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

Julie D. Thompson, Fridiric Plewniak and Olivier Poch.
Institut de Ginitique et de Biologie Moliculaire et Cellulaire,
(CNRS/INSERM/ULP), B.P. 163, 67404 Illkirch Cedex, France.
E-mail : julie@igbmc.u-strasbg.fr

In recent years improvements to existing alignment programs and the introduction of new iterative algorithms have changed the state-of-the-art in multiple sequence alignment. In spite of the wide variety of alignment programs now available, there have been few comparisons of the relative performance and reliability.

BAliBASE (1) is a database of manually-refined multiple sequence aligment specifically designed for the evaluation and comparison of multiple alignment programs. The reference alignments are categorised by core blocks of conservation, sequence length, similarity, and presence of insertions and N/C-terminal extensions. The sequences included in the database are selected from structural databases or from manually constructed alignments in the literature. The alignments are manually verified and adjusted, to ensure that conserved residues are aligned as well as the secondary structure elements.

BAliBASE (version 1.0) consists of 142 reference alignments, containing over 1000 sequences. The alignments are divided into 4 hierarchical reference sets, reference 1 providing the basis for construction of the following sets. Each of the main sets may be further sub-divided into smaller groups, according to sequence length and percent similarity. Reference 1 contains alignments of (less than 6) equi-distant, similar length sequences. Reference 2 aligns up to three "orphan" sequences (<25% identical) with at least 15 closely related sequences. Reference 3 consists of up to 4 sub-groups, with less than 25% residue identity between sub-groups. Reference 4 contains alignments including N/C-terminal extensions (up to 400 residues), or internal insertions (up to 100 residues).

A comparison of alignment programs (2) using the BAliBASE alignments has shown that the choice of an alignment program depends on the sequence set to be aligned, and no single *best9 program exists. It was demonstrated that the twilight zone still exists as a real barrier for all of the programs in the study, but that the best programs were still capable of aligning on average 47% of the residues below the twilight zone at 10-20% residue identity. The recently developed iterative programs often offered improved alignment accuracy although a heavy time penalty was incurred. A notable exception was the effect of introducing a single divergent sequence into a set of closely related sequences, causing the iteration to diverge away from the best alignment. Global alignment algorithms generally performed better than local methods except in the presence of large terminal extensions and internal insertions. In these cases, a local algorithm was more successful in identifying the most conserved motifs. Global programs which tend to favour a collinear alignment of the entire lengths of the sequences were less successful, often producing a total misalignment. None of the alignment programs in the study were capable of producing good, reliable alignments in all of the BAliBASE reference tests.

The results of the study should allow users to select the 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. The results also indicate guidelines for the future development of multiple alignment programs.

References

1. Thompson, J.D., Plewniak, F. and Poch,O. (1999) Bioinformatics 1, 87-88.
2. Thompson, J.D., Plewniak, F. and Poch,O. (1999) Nucl. Acids Res., (in press)