ISSN 0026-8933, Molecular Biology, 2009, Vol. 43, No. 3, pp. 480–484. © Pleiades Publishing, Inc., 2009.
Original Russian Text © A.B. Gordeev, A.V. Efimov, 2009, published in Molekulyarnaya Biologiya, 2009, Vol. 43, No. 3, pp. 521–526.

MATHEMATICAL AND SYSTEM BIOLOGY

UDC 577.150.2

Novel Structural Tree of (α+ β)-Proteins Containing abCd Units

A. B. Gordeev and A. V. Efimov

Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia;
e-mail: efimov@protres.ru

Received July 15, 2008
Accepted for publication September 30, 2008

Abstract—A database of 926 (α + β)-proteins and (α + β)-domains containing abCd units, including 401 nonhomologous, was compiled from the Protein Data Bank (total of 2636 PDB entries). A novel structural tree of this structural class of proteins was constructed to include 286 possible polypeptide chain folds. A structural classification of (α + β)-proteins containing abCd unit was developed on the basis of the structural tree. The database and the structural tree are available at http://strees.protres.ru/.

DOI: 10.1134/S0026893309030169

Key words: modeling, classification, database, protein folding, structural motif, structural similarity

INTRODUCTION

A structural tree of proteins is a set of all allowable intermediate and final 3D structures that can be derived from one root (starting) structure by consecutively adding other secondary structure elements. Possible ways of structural complication are shown with lines, which eventually integrate all structures in one tree. Secondary structure elements are added to complicating structures according to a set of rules inferred from the known principles of protein structural organization. A structural motif with a unique polypeptide chain fold is used as the root of a tree.

The earliest versions of structural trees were constructed more than a decade ago [1–4]. In 2008, a new supersecondary structure, φ-motif, was revealed in proteins, and a structural tree was constructed for proteins containing this motif [5]. The number of solved protein structures in the Protein Data Bank (PDB) has substantially increased over this decade. Hence, it is necessary to construct updated trees with all proteins available from PDB for the given class. We have constructed and analyzed a structural tree of β-proteins containing abcd units. Based on the updated tree, we compiled a hierarchic database of all β-proteins with abcd units. The database includes 528 proteins and domains and is available at http://strees.protres.ru/ [6].

Substantially larger sets of protein structures are also available for other structural trees. For instance, a structural tree constructed in 1997 [2] included approximately 50 known structures of (α + β)-proteins containing abCd units, while 926 proteins and domains have been compiled in a database for this class to date, suggesting the construction of an updated structural tree. Construction and analysis of a new tree makes it possible to study new pathways of structural complication, to search for new polypeptide folds, to identify new regularities in protein structures, and, eventually, to update the rules for structural tree construction. These problems were the focus of our work.

Structural trees provide a convenient and promising tool for solving many problems, e.g., to search for all possible (both known and still unknown) polypeptide folds in a compact spatial structure, to simulate the pathways of protein folding, to study the folding mechanism, to analyze the structural similarity of proteins, etc. An important application of structural trees is the development of a structural classification of proteins. Such a classification is based exclusively on the similarity of spatial structures and simulated folding pathways, thereby differing from other common protein classification systems [7–9], which utilize, to a certain extent, the data on the amino acid sequence homologies, functions, and evolutionary relationships of proteins (for a review, see [10, 11]).

SUBJECTS AND METHODS

A database of (α + β)-proteins containing abCd units was compiled using the Structural Classification of Proteins (SCOP) database, v. 1.73 (http://scop.mrclmb.cam.ac.uk/scop/). Proteins were manually selected. In total, we retrieved 926 proteins and domains containing abCd units, including 401 nonhomologous. Possible homologies were revealed by BLAST pairwise alignment (http://www.ncbi.nih.gov/BLAST/) [12]. Protein structures were visually

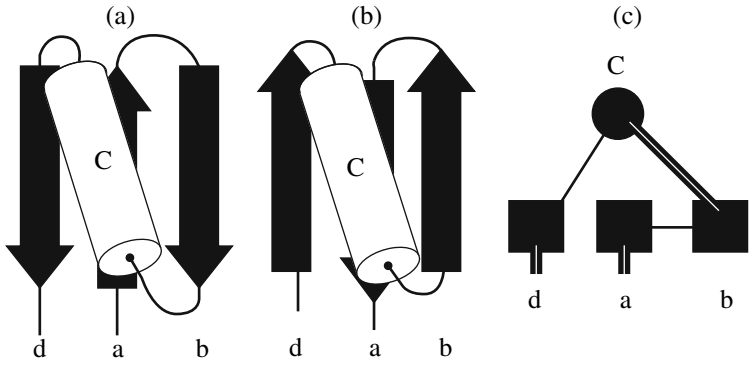

Fig. 1. Schemes of the abCd unit with a (a) direct or (b) reverse chain orientation. The β-strands are shown with arrows directed from the N to the C end. The α-helix is shown as a cylinder. (c) The abCd unit as viewed from the face. The β-strands are shown with squares; the α helix is shown with a circle; and the linkers directed toward and away from the viewer are shown with double and single lines, respectively. The β-strands and α-helix are designated with the corresponding letters.

examined using the RasMol molecular graphics program [13].

A structural tree was constructed according to known rules [1–4].

(1) As a root, we used an abCd unit, which is known to have a unique fold [14]. The abCd unit is a variant of the abcd unit, characteristic of β-proteins, and contains α-helix C in place of β-strand c. As in the case of abcd unit, strands b and d and helix C form a right-handed superhelix bCd. However, the structural similarity between (α + β)-proteins of this class and β-proteins containing abcd units is not restricted to this feature. The abCd units often occur at the margins of layer structures in (α + β)-proteins, as in β-proteins. Many proteins and domains of the two classes have fundamentally the same fold apart from the conformation of secondary structure elements. The abCd unit occurs in two variants, with a direct or a reverse orientation of the polypeptide chain (Figs. 1a, 1b). However, approximately 77% of the abCd units found in nonhomologous proteins have an oppositely oriented fold (see below). Like other structures, the abCd unit is simplified to omit the chain orientation (Fig. 1c) in tree construction, assuming that both direct and reverse orientations are possible in every case.

(2) Other α-helices and/or β-strands are added to the root and intermediate structures in a sequential mode, step by step, so that each previous structure was preserved in the next one. At each step, the secondary structure element closest to the growing structure in the polypeptide chain was the first to be added [2, 14].

(3) A structure is to be compact according to the principle of tight packing.

(4) Linker intercrosses [15] and knots [16] are prohibited.

(5) Each structural motif (not only those that are roots) has its characteristic chirality and polypeptide chain fold.

In (α + β)-proteins, all β-α-β units form a right-handed superhelix [14, 17].

RESULTS AND DISCUSSION

A fragment of the resulting structural tree is shown in Fig. 2. The complete tree of proteins containing abCd units is available at http://strees.protres.ru/. The designations used in the structural tree are the same as in the previous tree of β-proteins containing abcd units [6]. Several levels (rows) are distinguishable in the tree. Each level (row) comprises the possible polypeptide chain folds that have the same number of secondary structure elements (β-strands and/or α-helices). Secondary structure elements are shown in a simplified form: β-strands are shown with squares, α-helices are shown with circles, linkers directed toward the viewer are shown with double lines, and linkers directed away from the viewer are shown with single lines. All folds are numbered in the tree, and each fold has a specific identifier consisting of two numbers separated with a point. The first number shows the number of secondary structure elements added to the abCd unit. For instance, first-level folds contain one additional secondary structure element (a β-strand or an α-helix) and are designated with 1, second-level folds are designated with 2, etc. The second number is the ordinal number of the fold in the given row; folds are numbered from left to right. For instance, the third row includes 57 folds, designated from 3.1 to 3.57.

On the other hand, the structural tree has several branches. In each branch, a higher-level fold incorporates the folds of the lower levels. The folds from different branches incorporate the same fold that occurs at the branching point. The higher the branching point on the structural tree, the higher the structural similarity between the proteins and domains belonging to the corresponding branches.

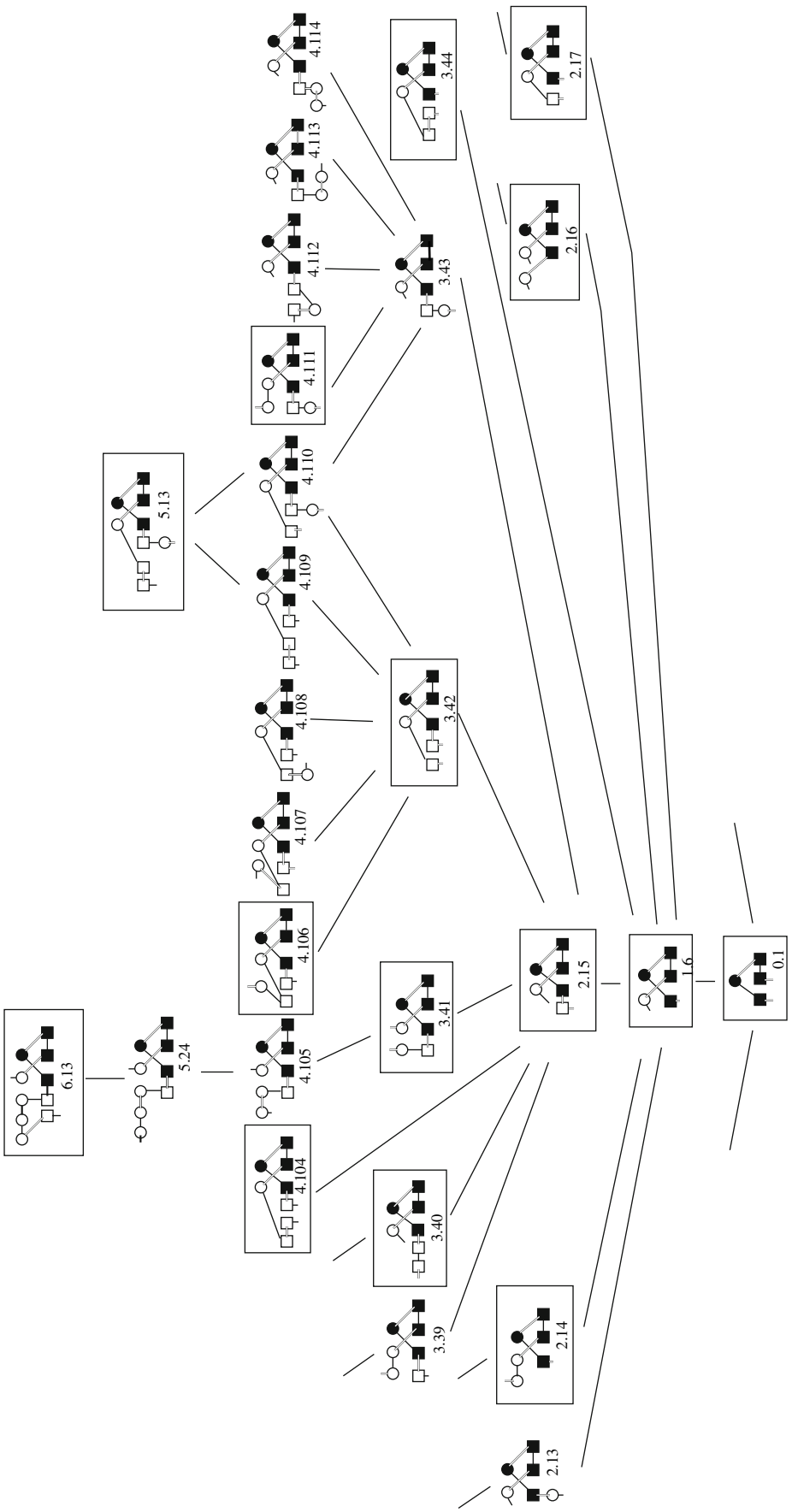

Fig. 2. Fragment of the structural tree of (α + β)-proteins containing abCd units. All structures are similarly oriented and are viewed from the face as in Fig. 1c. The folds actually found in proteins are framed.

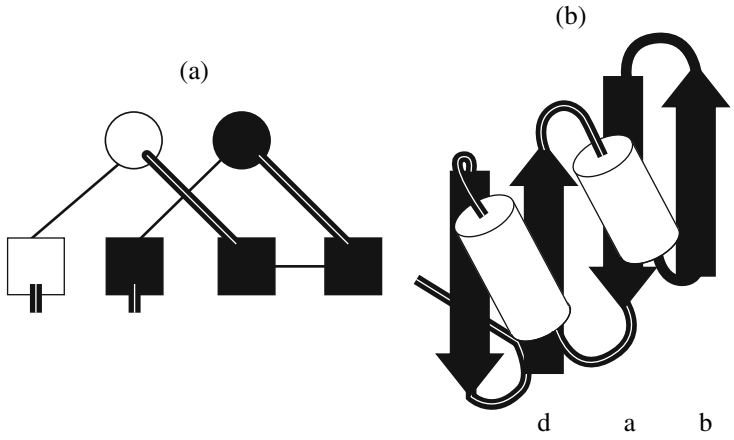

Fig. 3. Frequent fold, found in 256 proteins and domains. (a) The fold viewed from the face. (b) Scheme of the fold in ferredoxin (PDB ID 1ir0) used as an example.

It should be noted that certain folds can be derived from different parental structures via several different pathways. Consequently, it is rather difficult to determine the structural similarity between proteins or domains, and the tree is overloaded because some branches are repeated. To avoid such situations, we used an empirical approach when modeling the structural tree. First, a total tree was modeled using all theoretically possible folds and their complication pathways. Next, when two identical branches originating from different parental structures were found, we determined the abundance of proteins with known structures for each of the branches. The branch most abundant in proteins with known structures was left in the tree, while the identical branches were removed. This procedure made it possible to identify the preferential structural complication pathway in most cases.

Several limitations were used when constructing the structural tree of (α + β)-proteins. The most important limitations are described below. First, we disregarded the structures that consist of more than three layers, since the overwhelming majority of proteins (386 out of 401 of nonhomologous proteins) consist of two (α + β) or three (α + β + α) layers in this class. Second, the ratio between the number of α-helices in the α-layer to the number of β-strands in the β-layer was no more than 2 : 3, since incompact structures are otherwise obtained (see rule (3)). The complication pathways, including incompact structures, were disregarded when modeling the structural tree.

All theoretically allowable folds are included in the tree up to the level of four additional secondary structure elements. At the higher levels, the tree includes only the structural complication pathways that lead to known protein structures. One of the main causes of this is that the number of theoretically possible folds dramatically increases at higher levels, and only the branches containing known proteins are shown to avoid overloading. Another cause is that the abundance in known proteins rapidly decreases at higher level; i.e., proteins of this class are limited in size. For instance, we found only two proteins (among 401) that each had nine additional secondary structure elements.

On the other hand, the analysis of the tree showed that proteins consisting of an abCd unit and two or three additional elements are the most common (more than 60%); i.e., the majority of proteins are of an optimal size. As an example, Fig. 3 shows a fold that consists of an abCd unit and two additional elements, an α-helix and a β-strand. This fold occurs in 108 out of 401 nonhomologous proteins; i.e., its frequency is disproportionately high. In total, the updated tree comprises 286 folds. Of these, 80 are found in 401 nonhomologous proteins, and 25% of these proteins have the same fold. Note that a high frequency of this fold has already been observed [14, 18], although the statistics were not so impressive.

In addition, analysis of the updated tree showed that the reverse orientation of the polypeptide chain in abCd units of proteins is far more common than the direct orientation. Among 405 abCd units found in 401 nonhomologous proteins, 92 have a direct orientation, and 313 (77%) have a reverse orientation of the polypeptide chain. The cause of such predominance is unclear and needs further investigation.

As already mentioned, one of the most important applications of structural trees is related to a structural classification of proteins. All proteins and domains belonging to one tree can be assigned to one structural class or superfamily. Proteins and domains belonging to branches of a structural tree form subclasses. It is clear that such a classification is only based on the spatial structural similarity and common folding pathways simulated in the tree. The classification disregards the amino acid sequences, functions, and evolu-

tionary relationships of proteins. Such factors are considered, to some extent, in other known classifications [7–9].

Another application of structural trees concerns the relationship between the structure and functions of proteins. Several illustrative examples are known where proteins with the same folds perform the same function. For instance, many DNA-binding proteins have the same fold (THT motif) and group in the left branch of a structural tree of proteins containing α-α corners [1, 2], while the right branch of the tree includes Ca^{2+}-binding proteins, which have EF helix pairs. Proteins that contain the OB fold strongly tend to oligonucleotide and oligosaccharide binding [19] (see also the corresponding branch of a tree of proteins containing S-like β-sheets [4]). RNA-binding proteins group in the central part of the tree of (α + β)-proteins containing the abCd units (Fig. 2) [14, 18]. However, an analysis showed that an opposite picture is more common; i.e., proteins with similar or even identical folds have different functions (e.g., see the fold shown in Fig. 3), while proteins with the same functions have absolutely different polypeptide chain folds and belong to different structural classes. This is probably explained by the fact that the function of a protein depends not only on the general polypeptide chain fold (which is taken into account in the construction of structural trees), but also on the fine structure of the total protein and, especially, the fine structure of its active site. On the other hand, analysis of structural trees and structural motifs indicates that structural similarity of proteins is not based on their evolutionary relationships or similarity of their functions, but is underlain by general physicochemical regularities, which "select" the most advantageous polypeptide chain folds [1–5].

Based on the updated structural tree, we constructed a hierarchic database of all (α + β)-proteins containing abCd units. The database is available at http://strees.protres.ru/. The site includes pages with the hierarchic database of proteins of this group, the structural tree with folds actually found in proteins, 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. The pages are logically interconnected via context links. The PDB files of all proteins contained in the database (total 2636 PDB files for 926 proteins and their mutants) can be loaded and viewed using any molecular graphics program. In addition, the site includes a structural tree of β-proteins containing abcd units and the corresponding hierarchic database. Database compilation and construction of updated structural trees for other structural classes are currently in progress. All structural trees and databases will be available through the internet.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 07-04-00659).

REFERENCES

1. Efimov A.V. 1996. A structural tree for α-helical proteins containing α-α-corners and its application to protein classification. *FEBS Lett*. **391**, 167–170.
2. Efimov A.V. 1997. Structural trees for protein superfamilies. *Proteins*. **28**, 241–260.
3. Efimov A.V. 1997. A structural tree for proteins containing 3β-corners. *FEBS Lett*. **407**, 37–41.
4. Efimov A.V. 1998. A structural tree for proteins containing S-like β-sheets. *FEBS Lett*. **437**, 246–250.
5. Efimov A.V. 2008. Structural trees for proteins containing φ motifs. *Biokhimiya*. **73**, 29–35.
6. Gordeev A.B., Kondratova M.S., Efimov A.V. 2008. Novel structural tree of β-proteins containing abcd units. *Mol. Biol*. **42**, 323–326.
7. 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. *J. Mol. Biol*. **247**, 536–540.
8. Holm L., Sander C. 1999. Protein folds and families: Sequence and structure alignments. *Nucleic Acids Res*. **27**, 244–247.
9. 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.
10. Day R., Beck D.A.C., Armen R.C., Daggett V. 2003. A consensus view of fold space: Combining SCOP, CATH and the Dali Domain Dictionary. *Protein Sci*. **12**, 2150–2160.
11. Novotny M., Madsen D., Kleywegt G.J. 2004. Evaluation of protein fold comparison servers. *Proteins*. **54,** 260–270.
12. Tatusova T.A., Madden T.L. 1999. Blast 2 sequences: A new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett*. **174**, 247–250.
13. Sayle R.A., Milner-White E.J. 1995. RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci*. **20**, 374–376.
14. Efimov A.V. 1995. Structural similarity between two-layer α/β amd β proteins. *J. Mol. Biol*. **245**, 402–415.
15. Lim V.I., Mazanov A.L., Efimov A.V. 1978. Stereochemical theory of the spatial structure of globular proteins: 1. Highly coiled intermediate structures. *Mol. Biol*. **12**, 206–213.
16. Richardson J.S. 1977. β-Sheet topology and relatedness of proteins. *Nature*. **268**, 495–500.
17. Rao S.T., Rossman M.G. 1973. Comparison of supersecondary structures in proteins. *J. Mol. Biol*. **76**, 241–256.
18. Janin J. 1993. Shared structural motif in proteins. *Nature*. **365**, 21.
19. Murzin A.G. 1993. OB (oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. *EMBO J*. **12**, 861–867.