Bioinformatics Advance Access published February 27, 2009

PSIBLAST_PairwiseStatSig: Reordering PSI-BLAST Hits Using Pairwise Statistical Significance

Ankit Agrawal * and Xiaoqiu Huang

Dept. of Computer Science, Iowa State University, 226 Atanasoff Hall, Ames, IA 50011-1041, USA.

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Summary: We present an add-on to BLAST and PSI-BLAST programs to reorder their hits using pairwise statistical significance. Using position-specific substitution matrices to estimate pairwise statistical significance has been recently shown to give promising results in terms of retrieval accuracy, which motivates its use to refine PSI-BLAST results, since PSI-BLAST also constructs a position-specific substitution matrix for the query sequence during the search. The obvious advantage of the approach is more accurate estimates of statistical significance because of pairwise statistical significance, along with the advantage of BLAST/PSI-BLAST in terms of speed.

Availability: The implementation as a C library is freely available at www.cs.iastate.edu/∼ankitag/PSIBLAST_PairwiseStatSig.html

Contact: ankitag@cs.iastate.edu

1 INTRODUCTION

Database search is one of the most important applications of pairwise sequence alignment. The most popular heuristic-based methods for database search are the BLAST and PSI-BLAST programs (Altschul *et al.*, 1997). PSI-BLAST uses an iterative approach to BLAST using position-specific substitution matrices which are refined with every iteration, and its performance can be significantly better than BLAST. Another slightly slower but more accurate database search program than BLAST is FASTA (Pearson, 2000), which also employs heuristics to obtain a sub-optimal alignment. There also exists the SSEARCH program, which uses the full implementation of the Smith-Waterman algorithm (Smith and Waterman, 1981). Although more accurate, it can take many hours to days for a modest database search.

The hits of a database search are ranked according to statistical significance of the alignment scores rather than by alignment score themselves. An alignment score is considered statistically significant if it has a low probability of occurring by chance. The alignment score distribution (and hence statistical significance) depends on various factors like alignment program, scoring scheme, sequence lengths, sequence compositions (Mott, 2005). Accurate estimation of statistical significance of alignment scores is an important aspect of sequence comparison.

The methods to estimate the statistical significance of a pairwise alignment can be categorized into two primary methods. The statistical significance of the hits reported by database search programs is called database statistical significance, which is in general dependent on the size and composition of the database being searched. An alternative method to estimate statistical significance of a pairwise alignment independent of any database is to estimate pairwise statistical significance, which uses statistical parameters specific to the sequence-pair to estimate statistical significance.

In the last few years there have been considerable improvements to the BLAST and PSI-BLAST programs (Schäffer *et al.*, 2001; Yu and Altschul, 2005; Yu *et al.*, 2006), which have been shown to improve database search performance by using composition-based statistics and substitution matrix rescaling techniques, together with pre-computed statistical parameters for a wide range of alignment parameters. Recently, a study of pairwise statistical significance was conducted (Agrawal *et al.*, 2008). It compared various approaches to find that maximum likelihood fitting of an empirical distribution with censoring left of peak is most accurate for estimating pairwise statistical significance. Further, using position-specific substitution matrices to estimate pairwise statistical significance (Agrawal and Huang, 2008) gives the best results in terms of retrieval accuracy since it uses maximal sequence-specific information. Relevant details on pairwise statistical significance can be found in the supplementary notes.

2 PROPOSED APPROACH

The advantage of using position-specific substitution matrices (PSSMs) with pairwise statistical significance strongly motivates its use to refine PSI-BLAST results, since PSI-BLAST naturally constructs a PSSM for the query sequence, which can be used for estimating pairwise statistical significance. In this application note, we present an add-on to the BLAST and PSI-BLAST programs to refine their results using pairwise statistical significance. The proposed approach is implemented as a program named PSIBLAST_PairwiseStatSig which takes a query sequence, a database, the PSI-BLAST output file, and the PSI-BLAST constructed PSSM (if available), and gives the new pairwise statistical significance estimates.

To evaluate PSIBLAST_PairwiseStatSig, we used the same benchmark database (a non-redundant subset of CATH2.3 database of 2771 sequences, and its subset of 86 query sequences) as earlier used in (Sierk and Pearson, 2004; Agrawal *et al.*, 2008; Agrawal and Huang, 2008). For refining BLAST results, the BLSOUM62 matrix was used for the alignments as it is the default substitution matrix

*to whom correspondence should be addressed

for the BLAST program. For refining PSI-BLAST results, the PSI-BLAST constructed PSSM was used. To further take advantage of PSSMs, non-conservative pairwise statistical significance was also estimated (see supplementary notes). Note that for non-conservative pairwise statistical significance estimation with PSSMs, we would need PSSMs for both the sequences being aligned. But in general, after a PSI-BLAST run, we get a PSSM for only the query sequence, and not for the hits obtained. Therefore, here we use the standard substitution matrix BLOSUM62 instead of PSSM for second sequence, hoping that the PSSM for query sequence is significantly different from BLOSUM62 to take advantage of non-conservative pairwise statistical significance. The number of shuffles N was set to 1000.

The two evaluation methodologies used to compare the results are explained in detail in the supplementary notes. Here we only present the results using the standard methodology (earlier used in Brenner *et al.*, 1998; Sierk and Pearson, 2004) due to limited space. Fig. 1 shows the Error Per Query vs. Coverage curves for BLAST and PSI-BLAST with and without reordering their hits using pairwise statistical significance, depicting the improvement in performance using PSIBLAST_PairwiseStatSig (a curve more towards the right is better). The PSIBLAST_PairwiseStatSig program is tested to work with BLAST/PSI-BLAST output files for BLAST 2.2.17, but is expected to work for other versions as well.

Assuming that BLAST/PSI-BLAST output file is already available, the running time of the proposed method is dependent on the number of hits given by BLAST/PSI-BLAST. For a single sequence-pair, the pairwise statistical significance estimation time depends on the length of the two sequences. For typical protein sequence lengths (248 and 255), it took 0.45s to estimate pairwise statistical significance on a 2.8 GHz Intel processor.

An obvious disadvantage of the proposed approach is that its performance is upper-bounded by the number of true homologs detected by BLAST/PSI-BLAST. It can only reorder the hits with an attempt to rank the true homologs higher, but cannot recover any more homologs. However, considering this limitation, PSIBLAST_PairwiseStatSig has been demonstrated to give better results than BLAST and PSI-BLAST just by reordering the hits using pairwise statistical significance. It is also important to note that the proposed method is studied in context of protein database searches and not DNA, which may require substantial modification considering the arbitrary lengths of the DNA sequences.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Volker Brendel for helpful discussions and providing links to the data. Special thanks are due to the anonymous reviewers for their insightful comments, which made the manuscript stronger.

REFERENCES

Agrawal, A. and Huang, X. (2008). Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. under review.

Agrawal, A., Brendel, V., and Huang, X. (2008). Pairwise statistical significance and empirical determination of effective gap opening penalties for protein local sequence alignment. *IJCBDD*, **1**(4), 347–367.

Altschul, S. F., Madden, T. L., Schäffer, 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 Research*, **25**(17), 3389–3402.

Brenner, S. E., Chothia, C., and Hubbard, T. J. P. (1998). Assessing Sequence Comparison Methods with Reliable Structurally Identified Distant Evolutionary Relationships. *PNAS, USA*, **95**(11), 6073–6078.

Mott, R. (2005). Alignment: Statistical Significance. *Encyclopedia of Life Sciences*. doi.wiley.com/10.1038/npg.els.0005264.

Pearson, W. R. (2000). Flexible Sequence Similarity Searching with the FASTA3 Program Package. *Methods in Molecular Biology*, **132**, 185–219.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., and Altschul, S. F. (2001). Improving the Accuracy of PSI-BLAST Protein Database Searches with Composition-based Statistics and Other Refinements. *Nucleic Acids Research*, **29**(14), 2994–3005.

Sierk, M. L. and Pearson, W. R. (2004). Sensitivity and Selectivity in Protein Structure Comparison. *Protein Science*, **13**(3), 773–785.

Smith, T. F. and Waterman, M. S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, **147**(1), 195–197.

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**(7), 902–911.

Yu, Y.-K., Gertz, E. M., Agarwala, R., Schäffer, A. A., and Altschul, S. F. (2006). Retrieval Accuracy, Statistical Significance and Compositional Similarity in Protein Sequence Database Searches. *Nucleic Acids Research*, **34**(20), 5966–5973.

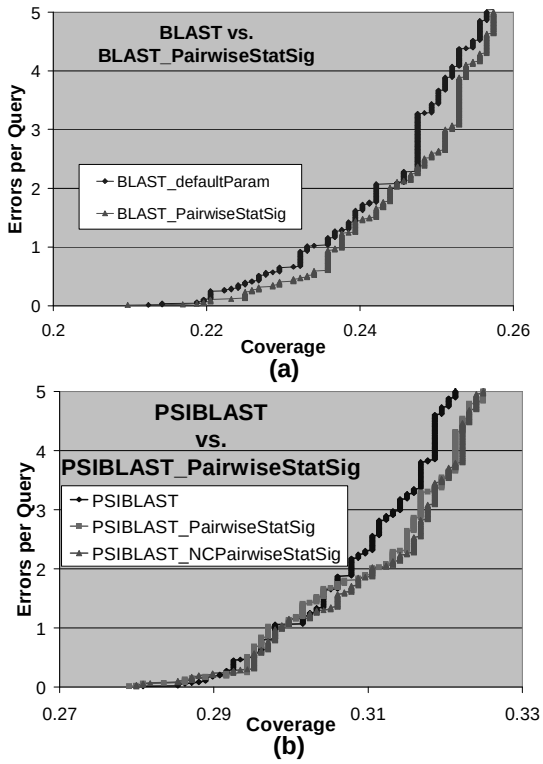

Fig. 1. Errors per Query vs. Coverage plots comparing the performance of (a) BLAST and BLAST_PairwiseStatSig (reordering BLAST results using pairwise statistical significance), and (b) PSIBLAST, PSIBLAST_PairwiseStatSig, and PSIBLAST_NCPairwiseStatSig (reordering PSI-BLAST results using non-conservative pairwise statistical significance). Reordering BLAST/PSI-BLAST hits using pairwise statistical significance leads to superior performance in terms of retrieval accuracy.