conformation generation. Answer sets may also be combined by using logical AND, NOT, and OR operations.

CONCLUSIONS

The system described above satisfies the prime requirements of a 3-D search system in that it

1. Is able to search effectively all accessible conformations of flexible molecules within realistic time scales.
2. Integrates both query construction and subsequent results processing with a modeling system.

It should be noted that no step in the process is dependent on another step having been taken previously, the suggested sequence of events being given for efficiency only.

As a consequence of the methods used, screen search times are independent of the number of conformations handled, as are the database disk storage requirements.

Subsequent papers in this series will give details of actual database configuration parameters and describe a typical case study.

REFERENCES

(1) Jakes, S. E.; Watts, N.; Willett, P.; Bawden, D.; Fischer, J. D. Pharmacophoric pattern matching in files of 3D chemical structures: evaluation of search performance. *J. Mol. Graphics* **1987**, *5*, 41-48.

(2) Sheridan, R. P.; Nilakantan, R.; Rusinko, A., III; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: A system for three-dimensional substructure searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255-260.

(3) Martin, Y. C.; Danaher, E. B.; May, C. S.; Weininger, D. MENTHOR, a database system for the storage and retrieval of three-dimensional molecular structures and associated data searchable by substructural, biologic, physical, or geometric properties. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 15-29.

(4) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructural searching of three-dimensional molecular structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225-251.

(5) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelick, T.; Hummelick-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. The Cambridge Crystallographic Data Center: Computer-Based Search Retrieval, Analysis and Display of Information. *Acta Crystallogr.* **1979**, *B35*, 2331-2339.

(6) Rusinko, A., III; Sheridan, R. P.; Ramaswamy, N.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. Using CONCORD to construct a Large Database of Three-Dimensional Coordinates from Connection Tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251-254.

(7) Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. CONCORD: *A program for the rapid generation of high quality approximate 3-dimensional molecular structure*; The University of Texas: Austin; and Tripos Associates: St. Louis, MO, 1988.

(8) Sheridan, R. P.; Nilakantan, R.; Dixon, J. S.; Venkataraghavan, R. The ensemble approach to distance geometry: application to the nicotinic pharmacophore. *J. Med. Chem.* **1986**, *29*, 899-906.

(9) Mayer, D.; Naylor, C. B.; Motoc, I.; Marshall, G. R. A unique geometry of the active site of angiotensin-converting enzyme consistent with the structure-activity studies. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 3-16.

(10) Lloyd, E. J.; Andrews, P. R. A common structural model for central nervous system drugs and their receptors. *J. Med. Chem.* **1986**, *29*, 453-462.

(11) Chem-X molecular modeling software, developed and distributed by Chemical Design Ltd., Oxford, England, 1990.

(12) Murrall, N. W.; Davies, E. K. In preparation.

(13) Ganellin, C. R. Chemistry and Structure-Activity Relationships of Drugs Acting at Histamine Receptors. In *Pharmacology of Histamine Receptors*; Ganellin, C. R., Parsons, M. E., Eds.; Wright-PSG: London, 1982; pp 35-37.

(14) Jakes, S. E.; Willett, P. Pharmacophoric pattern matching in files of 3-D chemical structures: selection of interatomic distance screens. *J. Mol. Graphics* **1986**, *4*, 12-20.

(15) Ganellin, C. R. Chemistry and Structure-Activity Relationships of Drugs Acting at Histamine Receptors. In *Pharmacology of Histamine Receptors*; Ganellin, C. R., Parsons, M. E., Eds.; Wright-PSG: London, 1982; pp 21 and 77.

(16) Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willett, P. *Communication, Storage and Retrieval of Chemical Information*; Ellis Horwood Ltd.: Chichester, U.K., 1985; pp 160-167.

(17) Dolata, P. D.; Leach, A. R.; Prout, K. WIZARD: AI in conformational analysis. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 73-85.

(18) Tarjan, R. E. Graph algorithms for chemical computation. *ACS Symp. Ser.* **1977**, *No. 46*, 1-19.

(19) MACCS—Molecular ACCess System, developed and distributed by Molecular Design Ltd., San Leandro, CA, 1989.

(20) OSAC—Organic Structures Accessed by Computer, developed and distributed by ORAC Ltd., Leeds, England, 1989.

(21) Ullman, J. R. An algorithm for subgraph isomorphism. *J. Assoc. Comput. Mach.* **1976**, *16*, 31-42.

(22) Brint, A. T.; Willett, P. Pharmacophoric pattern matching in files of 3D chemical structures: comparison of geometric searching algorithms. *J. Mol. Graphics* **1987**, *5*, 49-56.

(23) Ferro, D. R.; Herrmans, J. A different best rigid-body molecular fit routine. *Acta Crystallogr.* **1977**, *A33*, 345-347.

(24) Perry, N. C.; Davies, E. K. The use of 3D modeling Databases for Identifying Structure Activity Relationships. In *QSAR: Quantitative Structure-Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R. Liss Inc.: New York, 1989; pp 189-193.

(25) Marshall, G. R.; Barry, C. D.; Bossard, H. E.; Dammkoehler, R. A.; Dunn, D. A. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; ACS Symposium Series 112; American Chemical Society: Washington, DC, 1979, pp 205-226.

(26) Young, R. C.; Durant, G. J.; Emmett, J. C.; Ganellin, C. R.; Graham, M. J.; Mitchell, R. C.; Prain, H. D.; Roantree, M. L. Dipole Moment in Relation to H_2 Receptor Histamine Antagonist Activity for Cimetidine Analogues. *J. Med. Chem.* **1986**, *29*, 44-49.

(27) ORACLE: Relational Database Management System, distributed by ORACLE U.K. Ltd., Richmond, Surrey, England, 1989.

Automated Conformational Analysis and Structure Generation: Algorithms for Molecular Perception

ANDREW R. LEACH,* DANIEL P. DOLATA,‡ and KEITH PROUT

Chemical Crystallography Laboratory, University of Oxford, 9 Parks Road, Oxford OX1 3PD, U.K.

Received May 2, 1990

Many methodologies for performing automated conformational analysis require some means of "perceiving" a molecule to determine features of interest. Algorithms for finding rings, bond orders, and stereocenters and detecting the presence of substructural fragments have been developed. These algorithms are described, emphasizing their importance in conformational analysis.

INTRODUCTION

WIZARD and COBRA are two programs which use artificial intelligence techniques to construct low-energy conformations of molecules. One of the major objectives of this project was to develop a means by which low-energy conformations of a molecule could be rapidly generated, starting from a simple "two-dimensional" representation (i.e., connectivity and atom types). In this paper some of the algorithms that these programs use to "perceive" a molecule are described. First, however, a brief description of the method these programs use to perform a conformational analysis is given (further details can be found elsewhere[1]).

Starting from a definition of the molecule (which can be specified either via a datafile obtained from a molecular

[*] Address correspondence to this author at his present address: Computer Graphics Laboratory, School of Pharmacy, University of California, San Francisco, CA 94143-0446.

[‡] Present address: Dept. of Chemistry, University of Arizona, Tucson, AZ 85721.

0095-2338/90/1630-0316$02.50/0 © 1990 American Chemical Society

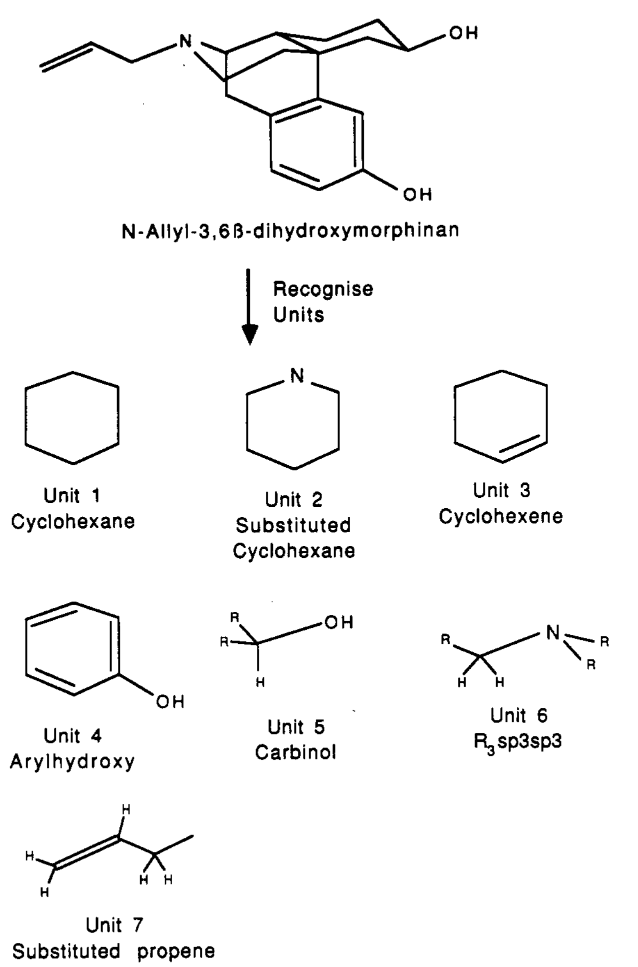

Figure 1. Units recognized in *N*-allyl-3,6β-dihydroxymorphinan.

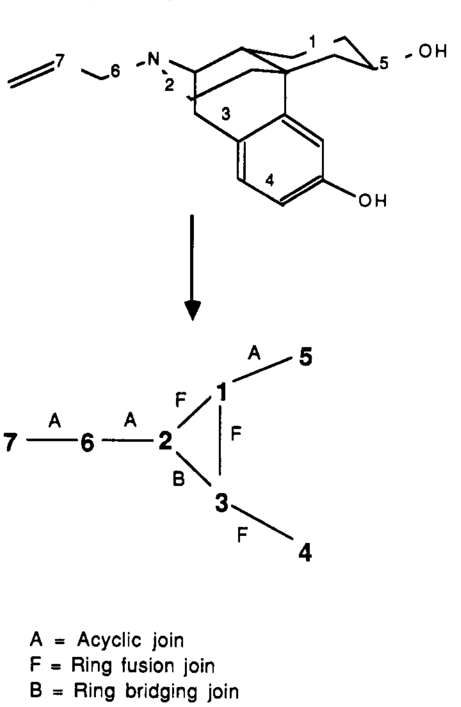

Figure 2. Unit graph for *N*-allyl-3,6β-dihydroxymorphinan.

graphics program or as a SMILES[2] string), the molecule is first analyzed to identify relevant features: the bond orders are determined, any rings present are identified, and aromatic rings are located. In addition, any atoms which have ambiguous stereochemistry (e.g., asymmetric carbon atoms) are identified. Finally, the presence of *conformational units* in the molecule is determined. A conformational unit is some group of connected atoms about which the system has "knowledge". As an illustration, Figure 1 shows the seven units recognized in *N*-allyl-3,6β-dihydroxymorphinan. A graph representation of the molecule is then constructed; this is based upon the units identified and how they are connected (the "unit graph"). Figure 2 shows the unit graph for *N*-allyl-3,6β-dihydroxymorphinan.

After identifying the conformational units which describe the molecule, the search of its conformational space commences. A *subconformation* is first assigned to each unit to give a high-level *suggested* conformation for the molecule. Each subconformation typically represents a minimum-energy conformation of the unit (for example, there are three subconformations for the butane unit, corresponding to the three minimum-energy staggered conformations of that molecule). The high-level suggestion is then examined to check that it has no problems (e.g., the subconformation energies are not too high and the suggestion does not contain a combination of subconformations found previously to give rise to a problem). If satisfactory it is then constructed by joining the unit subconformations one at a time, with checks after each join to ensure that the partially constructed conformation is acceptable. Should an unacceptable structure be detected, the construction is abandoned and the next suggested conformation is generated. A variety of different checks may be performed after each joining step. The quality of the join is checked to ensure that the unit subconformations fit adequately. Pairwise atomic close-contact ratios are tested to ensure that they do not exceed the currently defined allowed value. Additionally, it may be desired that the van der Waals' energy of the assemblage does not exceed a defined cutoff value, or that the interatomic separation of specified atom pairs falls within a user-defined range (e.g., to fit a receptor model). If the partially constructed molecule is acceptable, the next unit is joined, and the construction continues until the conformation is complete, whereupon its energy is calculated and output to disk. A "total" search of conformational space is performed by examining all possible combinations of unit subconformations. The programs have been used to examine a wide variety of organic molecules and have been extended to cover transition metal coordination complexes.

In this paper the initial analysis algorithms are described, covering ring perception, bond order determination, identification of ambiguous stereocenters, and discovery of conformational units. Various other programs (e.g., those for synthesis design or database searching) also use such algorithms, but in many cases we have found that these cannot be directly applied to conformational analysis. Thus, in the discussion below emphasis will be placed on the reasons for choosing a particular approach. There are some small differences between the implementations in COBRA and WIZARD. Where such differences exist the algorithm implemented in COBRA will be described.

PERCEPTION OF RINGS

Rings impose constraints on the possible conformations that can be adopted by a molecule. It is therefore important that the ring system of the molecule is correctly analyzed. When a molecule contains fused or bridged rings, it is possible to describe the ring system in a number of different ways; for example, to specify all possible rings. Consider bicyclo[4.3.1]nonane (hydrindane) which has a total of three rings, containing five, six, and nine atoms. The complete ring system of this molecule can be described using just the five- and six-membered rings as the nine-membered ring is a linear combination of these two.

Many algorithms have been described for determining the rings in a connected graph.[3] These algorithms do not all discover the same set of rings for a given ring system. Some find all possible rings; others find particular subsets such as the "chemically important rings". The minimum number of rings required to describe a ring system (the Frerejacque number) is given by

no. of rings = no. of bonds − no. of atoms + 1

One such fundamental set of rings is the smallest set of smallest rings (SSSR), first described by Plotkin.[4] All of the other rings in the system can be obtained by taking linear combinations of rings from the SSSR. There are a number of advantages in using the SSSR to describe the ring system of a molecule for conformational analysis. Of particular importance is that the conformation of the ring system is dictated by the conformational possibilities available to the smallest rings. For example, it would be impractical to try and construct conformations of hydrindane by using cyclononane templates as none of the stable conformations of hydrindane corresponds to a stable conformation of cyclononane.[5] Additionally, the conformational properties of smaller rings are more easily evaluated, and as they typically have fewer low energy conformations, the search of conformational space can be achieved more quickly. Further advantages in using the SSSR accrue from the easy recognition of aromatic ring systems and the fact that ring strain is more accurately estimated by summing the energies of the rings in the SSSR.

Consequently, it was decided to use the SSSR to describe the ring system of the molecule. A number of algorithms have previously been described for finding the SSSR. In Plotkin's method, all of the rings are found first. These rings are ordered according to the number of atoms they contain. The smallest ring is then assigned to the SSSR. Subsequent rings are only added if they are linearly independent of the rings already assigned. It would clearly be an advantage if the SSSR could be found more directly without first having to find all the rings present in the system. Two more recent papers describe algorithms for doing so.

Gasteiger and Jochum's algorithm[6] generates a spanning tree from an arbitrary atom and finds ring-closure bonds. For each of these, it traces back along the tree until ring closure is noted. This algorithm is successful in many cases, but for some ring systems the set of rings found is not the SSSR. They therefore use a measure of the "complexity" of the ring system to determine whether or not the procedure needs to be repeated using a different reference atom. In this case the SSSR is chosen from the entire set of rings found. However, this algorithm was found to fail (using the complexity criteria suggested) for certain types of ring systems which contain embedded rings. Roos-Kozel and Jorgensen have described an algorithm[7] which uses a "path-growing–ring-finding" method, starting at the most highly connected atoms in the ring system. The process continues until the SSSR has been found. Their basic algorithm is unable to deal with certain special cases which then have to be processed by a different method.

The objective when devising the algorithm for use in WIZARD and COBRA was that it should be as general as possible, correctly finding the SSSR for all ring systems, without having to rely upon the recognition of special situations or the use of any particular numbering scheme.[8] The algorithm devised, which shows some features common to the methods of both Gasteiger and Roos-Kozel, has the following steps:

1. Eliminate terminal atoms from the molecule so that only the ring atoms remain.
2. Create a spanning tree of the remaining ring system with an arbitrary atom as the root node.
3. Find ring-closure bonds.
4. For each ring-closure bond, find the smallest ring (or rings) of which it is a member.
5. Attempt to allocate a unique ring to each ring-closure bond, so that a linearly independent set is obtained. If this can be done, the set constitutes the SSSR. If this is not possible then the next smallest rings must be found for each ring-closure bond and another attempt is made to find a linearly independent set of rings. Continue the ring finding and testing until the SSSR is successfully generated.

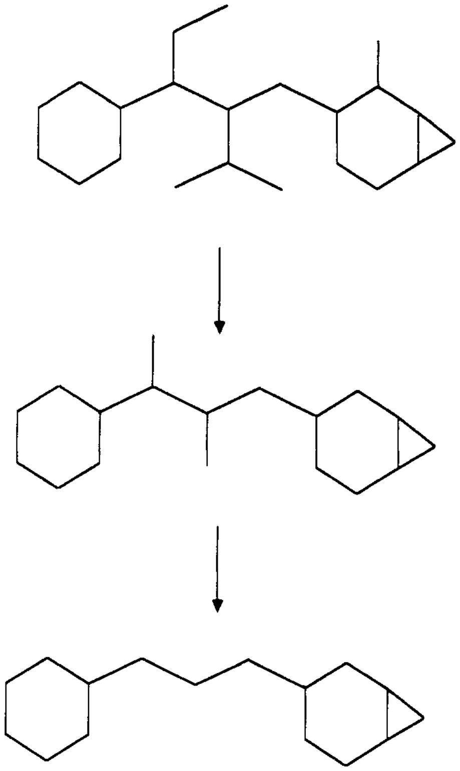

Figure 3. Elimination of terminal atoms prior to ring analysis.

Removing terminal atoms can greatly reduce the total number of atoms which have to be considered; Roos-Kozel and Jorgensen also use this technique to improve the efficiency of their algorithm. Each atom in the molecule which is bonded to a single other atom is identified and eliminated from the structure; the adjoining atom is now bonded to one fewer atom. This process is repeated until either all of the atoms have been eliminated (i.e., the molecule does not contain any rings) or until no more singly connected atoms remain. Figure 3 gives an illustration of how this pruning process operates; note that not every atom which remains is necessarily a member of a ring.

An arbitrary reference atom is then chosen from those which remain, and its *spanning tree* is created. The spanning tree of an atom is the molecular graph with the atom as the root node. The *i*th level of this tree thus contains the set of atoms whose minimal distance to the reference atom is *i*. Each atom which remains after pruning the terminal atoms (except of course the reference atom) will have at least one bond to an atom in a lower shell (i.e., closer to the reference atom). Any additional bonds to atoms in lower shells or any bonds to atoms in the same shell are *ring-closure bonds*. The number of ring-closure bonds in any ring system is equal to the number of rings in the fundamental set needed for its description. Moreover, *each of the ring-closure bonds can be associated with a unique ring in the SSSR* (a bond is *associated* with a ring if it is contained within the ring and a set of *unique* rings is a linearly independent one). Recognition of this fact is central to the method, for finding the ring to associate with each ring-closure bond will give the SSSR. (Here it is worthwhile to note the difference with the ring-assembly algorithm devised by Wipke and Dyott[9] which was based on the premise that every ring in the basis set contains one bond which

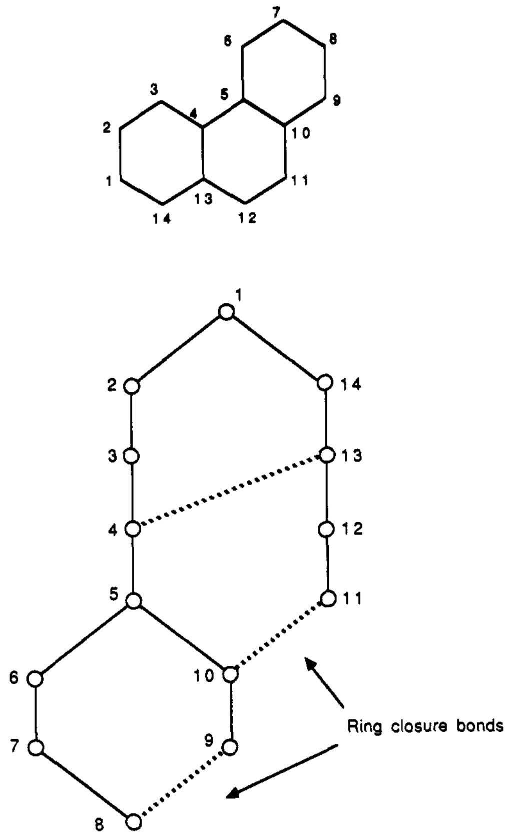

Figure 4. Ring system and associated spanning tree to illustrate operation of ring-finding algorithm.

is contained in no other ring in the basis set.)

In order to find the ring to be associated with a particular ring-closure bond, spanning trees are grown from the two atoms which comprise the bond. This continues until a common atom is discovered, indicating ring closure. To find the atoms in the ring the two spanning trees are retraced to the two atoms of the ring-closure bond by the shortest paths. Should more than one common atom be found, the path for each is retraced so that it is the smallest ring which is found. If the ring has not been discovered previously it is added to the list of rings found so far. If there is more than one ring of the minimal size, each is stored (this may occur in fused or bridged systems). Consider, for example, the molecule shown in Figure 4. Suppose atom 1 is chosen as the reference atom. When the spanning tree of this atom is constructed, three ring-closure bonds are found (bonds 4–13, 10–11, and 8–9; see Figure 4). Note that a different set could have been chosen; for example, atom 4 has bonds to two atoms in the previous level, either of which could be defined as the ring-closure bond. For the present purpose the ring-closure bond is defined as the bond to the higher numbered atom. For the bond 4–13, two rings (each with six atoms) are found when the spanning trees are constructed for atoms 4 and 13. For the bond 10–11 one ring is found, as is the case for bond 8–9.

When all of the ring-closure bonds have been processed, each of them will have one or more associated rings. It should be noted that a ring may be associated with more than one ring-closure bond. For example, ring [4,5,10,11,12,13] is associated with both bonds 4–13 and 10–11. An attempt is then made to try and obtain the SSSR from the rings found so far. This is done by selecting one ring for each ring-closure bond. The set so obtained is then examined to see if it satisfies the two criteria noted earlier. First, each ring-closure bond must be associated with a unique ring. Second, the set of rings must be linearly independent. A ring is linearly independent of the other rings in the set if it cannot be constructed by adding together any combination of the other rings in the set. In the hydrindane example above the nine-membered ring can be constructed by adding the five- and six-membered rings and so it is not linearly independent of these two rings.

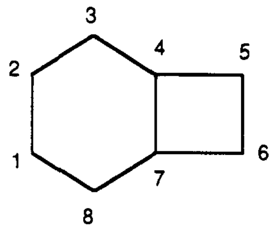

Figure 5. Bicyclo[4.2.0]octane to illustrate ring-finding algorithm.

If it is possible to choose a set of rings which satisfies these criteria, the set constitutes the SSSR and the ring system has been evaluated. For the molecule in Figure 4 this can be achieved by assigning ring [1,2,3,4,13,14] to bond 4–13, ring [4,5,10,11,12,13] to bond 10–11, and ring [5,6,7,8,9,10] to bond 8–9. However, it may not be possible to satisfy these requirements using the procedure described so far. In such cases it is necessary to search for additional rings in the molecule. A simple example is bicyclo[4.2.0]octane (Figure 5). Suppose atom 1 is taken as the reference atom, giving rise to the two ring closure bonds 4–7 and 5–6. When the smallest ring(s) is (are) found for these bonds the same four-membered ring is found for each. Hence it is not possible to assign a unique ring to each ring-closure bond. Consequently, each ring-closure bond is examined again, and the spanning trees extended to find additional rings, linearly independent of those already found. Further expansion of bond 4–7 gives the six-membered ring, which is added to the list of rings found. Bond 5–6 gives an eight-membered ring. However, this is rejected as it is not linearly independent of the other rings already found (i.e., it can be constructed from the four- and the six-membered rings). Bond 4–7 now has two associated rings (containing 4 and 6 atoms), while bond 5–6 still has just one associated ring (of size 4). It is now possible to take one ring from those found for each ring-closure bond to give the SSSR (i.e., the six-membered ring for bond 4–7, and the four-membered ring for bond 5–6).

Although different numbering schemes would have enabled the SSSR to be found in a smaller number of steps in the examples given above, the algorithm successfully discovers the SSSR for an "unhelpful" numbering scheme, as required. During testing many complex ring systems were analyzed, including all of the test cases of Gasteiger and Roos-Kozel, some of which are shown in Figure 6. In each case the required SSSR was correctly found.

DETERMINATION OF BOND ORDERS

COBRA must correctly assign an order to each bond in the molecule. Some file formats do allow (or even require) bond orders to be specified, but many do not. Consequently, it was decided to ignore any bond-order assignments that might be present in the input file (or in the SMILES string) and determine the bond orders solely from the atom type and connectivity information. This also provides a check on the consistency of a user-defined structure. For many molecules the task of assigning bond orders is straightforward. However, it can be more complicated for some types of conjugated systems, especially when they also contain unsaturated rings. It is important that the bond orders are correctly assigned in such molecules. For example, consider the conjugated system shown in Figure 7. This can adopt four different conformations (as indicated), which are interchanged by rotating about the two single bonds. If the bond orders were incorrectly assigned, incorrect results would be obtained. The algorithm

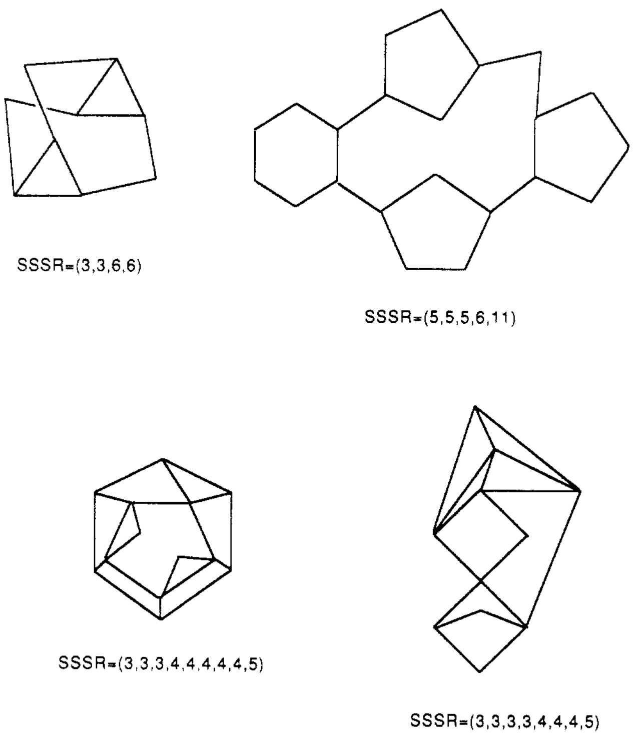

Figure 6. Some examples used to test the ring-finding algorithm.

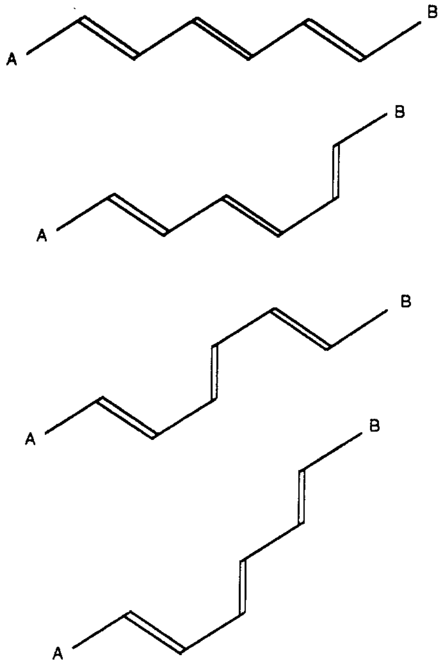

Figure 7. Representation of the four conformations of a conjugated system which requires the correct assignment of bond orders.

to determine the bond orders proceeds through the stages described below.

(1) Initialization. Each bond in the molecule is assigned a default order of one. A necessary, but not sufficient, requirement for an unsaturated bond is that it connects two atoms which are both unsaturated. The algorithm keeps a running check of the degree of unsaturation of each atom in the molecule; this is given by the number of unfilled valencies, NVAL(atom). The initial value of NVAL is the difference between an atom's valence and the number of connected atoms.[10] For example, the initial values of NVAL for the atoms in the molecule in Figure 8 are as indicated. When a bond is assigned, the values of NVAL for the two atoms involved are reduced in accordance with its order (i.e., for a double bond by 1; for a triple bond by 2). Hence, at any stage it is only necessary to consider bonds between pairs of connected atoms which both still have unfilled valencies. When all of the bonds have been correctly assigned, there will be no atoms left with unfilled valencies.

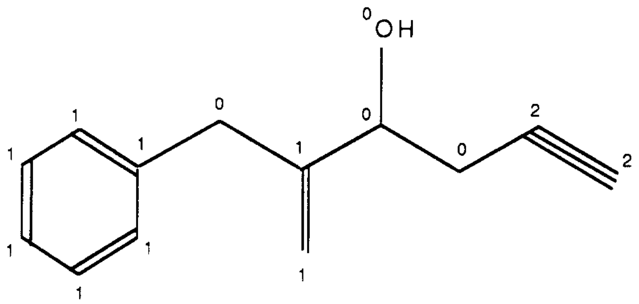

Figure 8. Initial values of NVAL (number of unfilled valencies).

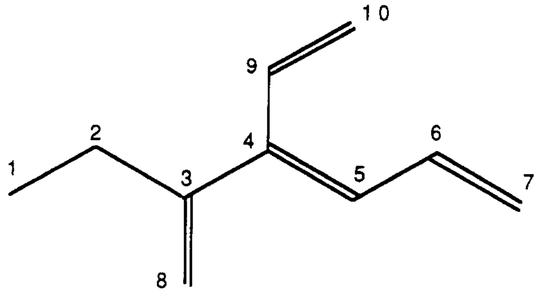

Figure 9. Molecule for which the repeated assignment of unambiguous bonds assigns all bond orders.

(2) Identification of Alkynyl and Allenyl Bonds. Alkynyl and allenyl bonds can be assigned directly from the atom types supplied. Alkyne bonds are found by searching for sp-hybridized atoms. Each such atom must be joined to another sp-hybridized atom in the molecule; the bond between them is given an order of 3. Similarly, an sp^2-hybridized atom which is bonded to just two other atoms, each of which is also sp^2 hybridized, is assumed to be part of a cumulene system, and the two bonds are assigned an order of 2. After assignment the number of unfilled valencies on the atom involved is then updated. For an alkyne bond the NVAL values are reduced by two. For a cumulene the number of unfilled valencies on the central atom is set to zero, and those of the adjoining atoms are each reduced by one.

(3) Assignment of Unambiguous Bonds. For acyclic molecules and for many cyclic molecules, the bond orders may be completely assigned by searching for additional bonds which can be unambiguously assigned. Such bonds exist between pairs of connected atoms which have both unfilled valencies, and additionally, one of the atoms has no other adjoining atoms with unfilled valencies. Because there are no other atoms to which this atom could form a multiple bond, the bond between them can be assigned. Its order is given by the number of unfilled valencies on this atom plus one (order = NVAL(atom) + 1).

Having assigned the bond an order, the numbers of unfilled valencies on the two atoms are updated; the new value for the first atom is set to zero and that for the second atom is set to the difference between its original value and the bond order just assigned. Each atom in the molecule is examined to see whether it is unsaturated and also has a single other adjoining unsaturated atom, and a count is made of the number of bonds assigned during the cycle. If at least one bond is assigned during the operation, the process is repeated again. The calculation is repeated until either no atom remains that has unfilled valencies or until no more bonds can be unambiguously assigned. Figure 9 shows a molecule for which the bond orders can be assigned by repeated application of this process. Here, bonds 3–8, 6–7, and 9–10 are assigned in the first application

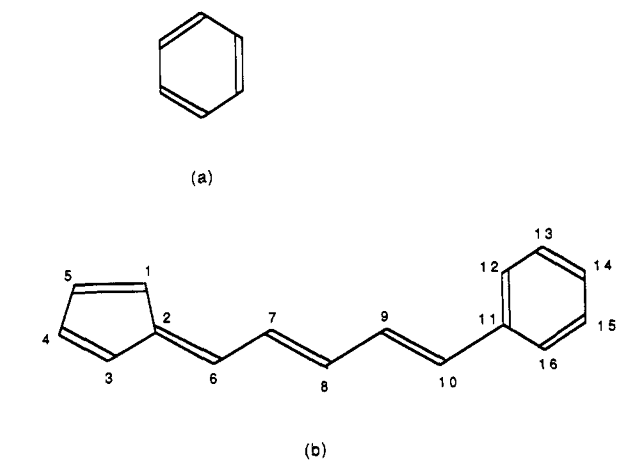

Figure 10. Two molecules to illustrate use of search method in assigning bond orders.

of the method and bond 4–5 in the second cycle.

(4) Assign Bonds Using Search Method. Should there still remain bonds to be assigned in the molecule, a search is made of the space of possible assignments to try and find a satisfactory combination of bond orders which leaves no atom with unfilled valencies. Figure 10 shows two molecules for which this situation can arise; as can be seen they both have conjugated π-systems, with at least one ring.

One atom is chosen from those which still have any unfilled valencies. Such an atom will have at least two adjoining atoms also with unfilled valencies. The bond between this atom and one of its adjoining atoms is then assigned an order of 2 (no other bond order is possible as alkyne bonds have been found earlier). Having made this assignment, the procedure described above in (3) is repeated to make any unambiguous assignments which are now possible. In the case of benzene (Figure 10a), if one of the bonds in the ring is defined as a double bond, then all of the remaining bonds can now be unambiguously assigned in this way. If it is not possible to assign all of the remaining bonds, another atom which still has unfilled valencies is chosen and one of its bonds arbitrarily assigned. Any additional unambiguous assignments are then made. This process of assigning an arbitrary bond and subsequent assignment of unambiguous bonds continues until there are no remaining atoms with unfilled valencies (i.e., there are no more bonds which still need to be assigned a bond order).

Before each arbitrary assignment is made, it is necessary to check that the proposed scheme of bond orders is satisfactory. This is determined by searching for atoms which have unfilled valencies but no adjoining atoms also with unfilled valencies. If such an atom (or atoms) exists, it is assumed that there is an error in the proposed scheme due to an incorrect choice for one of the previous bond assignments. The last atom for which such a choice was made is reexamined, a different assignment is made, and the process continues. Should it not be possible to make a different choice for this previous atom, then a different decision is taken for the atom before it. The algorithm backtracks using a depth-first search method until a satisfactory solution is found.

To illustrate further this search procedure, consider molecule b in Figure 10. Each atom in this molecule has one unfilled valence. Suppose atom 1 is taken as the first atom for which an arbitrary choice is to be made, and that the bond to atom 2 is assigned to be a double bond. Six other double bonds can then be unambiguously assigned (3–4, 6–7, 8–9, 10–11, 12–13, 14–15). However, atoms 5 and 16 still have unfilled valencies but no adjoining atoms to which they could form double bonds. Hence the last assignment decision is deemed incorrect, and a different choice is made. Thus the bond between atoms 1 and 5 is assigned an order of two. As a result, the bonds in the five-membered ring and the acyclic chain can now be unambiguously assigned. However, as none of the bonds in the benzene ring are yet assigned, a second arbitrary assignment is required. Suppose that this is to set the bond between atoms 11 and 12 as a double bond. The remaining bonds in the benzene ring will then be assigned, so completing the determination of bond orders in the molecule.

The backtracking search method could be used alone to assign the bond orders, but for many molecules the need to make any arbitrary choices can be eliminated and thus wasteful computation can be avoided. If it becomes necessary to search the space of possible bond orders, the algorithm described here keeps the number of arbitrary decisions to a minimum.

DETECTION OF STEREOCHEMISTRY AND LOCAL SYMMETRY

It is important that there is no ambiguity about the stereochemistry of any atom in the molecule. COBRA currently supports two means of entering the molecule, as indicated above. The first method is via a datafile produced from some molecular graphics program. Frequently such datafiles contain sufficient (*xyz*) coordinate information to allow the stereochemistry of each atom to be correctly deduced. However, this is not always the case; for example, only a two-dimensional "sketch" may be supplied which would not allow the relative stereochemistry at tetrahedral atoms to be determined. The second method uses the SMILES notation. The SMILES system does allow the precise stereochemical relationships at each atom to be specified (using "isomeric" SMILES), and where it is, the appropriate stereochemical relationships are maintained in all the conformations constructed. However, "isomeric" SMILES is frequently not used. Atoms with ambiguous stereochemistry form the basis for performing a combined *conformational/configurational* search. To do this the program generates all possible combinations of stereochemical assignments for the set of ambiguous atoms and searches the conformational space of each configuration so derived. It is thus necessary to be able to identify those atoms whose stereochemistry is ambiguous, and for this an algorithm to determine *topological equivalence* is used. Two atoms are topologically equivalent if their spanning graphs can be overlaid such that for every atom and bond in the graph of one atom there is an equivalent atom or bond in the graph of the other atom. We have implemented the method of Gasteiger and Jochum[11] with certain modifications to improve its efficiency.

Atoms can display different types of stereochemistry depending on their intrinsic geometry. For double bonds (or cumulenes with an odd number of double bonds), the correct cis/trans relationship between the substituents at the ends of the double-bond system must be defined, unless a pair of substituents from one end of the double-bond system are equivalent. Thus, for example, there is no ambiguity in 2-methyl-2-pentene because the two methyl groups at the end of the double bond are equivalent, and so there is only one possible configuration of this molecule. For 2-ethyl-2-pentene, however, COBRA would generate the two geometrical isomers.

Two tests are applied to atoms with tetrahedral geometry (and to allenes/cumulenes with an even number of double bonds, in which the substituents of the two atoms at the ends of the system show "pseudo-tetrahedral" geometry). First, any asymmetric centers are detected. An asymmetric tetrahedral atom has no pair of adjoining atoms that are topologically equivalent. The second test examines atoms that are members of more than one ring. Consider the atoms of fusion in decalin; in the absence of any stereochemical information *cis-* and

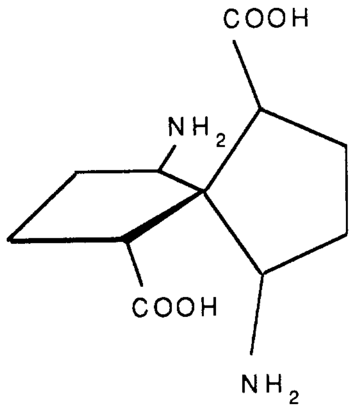

Figure 11. Chiral atom in spiro[4.4.1]nonane derivative is recognized by using the equivalent atom algorithm.

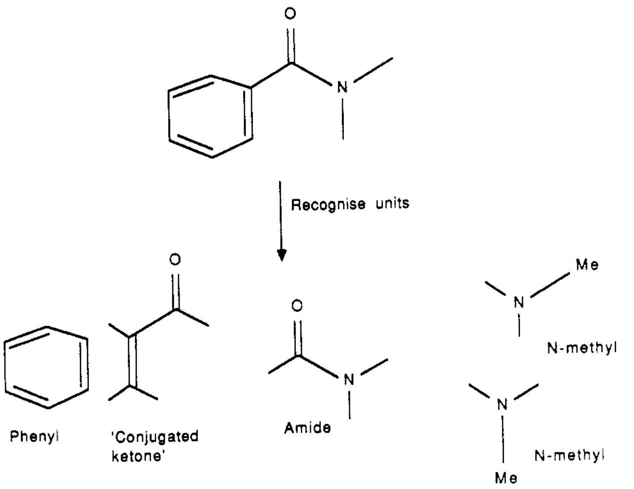

Figure 12. Units recognized in *N,N*-dimethylbenzamide.

trans-decalin cannot be distinguished. If an atom is a member of more than one ring, the two adjoining atoms which are also members of one of these rings are found. The remaining two adjoining atoms are then tested for equivalence; if they are not equivalent, then the atom is added to the list of undefined atoms. This enables certain types of nonasymmetric tetrahedral atoms to be recognized, such as the common ring atom in the spiro[4.4.1]nonane derivative shown in Figure 11.

The equivalence algorithm is also used to determine whether the presence of symmetry in the molecule can enable the conformational search space to be pruned and so reduce the time spent constructing structures. Currently, COBRA does not analyze the complete symmetry of a molecule, but it can recognize some types of local (intra-unit) symmetry. Consider *N,N*-dimethylbenzamide (Figure 12). This molecule is constructed from four units as shown. There are two subconformations available to the "amide" unit (Figure 13), but as the two substituents (labeled a and b in Figure 13) are equivalent, one of these subconformations does not need to be considered during the search of conformational space. For *N*-methyl, *N*-ethylbenzamide, however, both subconformations would be used as the methyl and ethyl carbon atoms are not equivalent.

DETERMINATION OF CONFORMATIONAL UNITS

The recognition of conformational units is the final task performed in the initial analysis of the molecule. COBRA uses a *mapping algorithm* to determine the units present in the molecule. The underlying problem here is closely related to substructure searching, in which a database (typically containing a sizeable number of entries) is searched to find all compounds within it which contain some structural fragment.

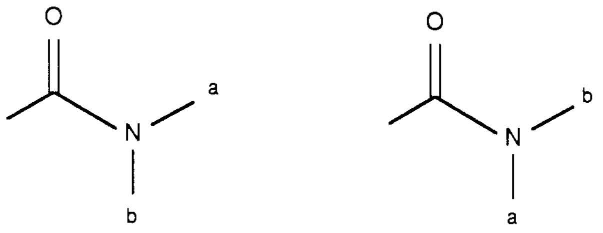

Figure 13. Two subconformations of the "amide" unit.

It is obviously preferable that the time taken for such searches should be as short as possible, as these databases can frequently contain hundreds of thousands if not millions of entries. Consequently, much effort has been expended in devising efficient searching algorithms.[12]

In contrast to a database search, the mapping algorithm used by COBRA tests for the presence of several substructures within a single molecule. The mapping algorithm must determine all occurrences of each unit in the molecule and return a complete list of stereochemically matching atom pairs. Many of the techniques devised for database searching rely upon efficient screening-out procedures to eliminate as many compounds as possible before a more detailed search is performed to identify exact atom-to-atom matches between the substructure and the molecule. This contrasts with our position, in which a few tens of units are mapped onto a molecule, frequently with a high success rate. Moreover, the search terminates when the molecule is completely defined (i.e., it may not be necessary to map all units onto the molecule). It is therefore important that the intensive atom-by-atom matching, which must be performed in order to determine all of the occurrences of a unit within a molecule, is done as efficiently as possible, and consequently, we have directed our efforts to the production of an efficient algorithm which nevertheless gives the desired flexibility.

Information about the units is stored in an ASCII file which is then converted into a direct-access file for use by COBRA. Consequently, it is easy for users to change the unit file when desired. The atom types, bond orders, and relative stereochemistry at each atom in the unit are used during the mapping process. In addition, each unit is designated as *cyclic* or *acyclic*. Cyclic units contain at least one ring, whereas acyclic units possess one or more bonds, each of which must match an acyclic bond in the molecule. The order in which the units are mapped onto the molecule is important, for COBRA should use the largest units possible to construct the molecule (for example, should a *trans*-decalin unit be present in the library this would be used in preference to cyclohexane units for analyzing a steroid). In COBRA the units are considered in an order determined by the number of "heavy" atoms, starting with the unit that has the largest number of such atoms.

Many of the units contain "vector" atoms, which give the system significantly greater flexibility. Their use is best understood by considering an example. The steroid skeleton shown in Figure 14 contains three cyclohexane rings, but they have different substituents. It would be inefficient to store a unique unit for every possible combination of substituents, as the cyclohexane ring adopts the same basic set of conformations in each. The cyclohexane unit therefore has vector atoms in place of the hydrogens in a cyclohexane molecule. A vector atom can match any atom type in the molecule, so when mapped onto the steroid skeleton, the vector atoms correspond to either hydrogen or carbon as appropriate. Further, it is possible that a vector atom may not correspond to any atom in the molecule (for example, if the cyclohexane unit matches a piperidine ring).

The mapping algorithm uses a breadth-first search to determine all the ways in which the graph of the unit could be

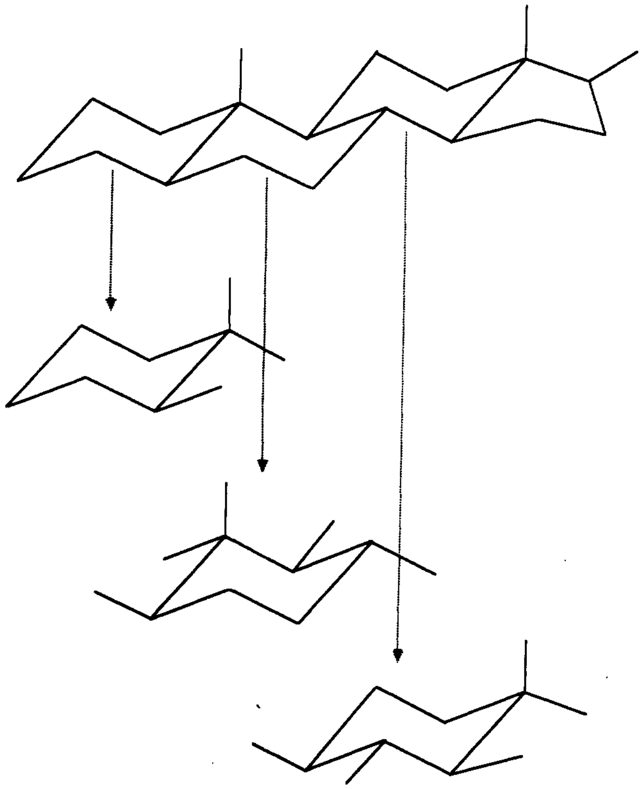

Figure 14. Presence of three cyclohexane units in steroid skeleton which have different substituents.

overlaid on the graph of the molecule. The use of a breadth-first search ensures that failures are found as quickly as possible and facilitates the backtracking necessary to find all solutions. Only heavy (non-hydrogen and non-vector) atoms are considered in the computationally intensive backtracking stage. In order to enhance the efficiency of the search, extensive use is made of filters, or screens, which aim to eliminate as many unproductive avenues as possible. These operate at different levels. For example, the highest level screens check that for each ring in the unit there is a ring of the same size in the molecule. Lower level screens check a variety of other properties, including atom type, bond order, and ring membership. These filters considerably restrict the search space that needs to be considered. When all of the heavy atoms have been matched to molecule atoms, the stereochemistry is checked and molecule atoms are assigned to the vector and hydrogen atoms in the unit (again ensuring a correct stereochemical relationship between the molecule and the unit). If the stereochemistries of the unit and the molecule do match, the sequence of matching atoms is stored and the algorithm backtracks to determine other solutions. Units are mapped onto the molecule until every atom and bond in the molecule has been matched to an atom or bond in one or more units.

The unit library may not contain all of the exact units required to completely describe the molecule. For example, our currently defined library does not contain an explicit piperidine unit, but rather uses the cyclohexane unit as the basis for its conformational behavior. This illustrates COBRA's use of *generalized units*, in which the atom types and bond orders in a unit are permitted to match onto different atom types and bond orders in the molecule. The types of generalization used are closely related to a series of *adjusting algorithms*. These employ a variety of simple geometrical techniques (no minimization) to alter the conformation of a unit so that it more closely approximates that actually required. For example, when forming conformations of piperidine from cyclohexane the bond lengths to the nitrogen would be shortened slightly to make them closer to the correct C–N bond length.

Table I. Currently Defined Atomic Geometry Classes in COBRA

classification	example
saturated tetrahedral	sp^3 carbon, sp^3 oxygen
unsaturated tetrahedral	sulfoxide
saturated trigonal	sp^3 nitrogen bonded to unsaturated atom (e.g., amide)
unsaturated trigonal	sp^2 carbon
unsaturated linear	sp carbon
saturated primary	any halogen
unsaturated primary	carbonyl oxygen
octahedral	cobalt(III)
square planar	nickel(II)

COBRA contains geometrical information about each atom type. For example, sp^3 carbon atoms adopt a tetrahedral geometry, whereas sp^2 carbons adopt trigonal geometry. These geometries form the basis for the different types of generalization which are used and are shown in Table I. COBRA makes up to four attempts to match units onto the molecule, employing successively greater degrees of generalization. In the first pass the exact atom types in each unit are used; here, for example, the cyclohexane unit would only match onto an exact cyclohexane ring in the molecule (as in a steroid). In the first level of generalization each unit atom is permitted to match onto atoms of the same geometrical class (e.g., an sp^3 carbon atom in the unit is allowed to match any saturated tetrahedral atom in the molecule and so the cyclohexane unit would match a piperidine ring). In the second level of generalization saturated tetrahedral atoms are permitted to match unsaturated tetrahedral atoms; saturated and unsaturated trigonal atoms can match each other; and square planar and octahedral atoms can match each other. In the third and final level of generalization saturated tetrahedral atoms are permitted to match both saturated and unsaturated trigonal atoms (for example, this would permit the cyclohexane unit to match a cyclohexanone fragment in the molecule).

The generalizations currently employed are intimately linked to the adjusting algorithms available, and to some extent are dictated by the units present in our template library. The adjusting algorithms give very good to reasonable structures when compared to the minimized conformations,[13] not surprisingly the quality of the structure is determined by the degree of adjustment.[14] The use of generalization means that a much smaller database is required to successfully analyze a wide variety of molecules.

ACKNOWLEDGMENT

A.R.L. thanks the SERC for a studentship; current financial support is provided through the SERC/NATO postdoctoral fellowship scheme. Initial support for the development of COBRA was provided by Glaxo Group Research (Greenford, U.K.). The use of the facilities of the Computer Graphics Laboratory, University of California San Francisco (NIH Division of Resources Grant RR-01081; R. Langridge PI) is gratefully acknowledged. For further details of COBRA, please contact Oxford Molecular, Terrapin House, South Parks Road, Oxford OX1 3UB.

REFERENCES AND NOTES

(1) (a) Dolata, D. P.; Leach, A. R.; Prout, C. K. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 73–85. (b) Leach, A. R.; Prout, K.; Dolata, D. P. *J. Comput. Chem.* **1990**, *11*, 680–693.
(2) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(3) For a comprehensive review see: Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172–187.
(4) Plotkin, M. *J. Chem. Doc.* **1971**, *11*, 60–63.
(5) For some ring systems the inclusion of additional rings may help to prune the combinatorial possibilities involved in the conformational search. An example of this would be the outer six-membered ring in

norbornane. However, the use of rings other than in the SSSR would not be applicable to all cases; as indicated above, the nine-membered cyclononane ring could not be used for this purpose when analyzing hydrindane.

(6) Gasteiger, J.; Jochum, C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43-48.

(7) Roos-Kozel, B. L.; Jorgensen, W. L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 101-111.

(8) One important point to note is that a molecule may have more than one SSSR. For example, cubane contains six four-membered rings, yet there are only five in the SSSR. In the algorithm described here, only one SSSR is determined; to date this has proved sufficient for our purposes. For some molecules it might be preferable to select one specific SSSR rather than use an arbitrary one. It would then be necessary to either extend the algorithm or use some other method.

(9) Wipke, W. T.; Dyott, T. M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140-147.

(10) The concept of valence is used in COBRA to mean the sum of the bond orders for a given atom type. COBRA also distinguishes between charged and neutral atom types-hence sp^3-hybridized oxygen has a valence of 2; whereas, negatively charged sp^3 oxygen has a valence of 1.

(11) Jochum, C.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113-116.

(12) See, for example: (a) Sussenguth, E. H. *J. Chem. Doc.* **1965**, *5*, 3643. (b) Randič, M. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 101-107. (c) Wipke, W. T.; Rogers, D. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 255-262. (d) von Scholly, A. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 235-241.

(13) Leach, A. R. D. Phil. thesis, University of Oxford, 1989 (Chapter 8).

(14) For example, the formation of an acyclic C-N bond from a C-C unit provides a structure very close to the minimum energy conformation for this fragment, but the formation of a cyclic double bond from a single bond leads to some distortion. The algorithms can perform more than one adjustment per unit (e.g., in the formation of morpholine from cyclohexane in which two ring carbons are substituted).

AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names. 1. General Design

J. L. WISNIEWSKI

Beilstein Institute, Varrentrappstrasse 40-42, D-6000 Frankfurt/Main 90, West Germany

Received May 9, 1990

The rules for assigning the systematic name to a structure are complex and frequently lead to ambiguous names. It is this difficulty in assigning names that can be overcome by a program which uniquely translates graphic structures into IUPAC-compatible text names and is readily available as a personal computer tool. The algorithm developed for AUTONOM analyzes the compound's structural diagram, input via a graphic interface, and generates the name purely on the basis of the resulting molecular connection table. This paper describes the design of AUTONOM, presents an analysis of important software and chemical nomenclature solutions adopted during the work on the system, and discusses the system's current accuracy and reliability.

INTRODUCTION

The structural information on a chemical compound can be represented and communicated by a variety of methods. The three most important categories are

chemical nomenclature used to name compounds
formulas and line notations used as shorthand representations of the content and orientation of compounds
structural diagrams used to represent complete graphic information on composition and topology of compounds

The structural diagram conventions are established as an international standard and transcend language barriers among chemists. Chemists are trained to communicate chemical information by using graphical images; however, to nonchemists they are only interesting shapes and strange configurations which convey little understandable information. Various line notations[1] focus on facilitating computer input and structure manipulation. With their human-unfriendly encrypted names and complex systems of rules and conventions (different for each different line notation), which have to be memorized, they create no alternative to either structural diagrams or chemical names. Names which can accurately describe the composition and format of the structure are still vital for a wider audience. In situations where chemical information on a compound needs to be communicated by the spoken or written word, structural diagrams are inappropriate and names are the only alternative. Names are also vital for institutions producing chemical information, such as the Beilstein Institute or Chemical Abstracts Service (CAS), as indexing tools and are important search-key fields for their databases.

Contrary to well-established, standardized, and internationally acknowledged structural diagram conventions, a complete comprehensive grammar for systematic chemical nomenclature does not exist to date. The system of recommendations[2] which have been developed by the Commision on Nomenclature of Organic Chemistry of the IUPAC has not become a universal standard, mainly because of the complexity of the recommended rules, frequent ambiguity in name assignment, and associated continuing use of much quasi-systematic and trivial nomenclature. There is also a reluctance by the chemical industry[3] to perceive the need for fully systematic nomenclature.

For the purpose of obtaining consistency in the selection of preferred names, both CAS and Beilstein devised nondocumented ad hoc subrules which only amplified the difficulty of uniquely naming organic compounds. These subrules were necessary since IUPAC rules frequently allow more than one name for a given chemical. As a result, both institutions revised the IUPAC system and created their own "systematic" IUPAC-compatible rather than IUPAC-sanctioned nomenclatures. In addition, trivial or trade names, being shorter and more concise, have successfully replaced systematic names for

0095-2338/90/1630-0324$02.50/0 © 1990 American Chemical Society