Sequence alignment of non-superposable beta-sheets

Evgeniy Aksianov

Belozersky Institute, Lomonosov Moscow State University,
Leninskie Gori 1, Moscow, Russia, evaksianov@gmail.com

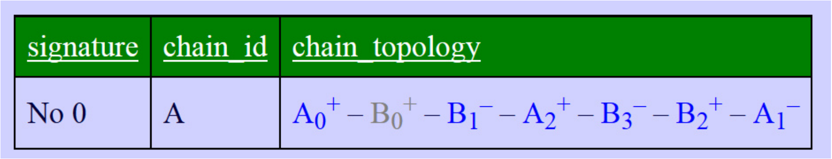
Andrey V. Alexeevski

Belozersky Institute, Lomonosov Moscow State University,
Scientific Research Institute for System Studies (NIISI RAN), Moscow, Russia
aba@belozersky.msu.ru

Protein 3D structures usually are more conserved than their sequences. Thus, structural comparison allows detecting homology between proteins with highly diverged sequences. Superimposition of 3D structures of proteins is commonly used for finding homologs with no detectable sequence similarity. The outputs of programs for hard (PDBeFold, CE) or flexible (FATCAT, MultAlign) superimposition include particularly sequences alignment of superimposed parts of protein structures. However, not all similarities of protein structures can be detected by superimposition.

We describe here new approach for revealing such kind of similarities. Previously, we developed web-service Proton (http://mouse.belozersky.msu.ru/proton) for automatic annotation and comparison protein structures of all-alpha and alpha/beta classes according SCOP classification. First, SheeP program represents each β-sheet in format of *sheet map* – a plain table, cells of which correspond amino acids within the sheet. Beta-strands are contained in sheet map rows. A pair of amino acids from adjacent strands is considered paired if it is surrounded by regular hydrogen bonds between backbone atoms. Crests are defined as chains of paired of amino acids. Crests are contained in sheet map columns.

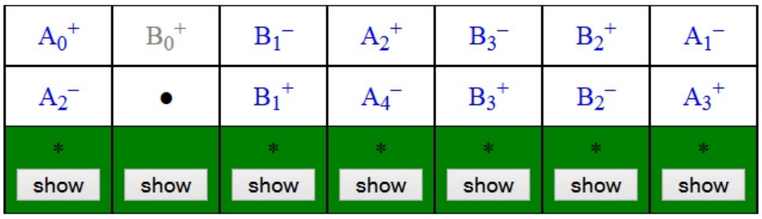
Second, MotAn program creates scheme of protein topology, which is as a sequence of strands and helixes. Each strand has additional marks. The marks indicate a sheet, a row in sheet map and an orientation of strand within sheet map (left to right or right to left), see Fig. 1.



signature	chain_id	chain_topology
No 0	A	$A_0^+ - B_0^+ - B_1^- - A_2^+ - B_3^- - B_2^+ - A_1^-$

Fig. 1. Topology scheme of 2CVV, chain A (N-terminal domain) generated by MotAn. The every strand is denoted by sheet name (A or B), index of row in the sheet map and sign (plus or minus) denoting the direction in the sheet map (left-to-right or right-to-left).

Third, ProTop program generates alignment of topologies of two proteins, i.e. optimal correspondence between nodes of topology descriptions relative certain score. This score particularly reflects similarities of orders of rows within beta-sheets (lower indices on topology scheme) . Fig 2 demonstrates an example of pairwise topology alignment.



A_0^+	B_0^+	B_1^-	A_2^+	B_3^-	B_2^+	A_1^-
A_2^-	●	B_1^+	A_4^-	B_3^+	B_2^-	A_3^+
* show	show	* show	* show	* show	* show	* show

Fig. 2. Alignment of topologies of 2CVV, chain A (N-terminal domain) and 1R75, chainA generated by ProTop. Insertion is denoted by bullet, matches – by asterisks.

We developed an algorithm, which takes on input alignment of topologies of two protein structures and return sequence alignment of corresponding β-strands within corresponding β-sheets. Sequence alignment of two strands is created by dynamic programming algorithm using the following scoring system. Let we chose one amino acid from strand of first topology, and another from second topology. This correspondence can be extended to correspondence of amino acids from crests containing chosen amino acids. Score for match of chosen amino acids is computed taking in account the size of common parts of crests and sequence similarities within them.

This work was partially supported by RFBR grants 14-04-31709_mol_a and 13-07-00969 and Russian Science Foundation grant 14-50-00029.