ISSN 0026-8933, Molecular Biology, 2008, Vol. 42, No. 2, pp. 285–288. © Pleiades Publishing, Inc., 2008.
Original Russian Text © A.B. Gordeev, M.S. Kondratova, A.V. Efimov, 2008, published in Molekulyarnaya Biologiya, 2008, Vol. 42, No. 2, pp. 323–326.

STRUCTURAL–FUNCTIONAL ANALYSIS OF BIOPOLYMERS AND THEIR COMPLEXES

UDC 577.150.2

Novel Structural Tree of β-Proteins Containing abcd Units

A. B. Gordeev, M. S. Kondratova, and A. V. Efimov

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

Received July 16, 2007
Accepted for publication August 16, 2007

Abstract—A database of 528 β-proteins and β-domains containing abcd units, including 244 nonhomologous ones, was compiled from the Protein Data Bank (total 1511 PDB entries). A novel structural tree of this class of proteins was constructed to include 153 possible polypeptide chain folds. A structural classification of β-proteins containing abcd units was proposed on the basis of the tree. The database and the structural tree are available at http://strees.protres.ru/.

DOI: 10.1134/S0026893308020155

Key words: modeling, classification, 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 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 a root of a tree. Possible ways of structural complications are shown with lines, which eventually integrate all structures in one tree.

The earliest versions of structural trees were constructed about a decade ago [1–4]. The number of solved protein structures in the Protein Data Bank (PDB) has substantially increased over this decade. For instance, a structural tree constructed for β-proteins containing abcd units in 1997 [2] included about 40 known protein structures. To date, we have compiled a database of 528 proteins and domains of this class, suggesting 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.

SUBJECTS AND METHODS

A database of β-proteins containing abcd units was compiled using the Structural Classification of Proteins (SCOP) database (http://scop.mrc-lmb.cam.ac.uk/scop/). Proteins were manually selected. In total, we retrieved 528 proteins and domains containing abcd units, including 244 nonhomologous. Possible homologies were revealed by BLAST pairwise alignment (http://www.ncbi.nih.gov/BLAST/). Protein structures were visually examined using the RasMol molecular graphics program [5].

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 [6]. The simplest abcd unit consists of β-strands a, b, c, and d, which are consecutively arranged in the polypeptide chain. Of these, three β-strands (a, b, and d) belong to one layer and the fourth one (c) belongs to another. Strands b, c, and d form a right-handed superhelix bcd. The abcd unit occurs in two variants, with a direct or a reverse orientation of the polypeptide chain (Figs. 1a, 1b). Like

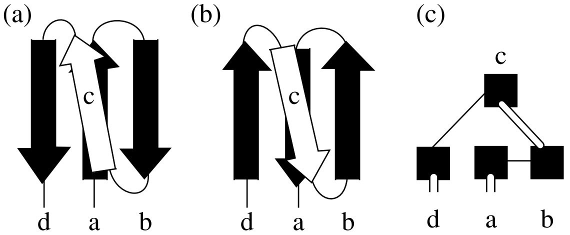

Fig. 1. Schemes of the abcd unit with a (a) direct or (b) reverse chain orientation. The β-strands are shown with arrows directed from the N to the C end. (c) The abcd unit as viewed from the face. The β-strands are shown with squares and designated with the corresponding letters; the linkers directed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

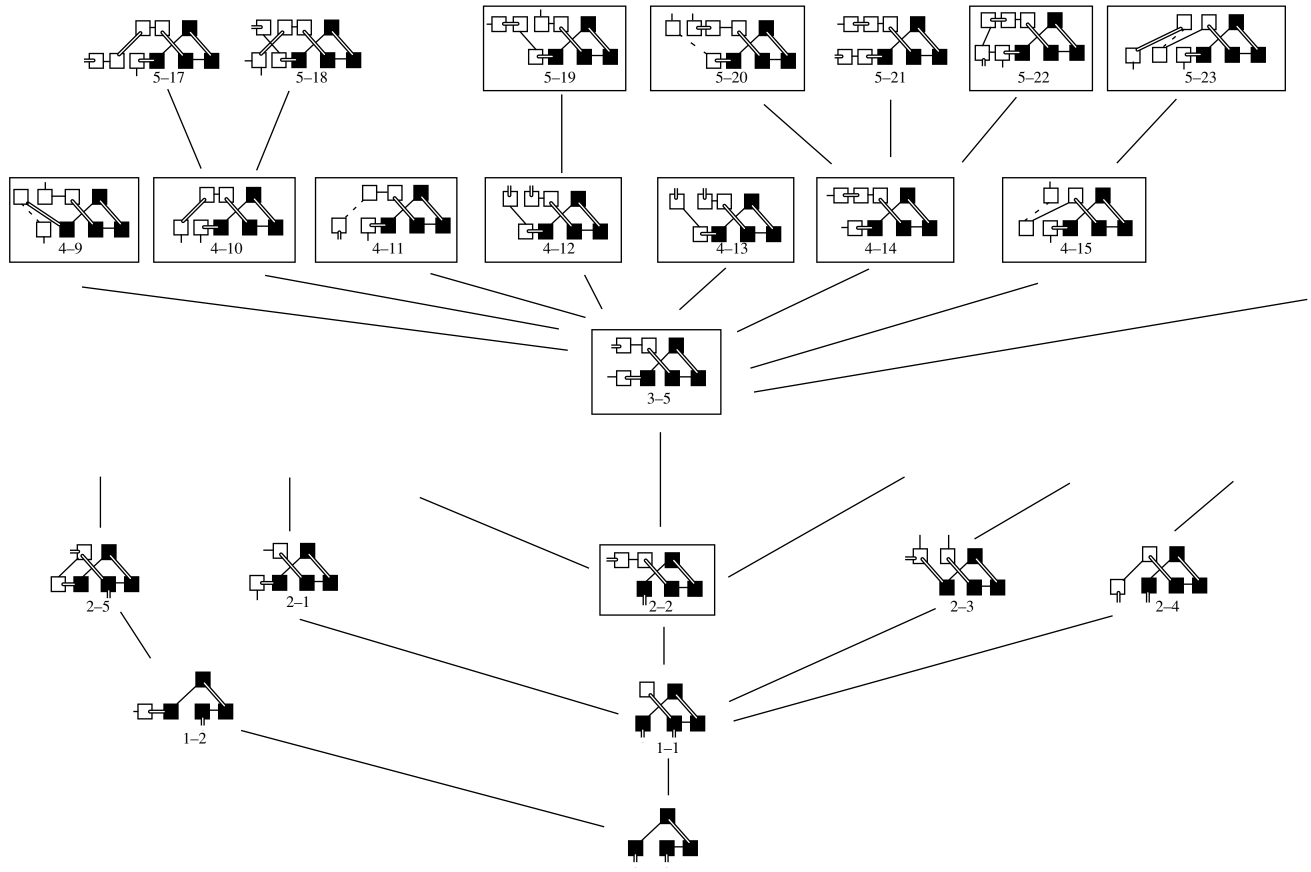

Fig. 2. Fragment of the structural tree of β-proteins containing abcd units. All structures are similarly oriented and are viewed from the face as in Fig. 1c. The folds actually found in proteins are framed.

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 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 β-strand closest to the growing structure in the polypeptide chain was the first to be added [2, 6].

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

(4) Linker intercrosses [7] and knots [8] are prohibited.

(5) Each structural motif (not only the root ones) has its characteristic chirality and polypeptide chain fold.

(6) In two-layered β-proteins, three β-strands located consecutively in the chain do not form a β–β–β superhelix where the first and third β-strands directly interact with each other to produce a parallel β-structure [2, 6]. The formation of a superhelix by three consecutive β-strands is allowed when at least one additional β-strand occurs between the first and third strand in the β-sheet (e.g., superhelix bcd and strand a in the abcd unit). As we demonstrate here, this rule works well when the closest three or four strands are added to the abcd unit and is violated in 25–30% of cases when the fifth, sixth, or seventh β-strand is added to complete the structure.

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/. As is seen, several levels, or rows, are distinguishable in the tree. Each level (row) comprises the possible polypeptide chain folds that have the same number of β-strands. All folds are numbered in the tree, each fold having a specific identifier consisting of two numbers separated with a point or a dash. The first number shows the number of β-strands added to the abcd unit. For instance, first-level folds contain one additional β-strand 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 left to right. For instance, the fourth row includes 36 folds, designated from 4–1 to 4–36.

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 of different branches incorporate the fold corresponding to 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 is.

In total, the updated tree includes 153 folds. All theoretically allowable folds are included in the tree up to the level of five additional β-strands. At the higher levels, the tree includes only the structural complication pathways leading to known protein structures. One of the main causes of this is that proteins and domains are limited in size and most (~75%) of them consist of an abcd unit with three to five additional β-strands. Another cause is that the number of possible folds dramatically increases at higher levels and only some folds are shown to avoid overloading. Even with these limitations, the updated tree comprises several times more possible folds than the earlier tree (153 vs. 36 [2]), while the number of known proteins and domains is almost one order of magnitude greater than in the earlier tree (528 vs. 40). The updated tree includes several new intermediate structural complication pathways, leading to new folds. Interestingly, among 18 folds included in the earlier tree but unknown for proteins at that time, seven have been revealed to date in actual protein structures. Thus, at least seven new folds were correctly predicted.

Analysis of the structural tree makes it possible to study the "maturation" of a protein globule or a domain and, first and foremost, to elucidate the factors preventing further structural complication. As already mentioned, proteins and domains are limited in size, those with four additional β-strands being the most abundant (about 37% of all nonhomologous proteins). The abundance of proteins rapidly decreases with increasing level, and we found only one protein (among 244) that has eight additional β-strands. Thus, a "mature" protein or a domain should have an optimal number of β-strands.

On the other hand, departures from the rules of adding secondary structure elements to growing structures mostly occur at the last steps of structural complication. For instance, rule (6) is violated in 20 proteins and domains belonging to seven different folds and a violation is observed at the last complication steps in all cases. Likewise, left-handed βαβ units (which are usually right-handed) in α/β-proteins are formed at the last steps of their growth [2]. Such violations are possibly a factor preventing further structural complication. Yet this problem needs further investigation.

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. 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 evolutionary relationships of proteins. Such factors are considered, to some extent, in other known classifications [9–11].

Based on the updated structural tree, we constructed a hierarchic database of all β-proteins containing abcd units. The database is available at http://strees.protres.ru/. The main page of the site gives access to pages with fragments of the structural tree and to a system retrieving a protein of interest by its PDB ID. The pages are logically interconnected via context links. The PDB files of all proteins contained in the database (total 1511 PDB files for 528 proteins and their mutants) can be loaded and viewed using any molecular graphics program. A guide page is included to facilitate working with the database.

Database compilation and construction of updated structural trees for other structural classes are in progress now. 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. Sayle R.A., Milner-White E.J. 1995. RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374–376.
6. Efimov A.V. 1982. Supersecondary structure of β-proteins. *Mol. Biol.* **16**, 799–806.

7. 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.

8. Richardson J.S. 1977. β-Sheet topology and relatedness of proteins. *Nature*. **268**, 495–500.

9. 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.

10. Holm L., Sander C. 1999. Protein folds and families: Sequence and structure alignments. *Nucleic Acids Res.* **27**, 244–247.

11. 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, 1093–1108.