Journal of Biomolecular Structure and Dynamics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tbsd20

Modeling of folds and folding pathways for some protein families of (α+β)- and (α/β)-classes

Alexey B. Gordeev [a] & Alexander V. Efimov [a]
[a] Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russian Federation
Version of record first published: 16 Jul 2012.

To cite this article: Alexey B. Gordeev & Alexander V. Efimov (2013): Modeling of folds and folding pathways for some protein families of (α+β)- and (α/β)-classes, Journal of Biomolecular Structure and Dynamics, 31:1, 4-16

To link to this article: http://dx.doi.org/10.1080/07391102.2012.691341

Journal of Biomolecular Structure and Dynamics
Vol. 31, No. 1, January 2013, 4–16

Modeling of folds and folding pathways for some protein families of (α + β)- and (α/β)-classes

Alexey B. Gordeev and Alexander V. Efimov*

Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region 142290, Russian Federation

Communicated by Vsevolod J. Makeev

(*Received 16 February 2011; final version received 3 February 2012*)

In this paper, updated structural trees for α/β-proteins containing five- and seven-segment (α/β)-motifs are represented. Novel structural motifs occurring in some families of (α + β)- and (α/β)-proteins are also characterized. Databases of these proteins have been compiled from the Protein Data Bank (PDB) and Structural Classification of Proteins (SCOP) and the corresponding structural trees have been constructed. The classification of these proteins has been developed and organized as an extension of the PCBOST database, which is available at http://strees.protres.ru. In total, the updated Protein Classification Based on Structural Trees database contains 11 structural trees, 106 levels, 635 folds, 4911 proteins and domains, and 14,202 PDB entries.

Keywords: structural motif; handedness; structural tree; classification

Introduction

The problem of protein folding remains one of the central problems in biochemistry and molecular biology to be solved. However, until now there are no experimental methods to observe protein folding in real time. Thus, theoretical modeling of protein folds and folding pathways is of particular value in solving the problem.

Our approach is based on the hypothesis that at the first step of protein folding a nucleus is formed and then the remaining part of the molecule or domain is folded around it (Levinthal, 1968; Ptitsyn, 1973; Wetlaufer, 1973). The structural motif having a unique overall fold and handedness is taken as the starting structure in modeling. The larger protein structures are obtained by stepwise addition of α-helices and/or β-strands to the growing structure taking into account a restricted set of rules inferred from the known principles of protein structure. A general scheme that represents the starting structural motif (root motif), all the intermediate and completed structures connected by lines showing allowed pathways of structure growth, is referred to as the structural tree (Efimov, 1995, 1996, 1997a, 1997b, 1998). A similar approach was used by Ptitsyn et al. (Ptitsyn, 1981; Ptitsyn, Finkelstein, & Falk (Bendzko), 1979; Ptitsyn & Rashin, 1975) in modeling of all β- and α-proteins. The main difference between their folding schemes and our structural trees concerns the starting structure. For example, Ptitsyn et al. used β- and α-hairpins as the starting structures which can exist in two forms, right turned or left turned, i.e. they do not have unique overall folds. In contrast, the root motifs of the structural trees have unique overall folds themselves and determine the place where the remaining part of the protein molecule is attached. On the other hand, in modeling, we take into account an updated set of general rules many of which were elaborated in our works (Efimov, 1995, 1996, 1997a, 1997b, 1998).

The first versions of structural trees were constructed more than a decade ago (Efimov, 1995, 1996, 1997a, 1997b, 1998). The increasing number of protein structures in the Protein Data Bank (PDB) (Bernstein et al., 1977) has prompted us to construct updated structural trees and their computer versions (Gordeev & Efimov, 2009; Gordeev, Kondratova, & Efimov, 2008). On the other hand, some novel structural motifs were found in proteins, which were used as root structures for construction of novel structural trees (Efimov, 2008; Kargatov & Efimov, 2010). Based on the structural trees, we have developed a novel Structural Classification of Proteins (SCOP) referred to as Protein Classification Based on Structural Trees (PCBOST) (Gordeev, Kargatov, & Efimov, 2010), which is

*Corresponding author. Email: efimov@protres.ru

ISSN 0739-1102 print/ISSN 1538-0254 online

http://dx.doi.org/10.1080/07391102.2012.691341
http://www.tandfonline.com

available at http://strees.protres.ru/. This classification is based primarily on the similarity of spatial structures and common folding pathways simulated with the trees, thereby differing from other known protein classifications like Structural Classification of Proteins (SCOP) (Murzin, Brenner, Hubbard, & Chothia, 1995), Class-Architecture-Topology-Homologous superfamily (CATH) (Orengo et al., 1997), and others (Dietmann et al., 2001; Sowdhamini, Rufino, & Blundell, 1996; Przytycka, Aurora, & Rose, 1999). Our classification disregards the amino acid sequences, functions, and evolutionary relationships of proteins which are taken into account in other known classifications.

In this paper, we present updated structural trees for (α/β)-proteins containing five- and seven-segment (α/β)-motifs. Several novel structural motifs commonly occurring in some families of (α + β)- and (α/β)-proteins are also characterized and corresponding structural trees are constructed. The classification of these proteins has been developed and organized as an extension of the PCBOST database.

Materials and methods

Databases for all the structural groups of proteins were compiled using the PDB (Bernstein et al., 1977) and the SCOP database http://scop.mrc-lmb.cam.ac.uk/scop/ (Murzin et al., 1995). Proteins were manually selected using the RasMol molecular graphics program (Sayle & Milner-White, 1995). The secondary structure assignment was done with the RasMol and Define Secondary Structure of Proteins (DSSP) (Kabsch & Sander, 1983). Possible homologies were revealed by Blastp (protein–protein BLAST) http://blast.ncbi.nlm.nih.gov/Blast.cgi (Tatusova & Madden, 1999).

Structural trees were constructed using a restricted set of known rules (Efimov, 1995, 1996, 1997a, 1997b, 1998). Among them, the following rules are the most important:

(1) Overall folds of protein molecules and intermediate structures are taken into account and details of the structures are ignored. Each structure in the trees for (α/β)-proteins can have both directions of the polypeptide chain.
(2) 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 it.
(3) 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; thus, an α-helix should be packed into the α-helical layer and a β-strand into the β-layer of a growing structure.
(4) The obtained structures should be compact; α-helices and β-strands should be packed in accordance with the rules that govern their close packing (Chothia, Levitt, & Richardson, 1977; Efimov, 1977, 1979).
(5) Crossing of connections (Lim, Mazanov, & Efimov, 1978) and formation of knots (Richardson, 1977) are prohibited.
(6) All the obtained structures should have the corresponding handedness. For example, all the β-α-β-units should have the form of a right-handed superhelix (Rao & Rossmann, 1973; Sternberg & Thornton, 1976).

Results and discussion

Construction and analysis of an updated structural tree for (α/β)-proteins containing five-segment (α/β)-motifs

The five-segment (α/β)-motif is a structural motif consisting of three β-strands and two α-helices folded

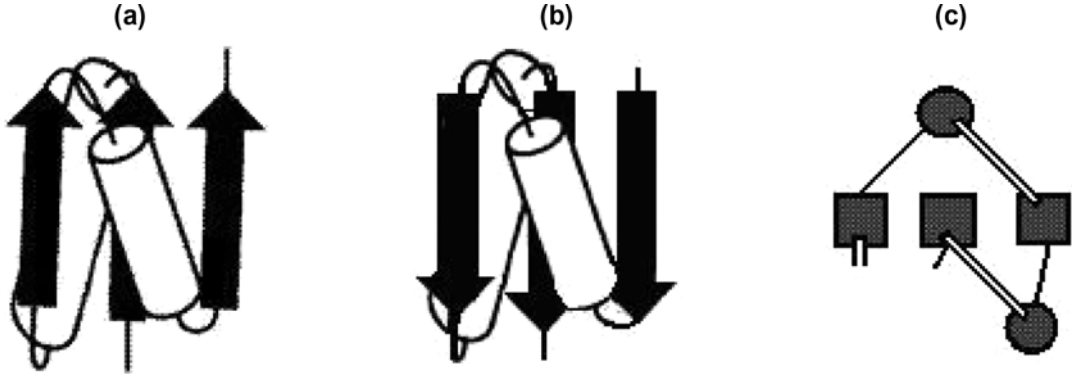

Figure 1. Schematic representation of a five-segment (α/β)-motif with a direct (a) and reverse (b) chain orientation. β strands are shown with arrows directed from the N to the C end. α helices are shown as cylinders. (c) The five-segment (α/β)-motif as viewed end-on. β strands are shown with squares; α helices are shown with circles and near and far linkers are shown with double and single lines, respectively.

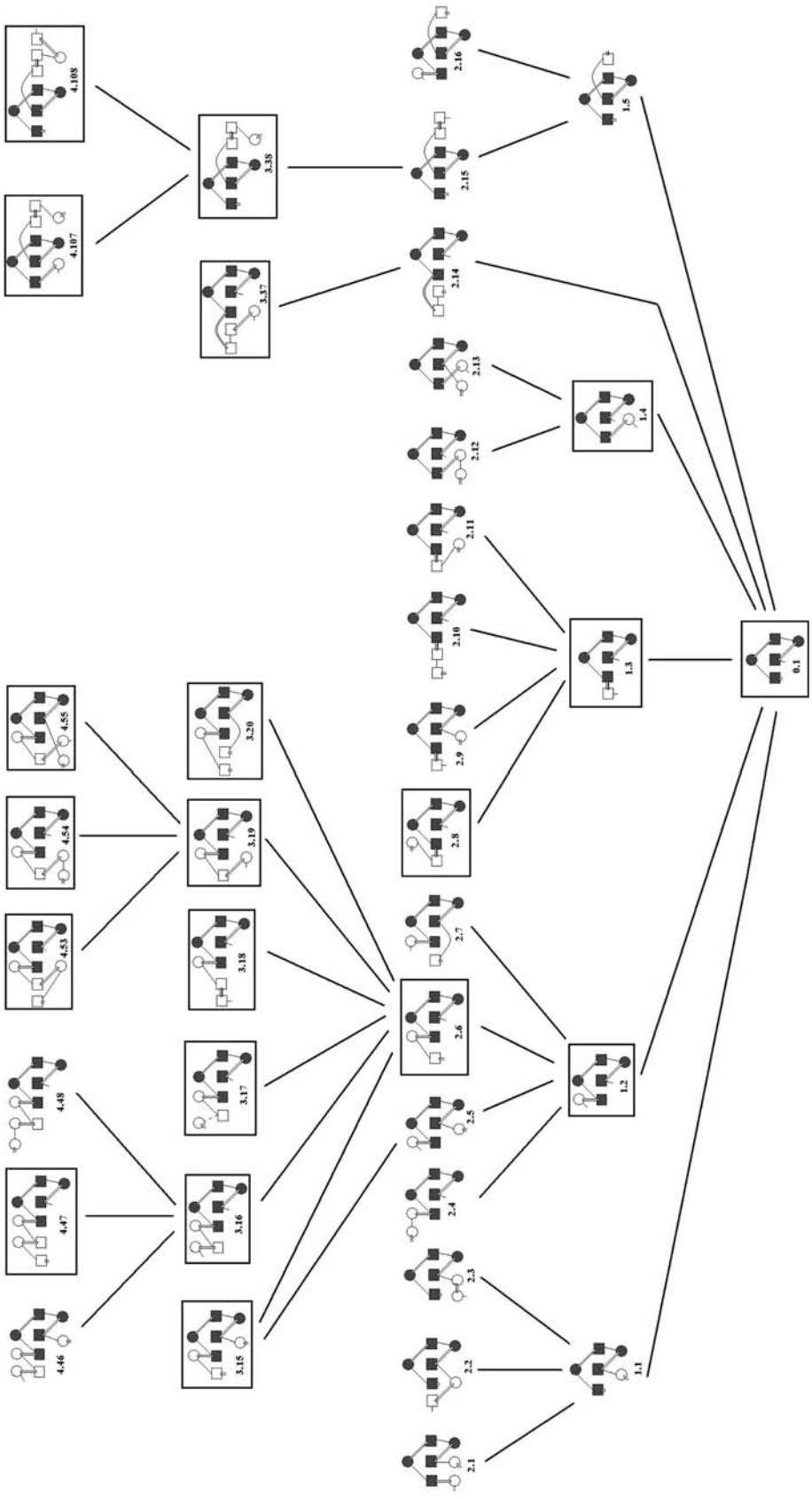

Figure 2. A fragment of the structural tree of (α/β)-proteins containing a five-segment (α/β)-motif. All structures are similarly oriented and are viewed end-on as in Figure 1(c). The folds actually found in proteins are framed.

into two βαβ-units and arranged into a three-layer structure having a parallel β-sheet and the α-helices packed on its both sides (Figure 1). It is widespread in (α/β)-proteins. We have found 1301 proteins and domains containing the five-segment (α/β)-motifs in the PDB (total 3926 PDB entries) and among them 388 are nonhomologous. Taking into account this database, we have constructed an updated version of the structural tree which is available at http://strees.protres.ru. A fragment of the structural tree is shown in Figure 2. The complete structural tree involves 494 folds of which 172 occur in known proteins. For comparison, the first version of the structural tree for proteins of this class included 39 folds (Efimov, 1997a).

Analysis of the updated tree shows that the five-segment (α/β)-motifs having a direct chain orientation (Figure 1(a)) occur more often than those with the reverse orientation (Figure 1(b)). Among 401 five-segment motifs found in 388 nonhomologous proteins, 323 have a direct (81%) and 78 (19%) the reverse orientation of the polypeptide chain. The reasons for different frequencies of occurrence of the versions are still poorly understood and will be further investigated. It should be noted that we have found only 11 of 388 nonhomologous proteins which contain stereochemically unfavorable left-handed βαβ-units (1KZY, 1WF6, 1NNS, 2C42, 1VL0, 1KBZ, 1E6U, 1QV9, and 2FRI, 1JU3, 1EA5) in which the polypeptide chain folds into the left-handed superhelix.

Construction and analysis of an updated structural tree for (α/β)-proteins containing seven-segment (α/β)-motifs

The seven-segment (α/β)-motif is a structural motif composed of four β-strands and three α-helices that are folded into βαβαβ- and αβ-units and arranged so that the β-strands form a parallel β-sheet and the α-helices are packed on both sides of the β-sheet (Figure 3). A database of proteins containing this motif has been compiled that includes 870 proteins and domains (total 2567 PDB-entries) of which 294 are nonhomologous. Taking into account the new database, we have constructed an updated structural tree that includes 310 folds of which 85 occur in known proteins (a fragment of the tree is shown in Figure 4). The left-handed βαβ-units have been found in 14 proteins (of 294 nonhomologous): 1P5H, 1I24, 1XGK, 1YDE, 2BLL, 1XKG, 1DHR, 1ZBQ, 1YO6, 1JTV, 2BKA, 1Y1P, 1QYC, and 2ARO. Similar to the five-segment (α/β)-motifs, the most seven-segment (α/β)-motifs have a direct orientation of the polypeptide chain (Figure 3(a)) and only 14 of 294 nonhomologous motifs have the reverse chain orientation.

Novel structural motifs and structural trees for (α/β)-proteins

An inspection of (α/β)-proteins enables us to find and characterize two novel commonly occurring structures referred to as a nine-segment (α/β)-motif (Figure 5(a) and (c)) and the analog of the seven-segment (α/β)-motif having an S-like β-sheet instead of the α-helix in the upper layer (Figure 5(b) and (d)). As seen the nine-segment (α/β)-motif is similar to the five- and seven-segment (α/β)-motifs. It is of interest to note that there are 388 nonhomologous proteins and domains containing a five-segment (α/β)-motif, 294 proteins and domains containing the seven-segment (α/β)-motif and only 34 proteins containing the nine-segment (α/β)-motif found in PDB. One more analogous structure in this row, the 11-segment (α/β)-motif is not found at all. The structural tree for the proteins containing nine-segment (α/β)-motifs has been constructed and is presented in Figure 6. Figure 7 shows

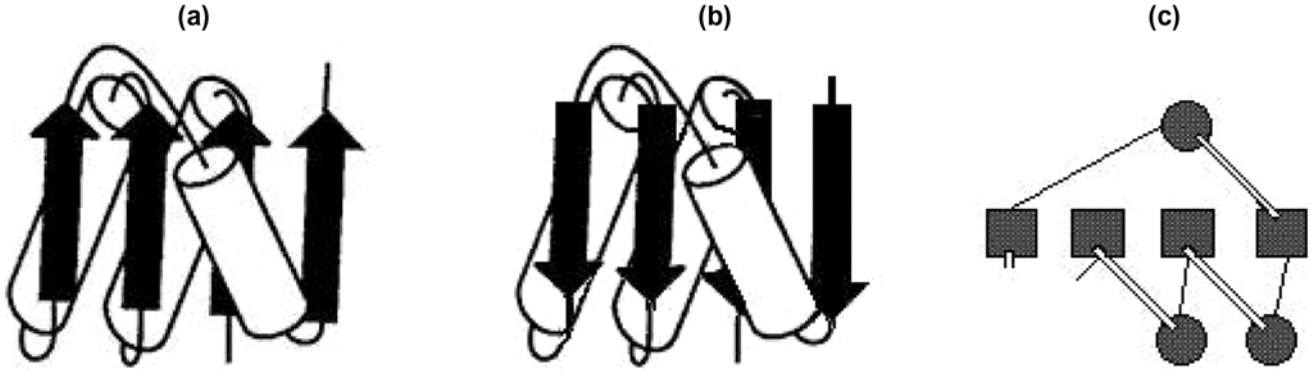

Figure 3. Schematic representation of a seven-segment (α/β)-motif with a direct (a) and reverse (b) chain orientation. (c) The seven-segment (α/β)-motif as viewed end-on.

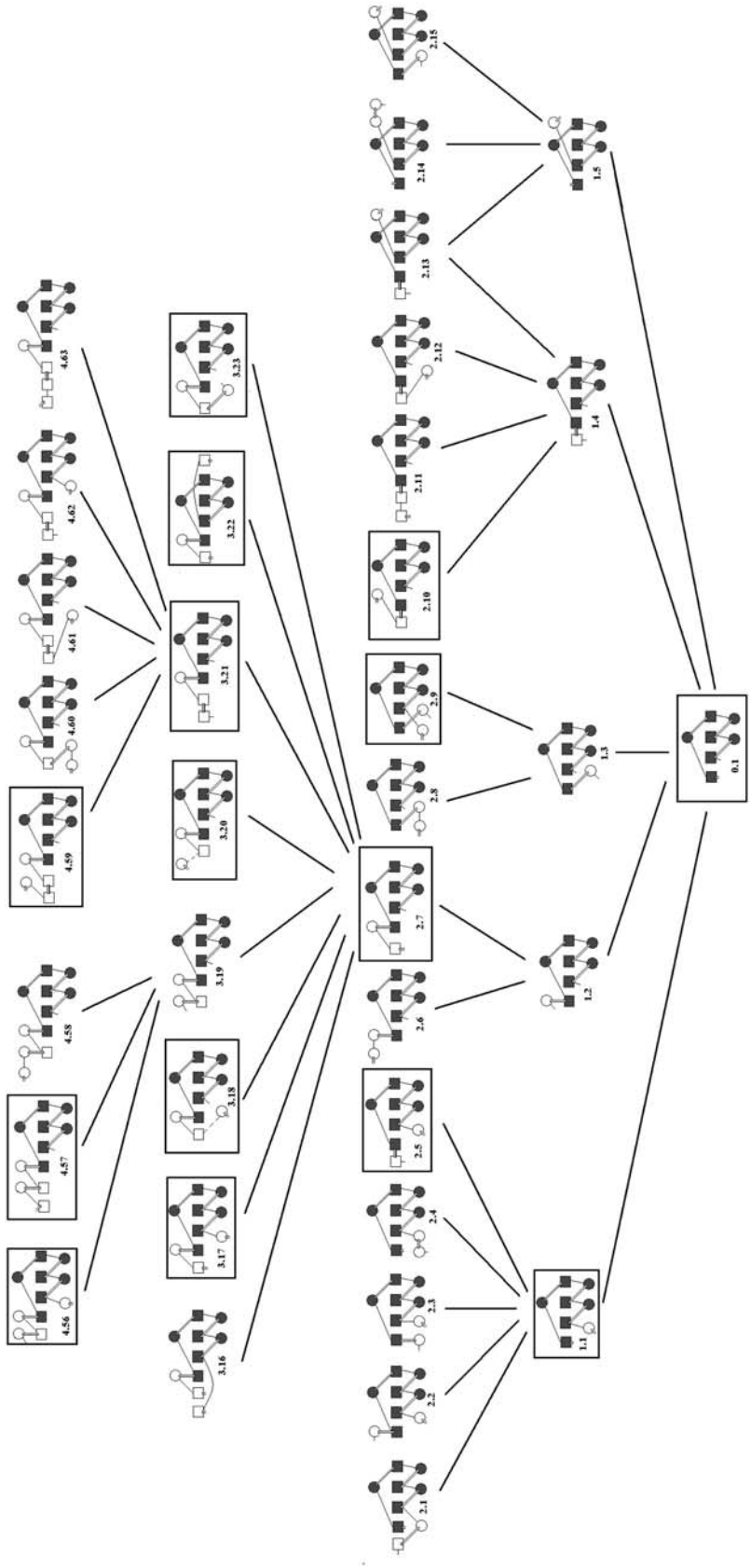

Figure 4. A fragment of the structural tree for (α/β)-proteins containing seven-segment (α/β)-motifs.

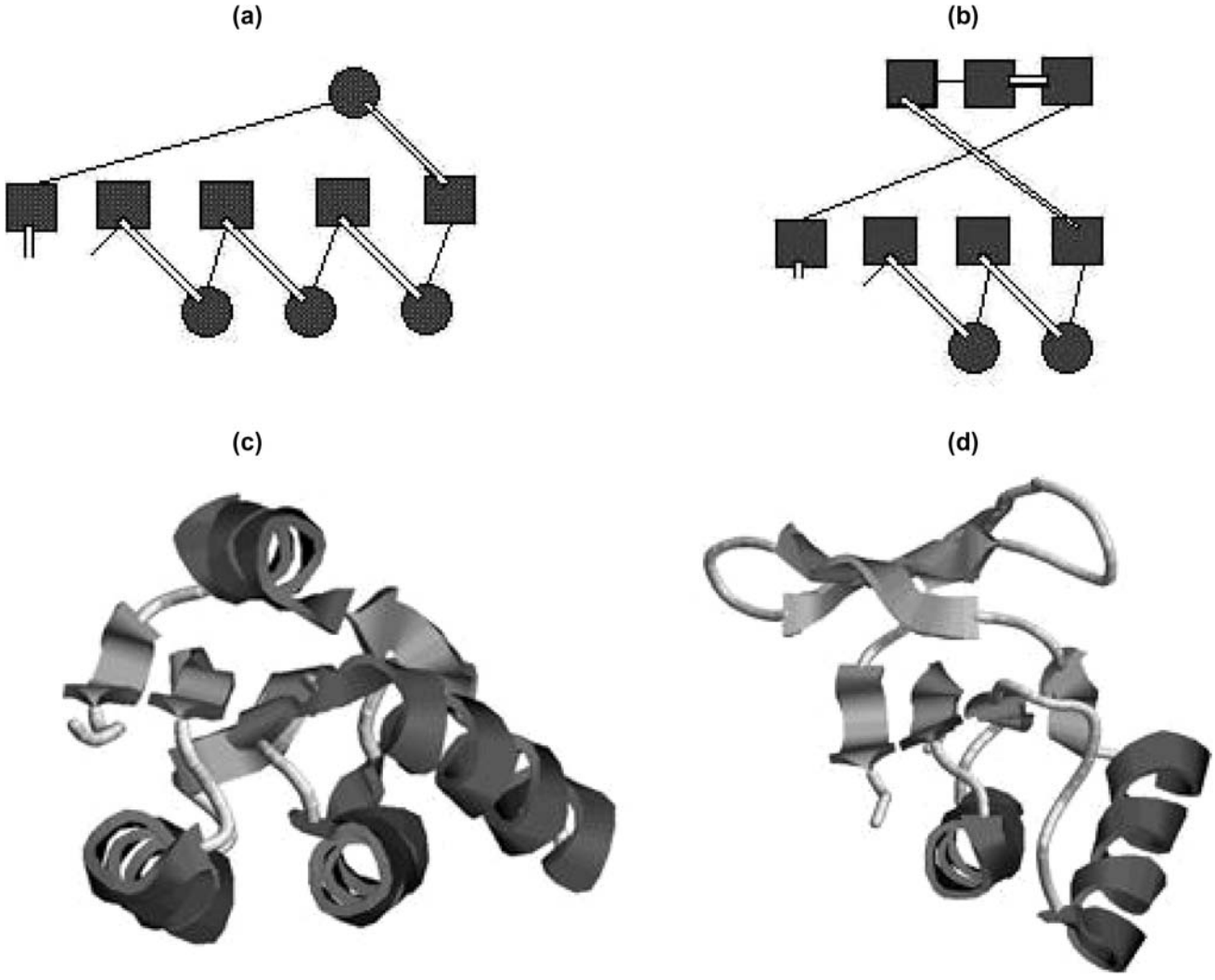

Figure 5. Schematic representation of a nine-segment (α/β)-motif (a) and the analog of the seven-segment (α/β)-motif containing an S-like β-sheet (b) viewed end-on. Representative structures of the nine-segment (c) and the analog of seven-segment (d) (α/β)-motifs are taken from 1Q0Q and 1TRB, respectively.

the structural tree for proteins containing the analog of the seven-segment (α/β)-motif with the S-like β-sheet. It involves 90 known proteins and domains.

Novel structural motifs and trees for (α+β)-proteins

In this section, some novel structural motifs analogous to the abCd-unit (Efimov, 1995, 1997a) are described. Previously, we have compiled a database of 926 (α+β)-proteins containing the abCd-unit and have constructed an updated structural tree for this group of proteins (Gordeev & Efimov, 2009). However, in (α+β)-proteins there are a number of other commonly occurring folding units which are quite similar to the abCd-unit. Their structures can be obtained by stepwise addition of β-strands to the abCd-unit in the β-sheet between its a- and d-strands as schematically shown in Figure 8. The additional β-strands joined to strand a of the abCd-unit are labeled here as a_1, a_2, … and the β-strands joined to strand d as d_1, d_2, … according to their distance from strands a or d in the polypeptide chain irrespective of the chain direction. Frequencies of occurrence of these structural motifs in known proteins and folds are presented in Table 1. In total, the database includes 485 (α+β)-proteins and domains (among them 199 nonhomologous) containing structural motifs analogous to the abCd-unit. As seen, the more additional β-strands the structural motif includes, the lower its frequency of occurrence in proteins. Structural trees for (α+β)-proteins containing a1abCd- and a2a1abCd-structures which occur most often are represented in Figures 9, 10a and 10b.

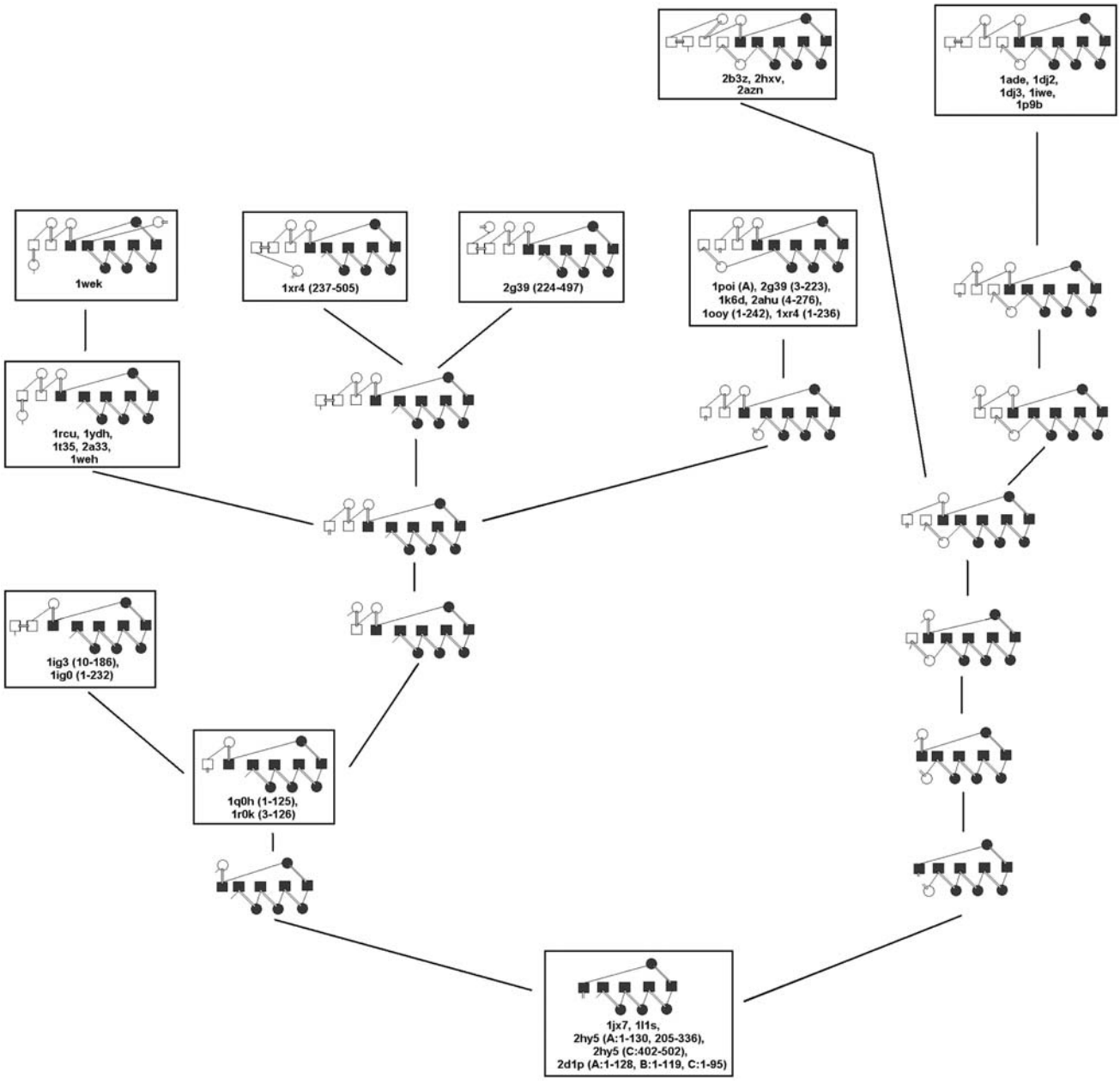

Figure 6. The structural tree of (α/β)-proteins containing nine-segment (α/β)-motifs. The folds that occur in known proteins and domains are framed and labeled with corresponding PDB-codes.

Structural classification of proteins based on structural trees

Previously, we had presented the classification of some structural groups of proteins based on structural trees (Gordeev & Efimov, 2009; Gordeev et al., 2008, 2010). Now the classification of the proteins described above was organized as an extension of the PCBOST database (Gordeev et al., 2010) and is available at http://strees.protres.ru. PCBOST is a WEB resource containing a hierarchically organized database of protein structures, constructed structural trees, a guide page to facilitate working with the database, a system for retrieving a protein of interest by its PDB ID, and the PDB files of proteins. It includes several hierarchical levels. Proteins and domains having the same root structural motif are combined into the STRUCTURAL TREE. Proteins and domains located at horizontal levels of the structural tree are grouped

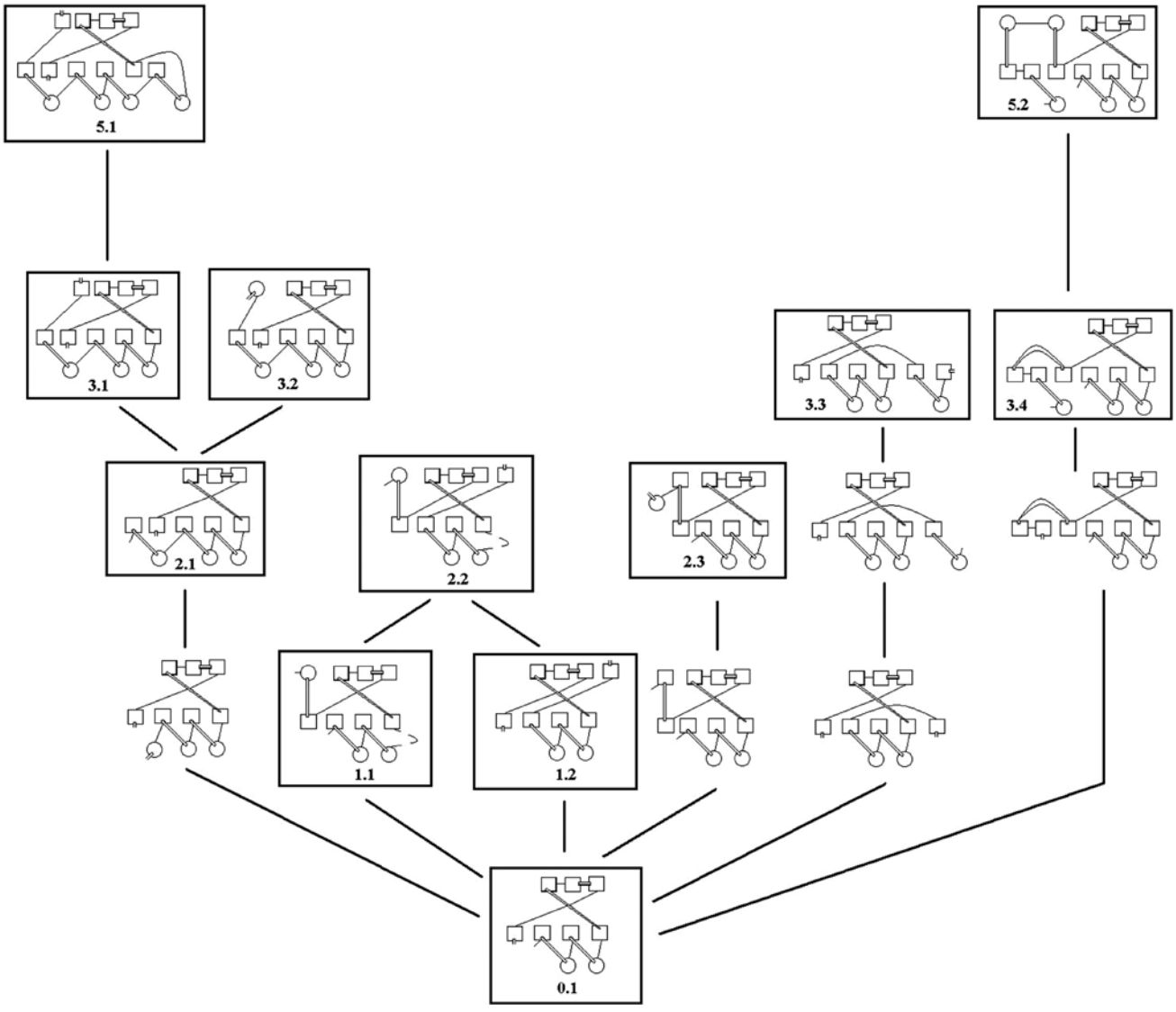

Figure 7. The structural tree of (α/β)-proteins containing the analog of a seven-segment motif containing an S-like β sheet. The folds found in known proteins are framed.

into LEVELS. All proteins and domains from one LEVEL having the same arrangement of secondary structure elements form FOLDS. Now, the PCBOST database consists of 11 structural trees, 106 levels, 635 folds, 4911 proteins and domains, and 14,202 PDB-files.

Discussion

In this paper, we present several novel structural trees for globular proteins. Some novel supersecondary structures that can be used as root structures of structural trees have been described. Respective structural trees have been constructed. Updated versions of the trees for two other well-known structural motifs have also been constructed. In our opinion, the structural trees are good tool for the searching of all possible folds of the polypeptide chain, for the modeling of folding pathways of proteins and their structures, for protein structure comparison. They can be used for a structural classification of proteins that should help us to analyze the information about their structures and use it for research.

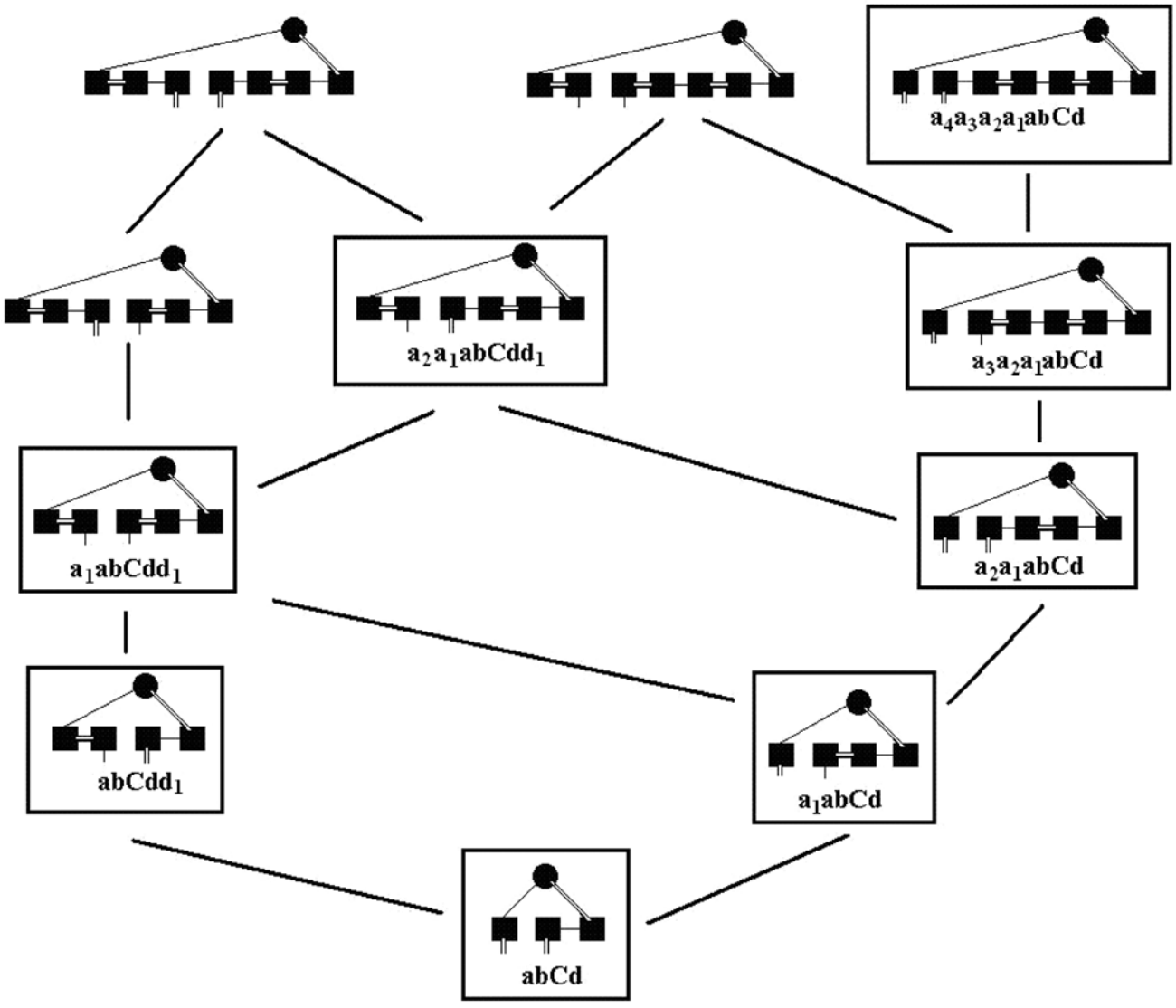

Figure 8. Modeling of the structures analogous to the abCd unit. The folds actually found in proteins are framed.

Table 1. Frequencies of occurrence of structural motifs analogous to the abCd unit in known proteins.

Name of the structure	Number of different chain folds	Total number of proteins	Number of nonhomologous proteins
$abCdd_1$	3	7	5
a_1abCdd_1	1	16	2
a_1abCd	39	283	104
a_2a_1abCd	24	84	52
$a_2a_1abCdd_1$	4	63	23
$a_3a_2a_1abCd$	1	2	1
$a_4a_3a_2a_1abCd$	4	30	12

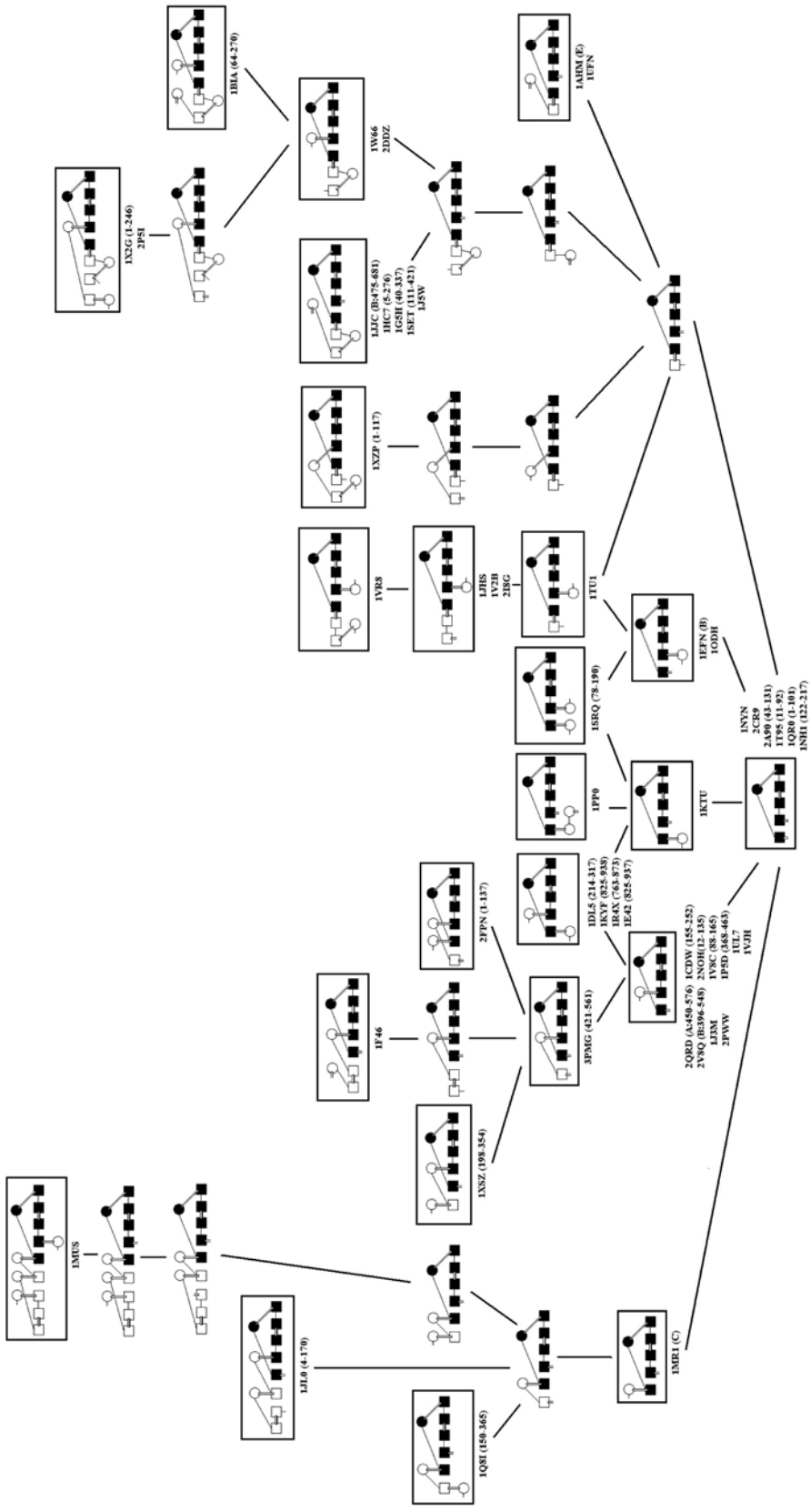

Figure 9. The structural tree for (α + β)-proteins containing $a_2a_1$abCd structures. The folds actually found in proteins are framed. PDB codes of only nonhomologous proteins are shown.

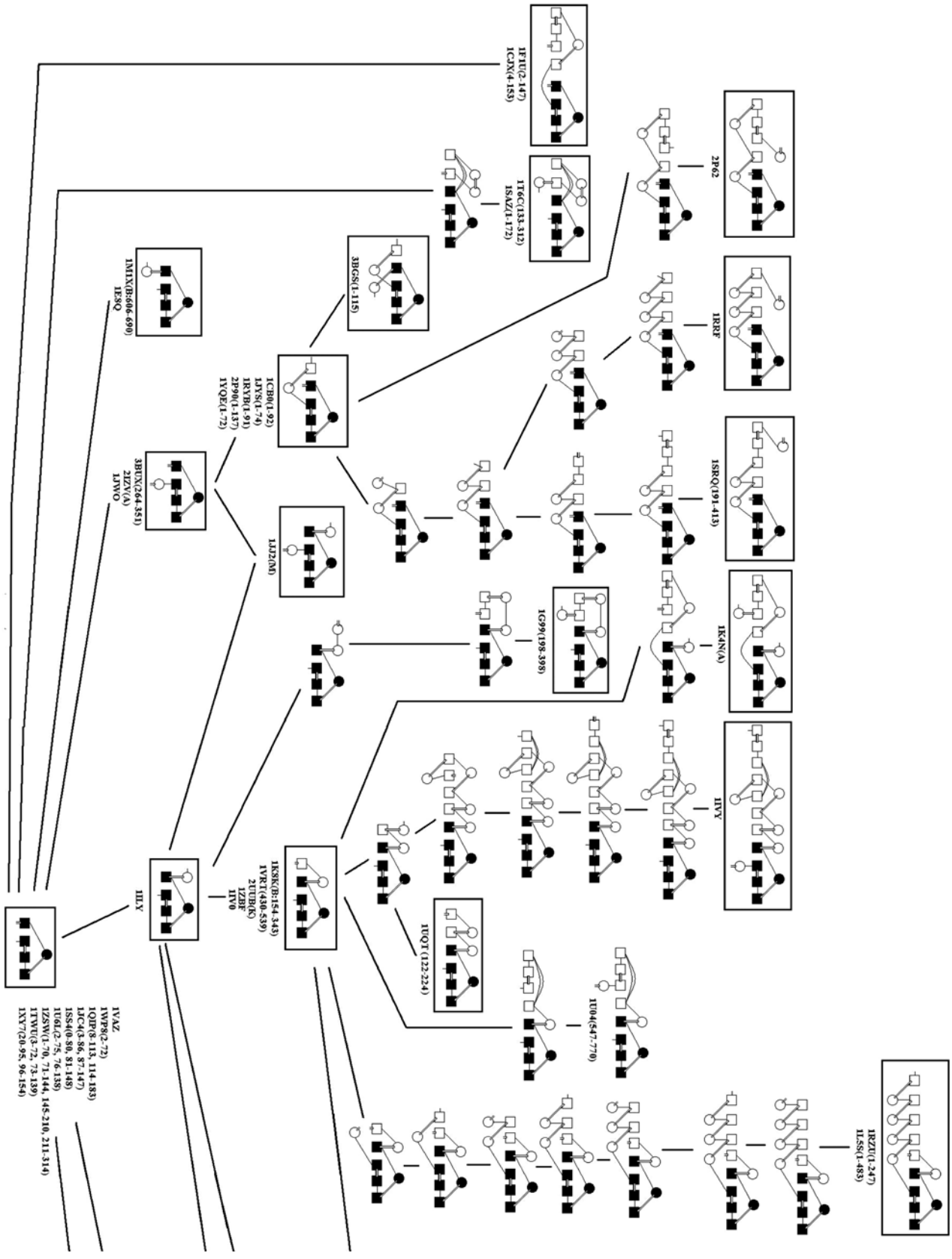

Figure 10a. The structural tree of (α + β)-proteins containing $a_1$abCd structures. The folds actually found in proteins are framed. PDB-codes of only non-homologous proteins are shown.

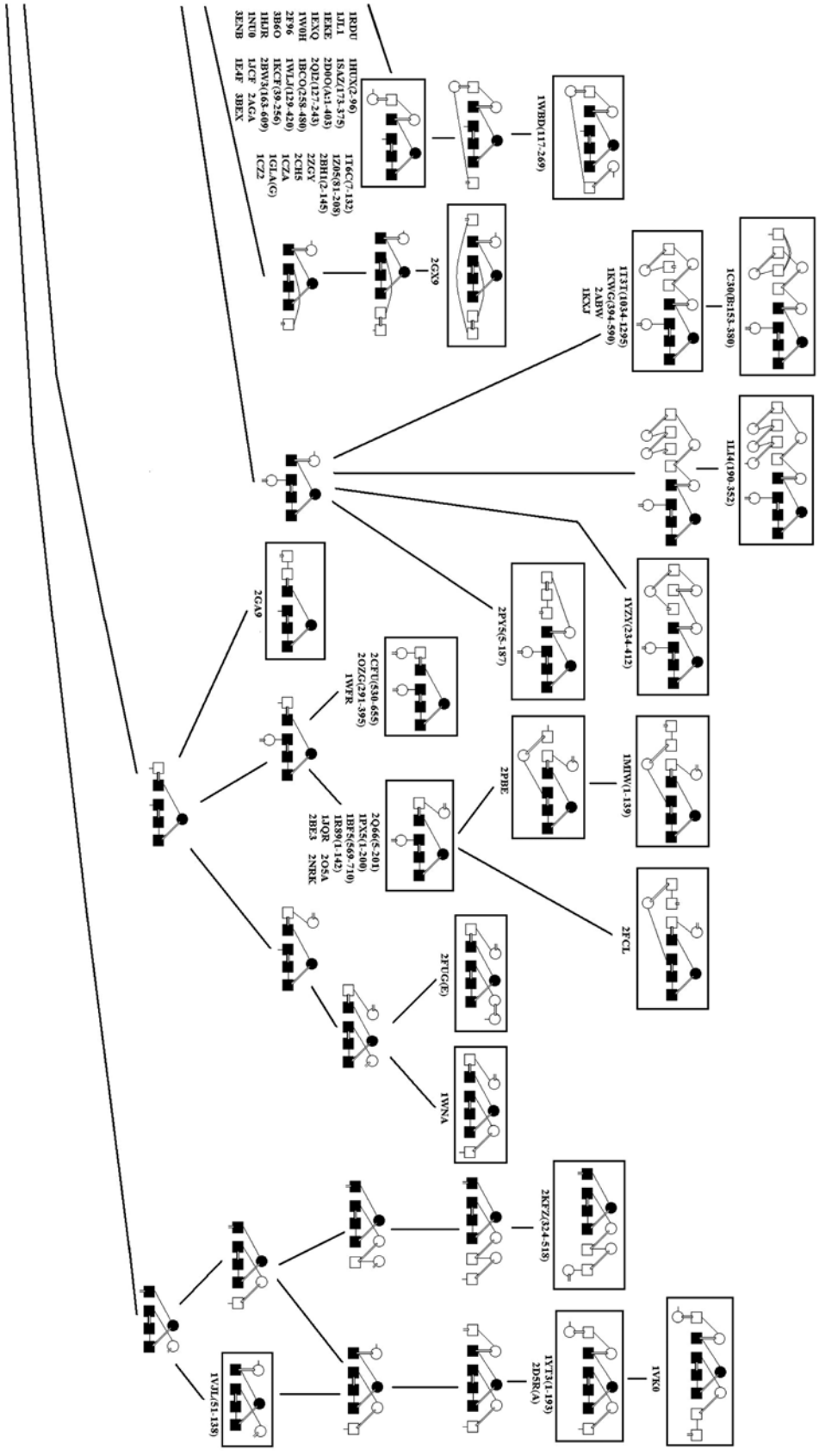

Figure 10b. The right part of the structural tree is shown in Figure 10a.

Acknowledgments

This work was supported by the Russian Foundation for Basic Research (Project No. 10-04-00727).

References

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., … Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology, 112*, 535–542.

Chothia, C., Levitt, M., & Richardson, D. (1977). Structure of proteins: Packing of alpha-helices and pleated sheets. *Proceedings of the National Academy of Sciences USA, 74*, 4130–4134.

Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M., & Holm, L. (2001). A fully automatic evolutionary classification of protein folds: DALI domain dictionary version 3. *Nucleic Acids Research, 29*, 55–57.

Efimov, A.V. (1977). Stereochemistry of the packing of alpha-spirals and beta-structure into a compact globule. *Doklady Akademii Nauk SSSR, 235*, 699–702.

Efimov, A.V. (1979). Packing of alpha-helices in globular proteins. Layer-structure of globin hydrophobic cores. *Journal of Molecular Biology, 134*, 23–40.

Efimov, A.V. (1995). Structural similarity between two-layer alpha/beta and beta-proteins. *Journal of Molecular Biology, 245*, 402–415.

Efimov, A.V. (1996). A structural tree for alpha-helical proteins containing alpha-alpha-corners and its application to protein classification. *FEBS Letters, 391*, 167–170.

Efimov, A.V. (1997a). Structural trees for protein superfamilies. *Proteins, 28*, 241–260.

Efimov, A.V. (1997b). A structural tree for proteins containing 3beta-corners. *FEBS Letters, 407*, 37–41.

Efimov, A.V. (1998). A structural tree for proteins containing S-like beta-sheets. *FEBS Letters, 437*, 246–250.

Efimov, A.V. (2008). Structural trees for proteins containing phi-motifs. *Biochemistry (Moscow), 73*, 23–28.

Gordeev, A.B., & Efimov, A.V. (2009). Novel structural tree for (alpha + beta)-proteins containing abCd-units. *Molecular Biology (Moscow), 43*, 480–484.

Gordeev, A.B., Kargatov, A.M., & Efimov, A.V. (2010). PCBOST: Protein classification based on structural trees. *Biochemical and Biophysical Research Communications, 397*, 470–471.

Gordeev, A.B., Kondratova, M.S., & Efimov, A.V. (2008). Novel structural tree for beta-proteins containing abcd-units. *Molecular Biology (Moscow), 42*, 285–288.

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

Kargatov, A.M., & Efimov, A.V. (2010). A novel structural motif and structural trees for proteins containing it. *Biochemistry (Moscow), 75*, 249–256.

Levinthal, C. (1968). Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique, 65*, 44–45.

Lim, V.I., Mazanov, A.L., & Efimov, A.V. (1978). Stereochemical theory of the 3-dimensional structure of globular proteins. I. Highly helical intermediate structures. *Molecular Biology (Moscow), 12*, 206–213.

Murzin, A.G., Brenner, S.E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology, 247*, 536–540.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., & Thornton, J.M. (1997). CATH – a hierarchic classification of protein domain structures. *Structure, 5*, 1093–1108.

Przytycka, T., Aurora, R., & Rose, G.D. (1999). A protein taxonomy based on secondary structure. *Nature Structural Biology, 6*, 673–682.

Ptitsyn, O.B. (1973). Stages in the mechanism of self-organization of protein molecules. *Doklady Akademii Nauk SSSR, 210*, 1213–1215.

Ptitsyn, O.B. (1981). Protein folding: General physical model. *FEBS Letters, 131*, 197–202.

Ptitsyn, O.B., Finkelstein, A.V., & Falk (Bendzko), P. (1979). Principal folding pathway and topology of all-β-proteins. *FEBS Letters, 101*, 1–5.

Ptitsyn, O.B., & Rashin, A.A. (1975). A model of myoglobin self-organization. *Biophysical Chemistry, 3*, 1–20.

Rao, S.T., & Rossmann, M.G. (1973). Comparison of super-secondary structures in proteins. *Journal of Molecular Biology, 76*, 241–256.

Richardson, J.S. (1977). Ƀ-Sheet topology and relatedness of proteins. *Nature, 268*, 495–500.

Sayle, R.A., & Milner-White, E.J. (1995). RasMol: Biomolecular graphics for all. *Trends in Biochemical Sciences, 20*, 374–376.

Sowdhamini, R., Rufino, S.D., & Blundell, T.L. (1996). A database of globular protein structural domains: Clustering of representative family members into similar folds. *Folding & Design, 1*, 209–220.

Sternberg, M.J.E., & Thornton, J.M. (1976). On the conformation of proteins: The handedness of the β-strand-α-helix-β-strand unit. *Journal of Molecular Biology, 105*, 357–382.

Tatusova, T.A., & Madden, T.L. (1999). Blast 2 sequences – a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters, 174*, 247–250.

Wetlaufer, D.B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proceedings of the National Academy of Sciences USA, 70*, 697–701.