Biochemical and Biophysical Research Communications 397 (2010) 470–471

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

PCBOST: Protein classification based on structural trees

Alexey B. Gordeev, Anton M. Kargatov, Alexander V. Efimov *

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

ARTICLE INFO

Article history:
Received 25 May 2010
Available online 2 June 2010

Keywords:
Protein structure comparison
Classification
Modeling

ABSTRACT

In this paper, we present the protein classification based on structural trees (PCBOST). This is a novel hierarchical classification of proteins that is primarily based on similarity of overall folds of proteins as well as on the modeled folding pathways of proteins. Amino acid sequences, functions of proteins and their evolutionary relationship are not taken into account in this classification. To date the database includes 3847 proteins and domains grouped into six categories having structural similarity and forming six structural trees (total 10,547 PDB-entries). The work on extension of the database and construction of novel structural trees is in progress. The service is free for all users and available at the URL <http://strees.protres.ru/>.

1. Introduction

Currently, there is a number of protein structure classification schemes and three of them, SCOP (structural classification of proteins) [1], CATH (classification–architecture–topology–homology) [2], and Dali Domain Dictionary [3], have found widespread use (for reviews, see, e.g., [4,5]). While the Dali Domain Dictionary is largely automated, CATH and SCOP combine manual classification with the automatic structural alignment programs. Both SCOP and CATH use homology for classification, SCOP in the first two hierarchical levels (family and superfamily) and CATH in the first level. In this paper, we present one more structural classification of proteins that is different from the above-mentioned schemes in some aspects. First of all, amino acid sequences, functions and evolutionary relationships of proteins are not taken into account in our classification. It is based on similarity of overall folds of proteins and domains and modeled protein folding pathways which are represented as structural trees.

2. Materials and methods

The structural tree for a group of proteins having structural similarity is a scheme that includes all the intermediate and final structures connected by lines showing possible pathways of stepwise growth of a starting structure [6]. The structural motif having a unique overall fold that occurs in all proteins of the structural group is taken as the starting structure in modeling or the root structure of the tree. Larger structures are obtained by stepwise addition of α-helices and/or β-strands to the root motif in accordance with a restricted set of rules inferred from known principles of protein structures [6–8].

Among them, the following rules are the most important:

(1) 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.
(2) 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.
(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]).
(4) Crossing of connections [11] and formation of knots [12] are prohibited.
(5) All the obtained structures should have the corresponding handedness. For example, all the β-α-β-units should be in the form of the right-handed superhelix [13,14].

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 starting and intermediate structures. Thus, the structural trees are a good tool for searching possible folding pathways and all allowed protein folds as well as for structure comparison and protein classification.

3. Results

The increasing number of protein structures in the Protein Data Bank (PDB) has prompted us to construct updated structural trees that include more known protein structures and show some novel folding pathways as compared with the first versions of the trees [6]. Based on the updated trees, we compiled a hierarchic database

* Corresponding author. Fax: +7 495 514 0218.
E-mail address: efimov@protres.ru (A.V. Efimov).

0006-291X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2010.05.136

of proteins of the corresponding structural classes. To date the classification database for six structural classes of proteins has been developed. The database includes: β-proteins containing abcd-Units [15], (α + β)-proteins containing abcd-Units [16] and α/β-proteins containing β-α-β-ψ-motifs, ψ-β-α-β-motifs [17], five-segment and seven-segment α/β-motifs. Six updated structural trees for these structural classes have been constructed. We have also constructed computer versions of the structural trees. Several other structural trees have also been constructed and published [6,18–20] but their computer versions are in progress now.

Based on the structural trees, we have developed PCBOST, the protein classification based on structural trees. This structural classification of proteins is only based on the spatial structural similarity and common folding pathways simulated with the trees. The classification disregards the amino acid sequences, functions, and evolutionary relationships of proteins which are taken into account in other known classifications. PCBOST is a hierarchically organized database of protein structures. It includes several hierarchical levels. Proteins and domains having the same root structural motif are combined into the STRUCTURAL TREE (Fig. 1). Proteins and domains located at horizontal levels of the structural tree are grouped into LEVELS. All proteins and domains from one LEVEL having the same arrangement of secondary structure elements form FOLDS. For example, the structure of SirA protein (PDB-code: 1DCJ) contains an abCd-Unit. So the protein belongs to "Mixed_alpha + beta_proteins_with abCd-Units" (STRUCTURAL TREE). The protein has two α-helices and four β-strands: four elements form the abCd-Unit, and two elements are added to it. So the protein belongs to "Mixed alpha + beta proteins with the abCd-Unit and two added elements" (the second LEVEL). In the second LEVEL of the corresponding STRUCTURAL TREE, it occupies the sixth position from the left denoted by an index number 2.6. It means that the protein belongs to "FOLD 2.6".

To represent structural trees we used the following rules. Protein structures are described in terms of simple schemes. β-Strands are shown as squares and α-helices as circles. The connections located near to the viewer are shown with double lines and the far connections with single lines. The crossovers of an element from one layer to the other are shown with dotted lines. The growth of the protein folds is realized by stepwise addition of secondary structural elements to the corresponding structural motif.

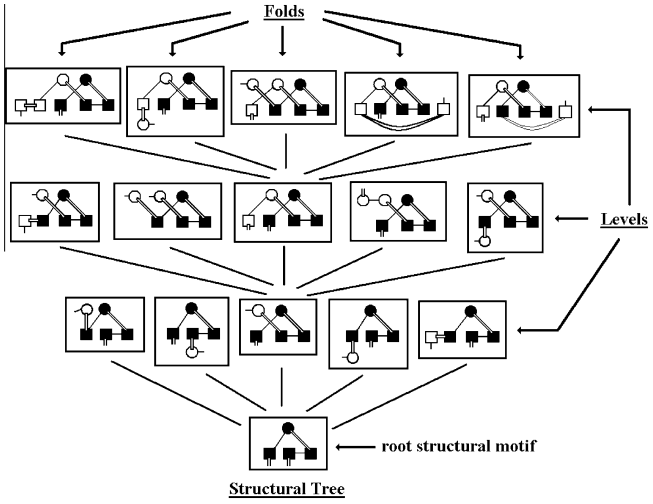

Fig. 1. A fragment of the structural tree of (α + β)-proteins containing the abCd-Unit. The structures are viewed end-on with α-helices shown as circles and β-strands as rectangles. Near connections are shown by double lines and far connections by single lines. Arrows show the root structural motif (abCd-Unit), levels and folds of the structural tree.

To date the PCBOST database includes six structural trees, 64 levels, 477 folds, 3847 proteins and domains and 10,547 PDB-entries. The site has a user-friendly interface and includes pages representing the hierarchic protein database and structural trees, a guide page, and a system retrieving a protein of interest by its PDB ID and PDB files. The pages are logically interconnected via context links. The PDB files of all proteins contained in the database can be downloaded and viewed using any molecular graphics program. The work on the construction of the updated as well as novel structural trees is now in progress.

PCBOST database is available at <http://strees.protres.ru/>.

4. Discussion

In our opinion, the structural classification of proteins should help us to analyze the information about their structures and to use it for research. Our approach shows structural relationship between nonhomologous proteins and suggests a mechanism of this relationship that is demonstrated with the structural trees of proteins. Thus, the structural trees are a good tool for searching of all possible folds of the polypeptide chain, for modeling of folding pathways of proteins and their structures, for protein structure comparison and classification etc.

Acknowledgments

This work was supported by the Russian Foundation for Basic Research (Project No. 10-04-00727) and by the Federal Agency for Science and Innovations (02.740.11.0295).

References

[1] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.
[2] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, CATH – a hierarchic classification of protein domain structures, Structure 5 (1997) 1093–1108.
[3] S. Dietmann, J. Park, C. Notredame, A. Heger, M. Lappe, L. Holm, A fully automatic evolutionary classification of protein folds: DALI domain dictionary version 3, Nucleic Acids Res. 29 (2001) 55–57.
[4] M. Novotny, D. Madsen, G.J. Kleywegt, Evaluation of protein fold comparison servers, Proteins 54 (2004) 260–270.
[5] R. Kolodny, D. Petrey, B. Honig, Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction, Curr. Opin. Struct. Biol. 16 (2006) 393–398.
[6] A.V. Efimov, Structural trees for protein superfamilies, Proteins 28 (1997) 241–260.
[7] A.V. Efimov, Structural similarity between two-layer α/β and β-proteins, J. Mol. Biol. 245 (1995) 402–415.
[8] A.V. Efimov, A structural tree for α-helical proteins containing α-α-corners and its application to protein classification, FEBS Lett. 391 (1996) 167–170.
[9] C. Chothia, M. Levitt, D. Richardson, Structure of proteins: packing of α helices and pleated sheets, Proc. Natl Acad. Sci. USA 74 (1977) 4130–4134.
[10] A.V. Efimov, Stereochemistry of packing of α-helices and β-structure in a compact globule, Dokl. Acad. Nauk S.S.S.R. 235 (1977) 699–702.
[11] V.I. Lim, A.L. Mazanov, A.V. Efimov, Stereochemical theory of spatial structure of globular proteins. I. Highly-helical intermediate structures, Mol. Biol. (Moscow) 12 (1978) 206–213.
[12] J.S. Richardson, β-Sheet topology and the relatedness of proteins, Nature 268 (1977) 495–500.
[13] S.T. Rao, M.G. Rossmann, Comparison of super-secondary structures in proteins, J. Mol. Biol. 76 (1973) 241–256.
[14] M.J.E. Sternberg, J.M. Thornton, On the conformation of proteins: the handedness of the β strand-α helix-β strand unit, J. Mol. Biol. 105 (1976) 357–382.
[15] A.B. Gordeev, M.S. Kondratova, A.V. Efimov, Novel structural tree of β-proteins containing abcd units, Mol. Biol. (Moscow) 42 (2008) 285–288.
[16] A.B. Gordeev, A.V. Efimov, Novel structural tree of (α + β)-proteins containing abCd units, Mol. Biol. (Moscow) 43 (2009) 480–484.
[17] A.M. Kargatov, A.V. Efimov, A novel structural motif and structural trees for proteins containing it, Biochemistry (Moscow) 75 (2010) 249–256.
[18] A.V. Efimov, A structural tree for proteins containing 3β-corners, FEBS Lett. 407 (1997) 37–41.
[19] A.V. Efimov, A structural tree for proteins containing S-like β-sheets, FEBS Lett. 437 (1998) 246–250.
[20] A.V. Efimov, Structural trees for proteins containing φ-motifs, Biochemistry (Moscow) 73 (2008) 23–28.