BUTTERWORTH HEINEMANN

NEW PROGRAMS

FOLD: Integrated analysis and display of protein secondary structure

Darren R. Flower*

Department of Physical Chemistry, Astra Charnwood, Loughborough, Leicestershire, UK

FOLD, a computer program for the definition and analysis of protein secondary structure, is described. Algorithms implemented in the software are reviewed. These include methods for the identification of simple features such as hydrogen bonds, α helices, β strands, β bulges, and β and ψ turns. Techniques are also described for the definition and analysis of higher-order structures, such as β hairpins, β sheets and their topology, and β barrels. In addition to considerable textual output the program supports visualization of protein secondary structure in either an atom-based display style or one reproducing the characteristics of a so-called ribbon drawing.

Keywords: Protein secondary structure, hydrogen bonding, β sheet, protein topology, graph theory

INTRODUCTION

The structures of ideal α helices and β strands, proposed by Pauling and Corey in the early 1950s, provide the basis for the definition and analysis of repeating secondary structure elements observed in protein crystal structures. Perhaps the commonest approach to secondary structure definition is the human visual inspection of protein structure. This can prove laborious, somewhat subjective and imprecise, and is potentially time-consuming; consequently, many attempts to automate secondary structure definition have been made.

Since ideal secondary structures have characteristic values for their main-chain torsion angles and give rise to characteristic patterns of hydrogen bonding, both can be used as the basis of structure definition. Of the two, hydrogen bonding is probably the less tendentious criterion by which to assess secondary structures. For example, many crystallographic refinement programs make use of energy refinement or restrain secondary structure elements, both of which bias torsion angle data.[1] In general, this is not true of hydrogen bonds, which are seldom restrained explicitly during crystallographic refinement. Also, it is well known that residues at the periphery of secondary structure elements do not always have ideal torsion angle values.[2]

A new computer program, called FOLD, which implements new or enhanced methods for the definition, and subsequent analysis, of hydrogen-bonded protein secondary and supersecondary structures, is described in this article.

Color Plates for this article are on page 355.

*Address reprint requests to Dr. Flower at the Department of Physical Chemistry, Astra Charnwood, Bakewell Road, Loughborough, Leicestershire, UK LE11 ORH.

Received 23 May 1995; revised 1 August 1995; accepted 8 August 1995.

METHODS

Overview

FOLD assigns secondary structures using pattern recognition methods based on matching different characteristic hydrogen-bonding patterns. This draws on the rigorous interpretation of such patterns given by Baker and Hubbard[3] and Kabsch and Sander.[4]

The protocol adopted in the program is hierarchical: it begins with objective identification of all main chain–main chain hydrogen bonds and proceeds to find within this set patterns characteristic of α and 3_{10} helices and β strands. This allows individual residues to be defined as a particular secondary structure type. Assigned residues are then grouped into elements: helices and strands. In the case of β strands, having defined a set of β strands and their connectivity, straightforward techniques drawn from graph theory are used first to partition these strands into sheets and then to analyze and express their topology.

Identifying hydrogen bonds

Except in unusual circumstances, such as very high resolution or neutron diffraction data, hydrogen atom positions

Journal of Molecular Graphics 13:377–384, 1995

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(95)00072-0

cannot be determined by crystallography and so a definition of hydrogen bonds based solely on heavy atom positions is used: a modified version of that adopted by Baker and Hubbard.[3] This geometry is shown in Figure 1. Given user-defined distance and angle bounds this procedure allows for the automatic definition of hydrogen bonds between main chain atoms. Within the program, a hydrogen bond is expressed as an entry in a hydrogen bond connectivity matrix **H**: $\mathbf{H}(i, j) = 1$ indicates that a hydrogen bond exists between the main chain carbonyl oxygen of residue i and the main chain nitrogen of residue j.

Identifying helices

Having identified hydrogen bonds, FOLD proceeds to assign individual residues to a secondary structure type by matching observed hydrogen-bonding patterns to those characteristic of ideal secondary structures. The two main types of helical secondary structure analyzed by the program are the α and 3_{10} helices. Following Kabsch and Sander,[4] these patterns can be described by simple logical conditions expressed in terms of the hydrogen bond connectivity matrix **H**. Residues are assigned to one of two basic states if they fulfill the relevant condition:

α Helix: $\mathbf{H}(i, i + 4) = 1$
3_{10} Helix: $\mathbf{H}(i, i + 3) = 1$

Identifying β strands

β Structure is found by matching observed hydrogen-bonding patterns to those characteristic of ideal parallel and antiparallel β strands. Following Kabsch and Sander,[4] these patterns are described by simple logical conditions expressed in terms of the hydrogen bond connectivity matrix **H**:

Parallel β bridge:
$\mathbf{H}(i - 1, j) = 1$ and $\mathbf{H}(j, i + 1) = 1$
Inner bridge

or

$\mathbf{H}(j - 1, i) = 1$ and $\mathbf{H}(i, j + 1) = 1$
Outer bridge

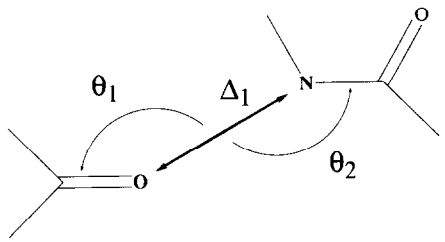

Figure 1. Hydrogen bond definition geometry. Definition of geometric parameters used to identify allowed hydrogen bonds within the program FOLD. Δ_1, Distance between donor and acceptor atoms forming hydrogen bond; θ_1 and θ_2, angles defining allowed orientation of participating peptide planes. For each parameter the user is able to set an upper and lower bound, the implicit tolerance tuned to the accuracy of the structure under study.

Antiparallel β bridge:
$\mathbf{H}(i, j) = 1$ and $\mathbf{H}(j, i) = 1$
Inner bridge
or
$\mathbf{H}(i - 1, j + 1) = 1$ and $\mathbf{H}(j - 1, i + 1) = 1$
Outer bridge

In the assignment process possible confusion is avoided by using a hierarchy of priorities, so that helical assignments take precedence over strands. Only residues i and j that are more than two residues distant in the sequence are checked for β-bridge connections.

Residues involved in a β bridge are said to be connected—the two residues being partners in the bridge. Note the distinction between inner and outer bridges made above: a given residue can make either an inner bridge, an outer bridge, or both. One can thus describe a residue in a β strand as being either doubly or singly connected, and optionally to describe whether a singly connected residue forms an inner or an outer bridge.

Element clustering: Formation of helices and strands

Having assigned all residues to a basic structural state, on the basis of matching a characteristic hydrogen-bonding pattern, FOLD uses these assignments to define secondary structure elements. The algorithm requires that at least two such assignments occur consecutively for a helix or strand to be defined. Individual elements, helices and strands, are identified in the program as continuous stretches of successive structural assignments.

FOLD supports two options for the definition of secondary elements. One uses the strict definition adopted by Kabsch and Sander.[4] The other uses the less restrictive definitions of IUPAC-IUB.[5] Helices are extended by one residue at each end relative to the Kabsch and Sander definitions. For strands, this option has two available modes. In the first, a strand is extended beyond its terminii, as defined by Kabsch and Sander, only if a terminal residue forms an outer β bridge. In the second a strand is extended, beyond these limits, if a terminal residue forms either an inner or an outer bridge.

Identifying β bulges

Once an initial identification of β strands has been made, it is possible to take account of interruptions in strands caused by β bulges. Bulges are a common form of defect or distortion observed in β sheets (reviewed by Richardson[2]), but also form important, characteristic features of β-hairpin loops.[6] Chan et al.[6a] have presented one approach to the automatic identification of β bulges. They show that two of the most common bulge structures are the classic and wide-type antiparallel β bulge; both of these structures have simple, but characteristic, hydrogen-bonding patterns, which are shown in Figure 2. FOLD identifies bulges explicitly in terms of simple hydrogen bond connectivity conditions, like those used to define repeating secondary structures, which can be generalized to situations in which there is more than

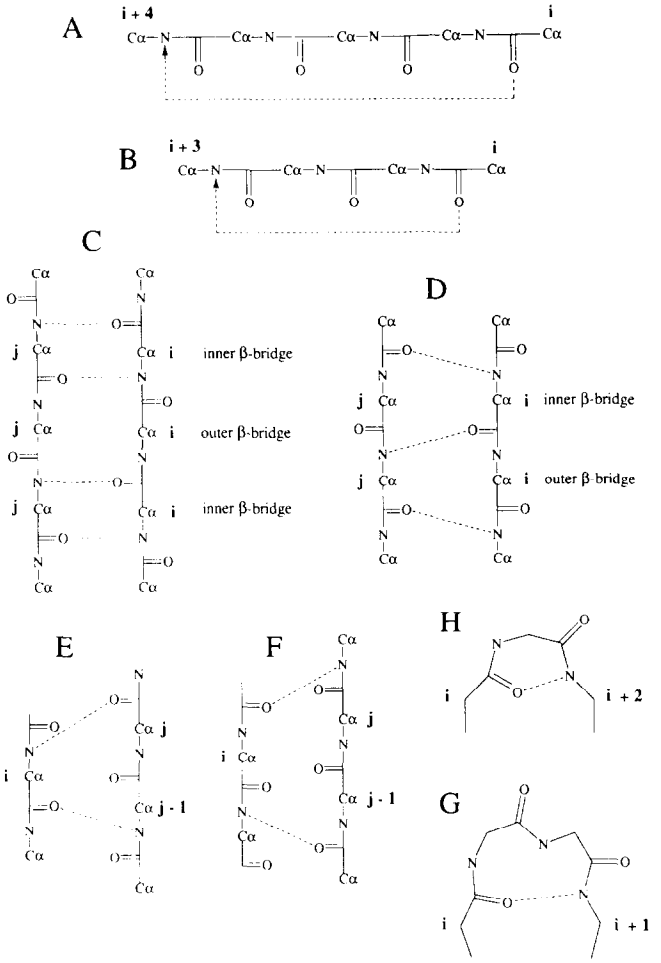

Figure 2. Hydrogen bond patterns of bridges and bulges. Schematic representations of main chain hydrogen bond patterns used to define repeating and nonrepeating protein secondary structures. Hydrogen bonds are shown as dashed lines. (A) α Helix; (B) 3_{10} helix; (C) antiparallel β bridge; (D) parallel β bridge; (E) classic β bulge; (F) wide-type β bulge; (G) β turn; (H) γ turn.

one residue in the bulge. For classic β bulges the corresponding logical expression is

$$\mathbf{H}(i, j - n) = 1 \quad \text{and} \quad \mathbf{H}(j, i) = 1$$

or

$$\mathbf{H}(j, i - n) = 1 \quad \text{and} \quad \mathbf{H}(i, j) = 1$$

and for wide-type bulges:

$$\mathbf{H}(j - 1, i + 1) = 1 \quad \text{and}$$
$$\mathbf{H}(i - 1, j + n + 1) = 1$$

or

$$\mathbf{H}(i - 1, j + 1) = 1 \quad \text{and}$$
$$\mathbf{H}(j - 1, i + n + 1) = 1$$

where n is the number of residues in the bulge. The maximum size of bulges found is set by the user.

FOLD allows residues participating in interstrand bulges to be reclassified as β structure if a bulge thus links two stretches of β residues. This causes disconnected sections of β strand separated by a bulge to be linked into one strand. Residues in such a bulge are referred to as unconnected rather than singly or doubly connected. Similarly, multiresidue sections of strands can be described as singly, doubly, or unconnected. Residues participating in bulges of a given size are reported separately by the program.

Identifying turns

Other well-studied features of protein structure are the β and γ turns.[7] Although some types of turn are classified by their conformation and lack a hydrogen bond, most turns are characterized by a particular hydrogen-bonding pattern. Again, these patterns can be expressed in terms of the hydrogen bond connectivity matrix **H:**

$$\gamma \text{ Turn:} \quad \mathbf{H}(i, i + 2) = 1$$
$$\beta \text{ Turn:} \quad \mathbf{H}(i, i + 3) = 1$$

Such structures are distinguished from helices by virtue of being isolated and do not form part of repeating structures. β Turns, in particular, have been studied extensively and many complicated, and often confusing, classifications for turn types have been proposed.[7] Rather than adopt any of these, FOLD uses a nomenclature, for both β and γ turns, inspired by that of Wilmott and Thornton.[8] Each of the residues comprising the turn, four and three residues for the β and γ turns, respectively, are denoted by a symbol corresponding to the region of the Ramachandran chart to which they belong (see Figure 3).

Identifying β sheets

Having identified a set of β strands, it is possible to determine the size and extent of the β sheets that they form. Following Koch et al.,[9] the structure of a protein β sheet can be expressed in terms of graph theory. The strands of a sheet correspond to the vertices of a graph and the hydrogen-bonded connection of strands to its edges. When viewed as such a graph, β-sheet identification can be seen as a clustering mechanism in terms of strand connectivity. The first step in the process involves forming an overall strand connectivity matrix **C.** This is constructed from knowledge of the β-bridge partners of each residue in a given strand. If some residue i of strand 3 is the partner of some residue j of strand 6, then strand 3 is connected to strand 6. At the same time, it is possible to determine whether two strands are parallel or antiparallel to each other from the nature of the bridges that link them. Given the minimum number of bridges between two strands required for them to be defined as part of the same sheet, the overall connection of strands can be determined. A value of 2 or 3 for this minimum value is typical. If strand i and strand j are connected then the corresponding entry in the strand adjacency matrix, $\mathbf{C_0}$, is given the value 1, that is, $\mathbf{C_0}(i, j) = 1$. A greedy algorithm, a variant of the shortest paths algorithm of Floyd,[10] is used to detect cliques in this graph by forming the strand connection matrix **C** from $\mathbf{C_0}$. Identifying these cliques allows partitioning of strands into sheets. The algorithm, if not its literal implementation, is well expressed by the following pseudocode:

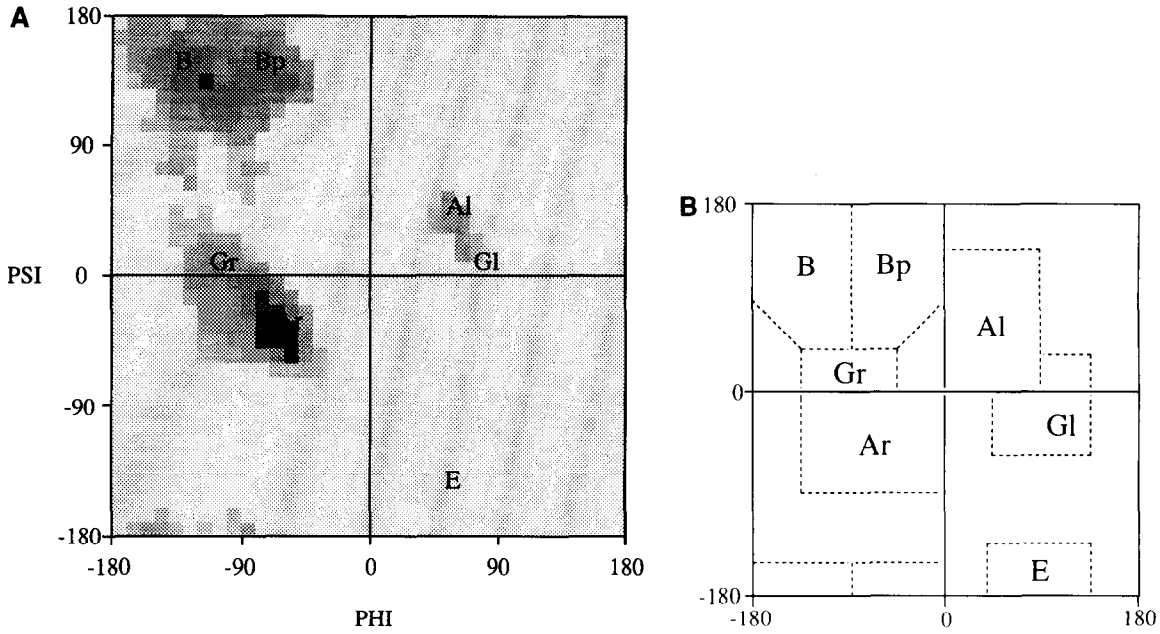

Figure 3. Definitions of regions of the Ramachandran chart. Ramachandran chart showing the labeling of distinct regions. These symbols (Ar, Bp, etc.) provide a short-hand for the corresponding conformation adopted by residues with main chain dihedral angles that fall into these regions. As described in the text, this notation is used by the program to describe the conformation of particular sections of backbone such as turns and bulges. (A) Ramachandran chart gray shaded to show the relative preference for discrete "allowed" regions exhibited by residues drawn from 69 high-resolution protein structures. (B) Definition of region boundaries. Each distinct region, corresponding to a discrete residue conformational state, is bounded by a dashed line and labeled with its short-hand symbol (Gl, E, etc.).

```
K ←1 to N_strand
I ←1 to N_strand
J1←1 to N_strand
J2←1 to N_strand

  if C(I,J1) = 1 and C(J2,I) = K or
    C(J2,I) = 1 and C(I,J1) = K then

    if C(J2,J1) = 0 then
      C(J2,J1) ← K + 1
    endif

  endif

end; end; end; end;
```

where N_{strand} is the total number of strands. After completion of this process, a nonzero entry, $\mathbf{C}(i, j)$, corresponds to the number of topological edges in the shortest path between strands i and j. Identifying cliques from this matrix allows partitioning of strands into sheets. If $\mathbf{C}(i, j)$ has a nonzero value then strands i and j are part of the same sheet. By passing through a row or column of the matrix and noting nonzero entries it is possible to determine the membership of a particular sheet in a single pass. Repeating this process with succeeding rows or columns, until all strands are accounted for, allows all sheets to be identified.

Identifying β hairpins

The strand connectivity matrix **C** can also be used in the analysis and classification of sheet structure and topology. For example, it is possible to identify so-called β hairpins: a feature of protein structures that has been scrutinized in some detail. Hairpins are loops that link two strands that follow each other directly in the sequence and are joined antiparallel. Sibanda et al.[11] proposed a nomenclature for β hairpins, based on the hydrogen bond pattern that closes the hairpin loop, and used it to classify manually large numbers of hairpins. It is possible to automate this classification. First, pairs of strands are identified for which $\mathbf{C}(i, j) = 1$ and $j = i + 1$. Second, by passing backward from the last residue of strand i, and examining the bridge partners of each of its residues in turn, the β bridge that closes the hairpin loop can be found. This corresponds to the β bridge, between strands i and j, with the smallest sequence separation between the two participating residues. If this is an outer β bridge then the hairpin is classified as $n - 2{:}n$, and if this is an inner β bridge it is classified as $n{:}n$,[11] where $n = j - i - 1$.

Identifying β barrels

When viewed as a graph, a β sheet is seen to be a complex topological object; like the atom/bond graphs of small molecules, a β sheet may be branched and can contain cycles, or rings. Having identified a sheet, it is possible to use properties of its connection matrix **C** to classify it meaningfully using a scheme that places all possible β sheets into one of four different classes.[12] This scheme defines a sheet as open or closed (mutually exclusive properties) and as either branched or unbranched (also mutually exclusive). Combination of these two characteristics gives four types: closed and unbranched, closed and branched, open and un-

branched, and open and branched. FOLD implements an automatic procedure for classifying sheets in this way.

Clearly, by definition, closed sheets contain cycles or rings. Such topological rings correspond, in the main, to β barrels. FOLD implements a method for identifying barrels based on ring perception.[13] The number of rings in a β-sheet graph can be determined from the total number of edges (connections) between vertices (β strands). If N_{ring} is the number of rings, then

$$N_{ring} = B_{edge} - N_{vertex} + 1$$

where N_{edge} is the total number of connections between strands and N_{vertex} is the number of strands in the sheet. Rings are familiar features of organic small molecules and their automatic perception, particularly in this context, has attracted a formidable literature. The many algorithms developed for automatic ring perception, and their theoretical basis, have been reviewed in detail by Downs et al.[14]

Thus if a β sheet is closed, the cycles it contains are found by the program using the method of Paton,[15] which has proved to be both fast and reliable. A depth-first path is traced through the strand adjacency graph $\mathbf{C_0}$, using a push-down stack of unused edges. When an edge is found to link two vertices already in the path this edge is flagged as ring closing. When the edge list is exhausted then the trace is terminated. The set of vertices forming the cycle associated with each ring-closing edge is found by backtracking through $\mathbf{C_0}$ using the path trace ordering to determine the shortest path between ring-closing vertices other than by their common edge. In this way a smallest set of smallest rings, each ring corresponding to a β barrel, is identified directly and unambiguously. Figure 4 gives an example of a β graph that contains a topological ring or barrel. The path trace and ring-closing connection are marked.

β Sheet topology

The topology of a protein β sheet, the relationship between the sequential ordering of strands and their hydrogen-bonded connectedness in space, is an important and well-studied property of such structures. Two systems of nomenclature have been proposed to describe β-sheet topology. One, first proposed by J. Richardson,[16] is based on following a path through the sequence order of strands and noting their spatial separation within the sheet. In this scheme only the connections between strands that follow each other in the sequence are considered, each connection having three properties: the physical separation within the sheet of the two participating strands, whether the strands are parallel or antiparallel, and whether the connection involves going forward or backward in the topology of the sheet. The other, lesser used approach,[9,12] is based on following a path through the connectedness of neighboring strands and noting their sequence separation. Rather than following the sequence order of strands a depth-first path is traced through the sheet and the labeled connections express the sequence separation between physically adjacent strands, whether it goes forward or backward in the sequence, and whether the strands are parallel or antiparallel. Figure 4 gives worked examples of both these types of topological nomenclature. It also gives the structure of the corresponding β graph.

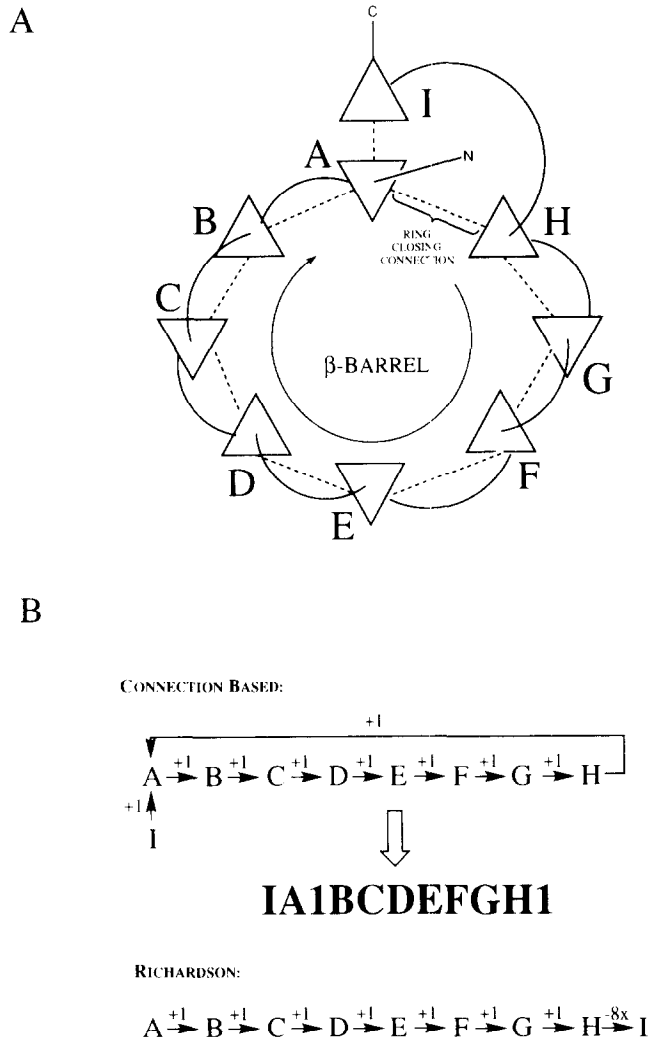

Figure 4. β Graph and topological nomenclatures. (a) The structure of the β graph, as generated with FOLD, corresponding to the structure of mouse major urinary protein (database code 1MUP[24]) is shown. It contains a topological ring; the path trace is marked, as is the ring-closing connection. The nine β strands of the antiparallel closed sheet, or barrel, of MUP are shown as triangles and labeled A–I. Triangles pointing downward indicate a strand direction into the plane of the paper and those pointing upward indicate a strand direction out of that plane. The hydrogen-bonded connectedness of strands is represented by a dotted line. Connecting loops are shown as solid lines. For simplicity of presentation, α helices are not shown. (b) Worked examples of Richardson and connection-based topological nomenclatures, for the structure 1MUP,[24] are given, together with the corresponding SMILES-like summary. These expressions correspond to the structure of the β graph shown above.

It is possible to automate the Richardson classification scheme bypassing forward through the sequence order of strands in a sheet. The separation between strands i and $i + 1$ corresponds to the number of edges in the shortest path between them, that is, the value of $\mathbf{C}(i, i + 1)$. Whether two nonadjacent strands are parallel or antiparallel can be found by creating the strand orientation matrix $\mathbf{V}$ in a man-

ner similar to that used to form the connectivity matrix **C**. Instead of incrementing elements of **V**, a truth table is used to assign values to an element. If strand a is parallel to strand b and strand c is parallel to b then strand a is also parallel to strand c. If strands a and c are both antiparallel to b, then strands a and c are again parallel to each other. If strand a is antiparallel to b and c is parallel to b, or vice versa, then strand a is antiparallel to c. **V** can be generated from $\mathbf{V_0}$, the initial parallel/antiparallel designation for adjacent strands, using the same greedy algorithm used above to generate **C** from $\mathbf{C_0}$. The generation of matrix **V** can be accomplished at the same time the matrix **C** is generated by including the truth table check of **V** within the innermost of the nested loops. Determining whether a connection goes forward or backward in the spatial topology of the sheet can be determined from inspecting strand separations, in **C**, relative to the first two strands in the sheet.

Equally, the connection-based nomenclature can be automated. A depth-first path is traced through the strand connectivity graph of a sheet and the sequence separation is noted between paired strands in path order. This is combined with an indication of whether the connection is parallel or antiparallel; this is obtained directly from $\mathbf{V_0}$ since only spatially adjacent connections are considered in forming this notation. Koch et al.[9] point out an ambiguity with regard to labeling: the separation in sequence can refer to the continuous numbering of strands in all sheets of the chain or only to strands of the same sheet, which they call a reduced notation. Both options are supported by FOLD and can be selected by the user.

New notation for β-sheet topology

β Sheets possess complex topological properties, including cycles and branches, which are not readily expressed by either form of consecutive notation. To overcome such limitations, a short-hand notation able to express β-sheet topology has been developed.[12] This system can be automated as a direct consequence of the internal graph representation of the sheet. In this notation, each strand is labeled with a letter corresponding to its sequence position within a protein chain, that is, the first strand is marked A, the second B, etc. This labeling can reflect either the continuous sequence numbering of strands within the same sheet or the continuous sequence numbering of strands through the whole chain, including all strands in all sheets. As a preliminary, all cycle-closing connections in the strand adjacency graph, if present, are removed to generate an acyclic graph or tree. The longest path through this tree is found and traced out marking each strand with its symbol (an A for the first strand, etc.). If present, ring-closing edges are marked with a single numeric label, incremented with each such edge, following the two strands (vertices) which form each closing edge. A ring is closed by the first matching digit and so closure numbers may be reused. In the unlikely event that more than 10 are open simultaneously then the number is preceded by a % sign. The sheet is denoted by passing through the depth-first path, in connection order, writing strand labels to form a string. When a branch point is encountered this branch of the tree is traced out in depth-first fashion. Such branches within the path are written as enclosures in brackets. To show the hierarchical treelike structure of the graph, brackets may be nested: branch points are marked with an open bracket. When a terminal node in the tree is encountered the most recent bracket is closed. The connection of strands is implied by the order of passing through the string. All connections are deemed to be antiparallel except where they are noted as parallel by the placing of an *x* in the path between connected strands or preceding a ring-closing digit. A worked example of this nomenclature is shown in Figure 4.

VISUALIZATION

FOLD also allows for the direct interactive visualization of the secondary structural features of proteins. The program has two display styles: an atomic mode (Color Plate 1) and a schematic mode (Color Plate 2). The atomic display style is detailed but straightforward, showing the backbone of the protein, color coded by secondary structure type, together with all of the hydrogen bonds identified by the program. Examination of this display, replete as it is with information, allows the user to assess carefully both the structure of a particular protein and the performance of the program for a given choice of parameters. By contrast, in its schematic mode, FOLD seeks to represent the overall structure of a protein in a highly simplified, but aesthetically pleasing, way. In common with most so-called ribbon drawings β strands are depicted as arrows, α helices as spiral ribbons, and other structure as a coiling rope or line. A number of programs have been developed that produce schematic line drawings, with or without shading, directly from atomic coordinates, in an essentially device-independent manner. The most popular of these programs have been, in turn, ARPLOT,[17] RIBBON[18] and its derivatives,[19] and MOLSCRIPT.[20] An alternative approach, exemplified by the programs RIBBONS[21] and SETOR,[22] uses hardware rendering to produce high-quality images. FOLD adopts this latter method, making use of the speed and power of the GL graphics library. This allows the display style and view to be manipulated interactively, although the definitions of secondary structure are those generated automatically by the program.

SOFTWARE

FOLD is written in standard Fortran 77 and was developed initially on a VAX running under VMS. The program was ported subsequently to run under UNIX on a series of Silicon Graphics workstations. FOLD is controlled via a simple command line interface through a set of keywords. The parameters used by the program, such as the geometric limits used to define hydrogen bonds, are fully configurable. FOLD is flexible in the type and quantity of output that it generates and can read protein structure coordinate data in PDB[23] and other formats. The program can operate in either of two modes, either reading, analyzing, and visualizing structures on an individual basis or automatically batch-processing sets of structures for the large-scale analysis of multiple proteins.

The Silicon Graphics version of FOLD supports GL-based visualization of hydrogen-bonding patterns and struc-

tural features. However, FOLD also generates a wealth of textual output. Data associated with each aspect of structure analysed (β bulges, β hairpins, sheet topology, etc.), as well as different types of partial or overall summary, are written to separate, self-naming files. An example of textual output is given in Figure 5 by way of illustration. This is taken from the sheet topology file for mouse major urinary protein.[24]

```
                    F O L D

            Summary of Secondary Structures

 Chain:  1 label: " "

 Sheet:  1 has  9 strands.

 Richardson Topology:

      Strand:   21   31
      Strand:   44   52   + 1.
      Strand:   54   64   + 1.
      Strand:   67   78   + 1.
      Strand:   83   88   + 1.
      Strand:   90   96   + 1.
      Strand:  103  113   + 1.
      Strand:  116  126   + 1.
      Strand:  151  155   + 2x.

 Topological summary:

1mup_A: +1, +1, +1, +1, +1, +1, +1, +2x.

 Topology based on Joining:

      Strand:  116  126
      Strand:  103  113   - 1.
      Strand:   90   96   - 1.
      Strand:   83   88   - 1.
      Strand:   67   78   - 1.
      Strand:   54   64   - 1.
      Strand:   44   52   - 1.
      Strand:   21   31   - 1.
      Strand:  151  155   + 8.

 Topological summary:

1mup_A: -1, -1, -1, -1, -1, -1, -1, +8.
1mup_A: H1GFEDCBA1I

 this sheet is CLOSED and BRANCHED.
  and contains 1 ring.

        ring number:   1. ring length:   8.

      Strand:   21   31   - 7.
      Strand:   44   52   + 1.
      Strand:   54   64   + 1.
      Strand:   67   78   + 1.
      Strand:   83   88   + 1.
      Strand:   90   96   + 1.
      Strand:  103  113   + 1.
      Strand:  116  126   + 1.

 Topological summary for ring:

1mup_A: -7, +1, +1, +1, +1, +1, +1, +1.

  sheet joining is antiparallel.
```

Figure 5. Example of textual output. An example of textual output is given by way of illustration. This is taken from the sheet topology file for mouse major urinary protein, database code 1MUP.[24] This output should be examined in conjunction with the β graph and worked examples given in Figure 4.

DISCUSSION

This article has described FOLD, a new computer program for the automated analysis of protein structures, and has elaborated a number of new or enhanced methods or nomenclatures, useful in such analyses, implemented in the software. Generally these methods represent an application of pattern recognition and graph theoretical methods to the study of an abstract representation of protein structure: in both developing and explaining these algorithms it is helpful to draw a comparison between the analysis of β-strand connectivity and the analysis of atomic connectivity in small molecules. Thus the work here represents an example of how graph theoretical methods developed in the study of small molecules can be easily transferred to the study of the three-dimensional structures of proteins. As such it forms a modest, but hopefully useful, addition to an increasingly important trend within the discipline.

FOLD can be obtained from the author and will be made available via QCPE.

ACKNOWLEDGMENTS

I thank the two, unfortunately anonymous, referees for their helpful, thoughtful, and knowledgeable comments. One referee, in particular, identified a number of errors and oversights in the original manuscript. The other referee made suggestions that have improved the display capabilities of the program.

REFERENCES

1 Morris, A.L., MacArthur, M.W., Hutchinson, E.G., and Thornton, J.M. Stereochemical quality of protein structure coordinates. *Proteins Struct. Funct. Genet.* 1992, **12,** 345–364

2 Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 1981, **34,** 167–339

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

4 Kabsch, W., and Sander, C. Dictionary of protein secondary structures: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, **22,** 2577–2637

5 IUPAC-IUB Commission on Biochemical Nomenclature. *J. Biol. Chem.* 1970, **245,** 6489–6497

6 Milner-White, E.J. β-Bulges within loops as recurring features of protein structure. *Biochim. Biophys. Acta,* 1987, **911,** 261–265

6a Chan, A.W.E., Hutchinson, E.G., Harris, D., and Thornton, J.M. Identification, classification, and analysis of β-bulges in proteins. *Protein Sci.* 1993, **2,** 1574–1590

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

8 Wilmott, C.M., and Thornton, J.M. β-Turns and their distortions: A proposed new nomenclature. *Protein Eng.* 1990, **3,** 479–493

9 Koch, I., Kaden, F., and Selbig, J. Analysis of protein

sheet topologies by graph theoretical methods. *Proteins* 1992, **12,** 314–323
10 Floyd, R.W. Algorithm 97 Shortest Path. *Commun. ACM* 1969, **12,** 345
11 Sibanda, B.L., Blundell, T.L., and Thornton, J.M. Conformation of β-hairpins in protein structures. *J. Mol. Biol.* 1989, **206,** 759–777
12 Flower, D.R. β-Sheet topology: A new system of nomenclature. *FEBS Lett.* 1994, **344,** 247–250
13 Flower, D.R. Automating the detection and analysis of protein β-barrels. *Protein Eng.* 1994, **7,** 1305–1310
14 Downs, G.M., Gillet, V.J., Holliday, J.D., and Lynch, M.F. Review of ring perception algorithms for chemical graphs. *J. Chem. Inf. Comput. Sci.* 1989, **29,** 172–187
15 Paton, K. An algorithm for finding a fundamental set of cycles of a graph. *Commun. ACM* 1969, **12,** 514–518
16 Richardson, J.S. β-Sheet topology and the relatedness of proteins. *Nature (London)* 1977, **268,** 495–500
17 Lesk, A.M., and Hardman, K.D. Computer generated pictures of proteins. *Methods Enzymol.* 1985, **115,** 381–390
18 Priestle, J.P. RIBBON—a stereo cartoon drawing program for protein structures. *J. Appl. Crystallogr.* 1988, **20,** 572–576
19 Flower, D.R. Improved ribbon drawing programs. *J. Mol. Graphics* 1991, **9,** 257–258
20 Kraulis, P.J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 1991, **24,** 946–950
21 Carson, M. Ribbon models of macromolecules. *J. Mol. Graphics* 1987, **5,** 103–106
22 Evans, S.V. SETOR: Hardware lighted three-dimensional solid modelling representation of macromolecules. *J. Mol. Graphics* 1993, **11,** 134–138
23 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The protein databank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542
24 Bocskei, Zs., Groom, C.R., Flower, D.R., Wright, C.E., Phillips, S.E.V., Cavaggioni, A., Findlay, J.B.C., and North, A.C.T. Pheromone binding to two rodent urinary proteins revealed by X-ray crystallography. *Nature (London)* 1992, **360,** 186–189