Ballast : a motif oriented Blast results post-processing program

F. Plewniak, J.D. Thompson & O. Poch

Institut de Génétique et de Biologie Moléculaire et Cellulaire (CNRS/INSERM/ULP)
BP 163, 67404 Illkirch Cedex, France.
plewniak@igbmc.u-strasbg.fr

In the process of genome annotation and protein function prediction the Blast[1] programs are widely used to search protein and nucleotide databases for similarity or homologs of query sequences. Though Blast programs are very efficient in finding relatively strong similarities, distantly related sequences scoring poorly are usually difficult to tell from the background of non-related ones. This problem is even more sensible in TBlastN searches and will become more challenging as genome portions and full cosmids are being incorporated into the databases.

A number of post-processing programs (BEAUTY[2], PowerBlast[3], MSPCrunch[4],...) have been developped to help the interpretation of Blast results but none actually examines the High-scoring Segment Pairs (HSPs) within a multiple sequence context to evaluate the quality of the proposed similarities. PsiBlast[5] on the other hand, iteratively builds a profile from aligned best HSPs and uses it to search the protein database until it converges and no new similar sequence is found. PhiBlast[6] limits the rate of false positives by highlighting those regions of homology that share *a priori* known functional motifs.

Here, we present Ballast, a Blast results post-processing program which extracts as much information as possible from stacked BlastP HSPs to help the interpretation of both BlastP and TBlastN searches at the same time. Ballast aligns the best HSPs with the query sequence and computes a profile of aminoacid conservation along the query sequence. It then suggests potential functional motifs which are used to re-estimate the quality of each HSP in both BlastP and TBlastN results. HSPs found within conserved regions are emphasized, thus reducing the rate of false positives, especially in TBlastN searches, while accentuating the functionally important regions. Tests on the BAliBASE[7] benchmark alignments database showed that the conservation profile and the predicted motifs are accurate even when very few sequences are available. We also demonstrate the Ballast ability to highlight distant relations which are not identified clearly by Blast. Furthermore, as it is not iterative, Ballast is very efficient in terms of CPU usage and is therefore particularly suitable for an integration into large scale projects such as automated genome annotation.

References :

1) S. F. Altschul, W. Gish, W. Miller, E. W. Myers & D. J. Lipman (1990), *Basic local alignment search tool.* J Mol Biol 215: 403-410
2) K. C. Worley, B. A. Wiese & R. F. Smith (1995) *BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results*. Genome Res 5: 173-184
3) J. Zhang & T. L. Madden (1997) *PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation*. Genome Res 7: 649-656
4) E. L. Sonnhammer & R. Durbin (1994), *A workbench for large-scale sequence homology analysis*. Comput Appl Biosci 10: 301-307
5) S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller & D. J. Lipman (1997), *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res 25: 3389-3402
6) Zhang, Zheng, Alejandro A. Schäffer, Webb Miller, Thomas L. Madden, David J. Lipman, Eugene V. Koonin, and Stephen F. Altschul (1998), *Protein sequence similarity searches using patterns as seeds*, Nucleic Acids Res. 26:3986-3990.
7) J. D. Thompson, F. Plewniak & O. Poch (1999), *BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs.* Bioinformatics 15: 87-88