Aǹ Interactive, Versatile, Three-Dimensional Display, Manipulation and Plotting System for Biomedical Research

RICHARD J. FELDMANN, STEPHEN R. HELLER,* and C. R. T. BACON
Division Of Computer Research and Technology,
National Institutes of Health, Public Health Service, Bethesda, Md. 20014

Received August 8, 1972

Computer graphics provides a valuable tool for the representation and a better understanding of structures, both small and large. Accurate and rapid construction, manipulation, and plotting of structures, such as macromolecules as complex as hemoglobin, are performed by a collection of computer programs and a time-sharing computer. The programs and techniques are described, and examples of the system are given.

Using the concept of a centralized time-sharing computer for interactive research with computers in chemistry, we have developed a number of systems for biochemical and biomedical research workers.[1] This present work describes an X-ray modeling system.

Techniques for the three-dimensional display of molecules have attracted considerable interest.[2-13] Levinthal and coworkers, thè first group to explore this area, have devised a large, sophisticated display system consisting of an AGT/50 Graphics Display Computer connected to a large high-speed computer. The AGT/50 is an expensive display terminal with special hardware which provides for the hardwired rotation of vectors displayed. Later work by Barry and coworkers also used a display with excellent hardwired manipulation controls and a large computer system. Perkins *et al.* and Meyer have developed systems which use a smaller display (with no hardwired capability) and a minicomputer. To compensate for a lack of both a versatile display and a large computer, Meyer uses a color display to help simulate a three-dimensional effect.

The approach described here uses computer programs with both kinds of display; those that have hardwired manipulation facilities and those which do not. The software, written originally in FORTRAN, is flexible so that the programs run on both an Adage AGT/30 display and a DEC 340 display. Some of the programs have been rewritten in SAIL to take advantage of its excellent string manipulation and dynamic storage allocation characteristics. Both types of displays are linked to a time-sharing DEC PDP-10 computer. The programs require 12–55,000 words of the 36-bit word PDP-10 computer, depending upon the size of the molecule displayed. At present, molecules of up to 4000 atoms can be displayed and manipulated.

METHODS

The options available with the system are given in Table I. The input for the programs comes from x,y,z cartesian coordinates, usually obtained from X-ray crystallography. Both the small molecule and protein libraries from the X-Ray Crystal Data Centre, Cambridge, England will be available for use in the system.[14] However, good approximate coordinates can be obtained using an x,y,z coordinate generator program.[15] In addition, the latter program allows a scientist to start with the X-ray coordinates of a basic structure and then add on atoms/groups at any desired site on the molecule. The file structure for a given molecule, such as a protein, contains the atom type (e.g., CA, x-carbon), the amino acid type (e.g., HIS, Histidine), the three coordinates (x,y,z), the atom number, and its connectivity (i.e., the atoms to which it is connected).

Molecular Rotation and Translation. A molecule can be rotated and translated in the x, y, and z axis by software programs on the DEC 340 display or by control knobs on the AGT/30. The real-time hardwired rotation/translation of the AGT/30 allows for the convenient and rapid orientation of a large macromolecule for viewing and for the plotting of a desired view. The real-time AGT/30 rotation gives an excellent perception of the three-dimension effect, without the need to display a stereo pair image.

Bond Rotation. The ability to rotate about a specific bond is very useful for a number of applications, including the measurement of possible van der Waals contact. To make changes in the conformation of a molecule, such as for energy state calculations using the CNDO/INDO program[16] or the MINDO program,[17] bond rotation is a necessity. The algorithm used is based on the method of

Table I. Options available in the DCRT/CIS Modeling System

Molecular rotation in the x, y, and z axis
Bond rotation around a specified bond
Molecular translation in the x, y, and z axis
Bond distance measurements
Bond angle measurements
Generation and alteration of bond connectivity tables
Clipping of structures
Saving modified structures
Atom lettering/numbering and group lettering/numbering
Calcomp plots of the display image
ORTEP plots of the display image
CNDO/INDO Energy State Calculations of a given conformation

* Author to whom correspondence should be addressed.

Sproull[18] and involves finding a transformation matrix which is applied to the coordinates of all the atoms in the "group" one wishes to rotate.

Measurements of Bond Distances and Bond Angles. Given the x,y,z coordinates of a pair of points, the simple geometric formula of square root of the sum of squares between the points gives the bond distance. Another geometric equation using three points will give the bond angle.

Connectivity of Atoms-Generation and Alteration. X-Ray crystallographic data give very precise information on the location of atoms, but say nothing as to what atoms are connected to one another. With bond connectivity information, structures would look like a confusing mass of points. To obtain bonds between atoms, a number of routines are available. The simplest requires minimum and maximum bond distances as input and then proceeds to generate a bond from every atom to every other atom within the specified range. This is done by computing the distance between two atoms using the usual geometric equation, and if the resulting distance is between the minimum and maximum distances, the connection table is changed to reflect this. The routine does have some notion of valence, but normally requires some fixing after being used the first time. A second method of bond generation uses van der Waal radii to link atoms to one another. For the case of a protein structure, a routine has been written which syntactically generates the correct bond connectivity, from a library of connectivities of amino acids.[19]

In those cases where there is an incorrect bond or a missing bond, the error can be corrected with a routine that alters the connection table of the file structure. To use this, one must know the atom numbers of the two atoms which need the removal or addition of the bond. Alternatively, the file can be modified and connections corrected using SOS, the standard PDP-10 text editor.

Labeling of Atoms and Groups. As indicated in the file structure, atoms and groups are both numbered and labeled. Thus a particular atom of Ribonuclease-S can have an atom type designation: CA-for alpha carbon; an atom number of 87 for its location in the file structure; a group type-HIS and a group number-12. Depending upon the particular aspect of the Ribonuclease-S enzyme one wishes to examine, either all of the 124 CA atoms may be labeled and shown on the screen, or all of the four Histidine residues may be designated. Only every tenth CA atom can be shown, or only every tenth atom number (1,11,21, . . ., 951) can be displayed. The lettering/numbering can be readily turned on and off and changed back and forth.

Clipping and Scaling of Structures. While the display of Ribonuclease-S (952 atoms) is quite impressive, one normally is interested in a particular section of the entire molecule, such as the active site. In addition to finding it difficult to look at a small section of a large molecule, the full structure takes up a great deal of computer space, computer time, and display capacity. Both cubic and spherical clippings are available on the AGT/30 while only the former is on the DEC 340 Display. The cubic clipping gives a much better perspective than the spherical clip; however, the spherical clip does convey a better feeling as to how far away atoms are from the center of the clipped view. To perform a clipping, one specifies the atom about which the clipping should occur. Figure 1 is a clipped section of Ribonuclease-S, clipped about the alpha carbon of HIS-12.

In addition to clipping a section of a large molecule, the window one uses to view a structure can be changed so that the structure can appear to be larger or smaller (i.e., closer or further away from the viewer).

Another useful option in the system is a routine that selectively goes through the file structure to clip or pick out substructures such as the backbone of a protein, all atoms of a particular acid type (e.g., all four Histidines in Ribonuclease-S), and all atoms of a given type (e.g., all amide nitrogens, all disulfide sulfer atoms). After the substructure is obtained, the file is cleaned up by generating the new connectivity and renumbering the atom numbers.

Saving Modified Structures. In almost all of the preceding operations, the resulting coordinates and related file structure information is different from that which it started out to be. Rather than keeping track of all the operations and repeating them every time a user started up the system, the modified file structure can be saved on the PDP-10 disk at any time. This allows the scientist to break off his work at any point, return at a later date and start up from where he finished. The new file can replace or write over the old file on the disk, or it can be given a new name and saved as a completely new file.

Plotting. The value of the system would be greatly reduced if there were no mechanism for producing a printed result of the scientist's work on the display. There are two routines for obtaining plots of the display image. The first employs a PDP-10 system routine which reproduces point for point a Calcomp plot of the CRT image. Because of lack of depth perception and suitability for scientific publication, a simple stick plot was insufficient. Hence, a simple routine to link the display image to the excellent ORTEP[20] plotting program was written. A comparison of a stick plot and an ORTEP plot for part of RNase-S is shown in Figure 2. A particular value of the ORTEP is the ability to produce a stereo plot. In addition to the plotting routines, one can take Polaroid pictures of the display screen. Also movies have been made of structures displayed and being rotated and tumbled on the screen for presentations at seminars or meetings.

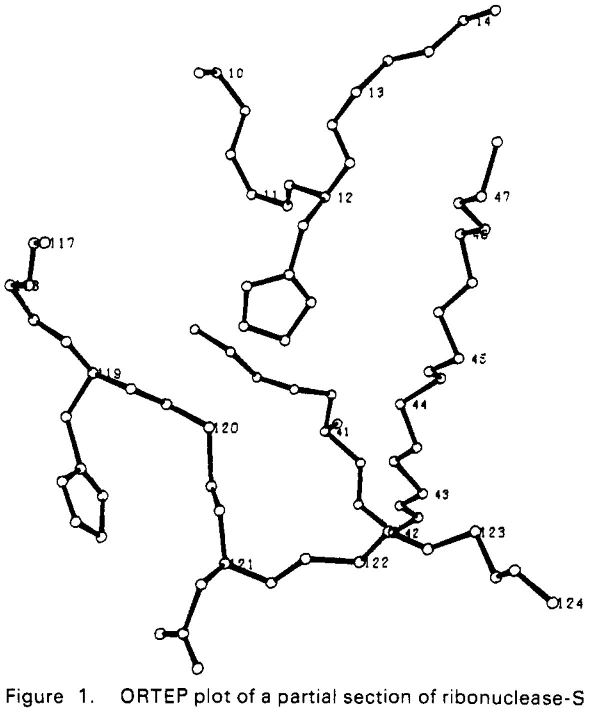

Figure 1. ORTEP plot of a partial section of ribonuclease-S

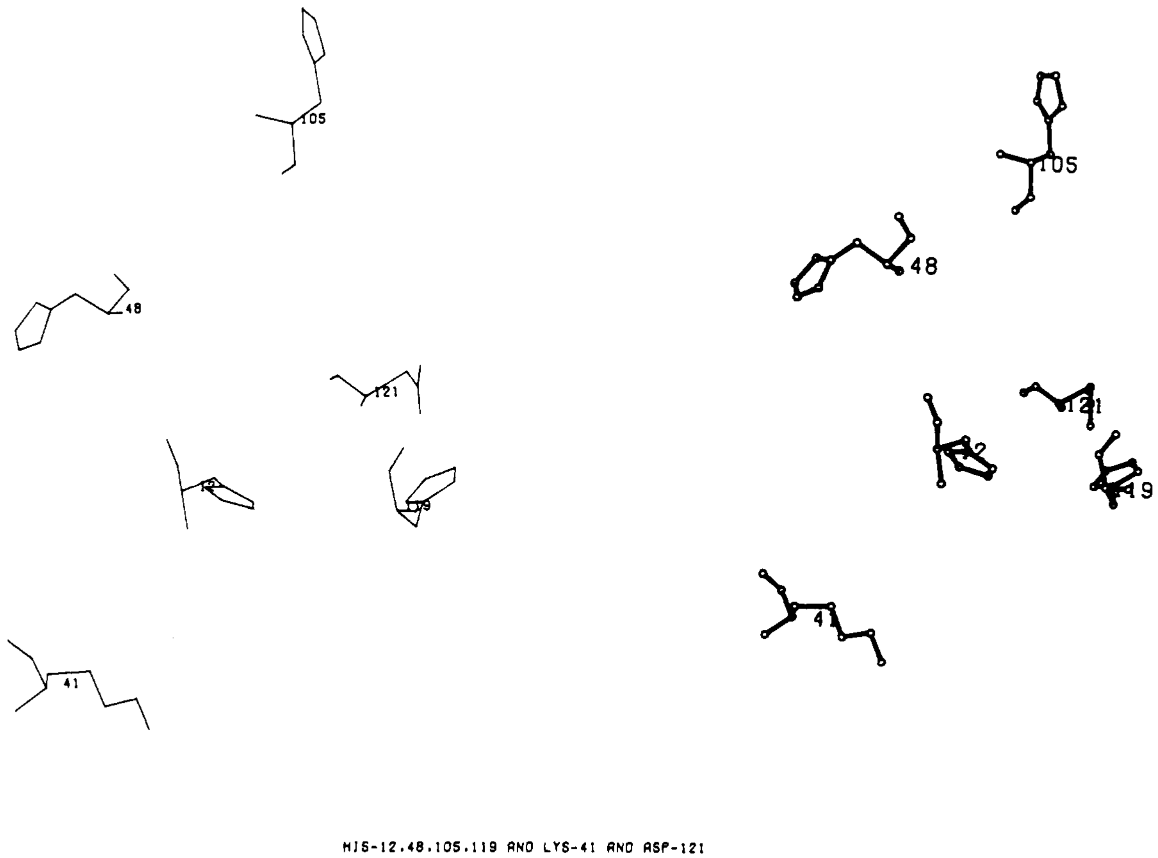

Figure 2. Comparison of ORTEP and standard plots for a section of ribonuclease-S

Energy State Calculations. For small molecules, scientists have found that the calculation of the energy of a given conformation of a molecule can lead to a reasonable estimate of electron densities, dipole moments, and energy states. The CNDO/INDO and MINDO energy state calculation programs have found wide use; thus, routines to generate input for the CNDO/INDO and MINDO programs from the display program were written. Both programs require x,y,z. Cartesian coordinates and a routine has been programmed to select a bond, perform a number of rotations, (e.g., 360 degrees by 30 degrees intervals) and generate an input data set for each rotation. Of course one can also specify just one particular angle and generate a CNDO/INDO or MINDO data set for that configuration. It is expected that as other similar programs are found to be of value, they will be interfaced with the system. (This application was materialized in collaboration with Robert Katz and Arthur E. Jacobson, NIAMDD, and applied by them to various biologically important molecules. These results will be described in a future publication.)

CONCLUSION

Physical models of molecules, in particular macromolecules, are extremely useful. The system described here for which an instruction manual is available,[21] allows the greatest flexibility in the display, manipulation, and plotting of three-dimensional structural data. The programs, which are used to maximum capability on the AGT/30 display, can be used with much less sophisticated and less expensive display terminals. With remote display terminals now available for about $10,000, it will be possible to do basic manipulations readily with the AGT/30 and make further minor modifications, plots, and energy calculations from such remote display terminals, connected to a central time-sharing PDP-10 computer over ordinary telephone lines.

ACKNOWLEDGMENT

We thank A. W. Pratt and W. A. Raub for their generous support and encouragement of this work.

LITERATURE CITED

(1) Feldmann, R. J., Heller, S. R., Shapiro, K. P., and Heller, R. S., "An Application of Interactive Computing—A Chemical Information System," *J. Chem. Doc.* **12,** 41 (1972).

(2) Levinthal, C., "Molecular Model Building by Computer," *Sci. Amer.* **214,** 42 (1966).

(3) Katz, L., and Levinthal, C., "Computer Graphics in Molecular Biology," in "Computers Graphics in Medical Research and Hospital Administration," R. D. Parslow and R. E. Green, Eds., pp. 56–70, Plenum Press, New York, 1971.

(4) Barry, C. D., Ellis, R. A., Graesser, S., and Marshall, G. R., in "Pertinent Concepts in Computer Graphics," M. Faiman and J. Nieuergelt, Eds., Univ. of Illinois Press, Urbana, Ill., 1969.

(5) Perkins, W. J., Piper, E. A., Tattan, F. G., and White, J. G., "Interactive Stereoscopic Computer Displays for Biomedical Research," *Comput. Biomed. Res.* **4,** 249 (1971).

(6) Meyer, E. F., "Towards an Automatic, Three Dimensional Display of Structural Data," *J. Chem. Doc.* **10,** 85 (1971).

(7) Meyer, E. F., "Three Dimensional Graphical Models of Molecules and a Time-Slicing Computer," *J. Appl. Cryst.* **3,** 392 (1970).

(8) Meyer, E. F., "Interactive Computer Display for the Three Dimensional Study of Macromolecular Structures," *Nature* **232,** 255 (1971).

(9) *Chem. Eng. News,* "TV Displays 3-D Structures," p. 25, May 17, 1971.

(10) Tometsko, A. M., "Computer Approaches to Protein Structure. I. Analysis of Atomic Distances," *Comput. Biomed. Res.* **3,** 229 (1970).

(11) Tometsko, A. M., "Computer Approaches to Protein Structure II. Model Building by Computer," *Ibid.,* **3,** 690 (1971).

(12) Tometsko, A. M., "Computer Approaches to Protein Structure. III. Transformation of Atomic Coordinates," *Ibid.,* **4,** 407 (1971).

(13) Portigal, L. D., and Minicozzi, W. P., "Computer Generated Display and Manipulation of a General Molecule," *J. Chem. Ed.* **43,** 790 (1971).

(14) Kennard, O., and Watson, D. G., "Molecular Structures and Dimensions," Vols. 1 and 2, Crystallographic Data Center, Cambridge, England, 1970.

(15) Quantrum Chemistry Program Exchange, Chemistry Department, University of Indiana, Bloomington, Ind., Program No. 178.

(16) Quantum Chemistry Program Exchange, Chemistry Department, University of Illinois, Urbana, Ill., Program No. 141.

(17) Quantum Chemistry Program Exchange, Chemistry Department, University of Illinois, Urbana, Ill., Program No. 137.

(18) Sproull, R. J., Stanford University "Topics in Computer Graphics," unpublished manuscript, p. 92-4, 1972.

(19) Kiefer, J., DCRT, NIH unpublished results, 1971.

(20) Johnson, C. K., "ORTEP," Oak Ridge Thermal Ellipsoid Program, 1964.

(21) Heller, S. R., and Feldmann, R. J., "DCRT/CIS X-Ray Modeling System Users' Manual," Division of Computer Research and Technology, Bethesda, Md., August 1972.

Substructure Search by Set Reduction*

JOHN FIGUERAS
Research Laboratories, Eastman Kodak Co., Rochester, N. Y. 14650
Received October 17, 1972

A PL/1 implementation of a substructure search system based on set reduction is described. The set reduction algorithm is based on set theory and Boolean algebra rather than the graph-theoretic approach described by Sussenguth [*J. Chem. Doc.* 5, 36 (1965)]. The use of ordered numerical codes for atom properties permits rapid list processing for fast rejection of non-matches, and the formulation of an efficient method for set generation. Time trials with a small file of organic chemical structures indicate that the algorithm can be economically used for substructure (or complete structure) sequential searches on a file containing 30,000-50,000 computer-coded structures.

The connection table is a widely used device for the compact storage of complete chemical structure information. It has the disadvantage, however, that structural relationships which are explicit in the original chemical structure are merely implied in the connection table, so that it is not possible to conduct straight-forward searches for fragments of structure. This paper describes an algorithm based on the use of sets which provides a route to structure matching when the structures are coded as connection tables. The algorithm is closely related to a graph-theoretic algorithm published by Sussenguth[1] in 1965, but is quite different in approach and organization and contains some novel features.

THE CONNECTION TABLE

A specially formulated connection table is used in order to expedite the substructure search program. The coding conventions employed follow closely the practice of *Chemical Abstracts*[2] with some modifications. The Nth row of the table refers to the atom with index N, and the entries in the row give atom indices and bond type for each atom connected to atom N. A typical table is shown in Table I for ethyl acetate. Entries in the body of the table (listed under "Connections and Bonds") are composite numbers, the last digit of each entry giving the nature of the bond joining two atoms; the remaining frontal digits give the index of the attached atom. For the connection table, the bonds are coded as follows:

1—single
2—double
3—triple
4—benzenoid
5—tautomer
6—charge delocalized

The definitions for "tautomer" and "charge delocalized" bonds can be found in *Chemical Abstracts* reports.[2] The use of composite numbers for joint representation of atom indices and associated bond types is a convenience that reduces the amount of information storage and the amount of subscripting required in the computer program. The composite numbers are easily resolved into their components by simple arithmetic. The table shows, for example, that atom 3 is connected to atom 4 by a double

*Presented at Division of Chemical Literature, 163rd Meeting, ACS, Boston, Mass., April 12, 1972.