ISSN 0006-3509, Biophysics, 2012, Vol. 57, No. 2, pp. 133–137. © Pleiades Publishing, Inc., 2012.
Original Russian Text © T.V. Astakhova, M.N. Lobanov, I.V. Poverennaya, M.A. Roytberg, V.V. Yacovlev, 2012, published in Biofizika, 2012, Vol. 57, No. 2, pp. 205–211.

MOLECULAR BIOPHYSICS

Verification of the PREFAB Alignment Database

T. V. Astakhova[a], M. N. Lobanov[b], I. V. Poverennaya[a,c], M. A. Roytberg[a], and V. V. Yacovlev[b]

[a] *Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia*
[b] *Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia*
[c] *Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119991 Russia*
E-mail: victor@lpm.org.ru

Received August 24, 2011

Abstract—Verification of the PREFAB database containing golden standard protein alignments was performed. It has revealed a significant number of differences between the sequences from PREFAB and PDB databases. It was shown that, compared with the sequences given in the PDB, 575 alignments referred to a sequence with a gap; such alignments were excluded. Furthermore, compared with the PDB sequences, single substitutions or insertions were found for 440 amino acid sequences from PREFAB; these sequences were edited. SCOP domain analysis has shown that only 502 alignments in the resulting set contain sequences from the same family. Finally, eliminating duplicates, we have created a new golden standard alignment database PREFAB-P based on PREFAB; the PREFAB-P database contains 581 alignments.

Keywords: amino acid sequence, golden standard, PDB structure, SCOP classification.

DOI: 10.1134/S0006350912020030

1. REFERENCE ALIGNMENT DATABASES

Building amino acid sequence alignments is one of the key tools in bioinformatics, molecular biology and genomic analysis. Alignments are used in constructing phylogenetic trees and assessing their quality, finding characteristic motifs and conserved residues in protein families, composing domain profiles and solving many other tasks. For the user of such programs, along with their computational complexity (operation time and memory requirements), very important is the biological adequacy of the output alignments. Thus, usually the work of each new program or algorithm is evaluated through comparison with that of the already existing programs by two parameters: alignment quality and speed. For such analysis, one must have so-called reference alignments, i.e., alignments deemed to be the most biologically correct.

Initially (in the 1980s to mid-1990s) the authors themselves chose alignments as reference ones for assessing program performance, based on their own criteria (see, e.g., [1–5]), but as a rule such samples were small. To add, the use of a large number of various sets of reference alignments in assessing different algorithms made the comparison not quite convenient. Among the works of that period, we can mark out the paper of McClure et al. [6] published in 1994. The authors tested various methods of multiple sequence alignment for the ability to find conserved motifs in protein families of hemoglobin, kinase, ribonuclease H, and aspartate-specific protease. For all these families the biologically important motifs were already known and studied; i.e., known were the alignments of sequences belonging to each family. It is such alignments that were taken as the golden standard; each family got its reference alignment database (benchmark). On the basis of the results obtained, the authors concluded that global alignment algorithms seek for conserved motifs better than local alignment algorithms. However, at that time the number and size of the available reference alignment databases were quite limited, so the analysis could not be complete and the conclusion could not be substantiated enough.

Since then, the amount of data on alignments has significantly increased, and today there are many different independent reference alignment databases for amino acid sequences. However, despite the success in developing reference alignment databases [7], still open is the main problem of how much these alignments can be trusted and whether they can be taken as golden standard. Currently more and more works appear devoted to testing such databases, for example, on the basis of domain homology or correspondence with the protein secondary structure [8].

In the present work we analyze the reference alignment database PREFAB [9], in particular, supplement it with information on the homology of aligned sequences from the standpoint of SCOP [10].

1.1. Principles of building reference alignment databases. Structural classification of proteins. Contem-

Editor's Note: I certify that this text exactly reproduces all factual statements and closely conveys the phrasing and style of the original publication. *A.G.*

porary reference alignment databases are as a rule built on structural alignments of proteins, i.e., alignments involving superposition of spatial structures. However, some reference alignment databases (see below) include alignments based only on sequence analysis. Different databases are distinguished by the choice of protein families, the algorithms of structure superposition, the method of refining the algorithmic structural alignments, which is usually done by experts.

Although the sequence alignments taking into account the corresponding structural alignment is held to be the most correct from the biological viewpoint, such an approach has a number of limitations. Firstly, attention must be paid to that the resolution of structures be high enough (less than 3 Å), otherwise a structural alignment may turn out to be simply meaningless. Secondly, different programs for structural alignments may build different structural alignments of the same sequences [11, 12], and it is often difficult to determine which of them is the more correct one. Therefore in analysis of reference alignment databases this must be taken into account. Alignment of different secondary structures, for example α-helix and β-strand, is deemed to be certainly incorrect. Of course the existence of different classifications [13] and the quite frequent disagreement between methods of secondary structure prediction introduce sometimes hardly solvable problems, yet such assessment of correctness of a reference alignment based on comparison of secondary structures is used quite often.

The most popular method of evaluating reference alignment databases is determination of the type of fold of the aligned structural domains and subsequent comparison of these types, because it is inexpedient to align domains from different families. The most known classifications of structural domains are SCOP [10] and CATH [14]. They differ both in the way of domain definition (manual or automatic) and the very system of classification. In the CATH database the procedure of domain isolation is automatic, up to 2003 they were isolated by three algorithms: DOMAK [15], DETECTIVE [16], and PUU [17]. In the case of discrepancy between the results of the algorithms the decision was made by experts. Since 2003 the main method of domain isolation is CATHEDRAL [18, 19]. Its principle consists in searching for similar domains among the already isolated ones. If no similar domain is found, use is made of the old procedure.

In the SCOP database, domains are isolated only by experts, without participation of special programs. The upper levels of classification: Class, Fold, Superfamily, Family. Here distinguished are four main classes (all α, all β, α + β, α/β), and also several special ones (membrane, small domains, etc.). The domains to have a common fold, they must have the same main secondary structure elements identically positioned both in space and in the protein chain. Relation to a superfamily means obvious signs of common origin, while proteins belonging to one family must have not less than 30% sequence similarity or very close structure and functions.

In our opinion, SCOP is preferable because every domain is analyzed by experts rather than a program.

1.2. Overview of reference alignment databases. The most popular databases at the given moment are BAliBASE [20–23], PREFAB, HOMSTRAD [24], OXBench [25], SABmark [26]. The most recently created PREFAB, the main subject of our interest, is considered in detail in section 2.3. In this section we describe the other above-listed databases.

1.2.1. BAliBASE is one of the first databases of multiple reference alignments. The alignments presented here were obtained by structural superposition with subsequent manual checking of the correctness of alignment for conserved amino acid residues. BAliBASE consists of nine sections. Each section reflects a certain class of situations that may be encountered by a multiple alignment program. Examples of such situations: a small number of remote sequences; sequences with extensive nonhomologous N/C-terminal regions or with large internal inserts; alignment of transmembrane proteins, domains with repeats and inversions, and even linear motifs of eukaryotes. The current version of the database contains 217 alignments of 4 to 142 sequences.

1.2.2. HOMSTRAD. Clustering of this protein domain database relies on sequence and structure similarity. Although it was not initially conceived as a reference alignment, many authors use it as such. HOMSTRAD presents data on protein sequence and structure from various databases including PDB [27], Pfam [28], and SCOP. The latest version of HOMSTRAD includes 1032 domain families represented by 2 to 41 sequences, and also 9602 families with a single representative (one sequence).

1.2.3. OXBench contains multiple protein alignments built using methods of both structural alignment and sequence alignment. The database comprises three sections. The first, master section consists of 673 alignments of protein domains with known 3D structure, from 2 to 122 sequences in each. The second, extended section was obtained on the basis of the master one by adding sequences of unknown structure. The third, full-length section is also based on alignments from the master section, but the entire sequence is aligned rather than one domain.

1.2.4. SABmark contains reference pair alignments of sequences with known 3D structure. It consists of two sections: Twilight (sequences with pair similarity Blast E-value ≥1) and Superfamilies (sequences with pair identity ≤50%). Both sections in turn are partitioned into groups according to SCOP: by fold (for Twilight) and by superfamily (for Superfamilies). The reference alignment for each sequence pair from the group was built with structural alignment programs SOFI [29] and CE[30].

2. DATA AND METHODS

2.1. PREFAB. *2.1.1. General info and structure.* PREFAB (Protein Reference Alignment Benchmark) was composed by R. Edgar in 2004 for testing the performance of multiple alignment programs.

PREFAB contains:

(1) A set of reference pair alignments.

(2) Samples of sequences for testing multiple alignment programs.

(3) A program of assessing the quality of work of multiple alignment programs.

Below, only pair alignments are considered.

The latest version of PREFAB – PREFAB v.4.0 [31] was published by R. Edgar in March 2005. In contains 1682 reference pair alignments.

Each alignment is contained in a separate file of FASTA format (or, more exactly, FSSP FASTA). The file name has the form NAME1_NAME2, where NAME1, NAME2 are the name of aligned sequences. The name of each sequence is simply its PDB identifier (four symbols) or the PDB identifier and chain (five symbols) in cases when it is explicitly specified. The sequence names in PREFAB correspond to the names of their FSSP structures. According to the FSSP FASTA format, capital letter in the alignment itself mark aligned positions, while lowercase ones, unaligned. In assessing the quality of the algorithmically built alignment, only the aligned positions are taken into account.

2.1.2. Constructing pair reference alignments. The alignments entered into PREFAB were obtained as described in [9]. First, pair alignments were taken from the test databases constructed and described by Sadreyev and Grishin [32] and by Edgar and Sjolander [33, 34]. The alignments entered in these databases were retrieved from FSSP [35] and then realigned with the structural alignment program CE. After this, only those alignments were selected for which FSSP and CE agreed in more than 50 positions. It is these alignments that constitute the sample of reference alignments in PREFAB.

2.2. Preprocessing of PREFAB. The two main stages of preprocessing the PREFAB pair reference alignments are verification of sequences and verification of reference alignments. In the former case this means mutual comparison of a sequence from the PREFAB alignment and the corresponding PDB sequence. By the PDB sequence we mean the sequence of protein from the corresponding PDB entry, such that the coordinates are known for all its amino acid residues. By reference alignment verification we mean comparison of the types of structure of compared domains according to SCOP.

2.3. Bringing files to a unified template. As already said in 2.1.1, files in PREFAB are named as NAME1_NAME2, where NAME1 and NAME2 are the names of aligned sequences. However such a file name does not at all signify that the sequences in the file appear in the same order as in the file name. Whereas for some programs this may prove important. Therefore all files in PREFAB were brought to a unified template: the order sequences in the file name corresponds to the order of sequences in the file itself. In the course of performing this stage we revealed so-called duplicate files, i.e. files named NAME1_NAME2 and NAME2_NAME1. Such files contain practically identical alignments. Since it is still unknown which of such alignments could be regarded as the more correct one, they were left for further analysis. At the same step we conducted a check of the sequence names for containing obsolete identifiers as compared with the current version of PDB. All obsolete identifiers were replaced with new ones.

2.4. Sequence verification. *2.4.1. Assignment of unique identifier.* Regretfully, in PREFAB one and the same sequence name may denote different sequences—different fragments of one protein chain. To add, one and the same sequence may be encountered in different alignments. Therefore, in order to avoid errors connected with further analysis, each sequence is assigned a unique identifier of form NAME.ALIGN_NAME, where NAME is sequence name and ALIGN_NAME is the name of the alignment from which this sequence was taken. Sequences with such identifiers we will call PREFAB sequences.

2.4.2. Obtaining PDB sequences. For every PREFAB sequence, the respective document was retrieved from PDB, the sequence of the required chain was retrieved from the ATOM fields. Therewith all modified residues were replaced with usual ones, e.g. formylmethionine with methionine, monoisopropyl phosphoserine with serine. Selenomethionine, which is often used in X-ray analysis, was also changed to methionine.

2.4.3. Building alignment between PREFAB sequences and PDB sequences. Every PREFAB sequence was aligned with a corresponding PDB sequence (a global alignment was build). The following variants are possible:

(A) The alignments have no inserts.

(B) The alignments have boundary inserts in PREFAB.

(C) The alignments have internal inserts in PREFAB.

(D) The alignments have boundary inserts in PDB.

(E) The alignments have internal inserts only in PDB.

(F) The alignments have internal inserts both in PDB and in PREFAB.

If the alignment built for a PREFAB sequence and a PDB sequence satisfies cases A and D, i.e., contains no internal inserts (or gaps) in PDB and no inserts in PREFAB, we accept the PREFAB sequence for further analysis. Cases B and C are treated as misprints,

Results of sequence verification

Parameters	Number of PREFAB sequences	Number of PREFAB alignments
Single substitutions	440	345
Insert in PREFAB sequence	34	31
Deletions in PREFAB sequence	580	575

The PREFAB alignments containing such a sequence are edited. While if the alignment contains inserts in PDB (cases E and F), i.e. if PREFAB presents an incomplete sequence, the PREFAB sequence and the corresponding PREFAB alignment are removed. Any replacements in the alignment thus built are regarded as misprints, the PREFAB sequence in the PREFAB alignment is edited according to its PDB sequence.

2.5. Domain determination. As the source of domain classification, we took SCOP v. 1.75. For every PREFAB sequence, all possible SCOP domains of the given protein chain are determined. Further identification of the SCOP domain(s) consists in comparing the corresponding coordinates, i.e. the coordinates of the domain and PREFAB sequence according to the protein sequence. For each domain its overlap with the PREFAB sequence is calculated. The overlap is calculated as the length of intersection of the given SCOP domain and the PREFAB sequence divided by the length of the PREFAB sequence. If the overlap is greater than 0.95 (95%), it is taken that the PREFAB sequence is uniquely specified by the given SCOP domain. Domains for this the overlap equals zero are excluded from consideration. If there are several possible SCOP domains, then every domain is first considered separately, and if the sequence is not uniquely determined by one of the putative domains, then we consider the sum overlap of the remaining domains. Domains are accepted if it is greater than 0.95 (95%). If for a PREFAB sequence not a single SCOP domain is determined, then such sequence and the corresponding alignment are removed.

2.6. Verification of alignment. At this step selection of alignments takes place by means of comparing the SCOP domains of the sequences aligned. An alignment is regarded as having passed verification if the SCOP classification of the compared domains coincides to a family. In case if for one of the PREFAB sequences several domains are determined, there appears an additional condition of selection: the number of domains in the other sequence must be the same. If this is fulfilled, then the domains are compared pair-wise according to their position in the chain, i.e., first domain with first, second with second, and so on. The alignment is accepted if each pair of the compared domains belongs to one family.

2.7. Checking for duplicates. The last step of preprocessing consists in selecting duplicate files. From every pair of files, chosen is only the one where the values of overlap of compared sequences are larger. If the values coincide, accepted is the first of the files in the list.

RESULTS

In replacing obsolete PDB identifiers it turned out that one entry (1bef) had been recognized as low-quality and deleted from PDB. Two alignments that contained this sequence were removed.

Sequence verification has shown (see table) that a significant part of PREFAB sequences have missing internal fragments. Such incomplete protein sequences are encountered in 575 PREFAB alignments. We found 34 cases of the presence of an insert in PREFAB sequence, at that practically always this was one additional amino acid residue in the beginning or in the end of the sequence. These cases were regarded as misprints, the sequences were edited. For 440 PREFAB sequences we identified single mismatches with the respective PDB sequences. Interestingly, of all amino acids, those most often replaced in PREFAB were methionine and cysteine. And while the change of methionine in PREFAB to any amino acid designated by symbol "X" can be easily explained by that in the PDB entry in this position there is selenomethionine, which in the course of preprocessing we have replaced with methionine, whereas with cysteine (and with any other amino acid) the matter is more complicated. At that apart of replacement with any amino acid residue, there occurred cases when polar uncharged cysteine was replaced, for example, with nonpolar alanine. Upon more detailed inspection it turned out that most often the replacements took place in the case of a modified amino acid residue in the PDB entry.

Also disclosed were cases when in a PREFAB sequence there are regions for which the coordinates are unknown in the PDB entry, which appears quite strange, since the reference alignment with the given sequence had been built basing on alignment of structures. In determination of the SCOP domain it came to light that the 1mfa sequence consists at once of two protein chains (L and H), each of which contains its own SCOP domain. This sequence and the corresponding alignment (1mfa_1neu) were excluded from consideration.

SCOP domain were determined for every PREFAB sequence. Comparison of their classifications has shown that in 581 alignments, i.e. in 31.2% of the entire PREFAB, aligned are homologous sequences whose domains belong to one family. At that 502 of such alignments contain sequences that are determined by one SCOP domain.

In PREFAB v4.0 we have disclosed 61 pairs of duplicate files. Upon completion of preprocessing

there remained 22 such pairs, one alignment of each pair was selected at random.

The database obtained as the result of the above-described work, PREFAB-P is available at http://server2.lpm.org.ru/static/prefab-p/.

Additional materials: http://server2.lpm.org.ru/~irina/supplementary.rar.

CONCLUSIONS

The work presents analysis of the reference alignment database PREFAB, including determination of homology of the aligned sequences based on SCOP.

We have conducted PREFAB preprocessing and selected only those alignments the sequences of which are homologous to each other. It has been disclosed that some PREFAB alignments present sequences for which the SCOP classification diverges not only at the family level but also at higher levels, such as superfamily, fold and even class.

On the basis of the conducted analysis we have created the database PREFAB-P. The next step of the work will be the assessment of the reliability of separate elements of reference alignments.

ACKNOWLEDGMENTS

The work was performed under State Contract 07.514.11.4004, code 2011-1.4-514-008-009 supported by the Russian Ministry of Education and Science.

REFERENCES

1. R. F. Smith and T. F. Smith, Protein Eng. **5**, 35 (1992).
2. E. Deperieux, G. Baudoux, P. Briffeuil, et al., Comput. Appl. BioSci. **13**, 249 (1997).
3. S. R. Eddy, ISMB **3**, 114 (1995).
4. B. Morgenstern, A. Dress, and T. Werner, Proc. Natl. Acad. Sci. USA **93**, 12098 (1996).
5. J. D. Thompson, T. J. Gibson, F. Plewniak, et al., Nucl. Acids Res. **24**, 4876 (1997).
6. M. A. McClure, Vasi T. K., Fitch W. M., Mol. Biol. Evol. **11**, 571 (1994).
7. M. R. Aniba, O. Poch, and J. D. Thompson, Nucl. Acids Res. **38**, 7353 (2010).
8. R. C. Edgar, Nucl. Acids Res. **38**, 2145 (2010).
9. R. C. Edgar, Nucl. Acids Res. **32**, 1792 (2004).
10. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).
11. H. Hasegawa and L. Holm, Curr. Opin. Struct. Biol. **19**, 341 (2009).
12. A. Godzik, Protein Sci. **5**, 1325 (1996).
13. C. Etchebest, C. Benros, S. Hazout, and A. G. de Brevern, Proteins **59**, 810 (2005).
14. C. Orengo, A. Michie, S. Jones, et al., Structure **5**, 1093 (1997).
15. A. S. Siddiqui and G. J. Barton, Protein Sci. **42**, 372 (1995).
16. M. B. Swindells, Protein Sci. **4**, 103 (1995).
17. L. Holm and C. Sander, Proteins **19**, 256 (1994).
18. A. Harrison, F. Pearl, R. Mott, et al., J. Mol. Biol. **5** (323), 909 (2002).
19. F. M. Pearl, C. F. Bennett, J. E. Bray, et al., Nucl. Acids Res. **31**, 452 (2003).
20. J. D. Thompson, F. Plewniak, and O. Poch, Bioinformatics **15**, 87 (1999).
21. A. Bahr, J. D. Thompson, J. C. Thierry, and O. Poch, Nucl. Acids Res. **29**, 323 (2001).
22. J. D. Thompson, P. Koehl, R. Ripp, and O. Poch, Proteins **61**, 127 (2005).
23. E. Perrodou, C. Chica, O. Poch, et al., BMC Bioinformatics **9**, 213 (2008).
24. K. Mizuguchi, C. M. Deane, T. L. Blundell, and J. P. Overington, Protein Sci. **7**, 2469 (1998).
25. G. P. Raghava, S. M. Searle, P. C. Audley, et al., BMC Bioinformatics **4**, 47 (2003).
26. I. Van Walle, I. Lasters, and L. Wyns, Bioinformatics **21**, 1267 (2005).
27. H. M. Berman, K. Henrick, and H. Nakamura, Nat. Struct. Biol. **10**, 980 (2003).
28. R. D. Finn, J. Mistry, J. Tate, et al., Nucl. Acids Res. **38**, 211 (2010).
29. N. S. Boutonnet, M. J. Rooman, M. E. Ochagavia, et al., Protein Eng. **8**, 647 (1995).
30. I. N. Shindyalov and P. E. Bourne, Protein Eng. **11**, 739 (1998).
31. PREFAB v. 4.0: http://www.drive5.com/muscle/prefab.htm
32. R. Sadreyev and N. Grishin, J. Mol. Biol. **326**, 317 (2003).
33. R. C. Edgar and K. A. Sjolander, Bioinformatics, DOI: 10.1093/bioinformatics/bth090 (2004).
34. R. C. Edgar and K. Sjolander, Bioinformatics, DOI: 10.1093/bioinformatics/bth091 (2004).
35. L. Holm and C. Sander, Nucl. Acids Res. **26**, 316 (1998).