Protein 3D architectures and structural motifs: automated annotation

E.A. Aksianov[1], A.V. Alexeevski[1,2]

[1] *Belozersky Institute, Lomonosov Moscow State University, Leninskie Gori 1, Moscow, Russia*

[2] *Scientific Research Institute for System Studies (NIISI RAN), Moscow, Russia*

evaksianov@gmail.com, aba@belozersky.msu.ru

Description of a protein 3D structure in terms of structural domains, architecture, topology and structural motifs is a hard problem. Indeed, there are continuous transitions between, for example, β-sandwiches and open β-barrels, and other pairs of folds; there is no standard means to describe arrangements of β-sheets and α-helixes in 3D space; minor structural elements (short α-helixes, β-hairpins, etc.) and uncertain secondary structure assignments complicate formal fold descriptions. These obstacles lead to significant discrepancies in classification of the same structures in popular databases SCOP and CATH [1].

Existing automatic tools for fold description (TOPS, PTGL, PROMOTIF and others) mainly rely on a protein topology (a scheme of contacts of consecutive β-strands and helixes) rather than architecture (spatial arrangement of β-sheets and α-helixes). In contrast, an expert examining a new structure perceives first an architecture of a protein; it takes more his efforts to recover topology.

Here we present a detector of the architecture of input protein 3D structure. In its algorithm we attempted to follow expert's decision making. The algorithm is implemented as **ProtOn** web-service (http://mouse.belozersky.msu.ru/proton). **ProtOn** consists of 3 main programs.

SheeP [2] is a detector of β-sheets. It enhances β-sheets recovered from secondary structure assignments made by **DSSP** or **STRIDE** detectors in order to obtain refined β-sheets more adequate to human judgments. SheeP presents β-sheets by *sheet maps* (Fig. 1B).

ArchiP is a detector of an architecture of a protein structure. It was designed for detection of architectures of all-β and α/β classes only. **ArchiP** identifies *architectural units* and constructs a graph of their contacts (Fig. 1C). Units are of types: β-barrels (which not necessarily coincide with whole sheet), not barrel β-sheets, parts of β-sheets, layers of α-helixes (α-layers) adjacent to the units of β-type. In the graph edges corresponding to main and side contacts are distinguished.

A core of an architecture is detected as a connected component with respect to the main edges only. Units that contacts with the core are considered as extension of the architecture.

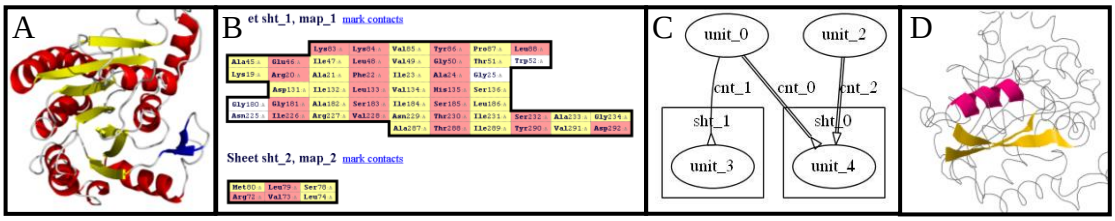

Fig. 1. Example of αβα-sandwich, PDB code 1CWU, chain A. A. Cartoon model. Small β-sheet is shown in blue. B. SheeP output: maps of β-sheets. Cells of the tables correspond to residues, rows correspond to strands. Residues located on the same side of the β-sheet are colored similarly. C. ArchiP output: architectural graph. Boxes correspond to β-sheets. All β-units are included in appropriate boxes. Vertices (ovals) outside boxes correspond to α-layers. A side of a β-unit contact is shown by a triangle at the end of corresponding edge. Main contacts are shown by double edges. Thus, core of the architecture consists of the β-sheet (unit_4) and two α-layers (unit_0 and unit_2) contacting with the β-sheet from opposite sides. The core is extended by small β-sheet (unit_3). D. MotAn output: one of detected βαβ-motifs.

Architectural graphs of a dozen of common architectures (β-sandwichs, β-barrels, β-propellers, TIM-barrels etc.) are of special type. If the graph of core units coincides with one of these graphs, then **`ArchiP`** describes the architecture in common terms. In other cases the output is architectural graph only. To check **`ArchiP`** we have detected architectures in all SCOP domains of 125 folds that have exact architectural annotations. Correct architecture assignment were obtained 74% of domains, from 66% for β-propellers to 99% for α/β-sandwiches.

The 3rd program, **`MotAn`**, detects structural motifs, namely, interlocks (87% correct assignments), jelly-rolls (86%), β-helixes (99%), βαβ-motifs (99%), meanders (80%). We checked that **`MotAn`** assignments are the same or substantially better than assignments of **`PTGL`** [3], another structural motif detector, for motifs common for both programs.

The work was partially supported by RFBR grants 10-07-00685-a, 11-04-91340-a.

1. G.Csaba et al. (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis, *BMC Structural Biology*, **9:**23.

2. E.Aksianov, A.Alexeevski (2012) SheeP: A Tool for Description of β-sheets in Protein 3D Structures, *JBCB*, **10:**2.

3. P.May et al. (2010) PTGL: a database for secondary structure-based protein topologies, *NAR*, **38**:D326-D330