Using of PREFAB for analysis of amino-acid sequence alignment algorithms.

Irina Poverennaya[1], Mikhail Lobanov[2], Victor Yacovlev[3], Mikhail Roytberg[3]

[1]*Moscow State University, Russian Federation*
[2]*Institute of Protein Research RAS,, Russian Federation*
[3]*Institute of Mathematical Problems in Biology RAS, Russian Federation, mroytberg@lpm.org.ru*

Reference alignments are essential for correct analysis of amino-acid sequence alignment algorithms, because their comparison with algorithmic alignments allows one to assess the quality of these methods. The protein reference alignment benchmark PREFAB [1] was used in many studies (e.g., see [2, 3]); it contains 1682 alignments which were obtained using 3D alignment of protein structures. Unfortunately, selection principles for aligned sequence pairs aren't described. At the same time, sequence names include only PDB ID and chain and it remains unclear what fragments of this chain were used.

The aim of our study is to find out correspondence between PREFAB sequence pairs and SCOP structural domain classification [4]. Briefly, we accept amino-acid sequence, if (1) it is a fragment of one chain in a PDB entry [5], and (2) this fragment match with at least one of domains which were describeded in SCOP. An alignment is accepted only if domains corresponded compared sequences belong to the same «family» of SCOP classification.

The analysis of PREFAB have shown the following results: PREFAB has been proved to include sequences that form domain with fragments from other chains of the same protein or contain more than one domain. Besides some chain fragments were found deleted in some one-domain PREFAB sequences. These fragments are likely to correspond with unstructural regions of protein. Consequently we select 1294 sequences (or 1115 PREFAB alignments) that satisfy mentioned conditions. 834 alignments of them satisfy the condition related to SCOP families. A quantity of insertions and sequence identity were computed for each selected alignment. In addition, new structural alignment was built for each PREFAB alignment using the program [6] and distances between the corresponding residues were calculated.

The refined dataset will be used to assess quality to alignments obtained with a method [2] (server address: [7]).

1. PREFAB benchmark: http://www.drive5.com/muscle/prefab.htm
2. Yacovlev V.V., Roytberg M.A. Increase of global amino-acid sequence alignment accuracy by alignment-candidate set construction // Biophysics. - 2010. - T. 55, N 6. - C. 965-975
3. Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97.
4. Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classi_cation of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.
5. Protein Data bank: http://www.pdb.org/pdb/home/home.do
6. Structural alignment: http://phys.protres.ru/~mlobanov/casp/prog.html#MaxSub
7. Server: http://server2.lpm.org.ru/bio/online/pareto/