Heuristic Approach for Displaying Chemical Structures

CRAIG A. SHELLEY

Research Laboratories, Eastman Kodak Company, Rochester, New York 14650

A heuristic procedure is described that displays chemical structures from connection tables. Most ring systems are displayed conventionally, oriented as a chemist would draw them. Novel heuristics avoid overlapping atoms and bonds and facilitate similar representations of the structural diagram for identical or similar structures.

INTRODUCTION

The increasing application of computers in chemical research is leading to an increased need for graphical communication of chemical structure between the chemist and the computer. Most efforts at improving this communication have been directed toward input of chemical structure to the computer (for an advanced system see ref 1). In many systems, the difficult problems encountered in displaying a structural diagram directly from the connection table can be circumvented by simply storing the essential information required to recreate the structural diagram in addition to the connection table.[1] However, many programs create connection tables or substantially change them. These programs require solution of the complex problem of directly displaying the structural diagram from a connection table.

Four characteristics of a good procedure to display chemical structures from a connection table can be identified: (1) Ring systems, if any, must be perceived and assigned coordinates that depict the ring system in an easily recognized format. Bonds should not be longer than necessary. (2) Coordinates should be assigned to the acyclic atoms of the structure that minimize atom crowding and the number of overlapping bonds. (3) Complete structures should be correctly oriented. Steroids are a good example, for which the ring system, **1**, is always

1
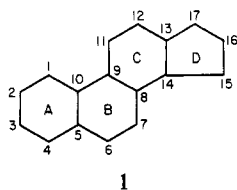
drawn with the A ring on the left and below the D ring. (4) Similar structures should be oriented the same way so that the chemist can quickly perceive differences and similarities in a list of structures.

One approach to display chemical structures, developed by Chemical Abstracts Service, is oriented toward chemical data base systems.[2] The program displays some groups of atoms as character strings, e.g., "CO$_2$H" for the carboxyl group. It has a routine to detect and avoid the overlap of these character strings. The program assigns coordinates to ring systems by use of a large dictionary of "template" ring systems, a method first used by Zimmerman.[3] The dictionary approach facilitates the uniform orientation of structures with common ring systems. The major disadvantage of this approach is lack of generality. Structures with ring systems not in the dictionary cannot be displayed.

Carhart developed an approach to structure display that does not require a dictionary of ring systems.[4] The program is designed to generate structural diagrams for alphanumeric terminals (not graphics terminals). Because of this limitation alone, the structures are frequently difficult for the chemist to interpret.

The Carhart procedure assigns two-dimensional coordinates to atoms by calculating three-dimensional coordinates with a simple model-building procedure. The three-dimensional image, as viewed from its most planar side, is then projected into two dimensions. The method is computationally simple and minimizes overlap problems. However, it is computationally expensive, ignores structure orientation, and often assigns poor coordinates to ring systems. Despite these limitations, it has been important in the use and development of many programs in chemistry.

Two other procedures to display chemical structures without use of a ring-system dictionary have been developed by Cox[5] and Zippel and co-workers.[6] Neither program uses the model-building technique of Carhart, but, as in the Carhart program, they cannot maintain similar orientations for similar structures or orient ring systems correctly.

The approach used in the structure-display procedure described in this paper uses novel heuristics to generate good coordinates for most chemical structures. It facilitates the orientation of diagrams in a consistent and conventional manner. The program contains six major steps: (1) initial feature perception, (2) ring-system perception and assignment of relative coordinates to atoms in each ring system, (3) assignment of absolute coordinates to atoms, (4) coordinate refinement, (5) additional feature perception and structure manipulation, and finally, (6) plotting the chemical structure. The structural diagrams are conventional "stick diagrams" with the contraction of long chains of methylene groups to character strings. The program handles Chemical Abstracts Service "tautomer" bonds by "denormalizing" them to display one tautomer of the connection table. The procedure is normally used with a raster graphics terminal, but it has also been interfaced to the Carhart structure-display program to produce output on character-oriented devices.

PROCEDURE

Initial Feature Perception. Methods to perceive cycles have been reviewed by Wipke and Dyott.[7] These procedures perceive cycles as sets of atoms without identifying the connectivity. Because structure display requires connectivity for the cycles, a modified procedure was developed. The procedure is as follows: (1) assign a spanning tree of the molecule (an acyclic connected subgraph that contains all atoms in the structure) to GRAPH; as the spanning tree is grown, put bonds not contained in the spanning tree (the ring bonds) into STACK; (2) get the top bond in STACK (if STACK is empty, all cycles have been perceived); (3) search depth first in GRAPH for all paths between the pair of atoms for this bond, adding each path to the list of cycles in the structure; (4) add the ring bond to GRAPH; go to step 2. The breadth-first perception of a spanning tree for the steroid ring system starting at atom 1 (see structure **1**) is illustrated in Figure 1. The last three steps of the cycle-perception procedure are illustrated in Figure 2 for the steroid ring system. The cycle-perception procedure produces a nonredundant and complete list of cycles for a structure.

Many cycles, however, cannot be used in assigning coordinates to cyclic atoms. Any cycle that contains all the atoms
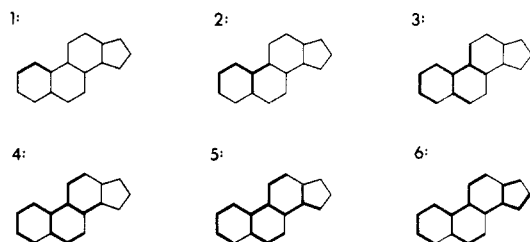
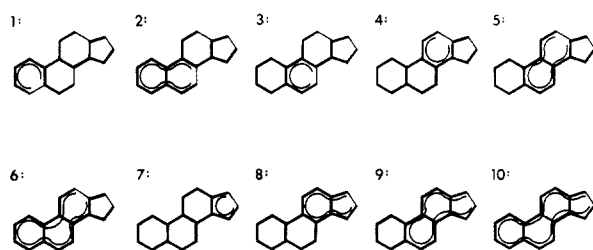**Figure 1.** Breadth-first assignment of a spanning tree.

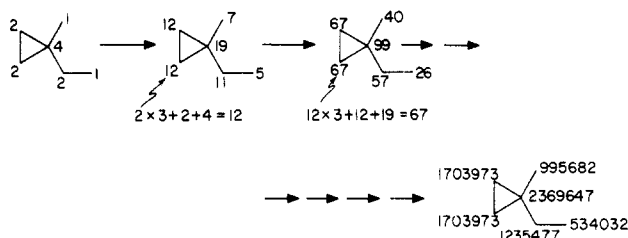**Figure 2.** Cycle perception using a spanning tree.

Figure 3. Calculating ATCD numbers.

of another cycle is removed from the list of cycles. In the decalin ring system, the two six-atom cycles are kept, and the ten-atom cycle is deleted. In the remainder of this paper, the term "cycles" refers to this reduced set of cycles.

In the next step of the feature-perception process, graph-invariant codes are generated for atoms, cycles, and ring systems that reflect, in part, their environment in the structure. (Features that are independent of the connection table's sequence numbers are called graph invariant.) These graph-invariant codes will be used to (1) minimize atom crowding and bond overlap in assigning absolute coordinates to the atoms, (2) correctly orient ring systems, (3) orient similar structures the same, and (4) generate ring-system coordinates independent of assigned sequence numbers.

The code assigned to each atom, ATCD, is related to the Morgan algorithm's connectivity number.[8] The ATCD numbers are generated by a procedure similar to that previously suggested for extending the Morgan code.[9] First, the ATCD for each atom is initialized to the number of attached atoms (degree). Next, a new code is calculated for each atom by multiplying the previous code of the atom by 3 and adding the sum of the ATCD numbers of adjacent atoms. This partitioning step is repeated seven more times to obtain the final ATCD numbers. The procedure is illustrated in Figure 3 for a simple structure. The multiplication factor of 3 is used to reduce the chance of assigning two atoms with different codes the same code in the next iteration. A small factor is used to avoid arithmetic overflow. The total number of iterations, 8, is based on the observation that the number of codes seldom increases for most structures as the number of iterations is increased above 8. Like the Morgan connectivity numbers, the ATCD numbers get larger for atoms located closer to the "center" of the structure. This characteristic of the ATCDs is used to minimize atom crowding. The ATCD numbers are also used to facilitate consistent orientation for structures.

The code CYATCD is generated for cyclic atoms. CYATCD numbers are generated in the same fashion as
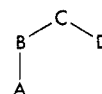

Figure 4. Cycle tree for the steroid ring system.

ATCD numbers except that the acyclic atoms are ignored. The CYATCD numbers are used to calculate codes for cycles, CYCD, and ring systems, RSCD. The CYCD of a cycle is the sum of the CYATCD numbers for all atoms in the cycle. The RSCD of a ring system is the sum of the CYATCD numbers for all atoms in the ring system. CYATCD, CYCD, and RSCD numbers are used to assign similar orientations for similar structures. CYATCD and CYCD numbers are also used to assign the same coordinates to a ring system even though the sequence numbers differ. RSCD numbers are also used to facilitate the correct orientation of ring systems.

Ring-System Perception and Assignment of Relative Coordinates. Two cycles are defined to be in the same ring system if they have a common atom. It follows that two cycles are also in the same ring system if they each have at least one atom in common with cycles in the ring system. Cycles do not need a common bond; e.g., the two three-member cycles in spiro-pentane are in the same ring system. By the above rule, cycles in the same ring system are grouped together. Next, a complexity number is assigned to each ring system. The complexity number n is given by eq 1, in which c is the number

$$n = c - b + a - 1 \qquad (1)$$

of cycles, b is the number of bonds, and a is the number of atoms in the ring system. The steroid ring system 1 contains 4 cycles (three six-atom cycles and one five-atom cycle; see cycle-perception above), 20 bonds, and 17 atoms. The complexity number is zero. Ring systems with a complexity number of zero will be referred to as "simple" ring systems. The complexity number of a ring system is used to select the procedure for generating its coordinates.

For simple ring systems, the coordinates of each cycle can be generated by use of a regular polygon as a "template". For the same ring system to be assigned the same coordinates, a processing order must be established that is independent of the sequence number assignment. This order is determined by the connectivity of the cycles in the ring system. Cycles in simple ring systems can have three relationships: fused (two atoms in common), spiro (one atom in common), or uncon-nected (no atom in common). The connectivity of cycles in simple ring systems can be defined by use of a tree in which each cycle is a node in the tree. The tree for the steroid ring system is illustrated in Figure 4 (see 1 for naming conventions of the steroid rings). The root of the tree is the cycle with the largest CYCD number. The tree is grown by a breadth-first procedure. The descendants of each node are ordered according to the magnitude of the CYCD numbers of the descendant cycles. The tree for a simple ring system is used to order the processing of the cycles during coordinate generation.

In addition to treating simple ring systems, the program generates coordinates for "bicyclic" and much more complex ring systems. Bicyclic ring systems such as norbornane have the following properties: (1) the complexity number must be 1; (2) two cycles in the ring system must have at least three atoms in common (a link); (3) removing the link from these cycles creates a third cycle that must also be contained in the ring system. When bicyclic ring systems are processed, the three links between the bridgehead atoms are extracted, and the three corresponding cycles are deleted from the ring system. The connectivity of cycles connected to each of the three links is perceived and represented as three cycle trees. The links and corresponding cycle trees will be used to generate relative coordinates for the bicyclic system.

DISPLAYING CHEMICAL STRUCTURES

J. Chem. Inf. Comput. Sci., Vol. 23, No. 2, 1983 **63**

Figure 5. Procedures for assigning coordinates to cycles.

Ring systems that are not simple or bicyclic are simplified, if possible, to a subring system that is either simple or bicyclic by removing "strategic" cycles. All strategic cycles contain a common atom. This atom must meet two criteria. First, it must be attached to exactly two cyclic atoms. Second, its removal from the ring system must produce only one subring system. If the ring system does not contain an atom meeting these conditions, the current version of the program cannot display the structure. Among atoms that meet these criteria, one is selected that is contained in as many cycles as possible and has the smallest CYATCD number. For example, consider ring system **2**. The structure contains four cycles: three

2
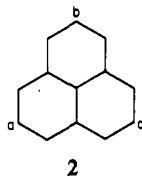
6-member cycles and a 12-member cycle that traces a path around the perimeter of the ring system. Atoms a–c meet the two criteria. The selection among them is arbitrary. Cycles containing the atom are removed from the ring system. The smallest cycle removed is put on a cycle stack. Cycles are removed until a simple or bicyclic subring system is encountered. In ring system **2** the 12-member cycle is deleted from the ring system, and a 6-member cycle is put on the cycle stack. This method may be viewed as "peeling" cycles from the ring system until a subring system is encountered that is simple, bicyclic, or cannot be processed further. Coordinates are generated for the "peeled" cycles in the stack after coordinates have been assigned to the other cyclic atoms.

Each cycle is assigned coordinates by using a regular polygon that contains an equal number of vertices. The coordinates of the regular polygon are generated by one of two procedures. One, SPIRO, is given the coordinate of one vertex and the angle from this vertex to the center of the polygon. The coordinates of the polygon are constructed in a clockwise direction from the vertex (see Figure 5). The other procedure, FUSE, is given the coordinates of two adjacent vertices. The coordinates are generated for the polygon in a clockwise direction starting with the first vertex and proceeding to the second (see Figure 5).

The layout of the steroid ring system is illustrated in Figure 6. The coordinates of the root cycle, cycle "C", are assigned by use of SPIRO. The sequence numbers of the C cycle are rotated until the first atom has the largest CYATCD (the largest ATCD is used to break ties). If the second atom does not have the larger CYATCD of the two atoms attached to the first, the cycle is inverted and rotated until the first atom has the largest CYATCD. Now, the sequence numbers of the cycles are in the order 8-9-11-12-13-14. The first atom, 8, is assigned the coordinate (0,0). The other atoms are assigned coordinates in a clockwise direction by SPIRO using an angle of 90° to the center of the hexagon. The other cycles in the tree (Figure 4) are assigned coordinates according to a depth-first traversal of the tree. Before coordinates are assigned to each fused cycle, the sequence numbers of the cycle are rotated and/or inverted so they are in clockwise order. This process is illustrated for the B, A, and D cycles in Figure 6. Before coordinates are generated for a spiro cycle, the cycle is rotated until the connecting atom is the first sequence
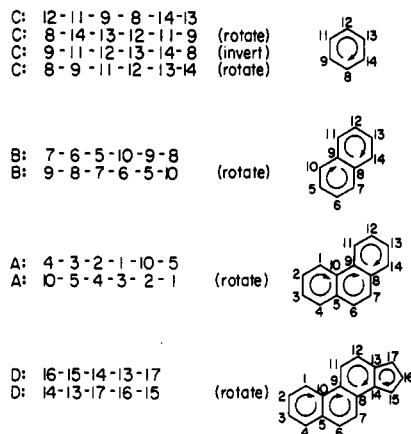

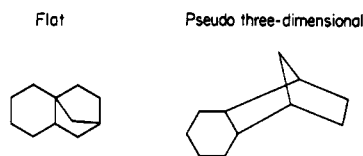**Figure 6.** Coordinate generation for the steroid ring system.

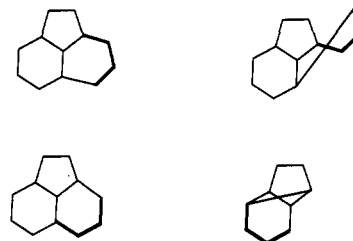**Figure 7.** Representations of bicyclic systems.

Figure 8. Assignment of coordinates to "peeled" cycles.

number. If the second atom does not have the larger CYATCD (the larger ATCD is used to break ties), the cycle is inverted and then rotated until the connecting atom is the first sequence number.

For bicyclic ring systems, coordinates are first assigned to the three links. If the two bridgehead atoms have exactly three cyclic atoms attached, the links are assigned pseudo-three-dimensional coordinates according to the procedure of Cox[5] (Figure 7); otherwise, the links are assigned a flat representation (Figure 7). The position of each link in the bicyclic network is selected by using the sum of the CYATCD numbers and the sum of the ATCD numbers for atoms in each link. For each cycle tree rooted at a link, the cycles are assigned coordinates depth-first. The layout procedure for these cycles is the same as in simple ring systems.

For complex ring systems, coordinates for the simple or bicyclic subring system previously perceived are first generated. Then coordinates are generated for atoms in peeled cycles. Peeled cycles are put back on the subring system in the reverse order of removing them. The program selects the preferred layout of a peeled cycle using the "energy" E of each layout. E is given by eq 2, in which r_{ij} is the distance between atoms

$$E = \sum_{i,j} 1/r_{ij}^2 \qquad (2)$$

i and j. Of the four potential layouts (Figure 8), the one with the lowest energy is used.

Assignment of Absolute Coordinates. Absolute coordinates are assigned breadth first to atoms by starting from an initiating nucleus. A procedure is used which causes atoms to flee each other. The fleeing tendency is increased for cyclic atoms, atoms located near the "center" of the molecule, and atoms that are constitutionally crowded. This eliminates atom

crowding and reduces bond overlap for most molecules.

The first step selects and assigns absolute coordinates to the initiating nucleus of atoms. For acyclic molecules, the nucleus is the atom with the largest ATCD number and the adjacent atom with the largest ATCD number of the attached atoms. For cyclic molecules, the initiating nucleus is the ring system with the largest RSCD (the sums of the ATCD codes for atoms in each ring system are used to break ties).

A nucleus ring system is assigned a "conventional" orientation by rotating and, in some cases, flipping the ring system. A ring system is flipped by changing the sign of the x coordinates. For the ring system to be assigned a "conventional" orientation, the user must have previously assigned the ring system a flip flag (yes or no) and a rotation angle (0–359°). The ring system (identified by its RSCD), and its associated rotation angle and flip flag, is stored in a small file, after which the ring system will be depicted with the preferred orientation.

A potential disadvantage of using the RSCD as a ring-system identifier is that different ring systems can have the same number. To estimate the probability for occurrence of this problem, a file of all possible ring systems containing eight atoms and two to four ring bond equivalents was created by use of the CASE program.[10] The RSCD number was determined for each of the 212 ring systems in this file. Four of the ring systems (**3–6**) contained the same RSCD number.

3   4   5   6
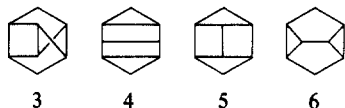
The sets of CYATCD numbers for the atoms in each of these four ring systems are also identical. Of the remaining ring systems, three had the same RSCD, 13 pairs of ring systems had the same RSCD number, and 179 had a "unique" RSCD. In other words, 18 of the 212 ring systems could not have their orientation uniquely defined. The frequency of RSCD collisions is low enough to justify its use in this application.

Previously entered ring-system codes are stored by use of a hash table and linked-list data structure. The RSCD modulo 251 is used as a pointer to a hash table with 251 indices (0–250). A hash-table entry of 0 indicates that the ring system is not in the file; otherwise, the entry is a pointer to the head of a linked list of ring systems that hash to the same index. Each linked list is sequentially searched for the RSCD number of the ring system. The ring system is then oriented according to the corresponding flip and rotation parameters.

After absolute coordinates are assigned to atoms in the nucleus, the remaining atoms are assigned coordinates breadth first. The fleeing heuristic is as follows. If there are more positions than atoms, then the low-energy positions are used. When two or more atoms compete for the same position, (1) cyclic atoms are always assigned lower energy positions than acyclic atoms (rule 1 precedes rule 2) and (2) the atom with the larger ATCD number is assigned the lower energy position (constitutionally crowded atoms have large ATCD numbers). The procedure is as follows: (1) atoms assigned absolute coordinates and adjacent to atoms not assigned coordinates are put on a list; (2) the list is ordered with cyclic atoms at the top of the list with descending ATCD numbers and acyclic atoms at the bottom of the list with descending ATCD numbers; (3.i) let $k = 1$ (top of the list); (3.ii) a list of atoms adjacent to atom k and not previously assigned coordinates is created and ordered with cyclic atoms at the top of the list with descending ATCD numbers and acyclic atoms at the bottom of the list with descending ATCD numbers; (3.iii) the "energy" P of each potential coordinate position is calculated from eq 3, in which A is the ATCD number normalized to a

$$P = \sum_i A_i/r_i^2 \qquad (3)$$

real number between 0.5 and 1.5 and r_i is the distance between atom i and the potential position; (3.iv) the adjacent atoms are assigned coordinates such that atoms at the top have lower energy positions; in addition, when a cyclic atom is assigned a coordinate, the corresponding ring system is also assigned coordinates to give the lowest energy R, as calculated from eq 4 (the ring system can be rotated around the connecting

$$R = \sum_{i,j} A_i A_j/r_{ij} \qquad (4)$$

bond by 180°); (3.v) let $k = k + 1$; (3.vi) repeat steps 3.ii–3.v until all atoms in the list have been processed; (4) repeat steps 1–3 until all atoms have been assigned absolute coordinates.

Coordinate Refinement. Although the fleeing heuristic eliminates crowding problems for most molecules, the coordinates of some structures can be further improved by rotating substituents, bending substituents, and stretching acyclic bonds. The choice among potential rotations, bends, and stretches is based on the energy of the structural diagram. The coordinate refinement procedure is implemented as follows: (1) set FLAG to "false"; (2) create a list of all acyclic bonds that reside in a path between each pair of close atoms (with 0.6 standard bond lengths); (3) if there are no bonds in the list, then the refinement process is complete; (4) calculate the energy M of the diagram according to eq 5; (5) if FLAG is "true", go to

$$M = \sum_{i,j} 1/r_{ij}^2 \qquad (5)$$

step 7; (6) rotate a substituent corresponding to each bond and save the new energy and coordinates if the energy decreased; set FLAG to "true"; go to step 2; (7) bend substituents for each acyclic bond and save the new energy and coordinates if the energy decreased; stretch each acyclic bond and save the new energy and coordinates if the energy decreased; set FLAG to "false"; go to step 2. FLAG is used to enable alternation between rotating substituents and bending or stretching substituents. Rotation is preferable to bending or stretching because rotation does not distort the structural diagram.
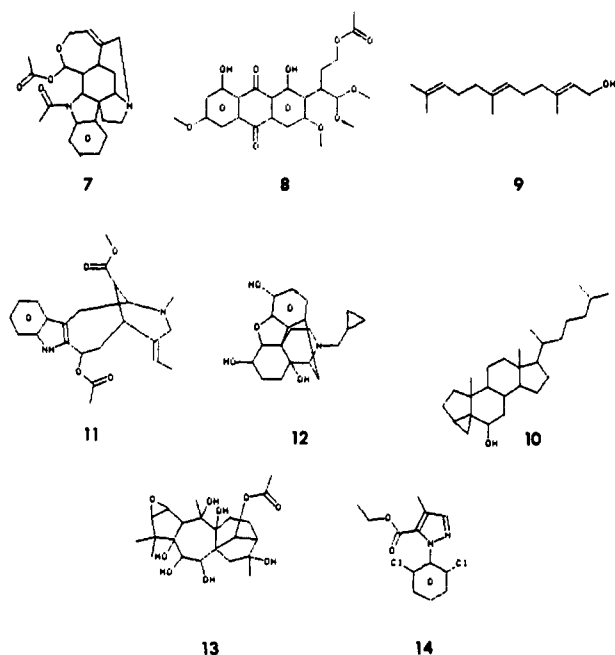
Additional Feature Perception and Manipulation. Connection tables submitted to the structure-display procedure may contain single, double, triple, aromatic, or "tautomer" bonds (as designated by Chemical Abstracts Service[11]). Six-member cycles containing "aromatic" bonds or alternating double and single bonds are detected to allow plotting the conventional circle in the center of the ring. CAS tautomer bonds are "denormalized" to single and double bonds. Oxygen has priority over nitrogen for double-bond placement. Acyclic chains of at least four methylene groups are contracted to reduce the size of some structural diagrams. The chains may be terminated by two methyl groups (normal alkanes), one methyl group (normal alkyl substituents), or no methyl groups (methylene links).

Plotting the Chemical Structure. The structure-display procedure plots bonds first and then heteroatom groups and methylene groups (including any terminating methyl groups). By plotting the features in this order, it is not necessary to plot short bonds to avoid overlap with the character strings. On a raster display, plotting over a previous image simply replaces the old image with the new one. The plotting routine selects a low-energy position to locate charge. However, the preferred location for charge is always above and to the left of the charged atom. Structures **7–14** illustrate several structural diagrams as they appear on the raster graphics display.

DISCUSSION

Because the structure display routine is a heuristic procedure, there is no guarantee that high-quality structural diagrams will be produced for all structures. However, high-

DISPLAYING CHEMICAL STRUCTURES

J. Chem. Inf. Comput. Sci., Vol. 23, No. 2, 1983 **65**

Chart I

7   8   9

11   12   10

13   14

quality structural diagrams are obtained most of the time. Some examples of structures created by the program appear as **7–14** (Chart I). The atoms of the terpene **9** are spread out because of the fleeing heuristic. The methyl groups appear on the same side of the chain for the same reason. The orientation of the steroid **10** is the same as the chemist would draw it. The morphine ring system in **12**, a rather complex one, is clearly shown. The pseudo-three-dimensional nature of the ring system in **11** is consistent with common conventions, as is the flat representation of the ring system in **13**. The acetate group in **13** is moved away from the other atoms during the refinement process. A bond angle in **14** is widened to remove overlap between the carbonyl and chloro groups.

On the basis of experience with the program, it appears that the most serious limitation of the procedure is its inability to generate good coordinates for some complex ring systems. For example, ring system **15** appears as **16**, which is misleading

15   16
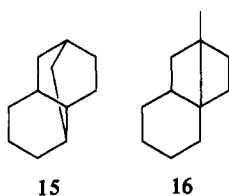
because of overlapping bonds. This limitation could be reduced by improving the code for generating ring-system coordinates, adding a small template dictionary for complex subring systems, or including a model builder that could generate coordinates for complex subring systems.

There are two minor limitations in the procedure's ability to display structures with good orientation. In some structures, e.g., nucleosides, two linked ring systems have a preferred orientation. Since the procedure orients the nucleus ring system independently of all other ring systems in the structure, a nucleoside may not be oriented as the chemist would normally draw it. Second, the orientation of structures is independent of the placement of heteroatoms and multiple bonds. For example, pyridine is displayed with the nitrogen at any of the six vertices in the ring, the position being dependent on the numbering of the connection table.

The structure-display procedure is written in FORTRAN. It requires about 256K bytes of memory when dimensioned for structures containing up to 255 nonhydrogen atoms. The procedure contains more than 50 FORTRAN subroutines and functions. Much of the I/O is dependent on the host computer (Data General MV/8000, FORTRAN 77, AOS/VS) and the raster graphics terminal (Hewlett-Packard 2647A). Contact the author regarding distribution of the program, which will be made available for noncommercial applications.

REFERENCES AND NOTES

(1) Howe, W. J.; Hagadone, T. R. "Progress Toward an On-line Chemical and Biological Information System at the Upjohn Company". In "Retrieval of Medicinal Chemical Information"; Howe, W. J., et al., Eds.; American Chemical Society: Washington, DC, 1978; *ACS Symp. Ser.*, No. 84, Chapter 8, p 107.

(2) Dittmar, P. G.; Mockus, J.; Couvreur, K. M. "An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 186–192.

(3) Zimmerman, B. L. "Computer-Generated Structural Formulas with Standard Ring Orientations". Ph.D. Dissertation, University of Pennsylvania, 1971.

(4) Carhart, R. E. "A Model-Based Approach to the Teletype Printing of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 82–88.

(5) Cox, B. D. "Computer Program STR3". Ph.D. Dissertation, Arizona State University, 1973.

(6) Zippel, M.; Mowitz, J.; Kohler, I.; Opferkuch, H. J. "SPEKTREN—A Computer System for the Identification and Structure Elucidation of Organic Compounds". *Anal. Chim. Acta* **1982**, *140*, 123–142.

(7) Wipke, W. T.; Dyott, T. M. "Use of Ring Assemblies in a Ring Perception Algorithm". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140–147.

(8) Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, *5*, 107–113.

(9) Shelley, C. A.; Munk, M. E. "Computer Perception of Topological Symmetry". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 110–113.

(10) Shelley, C. A.; Munk, M. E. "CASE, A Computer Model of the Structure Elucidation Process". *Anal. Chim. Acta* **1981**, *133*, 507–516.

(11) Mockus, J.; Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18–22.