Poster H-66

Analyzing the effects of generalizations implicit within the BLAST algorithm

Authors:
Mileidy Gonzalez (*University of Maryland Baltimore County*)
Stephen J. Freeland (*University of Maryland Baltimore County*)

Short Abstract: We have mapped the entire network of simplifying mathematical assumptions used by the BLAST algorithm (the most widely-used pairwise sequence comparison method) to identify the properties of unusual, natural sequences that could compromise the program's reliability. We present preliminary data that evaluates the magnitude of potential problems.

Long Abstract:
Biological sequence comparison is one of the most widely used techniques of modern biology. In particular, because this method can be used to make quantitative estimates of whether and how two sequences are homologous, its use is implicit within many fundamental bioinformatics techniques (e.g. phylogenetic tree construction, genome annotation, threading, protein family assembly, etc.). The underlying algorithm has been developed over the course of sixteen years and has gone through one major conceptual change [Altschul et al 1990, Altschul et al 1997]. Thus, although the use of local pairwise alignment algorithms such as BLAST is extremely widespread ("The number of BLAST queries sent to the server continues to increase, growing from about 100 000 per weekday at the beginning of 2002 to about 140 000 per weekday in early 2004."[McGinnis and Madden 2004]), performance has rarely been quantified other than by the developers themselves [Altschul et al 1990]. This history has left several key generalizations, implicit within the BLAST algorithm, largely untested for their magnitude of effect. For example, from a biological perspective we know that the parameters which influence protein evolution (particularly mutational bias [Sueoka 1988] and biochemical similarity of amino acids [Benner et al 1994]) vary considerably between different lineages. Implicitly, then, the quantitative patterns that describe homology will likewise vary (indeed, many researchers have documented that biased amino acid composition affects BLAST performance [Aoki et al 2005, Bastien et al 2005, Muller et al 2005, Ng et al 2000, Yu et al 2005]). However, no published evaluations of BLAST (or other local alignment programs, such as FASTA [Lipman and Pearson 1985]) have quantified the potential for this variation to influence the results of a search. In other words, most users simply know that “The E-value gives an indication of the statistical significance of a given pairwise alignment and reflects the size of the database and the scoring system used” [NCBI handbook]; they cannot know the extent to which such unusual properties of their sequences influence the derivation of E-values within a specific BLAST search.
To address this knowledge gap, we mapped the entire network of mathematical functions used by the BLAST algorithm to derive E-values. From this, we highlight the multiple points where key simplifying assumptions (e.g. of standard amino acid composition, alignment length, etc.) are embedded within BLAST’s derivation of significance scores. We use this knowledge to inform a series of tests, using query sequences with specific properties, to measure the effects produced by variations in these assumptions. Our method uses

standard, quantitative measures of BLAST performance [Pearson 1995] to explore the presence or absence of a correlation between amino acid composition and BLAST performance. We then quantify amino acid bias in various protein databases and combine this with our previous results to discuss the potential importance of deviations from BLAST's assumptions. We conclude that published evaluations of BLAST performance [Pearson 1995, Brenner et al 1998] are potentially biased toward overestimating software efficiency. Finally, we identify the specific properties of searches that will most likely compromise BLAST search efficiency, and provide a preliminary formula to estimate the effect for any given query sequence.

While these preliminary findings focus on evaluating the potential for 'unreliability' of results, they also allow us to start thinking about possible solutions that would minimize the limitations of current sequence comparison methods. Thus, the preliminary study that we present forms the first step in a program of research geared toward empowering the scientific user community to conduct more informed searches and to improve current methods of homology search performance evaluation.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool, J Mol Biol, 215, 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res, 25, 3389-3402.

McGinnis, S. and Madden, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools, Nucleic Acids Res, 32, W20-25

Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution, Proc Natl Acad Sci U S A, 85, 2653-2657.

Benner, S.A., Cohen, M.A. and Gonnet, G.H. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences, Protein Eng, 7, 1323-1332.

Aoki, K.F., Mamitsuka, H., Akutsu, T. and Kanehisa, M. (2005) A score matrix to reveal the hidden links in glycans, Bioinformatics, 21, 1457-1463.

Bastien, O., Roy, S. and Marechal, E. (2005) Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions, C R Biol, 328, 445-453.

Muller, T., Rahmann, S. and Rehmsmeier, M. (2001) Non-symmetric score matrices and the detection of homologous transmembrane proteins, Bioinformatics, 17 Suppl 1, S182-189.

Ng, P.C., Henikoff, J.G. and Henikoff, S. (2000) PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane, Bioinformatics, 16, 760-766.

Yu, Y.K. and Altschul, S.F. (2005) The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions, Bioinformatics, 21, 902-911.

Lipman, D.J. and Pearson, W.R. (1985) Rapid and sensitive protein similarity searches, Science, 227, 1435-1441

. The NCBI Handbook http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook

Pearson, W.R. (1995) Comparison of methods for searching protein sequence databases, Protein Sci, 4, 1145-1160.

Brenner, S.E., Chothia, C. and Hubbard, T.J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, Proc Natl Acad Sci U S A, 95, 6073-6078.