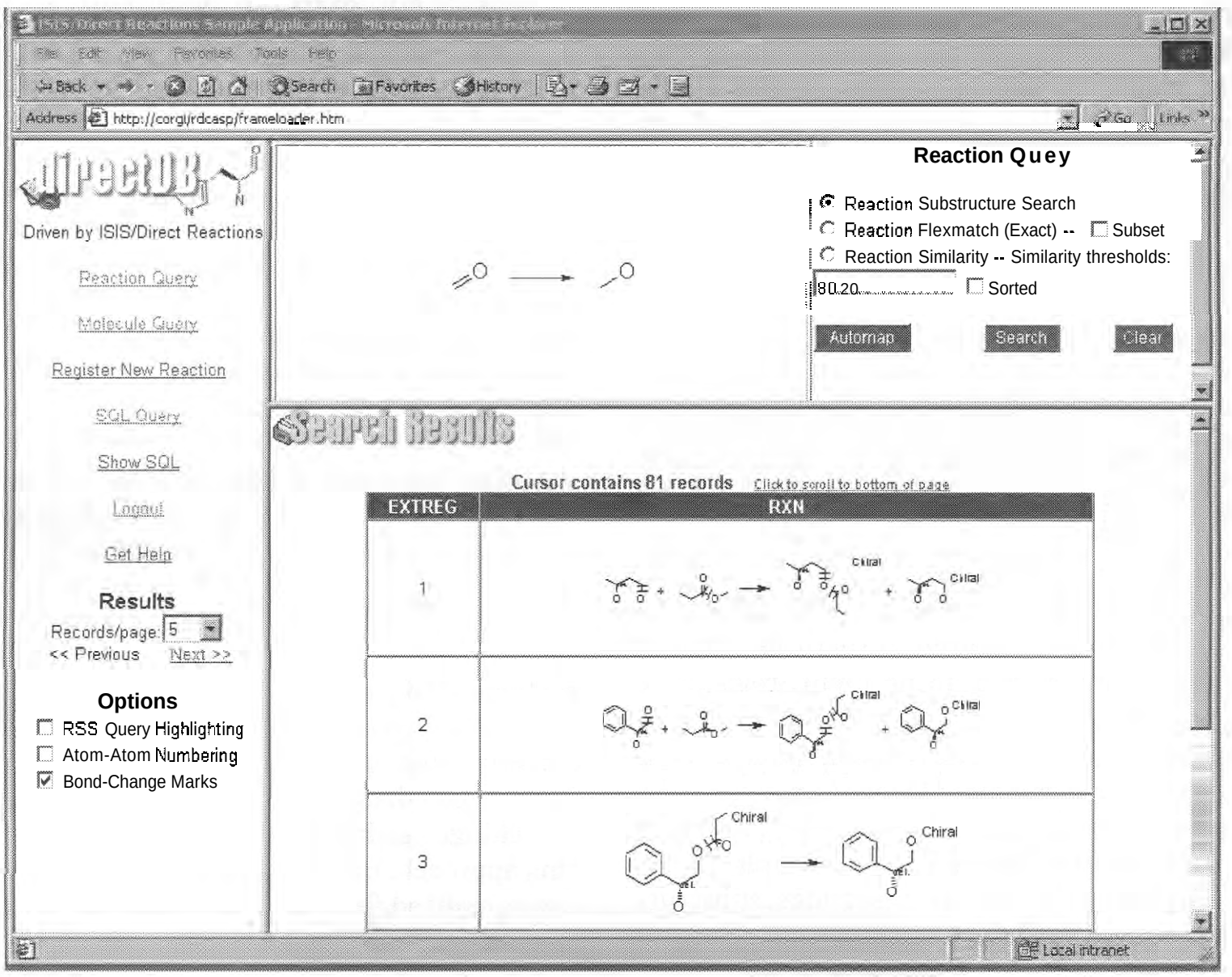


Figure 9.21. Web client for an application that searches a relational reaction database. SQL statements are used to select structures and reactions that satisfy the search query.

bound") because chemical structures are a highly "verbose" type of data. As chemical information systems integrate more with relational systems, they can take advantage of the parallel and distributed processing capability of the relational system. An important development is the "24-7" availability of data in chemical databases (24 h/d, 7 d/wk). This can only be accomplished by distributing and replicating databases across a network.

It would be pure speculation to estimate the impact of changes in hardware and operating systems on chemical information management. Presently, Sun Microsystems is probably the dominant Unix system in chemical database management, largely because of their network presence and their support of Java. Microsoft has released their Windows XP operating system, which merges the Windows 98 and Windows 2000 software streams and will give them continuing dominance in the PC market for a while. Linux is quickly catching on as an inexpensive alternative, and it has a strong foothold in the molecular modeling area, but it requires operating system expertise and lacks a business software base. Small handheld personal data assistants are becoming more capable, and wireless computing is on the rise. The standard desktop computer has at least a 2.0 GHz processor with 512 Mb or more of RAM, about a 30–72 Gb hard drive, and a combination read/write CD with DVD. A relational database of 1 million structures consumes about 3–4 Gb of disk space and can be substructure searched in a few seconds, returning a hit list containing thousands of structures. It takes much longer for a chemist to wade through the results of such a search or to process analytical or bioassay results from a single combinatorial chem-

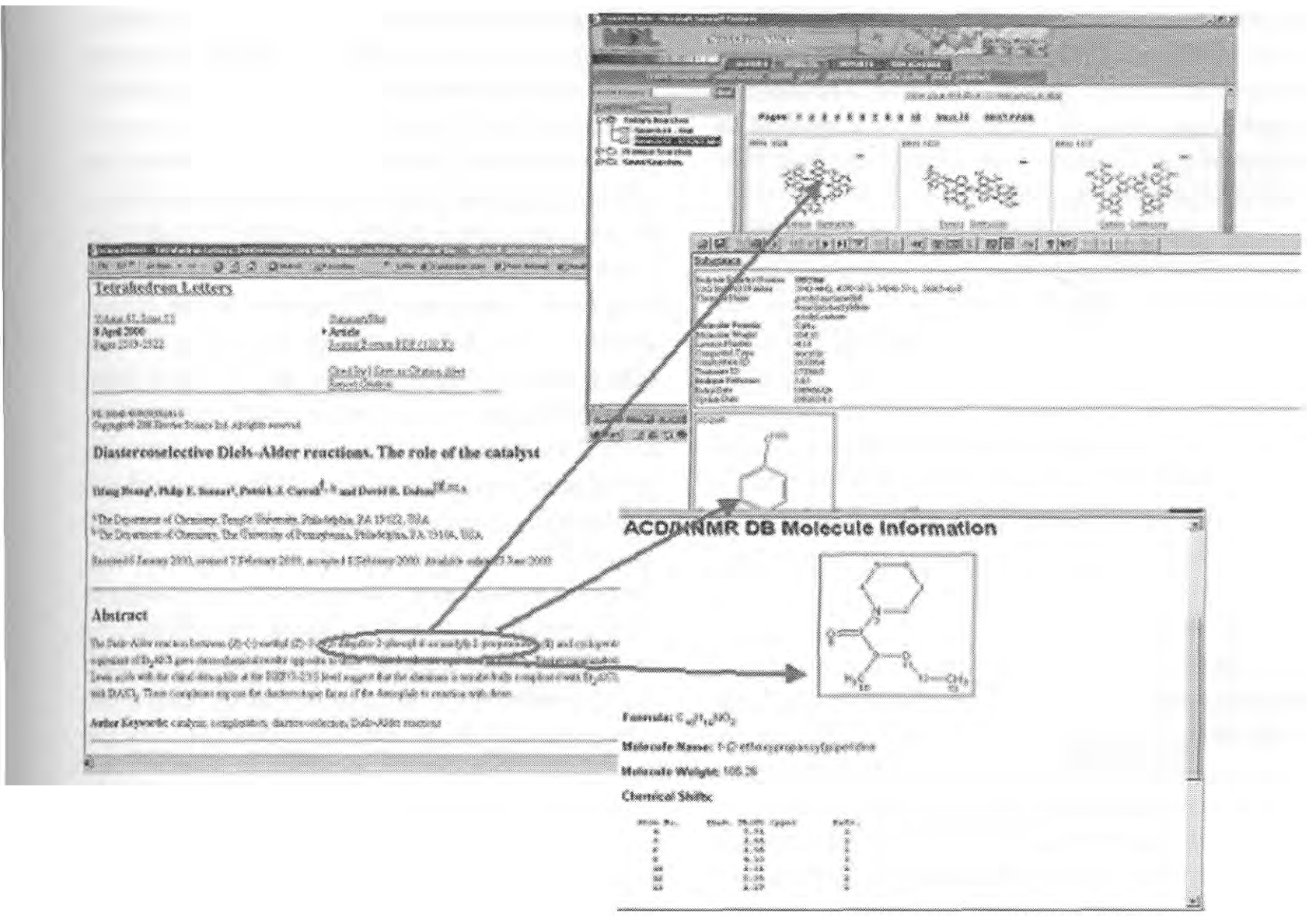


Figure 9.22. Using meta-layer searching to retrieve implicit information. A name search query is converted to a structure, which is then transparently searched to add structure-based search results to the literature citation.

istry experiment than to run most data searches. In light of this, it seems evident that the tools that will succeed are those that will best assist the chemist in extracting relevant, implicit knowledge from the data and deploy that knowledge for future benefit.

7 GLOSSARY OF TERMS

2D Query Feature. A structural feature added to a 2D substructure search query to generalize the query or make it more specific. An example atom query feature would be specifying a list of allowed atoms (Cl, Br, I) or limiting the number of attachments. A bond query feature would be allowing a single or double bond (S/D) or forcing the bond to have a particular stereochemistry. More complex query features can be used to specify which functional groups or substituents are allowed at a given position (*Rgroups*), specifying a range of chain length size (link nodes) and specifying atom, bond, or molecular data queries (Sgroups).

2D Structure. In terms of chemical information, a collection of information about atoms and bonds that can be displayed in a manner such that a chemist would recognize it as a chemical structure. The atom and bond types and connections are usually explicit. The layout of the atoms in the display may be explicit (x,y coordinates) or implicit—determined at the time of display. Hydrogen atoms may be fully or partially suppressed to save storage space.

3D Model. In terms of chemical information, all the information in a 2D structure plus at least one set of 3D atomic coordinates. This is a single conformation of the structure, which is typically a low energy conformation

or even a crystal or spectrometrically determined 3D structure. A 3D model may also contain information about multiple low-energy conformations and atom, bond, and molecular properties such as partial atomic charge, HOMO/LUMO energy, etc.

3D Query Features. Topographical features that relate atoms, bonds, and other 3D features to each other in a pharmacophore or 3D substructure search query. Typical features include (1) objects such as atoms, centers of rings, electron lone pairs, and regions of exclusion and (2) measurements of distance, angle, dihedral angle, radius of exclusion, etc. Measurements often have a range associated with them (e.g., distance between a carbonyl oxygen and a secondary amino nitrogen is 3.4–5.0 Å).

Agent. A computer program that can run autonomously and on a schedule to perform database searches, maintenance, and reporting activities that a chemist would otherwise have to do manually. An example would be an Internet notification service that sends the chemist an e-mail notification whenever a particular database has been updated.

Application Tier. In a multi-tier architecture, the collection of programs that run on the chemist's client or workstation machine. It is the tier of programs "closest" to the chemist in the architecture. Typically this may be a Web client program or other program with a GUI that allows the chemist to interact with the architecture.

Artificial Intelligence. A branch of information science that attempts to use computer programs to perform or simulate human mental activity. Applications in chemistry include perceiving chemical structures, designing structures to fit topological or topographical criteria, designing 'novel' structures, etc. Many of the activities of AI overlap with, or contain elements of pattern recognition and data mining.

ASCII. American Standard Code for Information Interchage—a widely used system of encoding alphanumeric information into eight-bit bitsets (bytes). The expansion of information to include non-English characters requires the use of larger (16- or 32-bit) character sets such as Unicode.

Atom List. In a substructure search query, a list of allowed (or perhaps disallowed) atom types. Often represented within brackets: [Cl,Br,I].

Atom Stereochemistry. Usually refers to tetrahedral stereochemistry at a given atom, which must be a chiral or prochiral center. The stereochemistry may be local (or relative) or global (based on CIP conventions). If it is local, it usually is termed "parity" or some other nonspecific term, to distinguish it from true global stereochemistry (*R*,*S*). Local atom stereochemistry is a property of the atom and its nearest attached atoms. Global atom stereochemistry depends on the entire molecule and the stereochemistry at other chiral atoms. In some systems, the atom stereochemistry is perceived from the drawing of the structure, using "up" (wedged) and "down" (dashed) bond marks as cues. In linear notations, characters in the string can be used to specify the counterclockwise (@) or clockwise (@@) orientation of attachments at a given center.

Atom-Atom Mapping. The procedure of assigning each atom in a substructure query to a given atom in a candidate structure. The assigned structure atoms must match the query atom in all characteristics, including atom type, stereochemistry, charge, attachments, etc. In some structures, a query may map onto the structure in many ways (multiple mappings). Additionally, these mappings may overlap each other in terms of atoms and bonds, or they may be non-overlapping. Some search systems stop after the first mapping, whereas other perform exhaustive mapping, until no further mapping can be found.

Automap. A feature implemented in reaction indexing programs like REACCS, which attempts to automatically "discover" which atoms and bonds are involved in a reaction transformation (the reacting center atoms and bonds). The chemist draws the reaction or the reaction query as a set of reactants leading to a set of products, then invokes the Automap feature, which causes the reacting atoms and bonds to be marked and identified. When reacting center atoms and bonds are specified in both the query and the reactions in the database, reaction substructure searching is faster and gives fewer false hits.

Backtracking. One process that is used in mapping substructure atoms and bonds to the corresponding atoms and bonds in a candidate structure. Given a certain query—say, an amide group [—C(=O)N—], a backtracking algorithm searches first for a carbon atom, then for an oxygen atom, then checks to see if they are doubly-bonded, and finally checks to see if a nitrogen is singly-attached. At any step in the process, if the check fails (e.g., with an ester, the final check would fail), the program "backtracks" to the last successful step and examines another eligible atom or bond. If no eligible atoms or bonds are found at that step, it backtracks to the next previous step in turn. This procedure is guaranteed to find a mapping, but it can be slow, especially with large or highly symmetric queries or structures, where a multitude of similar paths must be examined. Alternative approaches that use an indexed tree can be faster, especially for large databases.

BCUT Descriptors. Descriptors of chemical structure that are derived from an eigen analysis of the connection table of the structure. The class of BCUT descriptor depends on the quantities that are stored in the table (simple connection information versus electronic or steric interaction values). BCUT descriptors have found value in molecular diversity and chemical library design.

Binary Data. Data stored in a file or database that is not chemist-readable, and usually cannot be converted to printable characters. Examples include connection table storage in a database, substructure search keys, and a graphics image of a structure. Note that some other data that is also not chemist-readable, like certain linear notations (e.g., a Chime string), may be made up of printable characters and is not strictly binary data.

Bioinformatics. The application of statistical and mathematical techniques to turn sequence data into useful biological information. The general goal of bioinformatics is to define the structure, location, and function of the proteins and nucleic acids that are the products of the processing of a genome. The application of bioinformatics in drug discovery is primarily the identification of new therapeutic targets.

Biological Data. This includes the results of in vitro and in *vivo* assays, toxicology and metabolism studies, DNA and protein array data, etc. It complements the chemical data, and increasingly, both chemical and biological data are being stored in large corporate relational databases. At any given stage in the drug discovery process, obtaining and analyzing the biological data has traditionally been considered the more complex and rate-limiting step in the process. The application of high-throughput methods to screening and pharmacokinetic analysis is yielding considerable benefit in the collection and processing of biological data.

Bitset. A contiguous set of binary digits (bits, 011) in computer memory. Bitsets are often used in chemical information to store collections of yes/no, presence/absence, and activelinactive responses in a compact form. Bitsets are used to store substructure search keys for each structure (fingerprints), which are used in similarity calculations. Bitset indexes are common features of relational databases, where a collection of bits, one for each structure in the database, can store the presence or absence of a given piece of data for each structure, or, in the case of a substructure search, the a compact representation of the result set from the search. The advantage of bitset representation is that computers can perform very fast logical operations (union and intersection) on bitsets, which enables filtering and subsetting of large lists of structures and data.

BLOB. A Binary Large Object data type. This data type is used in Oracle, for instance, to store large amounts (e.g., up to several Gigabytes) of binary data. Storage of the connection table and all the perceived structural information for a registered structure is one example. Another example is the storage of the entire fastsearch index for a database, which can be accessed as a single object by the Oracle data storage and retrieval routines.

Bond Stereochemistry. This complements the atom and molecule stereochemistry of a structure. A given double bond can be assigned Z or E, or cis or trans stereochemistry based on the attachments. If the stereochemistry is unknown or is a mixture, it can be assigned a value of "either." In a substructure

search, the bond stereochemistry can be specified in the query to limit the scope of the search. In some registration systems, the bond stereochemistry of a given structure is perceived from the input drawing of the structure. In the case of linear notations, it can be specified by characters in the string (e.g., Cl\C=C\Cl specifies ***trans*** dichloroethene).

BRN. The Beilstein Registry Number, which can be used to access structures in the Beilstein database.

Business Rule. An established convention for the representation of data in a given company or laboratory. In the case of chemical structures, an example of a chemical business rule would be "all nitro groups should be drawn as —N(=O)(=O)— and not the charge separated form —N^+(—O^-)(=O)." In the case of biological data, a business rule might enforce the units in which a given piece of test data is reported (e.g., dosage in mmol/kg). Business rules can be enforced by preprocessing data before it enters the database, or in the case of multiple, diverse data sources feeding into a data warehouse or data mart, the data can be transformed to the correct form before storage in the warehouse.

Canonical Numbering. Reordering the numbering of atoms in a structure to a unique order, based on the extended counting of the number of attachments at each center the atom and bond types, etc.

CAS Number. Chemical Abstracts Registry identification number—very widely used to identify chemical structures.

Chem(o)Informatics. By analogy with bioinformatics, this is the application of statistical and mathematical techniques to turn chemical structure data into useful chemical and biological information. It makes use of techniques from statistics, pattern recognition, artificial intelligence, and data mining to derive useful predictive relationships between structures and their biological or physicochemical properties. Broadly considered, cheminformatics also includes the input, storage, management, and searching of chemical structure information.

Chemical Library. A collection of structures, real or virtual, that is the current starting point for high-throughput screening or analysis. A library may be all the structures in a database, or more commonly, a subset of these. It might consist of diverse structure types or it might represent the enumeration of one or a few generic structures. Libraries can be classified according to the stage of discovery—i.e., diverse libraries for lead discovery, focused libraries for lead development, and optimized libraries for lead optimization.

Chemical Space. A loosely defined concept that all the known or possible chemical structures define some multidimensional syace in which the structures are points. Structures that are topologically or topographically similar to each other (i.e., look similar), cluster in chemical space, and by the principle of chemical similarity, should show similar physicochemical and biological properties. This is the basis for diversity analysis of chemical libraries. The challenge is to select or discover properties of the structures that define the chemical space and can be used.

CIP Stereochemistry. Cahn-Ingold-Prelog stereochemistry conventions. An IUPAC approved and widely used set of rules for assigning stereoisomers based on atom and group priorities (see http://www.chem.qmw.ac.uk/iupac/stereo/).

Cleaning and Transforming Data. When importing data from diverse data sources (files, databases, spreadsheets, LIMS systems, etc.) into a database or data warehouse, the data usually needs to be standardized, checked, and sometimes transformed to some common format and content. This allows faster search and retrieval, and serves as a check of data integrity. The rules that define the cleaning/transformation process are often termed "business rules," and in the case of chemical data, they may include checking and modification of chemical structures.

Client-Server Architecture. A computer architecture in which a "server" computer (usually a larger and faster machine at a central location) runs programs that communicate over a network with numerous workstations or "client" machines that reside in offices and laboratories. The server computer performs heavy duty computing tasks such as database searching and molecular and data modeling, in response to commands from the users of the client comvuters. It then communicates the results back to the client machines. There, depending on whether the client is "thick" (a

relatively large and capable application that can display and manipulate data and structures), or "thin" (a small program, possibly running in an Internet browser), the data is displayed, manipulated, and reported. Client-server architecture is two-tier, and is being supplanted by more versatile multi-tier approaches.

Clipping. The computer application of a chemical transformation to a set of structures. One example would be the conversion of a set of o-subsituted phenols to a generic representation with the ortho substituents collected into an Rgroup attached to the parent phenol structure. The reverse process, going from the generic structure to all the specific non-generic structures, is termed ***enumeration.*** Clipping also includes functional group transformations, such as converting a ketone to an alcohol. In the process of cleaning and transforming chemical structure data, clipping may be involved when chemical business rules are "enforced."

CLOB. A Character Large Object data type. This data type is used in Oracle, for instance, to store large amounts (e.g., up to several Gigabytes) of character data. Storage of structures in a relational database in molecule-file format is an example.

Cluster Analysis. The process of discovering "natural" groupings of points in the space of some measurements or descriptors. In chemical information management, one often clusters chemical structures for diversity analysis or to subset the results of a search. Structures are most often clustered using functional group fingerprints as the descriptors. Clustering methods usually consist of either partitioning methods like k-means and Jarvis-Patrick, or hierarchical methods, which may work by successively dividing the points (divisive clustering) or by successively aggregating points (agglomerative clustering). Cluster analysis is an important part of ***unsupervised*** data ***mining*** and pattern recognition.

CML. The Chemical Markup Language. Based on XML and HTML, it provides a standard self-documenting molecule file and information interchange format. Information is described by tags and values. A CML document can be "parsed" by a freely available computer program that can return the structural information on demand.

Combinatorial Chemistry. The application of high-throughput, parallel methods to the synthesis, analysis, screening, and testing of materials. This approach relies on robotics and computer-assisted methods to generate and analyze the results. Synthesis, analysis, and testing of samples occurs in the wells of microtiter plates, which may contain as few as 96 samples or as many as a few thousand. Solid-phase and solution methods are used, and samples may be "one-bead-one-compound" or they may contain mixtures, which require "deconvolution" to determine which component is responsible for observed activity.

CONCORD. Rapid 2D to 3D conversion program introduced by Robert Pearlman's group in 1987. It generates low energy–approximate 3D models from 2D connection tables. It can also do stereo "multiplexing," where multiple configurations of stereochemically ambiguous structures are generated. Marketed by Tripos, Inc.

Concordance. A data warehouse architecture used in MDL relational chemical and reaction databases. The central "fact" table of a concordance has a record for each unique structure in the database, with pointers to the instances of the structure in various "source" databases.

Connection Table. A table or matrix containing topological information about a chemical structure. A structure can be considered a "graph" in 2D space, with atoms as "nodes" and bonds as "edges." The atom connection table has one row and one column for each atom. The diagonal elements of the table are usually the atomic number, and the off-diagonal elements have a zero or null if two atoms are not connected; otherwise they contain the order (1, 2, 3, aromatic, etc.) of the bond connecting the row and column atom. A less common connection table is the ***bond*** connection table, in which the rows and columns are the bonds in the structure, the diagonal elements are the bond order, and the off-diagonal elements contain information about the atoms at the ends of the bonds.

CONVERTER. A rapid 2D to 3D conversion program marketed by Accelrys. It uses a distance geometry approach to modeling, which covers a wider range of conformations than other methods.

CORINA. A rapid 2D to 3D conversion program developed by the Gasteiger research group at the University of Erlangen. It can handle macrocyclic ring structures, which can be problematic in other conversion programs. Chemists can access CORINA online (http://www2.chemie.uni-erlangen.de/software/corina/free_struct.html).

Daemon. From Unix, a program that runs continually as a background process to perform routine functions on demand or on a schedule. In the context of a chemical data warehouse, an example would be a registration program that periodically checks input databases to see if there are any new structures that need to be added to the warehouse. If there are, the daemon extracts the structures from the source databases, transforms and "cleans" them if needed, and registers them to the warehouse.

Data Cartridge. A popular term for user-customizable search "operators" that can be added to the SQL language of a relational database system. An example in chemical information is the addition of a substructure search (SSS) operator to integrate this type of search directly into a relational database search. One advantage of this approach is that the search "strategy" that the relational search program applies can take the complexity of the custom operator into account (the "cost") when performing the various search operations.

Data Compression. The process of transforming a potentially large amount of data into a smaller dataset, in such a way that reversing the transformation results in no loss of information in the original data. A simple example of a compression operation is to replace a string of blanks with a count plus a number that designates the character to be repeated. Compression programs include personal computer programs like PKZIP and WINZIP, and Unix utilities like gunzip. Compression methods are often used before storing data in a database and before transmitting data over a network. When the data is retrieved from the database or received on the network, it must then be "decompressed" by reversing the steps in the compression process. A chemical information example of compression is the conversion of an MDL molfile to a Chime string, which uses ZIP file compression methods.

Data Mart. Typically, a one-dimensional data warehouse—collecting data from multiple sources, extracting/cleaning/transforming/loading (ECTL) the data, and then indexing it for analytical (OLAP) and data mining purposes, to be used by a given group or department. In chemistry, an example would be an inventory database, with structures, location, and purchasing information from many vendors for use by synthetic chemists. Like a data warehouse, a data mart often has a central ***fact*** table with each record containing pointers into other ***dimension*** tables that contain relational data about the items in the fact table. The fact and dimension tables are connected in an organization called the ***star schema,*** which is a common design for data marts and warehouses. Data marts are often subsets of a data warehouse.

Data Mining. The extraction of previously unknown predictive relationships from a large data set or database. Data mining makes use of descriptive ***unsupervised*** methods such as association and cluster analysis, as well as predictive ***supervised*** methods such as decision trees, curve fitting, neural networks, and Bayesian methods. Data mining was once considered "data snooping" and had a poor reputation. The need to analyze huge volumes of data and the success of these methods in marketing and finance have prompted scientists and statisticians to reconsider its use.

Data Warehouse. A large relational database that collects data from multiple diverse sources and organizes it for optimal analytical searching and reporting (OLAP). A data warehouse is a superset of a data mart, containing archival and unchanging data that is important to several groups of researchers (i.e., multidimensional). Data that enters a data warehouse does not usually come from original sources (i.e., chemists, instruments, or assays). It usually comes from intermediate data sources and undergoes cleaning and transformation (ECTL) before registry into the data warehouse. Typically, data is not deleted from

a data warehouse, because historical trends are important. For this reason, the warehouse grows very large over a long period of time, and thus its organization and indexing are crucial considerations. An example in chemistry would be a single database containing structures, models, reactions, and data, all cross-referenced, and used by chemists, biologists, and modelers. Typically, each group would extract their own data mart from the warehouse, containing information relevant to their needs. Data warehouses are often used in decision support systems (DSS) to provide data on which to base important corporate decisions.

Database Tier. In a three-tier programming architecture, the database tier resides on a server computer with access to the databases and the programs that manage them.

Deduplication. When registering into a chemical structure database, the process of finding whether the given structure already exists in the database. This usually involves performing an ***exact match*** search with the given structure as the search query. Note that the definition of exact match may vary with the database, and it may even be configurable. For example, some databases may consider tautomers to be acceptable as exact matches, whereas others may require a more strict definition.

Dimension Tables. In a data mart or warehouse, the dimension tables store non-redundant information about the entries in the ***fact table*** of the database. For the chemical example of an inventory data mart, the fact table stores the various source database identifiers of each unique structure in the data mart. A dimension table of molecular formulas would store the formula for the unique structure in the mart, rather than storing the same formula for each occurrence of that structure in the various source databases.

Drill-Down. Accessing data with increasing amounts of detail. When examining and browsing the results of a database search, a chemist can often request further information about a structure, even though that information was not included in the search. The process of accessing further information, often stored in a hierarchical manner, is termed drill-down. The opposite process, which aggregates data, is termed ***roll-up.***

ECTL. The process of ***Extracting, Cleaning, Transforming,*** and ***Loading*** data into a data mart or data warehouse. The data in a mart or warehouse should be standardized, complete, unambiguous, etc. Raw data from files, instruments, databases, the Internet, etc., must usually be preprocessed before it is "clean" enough to be used in decision making. Structures present special problems because tautomers, isomers, salts, etc. may all represent valid forms. The use of chemical processing languages, which can search for substructures and make modifications of specific atoms and bonds, enables the enforcement of ***chemical business rules*** during the ECTL process.

Encryption. The conversion of data in a readable or decipherable code into another, possibly undecipherable, code. The most common encryption involves sensitive pieces of data like passwords and identification numbers. In chemistry, it is sometimes necessary to encrypt larger pieces of information, such as chemical structures and the results of assays—at least during passage of such information over networks or the Internet. Decryption of the information typically requires one or more ***keys,*** which are often built into the encryption and decryption software.

Enumeration. The systematic substitution of all the Rgroup members in a generic structure, giving each possible specific structure the generic structure represents. If some of the Rgroups are not converted in the process, it is ***termedpartial enumeration.***

Equivalence Class. In the canonicalization of structures that have some element of symmetry, certain atoms that are topologically equivalent may yield the same canonical number. These atoms are considered to be in the same equivalence class. The concept of equivalence class is used, for example, in the Daylight Chemical Information Systems handling of reactions, to examine equivalent atoms when mapping reactant and product atoms.

Exact Match Search. One type of structure searching in which a query molecule is searched for in a database of structures. To exactly match the query, the target structure must be topologically identical and not be a substructure or superstructure of the query.

Extended Stereochemistry. A type of tetrahedral or higher level stereochemistry that appears, for example, in allenes, where the stereochemical center is not a single atom but a system of atoms and bonds that can be conceptually collapsed to a single atom to yield a stereo center.

External Registry Number. A unique "external" identifier assigned to a structure, reaction, 3D model, assay, etc. The external registry number is usually unique across databases, and it can be used as a key to link data from one database or table to another. Any given database may have its own "internal" identifiers that may not be unique across databases.

Fact Table. A central table in a data warehouse whose rows each represent one unit of primary importance in the warehouse. In a chemical warehouse, the rows of the fact table might correspond to unique structures in the database. In a biology data warehouse, each row might correspond to a single experiment. The fields in the fact table are mainly pointers to information stored in other tables, or they contain data that may be repeated in other tables but is stored in the fact table (i.e., denormalized) for rapid access. The fact table connects to other "dimension" tables in the warehouse that contain specific information that is not duplicated.

Fastsearch Index. Term used in MDL databases for a tree-structured index of all the structures in the database. The nodes in the tree represent increasingly complex substructures or properties, where all the structures at or below a given node have in common. The fastsearch index can be very large, but it makes possible very rapid substructure searching.

Field. In database terminology a column of data in a table. Fields are commonly selected in searches of the database, such as "SELECT MOLSTRUCTURE, MOLWEIGHT FROM MOLTABLE WHERE SSS(MOLSTRUCTURE, 'query.mol')=1 AND MOLWEIGHT<500.0". Here, MOLSTRUCTURE and MOLWEIGHT are fields in the MOLTABLE table. SSS is a function that operates on the MOLSTRUCTURE field to find molecules that contain the structure contained in the file query.mol, as a substructure.

Filter. A query or set of criteria designed to select a subset of a given set of data or results. Filters are usually applied to limit the number of hits from a search, or to limit the input to some analysis. Sometimes filters are designed to remove invalid rows of data, to randomly select a subset, or to remove rows based on the values of certain fields. A common application of filtering is in reagent selection, where reactive groups, multi-functional structures, or cost criteria may be applied to the selection of compounds for reactions. A filter can usually be expressed as a SELECT statement in a database search.

Fingerprint. A set of measurements or descriptors, usually binary, that can be used to characterize and identify an object. In the case of a chemical structure, a common fingerprint is a set of substructure keys that represent the presence or absence of specific functional groups. Such keys can be used to compute the topological similarity of the structure to other structures, and can be used as filters in database searching. Other common fingerprints include IR and mass spectral fingerprints, and fingerprints of how a structure behaves in a set of biological assays.

Flat Database or File. Essentially a spreadsheet of data, in which a given row contains all the data about a structure. There are no hierarchical relationships in a flat database. Many older and proprietary structure databases were flat in structure. These are in contrast to relational databases that are more commonly used at present.

Flexmatch Search. Term used in MDL structure searching to allow "relaxed" exact match searching of structures. One can specify, for instance, that everything must match except bond orders, or stereochemistry, or valence at atom centers, etc. By turning on or off various flags, one can for a given structure query, retrieve isomers of various types, salts of a the structure, or instances of the structure that may contain different values of certain types of attached data.

Generic Structure. A structure convention that allows representation of, say, a combinatorial library, as a single, generalized structure. The fixed parts of the structure are represented by the "root" or "parent" structure, and variable parts are represented by

"Rgroups" or "substituent groups" that can each contain multiple substituents or fragments.

Gigabyte. One thousand megabytes, or 10^9 bytes of data. The largest chemical structure databases presently contain a few tens to hundreds of gigabytes of data. A typical structure in a database may require a few thousand bytes of data to store the connection table, coordinates, and other structure-specific data.

Graph and Subgraph Isornorphisrn. In chemistry, the mapping of a structure or substructure query to a target structure. All the atoms and bonds in the query (the nodes or vertices and edges of the "graph" of the query structure) must be mapped to corresponding atoms and bonds in the target structure to generate a hit.

Hash Code. Converting a set of numeric or character properties into a single, mostly unique, number, for the purpose of rapid lookup and retrieval. For example, in the case of chemical structures, it is common to generate and store a hash of the molecular formula, so that when a user requests a formula search, the search query typed by the user is converted to the same hash number, and a single lookup in the index gives all the structures that correspond to the given formula. A hash code is often generated as a linear combination of the possible values of each of the properties (e.g., $n_1P_1+n_2P_2+\cdots$, where the n's are selected such that the products never overlap). If several structures have the same hash code, they are termed "collisions," and typically require further processing—like substructure searching—to differentiate them.

Hierarchical Clustering. One of three main types of clustering applied to chemical structures (hierarchical, partitioning, and fuzzy clustering). In hierarchical clustering, a tree or dendrogram is constructed, with one structure at each of the leaves of the tree. By "trimming" the tree at a given level, one can collect structures into a given number of clusters, such that all the structures in a single cluster have some level of similarity to each other.

High-Throughput Chemistry. Application of parallel processing to the synthesis, analysis, and screening of structures. A subset of high-thoughput chemistry is combinatorial chemistry.

Hit List. Older term for a list of identifiers of structures or other objects obtained from a database search. A more modern term is "result set."

HTML. HyperText Markup Language. The most commonly used specification language of the Internet. Other markup languages of interest in chemistry include XML (extensible Markup Language—information in general), CML (Chemical Markup Language chemical structures), VRML (Virtual Reality Markup Language—3D visualization), and PMML (Predictive Model Markup Languagedata mining).

Index. A secondary data field generated from one or more primary data fields, to enhance the searching and retrieval of the primary data. An index in a chemical database may be a characteristic of the database, such as Oracle indexes, or it may be a chemistry-specific index such as a tree index for substructure searching. Indexes require extra space, and they typically must be created and maintained by some administrative process in the database.

Inventory Data. Typically, information about the availability of reagents for chemical synthesis. This includes the suppliers, package sizes, purity, and cost of commercial reagents, and the location, owner, and availability of in-house reagents. Increasingly, this information is being integrated with chemical structure databases and warehouses and with automated ordering and procurement programs.

Inverted Keys. When substructure search keys are generated for a structure, they may be stored in normal order (where each record represents a structure, and the bits or fields for that structure represent the keys). Alternatively, they may be stored in inverted or pivoted order, where each record represents a given substructure key, and the bits represent structures that have that particular key set. This type of storage benefits key searching, where a user wants all the structures that have a particular key set.

Isomer and Tautorner Search. A search types where bond order, hydrogen counts, certain atom valences, and bond or atom stereochemistry may be allowed to vary from those specified in the query. Such searching allows re-

trieval of keto-enol tautomers, cis-trans isomers, etc. The generalization of the search may be a function of the query of the search process (see Flexmatch search) or both.

Java. Currently the most popular computer language for Internet and middle-tier programming. Java is an object-oriented language developed by Sun Microsystems, that runs on multiple platforms and contains built-in features for networking, database access and security, graphics, etc. Other languages widely used in chemical information programming include C, C++, and Perl. Java and C++ are called "object oriented" languages that focus on the "business objects" of the application—like molecules and reactions. C and Perl are more "procedure oriented" languages that focus on the things the objects do and the process that manages them.

Join. The retrieval of data from more than one table in a relational database, into a single result set. Depending on the structure of the data in the various tables, and the nature of the search query, extra data may need to be added to the result set to fill in certain fields, or the fields may be unpopulated in the result set.

KDD. Knowledge Discovery in Database* application of analysis and data mining techniques to discover "knowledge" that may be implicit but undiscovered in large amounts of data.

Key Field. Field in a table that uniquely identifies rows in the table (primary key) or contributes to uniquely identifying the rows (secondary or composite key), or that connects the given row to data in another table (foreign key). Key fields are usually indexed for rapid lookup and retrieval.

Kmeans Clustering. Type of partitioning cluster analysis in which an object, such as a chemical structure, is placed into one of K clusters, based on how similar the structure is to the average value (or centroid) of each cluster. The average of the cluster may be an actual structure itself, in which case the technique is referred to as K-medoids clustering.

Linear Notation. Representation of a chemical structure using a linear string of numbers and letters. A linear notation is designed to be interpreted by a chemist. Thus, a SMILES string represents a linear notation, while a Chime string represents a compressed notation.

Linux. Microprocessor version of Unix developed by Linus Torvald. Linux is presently used mostly in network servers and in clusters of microcomputers used for large-scale parallel computation. It is gaining status as an alternative to Microsoft and to Unix, for database applications, because Oracle and other vendors provide Linux versions.

Logic in Query Features. Using AND, OR, and NOT as modifiers on the application of query features. For example, one could run a search to select structures that contain "halogen and not primary or secondary amine, or not halogen and any amine." The logic can be a part of the query substructure, as with Markush queries, or it can be part of the SELECT statement.

Markush Structure. Essentially a generic structure, in which a root or parent structure plus Rgroups and their members can represent an entire combinatorial library. Markush structures were developed for patent purposes, and the specification of substituents are often more general than in the case of generic structures or database queries. Markush structures are also used to represent generic reactions, in which the reactants and products are represented by generic structures.

Member. A single substituent or moiety in an Rgroup collection. If R_1 consists of the substituents (Me, Et, Pr, and Ph), these are termed the members of R_1.

Metadata—"Data About Data". In a database, metadata describes the structure, format, storage, access, and various properties of other stored data, or it describes the characteristics of the database itself. Metadata is sometimes stored in tables termed "dictionaries." As an example, chemical metadata might include the tablename, fieldname, and format of the field containing the chemical structures. This metadata might be stored in a master table in the database and be stored as property_name and property–value pairs, which the application that accesses the database can understand.

Middle Tier. Large-scale database applications often consist of three "tiers" of programs: (1) the application tier, which interacts

with the user, (2) the database tier, which interacts with the database management system, and (3) the middle tier, which sits between the user application and the database management. The functions of the middle tier include (1) receiving, transforming, and passing queries and commands from the application tier to the database tier, (2) receiving, consolidating, and formatting data from the database tier and passing it to the user, and (3) performing tasks that may be required by the application but are not available in the database tier, such as managing hit lists and queries. Middle tier programs are often written in Java, a language that runs on many platforms and contains Internet features to communicate with the application tier and database features (JDBC—Java Database Connectivity) to connect to the database tier.

Molecular Connectivity. A class of molecular descriptors derived from the connection table of a structure. For increasing path lengths (1-, 2-, 3-bonds, etc.), the molecular connectivity values are computed as the sum of functions of the connectivity values (number of attachments) of the atoms in the path. Molecular connectivity descriptors can be used to distinguish structures. As such, they can be correlated with physicochemical properties that are functions of structure size, linearity, and degree of branching.

Morgan Algorithm. A procedure for finding a mostly unique (canonical) ordering of atoms in a structure. It involves an iterative process that begins by assigning each atom a score, which is initially computed by counting the number of neighboring atoms attached. In successive iterations, the score at a given atom is computed by summing the previous iteration scores of all the atoms to which it is attached. Eventually, the ***order*** of the scores becomes invariant (i.e., does not change with further iteration). At that point, atoms that are topologically equivalent have the same score. Atoms in the structure are then renumbered by the order of their Morgan number. One advantage of canonical numbering is that a given structure, drawn two different ways (i.e., different ordering of the atoms) can be reduced to the same Morgan numbering, and thus be matched quickly.

Multidimensional Database. A relational database in which multiple general types of data are stored, indexed, and cross-referenced, for use by several different groups. In chemistry, an example would be a database containing reactions, 2D structures, perhaps generic structures or libraries, and 3D models. Such a database would be used by synthetic, chemical informatic, and molecular modeling scientists. A data warehouse is often a multidimensional database, whereas a data mart is usually single-dimensional.

Multi-tier Architecture. An expansion of a client-server architecture to include a middle layer of software. The middle tier may run on a computer different from either the client or server computers. The middle tier isolates the client and server programs, so that changes in either of them do not require corresponding changes in the other. The middle tier acts as to receive, authenticate, and transform data as it passes between client and server computers. To make middle tier software easy to change and maintain, it is often written in Java, a modern object-oriented computer language that is available free on most of today's computer platforms.

Object-Oriented Language. Procedure-oriented languages like Fortran, Basic, and C operate by calling functions or subroutines with "arguments" that tell the program what to do with certain variables in memory—such as calculate the molecular weight of a collection of atoms. In object-oriented (OO) languages like C++ and Java, an object such as a molecule, has a molecular weight "method" that is specific to that object and is stored with the object in memory. In this way, a slightly different object, like a mixture, could have its own molecular weight method—perhaps a weighted average.

Object Relational Database. A relational database in which data can be collected and combined with methods to fit an object oriented model. Searching in the database is conducted in the context of the objects and their methods, rather than the raw data fields and stored procedures. There is usually considerable overhead in building, maintaining, and using an object relational database, so this type of organization has not so far been widely used in chemical or biological databases.

OLAP. OnLine Analytical Processing. An activity that involves routine searching, analysis, and reporting on data stored in a large database. The database, which may have a data mart or data warehouse organization, is optimized for the kinds of searches and reports that it supports. It may not be optimal in organization for transaction processing (OLTP), which may involve registration of small amounts of data on an irregular schedule, or for data mining, which involves the retrieval and analysis of large volumes of data. In chemistry, an example of OLAP might be an inventory application in which the chemist draws in a structure or substructure, runs one or more filters on the resulting result set, and retrieves and prints a report of structures and inventory data.

OLTP. OnLine Transaction Processing. An activity that involves registration, update, or simple searching in a database of transactions. In chemistry, this might be the routine registration of a new structure and analytical data into a chemical database. Such a database is optimized for registration and may not be suitable either for more analytical types of searching and reporting (OLAP) or for data mining.

Parallel Processing. A technique whereby a given computer task is distributed among several central processing units (CPUs). The CPUs may be part of the same computer (e.g., a multiprocessing computer in which several CPUs share common memory and physical devices), or they may consist of several single-processor computers (a "cluster") that are networked to rapidly share information and disk space. In database management, it is becoming increasingly common to have parallel copies (replications) of a given database at several sites, perhaps worldwide. Special database and networking software provides rapid updates of certain information like data and periodic updates of other information like search indexes.

Pattern Recognition. The application of computers to build descriptive or predictive models (i.e., find patterns) of information from input datasets. The techniques of pattern recognition overlap those used in statistics, chemometrics, and data mining, and include data display, description, and reduction, unsupervised methods such as cluster analysis, and supervised methods such as curve fitting and classification. Engineering applications of pattern recognition include recognizing objects in pictures, and character and voice recognition. Chemical applications are found in the fields of drug discovery, analytical chemistry, and chem/bioinformatics.

Petabyte. One thousand terabytes of data (10^{15} bytes). At present, the largest databases of chemical and biological information are gigabytes (10^{9} bytes) in size.

Pharmacophore. The minimum amount of chemical functionality needed in a drug to elicit a given biological response. This functionality is defined in terms of atoms and functional groups and their geometric relationships to each other, including distances, angles, etc. A pharmacophore ***query*** is the representation of a pharmacophore in a format that can be used to search a chemical database for structures that can satisfy the pharmacophore and may elicit the desired response. Pharmacophore searching is usually conducted on a 3D structural database using search software that combines 2D searching with conformational analysis to find structures that can, by rotating about single bonds, adopt a conformation that satisfies the pharmacophore.

Pharmacophore Keys. Originally designed to speed pharmacophore query searching, pharmacophore keys are bitset fingerprints that indicate the presence or absence of given 3- or 4-point pharmacophores in a structure stored in a database. The 3-point pharmacophore keys represent triangular arrangements of atoms and functional groups separated by given distances or distance ranges. The 4-point pharmacophore keys represent tetrahedral arrangements of atoms and functional groups. When a structure is registered into a 3D database, a rapid conformational analysis is performed involving key single bonds in the structure. From the interatomic distance ranges between given atoms and functional groups in the structure, the various bits in the pharmacophore keys are set. These keys can be used as filters in pharmacophore searching, or increasingly, as filters before docking the structures into a known receptor (virtual screening). Pharmacophore keys have also

been used less commonly as descriptors in QSAR and data mining.

Physicochemical Properties. Originally these were just measured properties like melting point, p*K*a, solubility, and octanol/water Log*P*. Increasingly, they are obtained from programs that can calculate them from the 2D or 3D structure. In QSAR, the classical triad of steric, electronic, and lipophilic properties is still widely used, but it has been enhanced to include enery-based descriptors, measures of binding interaction, 3D-QSAR multivariate descriptors (CoMFA), and others. Quantum mechanical calculations are being used increasingly to estimate physicochemical properties. Once calculated, the properties are used to filter structures which may have undesirable ADMET criteria (e.g., the "rule of five"), or they may be used directly in models to estimate the type or level of biological activity (QSAR).

Pivoting Data. Changing data from row to column values or vice versa. This technique can be a very useful tool for summarizing data. One example of pivoting is to convert substructure keys that are stored by structure (with a bit turned on for each key the structure contains), to storage by key (with one bit turned on for each structure that has the given key). Another example is to convert assay data that is stored by structure, to data that is stored by assay. In the process of pivoting data, it is common to consolidate values, for example, converting raw assay results to ED_{50} values, or taking the average of some physicochemical property.

Proteomics. The conversion of protein sequence data into useful biological information. In general, the goal of proteomics is to characterize a gene product—i.e., protein—as to its structure, subcellular location, and function. Additional information includes how a protein interacts ("networks") with other reactions and cell processes.

QSAR. Quantitative structure-activity relationships—the science of deriving quantitative linear or nonlinear mathematical relationships between physicochemical and topological/topographical properties of chemical structures and their biological activity. Originally, regression analysis was the only tool used to derive QSAR equations. More recently, tools such as partial least squares (PLS), neural networks, and a variety of data mining methods like decision trees and support vector machines have come into use.

Reacting Center. An atom in a reactant which is modified during the course of a reaction. Specifying reacting center information when searching for reactions can speed the search and reduce the number of incorrect hits.

Reaction Scheme. A series of one-step reactions that lead from a given reactant to a given product, by way of intermediate steps. A reaction search system should be able to find reactant/product combinations that span several intermediate reactions.

Refine a Search Query. The process of adding or modifying constraints of a search query to reduce or increase the number of hits. Constraints may be added, removed, relaxed, or tightened to achieve the desired search results.

Registry Number. A unique identifier assigned to a chemical structure or other piece of data when it is registered into a database. The registry number may be internal, primarily for use by the database search system, or external, to be used by chemists and to link the data to other databases and files.

Relational Database. A common database. architecture in which related data items are stored in separate tables, accessed by key fields, and indexed for rapid search and retrieval. The dominant relational database systems used in pharmaceutical discovery include Oracle, Microsoft Access and SQL-Server, and IBM DB2.

Result Set. A list of records resulting from a database search. The result set commonly consists of a list of record identifiers (sometimes called a cursor), which can be navigated to select records. In some systems, a result set may also contain related data for each record.

Retrosynthetic Analysis. An approach to computer-assisted synthesis design that starts with the products of a reaction or sequence of reactions and works backwards toward the reactants. An example program that implements retrosynthetic analysis is the LHASA program of E. J. Corey's group.

Rgroup. In a generic or Markush structure, generalized substituents or moieties are given

the representation R_1, R_2, etc. These Rgroups represent collections of specific substituents or moieties (members) that can be replaced at the given position.

Roll-up. The agglomeration, summarization, or consolidation of data into a summary presentation. Roll-up often involves summarizing data at a given level in a data hierarchy. Examples would include the average of several ED_{50} values, or a simple yes/no indication that toxicity data for a given structure exists somewhere in the database.

Root Structure. The invariant portion of a Markush or generic structure. The attached Rgroups contain the substituents that vary from one specific structure to the next. Sometimes termed a parent structure.

ROSDAL. Linear notation scheme devised by the Beilstein Institute. It can contain just connection table information, or it may also contain atom coordinates. Several chemical information systems can convert ROSDAL strings to other structure file formats.

Sgroup Data. In MDL structure storage, the attachment of structure-differentiating data directly to the structure. Such data may relate to the structure as a whole, or to atoms, bonds, fragments, or collections of atoms and bonds. Examples would include atomic partial charges on 3D models or percent composition attached to components of a formulation.

Similarity Search. A type of "fuzzy" structure searching in which molecules are compared with respect to the degree of overlap they share in terms of topological and/or physicochemical properties. Topological descriptors usually consist of substructure keys or fingerprints, in which case a similarity coefficient like the Tanimoto coefficient is computed. In the case of calculated properties, a simple correlation coefficient may be used. The similarity coefficient used in a similarity search can also be used in various types of cluster analysis to group similar structures.

SLN—Sybyl Line Notation. Linear notation used in conjunction with Tripos SPL (Sybyl Programming Language) to manipulate chemical structures. It is similar in syntax to SMILES notation.

SMILES. Simplified Molecular Input Line Entry System—linear notation used in Daylight Chemical Information Systems software and widely supported by other systems.

SQL—Structured Query Language. The standard query specification language for searching relational databases. Most database systems support the SQL standard but then add extensions particular to their implementation.

Star Schema. A standard data warehouse architecture, characterized by Ralph Kimball, in which a central "fact" table is connected to various "dimension" tables.

Structural Data Mining. Application of data mining methodology to chemical structure and reaction databases. Currently in its infancy, it remains to be seen whether a "data snooping" approach to information and knowledge discovery can be as useful in drug discovery as it has proven to be in finance, marketing, and merchandising.

Substance. In some structure databases, an entry that lacks a structure completely (a "nostructure"), is only partially characterized, or is an unspecified mixture of known structures. Substances pose obvious problems in database searching.

Substructure Search (SSS) Keys. Originally developed to facilitate substructure searching, these consist of a string of bits that represent a fingerprint of the structure with respect to either (1) a set of known and defined functional groups (e.g., MDL), or (2) a set of discovered atom-bond paths that the structure contains (e.g., Daylight). In MDL systems, the substructure keys are currently either 166 or 960 bits in length. In Daylight systems, the substructure keys are of varying length and can be "folded" to achieve a higher density of bits turned on. Although SSS keys were originally developed to screen candidates for substructure searching, they are currently used more for similarity calculations.

Substructure Search. Application of "subgraph isomorphism" search to chemical structures. This consists of finding a particular arrangement of atoms and bonds as they are embedded in a chemical structure. The arrangement being searched for is termed the query substructure, the structures being searched are termed the candidates, and any particular structure in that set is termed a target structure. If the query substructure is

found in the target, the target structure is added to the result set (or hit list). In displaying the results of the search, the atoms and bonds in the substructure as mapped onto the hit may be highlighted (shown darkened or in a different color) in the structure display. In general, more than one occurrence of a substructure may be found in a given structure, and substructure mappings may be overlapping or non-overlapping.

Superstructure Search. Modification of substructure search in which the substructure query becomes the target structure, and the target structure in the database becomes the substructure search query. The search finds structures in the database that are substructures of the query. A similar extension to structure similarity searching yields superstructure similarity searching.

Supervised Data Mining. Searching large volumes of data for hidden predictive relationships. Supervised analysis requires one or more "dependent" or response variables, to be predicted from a set of "independent" or predictor variables. The techniques used include various classification methods (decision tree, support vector, Bayesian) and various estimation methods (regression, neural nets).

Tanimoto Coefficient. Standard coefficient for computing the similarity of chemical structures. If structure A has 20 bits turned on in a fingerprint, and structure B has 30 bits turned on, and the two structures have 10 bits in common, the Tanimoto coefficient is $10/(20 + 30 - 2 \times 10)$ or 0.33. Its value can range from 0 (no similarity) to 1.0 (perfect match). Other similarity coefficients are also used, and in some systems (such as MDL) the various bits are weighted inversely according to their occurrence in the database, so that very common substructures do not contribute much to the similarity.

Terabyte. One thousand gigabytes, or 10^{12} bytes of data. The largest relational databases of any kind are currently a few tens of terabytes in size. At present, the largest databases of chemical and biological information are gigabytes (10^9 bytes) in size.

Thick or Thin Client. A thick client architecture is one in which a significant amount of computing is done on the user's workstation. This is appropriate for user-interface-intensive activity like graphics and calculations on individual molecules. The alternative thin-client architecture either does not require much local computing, or it uses a built-in resource like an Internet browser as a client.

Toolkit. A collection of computer routines that each perform one or a small number of information management tasks. The routines are provided as a library and they can be incorporated into custom user-written application programs to carry out tasks that ordinary application programs may not perform. The interface between the toolkit routines and the user-written program is referred to as the Application Programming Interface, or API.

Topographical. Structure data that is based on the connection table and the 3D structure of a molecule. Examples include surface area and volume and pharmacophore distances between atoms.

Topological. Structure data that is based only on the connection table of the structure, without regard to 2D or 3D coordinates of the atoms. Examples include molecular weight and formula, counts of substructures, and indices like molecular connectivity.

Tree. A data structure that is widely used in chemical information storage. Commonly viewed with the root of the tree at the top (or to the left), successive levels of branching lead to the "nodes" of the tree and ultimately to its "leaves" (terminal nodes). Depending on how they split at a node, trees may be binary or n-nary, and depending on how their nodes are distributed, they may be balanced or unbalanced. A tree is usually traversed from the root to the leaves, and this traversal can be depth-first (followinga single path until a leaf node is reached), or breadth-first (looking at all the nodes at a given level). An example of a tree data structure is the fastsearch index used in MDL substructure searching.

Unicode. A 32-bit successor to the ASCII character set. With Unicode, foreign alphabets and special characters can be encoded.

Unix. Widely used operating system for workstations and server computers. Various computer vendors supply their version of Unix, which typically descends from either the Bell Labs or Berkeley versions. A microcomputer version of Unix is Linux, which is rapidly growing in acceptance.

Unsupervised Data Mining. Searching large volumes of data for hidden descriptive relationships. Unlike supervised data mining, no response variables are used. The techniques used include various display and data reduction methods, as well as cluster analysis and association analysis.

VARCHAR, VARCHAR2. SQL data types used to store character data in a relational database system. Storage is limited to about 4000 characters, so larger pieces of data must be stored as CLOB data.

Virtual Screening. Using computer modeling to screen leads for activity. The screening may be through some QSAR or data mining model, which typically requires only 2D structures and data, or it may involve 3D molecular modeling and docking with a known or putative receptor. The speed and increasing accuracy of virtual screening make it a vital step in the drug discovery process.

XML. Extensible Markup Language––a widely used standard for producing self-documenting text. Documents that subscribe to the XML standard can be freely exchanged over networks and between applications, using standard parsing programs to interpret the document. CML is an extension of XML that can be used to transport structures, reactions, and chemical and biological data. A query language, XMLQuery, is being developed to allow searching of XML documents in a manner similar to the use of SQL to search relational databases.

8 ACKNOWLEDGMENTS

Grateful acknowledgement is made to Guenter Grethe, Stephen Heller, Lingran Chen, and Tim Hoctor of MDL Information Systems, Inc. for useful discussions during the preparation of this chapter.

REFERENCES

1. *J. Knight, Nature,* **412, 571 (2001).**
2. *Some examples include Drug Discovery World, New Drugs, Nature Reviews Drug Discovery, Current Drug Discovery and Current Opinion in Drug Discovery and Development.*
3. *The Chemical Heritage Foundation, available online at* http://www.chemheritage.org/HistoricalServices/cheminfo.htm, *accessed on September* **10, 2002.**
4. *W. V. Metanomski, J. Chem. Inf. Comput. Sci.,* **35,** 173–174 **(1995).**
5. *The 5th International Conference on Chemical Structures, June* **6-10, 1999,** *Leeuwenhorst Congress Center, Noordwijkerhout, The Netherlands, available online at http://www.* chemweb.com/conference/5iccs/5iccs.html, *accessed on September* **10, 2002.**
6. *The* **2001** *Gordon Research Conference on Quantitative Structure Activity Relationships, August* **5-10, 2001,** Tilton *School,* Tilton, *NH, available online at http://www.grc.uri.edu/* programs/2001/qsar.htm, *accessed on August* **28, 2001.**
7. *The* **2001** *International Chemical Information Conference,* October **21-24, 2001,** *Nimes, France, available online at http://www.* infonortics.com/chemical/index.html, *accessed on August* **26, 2002.**
8. *H. R. Collier, Chemical Information,* Springer-*Verlag, New York,* **1990;** *H. R. Collier, Ed., Recent Advances in Chemical Information, Royal Society of Chemistry, London,* **1993;** *H. R. Collier, Ed., Further Advances in Chemical Information, Royal Society of Chemistry, London,* **1994.**
9. *Y. C. Martin and P. Willett, Eds., Designing Bioactive Molecules: Three-Dimensional Techniques and Applications, American Chemical Society,* Washington, *DC,* **1998;** *P. Willett, Three-Dimensional Chemical Structure Handling, vol.* **1,** *Wiley, New York,* **1991.**
10. *W. Warr and C.* Suhr, *Chemical Information Management, Wiley, New York,* **1992;** *W. Warr, Ed., Chemical Structures. The International Language of Chemistry,* Springer-Verlag, *New York,* **1988;** *W. Warr, Ed., Chemical Structures 2. The International Language of Chemistry, Springer-Verlag, New York,* **1993.**
11. *G. D.* Wiggins *and K. Emry, Eds., Chemical Information Sources,* McGraw *Hill, New York,* **1991.**
12. *R. E. Maizell, How to Find Chemical Information, Wiley, New York,* **1998.**
13. *J. E.* Ash, *W. A. Warr, and P. Willett, Chemical Structure Systems: Computational Techniques for Representation, Searching, and Process of Structural Information, Ellis Horwood, New York,* **1991.**
14. *D. D. Ridley, Online Searching: A Scientist's Perspective; A Guide for the Chemical and Life Sciences, John Wiley & Sons, Chichester, New York,* **1996.**
15. *Computational Informatics Research Group, available online at* http://www.shef.ac.uk/~is/research/cirg.html, *accessed on June* **19, 2002.**

16. J. Gasteiger Research Group, available online at http://www2.ccc.uni-erlangen.de, accessed on March 6, 2002.
17. G. D. Wiggins, available online at http://www.indiana.edu/~cheminfo/, accessed on September 16, 2002.
18. Cambridge Health Institute, Cheminformatics Glossary, available online at http://www.genomicglossaries.com/default.html, accessed on August 21, 2002.
19. The Chemical Structure Association, available online at http://www.chem-structure.org/, accessed on September 10, 2002.
20. The Ohio State University, Computational Chemistry Listserver, available online at http://www.ccl.net/chemistry/, accessed on June 26, 2002.
21. The Molecular Graphics and Modeling Society, available online at http://www.mgms.org/, accessed on September 10, 2002.
22. The Open Molecule Foundation, available online at http://www.xml-cml.org/, accessed on August 14, 2002.
23. The QSAR and Modeling Society, available online at http://www.ndsu.nodak.edu/qsar_soc/, accessed on September 17, 2002.
24. The Royal Society of Chemistry Chemical Information Group, available online at http://www.rsc.org/lap/rsccom/dab/scaf001.htm, accessed on September 10, 2002.
25. The UK QSAR and Cheminformatics Group, available online at http://www.iainm.demon.co.uk/indexnew.htm, accessed on February 1, 2002.
26. M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, Computer Handling of Chemical Structure Information, MacDonald, London, 1971.
27. W. J. Howe, M. M. Milne, and A. F. Pennell, Retrieval of Medicinal Chemical Information, ACS Symposium Series 84, American Chemical Society, Washington, DC, 1978.
28. Chemical Abstracts Service, available online at http://www.cas.org, accessed on September 16, 2002.
29. W. A. Warr and A. R. Haygarth Jackson, J. Chem. Inf. Comput. Sci., 28, 68–72 (1988).
30. D. E. Meyer, W. A. Warr, and R. Love, Eds, Chemical Structure Software for Personal Computers, American Chemical Society, Washington, DC, 1988.
31. E. K. F. Ahrens in W. A. Warr, Ed., Chemical Structures, Springer-Verlag, Berlin, 1988, pp. 97-111.
32. P. Willett, Ed., Modern Approaches to Chemical Reaction Searching, Gower Press, Brookfield, VT, 1986.
33. S. R. Heller, Ed., The Beilstein Online Database: Implementation, Content, and Retrieval, American Chemical Society, Washington, DC, 1990.
34. R. S. Pearlman, Chem. Design Automation News, 1, 5-7 (1987).
35. Y. C. Martin, M. G. Bures, and P. Willett in K. B. Lipkowitz and D. B. Boyd, Eds., Reviews in Computational Chemistry, VCH Publishers, New York, 1990, pp. 213-264.
36. R. D. Cramer, D. E. Patterson, R. D. Clark, F. Soltanshahi, and M. S. Lawless, J. Chem. Inf. Comput. Sci., 38, 1010-1023 (1998).
37. For discussions of relative stereochemistry, see the documentation for various chemical database systems. For example, MDL stereochemistry is described online at http://www.mdli.com/downloads/literature/ctfile.pdf, and Daylight conventions are described at http://www.daylight.com/release/f_manuals.html. CAS stereochemical conventions have been described in L. M. Staggenborg in H. Colier, Ed., Recent Advances in Chemical Information, Royal Society of Chemistry, Cambridge, UK, 1993, pp. 89-112.
38. R. S. Cahn, C. K. Ingold, and V. Prelog, Angew. Chem., 78, 413-447 (1966); V. Prelog and G. Helmchen, Angew. Chem., 94, 614-631 (1982); D. Seebach and V. Prelog, Angew: Chem., 94, 696-702 (1982).
39. J. Sadowski and J. Gasteiger, Chem. Rev., 93, 2567-2581 (1993).
40. M. F. Lynch, J. M. Barnard, S. M. Welford, J. Chem. Inf. Comput. Sci., 21, 148-150 (1981); J. D. Holliday, G. M. Downs, V. J. Gillet, and M. F. Lynch, J. Chem. Inf. Comput. Sci., 33, 369477 (1993).
41. G. M. Downs and J. M. Barnard, J. Chem. Inf. Comput. Sci., 37, 59-61 (1997).
42. A. J. Gushurst, J. G. Nourse, W. D. Hounshell, B. A. Leland, and D. G. Raich, J. Chem. Inf. Comput. Sci., 31, 447-454 (1991).
43. J. L. Schultz and E. S. Wilks, J. Chem. Inf. Comput. Sci., 37, 436-442 (1997).
44. ChemDraw from CambridgeSoft: http://www.cambridgesoft.com/; ChemWindow from Bio-Rad: http://www.bio-rad.com; ChemSketch and Structure Drawing Applet from ACDLabs: http://www.acdlabs.com; Peter Ertls Java molecular editor: http://www.elsevier.com/inca/homepage/saa/eccc3/paper6/; ISIS/Draw from

MDL: http://www/mdl.com; JChemPaint from the JChemPaint Project: http://sourceforge.net/projects/jchempaint; and Marvin from ChemAxon: http://www.chemaxon.com/marvin/.

45. A comparison of chemical drawing programs, available online at http://dragon.kite.hu/~gundat/rajzprogramok/dprog.html, accessed on September 13,2002.
46. O. F. Giiner, Ed., Pharmacophore Perception, Development, and Use in Drug Design, International university Line, San Diego, CA, 2000.
47. W. J. Wiswesser, J. Chem. *Inf.* Comput. Sci., 25,258–263 (1985).
48. J. M. Barnard in P. V. R. Schleyer, Ed., The Encyclopedia of Computational Chemistry, vol. 4, Wiley, New York, 1999, pp. 2818–2826.
49. D. Weininger, A. Weininger, and J. L. Weininger, *J.* Chem. *Inf.* Comput. Sci., 29, 97–101 (1989); M. A. Siani, D. Weininger, C. A. James, and J. M. Blaney, J. Chem. *Inf.* Comput. Sci., 35,1026–1033 (1995).
50. R. G. A. Bone, M. A. Firth, and R. A. Sykes, J. Chem. *Inf.* Comput. Sci., 39, 846–860 (1999).
51. D. Weininger, J. Chem. *Inf.* Comput. Sci., 30, 237–243 (1990).
52. For example, the University of Erlangen TORVS research team offers 3D models, physicochemical properties, IR and Raman spectral simulation, etc., available online at http://www2.chemie.uni-erlangen.de/services/index.htm, accessed on September 6,2002.
53. Daylight Chemical Information Systems, Inc., Daylight Reactions, available online at http://www.daylight.com/dayhtml/doc/theory/theory.rxn.html, accessed on September 6,2002.
54. W. C. Herndon and S. H. Bertz, J. Comp. Chem., 8, 367–374 (1987); S. Ash, M. A. Cline, R. W. Homer, T. Hurst, and G. B. Smith, J. Chem. *Inf.* Comput. Sci., 37, 71–79 (1997).
55. A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer, J. Chem. *Inf.* Comput. Sci., 32, 244–255 (1992).
56. Peter Murray-Rust and Henry S. Rzepor, Chemical Markup Language, available online at http://www.xml-cml.org, accessed on August 14, 2002.
57. For example, Chemeleon from Exographics Inc., P. O. Box 655, West Milford NJ 07480, and Mol2Mol, http://www.compuchem.com.
58. Babel Chemical File Conversion Program, http://www.smog.com/chem/babel.
59. OpenBabel Project, http://openbabel.sourceforge.net.
60. Converting structures to IUPAC names: Nomenclator from Cheminnovation, http://www.cheminnovation.com; Autonom from MDL GmbH, http://www.beilstein.com/products/autonom.
61. Converting names to structures: NamExpert from Cheminnovation, http://www.cheminnovation.com; CD/Name from ACD Labs, http://www.acdlabs.com.
62. T. R. Hagadone and M. S. Lajiness, Tetrahedron Comp. Methodol., 1, 219–230 (1988).
63. B. D. Christie, B. A. Leland, and J. G. Nourse, J. Chem. *Inf.* Comput. Sci., 33,545–547 (1993).
64. M. Z. Nagy, S. Kuzics, T. Veszpremi, and P. Bruckin W. A. Warr, Ed., Chemical Structures, Springer-Verlag, Berlin, 1988, pp. 127–130.
65. Daylight Thor database system: http://www.daylight.com.
66. Accelrys Accord and RS^3 database systems: http://www.accelrys.com.
67. Tripos Unity database system: http://www.tripos.com.
68. MDL ISIS/Base: http://www.mdli.com
69. Afferent database system for Access: http://www.afferent.com.
70. H. L. Morgan, J. Chem. Doc., 5, 107–113 (1965); J. Figueras, J. Chem. *Inf.* Comput. Sci., 33,717–718 (1993).
71. T. E. Moock, J. G. Nourse, D. Grier, and W. D. Hounshell in W. E. Warr, Ed., Chemical Structures 2, Springer-Verlag, New York, 1993, pp. 303–313; Daylight Theory Manual, Part 7, Reaction Processing, available online at http://www.daylight.com, accessed on September 6, 2002.
72. B. A. Leland, B. D. Christie, J. G. Nourse, D. L. Grier, R. E. Carhart, T. Maffett, S. M. Welford, and D. H. Smith, J. Chem. *Inf.* Comput. Sci., 37, 62–70 (1997).
73. T. E. Moock, D. L. Grier, W. D. Hounshell, G. Grethe, K. Cronin, J. G. Nourse, and J. Theodosiou, Tetrahedron Comput. Methodol., 1, 117–128 (1988).
74. B. Christie and T. Moock in W. E. Warr, Ed., Chemical Structures 2, Springer-Verlag, Berlin, 1993, pp. 469–482.
75. E. J. Corey, Chem. *Soc.* Reu., 17, 111–133 (1988).
76. Program CLASSIFY: InfoChem reaction classification, InfoChem GmbH, Munich, http://www.infochem.de/classify.htm; H. Satoh, O.

Sacher, T. Nakata, L. Chen, J. Gasteiger, and K. Funatsu, J. Chem. *Inf.* Comput. Sci., 38, 210–219 (1998).

77. D. R. Henry and A. G. Ozkabak in P. V. R. Schleyer, Ed., The Encyclopedia of Computational Chemistry, vol. 3, Wiley, New York, 1999, pp. 543652.
78. J. M. Barnard, J. Chem. *Inf.* Comput. Sci., 33, 532–538 (1993).
79. G. W. A. Milne, J. Chem. *Inf.* Comput. Sci., 33, 531–648 (1993).
80. M. G. Hicks and C. Jochum, J. Chem. *Inf. Comput.* Sci., 30,191–199 (1990).
81. P. Willett, J. Chem. *Inf.* Comput. Sci., 38, 983–996 (1998).
82. P. Willett, V. Winterman, and D. Bawden, J. Chem. *Inf.* Comput. Sci., 26, 36–41 (1986).
83. G. Grethe and T. E. Moock, J. Chem. *Inf. Comput.* Sci., 30, 511–520 (1990).
84. W. Fisanick, A. H. Lipkus, and A. Rusinko III, *J.* Chem. *Inf.* Comput. Sci., 34, 130–140 (1994).
85. H. Briem and U. F. Lessel, Perspect. Drug Disc. Des., 20,231–244 (2000).
86. J. E. Mills, C. A. Manryanoff, K. L. Sorgi, R. Stanziione, L. Scott, L. Herring, J. Spink, B. Baughman, and W. Bullock, J. Chem. *Inf.* Comput. Sci., 28,155–159 (1988).
87. J. B. Hendrickson and T. L. Snyder, WebReactions—Organic Reaction Retrieval System, available online at http://webreactions.net,accessed on September 6,2002.
88. Literature searching—for example: Chemical Abstracts Service: http://www.cas.org; LitLink literature server: http://www.litlink.com; Science Citation Index: http://www.isinet.com; Scirus: http://www.scirus.com; Thomson/ISI Web of Science: http://www.isinet.com/isi/products/citation/wos/.
89. H. Tokuno, *J.* Chem. *Inf.* Comput. Sci., 33, 799–804 (1993).
90. For example, Vermont SIRIMSDS: http://www.hazard.com/msds; Cornell University: http://msds.pdc.cornell.edu/msdssrch.asp.
91. For examples, see the Washington University Libraries Chemical Supplier Catalogs: http://library.wustl.edu/subjects/chemistry/chem/chemical_supplier.html.
92. Accord ChemExplorer: http://www.accelrys.com; CambridgeSoft ChemFinder for Word: **http://www.cambridgesoft.com.**
93. Accelrys: http://www.accelrys.com.
94. The Beilstein Institute: http://www.beilstein-institut.de.
95. Beilstein Distribution: http://www.beilstein.com.
96. S. R. Heller, Ed., The Beilstein System: Strategies for Effective Searching, American Chemical Society, Washington, DC, 1997.
97. Chemical Abstracts Service: http://www.cas.org.
98. Daylight Chemical Information Systems, Inc.: http://www.daylight.com.
99. Dement: http://www.derwent.com/userguides/chem.html.
100. Merged Markush Service: http://www.i-mms.com/; MicroPatent: http://www.micropat.com, and ISTA: http://www.paterra.com.
101. MDL Information Systems, Inc: http://www.mdli.com.
102. Tripos, Inc.: http://www.tripos.com.
103. R. M. Woodsmall and D. A. Benson, Bull. Med. Libr. *Assoc.*, 81,282–284 (1993).
104. D. B. Emmert, P. J. Stoehr, G. Stoesser, and G. N. Cameron, Nucleic Acids Res., 22, 3445–3449 (1994).
105. A. D. Baxevanis and B. F. F. Ouellette, *Bioinformatics*—A Practical Guide to *the Analysis* of Genes and Proteins, Wiley-Interscience, New York, 1998.
106. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res., 28, 235–242 (2000).
107. The Cambridge Crystallographic Data Center: http://www.ccdc.cam.ac.uk.
108. T. R. Hagadone and M. W. Schulz, *J.* Chem. *Inf.* Comput. Sci., 35, 879–884 (1995).
109. S. V. Trepalin and A. V. Yarkov, J. Chem. *Inf.* Comput. Sci., 41, 100–107 (2001).
110. CambridgeSoft: http://www.camsoft.com; ACDLabs: http://www.acdlabs.com; Softshell: www.softshell.com; Cybula Inc.: http://www.cybula.com.
111. L. H. Hall and L. B. Kier in D. B. Boyd and K. Lipkowitz, Eds., Reviews of Computational Chemistry, vol. 2., VCH Publishers, New York, 1991, pp. 367–422.
112. R. S. Pearlman and K. M. Smith, Perspect. Drug Disc. Des., 9, 339–353 (1998).
113. J. M. Barnard and G. M. Downs, *J.* Chem. *Inf.* Comput. Sci., 37, 141–142 (1997);http://www.bci.gb.com/.
114. R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, John Wiley & Sons, New York, 2001.

115. L. B. Kier and L. H. Hall, Molecular Structure Description: The Electrotopological State, Academic Press, New York, 1999.
116. Biobyte, Inc: http://www.biobyte.com.
117. Syracuse Research Corporation: http://www.syrres.com.
118. CompuDrug Ltd: http://www.compudrug.com.
119. ACD Labs, Inc: http://www.acdlabs.com.
120. Peking University: http://cheminfo.pku.edu.cn/calculator/xlogp/.
121. EduSoft LC: http://www.eslc.vabiotech.com.
122. SciVision, Inc: http://www.scivision.com/.
123. A. M. Zissimos, M. H. Abraham, M. C. Barker, K. J. Box, and K. Y. Tam, J. Chem. Soc., Perkin Trans., 2, 1–9 (2002).
124. LION Bioscience, Inc: http://www.lionbioscience.com/.
125. V. V. Poroikov, D. A. Filimonov, V. B. Yu, A. A. Lagunin, A. Kos, J. Chem. Inf. Comput. Sci., 40,1349–1355 (2000).
126. Daylight Chemical Information online calculator: http://www.daylight.com/cgi-bin/contrib/pcmodels.cgi.
127. ACD Labs online calculator: http://www2.acdlabs.com/ilab/.
128. Syracuse Research online calculator: http://esc.syrres.com/interkow/kowdemo.htm.
129. Molinspiration: http://www.molinspiration.com/.
130. Alogp—VCCLab online LogP calculation: http://vcclab.org/lab.
131. PETRA: http://www2.chemie.uni-erlangen.de/services/petra/.
132. USEPA Suite: http://www.epa.gov/opptintr/exposure/docs/episuite.htm.
133. R. Kimball and M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd ed., Wiley, New York, 2002.
134. M. G. Axel and I.-Y. Song, "Data Warehouse Design for Pharmaceutical Drug Discovery Research," in Proceedings of the 8th International Workshop in Database and Expert Systems Applications, IEEE Press, Piscataway, NJ, September 1997.
135. D. Normile, Science, 293, 787 (2001).
136. S. B. Singh, R. P. Sheridan, E. M. Fluder, and R. D. Hull, J. Med. Chem., 44, 1564–1575 (2001).
137. R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, Applications of Artificial Intelligence for Chemistry: The DENDRAL Project, McGraw-Hill, New York, 1980.
138. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uythurusamy, Eds., Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, 1996.
139. M. J. A. Berry and G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, Wiley, New York, 1997.
140. M. J. Zaki, H. T. T. Toivonen, and J. T. L. Wang, Eds., Proceedings of BIOKDD'01: Workshop on Data Mining in Bioinformatics, 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'01, San Francisco, CA, Association for Computing Machinery, New York, August 26, 2001.
141. R. W. Snyder, "Symposium on Structure-Based Data Mining", Abstracts, American Chemical Society 221st National Meeting, San Diego, CA, April 1–5, 2001, American Chemical Society, Washington, DC, 2001.
142. J. F. Cargill and N. E. MacCuish, Drug Discov. Today, 3,547–551 (1998).
143. B. de Ville, Microsof? Data Mining, Digital Press, Boston, MA, 2001.
144. T. Hoctor, "Linking Context-similar Information", Abstracts, American Chemical Society 222nd National Meeting, Chicago, IL, August 2630, 2001, American Chemical Society, Washington, DC, 2001.
145. D. R. Swanson and N. R. Smalheiser, Artificial Intelligence, 91, 183–203 (1997).
146. N. Farmer, J. Amoss, W. Farel, J. Fehribach, and C. Eidner in W. A. Warr, Ed., Chemical Structures, Springer-Verlag, Berlin, 1988, pp. 283–295.