ELSEVIER

FEBS Letters 344 (1994) 247–250

FEBS 14007

FEBS LETTERS

β-Sheet topology

A new system of nomenclature

Darren R. Flower*

Department of Physical Chemistry, Fisons Plc, Pharmaceuticals Division, R & D Laboratories, Bakewell Rd., Loughborough, Leicestershire, LE11 ORH, UK

Received 17 March 1994

Abstract

The topology of a protein β-sheet, the relationship between the sequential ordering of strands and their connectedness in space, is an important and well studied feature of protein structures. The prevalent nomenclature for describing β-sheet topologies is based on following a path through the sequence order of strands and noting their separation in space. Although powerful, this approach can be usefully complemented by a notation based on following a path through the connectedness of neighbouring strands and noting sequence separation. This leads in turn to a short hand expression of sheet topology, based on a method for describing the covalent structure of small molecules, which is able to express concisely the complex non-linear topological relationships of β-sheets, including bifurcations and closed structures, in a clear and natural manner. Using this novel system of notation it is possible to follow simultaneously the sequence and hydrogen bonded connectedness of strands within the topology of a sheet.

Key words: β-Sheet; Protein topology; Nomenclature; Protein structure representation; Graph theory

1. Introduction

The overall structural pattern, or fold, of a globular protein is typically dominated by elements of repeating secondary structure: α-helices and β-strands [1]. While α-helices may exist in isolation, β-strands exist only as part of cooperative structures: β-sheets. The topology of a protein β-sheet, the relationship between the sequential ordering of strands and their hydrogen bonded-connectedness in space, is one of the most important and best studied properties of such structures [2–5]. Most recently, Woolfson et al. have shown that simple topological constraints underlie much of the apparent diversity and complexity of β-structures [6].

The nomenclature most often used to describe and classify the topology of β-sheets is that proposed by Jane Richardson [2]: it is based on following a path through the sequence order of strands and noting their separation in space. However, as shown below, this powerful approach can be usefully complemented by a system of notation based on following a path through the connectedness of neighbouring strands and noting their sequence separation. Moreover, this alternative approach leads directly to a short hand expression of sheet topology, based on a method for describing the covalent structure of small molecules, which is able to fully express the complex topological relationships in β-sheets in a complete, but concise, way.

2. Materials and methods

2.1. Topological classification of β-Sheets

Following Koch et al. [4] the topology of a protein β-sheet, the relationship between the ordering of strands and their connectedness, 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, a β-sheet is seen to be a complex topological object possessed of properties and characteristics not readily expressed by Richardson's consecutive notation [2]. Like the atom/bond graphs of small molecules [7], a β-sheet may contain cycles, or rings, and be bifurcated, or branched. Topological rings in β-sheets often correspond to well studied geometrical features of protein structures, so-called β-barrels [8]. Bifurcation can result from the packing defects sometimes observed in β-sheets [9], such as β-bulges and β-blowouts.

It is possible to classify all possible β-sheets into one of four different classes. This scheme defines a sheet as open or closed (which are mutually exclusive) and as either branched, or unbranched (also mutually exclusive). A closed sheet, either branched or un-branched, contains one or more cycles or rings, an open sheet does not. Combination of these characteristics gives four types: open and branched, open and unbranched, closed and branched, and closed and unbranched. Examples of β-sheets which fall into each of the four types are shown diagrammatically in Fig. 1.

*Corresponding author. Fax: (44) (509) 266 738.
E-mail: FLOWER_DR@FISONS.COM *or*
FLOWER_DR@FISONS-PHARM.CO.UK.

Abbreviations: DHFR, dihydrofolate reductase; MUP, mouse major urinary protein; FKBP, FK506 binding protein; GOX, glycolate oxidase.

SSDI 0014-5793(94)00393-A

2.2. A connection based nomenclature

Richardson [2] has proposed a nomenclature for classifying the topology of β-sheets based on following a path through the sequence order of strands and noting their separation in space. In this scheme 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.

This procedure can be usefully complemented by an approach based on following a path through the connectedness of neighbouring strands and noting sequence separation. This alternative topological classification scheme is based on the hydrogen bonded connectivity of the sheet. It is closely analogous to that of Richardson's, but 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 it goes forward or backward in the sequence, and whether the strands are parallel or antiparallel. Such a topological representation has been described before, for example by Koch et al. [4], but its value has not been appreciated and remains little used.

Koch et al. also point out an ambiguity with regard to labelling, 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. They also note that a notation based on lists of consecutive connections deals badly with bifurcations and sheet closures. In the next section a notation is developed able to deal with these problems in a natural and lucid fashion.

2.3. A new notation

Following from the above discussion, the graph theoretical analysis of small molecule immediately suggests a short hand notation able to express β-sheet topology both concisely and completely. This notation is an adaptation of the SMILES nomenclature system for expressing the covalent structure of organic compounds: for full details of this system the reader is referred to the original paper [10]. In brief, it is a line notation able to express the connectivity of any chemical structure as a string of alphabetic characters (denoting atoms) and special characters (denoting topology, bonding, etc.).

The rules of the modified sheet nomenclature system are described below. They are simpler than SMILES and somewhat different from it: strand labels are used rather than element symbols and rather than bond order symbols the parallel/antiparallel sense of strand connectedness are shown, but the means of representing branching and cycle closure are retained. Two worked examples of this nomenclature are presented in Fig. 2. This shows an open, branched mixed parallel/antiparallel sheet (DHFR) and a closed antiparallel β-barrel (MUP): examples which will be referred to below. In our notation, each strand is labelled with a letter corresponding to its sequence position within a protein chain, i.e. the first strand is marked A, the second B, etc. This labelling 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. If a sheet is closed, as MUP in Fig. 2, then ring closing edges are perceived, broken, and marked by a single digit following the two strands (vertices) which form this closing edge, i.e. between strands A and H in MUP. 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 then denoted by passing through the depth-first path, in connection order, writing strand labels to form a string. Branches within the path, such as involving strands G,H, and I in DHFR, are written as enclosures in brackets, i.e. I(G)H. To show the hierarchical tree-like structure of the graph, these enclosures may be nested, or stacked, to form branches within branches, brackets within brackets. 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, including preceding a ring closing digit, such as between strands D,C, and E in DHFR which become the string DxCxE.

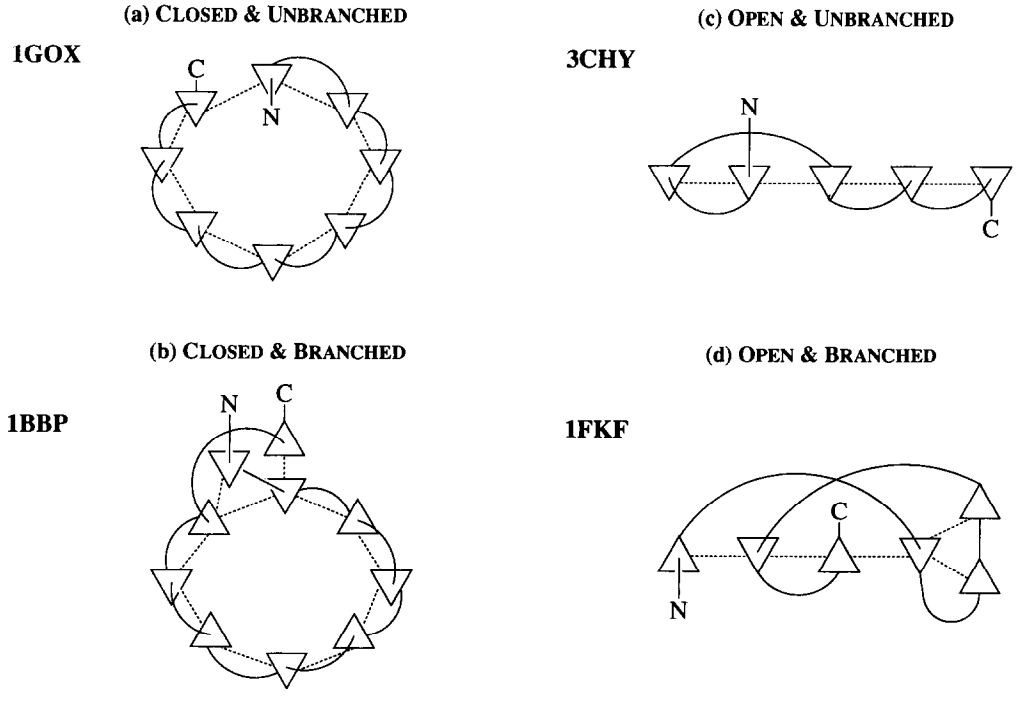

Fig. 1. Definitions of sheet classes. β-Strands are shown as triangles, after the diagrammatic style of Sternberg and Thornton [11], with triangles pointing downwards indicating a strand direction into the plane of the paper and those pointing upwards indicating a strand direction out of the plane of the paper. The hydrogen bonded connectedness of strands is represented by a connecting dotted line. Connecting loops are shown as solid lines. For simplicity of presentation, α-helices and other non participating strands are not shown. (a) Closed and unbranched; an 8 stranded parallel sheet from Glycolate Oxidase (database code 1GOX, [14]). (b) Closed and branched; a ten stranded antiparallel sheet from Bilin Binding protein (code 1BBP [15]). (c) Open and unbranched; a five stranded parallel sheet from *E. coli* CheY protein (code 3CHY [16]). (d) Open and branched; a six stranded antiparallel sheet from human FK506 Binding protein (code 1FKF [17]).

3. Results

The systems of notation outlined above have been used to describe the topological structure of representative sample of dissimilar β-sheets drawn from the Brookhaven protein databank [12]. Table 1 lists different topological expressions for both the Richardson and connection based topology of the sheets in our sample, as well as the corresponding SMILES-like topological summaries for each sheet. In deriving the connection-based topologies and summary strings, within-sheets or reduced labelling is used through out for simplicity. The four sheets from Fig. 1 are also included so that these expressions may also be compared with a corresponding diagrammatic representation of sheet topology.

It is clear from comparison of these forms of expression that while Richardson's method does not suffer from an ambiguity with regard to within-sheet or within-chain numbering of strands, it can label connections as parallel even in purely antiparallel sheets, and vice versa, since it does not deal in direct connections and so can prove confusing to the untrained eye. Also both fail to show clearly more complex, non-linear topological structures such as cycles and branches: although both have strengths it is clear than neither form is without weaknesses. Compared to the other two forms, summary strings are more compact and yet more complete in the information they display: all spatial connections, and their parallel or anti-parallel nature, are shown explicitly while all sequential connections are also shown through the labelling of strands. Moreover, the expression of features such as sheet closure and bifurcation is both clear and self evident within these strings.

4. Discussion

A new SMILES-like nomenclature for describing β-sheet topology has been described above. This scheme has a number of useful qualities. Because it is based on the physical connections between strands it contains all the information needed to reconstruct the topological structure of a sheet and yet is able to store this information in an extremely compact way. However this does not lead to an impenetrable and confusing terseness. Its implicit simplicity, and its reassuring consistency with extant nomenclatures [2,6] make it an accessible and user-friendly way to represent the potentially complex structures of β-sheets; while its origins in graph theory and the computer representation of small molecule structures make it, like its counterpart SMILES, very computer-friendly. This property may make it an effective medium for the computer storage of topological patterns. A number of workers have addressed the problem of searching known protein structures for particular β-sheets topologies [4,13]. They have employed graph theoretical methods to identify recurrent patterns of topology in databases of sheet structures; the notation described here provides a convenient, compact means to store this information in a computer readable form which remains amenable to direct human interrogation.

The present work has aimed to extend existing nomen-

Table 1
A sample of β-sheet topologies

Sheet label	Richardson	Connection	Summary
1gox:	+ 1x, + 1x, + 1x, + 1x, + 1x, + 1x, + 1x.	+ 1x, + 1x, + 1x, + 1x, + 1x, + 1x, + 1x.	Ax1xBxCxDxExFxGxHx1
1bbp:	+ 2x, + 1, + 1, + 1 + 1, + 1, + 1, + 1, + 2x.	+ 8, −1, −1, −1, −1, −1, −1, −1, + 8.	A11HGFEDCB1J
3chy:	+ 1x, −2x, −1x, −1x.	−1x, + 2x, + 1x, + 1x.	BxAxCxDxE
1fkf:	+ 3, + 1, + 1x, −3, + 1.	+ 4, + 1, −4, + 1, + 2.	AEFB(D)C
1fnr_A:	+ 2x, + 1, −3, −1, + 5, −3.	−1, + 3, −6, −5, + 1, + 3.	EDG(A)BCF
2glsB_D:	+ 1, −3, + 1, −2x.	−1, + 4, + 3, −1.	BA(E)DC
5p21_A:	+ 2x, −1, + 2, + 1x, + 1x.	+ 1, −2, + 3x, + 1x, + 1x.	BCAxDxExF
1ezm_B:	+ 1x, + 2, −1.	+ 1x, + 2x, −1.	AxBxDC
4gpd2_B:	+ 1, −2x, −2x, −2x, + 1.	−1, + 2, + 3, −1, −2.	BACF(E)D
3sc2A_A:	+ 1, + 2, −1x, + 2x, −4x, −1x.	+ 1, + 3, + 2, −1x, + 3x, + 1x.	AB(E)DxCxFxG
1r1a1_A:	+ 2x, + 1, −2x, −2x.	+ 5, −4, −3, + 2, + 1.	AF(B)C(E)D
1ximA_A:	+ 1x, −3x, + 1x.	−1x, + 3x, −1x.	BxAxDxC
2bpa2_A:	+ 4x, −3, −2x, −2x, + 2x, −1, + 2x.	+ 2, + 5, −4, −2, −3, + 2, −1.	ACH(DB)EGF
1eaf_A:	+ 4, + 1, −2x, −2, −1, −2x, + 1.	+ 7, −1, −2, −1, −3x, + 2, + 1.	AH(G)FExB(D)C
2dpv_B:	+ 1, + 2x, + 2x, −3, +4x, −2x.	+ 1, + 3, −2, + 4, −3, + 2.	ABECGD
4sbvA_A:	+ 2x, + 1, −2x, −2x.	+ 4, −3, + 2, + 1.	AEB(D)C

Richardson and connection-based topologies of the sheets in a sample of 16 dissimilar β-sheets drawn from the Brookhaven protein databank [12]. The corresponding SMILES-like topological summaries for these sheets are also shown. The first four sheets correspond to the examples presented in Fig. 1. Sheet labels are formed from the four letter Brookhaven code, optionally followed by the chains label, and suffixed by the sheet code with its chain.

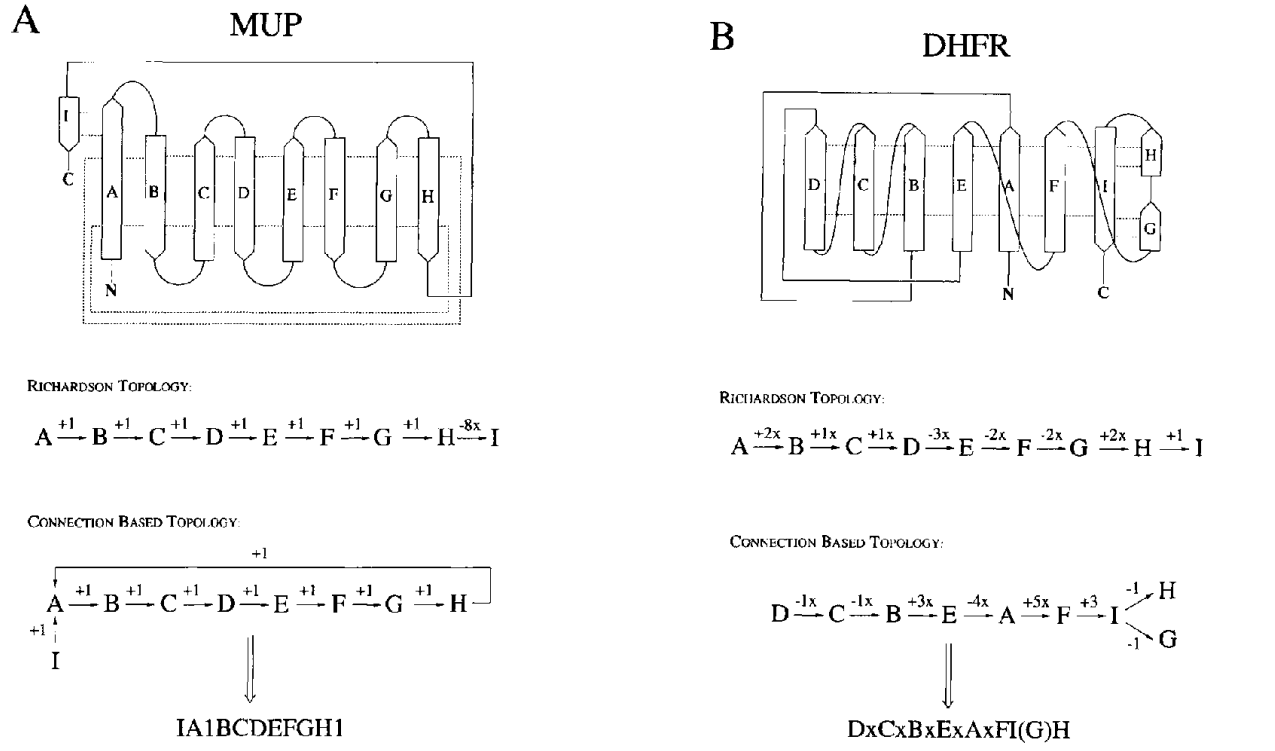

Fig. 2. Worked examples of sheet topologies. In each diagram the hydrogen bonded connection of two strands is indicated by a pair of dotted lines between them. Connecting loops are shown as solid lines. For simplicity of presentation, α-helices and other non-participating strands are not shown. For each the Richardson and connection-based topological expressions are given together with the corresponding SMILES-like summary. (A) The nine β-strands of the antiparallel closed sheet, or barrel, of Mouse Major Urinary protein [18] are shown as arrows and labelled A–I. (B) The nine β-strands of the open mixed parallel and antiparallel sheet of dihydrofolate reductase [19] are shown as arrows and labelled A–I.

clature for describing the topology of protein β-sheets. It evinces a system based on the hydrogen bonded connectedness of strands, complementary to the sequence based approach most commonly used to describe topology [2,4], which leads directly to a novel system of topological notation which is, unlike other schemes, able to represent the complex non-linear topology of β-sheets, including bifurcations and closed structures, in a clear and natural manner. Using this short-hand notation it is possible to follow simultaneously the sequence and hydrogen bonded connectedness of strands within the topology of a sheet, and it is this property which invests our notation with the power to fully represent a sheet structure in a clear, simple, yet non-diagrammatical way.

References

[1] Chothia, C. and Finkelstein, A.V. (1990) Annu. Rev. Biochem. 59, 1007–1039.
[2] Richardson, J.S. (1977) Nature 268, 495–500.
[3] Chirgadze, Y.N. (1987) Acta Cryst. A43, 405–417.
[4] Koch, I., Kaden, F. and Selbig, J. (1991) Proteins 12, 314–323.
[5] Stirk, H.J., Woolfson, D.N., Hutchinson, E.G. and Thornton, J.M. (1992) FEBS Lett. 308, 1–3.
[6] Woolfson, D.N., Evans, P.A., Hutchinson, E.G. and Thornton, J.M. (1993) Prot. Eng. 6, 461–470.
[7] Trinajastic, N. (1983) Chemical Graph Theory, CRC Press, Boca Raton, FL.
[8] Lasters, I., Wodak, S.J., Alard, P. and van Cutsem, E. (1988) Proc. Natl. Acad. Sci. USA 85, 3338 3342.
[9] Richardson, J.S., Getzoff, E.D. and Richardson, D.C. (1978) Proc. Natl. Acad. Sci. USA 75, 2574–2578.
[10] Weininger, D. (1988) J. Chem. Inf. Comput. Sci. 28, 31–36.
[11] Sternberg, M.J.E. and Thornton, J.M. (1977) J. Mol. Biol. 110, 269–283.
[12] 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. (1977) J. Mol. Biol. 112, 535–542.
[13] Artymiuk, P., Grindley, H.M., Poirrette, A.R., Rice, D.W., Ujah, E.C. and Willett, P. (1994) J. Chem. Inf. Comp. Sci. 34, 54–62.
[14] Lindqvist, Y. (1989) J. Mol. Biol. 209, 151 165.
[15] Huber, R., Schneider, M., Mayr, I., Muller, R., Deutzmann, R., Suter, F., Zuber, H., Falk, H. and Kayser, H. (1987) J. Mol. Biol. 198, 499–513.
[16] Volz, K. and Matsumara, P. (1991) J. Biol. Chem. 266, 15511–15519.
[17] Van Duyne, G.D., Standaert, R.F., Karplus, P.A., Schreiber, S.L. and Clardy, J. (1991) Science 252, 839–841.
[18] Bocskei, Z.S., Groom, C.R., Flower, D.R., Wright, C.E., Phillips, S.E.V., Cavaggioni, A., Findlay, J.B.C. and North, A.C.T. (1992) Nature 360, 186–189.
[19] Matthews, D.A., Bolin, J.T., Burridge, J.M., Filman, D.J., Volz, K.W., Kaufman, B.T., Beddell, C.R., Champness, J.N., Stammers, D.K. and Kraut, J. (1985) J. Biol. Chem. 260, 381–391.