SHORT COMMUNICATION

A topological nomenclature for protein structure

Darren R. Flower

Department of Physical and Metabolic Sciences, Astra Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, UK

Keywords: α-helix/β-sheet/graph theory/protein structure representation/protein topology/topological nomenclature

In recent years, technical advances in X-ray crystallography and multi-dimensional NMR, coupled with comparable advances in protein engineering and computing, have led to an unprecedented growth in the number of solved protein structures, their number at least doubling every few years. This embarrassment of riches has, in turn, led to an increasing interest in the molecular systematics of protein structures (Wodak, 1996).

The remarkable commonalities of protein structure mean that the search for structural relationships is a rich and rewarding avenue of investigation. With only relatively few exceptions, the three-dimensional structure of proteins is characterized by patterns of repeating secondary structure: α-helices and β-sheets (Chothia and Finkelstein, 1990). Structural topology—the relationship between the sequential ordering of such secondary structure elements and their spatial organisation—is one of the principal means by which protein structures and their commonalities can be classified, categorized and compared. The topologies exhibited by β-sheets—the relationship between the sequential ordering of strands and their hydrogen-bonded connectedness in space—have been studied particularly well. Two systems of nomenclature have been proposed to describe β-sheet topology.

The first and better known of these is that developed by Richardson (1977): it is based on following a path through the sequence order of strands and noting their separation within the β-sheet. Only the connections between strands which 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, less widely known, approach is based on following a path through the connectedness of neighbouring 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 labelled connections express the sequence separation between physically adjacent strands, whether this goes backwards or forwards in the sequence and whether the strands are parallel or antiparallel.

Several authors have applied graph theoretical methods to facilitate the study of β-sheet topologies. Koch et al. (1992) introduced the problem and formulated approaches to searching and comparing different topologies. They noted that protein structures are complex topological objects possessed of properties and characteristics not readily expressed by a consecutive notation, such as that of Richardson (1977). A β-sheet, for example, may contain cycles and be branched. Topological cycles in β-sheets correspond to well studied geometrical features of protein structures, so-called β-barrels (Flower, 1994b). A topological notation based on listing consecutive connections will not deal with branching and sheet closures. A nomenclature for β-sheet topology, based on ideas from graph theory which resolves these problems, has been described previously (Flower, 1994a). Likewise, graph theoretical methods first developed to identify topological rings in small molecules can be applied to the automated detection and topological analysis of protein β-barrels (Flower, 1994b). Most recently, I have explored how graph theoretical techniques can be combined with established pattern recognition methods to help automate the analysis of super-secondary structures (Flower, 1995a).

The topology of a protein structure, the relationship between the ordering and connectedness of secondary structure elements, can, in a more general way, also be expressed in terms of graph theory. Individual helices and strands correspond to the vertices of a graph and their connection to its edges. In this context, connection means some form of significant non-bonded contact. This can be either the hydrogen-bonded connectivity between β-strands or the side-chain interactions which characterize the packing between any secondary structure element: helix–helix, helix–strand or strand–strand. In practice, one might wish to place a minimum value on the number of side-chains in van der Waals contact or on the surface area lost on forming the association.

The work presented here extends the application of ideas drawn from graph theory to the analysis of protein structure and a new nomenclature for protein structural topology is described. The proposed notation is based on representing a set of secondary structure elements and the relationship between them in the form of a labelled tree. As originally noted in 1857 by the English mathematician Arthur Cayley, the branching structure of a tree can be represented as nested parentheses. There exist a number of line notations which use this concept to represent the complex and highly non-linear topology of chemical structures in a linearized form (Weininger, 1988; Ash et al., 1997).

In the notation presented here, each secondary structure element is labelled with a letter corresponding to its sequence position within a protein chain: the first element is marked A, the second B, etc. Unlike the sheet nomenclature described earlier, the present notation does not give rise to ambiguities of continuous or discontinuous labelling. The nature of an element (helix or strand) is distinguished by a subsidiary symbol. If a structure contains closures, then, as a preliminary, all cycle closing connections are broken and marked by a single numeric label—incremented with each such edge—following the two elements (vertices) which form the edge. A cycle is closed by the first matching digit and so cycle numbers may be reused. In the event that more than 10 cycles...
Worked examples of topological nomenclature. (a) A series of steps indicating how a structure is reduced in stages to a labelled graph and then linearized. The first part of the diagram shows a simple protein structure drawn according to a common convention: β-strands are shown as triangles, after the diagrammatic style of Sternberg and Thornton (1977b), with triangles pointing downwards indicating a strand direction into the plane of the paper and those pointing upwards indicating a strand direction out of the plane of the paper. The hydrogen-bonded connectedness of strands is represented by a connecting dotted line. Connecting loops are shown as solid lines. α-Helices are shown as circles, with open circles indicating helices pointing into the plane of the paper and circles with an inner dotted circle pointing outwards. The second part shows connections—hydrogen-bonded or side-chain contacts—between secondary structure elements. The third part of the diagram shows the corresponding graph with vertices labelled. The last part shows the disconnection of cycles within the graph and the corresponding relabelling of vertices. The last part of the diagram shows topological summaries corresponding to the acyclic form of the graph with labelling of graph edges. The set of final topological summaries are equivalent and are presented to illustrate the flexibility of style inherent within the system. A plain text equivalent might appear as follows:

\[
A : B \ (C^1 : D^1 (E^2) : F^2; G^β) \\
A^α : B^β \ (C^α 1; D^β 1(E^α 2); F^β 2; G^β) \\
B^β A^γ B^β \ (α^α 1; D^β 1(E^β 2); F^β 2; G^β) \\
B^β (α 1; D^β 1(E^β 2); F^β 2; G^β) \\
A : B \ (C^1 α β 2) : D^β 1(E^α 2); F^β 2; G^β \\
A^α : B^β 1(D^β 2(C^α 1); F^β (E^α 2); G^β)
\]

(b) An illustration, using the structure from (a), of how the process of removing different edges from cycles and following different paths through the acyclic graph can give rise to alternative ways of writing the same topological summary. Starting with the same structure (top), we proceed through two alternative routes to yield two different, but equivalent, summary strings for that structure. The final convergence is to a canonical form where the sequence order of vertices is used to avoid ambiguity and remove redundancy.

are open simultaneously then the number is preceded by a % sign.

A structure is effectively reduced by this process to an acyclic graph or tree. The structure is then denoted by walking through the depth-first path, in connection order, to form a string, marking each secondary structure element with its label (‘A’ for the first element, ‘B’ for the second, etc.). Branches within the path are traced out in depth-first fashion. Such branches are written as enclosures in brackets: branch points are marked with an open bracket; when a terminal node in the tree is encountered the most recent branch is closed. To show the hierarchical tree-like structure of the graph, these enclosures may be stacked or nested, to form brackets within brackets, branches within branches.

The connection of secondary structure elements is implied by the order of passing through the string. All connections are assumed to be antiparallel except where they are noted otherwise. Parallel connections are marked by placing an ‘x’
A topological nomenclature for protein structure

Fig. 2. Extensions to the nomenclature. This figure shows some useful extensions to the nomenclature allowing us to represent domains and complexes. A protein structure with two domains—one of which is helical in nature, the other being an antiparallel β-barrel—complexed with a small peptide forming a β-hairpin. Isolating each of these three nominal components allows three sub-strings to be generated easily. We can then place these together, using a comma to separate the discrete domains and a period to delimit the separate partners in the complex.

in the path between connected secondary structure elements; this includes preceding a ring-closing digit. Two different kinds of connection are recognized: the hydrogen bonded connection of two strands and the non-bonded contact of two secondary structure elements characterized by significant side-chain packing. All connections are deemed to be of the second, more general, type except where they are shown to be of the first type by a colon in the path between connected elements; this includes ring-closing digits. A worked example of the nomenclature is presented in Figure 1.

In the general case, the topological representation of protein structure described above allows for a number of different orderings of labels within the string. However, a simple extension of the approach allows for straightforward canonicalization. Of the possible closing edges within a topological cycle, the edge lowest in sequence order is chosen to be cut. Likewise, when traversing the acyclic graph, the sequence order of secondary structure elements determines a unique order for tracing out the path: beginning with the terminal node lowest in sequence, a path is followed such that at each branch point branches are chosen in ascending sequence order.

Another extension to the notation allows other information to be associated with each secondary structure element. Each element label can be suffixed by a ‘{ }’ which can contain one or more data items separated by a semi-colon. These data items might be the residue range of the element or some functional flag, etc. For plain text versions of the notation, such as ASCII, without the ability to subscript fonts, these brackets contain symbols denoting the type—helix or strand—of each secondary structure element (see caption to Figure 1). Distinct domains can be identified by separating discrete sets of secondary structure elements which form a domain by a comma. Separate protein chains, in a protein–protein complex, for example, can be separated by a full stop or period. Figure 2 gives an example of these extensions to the nomenclature.

One aspect of protein structure with which this notation does not deal is chirality or ‘handedness’. Several approaches have been taken to characterizing the chirality of or within protein structures. Some have looked at their covalent chirality (Kikuchi et al., 1989; Liang and Mislow, 1994), while others have looked at the chirality of local structures, such as the handedness of crossover connections (Sternberg and Thornton 1976, 1977a). As yet no general method for specifying chirality and no automated method for its perception have been developed (Slidel and Thornton, 1995).

In line with standard practice (Richardson, 1977; Lesk, 1995), I have concentrated here on describing the present system of nomenclature, without attempting to bias readers in their choice of potential applications. However, Figure 3 contains a few simple examples which help to illustrate how this nomenclature can be used in the analysis of protein topology. In assessing its utility, a distinction should be drawn between the notation itself, the graph theoretical representation of protein structure which underlies it and software needed to compare it. For example, it should be possible, using techniques from computational graph theory, to generate these topological strings in an automated fashion, allowing incorporation of the technique into programs for the automated analysis of protein structure. There are many such programs, including DSSP (Kabsch and Sander, 1983), FOLD (Flower, 1995a) and PROMOTIF (Hutchinson and Thornton, 1996). The notation is itself part of a larger programme to apply graph theory to protein topology (Koch et al., 1992; Flower, 1994b, 1995a) and should be viewed in that context.

Within the arena of protein structure analysis, there have been other recent attempts to develop simple linear topological
summaries akin to that presented here. For example, Lesk (1995) used Forsyth notation (most familiar as a way of representing the location of pieces on a chess board) to summarize a tableau representation of protein structure, while Groß (1996) used linguistic trees to represent folding topology. The work of Groß is of special interest because it stresses the link between natural language and the systematic representation of structure. Human language is a redundant means of communicating complex information, where its own internal structures can obscure meaning: what we say is lost in how we choose to say it. Systems of nomenclature—good ones at least—escape the confinement of language and, through conciseness and specificity, become instruments not only of communication but of reasoning also. Abstract structures are, in a sense, encoded by the notation, but that encoding is both transparent, simplifying and potentially illuminating. Mathematics is an obvious example, but music and dance have also developed effective but almost entirely abstract or symbolic notations which function better than more literal representations (Hutchinson Guest, 1984).

In this paper I have described a short-hand notation which extends existing nomenclature for describing the topology of protein structures. The system described has many desirable features. Because the notation is based on the connectedness of secondary structure elements, it contains all the information needed to reconstruct the topological structure of a particular protein fold and yet it is able store this information in a very concise way. For example, such topological summaries contain all necessary information for input to automatic topology visualization programs, such as TOPS (Flores et al., 1994).

However, this does not lead to an impenetrable and confusing terseness. Although the number of rules defining this system of nomenclature is relatively small, it is difficult to describe them in a lyrical manner; nonetheless, once understood, the
A topological nomenclature for protein structure

Fig. 3. Example applications. (a) Topological analysis of α/β proteins. Three related structures are compared and contrasted with another, superficially similar structure. It is clear from the topological summaries that the three top structures (the N-terminal domain of aspartate carbamoyltransferase (pdb code: 8atc), S6 ribosomal protein (1ris) and the N-terminal domain of the apatase fragment of the 70 kDa heat shock protein) share, in common with many other proteins not shown here, a topological identity. In contrast, the structure of chorismate mutase (2chs) shares only a gross similarity and, as shown by the nomenclature, is topologically distinct. (b) Topological analysis of up-and-down β-barrels. Three closely related eight-stranded antiparallel β-barrel structures are compared and contrasted with another, superficially similar, structure. It is clear from the topological summaries that the lipocalin and avidin structures are topologically identical and are closely related to the trinub structure. These similarities are in line with the results of close structural comparison (Flower, 1993, 1995b, 1996; Flower et al., 1993; Fuentes-Prior et al., 1997). Cyclophilin, while falling into the same class, is clearly topologically quite distinct. (c) Topological analysis of topological features. An example is given of using a topological summary to help enumerate topological motifs within a structure, in this case cyclophilin. Two simple motifs are used: β-hairpins and ‘ψ-loops’.

notation is able to represent the complex non-linear topology of protein structure, including bifurcations and closed structures, in a clear and natural way. Using the notation it is possible simultaneously to follow sequential and spatial connectedness within the topology of a given structure. This property invests the notation with the ability to represent fully the organization of a protein structure in a non-diagrammatic way which is straightforward and easily understood.

The implicit simplicity of this nomenclature and its reassuring consistency with extant nomenclatures make it an accessible and user-friendly way to represent potentially complex protein structures, yet its origins in graph theory and the computer representation of small molecule structures make it equally computer-friendly. These properties may make it an effective medium for the computer storage of topological patterns—the notation provides a convenient, compact means to store this information in a computer-readable form which remains amenable to direct human interrogation. This system, combining simplicity with completeness and generality, should, potentially, be a powerful tool for the comparison, classification and analysis of protein structures.

References


Received October 2, 1997; revised February 9, 1998; accepted March 3, 1998