Interactive program for investigation of protein structures based on ^{1}H NMR experiments

M Billeter*, M Engeli† and K Wüthrich*

*Institut für Molekularbiologie, ETH Zürich, Switzerland
†FIDES Treuhandgesellschaft, Zürich, Switzerland

An interactive molecular graphics program running on an Evans & Sutherland picture system is presented. It allows work on the conformations of a maximum of five molecules with a combined total of up to 1 000 atoms. Conformation changes are executed by real-time variations of dihedral angles about single bonds. Up to eight torsion angles can be selected anywhere in the molecule and changed simultaneously by turning dials. The program aims to establish geometric structures of molecules based on conformation-dependent ^{1}H NMR data; primarily upper limits on distances between protons. These may be located anywhere in the sequence of amino acids, which requires that the user is given access to the whole structure, i.e. all atoms and all dihedral angles, at any time. Violations of the upper limits in the actual conformations are displayed graphically to guide the user in search of conformations consistent with the ^{1}H NMR data. Applications discussed include various aspects of the previously published[1] determination of an approximate, topologically correct structure of the lac *repressor DNA-binding domain from* E. coli, *and interactive steps in the energy refinement of the solution conformation of a protease inhibitor from bull seminal plasma, which was obtained from distance geometry calculations with NMR data. Furthermore, the present graphics program proved to be a useful tool for detailed examinations and comparisons of protein structures obtained either with crystallographic methods or with the combined use of NMR and distance geometry.*

Keywords: proteins, molecular graphics, biopolymer conformation

received 4 March 1985, revised 1 April 1985

Recent work has provided evidence that spatial structures of biopolymers, which had hitherto been accessible only with scattering experiments in single crystals[2,3], could also be determined by distance geometry calculations using nuclear magnetic resonance (NMR) data from solution studies as input[4-8]. The present paper describes an interactive computer graphics program which can be used in certain situations as an alternative to distance geometry calculations for the structural interpretation of NMR experiments.

NMR data for spatial-structure determination include distance information from nuclear Overhauser enhancement (NOE) experiments, torsion angles about covalent single bonds and delineation of the solvent accessibility of labile hydrogen atoms. The most important parameters are the distance constraints between specified hydrogen atoms[6]. Using 2D spectroscopy, a network of distance constraints can be obtained which may extend over the entire macromolecular structure[9]. So far, use of the NOE experiments has been limited to obtaining upper bounds for the distances considered[4,5,8]. Therefore, the problem to be solved in the structural interpretation of NOE data is to find molecular conformations of the polymer chains which satisfy these constraints on the upper bounds of numerous intramolecular proton–proton distances and contain no violations of the minimum interatomic distances imposed by the van der Waals atomic volumes. The use of the presently described, interactive computer graphics approach to this problem is illustrated with practical applications to small proteins.

GENERAL OVERVIEW

The program presented here, Confor*, is a specialized tool for solving problems of the type described above using molecular graphics. Short distances accessible to ^{1}H NMR measurements connect specified pairs of atoms, which may be located anywhere in the sequence. These distances depend on all dihedral angles between the two atoms. Therefore, a determination of the 3D structure using such a list of upper limits on specified distances means that the computer program must have the entire structure available at any time. The program can display and store several molecules with a combined total of up to 1 000 atoms, and all dihedral angles defined in these molecules are accessible for conformation changes at any time. Violations

*The program is available on request. It has been written in FORTRAN-4-PLUS for the RSX-11M operating system of the PDP 11, but the current version uses the standard FORTRAN-77 compiler. It drives an Evans & Sutherland Multi Picture System.

of upper distance limits obtained by ^{1}H NMR or of lower limits imposed by the van der Waals volumes are visualized in the protein structure and guide the user in making conformation changes. The user may vary simultaneously any combination of up to eight dihedral angles, thus performing complex changes of the conformation in real time.

To simplify the picture on the screen, it is possible to draw only parts of the molecule by selecting either a segment of the protein sequence or a subset of the atoms, e.g. the backbone. It is also possible to use only a subset of the list of (up to 100) distance constraints according to different criteria. Finally, Confor allows a visual comparison of different structures obtained from ^{1}H NMR with each other or with protein structures from the available databanks (see for example Reference 10).

DESCRIPTION OF CONFOR

The Confor program fully uses the addressable memory capacity (32 kwords) of the PDP11/34. Different versions are distinguished by different configurations of principal arrays. The program can be customized to suit particular needs, for example one version of the program accepts 1 000 atoms, which can be distributed among up to five molecules, leaving memory space for 100 NOE definitions. Another version accepts five molecules with 57 residues each but no dihedral angles or NOE definitions. On the left side of the screen a list of commands always appears; there are four different command lists available (called menus). Colour plates 1–3 and Figure 1 show the four menus, together with molecular information as they might appear during the work with Confor.

The hardware consists of a PDP11 host computer together with an Evans & Sutherland picture system including a colour screen, a tablet, two sets of eight dials for analogue input, and a box with 16 switches. All menu commands are initiated by use of the tablet. The top four commands are common to all menus and allow switching between them.

Menu 1 (Colour plate 1) serves input and output purposes. First, the covalent geometry of all amino acid residues is read in. Then protein structures can be input as coordinate lists or as a list containing the sequence and the dihedral angles. Output files are written in the same format. In addition, plots can be directly initiated. The 'CONFOR' commands read or write binary files containing all parameters of the program and allow interruption and restart of the program in exactly the same state. Further commands recompute structural parameters in the case of rounding errors that occur mainly in the picture system transformation matrices, choose the display speed (important for the colour screen) and terminate the program. Colour plate 1 shows one mode of presentation by Confor, i.e. a schematic drawing of the polypeptide backbone in a protein by a smooth curve through the C^α positions.

Menu 2 (Colour plate 2) allows 'synthetising' of molecules and defining their visible parts for use in menu 3. The first command called 'DRAW' has various functions: it can be used to move molecules from the background to the foreground or to prevent them from being drawn in menu 3. 'Foreground' implies full access to all parameters of the molecule, 'background'

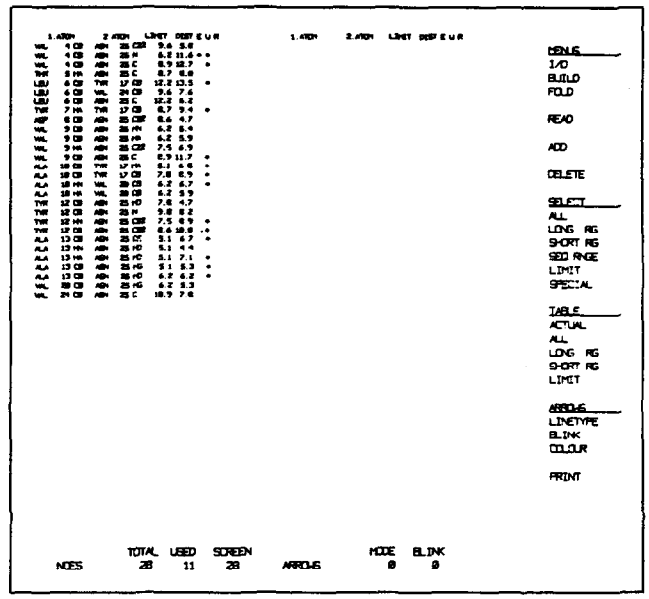

Figure 1. Picture system screen when Confor is in menu 4 showing command list of menu 4 and list of NOEs used. Each line gives: two atom identifications, an upper distance limit, the distance between the two atoms in the actual structure, an asterisk in 'E' column when the violation is larger than 5 Å, an asterisk in 'U' column when the violation is used for the computation of the arrows drawn in menu 3, and the number of residues, by which the two atoms are separated with respect to the primary structure (only numbers less than five would be shown).

allows no access but the molecule is still visible. DRAW also allows the definition of visible segments of the chain and colouring of entire molecules, chain segments or residues according to the colours attributed to each residue type in the corresponding list which appears on the lower part of the screen.

A final use of the DRAW command combines several molecules into one group; these molecules are then considered as a single unit when rotations and translations are applied. For example, consider two strands of a DNA double helix plus a DNA-binding molecule; the two strands can form a molecule group so that the DNA double helix remains intact during all operations used to change the relative positions and orientations of the DNA and the DNA-binding molecule.

Other commands under menu 2 allow construction of new molecules by addition or subtraction of complete residues. Also, individual molecules can be deleted from a group of structures or the order in which molecules are stored in the computer memory can be changed (sometimes necessary for the 'CONNECT' command). 'BEST FIT' moves one molecule on top of another by translations and rotations, so that the following sum becomes minimal[11]:

$$\mathrm{RMSD} = \left[1/n \sum_{i=1}^{n} d(C_i^{\alpha'}, C_i^{\alpha})^2 \right]^{1/2}$$

where: n is the number of residues in each of the two molecules; and d is a distance function. The atoms $HC^{\alpha'}$

belong to the molecule being moved and atoms C^{α} are part of the second molecule.

Colour plate 2 illustrates how different segments of a polypeptide chain can be treated separately in menu 2. The chain is divided into two pieces, 'molecule 1' and 'molecule 2'. Helical residues are green, yellow or red, nonhelical residues are blue. The N-terminal blue part in the first piece and the whole second piece are not underlined, which indicates that the user does not want to see these parts when switching to menu 3. On the lower half of the screen is a list of all amino acid residues that can be used in the structures. Below this list, the complete spectrum of colours available with the picture system is given. It is used for the selection of the colours in the presentations of the structures.

Menu 3 (Colour plate 3) allows work on the conformations of the molecules and is thus effectively the heart of the program. The first few commands display values of structural parameters: coordinates of atoms, dihedral angles and atom–atom distances. The next commands are for conformational changes using the dials, allowing the definition of up to eight dihedral angles, selected anywhere in the visible part of the molecule, and their association to a dial. By turning a dial, the corresponding dihedral angle is changed and the structure on the screen updated in real time. Two dihedral angles can be connected to the same dial. Turning this dial then results in the change of both dihedral angles by the same amount, although the changes can have either the same or opposite signs. This feature is particularly useful for work with proteins, when a ψ-angle and the ϕ-angle of the following residue are connected to one dial so that concerted changes occur with opposite signs. This allows the local conformation to be altered while keeping the global conformation intact.

The 'UPDATE' and 'ACCEPT' commands update the coordinates in memory. 'UPDATE' automatically gives back the old choice of dihedral angles attached to the dials. By setting appropriate switches, the dials can also be programmed to perform rotations, translations and zooming of entire molecules or molecule groups. Other switches select the drawing mode: all atom presentation, backbone only, C^{α}-atoms only, smooth curves through the C^{α} positions (Colour plate 1), individually selected subsets of atoms (Colour plate 3) or ribbon-like representation of the backbone. All of these can be shown in stereo pairs. Two further commands in this menu allow displaying of lines and text of various types and colours anywhere on the screen which may be useful when preparing slides directly from the screen. Menu 3 also includes a display of the distance information obtained from ^{1}H NMR or other sources (see also menu 4). It consists of arrows attached to individual atoms pointing in the direction in which the atoms should be moved to satisfy all presently used (see below) violated constraints involving this atom. These arrows result from the following vector sum:

$$a_i = \sum_j \frac{x_j - x_i}{|x_j - x_i|} * (d_{ij} - u_{ij})$$

where: a_i is the arrow attached to atom i; x_i and x_j are atom coordinates; d_{ij} is the distance in the actual structure between atoms i and j; and u_{ij} is the upper limit measures over this distance; j runs over all atoms connected by a NOE to atom i.

Colour plate 3 shows the screen during operation of Confer in menu 3. In a peptide segment of two helices (green and yellow) and three intervening residues (blue), red lines indicate violations of experimental NOE distance constraints. They are attached to the atoms involved in such a violation and point in the direction in which the atoms should be moved to meet the NOE-constraints.

In addition to the display of violations of upper limits on distances, an individually selected number of critical atom pairs are checked for possible van der Waals contacts. The distances within these atom pairs can be displayed together with the identification of the atoms in the upper left corner of the screen. In case of a collision between two atoms, the centres of these atoms are connected by a line of user-defined type and colour. The commands 'UPDATE' and 'ACCEPT' will also update the arrows and the information on lower distance bounds.

Menu 4 (Figure 1) allows the manipulation of a list of NOEs. Up to 100 NOE-constraints, i.e. upper limits for individual atom–atom distances can be read in and modified. This list can then be varied by selecting a subset according to one of the following criteria: 1. all NOEs; 2. only long-range NOEs, i.e. NOEs connecting atoms that are more than five residues apart; 3. only short-range NOEs; 4. NOEs connecting atoms that lie in a certain segment of the chain; 5. NOEs that are violated by more than a certain value; 6. each NOE can be individually included or eliminated from a subset. Thus only an individually selected subset of NOEs is used for the computation of the arrows in menu 3. The screen in menu 4 shows the NOE list (or a subset thereof) including for each element the two atom identifications and the limit defining the NOE, the corresponding distance in the actual structure and information about the violation, the use in menu 3 and the number of residues, by which the two atoms are separated with respect to the primary structure (Figure 1).

APPLICATIONS

Topology of the three helices in *E. coli lac* repressor DNA-binding domain

In a joint project with Dr R Kaptein and his group at the University of Groningen, the overall folding of the polypeptide chain in *E. coli lac* repressor DNA-binding domain 1-51 was determined. There is no single-crystal structure available for this protein. From the sequential resonance assignments and additional NMR experiments[12,13] it is known that there are three helices at the approximate residue positions 6-13, 17-25 and 34-44. One of the tools used for the determination of the relative spatial locations of these three helices was CONFOR. The results have been published elsewhere[1]; here some details of the application of CONFOR to this particular problem are described.

Compared with work on other proteins of comparable size (e.g. Reference 8) only a relatively small number of NMR constraints had been collected for *lac* repressor 1-51[1]. For the structural interpretation the number of variables was therefore reduced by the following assumptions. The three helices were kept in the

standard α-helix conformation and all NOEs involving side-chain protons were referred to the corresponding β-carbon atom positions, using appropriate corrections for the distance constraints (see below). With these assumptions the structure analysis was reduced to a determination of the spatial topology of the three helices, and did not extend to the conformations of the non-helical backbone segments and the amino acid side chains.

The input used to solve this problem with CONFOR consisted of two files: the first file contained the amino acid sequence and a set of values for the dihedral angles describing the starting conformation. All initial dihedral angles were set to 180°, except for the ϕ and ψ angles in the three helical segments which were set to the standard values for α-helices, i.e. $-57°$ and $-47°$, respectively. The second input file contained a list of 28 NOE upper-distance constraints of 4.0 Å between specified pairs of protons. All H constraints involving side-chain protons were referred to the C^β atom of the side chain, and were consequently corrected by addition of the maximum sterically allowed distance from the observed proton to the C^β position. Figure 1 shows this second input file as it appeared in menu 4. Since no NOEs were observed for residues 1–3 and 48–51 only the residues 4–47 were used for the structure determination.

In a first step only the segment 6–25, which contains the first two helices, was considered (Colour plates 2 and 3). In this segment there were seven NOEs to be satisfied (Figure 1) and the only conformational variables were the six backbone dihedral angles ϕ and ψ of the three residues between the two helices. These six dihedral angles were attached to six dials and changed simultaneously. Frequent use of the 'UPDATE' command helped to keep track of the improvements to the structure with respect to the seven NOEs. Colour plate 3 was a display encountered during this phase of the project. It shows that the distance constraints between the following pairs of atoms were still violated: LEU 6 C^β and TYR 17 C^β, TYR 7 H^α and TYR 17 C^β, ALA 10 C^β and TYR 17 H^α, and ALA 10 C^β and TYR 17 C^β. These violations are indicated by red lines in Colour plate 3.

Once a satisfactory conformation was obtained for the first two helices, the third helix and residues 26–33 were also included. The relative orientation of the third helix with respect to the first two was then adjusted by variation of the torsion angles ϕ and ψ for residues 26–33. In a third step the helix topology obtained was checked for compatibility with the NOEs involving residues 4, 5 and 45–47. Colour plate 1 shows the resulting helix topology, where the polypeptide chain is represented by a smooth curve through the C^α atoms. In this presentation the first helix (residues 6–13) runs from back to front, forming an angle of $\approx 50°$ with the projection plane; the other two helices (residues 17–25 and 34–44) are oriented parallel to the projection plane, with the C-terminal helix in front and the middle helix in the back. This topology[1] showed no residual violations of NOE distance constraints.

Van der Waals contacts for the backbone atoms, including the backbone protons, and for the C^β atoms were considered in all steps of this topology determination. For this purpose a small number of atom pairs was selected for which the distance was in a critical range. Such a pair would for example consist of an atom on each of two helices and be located in a contact region between the two helices. As described above CONFOR automatically keeps track of these distances and displays violations of van der Waals volumes directly in the structures.

Use of Confor in conjunction with structure determinations by distance geometry and energy minimization

The study of the *lac* repressor DNA-binding domain illustrates that Confor can be employed independently of other procedures for the structural interpretation of NMR data representing constraints on the conformation of polypeptide chains in solution. In practice, however, Confor has so far been used more extensively for the presentation and inspection of structures computed with distance geometry calculations, for interactive steps in the procedures used for refinement of these structures by energy minimization and generally as a complementary tool to optimization programs for the elucidation of polypeptide structures from experimental distances and torsion angles. These uses are illustrated here with work on a protease inhibitor from bull seminal plasma, BUSI IIA[8].

Colour plate 4 shows five conformations of BUSI IIA which were obtained from the same experimental NMR data in five subsequent distance geometry calculations. The command 'BEST FIT' was used to superimpose the four other conformers for the best fit with the green structure (the average RMSD value between any two structures is ≈ 2.0 Å). In Colour plate 5, one of the BUSI IIA structures in Colour plate 4, obtained from ^{1}H NMR measurements in solution, is compared with two homologous proteins from the protein databank, porcine secretory trypsin inhibitor and the third domain of Japanese quail ovomucoid. For the superpositions with the 'BEST FIT' command only residues 23–57 from BUSI IIA and residues 22–56 from the two homologues were used, since the similarity of the structures does not extend over the N-terminal part of the sequence.

Out of the 5 BUSI IIA conformations obtained with distance geometry the one with the smallest violations left was selected for a pilot study on restrained energy minimization to improve local conformation. The full refinement procedure is described in the appendix of Reference 8. Here only those steps are mentioned where Confor proved to be a useful tool in alleviating remaining NOE violations or avoiding energetically unfavourable conformations. In one of the first steps Confor was used to adjust long side chains on the exterior of the molecule thus eliminating eclipsed conformations and isolated contacts with the adjacent backbone. Since NOE-constraints are monitored simultaneously, changes could be made without introducing new violations of NOEs. In a later step, i.e. after further automatic cycles of optimization, CONFOR was again helpful in relieving remaining large violations, e.g. by rotating a disulphide bond over its 10 kcal torsional barrier.

For the evaluation of structural data obtained from NMR and distance geometry it is generally essential to

have a wide range of different representations for the proteins available. CONFOR performs real time rotations, which can be executed simultaneously with conformational changes, produces stereo pictures and allows various uses of different colours, which all helps in an understanding of complex situations. Switching between different representations, for example 'all atoms' or 'backbone only' in <1 s helps to investigate local problems while keeping track of the global situation. Confor can also generate structures from sets of dihedral angles, and the coordinates thus obtained can then be used, for example, as starting conformations in refinements programs. For all these potential uses it is important that Confor accepts different input formats, which makes communication with other programs quick and easy.

DISCUSSION

Confor can be used to advantage as a complement to automatic programs for distance geometry calculations[7] and/or simple optimizing algorithms. Nevertheless it has been set up to allow all steps of structure determination independently of these other techniques. Based on the experience obtained with the problems discussed above, as well as with other applications, the following suggestions can be made for profitable, practical uses of Confor:

- Collect and enter as much of the readily obtainable structural information as possible into the starting structure. When working with ^{1}H NMR, regular secondary structures like helices and β-sheets can usually be identified from inspection of the spectra without needing any computations[14,15]. Such structural input may be a useful approximation of the real structure. For example, regular helices or β-strands defined by the standard dihedral angles (e.g. Reference 3) are used as elements in the starting structure.
- Split the problem into small sub-problems and try to proceed step by step. For example, take individual β -strands as building blocks to form β-sheets or, as in the *lac* repressor problem, start by first fitting two helices and then add further structure elements. In this way even large molecules with 50 or more residues and with 100 or more NOEs may be handled efficiently, since the sub-problems are less complex. One of the main advantages of an interactive program compared with an automatic program is the fact that the user can more readily adapt the choice of subproblems and of the order in which they are treated to a specified protein and a particular data input.
- An interesting and potentially useful application of Confor is for obtaining approximate but topologically correct structures, which may then be handed over to an automatic optimizing program. If necessary the resulting structure can again be inspected and locally improved with Confor, as described for the BUSI IIA problem.
- The interactive structure determination enables the use of additional conformational information, such as data on the structures of homologous proteins or ^{1}H NMR data other than distance measurements (spin–spin couplings, chemical shifts or amide proton exchange rates)[16].

ACKNOWLEDGEMENTS

The authors thank Drs R Kaptein and ERP Zuiderweg for collaboration on the DNA-binding domain of *E. coli lac* repressor, and Drs T Havel and M P Williamson for the data on BUSI IIA. Financial support was obtained from the Schweizerischer Nationalfonds (project 3.284.82). Use of the facilities at the Zentrum für Interaktives Rechnen at the ETH (ZIR) is gratefully acknowledged.

REFERENCES

1 **Zuiderweg, E R P, Billeter, M, Boelens, R., Scheek, R M, Wüthrich, K and Kaptein, R** *FEBS Lett.* vol 174 (1984) pp 243–247
2 **Blundell, T L and Johnson, L N** *Protein Crystallography* Academic Press, USA (1976)
3 **Richardson, J** *Adv. Prot. Chem.* vol 34 (1981) pp 167–335
4 **Braun, W, Bösch, C, Brown, L R, Go, N and Wüthrich, K** *Biochim. Biophys. Acta* vol 667 (1981) pp 377–396
5 **Braun, W, Wider, G, Lee, K H and Wüthrich, K** *J. Mol. Biol.* Vol. 169 (1983) pp 921–948
6 **Wüthrich, K, Wider, G, Wagner, G and Braun, W** *J. Mol. Biol.* Vol 155 (1982) pp 311–319
7 **Havel, T and Wüthrich, K** *Bull. Math. Biol.* Vol 46 (1984) pp 673–698
8 **Williamson, M, Havel, T and Wüthrich, K** *J. Mol. Biol.* Vol 182 (1985) pp 295–315
9 **Anil Kumar, Ernst, R R and Wüthrich, K** *Biochem. Biophys. Res. Comm.* Vol 95 (1980) pp 1–6
10 **Bernstein, F C, Koetzle, T F, Williams, G J B, Meyer, E F Jr, Brice, M D, Rodgers, J R, Kennard, O, Shimanouchi, T and Tasumi, M** *J. Mol. Biol.* Vol 112 (1977) pp 535–542
11 **McLachlan, A D** *J. Mol. Biol.* Vol 128 (1979) pp 49–79
12 **Zuiderweg, E R P, Kaptein, R and Wüthrich, K** *Eur. J. Biochem.* Vol 137 (1983) pp 279–292
13 **Zuiderweg, E R P, Kaptein, R and Wüthrich, K** *Proc. Natl. Acad. Sci. USA* Vol 80 (1983) pp 5837–5841
14 **Wüthrich, K, Billeter, M and Braun, W** *J. Mol. Biol.* Vol 180 (1985) pp 715–740
15 **Pardi, A, Billeter, M and Wüthrich, K** *J. Mol. Biol.* Vol 180 (1985) pp 741–751
16 **Wüthrich, K** *NMR in Biological Research: Peptides and Proteins* North-Holland, Netherlands (1976)