"Making a Sound" in Chemical Information: The Importance of a Structure Editor in Information Retrieval

Judith Currano
Graduate School of Library and Information Science
University of Illinois, Urbana-Champaign

The importance of chemical structure in information retrieval, and conventions for representing three-dimensional structures on a flat surface form the background for this study of methods of retrieving structures in electronic databases. Storage is a particular problem. Fragmentation codes are ambiguous, (incapable of specifying the order of functional groups in a molecule), inflexible, and require a large amount of user training. Linear notation, while being unambiguous and more flexible, still involves a fair amount of user training. Connection tables, the easiest method to use, accurately specify the order and connectivity of atoms but omit the important feature of stereochemistry. Thus far, object-oriented programming offers the only method of specifying stereochemistry, although a recommendation is made for a text-based system that combines connection tables with the Cahn-Ingold-Prelog rules for assigning stereochemistry, creating a more specific and universal system. This article discusses the problem of using chemical structure for information retrieval and the difficulty of representing chemical structure in a machine-readable form.

INTRODUCTION

"If a tree falls in a forest and nobody is around to hear it, does it still make a noise?" This question often seems hackneyed, but it contains a legitimate query. One common answer to the question is that it does not matter whether or not the tree makes a noise since there is no person around for the noise to disturb. A similar thing can be said about most areas of science. If a scientific discovery is made and nobody is told about it, it does not matter if the discovery has "made a noise."

Science builds upon previous research, and it progresses much more quickly if the scientists do not have to continually reinvent the wheel. Since scientific communication is the key to new research, it is very interesting that one of the greatest challenges for scientists has always been the means of communication of ideas and research to their colleagues. The scientific paper, generally published in a peer review journal, has gradually become the most prevalent means of disseminating scientific information. The advent of electronic journals and other electronic media makes it even easier for the scientists to present their work.

Unfortunately, the improved technological facilities available to scientists carry a price along with their benefits. The increased number of journals and publications make information retrieval more and more difficult. Often, searching for that one article that can make or break an experiment is like searching for one particular tree in a vast forest. Without forms of information retrieval suited to the current technical environment, an individual scientist's paper becomes that tree, and it is doubtful if anyone will ever hear it if it falls.

CHEMICAL STRUCTURE AS A COMMUNICATION AID

Organic chemistry is a very quickly growing field. With advances in modern medicine and the growing demand for pharmaceutical products, synthetic organic chemistry, biochemistry, inorganic chemistry, and related branches of chemistry have increased in importance while developing in leaps and bounds. Since groundbreaking work is continually advancing the field, it is imperative for chemists to have access to the most current publications, while synthetic considerations mandate access to the oldest publications as well. It is, therefore, necessary for chemists to have an information retrieval system that addresses these needs and allows them to specify the exact properties of the substance or reaction they wish to research. Ash, Chubb, Ward, Welford, and Willett (1985) explain that "for centuries, chemists have sought means to represent the nature of chemical substances,

Katharine Sharp Review [ISSN 1083-5261] No. 8, Summer 1999 [http://www.lis.uiuc.edu/review/8/currano.pdf]

and the long history of chemical investigation is reflected in part in the multitude and diversity of notational systems" (p. 128).

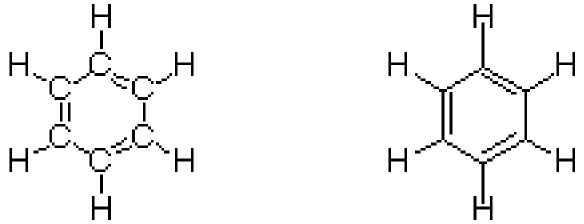
From IUPAC (International Union of Pure and Applied Chemistry) nomenclature to common names, and from complex descriptions of reactions to simply calling the reactions by the names of their creator, there are many different methods for identifying and specifying chemicals. However, a challenge arises in that the areas of chemistry that are growing the most quickly are also among the most visually oriented. Molecules exist in three-dimensional space and as such, have properties that are difficult to accurately describe on a two-dimensional page. Often, the names of large or complex organic compounds are equally large and complex, and while the vocabulary accurately describes the substituents and their order in the molecule, it does not always show their orientation in space. In order to do this, it is necessary to resort to a three-dimensional sketch of the molecule in question.

No single method, therefore, has been as effective and universally comprehensible as the use of the chemical structure in organic chemistry. It is much easier simply to draw the structure in question than to write out its long and bulky name and then explain the orientation of each ambiguous atom. Likewise, when describing a particular reaction between several of these compounds, a carefully drawn reaction mechanism can save paragraphs of descriptive language. For this reason, chemists have taken to relying heavily upon molecular sketches, often numbering them within a document and, when it becomes necessary to name the molecules in question in the body of the article, addressing them simply by their numbers.

OVERVIEW OF CHEMICAL STRUCTURE

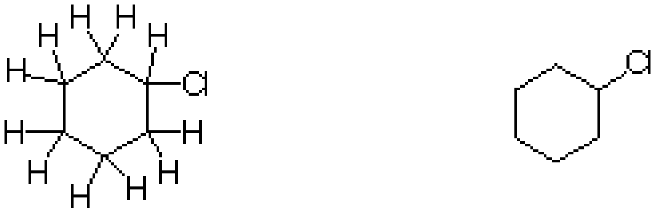
In order to understand clearly the importance of chemical structure to chemical communication and information retrieval, it is necessary to understand a few things about organic chemistry and chemical structure in general. Organic chemistry is the chemistry of carbon. The most common atoms found in organic molecules are carbon, hydrogen, oxygen, sulfur, silicon, nitrogen, and halogens, such as fluorine, chlorine, bromine, and iodine. The electronic structure of each of these atoms has an incomplete outermost, or valence, shell, which causes the atoms to form covalent bonds in an effort to complete the shells and increase stability. It is the number of electrons missing from the valence shell that dictates the number of bonds each atom will form. For example, carbon needs four electrons to complete its valence shell, so it forms four bonds. These four bonds can be any combination of single bonds, represented by a single line between atoms, double bonds, the chemical equivalent of two single bonds and represented by a double line between atoms, and triple bonds, represented by three lines. There is no quadruple bond.

Since organic chemicals are composed primarily of carbon and hydrogen, chemists have adopted certain shorthand notations in the drawing of chemical structures. Instead of writing out each carbon in the structure (Figure 1a), they choose to represent a carbon as a joint between two or more lines (Figure 1b). If a non-carbon, or heteroatom, appears in the structure, the atomic symbol of that atom is included in the structural representation.



FIGURES 1a. and 1b. A chemical structure (1a) and its "shorthand" equivalent (1b).

The case of the hydrogens is a bit different. Hydrogen bonds to only one atom and does not have any special bonding characteristics. Since the number of bonds formed by each component atom in an organic molecule is known, the convention has become to depict only bonds between the larger atoms and leave off the hydrogens. The two drawings in Figure 2 are therefore equivalent, and it takes much less time to draw the second than the first.



FIGURES 2a. and 2b. Two representations of a molecule depicting, (2a) all of the component atoms, and (2b) only the non-hydrogen component atom.

Finally, it is important to note that molecules exist in three-dimensional space. If an atom bonds with two other atoms, the three together form an angle of a particular measurement. For example, when carbon forms two single bonds with other atoms, the angle produced measures approximately 109 degrees. However, if the carbon forms two double bonds to other atoms, as in carbon dioxide, the resulting molecule is linear because the bond angle for double bonds to carbon is 180 degrees (Figure 3).

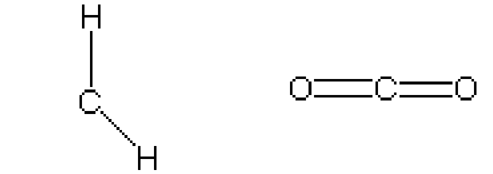


FIGURE 3.

As a result of these differing bond angles, even molecules that can be represented by flat shapes such as hexagons and pentagons are not completely flat. This is important to note because it is the three-dimensionality of the molecules that gives rise to a critical, yet incredibly difficult property to represent: stereochemistry. If cyclohexane (six carbons in a ring) has one of its hydrogens replaced by a chlorine atom, there are two possible configurations of the atoms. The molecule will have different reactivity if the chlorine is sticking straight up in the air than if it is sticking out to the side (Figure 4).

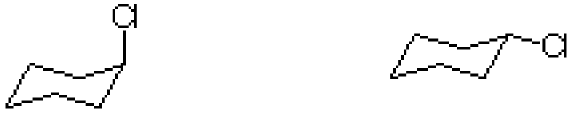


FIGURE 4. A substituted cyclohexane molecule possesses different reactivity depending on the orientation of the nonhydrogen substituent.

Probably the most important use of stereochemistry is in determining whether or not two molecules composed of identical atoms are actually the same substance. In looking at the substituted carbons in Figure 5, it is clear that the empirical formulae for the two are identical: CHFIBr. An examination of the orientation in space of the two molecules' atoms reveals that they are like a person's right and left hands; they are mirror images of one another but cannot be superimposed upon one another. Such molecules are called enantiomers, and while they may have the same solubilities and melting or boiling points, they are not molecules of the same substance. This type of determination is especially important in drug manufacturing since receptors in the body react differently to each member of an enantiomeric pair. For instance, one member of a pair — the "right hand" — might be a useful drug, while the other — the "left hand" — could be highly toxic.

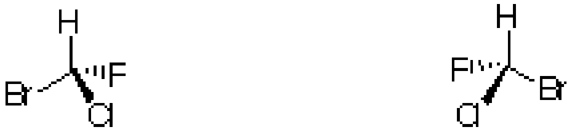


FIGURE 5. (Note: these molecules have been created for the purposes of demonstration and may not exist in nature.)

TRADITIONAL METHODS OF STRUCTURAL REPRESENTATION

In order to better describe the molecules that they are synthesizing or reacting, chemists use many different methods of structural representation. Using new computer software, chemists have been able to perform complex simulations of crystal structures, representations of large bio-organic or inorganic molecules, or complexes characterized by accurate relative atom size, bond lengths, bond angles, and three-dimensional atomic orientation. Less complex software allows chemists to draw various projections of chemical structures, including Newman projections, Fisher projections, sawhorse projections, and the "two-dimensional" molecular representations common to scientific publications. Each type of projection fulfills different needs.

The Fisher projection, shown in Figure 6, is probably the simplest of all of the projections to draw. It is used most frequently in drawing sugars and other hydrocarbons, and presents the molecule's carbon backbone and the substituents attached to the backbone. It is, in essence, an aerial view of the molecule, looking straight down on the carbon backbone. Using a Fisher projection, it is very easy to see enantiomeric differences between two different molecules. However, Fisher projections are not useful for cyclic structures or more complex structures with side chains attached to the main backbone.

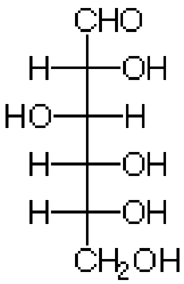
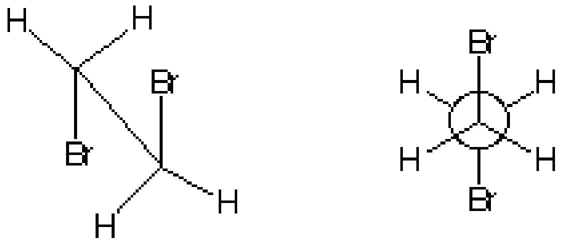


FIGURE 6. Fisher projection of the D-Glucose molecule.

The sawhorse projection becomes slightly more useful in such a case (Figure 7a). It shows the bond angles between the atoms and clearly presents the relative configuration of the atoms so it is easy to locate stereocenters, carbons with multiple substituents that have potential to define an enantiomeric pair. While cyclic and complex acyclic compounds can be drawn with a sawhorse projection, the drawing is often complex and time consuming.

The Newman projection, shown in Figure 7b, is a particularly interesting method of representing chemical structure. In a Newman projection, one looks directly down one of the carbon-carbon bonds in a molecule and then arranges the substituents around the carbons. Again, this is a good method for determining angles of attack in a chemical reaction or assigning stereochemistry to a center, but it is far too cumbersome to be a general means of communicating structure.



FIGURES 7a. and 7b. The same molecule expressed in a sawhorse projection (7a) and a Newman projection (7b).

The most widely used method of structural representation is the "two-dimensional" convention. In this convention, the thickness of the lines describes the positioning of the atoms in space. A thin, straight line describes atoms that are in the plane of the paper. A thick wedge represents an atom that is sticking out of the paper, and a dotted wedge represents an atom sticking back behind the paper. This convention is demonstrated with the molecule in Figure 8.

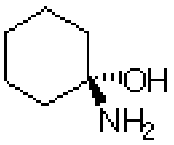


FIGURE 8. Two-dimensional convention for representing three-dimensional structures. The solid wedge which binds the NH2 group indicates that it is sitting "in front of" the page; the dotted line shows that the OH group is "behind" the page. (Note: this molecule has been created for the purpose of demonstration and may not exist in nature.)

CHEMICAL STRUCTURE IN INFORMATION RETRIEVAL

With the arrival of computerized information, many new products have emerged to answer the question of how best to retrieve all relevant information on a given subject. Since the drawing of chemical structures is quickly becoming the universal language of organic chemistry, complete and accurate information retrieval in the field of organic chemistry becomes increasingly reliant on matches between structures in a search query and those published in the literature. Therefore, it is imperative to have accurate structure editors that draw the molecules of interest and retain proper chemical significance. In addition, the use of chemical structure in information retrieval has some very important advantages over keyword or nomenclature searches. As has already been described, large organic molecules often have long and complex names. For example, one antidepression drug — known commercially as Trimipramine — is called 5-(3-dimethylamino-2-methylpropyl)-10,11-dihydro-5H-dibenz(b,f)azepine acid maleate (racemic form) (Windholz, 1983). Rather than write out this incredibly complex name that is almost as impossible to remember as it is to pronounce, a chemist vastly prefers to simply draw the structure (Figure 9).

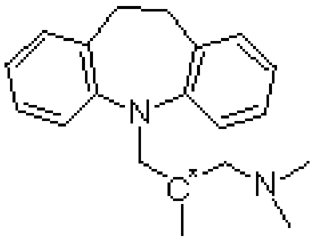


FIGURE 9. Trimipramine

However, long and complex names are not the only reason that structure is an important tool in information retrieval. Many compounds have multiple names. The IUPAC nomenclature system used to name the above antidepressant can assign multiple names to a single compound, and, with the addition of common names, a compound may be referred to by as many as three or four different titles. For example, a popular solvent in organic chemistry is CH_2Cl_2. This compound is commonly called both dichloromethane and methylene chloride, occasionally even interchangeably in the same text or paper. For an information retrieval system to work efficiently and to get the best possible recall, it is necessary that only one name be associated with each substance. If a scientist were to enter a search for "methylene chloride," all references about "dichloromethane" would be left out of the hit list.

One of the most important reasons for the use of structure searching in information retrieval is the fact that often chemists are performing reactions and synthesizing molecules that have never been attempted before. If a synthetic organic chemist is attempting to create a new molecule or find new pathways in the synthesis of a natural product, the intermediates and products that he desires to create will not be reported in the chemical literature. Since similar molecules react in similar manners, the chemist would be well served to search the literature for molecules with similar structures to the one in which he is interested. Therefore, these generic structures would not have names aside from the class name of the class of compounds being researched, and keyword searching would be impossible. An excellent example of this is the chemist who desires to make a particular kind of lactam. Structure A in Figure 10 shows the molecule that the chemist hopes to synthesize (Boeckman, Currano, Govek, & Hall, 1996-1999). A literature search indicates that no such compound has ever been made. However, structure B has been reported by Wanner (1988). Therefore, in order to retrieve this structure as well as other similar structures, the chemist needs a method of input such as structure C. In this manner, he details exactly which substituents must be present on the molecule (fixed sites) and allows sites in which it is unnecessary to specify an exact substituent to be open-ended (wildcard sites).

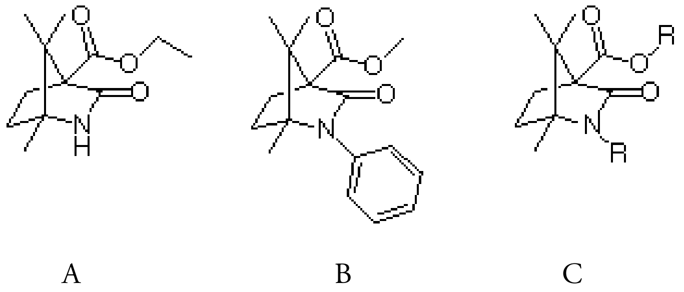


FIGURE 10.

METHODS OF STRUCTURE SEARCHING

The method most often adopted for performing chemical structure searches in information retrieval is a direct comparison of the queried structure to the structures present in the literature. In the past, it was found that it is more efficient to store the structures as text files than as graphic files for several reasons. Baumgras and Rogers (1995) enumerate several of these:

> Use of a standard text-based format offers two other advantages over moving images and other non-text formats. It is less subject to translation error during network transmission and file conversion. Non-text formats can include special characters which do not survive transmission and/or file translation well. Second, the file size of a [text-based format file] is substantially less than an image file. When sending across networks, file size is an important consideration. (p. 626)

Three types of text storage have become standard for chemical structures: fragmentation codes, linear notation, and connection tables. Each has its advantages and drawbacks, performing slightly different functions of comparison.

Fragmentation Codes

> Fragmentation codes are alphanumeric systems that provide the indexing needed to assist in searching for the structural aspects that are especially crucial for patents. They are intended to help provide the most complete recall possible, a feature important to searching of chemical patents. (Maizell, 1987, p. 173)

The concept of the fragmentation code dates back to the days of computer punch cards. A punch would be made on a card in an area representing a particular functional group, or fragment of a molecule. The identity of all the punches on the card would detail exactly which groups the molecule contained, and cards could be sorted by stacking them on pins positioned under the holes representing the desired structural elements. The modern computer uses fragmentation codes in a very similar way. Each possible functional group is given an alphanumeric code that uniquely identifies it. When a search of a compound is desired, the codes for the fragments are entered into the computer and a search is performed for substances containing matching codes. Two different types of codes exist: fixed (closed) codes and open-ended codes. Fixed codes contain a set number of fragments, while "[t]he fragments of an *open-ended* code are constructed from a structure representation by a computer algorithm according to a particular set of rules" (Ash et al., 1985, p. 130).

Fragmentation codes are referred to as "ambiguous" forms of representation because the codes are not searched in any particular order. Therefore, if four codes are entered into the system, the computer retrieves every structure in the database that contains those four codes regardless of order. Given the fact that conformation of atoms is very important in organic and related branches of chemistry, retrieving these varied structures can be problematic. For example, if one is searching for structure A in Figure 11 and enters the codes for its four fragments, many possible structures can be retrieved, including structures B, C, and D. Since the synthesis of a molecule involves finding a process by which a particular substituent can be attached to a particular site, papers on B, C, and D would not be at all useful to the chemist.

There are other drawbacks to fragmentation codes as well. Fixed codes are "not easily modified if new features need to be included. If a new fragment is invented, all previously processed structures will need recoding" (Warr & Suhr, 1992, p. 96). Although the concepts behind this method are clear-cut, it is often quite difficult to learn all of the codes in a system. In addition, it is impossible to export codes from one

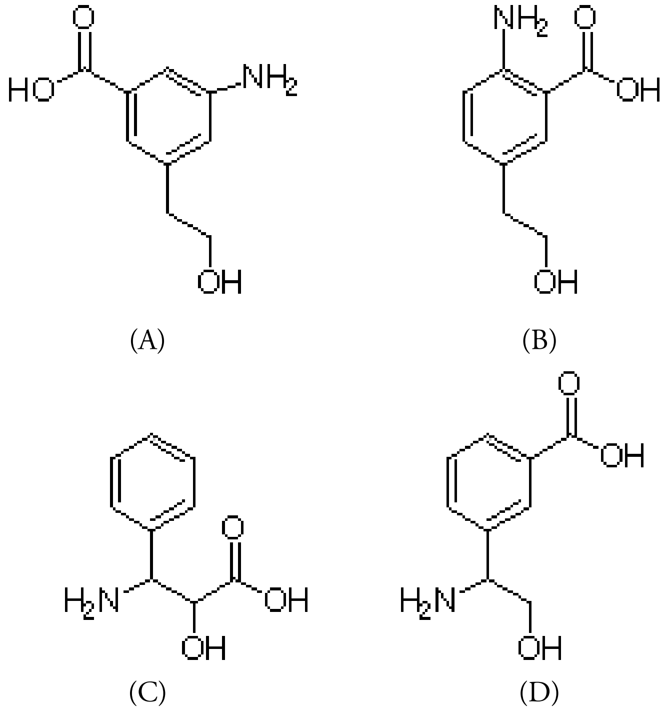


FIGURE 11. Various incorrect search results (B, C, D) based on a search for A due to ambiguity in fragmentation codes. (Note: Molecules have been created for the purpose of demonstration and may not exist in nature.)

system to another. Thus, for each system that the chemist uses, he must learn a new set of codes. Currently, one of the largest fragmentation code systems is the IDC-GREMAS (International Documentation in Chemistry - Genealogische Recherche mit Magnetband-Speicherung) code.

In the GREMAS system, each fragment is given a three-letter code. The first letter refers to the "genus" of the fragment. This letter defines the class of compound very broadly — carboxylic acids have one code, amines have another code, alcohols yet another. The second letter is the "species," which tells whether the fragment is part of a pure substance or a derivative of a substance. Finally, the third letter denotes the "subspecies," which gives additional information about groups surrounding the fragment in question (Ash et al., 1985). Although the codes are split up into degrees of specificity, there are still a large number of codes to learn. GREMAS has more than 700 different combinations of genus and species. In addition, two very different structures can end up having identical codes. Therefore, the system is complex to learn and rather ambiguous as well.

Linear Notation

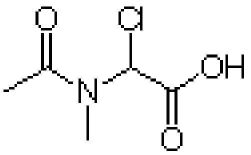
"Notation systems typically use combinations of letters, numbers, and punctuation marks to represent chemical formulas in simplified form" (Maizell, 1987, p. 174). Linear notation, therefore, is a cross between keyword searching and chemical structure searching. It provides an unambiguous method of searching for a structure because the fragments are presented in the order in which they appear in the molecule. The most prevalent linear notation is the Wiswesser Line-formula Notation (WLN). WLN takes on aspects of a fragmentation code because common functionalities are represented by single letters. Bond types such as single and double bonds are represented by symbols as well. For example, if a carbon is joined to another carbon with a double bond, the structure might be described as CUUC, where U represents a bond between the two atoms. If the carbons are triple bonded, the code changes to CUUUC (Ash et al., 1985).

Linear notation has definite advantages over fragmentation codes. Although it is necessary to learn the specific codes for common functionalities, the degree of memorization is less than that of fragmentation codes. Linear notations are, on the whole, much easier to learn and require less training. They are also more intuitive. It is easy to remember that if a triple bond is desired, three U symbols are needed, while a double bond only requires two. In addition, the linear notation method shows the order of the atoms as well as the connectivity between them. This makes it much more difficult to retrieve multiple hits for a given query and thus makes linear notation a good method of registering compounds. Unfortunately, "WLN has some of the disadvantages of systematic nomenclature; it is not suitable for handling tautomerism, mesomerism, and stereochemistry" (Warr & Suhr, 1992, p. 97). Although the connections between the atoms are clearly seen, there is no way to determine whether or not a given substituent on a carbon center is sticking out of the paper or into the paper. All that can be determined is that this substituent is attached to the carbon.

Connection Tables

The connection table is probably the best means of unambiguous representation available. In a system using connection tables as a means of storing chemical structures, each atom in the structure is assigned a number. The table contains three different types of information for each atom in the structure. The atoms are named using their chemical symbols and are assigned a number in the structure. The bond type between the atom and another atom in the structure is described numerically using 1 for a single bond, 2 for a double bond, and 3 for a triple bond, and the number of the atom or atoms to which it is bonded is presented. An example of a connection table can be seen in Figure 12.

The connection table in Figure 12 is called a redundant table because each bond is described in the row of each atom that participates. This is not, however, a very space-efficient



Atom Name	Connection	Bond	Connection	Bond	Connection	Bond
1C	2	1				
2C	1	1	3	2	4	1
3O	2	2				
4N	2	1	5	1	6	1
5C	4	1				
6C	4	1	7	1		
7C1	6	1				
8C	6	1	8	2	9	1
9O	7	2				
10O	7	1				

FIGURE 12. Redundant connection table for the hypothetical molecule above.

Atom Name	Connectivity	Bond
1C	-	-
2C	1	1
3O	2	2
4N	2	1
5C	4	1
6C	4	1
7C1	6	1
8C	6	1
9O	7	2
10O	7	1

FIGURE 13. Nonredundant connection table.

means of storing the information about the bonds. Therefore another method, known as a nonredundant table, was introduced. In a nonredundant table each bond is only described once. Figure 13 shows the nonredundant table for the structure described in Figure 12.

As was shown in conventional methods of drawing chemical structures, information about the hydrogens in the molecule is omitted. This is done for two reasons. Adding the hydrogen bonds into the structure would make the table much more complex and larger, mandating the use of more disk space to store it. In addition, it makes the system easier to remember for the chemists who do not generally draw the hydrogen atoms when they are drawing structures.

Connection tables meet several of the needs ignored by linear notation and fragmentation codes: it is very easy for a chemist to understand and to create a fragmentation code; no special nomenclature or symbols need be learned; and, in addition, by providing the chemist with specific information about the order and bonding of all of the atoms in the molecule, the table leaves no ambiguity about the topology of the atom. Unfortunately, the table does not address the problem of stereochemistry. Again, although it is clear to which carbon each substituent is bonded, the orientation in space is unclear. The connection tables in Figures 12 and 13 clearly specify that the chlorine atom is bonded to carbon 6, but they do not specify whether it is pointing out of or into the paper. In this case, the program would retrieve both forms of the molecule, which happen to be enantiomers of one another (Figure 14).

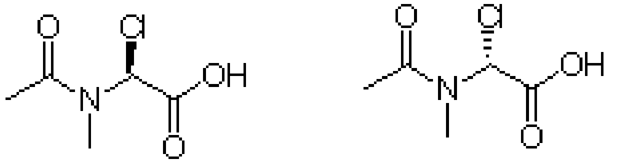


FIGURE 14. Possible enantiomeric molecules represented by the connection tables in Figures 12 and 13. Lack of stereospecificity in the tables can lead to false hits.

SUBSTRUCTURE SEARCH AND APPLICATIONS TO INFORMATION RETRIEVAL

The best method of locating structures with particular functionality is through substructure searching. Loosely defined, substructure is a particular combination of atoms or functional groups that makes up part of a larger structure. Each fragment in a fragmentation code, for example, is a substructure. For this reason, substructure searching is very useful in locating related compounds. If an organic or organometalic chemist is performing a synthesis of a compound or ligand (an organic compound that is bonded to a metal, usually for use in a catalyst) that has never been made before, a substructure search can help him to locate all molecules reported in the literature that contain the substructures in which he is interested. In this way, it is possible to state fixed sites while leaving wildcard sites open:

> Substructure search becomes essential where the interest of the searcher lies in the chemistry of the substructure itself; for example, in its involvement in certain types of chemical reaction . . . or in its contribution to the biological activity of a series of compounds . . . . In this case, substructure search retrieves from a file every structure which contains the specified combination of substructural features, irrespective of any other features which may be present. (Ash et al., 1985, p. 157)

Substructure searches can be made more or less specific with the addition of the Boolean operators AND, OR, and NOT. In addition, the degree of search can be specified. A substructure search can be so specific as to define a substructure as a single atom or general enough to specify a substructure consisting of several functional groups.

In the realm of single atoms as substructures, Xiao, Qiao, Shang, Lin, and Zhang (1997) describe an algorithm for a molecular representation code in which one starts with a single atom and "circles out" in the following manner. The first layer of code is the atom in question; the second layer encompasses all atoms bonded to the first-layer atoms; the third layer represents all atoms bonded to second layer atoms; and so on. Each layer is coded using the combined atomic weights of all substituents in that layer. Thus, the chances of achieving duplicate structures are slim. However, this type of substructure search is cumbersome for the user (especially the user who is not aware of the algorithm for determining the codes) and is an ineffective method of information retrieval due to the fact that it is nearly impossible to leave sites as wildcards. These problems can be combated, to a certain extent, by combining this type of substructure search with the connection table method of storing chemical information. *CAS Online* uses this type of structure search (Ash et al., 1985).

THE PROBLEMS OF STEREOCHEMISTRY, AND SOLUTIONS USING OBJECT ORIENTED PROGRAMMING AND CONNECTION TABLE INFORMATION

Thus far, no single method of substructure searching or information storage has adequately addressed the problem of stereochemistry, which is surprising given the importance of stereochemistry to chemical information retrieval. Over the years, several proposals have been made detailing possible methods for including stereochemistry in structure searching. Most, if not all of these procedures involve the use of object oriented programming and graphical information storage.

In 1990, Hanessian, Franco, Babnon, Laramee, and Larouche proposed a program specifically aimed at creating and identifying compounds with correct stereochemistry. Called CHIRON, it consists of four subprograms that deal with drawing structures: CARS-2D (Computer-Assisted Reaction Schemes and Drawing), CARS-3D (Simulation of 3-D Perception by Simultaneous Projection on Two Planes and Stereoscopic Viewing in Real-Time Mode), CASA (Computer-Assisted Stereochemical Analysis), and CAPS (Computer-Assisted Precursor Selection). The structure in question is entered graphically, building it on a grid of hexagons and specifying stereochemistry and orientation using the thick and dotted line convention previously described. Once the structure is input, it is possible to convert it from a two-dimensional drawing to a sawhorse projection and observe it from all angles by "suspending" it in the center of a virtual box and projecting it onto three of the walls. It can also be commanded to turn the bond between any two atoms into the axis for a Newman projection, which also helps the scientist to understand the stereochemistry of the molecule.

Using this technology, the program has the capability of taking two molecules and attempting to superimpose one over the other. "By activating the Ident command, CHIRON will find the identical (superimposable) segments consisting of four carbon atoms or more within a given structure or between two different structures, and it will display them on the screen in a visually clear manner" (Hanessian et al., 1990, p. 417). In addition, the program's Match Option will match two molecules without taking stereochemistry into account. Such capabilities would be very useful in an information retrieval system. A preliminary search using a "match" function would allow for a rough scan of the structures in a database. Following this with the more specific stereochemical search would provide an exact match.

While the extensive capabilities of the CHIRON system make it ideal for computer-aided synthesis design because of the easy manipulations of structures, the amount of programming needed to turn this technology into a database retrieval system seems to call for a more expedient method of specifying stereochemistry. In actuality, the result can be found by examining the methods by which chemists assign stereochemistry to molecules. The only place that stereochemistry exists is on a carbon atom that is bonded to four other atoms, at least three of which are unique. A set of rules known as the Cahn-Ingold-Prelog (CIP) convention are used for assigning stereochemistry to a center and for determining in which direction the molecule rotates light. This is done by drawing a Newman projection of the carbon center, usually viewing the center down a carbon-hydrogen or carbon-carbon bond, and ranking the other three substituents using the CIP rules. According to CIP, the higher the atomic number, or mass, the higher the rank of the atom, and usually this is sufficient for determining the stereochemistry of the centers. Next, the numbers are connected with an arrow, as shown in Figure 15. If the arrow is pointing counterclockwise, the stereochemistry is termed S. If the arrow points clockwise, the stereochemistry at the center is R.

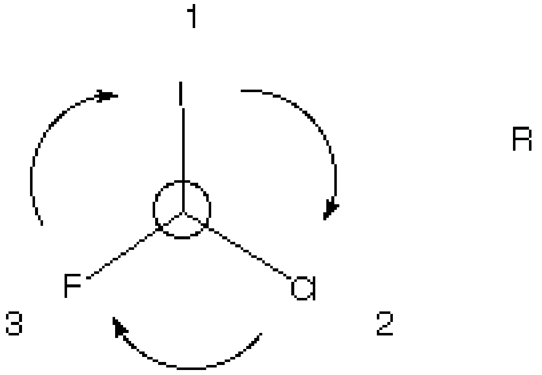


FIGURE 15. Assignment of stereochemistry of a hypothetical molecule according to the Cahn-Ingold-Prelog convention.

The CIP rules have been implemented in several programs, including the CHIRON program discussed previously and another program called LHASA (logic and heuristics applied to synthetic analysis). The method of implementation in LHASA seems particularly well suited to information retrieval and can even be used in a text-based storage and retrieval system "to achieve a complete identification and specification of the most important stereogenic units in organic chemistry and to produce a solid basis for future extensions of the module" (Mata, Lobo, Marshall, & Johnson, 1994, p. 491).

The LHASA program stores information in connectivity tables. When analyzing the stereochemistry of a molecule, it identifies each stereocenter present and divides them into two categories: those for which all classification information are present and those that are missing information. The program begins by analyzing the centers for which all pertinent information is present. Mata et al. (1994) explain that in the LHASA program, "if the comparison of a pair of ligands [substituents] cannot be concluded because it requires information about centers that have not yet been specified, it is suspended and resumed later. If the ligands can be compared, a decision must be made as quickly as possible about whether the center is stereogenic or prochiral, and if it is a stereogenic center, it is specified . . . . This cycle is repeated as many times as required until all possible centers are analyzed" (p. 494).

This type of test for stereochemistry is invaluable for information retrieval purposes for a number of reasons. The information found in the connection tables of many databases can be used without alteration. It is simply necessary to input a program that prompts the retrieval system to examine each structure according to the CIP rules. If CIP information about each of the structures is added to the databases, it is possible, after performing a stereochemical test of the query molecule, to compare the R and S configuration of that molecule to those in the database. This should, therefore, be the final step in any information search. This is much simpler than converting all of the structures in the database to three-dimensional structures that can be turned and manipulated, as the CHIRON system does. All of the information needed to assign a stereochemistry to a center is present in the connection table because the only necessary requirements are that the substituents of each center and the stereochemistry of adjacent centers are known. In addition, if the center is not optically active (it has neither an R or S configuration and, therefore, does not rotate light), such a method would immediately recognize this fact and would eliminate any optically active compounds in the databases from the list of retrieved hits.

As computer technology improves and storage space considerations become less of an issue, it becomes increasingly popular to eliminate all of the problems that exist with text-based information retrieval systems by switching to graphically based systems. CS ChemFinder, MDL's Chemscape, and Beilstein CrossFire all use object oriented programming to create structure search engines that allow the chemist to draw a structure using his mouse and search the literature for corresponding structures. Using the two-dimensional convention in which thick and dotted lines aid in describing stereochemistry, ambiguity of any type is effectively eliminated. Unfortunately, the use of graphic information storage comes equipped with a set of problems that is entirely its own: "[T]he major limitation of both [ChemFinder and MDL's Chemscape] is their need to have the sub-structure queries drawn in ChemDraw or ISIS/Draw. This requires ChemDraw or ISIS/Draw to be loaded on the client's machine. If a user does not have the software, no structural searches can be done" (Lo, 1998, p. 8).

This problem is very pertinent because of the means of graphical storage. Since the functional groups of chemical structures must have recognized chemical functionality in order to compare to similar functionalities present in the literature, the chemical structures cannot just be pretty pictures. Two lines between atoms must always represent a double bond, the letter 'N' must always represent a nitrogen atom, and the '+' and '-' symbols must always show charge.

A chemical structure editor must also have the ability to apply and interpret chemical shorthand. For example, chemists often like to use shorthand notation when their molecules have lengthy side chains or particular functional groups such as alcohols or carboxylic acids. Whenever a chemist attaches the group COOH to a free site in his molecule, the functionality of the carboxylic acid must be recognized. A problem with current structure editors is that functionality is often lost when a structure is transferred between one editor and another. Baumgras and Rogers (1995) explain that,

> [m]ost tools do not understand the chemical significance of a structure drawn using another tool, and the structure often can not be edited using a different tool from the one that created it. Also, it cannot be used for searching or registration without the underlying chemistry.
>
> A brief understanding of the cut and paste function explains this. When any picture is cut and pasted, an image of the picture is captured onto the clipboard. On a given platform, a common image format is used, which is why the picture of the structure usually transfers well from package to package. The software vendor . . . attach[es] a file describing the structure onto the end of the image file . . . . Generally speaking, one vendor does not interpret another's format, and the chemistry is lost when cutting and pasting to a different drawing tool. (pp. 624-625)

The result of this incompatibility is often "enormous atoms" — a circumstance that occurs when a side chain is interpreted as a single atom. There are also problems in interpreting the type of bonds and the charge that often appears on atoms within the molecule.

One method of circumventing this problem is to use a structure editor called WebSketch, which "uses JAVA applets to match a 2D query against the structures and [then] display the structures" (Lo, 1998, p. 12). This program is beneficial because it does not require any additional software and it allows for searching via the Web. Unfortunately, many institutions prohibit employees from downloading applets or other free software because of the risk of viruses. In addition, it is unclear whether or not this would truly solve the problem of misinterpretation of chemical significance.

SUGGESTIONS FOR A TEXT-BASED INFORMATION RETRIEVAL SYSTEM

Since the use of substructure searching applies both to the comparison of structural elements and the determination and comparison of stereochemistry, it is best to begin the design of a structure retrieval system with a substructure system. One problem with chemical substructure searching is that it takes a great deal of time to perform the most detailed atom level searches that are often needed to find exact or partial matches to a queried compound. The larger the database being searched, the greater the length of time and amount of computing resources needed to perform the search. Therefore, it is useful to be able to break a substructure search into several different stages. This technique is often referred to as screening. Ash et al. (1985) define several "practical considerations" for designing a screening system for structural retrieval:

1. the nature of the structure file;
2. the probable nature of the queries;
3. the characteristics of the use of the file;
4. the size of the file; and
5. the file organization. (p. 160)

Taking these considerations into account, it is possible to propose a text-based structure retrieval system that answers the needs of most chemists.

Using a connection table format, the chemist is able to input information about the structure he desires to locate. If the chemist desires to leave sites wildcard, he simply connects a generic R group to the unspecified site. When the structure is entered into the computer, the search begins. The first stage of the search very much resembles a fragmentation code search. The information retrieval program examines the identity of the atoms in the structure, counts the numbers of each atom type present, and compiles a molecular formula. This formula is entered into the database, and an initial hit list is generated, containing only those compounds whose molecular formula matches that in the query. Next, the program analyzes the connections in the query structure. Beginning with the first atom, it runs a program that compares the first three or four atoms and their connectivity to the structures retrieved from the database in the molecular formula search. As soon as a molecule from the hit list fails to match the query molecule, it is removed from the set of hits. The next comparison examines the next three atoms in a similar manner, and the process continues until all atoms have been compared. Since a structure is rejected the moment it fails to match the query and increasingly smaller groups of compounds are being compared, time is saved in the search process. A final hit list consists of only molecules that match the structure as entered. This final hit list would be compared using LHASA-style stereochemical analysis to locate molecules that resemble the query as closely as possible.

CONCLUSION

Although the benefits of object oriented programming for use in chemical structure searches seem vast, until technology produces a universal structure editor, it is much safer to stick to the tried-and-true text representations. Through the use of an information retrieval system such as the one described above, it would be possible for chemists not only to find their way through the vast forest of chemical literature, but to find the particular tree, or structure, for which they are searching.

REFERENCES

Ash, J., Chubb, P., Ward, S., Welford, S., & Willett, P. (1985). *Communication, storage, and retrieval of chemical information.* Chichester, England: Ellis Horwood Limited.

Baumgras, J. L., & Rogers, A. E. (1995). Chemical structures at the desktop: Integrating drawing tools with on-line registry files. *Journal of the American Society for Information Science, 46*(8), 623-631.

Boeckman, R. K., Currano, J. N., Govek, S., & Hall, D. (1996-1999). Unpublished results. Department of Chemistry, University of Rochester, Rochester, NY.

Hanessian, S., Franco, H., Babnon, G., Laramee, D., & Larouche, B. (1990). Computer-assisted and perception of stereochemical features in organic molecules using the CHIRON program. *Journal of Chemical Information and Computer Science, 30*, 413-425.

Lo, M.-L. (1998). Recent strategies for retrieving chemical structure information on the Web. *Science and Technology Libraries, 19*(1), 3-17.

Maizell, R. E. (1987). *How to find chemical information: A guide for practicing chemists, educators, and students* (2d ed.). New York: John Wiley & Sons.

Mata, P., Lobo, A. M., Marshall, C., & Johnson, A. P. (1994). Implementation of the Cahn-Ingold-Prelog System for stereochemical perception in the LHASA program. *Journal of Chemical Information and Computer Science, 34*, 491-504.

Wanner, K. T. (1988). New chiral auxiliary groups: Formation of a phenylaza analog of camphoric acid after change of N/O-regioselectivity of a ring-contraction reaction. *Liebigs Annalen der Chemie, 6*, 603-604.

Warr, W. A., & Suhr, C. (1992). *Chemical information management.* Weinheim, Germany: VCH Publishers.

Windholz, M. (Ed.). (1983). *The Merck Index: An encyclopedia of chemicals, drugs, and biologicals* (10th ed.). Rahway, NJ: Merck & Co., Inc.

Xiao, Y., Qiao, Y., Shang, J., Lin, S., & Zhang, W. (1997). A method for substructure search by atom-centered multilayer code. *Journal of Chemical Information and Computer Science, 37*, 701-704.

ADDITIONAL READING

Bauerschmidt, S., & Gasteiger, J. (1997). Overcoming the limits of a connection table description: A universal representation of chemical species. *Journal of Chemical Information and Computer Science, 37*, 705-714.

Dietz, A. (1995). Yet another representation of molecular structure. *Journal of Chemical Information and Computer Science, 35*, 787-802.

Heller, S. R. (Ed.). (1990). *The Beilstein Online Database: Implementation, content, and retrieval.* Washington, DC: The American Chemical Society.

Holliday, J. D., & Lynch, M. F. (1995). Computer storage and retrieval of generic chemical structures in patents. 17. Evaluation of the refined search. *Journal of Chemical Information and Computer Science, 35*, 659-662.

Leung, K. M., Chau, F. T., Kwok, P. H., & Lau, W. T. (1997). ChemISTools: A computer software for chemical information systems. *Computers and Chemistry, 21*(3), 161-166.

Schoch-Grüber, U. (1990). (Sub)structure searches in databases containing generic chemical structure representations. *Online Review, 14*, 95-108.

Wiggens, G. (1991). *Chemical information sources.* New York: McGraw-Hill.

ABOUT THE AUTHOR

Judith Currano graduated from the University of Rochester with bachelor's degrees in chemistry and English. While at Rochester, she worked on an independent organic synthesis project and taught in the freshman chemistry labs, gaining first-hand experience with the problems of information retrieval in organic chemistry. Her time at the University of Illinois at Urbana-Champaign's Graduate School of Library and Information Science was spent learning as much as possible about chemical information. She received her MLIS in August and is currently the chemistry librarian at the University of Pennsylvania.