NEW PROGRAMS

Displaying inter-main chain hydrogen bond patterns in proteins

Khaled Belhadj-Mostefa,* Ron Poet* and E. James Milner-White

Departments of Computing Science and Biochemistry, University of Glasgow, Glasgow, UK*

Two computer graphics techniques for displaying hydrogen bonds between the main chains of different proteins are described, and illustrated for two thiol proteases. (The X-ray crystallography was performed by Kamphuis et al. in 1984,[1] and by Baker and Dodson in 1980.[2]) One is a three-dimensional model that can be manipulated in space; the hydrogen bonds are represented with the smoothed α-carbon plot of the polypeptide chain. In the other type of display, hydrogen bonds are viewed in relation to the one-dimensional sequence. Both types of picture facilitate visualization of hydrogen bond patterns, such that loop motifs, as well as α-helices and β-sheets, can be examined easily. We suggest that such displays are useful as a general means of displaying whole proteins and whole domains because they reveal more information than do conventional simplified pictures of proteins, which focus exclusively on α-helices and β-sheets. These techniques can be implemented on a UNIX-based computer graphics workstation. (UNIX is a trademark of Bell Telephone laboratories.)

Keywords: hydrogen bonds, proteins, α-helix, β-sheet, loop motifs, papain, actinidin

Address reprint requests to Dr. Milner-White at the Department of Biochemistry, The University of Glasgow, Glasgow G12 8QQ, UK.
Received 11 September 1990; accepted 6 November 1990

INTRODUCTION

When pictures of whole proteins are displayed in the chemical and biological literature, they are almost invariably represented as chains with the α-helices and strands of β-sheets singled out in some way.[3–7] Programs are available to do this. It is a convenient representation, but it does make the assumption that these particular features are the most important ones. Hydrogen bonds between main chain atoms are well known to be the basis for α-helices and β-sheets. However, other inter-main chain hydrogen bonds occur in proteins, many of which appear to stabilize loop motifs.[8,9] Although loop motifs are nonrepetitive (unlike α-helices and β-sheets), we argue that they are still relatively well conserved in evolution, somewhat less than secondary structure but more than the sequence. We therefore suggest that it is useful to provide representations of whole proteins that display all inter-main chain hydrogen bonds and not just those in secondary structure elements. Two display techniques for relating such hydrogen bonds to primary structure are presented: three-dimensional (3D) plots and one-dimensional (1D) plots (one-dimensional because bonds are displayed relative to the primary structure). The 3D plots can be rotated in three-dimensional space and provide a quasi-realistic view of a protein, while the 1D plots are a computer-generated version, in color and with several modifications, of the black and white diagrams that are often drawn by protein crystallographers to show how the inter-main chain hydrogen bonding relates to the sequence.

In the present work we compare the inter-main chain hydrogen bonds of actinidin from the chinese gooseberry and papain from the papaya fruit. They are members of the thiol protease family that exhibit 50% sequence identity and whose three-dimensional structure is known at high resolution.[1,2] They possess catalytically important hydrogen-bonded thiol–imidazole pairs (CYS 25–HIS 159 in papain, CYS 25–HIS 162 in actinidin). Each has two domains, an N-terminal one that is mainly β-sheet and a C-terminal one that is mostly α-helix. The polypeptide binding site, including the thiol–imidazole pair, lies between the two domains. The Figures show that the hydrogen bond patterns are more highly conserved than the sequences.

DISPLAY METHODS

For a given protein, data from the Brookhaven Protein Data Bank are processed and a list of the inter-main chain hydrogen bonds in the protein is stored. The bonds are calculated by a procedure similar to the one used by Baker and Hubbard.[10] We use the procedure that they recommended, rather than the one that they actually used. (Briefly, the angle between the N,H and O atoms has to be more than 120°; the angle between the H,O and C atoms has to be more than 90°; and the distance between H and O atoms has to be less than 0.25 nm.)

3D plots

3D display techniques we used are:

- Depth sorting: Alpha-carbon atoms and hydrogen bonds (midpoints) are

drawn in depth order with the most distant element being drawn first.

- Intensity depth cueing: Atoms near the viewer appear brighter.
- Brush shape: A circular "brush" is used to draw the thicker lines; this means that the line begins and ends in a semicircular shape, which is specially useful for representing bonds between atoms.
- Tramlines: The relative depths of two line segments of different atoms are indicated by a bridge effect when they cross. This is done by drawing two parallel lines in the background color, one on either side of all line segments of the α-carbon atoms. We use the term "tramlines" to refer to this effect. It provides an indication of depth when line segments have the same, or nearly the same, color intensity. A new technique that uses the circular brush to draw tramlines is presented below and in Figure 1. It is faster than the previous method for drawing tramlines, which involved polygon filling.[11]

In the 3D-plots the main chain is drawn as a smoothed α-carbon plot. Smoothing is carried out[8] by taking the average of the coordinates of five consecutive α-carbon atoms. The lines joining α-carbons are drawn as described below, with tramlines, but the hydrogen bonds are drawn simply as thick or thin lines, without tramlines. The smoothed α-carbon atoms and the midpoints of the hydrogen bonds are sorted according to their depth, and the appropriate shapes are plotted in order (those at the back first). The depth value for the midpoint of each hydrogen bond is decreased by 0.4 nm; this offset ensures that bonds in an α-helix do not obliterate the main chain of that helix.

Bonds joining α-carbon atoms are drawn as line segments that run from the atom in question to a point halfway between the two atoms being joined. Hence each bond consists of two segments, one contributed by each atom being joined. Figure 1 shows how consecutive α-carbon atoms are drawn. An α-carbon segment refers to a line segment going from one α-carbon halfway towards the next. Of the atoms shown, *A* and *D* are closest to the viewer, *C* is next, and *B* is the furthest away. In Figure 1a a circle is first drawn in the background color centered at *B*, with a radius equal to that of the one used to plot the α-carbon segments. The brush used to draw this circle has a diameter of 3 pixels. The tramlines are then drawn in the background color (with a brush of the same thickness) from points on the circle to points halfway between *BA* and *BC*. The next step (Figure 1b) is to draw the α-carbon segments (*B* towards *A* and *B* towards *C*) according to the rule stated earlier. The shading used depends on the depth of *B*. The same technique is performed for atom *C* (Figure 1c and 1d), which has two connecting bonds (*C* towards *B* and *C* towards *D*). The shading is lighter as *C* is in front of *B*. The process is repeated for every α-carbon atom, in depth order, until the entire polypeptide chain is drawn.

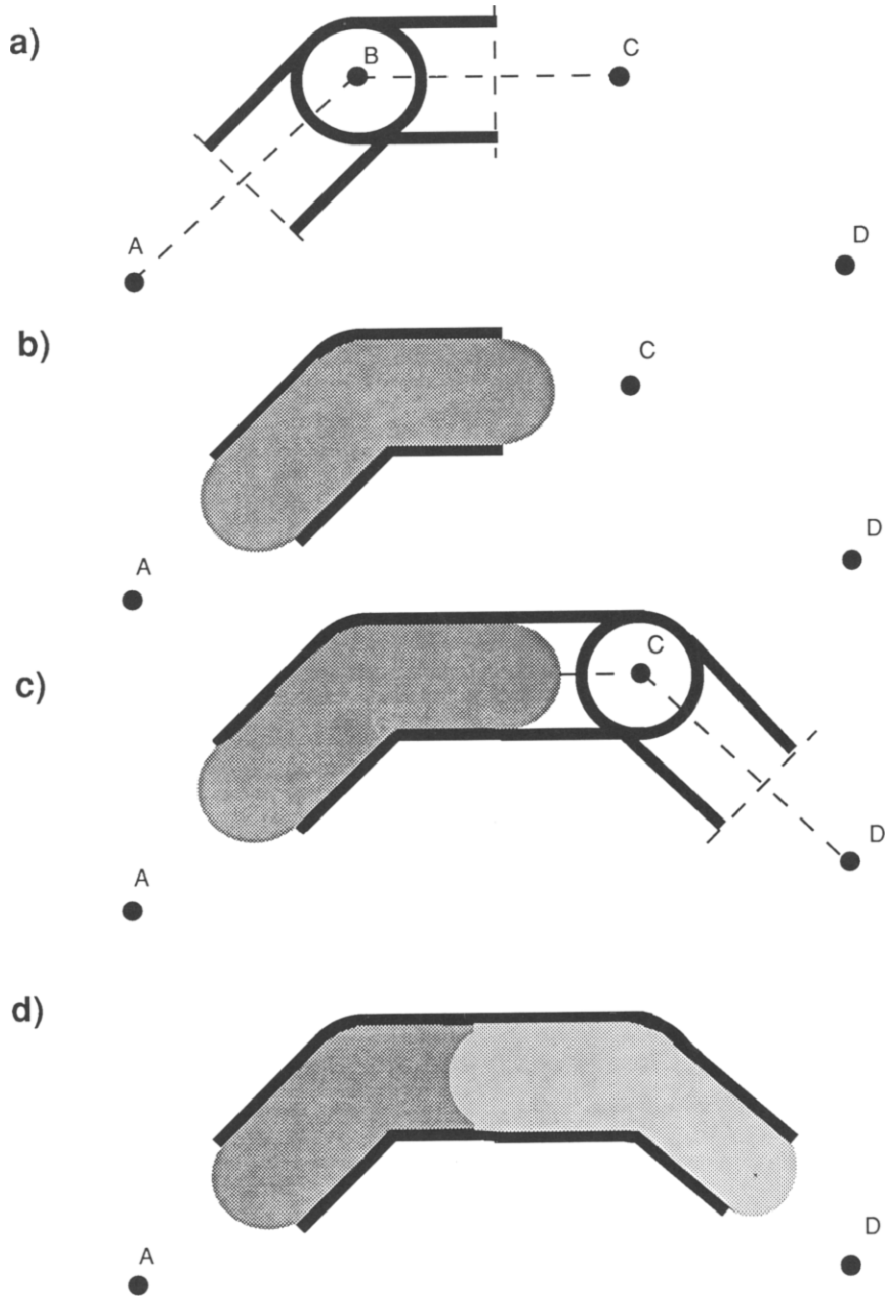

Figure 1. How consecutive α-carbon atoms are drawn. There are four consecutive α-carbon atoms: A, B, C *and* D. *The tramlines are drawn first, in black, followed by the virtual bond line segments joining α-carbons, which are shown shaded. In the Color Plates the background is black rather than white as in this Figure, so the tramlines are often invisible.*

1D plots

In 1D plots the sequence is represented as a horizontal row of boxes. For short proteins, the single-letter code for amino acids can be included within the boxes. Hydrogen bonds are drawn as pipe-like features joining the boxes. Single hydrogen bonds are drawn with a thin line at one side of the box to represent either NH or CO group involvement. Where there is a pair of hydrogen bonds between both sets of CO and NH groups, a thick line is drawn from the middle of each appropriate box. Such thick pipe-like lines are an indication of hydrogen bonds involved in antiparallel β-sheets. To ensure that the hydrogen bonds can be clearly differentiated in all pictures, they are drawn with tramlines as described earlier for the line segments representing virtual bonds between α-carbons in the 3D plots.

A further means of identifying hydrogen bonds involved in parallel β-sheets is included, a rectangular box drawn at the center of the set of two or more bonds joining two parallel strands of β-sheets. The length and height of the rectangle are variable and depend on, respectively, the distance separating a pair of strands and the number of bonds involved. The more amino acids that separate the strands the longer the rectangle, and the more bonds there are the fatter it is.

Depth sorting is applied to all of the hydrogen bond lines. The depth order depends primarily on the 'sequence difference' (the number of amino acids separating the hydrogen-bonded amino acids), with the longest drawn first. Note that if depth sorting were the other way around, some short vertical pipe-like features would be obliterated by fatter ones. Bonds with the same sequence difference are ordered so that those with the highest residue numbers are drawn first. Horizontal overlapping is avoided by allocating a level variable to every hydrogen bond that records the horizontal level in which the corresponding line is to be drawn. To do this, a linear integer vector list *chainette* is created. The number of entries is twice the total number of residues constituting the protein because it is necessary to differentiate between the lefthand (NH group) and righthand (CO group) side of a box. Sometimes the CO and NH groups of the same amino acid are both hydrogen-bonded. The bonds may be drawn at the same horizontal level, as there is no expectation of overlap. The algorithm for allocating levels to bonds is described below.

Initially all elements of the vector list are set to zero. The hydrogen bonds are examined in the same order as for the depth except that bonds with the same 'sequence difference' are ordered with the lowest residue numbers first. For every line read, the residue numbers of the bonded NH group and CO group are noted. For an NH connection at residue i the corresponding element is indexed as *chainette*$[(2*i)-1]$; a CO connection at residue j is indexed as *chainette*$[2*j]$. All contiguous elements of *chainette* between and including those corresponding to the NH and CO groups of the bond are checked. The highest of those existing vector elements, incremented by one, gives the new bond level, and all vector elements bracketed by the bond in *chainette* are altered to the same level value before the next bond is processed.

An example of the 1D plot algorithm

Suppose that the number of residues constituting the protein is 8 and that the input hydrogen bond file contains the following 3 bonds: 1NH → 3CO, 4CO → 7NH and 5CO → 8NH. Initially, *chainette* contains 16 elements (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0). When the first bond is processed (1NH → 3CO), all elements of *chainette* between *chainette* (1) and *chainette* (6) are checked. Because all of the elements are zero, the bond is assigned level 1 and *chainette* is updated to (1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0). To find the level for the next bond all of the elements between *chainette* (8) and *chainette* (13) are checked. Because these are all zero the bond is assigned level 1 and *chainette* is updated accordingly to (1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0). For the last bond elements between *chainette* (10) and *chainette* (15) are checked. As the highest content of any of these elements is 1, the bond is assigned level 2 and *chainette* is updated to (1, 1, 1, 1, 1, 1, 0, 1, 1, 2, 2, 2, 2, 2, 2, 0).

Hardware and software

The software was written in the C programming language for a Whitechapel CG-1 color graphics workstation. The Color Plates are directly photographed from this workstation. The display device generates a range of 64 colors from a pallette of 256. The CG1 runs on GENIX (a variant of the UNIX operating system) that provides a multi-window UNIX environment. Whitechapel hardware is no longer available but we have recently acquired the use of a SUN-4 and are adapting the software to this workstation.

DISCUSSION

A 3D plot for actinidin is presented in Color Plate 1. The α-carbon plot is smoothed by averaging 5 successive α-carbon atoms; this has the effect of making α-helices look like wavy lines when viewed from the side. Inter-main chain hydrogen bonds are drawn as colored lines joining appropriate smoothed α-carbon atoms. Where both the NH and CO groups of a pair of amino acids are bonded to each other, a single thick line is drawn. In actinidin, the lower domain has a number of red bonds because it is mainly α-helical, and the top domain, being mainly composed of β-sheets, has a majority of yellow bonds. (See color key.)

Color Plates 2–4 are 1D plots. The main chain is shown as a horizontal row of boxes and hydrogen bonds are drawn in a pipe-like representation, colored as in Color Plate 1. Hydrogen bonds are drawn at either side of each box to represent NH or CO groups, except for the pairs of bonds mentioned below, which are drawn in the middle. Where NH and CO groups of a pair of amino acids are bonded to each other, a single thick line is drawn, again as in the 3D plots. Such pairs of bonds are characteristic of antiparallel β-sheets. Parallel β-sheets are displayed by means of a hollow box that groups the relevant bonds; one is seen at the bottom of Color Plate 4.

Color Plate 2 provides a comparison between the inter-main chain hydrogen bond patterns of these thiol proteases. The similarity is remarkable, considering that they are only 50% identical in sequence. The five conserved α-helices can be recognized by the sets of red bonds. A few purple $(i, i+3)$ bonds appear to be associated with these helices, which is typical of α-helices in general. There are some 3_{10}-helices, distinguished by the exclusively purple

bonds: three in actinidin and two in papain. They are shorter and not as well conserved as the red α-helices. One α-helix encompassing residue 100 in actinidin is a 3_{10}-helix in the corresponding region of papain. Beta-sheets are drawn as regularly arranged yellow bonds; sets of these bonds can be seen that link adjacent strands. There are nine sets of antiparallel bonds plus one set of parallel bonds that are all conserved. One set of yellow parallel bonds in actinidin is not fully conserved in papain.

Loop motifs are seen especially well in these pictures. The purple and orange bonds at residue 10 in both proteins are typical paperclip loops.[8,12] The thick turquoise bonds at the right of the picture are both class-2 β-hairpins (using the hairpin terminology of Milner-White and Poet[8]; they could also be called 2:2-residue hairpins, using the nomenclature of Blundell et al.[13]). There is another thick turquoise class-2 hairpin at residue 60 in actinidin that is absent in papain; this is mentioned below. To the right of this, also seen in Color Plate 3, are the characteristic green and purple colored bonds of the shorter of the two kinds of β-bulge loop[8]; they are conserved, though not exactly, in the two proteins. Each protein also has a short white bond, seen best in Color Plate 4. These are inverse γ-turns,[14] which can be thought of as being like β-turns, but shorter. In these proteins they occur in nonhomologous positions.

Color Plates 3 and 4 show that information relating to the situations of insertions and deletions (referred to as "indels"[15]) can be gained, beyond that predicted merely by aligning sequences. The thick turquoise hairpin mentioned above in actinidin corresponds to the position of an indel of 2 residues at residue 59 or 60, such that the hairpin is absent in papain. There is also an indel of 1 residue that is seen, also in Color Plate 3, to be at the C-terminal end of the longer α-helix. Hence there is an orange bond there in actinidin, but not in papain. Another indel occurs in the middle hairpin of Color Plate 4. Examination shows that there appears to be an indel of 4 residues, 2 at opposite positions in both strands, giving rise to a further pair of hydrogen bonds within the hairpin. This is just the sort of change to be expected in this situation, because the insertion of 4 such residues causes minimal disruption to a β-hairpin[8].

CONCLUSION

This paper is concerned with the way whole proteins are viewed. Such displays have to be simplified to be appreciated. Typically, the main chain is drawn as a series of helices and strands of β-sheets. We argue that this method selects certain hydrogen-bonded features in an arbitrary way, and that it is better, instead, to depict all of the inter-main chain hydrogen bonds. The pictures in Color Plates 1–4 provide a means of doing this. The programs that produced them (in a UNIX environment) are described. They are called '3D plots' (Color Plate 1) and '1D plots' (Color Plates 2–4). Pictures similar to the 3D plots have been described before,[8,11] although the display methods have been substantially altered and improved. The 1D plots are novel and provide a convenient means of describing three-dimensional features (namely, hydrogen bond patterns), of proteins in the context of the sequence. Because of the complexity of these patterns, new techniques for presenting them have been developed to ensure that the bonds overlap in the most meaningful way.

REFERENCES

1 Kamphuis, I.G., Kalk, K.H., Swarte, M.B.A. and Drenth, J. Structure of papain refined at 1.65 Å, *J. Mol. Biol.* 1984, **179**, 233–256

2 Baker, E.N. and Dodson, E.J. Crystallographic refinement of the structure of actinidin at 1.7 Å by a fast Fourier least-squares method. *Acta Crystallogr. Sect. A* 1980, **36**, 559–598

3 Richardson, J.S. Protein anatomy. *Adv. Protein Chem.* 1981, **34**, 167–339

4 Lesk, A.M. and Hardman, K.D. Computer-generated pictures of proteins. *Methods Enzymol.* 1985, **115**, 381–390

5 Burridge, J.M. and Todd, S.J.P. Protein secondary structural representations using real-time interactive computer graphics. *J. Mol. Graphics* 1986, **4**, 220–222

6 Carson, M. Ribbon models of macromolecules. *J. Mol. Graphics* 1987, **5**, 103–106

7 Priestle, J. *J. Appl. Crystallogr.* 1988, **21**, 572–576

8 Milner-White, E.J. and Poet, R. Loops, bulges, turns and hairpins in proteins. *Trends Biochem. Sci.* 1987, **12**, 189–192

9 Sibanda, B.L., Blundell, T.L. and Thornton, J.M. Conformation of β-hairpins in protein structures. *J. Mol. Biol.* 1989, **206**, 759–777

10 Baker, E.N. and Hubbard, R.E. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* 1984, **44**, 97–179

11 Poet, R. and Milner-White, E.J. Displaying relevant features of protein molecules. *Comp. Graphics Forum* 1986, **5**, 211–215

12 Milner-White, E.J. Recurring loop motif in proteins that occurs in right-handed and left-handed forms. *J. Mol. Biol.* 1988, **199**, 503–511

13 Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 1987, **326**, 347–352

14 Milner-White, E.J., Ross, B.M., Ismail, R., Belhadj-Mostefa, K. and Poet, R. One type of γ-turn, rather than the other, gives rise to chain reversal in proteins. *J. Mol. Biol.* 1988, **204**, 777–782

15 Collins, J.F. and Coulson, A.F.W. Molecular sequence comparison and alignment. In *Nucleic Acid And Protein Sequence Analysis—A Practical Approach* (Bishop, M.J. and Rawlings, C.J., Eds.) IRL Press, Oxford, 1987, pp. 323–358