FEBS Letters 437 (1998) 246–250 FEBS 21003

# A structural tree for proteins containing S-like β-sheets

Alexander V. Efimov*

*Institute of Protein Research, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia*

Received 21 July 1998; received in revised form 1 September 1998

**Abstract A structural tree for proteins and domains containing S-like β-sheets has been constructed. An S-like β-sheet is taken as a starting structure in modelling or as a root structure of the tree. Larger structures are obtained by a stepwise addition of β-strands and/or α-helices to the root S-like β-sheet in accordance with a restricted set of rules inferred from known principles of protein structure. Applications of the structural tree to structure comparison, protein classification and protein folding are described.**



*Key words:* Protein classification; Protein structure modelling; Stepwise folding; Structural motif; Structural similarity

## 1. Introduction

The structural tree for a protein superfamily is a scheme that includes all the intermediate and final structures connected by lines showing possible pathways of stepwise growth of a starting structure. The structural motif having a unique overall fold that occurs in all proteins of a superfamily is taken as the starting structure in modelling or the root structure of the tree. Other α-helices and/or β-strands are added to the corresponding root structure step-by-step in accordance with a restricted set of rules inferred from known principles of protein structure. The number of possible overall folds that can be obtained from one root structural motif is limited since the rules drastically reduce the number of allowed pathways of growth of the starting and intermediate structures. Thus, the structural trees are a good tool for searching possible folding pathways and all possible protein folds as well as for structure comparison and protein classification. Previously, six structural trees have been constructed and published [1,2]. This paper describes one more structural tree for proteins containing S-like β-sheets.

Up-and-down β-sheets formed by three consecutive β-strands can be of two types, S-like and Z-like β-sheets, when viewed from the same side (e.g. from the hydrophobic core of a protein). In theory, it is rather difficult to find preference for one form of the β-sheets if they are considered as isolated flat structures. However, the preference is observed at the level of super-secondary structures of higher order that include the β-sheets or if the triple-strand β-sheets fold into three-dimensional structures themselves. A distinctive feature of such structures of a higher order is that some of them can include only S-like β-sheets and the others only Z-like β-sheets [3]. For example, a Z-like β-sheet can fold onto itself so that it forms the so-called 3β-corner. Proteins containing this structural motif have been represented in the corresponding structural tree [2]. Some protein structures containing S-like β-sheets have also been described previously [3,4]. With a growing number of protein structures now available, it is fruitful to reinvestigate the proteins containing S-like β-sheets. The structural tree presented here includes more protein structures and shows some novel pathways of growth of the root S-like β-sheet as compared to the results of our previous analysis.

## 2. Construction and analysis of the structural tree

An S-like β-sheet (when viewed from the hydrophobic core of a protein or domain) is a structural motif widespread in many proteins and domains irrespective of whether these proteins and domains are homologous or not. As a rule, an S-like β-sheet together with the flanking elements of secondary structure form different kinds of right-handed superhelices with the exception of the β-sheet flanked with two α-helices which can fold into both right- and left-handed superhelices [3]. In most proteins, an S-like β-sheet occurs as a part of higher-order structures. Midkine [5] is the only protein found yet whose structure consists of a twisted and coiled S-like β-sheet flanked with short irregular 'tails'. The chromatin binding domain from mouse modifier protein 1 [6], composed of an S-like β-sheet and a short α-helix, and the N-domain of transcription initiation factor TFIIB [7], consisting of an S-like β-sheet and a short β-hairpin, are other examples of small domains containing S-like β-sheets but having additional elements of secondary structure. Apparently, an S-like β-sheet is not stable enough to exist in an isolated form. However, it may be quite stable to be a nucleus in protein folding. As shown below, larger protein structures can be obtained by a stepwise addition of α-helices and/or β-strands to an S-like β-sheet, taking into account a restricted set of simple rules. These are the main reasons why an S-like β-sheet is taken as a starting structure in protein modelling and a root structure in constructing the structural tree.

The structural tree has been constructed taking into account a restricted set of general rules that have been derived from analysis of the structural features observed in globular proteins:

1. Overall folds of protein molecules and intermediate structures are taken into account and details of the structures are ignored. For space economy, only the pathways leading to known protein structures are represented.
2. Larger protein and intermediate structures are obtained by a stepwise addition of β-strands and α-helices to growing structure so that a structure obtained at the preceding step is maintained (it can be slightly modified). At each step, the β-strand or α-helix nearest to the growing structure along the polypeptide chain is the first to be attached to the growing structure [1,2,8].
3. The obtained structures should be compact; α-helices and β-strands should be packed in accordance with the rules that govern their close packing (see, e.g., [9,10]).

*Fax: (7) (95) 9240493. E-mail: efimov@protres.ru


PII: S0014-5793(98)01244-7

4. α-Helices and β-strands cannot be packed into one layer because of dehydration of the free NH and CO groups of the β-strands [10,11]. Thus, the next β-strand should be packed into a β-layer and the next α-helix into an α-layer or into a hydrophobic concavity (it can be formed, for example, by a strongly twisted and coiled β-sheet or between two β-sheets that splay apart) of a growing structure. The next α-helix or β-strand can also form a new α- or β-layer in the growing structure.
5. Crossing of connections [12] and formation of knots [13] are prohibited.
6. The obtained structures should have the corresponding handedness. First of all, if structural motifs are formed they should have the corresponding handedness and overall folds, for example, β-α-β-units should be in the form of the right-handed superhelix [14,15]. β-Strands that covalently link the two β-sheets at close corners and bend through 90° when passing from one β-sheet to the other should form the so-called right-handed bends [16]. So if an S-like β-sheet is flanked by two such bends, the overall fold of the obtained β+S+β-structure can be represented as a right-handed superhelix (for easier chain tracing, the S-like β-sheet may be replaced by one imaginary strand; see Fig. 1 and [3]). The S-like β-sheet is denoted here and below by the letter S; the sign '+' indicates that a flanking element of secondary structure is oriented orthogonally to the β-sheet and packed in the other layer; the sign '−' indicates an aligned packing of an adjacent element in the other layer, e.g. S-α; Sβ shows that all the β-strands are packed in one β-layer (for other examples, see Fig. 1).

Analysis shows that, similar to the β-strand forming the right-handed bend [16], the polypeptide chain bends through ~90° in the right-handed direction when passing from the S-like β-sheet to the other layer in a number of other structures, such as S+α, S+β, S+c, S+*l*, α+S, β+S, c+S and *l*+S (Fig. 1), where α, β, c and *l* are α-helical, β-structural, coiled (or irregular) and linker regions, respectively. So the overall folds of structures, α+S+α, β+S+α, β+S+β, β+S+c and c+S+c, can be represented as right-handed superhelices if the S-like β-sheet is replaced by one imaginary strand.

An S-like β-sheet can be included into right-handed superhelices formed by the βαβ- or βcβ-units [14,15]. They are denoted as S-α-β and β-α-S in Fig. 1B. Superhelix β-c-S found in aspartyl and HIV-1 proteases is similar to superhelix β-α-S but contains irregular regions c instead of the α-helix.

The structural motif denoted as the abSd-unit in Fig. 1A can be considered as a variant of the abcd-unit [1,8,11] in which region c is replaced by the S-like β-sheet. Like the abcd-unit that contains right-handed superhelix bcd, the abSd-unit has right-handed superhelix bSd which can also be denoted as superhelix β-S-β. A variant of the abSd-unit observed in the N- and C-domains of ribosomal protein L6 has a slightly different orientation of the β-sheets as compared with that of the abSd-units found in region 170–230 of cytochrome *f*, biotin carboxyl carrier protein and others (see Fig. 1A). The overall folds of region 82–132 of Umu D′ protein and DNA binding domain of HIV-1 integrase can also be referred to as the abSd-unit which has the orthogonal packing of the β-sheets. It contains right-handed superhelix β+S+β (Fig. 1C).

The 'Greek key' motif (Fig. 1A) can have both directions of the polypeptide chain and different length and conformation of the linker but the handedness of the S+*l* or *l*+S parts is always the same. The simplified depiction of this motif according to Richardson [13] results in the so-called Greek key topology with a clockwise swirl (when viewed from the outside). For comparison, such depiction of the abcd-unit (this structural motif recurs within two-layer β-proteins having the aligned β-sheet packing; see [1,8,11]) results in the Greek key topology with a counterclockwise swirl.

As mentioned above, an S-like β-sheet flanked with two α-helices can fold into both left-handed and right-handed super-

→

Fig. 1. A structural tree for proteins and domains containing S-like β-sheets. A, B, C and D: Different branches of the tree. β-Strands are shown as arrows directed from N- to C-ends, α-helices as cylinders, connection regions between elements of secondary structure located in one layer as single lines and those between elements located in different layers as double lines. Long loops are simplified and drawn by dashed lines. Thick lines between structures represent possible pathways of their growth. Abbreviations and references (PDB codes or original papers) are as follows: MBP: mannose-binding protein (1RTM); DNA Pol III: β-subunit of DNA polymerase III (2POL); Capping enzyme: mRNA capping enzyme (1CKN); S17: ribosomal protein S17 (1RIP); GSH: glutathione synthetase (1GLT); T7 DNA ligase: DNA ligase from bacteriophage T7 [17]; DNA topoisomerase I (1VCC); dehydrase (1MKB); glutathione reductase (3GRS); Con A: concanavalin A (2CNA); gelatinase: C-domain of gelatinase A (1GEN); $G_A$: β-propeller domain of $G_A$ protein (1TBG); neuraminidase (1KIT); GNA: *Galanthus nivalis* agglutinin (1MSA); S-lectin (1SLT); CLC protein: Charcot-Leyden crystal protein (1LCL); hCRP: human C-reactive protein (1GNH); arcelin-5 (1IOA); BGLL: 1,3-1,4-β-glucanase from *Bacillus licheniformis* (1GBG); EGI: endoglucanase I (1EG1); isolectin B4 [18]; E2o domain: E2o lipoyl domain [19]; BCCP: biotin carboxyl carrier protein (1BDO); cytochrome *f* (1CTM); III Glc: phosphocarrier protein III Glc (1F3G); H-Pro: H-protein (1HTP); L6: ribosomal protein L6 [20]; HLA-A2: human leukocyte antigen (1HLA); T4 Lysozyme: bacteriophage T4 lysozyme (2O6L); LALF: *Limulus* anti-LPS factor [21]; β-lactamase: TEM1 β-lactamase (1BTL); dsRBD: dsRNA binding domain [22]; Rec A: Rec A protein (1AA3); Ada C: Ada-C protein (1SFE); hexokinase: yeast hexokinase (1HKG); Rnase H: ribonuclease H (2RN2); ASV integrase: avian sarcoma virus integrase (2ASV); Glyox I: glyoxalase I (1FRO); DHBD: biphenyl-cleaving extradiol dioxygenase (1HAN); SRP 9/14: signal recognition particle [23]; KF: Klenow fragment of DNA polymerase (1DPI); EPNP: purine nucleoside phosphorylase (1ECP); ARF-1: ADP-ribosylation factor-1 (1RRG); Pol β: DNA polymerase β (1RPL); creatinase: creatine amidinohydrolase (1CHM); HIV-1 Nef (1EFN); pilin [24]; Rm lipase (1TIA); HIV-1 aspartyl protease (2HVP); aspartyl protease: penicillopepsin (3APP); ricin A chain (2AAI); SH2: SH2 domain (1LCJ); HAP1: human apurinic/apyrimidinic endonuclease [25]; H-subunit: H-subunit of the photosynthetic reaction center (1PRC); ε-subunit: ε-subunit of ATP synthase [26]; Umu D′: Umu D′ protein (1UMU); IN-DBD: DNA-binding domain of HIV-1 integrase [27]; papain (9PAP); head-binding domain (1LKT); SH-3: SH-3 domain (1SHG); Psa E: photosystem I protein (1PSE); g3p-D1: minor coat protein g3p (1FGP); PDZ domain (1PDR); TFIIB: transcription initiation factor TFIIB (1PFT); F1-G pair (1TPG); PH: pleckstrin homology domain (2DYN); MK: midkine [5]; RNase $T_1$: ribonuclease $T_1$ (1RNT); Rnase St: ribonuclease St (0RST); Rnase $C_2$: ribonuclease $C_2$ [28]; Sac Y: antiterminator protein Sac Y (1AUU); MoMOD1-N: mouse chromatin modifier protein 1 [6]; Cyt c peroxidase: cytochrome *c* peroxidase (2CYP); IL-8: interleukin 8 (3IL8); PF-4: platelet factor 4 (1PLF); HP1 integrase: bacteriophage HP1 integrase (1AIH); λ integrase (1AE9); Cre recombinase [29]; Xer D recombinase [30]; CMTF: aspartate carbomoyltransferase (8ATC); neurophysin: bovine neurophysin II (1BN2); Sac 7d: DNA-binding protein Sac 7d (1SAP); Sso 7d: DNA-binding protein Sso 7d (1SSO); PDF: peptide deformylase (1DEE); OB-fold [31]; PT: pertussis toxin (1PRT); Lys RSTT: lysyl-tRNA synthetase [32]; RPA: replication protein A (1JMC); $SSB_c$: single-stranded DNA-binding protein (1KAW); G5BP: gene 5 DNA-binding protein (2GN5); CspB: cold-shock protein B (1CSP); PNP: polynucleotide phosphorylase (1SRO); TSST-1: toxic shock syndrome toxin-1 (2QIL); PPase, inorganic pyrophosphatase (1JFD).

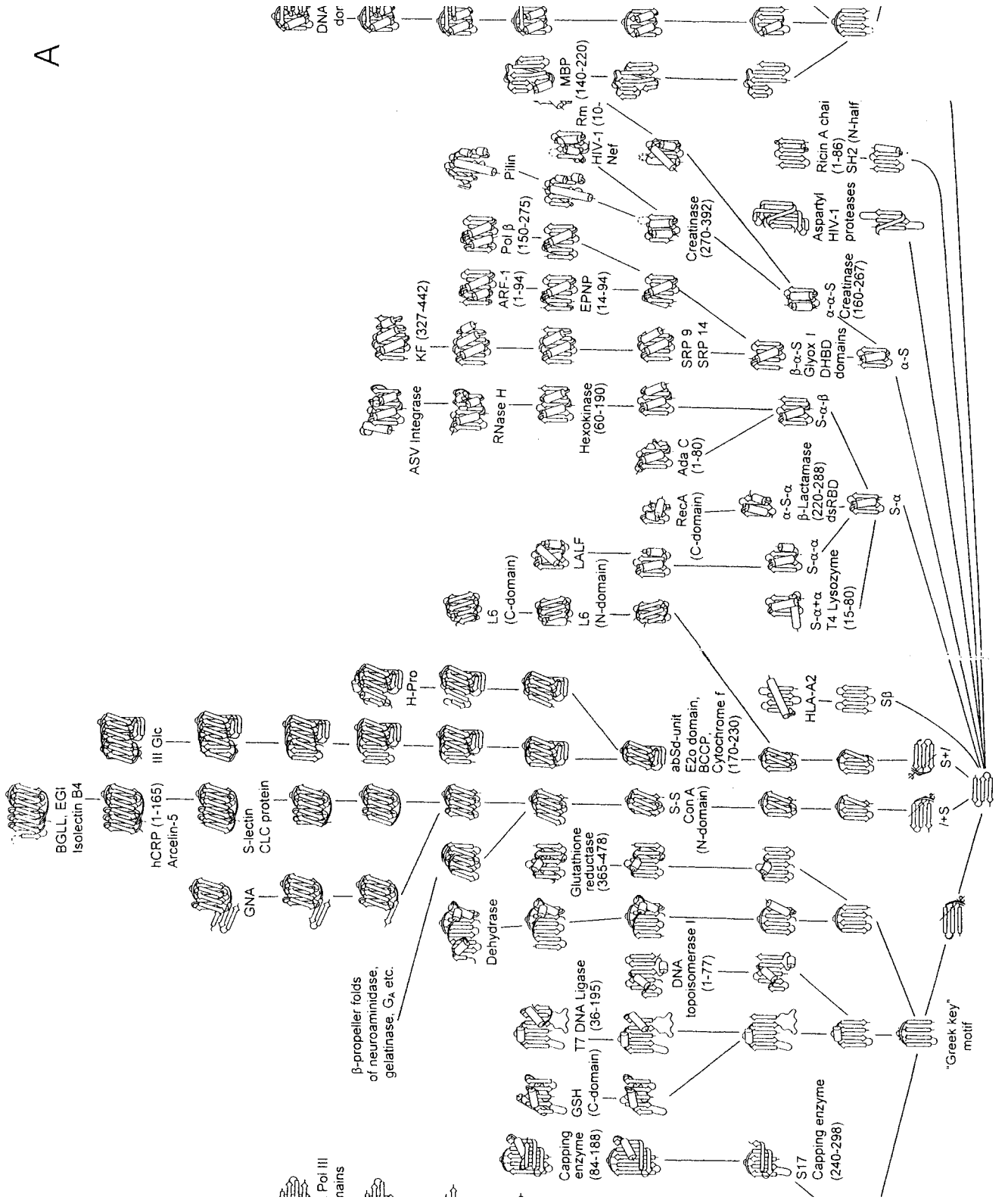


Fig. 1.

helices. A right-handed superhelix is formed when α-helices are packed approximately perpendicular to the S-like β-sheet (see structures containing superhelix α+S+α in Fig. 1D). A left-handed superhelix is formed in the case of an aligned packing of the α-helices onto the β-sheet (see structures containing superhelix α-S-α in Fig. 1B).

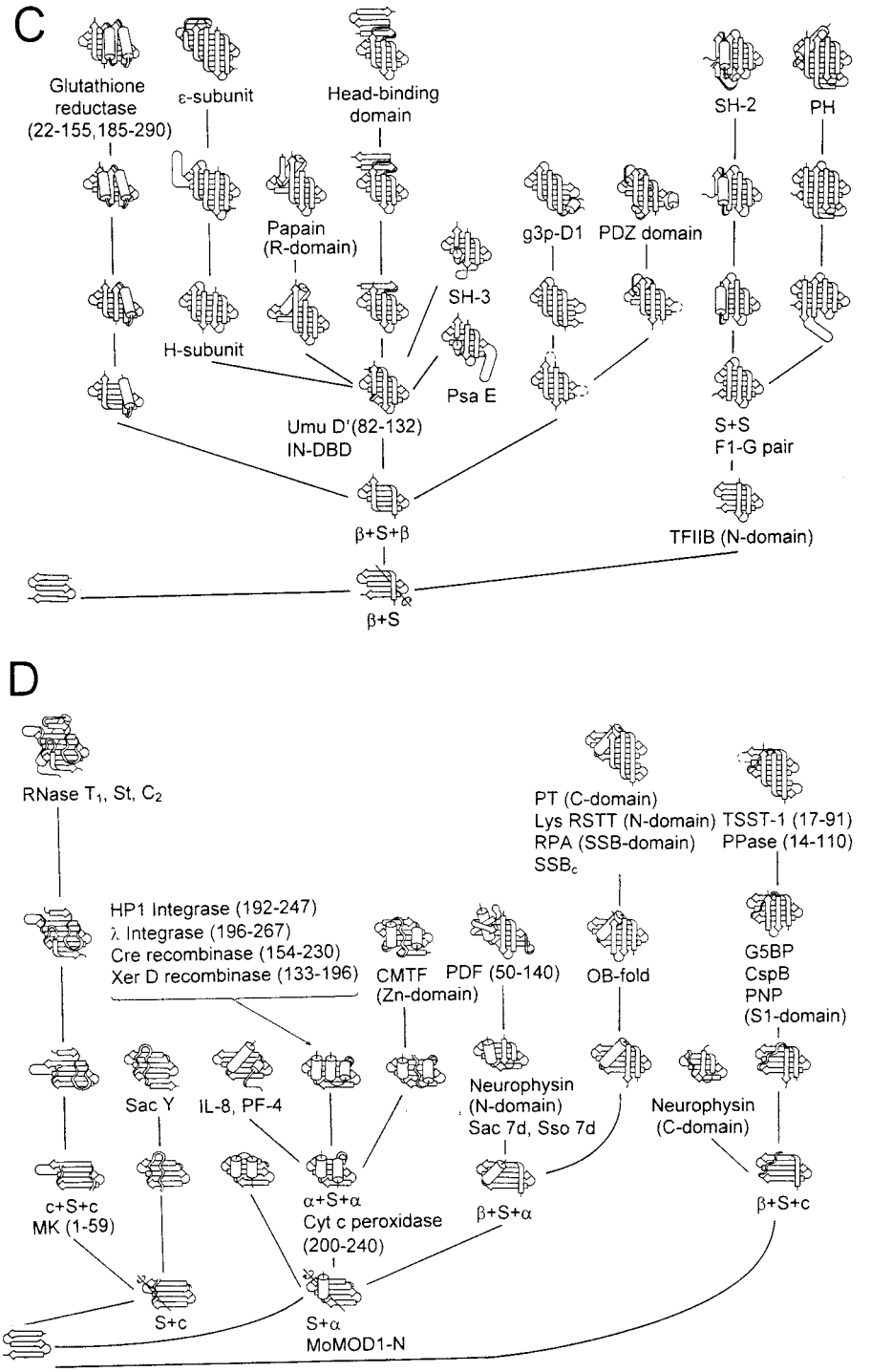


Fig. 1 (*continued*).

Fig. 1 represents a structural tree for proteins containing S-like β-sheets constructed in accordance with the rules. As can be seen, an addition of a β-strand, an α-helix or a c-region to the root S-like β-sheet at the first step can be done in different ways and results in formation of the structures shown in the bottom row of the tree. The next row represents structures obtained at the second step etc. All the pathways of stepwise growth of the root β-sheet that lead to known protein structures are shown with thick lines. Since the root β-sheet and many intermediate structures can grow in two or more ways, there are several branches in the structural tree.

## 3. Application of the structural tree to structure comparison, protein classification and protein folding

Levels of structural similarity between different proteins and domains can easily be observed by visual inspection of the tree. Within one branch, structures having a higher position in the tree include the structures located lower. Proteins and domains of different branches have a common fold located in the branching point. The higher the branching point is located in the tree, the higher the level of structural sim-

ilarity between proteins and domains of the corresponding branches is observed.

Although all the proteins and domains shown in the structural tree contain an S-like β-sheet, this level of structural similarity between them is hardly enough to group them into one structural class. However, proteins and domains found within branches of the tree, for example, those within relatively large branches, can be grouped into structural families. The above mentioned superhelices and other structural motifs commonly occurring within branches may be taken as structural determinants of such families. Thus, it is possible to recognize structural families whose proteins and domains contain 'Greek key' motifs (Fig. 1A), S-S-structures (Fig. 1A), abSd-units (Fig. 1A,B), α-S-α-superhelices, S-α-β-superhelices, β-α-S-superhelices, α-α-S- and S-α-α-structures (Fig. 1B), β+S+β-superhelices, S+S-structures (Fig. 1C), c+S+c-superhelices, α+S+α-superhelices, β+S+α-superhelices, β+S+c-superhelices (Fig. 1D). It should be noted that proteins and domains containing superhelices, β+S+α and β+S+c, have very similar overall folds and can be considered as one structural family. Proteins containing structures, α-α-S and S-α-α, may be grouped into one family if the polypeptide chain directions are ignored. The same may be done for proteins containing superhelices, S-α-β and β-α-S.

As seen, this classification is different from those suggested by other authors [33–36] in some aspects. First of all, amino acid sequences and functions of proteins are not taken into account in this classification. It is primarily based on similarity of overall folds of proteins and domains and modelled pathways of stepwise growth of the structural motif or, in other words, on the structural trees of proteins (for other structural trees and their application to structural classification of proteins, see [1,2]). This approach permits the structural classification of both homologous and non-homologous proteins and offers a stimulating perspective regarding their folding pathways.

Analysis of all the data presented above as well as results on other structural trees [1,2] have led us to a hypothesis that root structural motifs can be formed at first steps of protein folding and can act as nuclei and the pathways of their stepwise growth can be considered as possible folding pathways of the proteins. However, it should be noted that the order of events in protein folding may be different from that presented in the trees. One of the reasons is that, at some steps of modelling, it is difficult to determine what segment is the first to be attached to the growing structure. The other reason is that, in some growing structures, there can be two sites of growth and the corresponding parts of such proteins can fold independently of each other and simultaneously (which is not possible to show in the tree) or one after another. Examples of such cases have been described previously [1]. It is quite interesting to note that different structural motifs can coexist in the same protein molecule or domain. So all the protein structures but one (glutathione reductase) shown in Fig. 1C contain both an S-like β-sheet and a 3β-corner. In theory, all these structures can be obtained by a stepwise addition of β-strands and α-helices to the S-like β-sheet (Fig. 1C) or to the 3β-corner [2]. This results in some uncertainty in protein classification but may be important for reliable protein folding in nature. This seems to be consistent with the main conclusion of Viguera et al. [37] that different folding transition states may result in the same native structure.

*Acknowledgements:* This work was supported in part by the Russian Foundation for Basic Research (RFFI Grant no. 98-04-48252).

## References

[1] Efimov, A.V. (1997) Proteins 28, 241–260.
[2] Efimov, A.V. (1997) FEBS Lett. 407, 37–41.
[3] Efimov, A.V. (1993) FEBS Lett. 334, 253–256.
[4] Efimov, A.V. (1994) Structure 2, 999–1002.
[5] Iwasaki, W., Nagata, K., Hatanaka, H., Inui, T., Kimura, T., Muramatsu, T., Yoshida, K., Tasumi, M. and Inagaki, F. (1997) EMBO J. 16, 6936–6946.
[6] Ball, L.J., Murzina, N.V., Broadhurst, R.W., Raine, A.R.C., Archer, S.J., Stott, F.J., Murzin, A.G., Singh, P.B., Domaille, P.J. and Laue, E.D. (1997) EMBO J. 16, 2473–2481.
[7] Zhu, W., Zeng, Q., Colangelo, C.M., Lewis, L.M., Summers, M.F. and Scott, R.A. (1996) Nat. Struct. Biol. 3, 122–124.
[8] Efimov, A.V. (1982) Mol. Biol. (Moscow) 16, 799–806.
[9] Chothia, C., Levitt, M. and Richardson, D. (1977) Proc. Natl. Acad. Sci. USA 74, 4130–4134.
[10] Efimov, A.V. (1977) Dokl. Akad. Nauk SSSR 235, 699–702.
[11] Efimov, A.V. (1995) J. Mol. Biol. 245, 402–415.
[12] Lim, V.I., Mazanov, A.L. and Efimov, A.V. (1978) Mol. Biol. (Moscow) 12, 206–213.
[13] Richardson, J.S. (1977) Nature 268, 495–500.
[14] Rao, S.T. and Rossmann, M.G. (1973) J. Mol. Biol. 76, 241–256.
[15] Sternberg, M.J.E. and Thornton, J.M. (1976) J. Mol. Biol. 105, 367–382.
[16] Chothia, C. and Janin, J. (1982) Biochemistry 21, 3955–3965.
[17] Subramanya, H.S., Doherty, A.J., Ashford, S.R. and Wigley, D.B. (1996) Cell 85, 607–615.
[18] Osinaga, E., Tello, D., Batthyany, C., Bianchet, M., Tavares, G., Durán, R., Cervenansky, C., Camoin, L., Roseto, A. and Alzari, P.M. (1997) FEBS Lett. 412, 190–196.
[19] Ricaud, P.M., Howard, M.J., Roberts, E.L., Broadhurst, R.W. and Perham, R.N. (1996) J. Mol. Biol. 264, 179–190.
[20] Golden, B.L., Ramakrishnan, V. and White, S.W. (1993) EMBO J. 12, 4901–4908.
[21] Hoess, A., Watson, S., Siber, G.R. and Liddington, R. (1993) EMBO J. 12, 3351–3356.
[22] Kharrat, A., Macias, M.J., Gibson, T.J., Nilges, M. and Pastore, A. (1995) EMBO J. 14, 3572–3584.
[23] Birse, D.E.A., Kapp, U., Strub, K., Cusack, S. and Åberg, A. (1997) EMBO J. 16, 3757–3766.
[24] Parge, H.E., Forest, K.T., Hickey, M.J., Christensen, D.A., Getzoff, E.D. and Tainer, J.A. (1995) Nature 378, 32–38.
[25] Gorman, M.A., Morera, S., Rothwell, D.G., de La Fortelle, E., Mol, C.D., Tainer, J.A., Hickson, I.D. and Freemont, P.S. (1997) EMBO J. 16, 6548–6558.
[26] Wilkens, S., Dahlquist, F.W., McIntosh, L.P., Donaldson, L.W. and Capaldi, R.A. (1995) Nat. Struct. Biol. 2, 961–967.
[27] Eijkelenboom, A.P.A.M., Lutzke, R.A.P., Boelens, R., Plasterk, R.H.A., Kaptein, R. and Hård, K. (1995) Nat. Struct. Biol. 2, 807–810.
[28] Polyakov, K.M., Vagin, A.A., Tishchenko, G.N., Bezborodova, S.I. and Vainshtein, B.K. (1983) Dokl. Akad. Nauk SSSR 273, 1383–1386.
[29] Guo, F., Gopaul, D.N. and Van Duyne, G.D. (1997) Nature 389, 40–46.
[30] Subramanya, H.S., Arciszewska, L.K., Baker, R.A., Bird, L.E., Sherratt, D.J. and Wigley, D.B. (1997) EMBO J. 16, 5178–5187.
[31] Murzin, A.G. (1993) EMBO J. 12, 861–867.
[32] Cusack, S., Yaremchuk, A. and Tukalo, M. (1996) EMBO J. 15, 6321–6334.
[33] Rufino, C.D. and Blundell, T.L. (1994) J. Comput. Aided Mol. Design 8, 5–27.
[34] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) J. Mol. Biol. 247, 536–540.
[35] Holm, L. and Sander, C. (1996) Science 273, 595–602.
[36] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) Structure 5, 1093–1108.
[37] Viguera, A.R., Serrano, L. and Wilmanns, M. (1996) Nat. Struct. Biol. 3, 874–880.