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.


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
**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**


