BIOINFORMATICS APPLICATIONS NOTE Vol. 15 no. 1 1999
Pages 87–88

BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs

Julie D. Thompson, Frédéric Plewniak and Olivier Poch

Institut de Génétique et de Biologie Moleculaire et Cellulaire (CNRS/INSERM/ULP), BP 163, 67404 Illkirch Cedex, France

Received on October 15, 1998; revised and accepted on November 18, 1998

Abstract

Summary: *BAliBASE is a database of manually refined multiple sequence alignments categorized by core blocks of conservation sequence length, similarity, and the presence of insertions and N/C-terminal extensions.*

Availability: *From http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/index.html*

Contact: *julie@igbmc.u-strasbg.fr*

The alignment of protein sequences is a crucial tool in molecular biology and genome analysis. Historically, the quality of new alignment programs has been compared to previous methods using a small number of test cases selected by the program author (e.g. Smith and Smith, 1992; Eddy, 1995; Morgenstern *et al*., 1996; Deperieux *et al*., 1997; Thompson *et al*., 1997). Recently, some comparisons have been made using a set of alignments selected from structural databases (Gotoh, 1996; Notredame *et al*., 1998). However, the alignment databases currently available (3D-ali, Pascarella *et al*., 1996; FSSP, Holm and Sander, 1998; HOMSTRAD, Mizuguchi *et al*., 1998; CAMPASS, Sowdhamini *et al*., 1998; Pfam, Sonnhammer *et al*., 1998; SMART, Schultz *et al*., 1998) assemble proteins into homologous families. The alignments are not structured and classified specifically for the systematic evaluation of multiple alignment programs.

A comprehensive evaluation and comparison of alignment programs requires a large number of accurate reference alignments which can be used as test cases. It has been shown (McClure *et al*., 1994) that the performance of alignment programs depends on the number of sequences, the degree of similarity between sequences and the number of insertions in the alignment. Other factors may also affect alignment quality, such as the length of the sequences, the existence of large insertions and N/C-terminal extensions, and over-representation of some members of the protein family. We have constructed BAliBASE (Benchmark Alignment dataBASE) containing high-quality, documented alignments to identify the strong and weak points of the numerous alignment programs now available.

A frequent problem encountered when using reference alignments has been the effect of ambiguous regions in the sequences which cannot be structurally superposed. Very distantly related sequences often have only short conserved motifs in long regions of low overall similarity. These regions can only be aligned arbitrarily in the reference and may lead to a bias in the comparison of programs. In BAliBASE, we have annotated the core blocks in the alignments that only include the regions that can be reliably aligned.

The sequences included in the database are selected from alignments in either the FSSP or HOMSTRAD structural databases, or from manually constructed structural alignments in the literature. When sufficient structures are not available, additional sequences are included from the HSSP database (Schneider *et al*., 1997). The VAST Web server (Madej *et al*., 1995) is used to confirm that the sequences in each alignment are structural neighbours and can be structurally superimposed. Functional sites are identified using the PDBsum database (Laskowski *et al*., 1997) and the alignments are manually verified and adjusted, in order to ensure that conserved residues are aligned as well as the secondary structure elements. Alignments are also verified using the SAP protein structure comparison program (Taylor, 1998).

BAliBASE (Version 1.0) consists of 142 reference alignments, containing >1000 sequences. Of the 200 000 residues in the database, 58% are defined within the core blocks. The remaining 42% are in ambiguous regions that cannot be reliably aligned. Figure 1B shows the WWW display of a typical alignment. The alignments are divided into four hierarchical reference sets, reference 1 providing the basis for construction of the following sets. Each of the main sets may be further subdivided into smaller groups, according to sequence length and per cent similarity. The alignments are provided in either RSF format or MSF format with an attached text file containing a list of the known secondary structure elements for each sequence. Each alignment is also associated with an annotation file containing a description of the alignment.

In the future, new reference sections could be added to BaliBASE to include other criteria that affect the performance of alignment programs, particularly transmembrane proteins and other sequences showing a compositional bias.

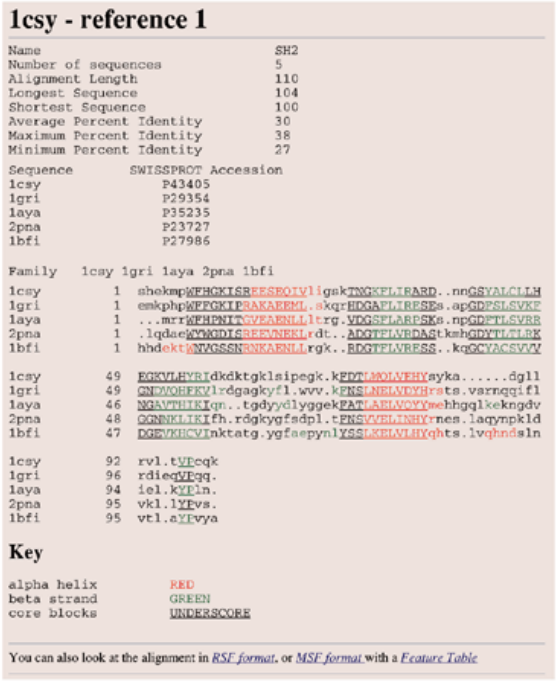
A comparison of alignment programs (work in progress) suggests that not all of the alignment algorithms react in the same way to the problems presented in the BAliBASE alignments. The results of the study should allow users to select the

A

Reference 1	<100 residues	200<300 residues	>500 residues
<25% identity	7	8	8
20-40% identity	10	9	10
>35% identity	10	10	8
Reference 2	9	8	7
Reference 3	5	3	5

	Extensions	Insertions
Reference 4	12	12

B



1csy - reference 1

```
Name                          SH2
Number of sequences           5
Alignment Length              110
Longest Sequence              104
Shortest Sequence             100
Average Percent Identity      30
Maximum Percent Identity      38
Minimum Percent Identity      27

Sequence       SWISSPROT Accession
1csy               P43405
1gri               P29354
1aya               P35235
2pna               P23727
1bfi               P27986

Family   1csy 1gri 1aya 2pna 1bfi

1csy         1  shekmpWFHGKISREESEQIVligskTNGKFLIRARD..nnGSYALCLLH
1gri         1  emkphpWFFGKIPRAKAEEML.skqrHDGAFLIRESEs.apGDFSLSVKF
1aya         1  ...mrrWFHPNITGVEAENLLtrg.VDGSFLARPSKs.npGDFTLSVRR
2pna         1  .lqdaeWYWGDISREEVNEKLrdt..ADGTFLVRDAStkmhGDYTLTLRK
1bfi         1  hhdektWNVGSSNRNKAENLLrgk..RDGTFLVRESS..kqGCYACSVVV

1csy        49  EGKVLHYRIdkdktgklsipegk.kFDTLWQLVEHYsyka......dgll
1gri        49  GNDVQHFKVlrdgagkyfl.wvv.kFNSLNELVDYHrats.varnqqifl
1aya        46  NGAVTHIKIqn..tgdyydlyggekFATLAELVQYYmehhgqlkekngdv
2pna        48  QGNNKLIKIfh.rdgkygfsdpl.tFNSVVELINHYrnes.laqynpkld
1bfi        47  DGEVKHCVInktatg.ygfaepynlYSSLKELVLHYqhts.lvqhndsln

1csy        92  rvl.tYPcqk
1gri        96  rdieqVPqq.
1aya        94  iel.kYPln.
2pna        95  vkl.lYPvs.
1bfi        95  vtl.aYPvya

Key

alpha helix          RED
beta strand          GREEN
core blocks          UNDERSCORE

You can also look at the alignment in RSF format, or MSF format with a Feature Table
```

Fig. 1. (**A**) Current status of the database, showing the number of alignments in each reference set. Reference 1 contains alignments of (<6) equi-distant sequences. Reference 2 aligns up to three 'orphan' sequences (<25% identical) from reference 1 with at least 15 closely related sequences. Reference 3 consists of up to four subgroups, with <25% residue identity between groups. Reference 4 contains sequences with N/C-terminal extensions (up to 400 residues) or insertions (up to 100 residues). (**B**) The Web page of an alignment.

most suitable program depending on the set of sequences to be aligned, thus improving the accuracy of the automatic alignment and reducing the manual refinement required to obtain the final, optimal alignment.

Acknowledgements

We would like to thank M.Bergdoll, L.Moulinier and J.-M.Wurtz for their structural alignments D.Moras and J.-C.Thierry for their support during this work, and F.Jeanmougin and T.Gibson for useful discussions. The work was supported by institute funds from INSERM, CNRS, H.U.S. and Bristol-Myers Squibb.

References

Deperieux,E., Baudoux,G., Briffeuil,P., Reginster,I., De Bolle,X., Vinals,C. and Feytmans,E. (1997) MATCH-BOX server: a multiple sequence alignment tool placing emphasis on reliability. *Comput. Applic. Biosci.*, **13**, 249–256.

Eddy,S.R. (1995) Multiple alignment using hidden Markov models. *ISMB*, **3**, 114–120.

Gotoh,O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.*, **264**, 823–838.

Holm,L. and Sander,C. (1998) Touring protein fold space with DALI/FSSP. *Nucleic Acids Res.*, **26**, 316–319.

Laskowski,R.A., Hutchinson,E.G., Michie,A.D., Wallace,A.C., Jones,M.L. and Thornton,J.M. (1997) PDBsum: A Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.*, **22**, 488–490.

Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **3**, 356–369.

McClure,M.A., Vasi,T.K. and Fitch,W.M. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.*, **11**, 571–592.

Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, in press.

Morgenstern,B., Dress,A. and Werner,T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl Acad. Sci. USA*, **93**, 12098–12103.

Notredame,C., Holm,L. and Higgins,D.G. (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **5**, 407–422.

Pascarella,S., Milpetz,F. and Argos,P. (1996) A databank (3D-ali) collecting related protein sequences and structures. *Protein Eng.*, **9**, 249–251.

Schneider,R., de Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.

Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.

Smith,R.F. and Smith,T.F. (1992) Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng.*, **5**, 35–41.

Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **24**, 17–20.

Sowdhamini,R., Burke,D.F., Huang,J., Mizuguchi,K., Hampapathalu, Nagarajaram,A., Srinivasan,N., Steward,R.E. and Blundell,T.L. (1998) CAMPASS: a database of structurally aligned protein superfamilies. *Structure*, **6**, 1087–1094.

Taylor,W.R. (1998) Protein structure comparison using iterated double dynamic programming. *Protein Sci.*, in press.

Thompson,J.D., Gibson,T.J, Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence aligment aided by quality analysis tools. *Nucleic Acids Res.*, **24**, 4876–4882.