Combining structure and function information in a local alignment search tool for sequence-sequence comparison.

Maricel Kann[1], Paul Thiessen[1], Anna Panchenko[1], Alejandro Schaffer[1], Stephen F. Altschul[1] and Stephen H. Bryant[1].

Keywords: protein sequence alignment, alignment algorithms, statistical significance, similarity search.

1 Introduction.

With thousands of recently sequenced proteins, sequence-sequence comparison methods have become the most widely used tool in bioinformatics and related fields. Several popular methods (i.e., PSI-BLAST and IMPALA) [1, 2] implement the search using a position-specific scoring matrix (PSSM) as a scoring scheme, which captures important information about the protein sequence, and achieves a fast and effective search for sequence homologies. However, explicit information about the structure and function of proteins in the databases during the search is difficult to include. One could incorporate such information by modeling the indels, represented by gaps in the alignment, and/or by correctly aligning all biologically relevant residues. In most of the algorithms for sequence comparison, however, the choice of the gap penalties is determined ad-hoc, gaps occur anywhere, and key residues are aligned only if this increases the total score for the alignment. In this paper, we introduce an algorithm, Structure-based Local Algorithm Method or SLAM; that uses a new approach for the placements and penalization of gaps with a novel approach for the definition of aligned regions. This algorithm produces a local alignment between a query protein sequence and a database of PSSMs, and uses a scoring scheme based on the differences in conservation along the protein sequences. Highly-conserved regions (or blocks) will be aligned from start to end without any gaps and without penalties for the insertion of gaps (up to a maximum length) between those blocks. For the classification of the families, definition of the blocks or conserved domains (CDs) and PSSMs, we use an approach developed by our group in which multiple sequence alignment of related sequences have been manually curated to represent the function of that family. The scores obtained using SLAM follow an extreme value distribution which allows the correct estimation of their statistical significance. The fact that CD database have been manually curated with key residues necessary for the specific function of each of the families and inter-block regions carefully selected, suggests that SLAM's alignments are highly biologically significant. SLAM's performance in the search for biological relationships, alignment accuracy and speed is very similar to IMPALA's. Based on these results, SLAM has been successfully included into Cn3D (tool for visualization of protein structures) as an alternative choice for the alignment of new sequences.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 17: pp. 3389-3402.

[1] Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.

E-mail: kann@mail.nih.gov

[2] A. A. Schaffer, Y. I. Wolf, C. P. Ponting, E. V. Koonin, L. Aravind and S. F. Altschul. 1999. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics. 12: pp. 1000-1011.