eBLAST: Building a Better BLAST

A. Pearson, H. Peri, O. Jabado, O. Wood

Abstract

The homology detection that is conducted by sequence comparison algorithms enables biologists to characterize unknown proteins by their similarity to ones that have been described. Although BLAST is an invaluable resource, the alignments it produces are heuristically determined. A more mathematically rigorous alignment such as Smith-Waterman would be desirable, though computationally unfeasible for genome size searches. We have constructed a new algorithm, eBLAST, which improves the homology detection of BLAST by incorporating Smith-Waterman alignments.

Introduction

DNA contains the biological template for the development and maintenance of organisms. The primary sequence of DNA provides enough information to direct protein expression and regulation, in concert with existing proteins in the cell. Transcribed proteins in the cell have functions determined by three-dimensional structure, which in turn is determined by their DNA sequence. Unfortunately, the ability to predict structure from nucleotide sequence is very difficult. This occurs because there are several significant biological modifications that occur between the translation, transcription and final operation of proteins in the cell. The ability to characterize the functional significance of DNA sequences is a fundamental task of computational biology.

Instead of attempting to predict function directly from sequences, most biological researchers compare DNA or protein sequences from genes with known functions to unknown sequences. Generally, this is termed homology detection and occurs by alignment of the sequences according to some scoring scheme. This way, unknown proteins can be characterized by their relatedness to known proteins (Durbin 1998). This has biological relevance because proteins generally have short stretches that are critical to the function and tend to be evolutionarily conserved across species. Examples would be binding pockets for enzymes or transmembrane domains for cell signaling receptors, which can be found with high similarity across species (Alberts 1994).

This kind of alignment based homology detection can be accomplished by a variety of methods. Currently, online tools exist using pairwise sequence comparison with scoring methods for amino acid or nucleotide searches, others use profile or position specific weight matrixes. Additionally, Hidden Markov Models (HMMS) are also in use (Jaakkola et al., 1999). Our investigation centers on the most popular online search tool, BLAST at the NCBI (Altschul 1997). The BLAST algorithm attempts to quickly align a query sequence with a multiple genome size database, while computing an index of similarity. It is actually a heuristic and does not find the provably optimum alignment between sequences, but one that is probably close.

The core of the heuristic is that similar sequences will probably have at some point a short stretch of identities. BLAST attempts to find short identical matches and use them as ‘seed’ to find high scoring extensions (Altschul 1997, Durbin 1998). BLAST then compares the scores of the alignments to a cutoff value and reports the significant alignments. The new version of BLAST creates gapped alignments, alignments between different length sequences, for which rigorous statistical evaluations have not been developed (Altschul 1997). Accordingly, the E score is only a rough measure of relatedness.

The actual calculation has three parts. The alignment is done by dynamic programming and given a particular score based on substitution matrices of DNA or amino acids. Then, the probability of that alignment score occurring by chance is factored into the E score. Finally, the scores are normalized into bits to allow easy comparison. In addition the algorithm also considers the chance of a query sequence, length n, having an alignment in a database of length N. BLAST accounts for this by multiplying N/n to the final E score, thus correcting for the high probability of matching a small sequence in a large database (Altschul et al., 1997). The main advantage of using BLAST is the speed at which it returns results. The disadvantage is that BLAST may miss remote homologies and may not find the maximum scoring alignment, both of which are important concerns to biologists looking to characterize unknown sequences which may only be marginally related to known families (Pearson 2000).

The more rigorous approach to alignment is the Smith-Waterman dynamic programming algorithm (Durbin et al., 1998, Pearson 2000). It computes the maximum local gapped alignment for two sequences, given a substitution matrix by which to weigh mismatches and matches. The Smith-Waterman is the most attractive algorithm to use from a mathematical perspective, because it provably finds the optimal gapped alignment, in contrast to BLAST, which heuristically finds only good alignments. The problem with instituting Smith-Waterman on a large scale is the time it take to process long queries, since must align the query to each entry in the database. Figure 1 shows our own estimation of BLAST running time versus Smith-Waterman in a genome size database.

Heuristics of this have been developed to speed the search, primarily the FASTA search engine (Pearson 2000). The current FASTA algorithm only computes a dynamic programming matrix for a few, probably high scoring database entries. It determines the canidate scores by a similar method to BLAST's 'seed' search. The FASTA classifies the query sequence into short sequence 'words' given by a *ktup* value (usually 1 or 2 for amino acids and 4-6 for DNA). The 'words' are then used to find canidate sequences by looking for identities in the database, it then attempts to find supporting matches within sequences. Finally it performs dynamic programming on all the candidates (Pearson 2000).

Algorithmic performance for database queries can be measured by the ability to classify known proteins in the correct structural/functional category. The SCOP database, covering 800+ superfamilies of proteins has been used a metric to determine remote sequence homology detection (Hubbard 1997, Jaakkola 1999). The database is heirarchially structured into Protien Folds, Superfamilies, Families, Species and finally protein sequences. The members of individual families in the SCOP database have, due to structure, highly redundant sequences. To use the database and associated classification hierarchy, the redundant entries must be removed for algorithms to find sequences outside of a single family. Additionally, since the SCOP is a domain database, results from similarity between unclassified and uninteresting (from a biological standpoint) of proteins will be minimal (Brenner 2000, Jaakkola 1998).

Homology detection is a key part of the biologists toolkit for determining protein function from newly discovered sequences. The problem seems largely solved by programs such as BLAST and FASTA, but these tools can be much further enhanced to produce much more refined and discriminative output. We report that our modified version of BLAST, enhanced BLAST (eBLAST) is able to return Smith-Waterman like results on a time scale comparable to BLAST.

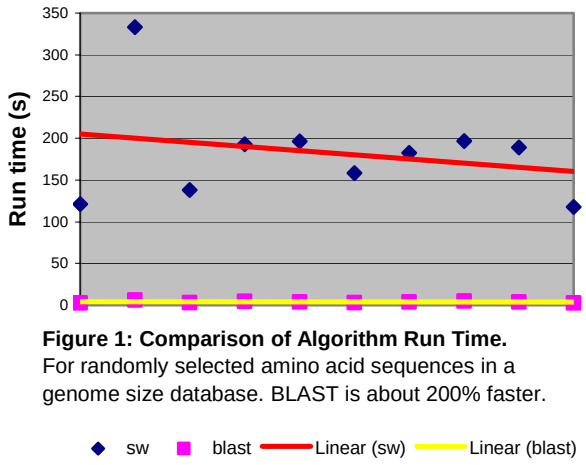

Figure 1: Comparison of Algorithm Run Time. For randomly selected amino acid sequences in a genome size database. BLAST is about 200% faster.

Methods

eBLAST

This algorithm involves feeding the results of a run of BLAST algorithm to the Smith-Waterman algorithm. For a given sequence, and a given database, we run the BLAST algorithm on that database (using the "blastall" program). The purpose of this initial BLAST run is to remove obvious false positives from the results. In addition, we make BLAST return sequences with significantly high E-values (low homology). BLAST, being a heuristic and not a provably optimal algorithm cannot be trusted for its accuracy. Hence, by returning a large set of results, we reduce the number of false negatives that will otherwise be eliminated by BLAST. Once this set of BLAST results (ordered by E-value) is obtained, we choose the top 10% of these results and use those to create a new "database" of sequences that will be used in the Smith-Waterman search. The decision to use 10% of the results is based on the results of running Smith-Waterman on various percentages of BLAST results. The final step in the algorithm is to run the Smith-Waterman algorithm (using the "ssearch33" program) on the shrunken "database" with the original query sequence.

Algorithm Performance—Search Relevance Heuristic (SRH)

The relative performance of the three algorithms was compared on a large set of test sequences. These test sequences were obtained from the SCOP database (95% sequence dissimilarity between families). Since

this database organizes protein sequences by structural/functional families, we chose one sequence from each family to make our set of test candidates. We used two versions of the SCOP PDB-ISL database (Brenner 2000) the pdb_isl90 and psb_isl95. These are amino acid databases heirarchially ordered in the SCOP scheme. We removed any families having less than 5 sequences, to avoid skewing the data by inappopriately weighting small clusters of homologous sequences. The sequences within each family are either 90% or 95% dissimilar. The superfamily level of the 95% database contained 131 entries, the 90% contained 220. The three algorithms: BLAST, EBLAST and Smith-Waterman, were run for each of these test sequences.

We implemented a metric, which we called the SRH, to compare the relative performance of each algorithm. The SRH was a measure of practical relevance of the results. Given a sequence and a database, each algorithm produced a list of sequences similar to the query sequence sorted in ascending order by E-value. Beginning with the resulting sequence with the lowest E-value, the number of consecutive sequences returned which were classified as related to the query sequence at the superfamily level by SCOP were counted. This number was then expressed as a percentage of the total number of sequences returned by that algorithm. To allow for some noise, small runs of 5 or less unrelated sequences were allowed before counting stopped. The rationale for this metric was that, when a biologist performs a search using this database, they would begin examining the results starting with the sequence with the lowest E-value and continuing on with each subsequent result as long as it appeared related; after too many false positives, the biologist could safely assume that the remaining results were no longer of interest. The results from Smith-Waterman were considered control values (since Smith-Waterman produces provably optimal alignments) to which the SRH's of BLAST and eBLAST were compared.

Results

We compared the performance of the three search algorithms with the SCOP pdb_isl90 and pdb_isl95 database. The database has heiracrhially organized amino acid sequences corresponding to protein domains. There is 95% or 90% sequence dissimilarity in between families, thus the ability of an algorithm to find members of a superfamily is tested more rigorously. We chose one member per family, ran the alignment against the databases then compared the abilities of Smith-Waterman, BLAST and eBLAST with our SRH metric. Figure 2 shows the result of the algorithms classifying the 131 and 220 superfamilies in the database. For each superfamily, BLAST and eBLAST were normalized to the Smith-Waterman alignment performance, since it is the most statistically rigorous algorithm and is not based on heuristics. In the graph, the SRH value of 0 corresponds to the Smith-Waterman performance. Blast deviated most from 0, average 0.0934, while eBLAST's average was 0.0137. The X axis in this case is increasing family size. For the 90% database, there were much fewer cases where Smith-Waterman obtained the same SRH score as the two algorithms. In figure 2b, the BLAST average SRH was 1.905 while eBLAST averaged 1.225.

Figure 3 compares BLAST and eBLAST by normalizing SRH scores to the Smith-Waterman. The square area is a two dimension space where the X axis is the 'eBLASTness' of the search and the Y axis is the 'BLASTness'. Search results where Y=0 is a point where eBLAST had an SRH score exactly the same as Smith-Waterman, while X=0 meant BLAST returned a score equal to Smith-Waterman. Scores on the diagonal X=Y were searches where both BLAST and eBLAST were an equal distance away from the Smith-Waterman score, essentially tying. For the 95% database (fig 3a), most points are Y=0, showing in the majority searches eBLAST scores the same as Smith-Waterman, BLAST scores higher than eBLAST only 5 of 131 times. In the 90% database (fig 3b), there is a larger distribution of scores. Most scores are in the lower triangle, showing that eBLAST performed more like Smith-Waterman than BLAST.

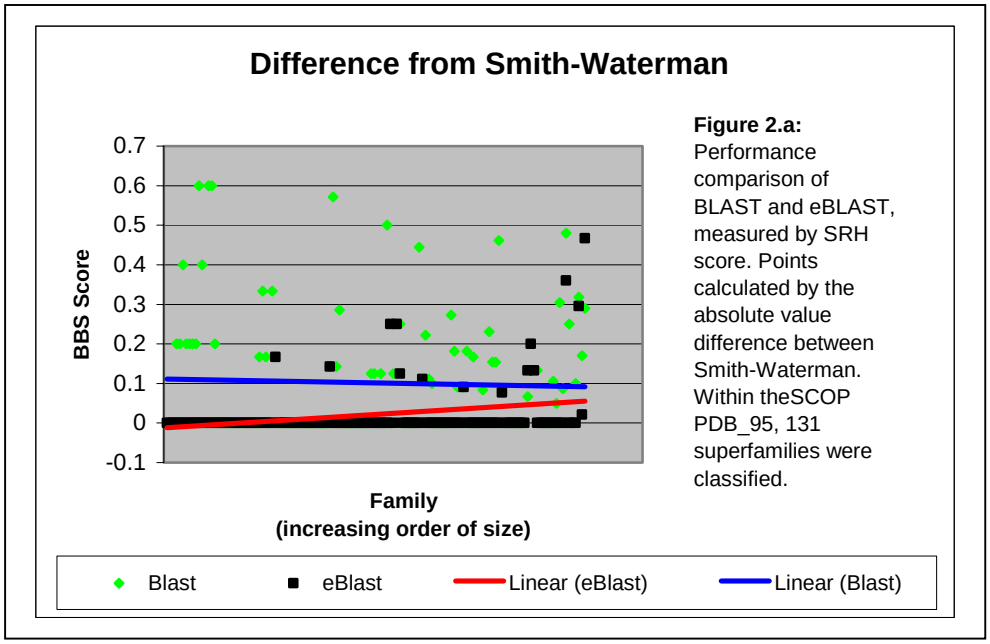

Figure 2.a: Performance comparison of BLAST and eBLAST, measured by SRH score. Points calculated by the absolute value difference between Smith-Waterman. Within theSCOP PDB_95, 131 superfamilies were classified.

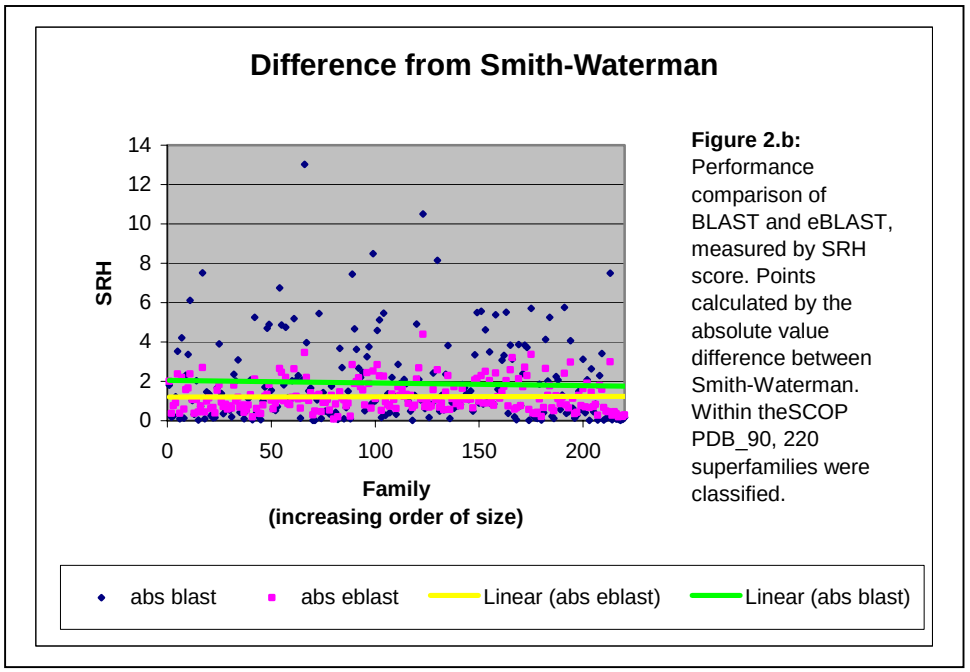

Figure 2.b: Performance comparison of BLAST and eBLAST, measured by SRH score. Points calculated by the absolute value difference between Smith-Waterman. Within theSCOP PDB_90, 220 superfamilies were classified.

Discussion

We report that eBLAST performed comparabably to Smith-Waterman alignment, within a much smaller time period. In broad terms, BLAST casts a large net and returns many sequences, most of which have high E scores (low similarity). eBLAST is passed this set of pre-filtered sequences, then conducts a Smith-Waterman alignment, essentially resorting the BLAST reported hits. eBLAST reports a small percentage of the alignments it created from the database it was passed, though the alignments are optimal local alignments. Our improvement over BLAST is twofold, we are able to reproduce Smith-Waterman like results in much less time, second, we are able to sort the hits so more relevant hits occur earlier in the results list.

By our rough estimates, Smith-Waterman is 200% slower than BLAST at queries on large databases (fig 1). We were not able to directly test the speed of eBLAST since we were using separate search programs that had to be executed serially, using temporary files to pass sequences between the programs. Generally, the time eBLAST will take to return a set of alignments will be linearly determined by the percentage of genome database passed to it. The percentage of the genome database BLAST passes will also affect the accuracy of eBLAST's results. We found 10% to be a suitable compromise for the SCOP databases we searched, though searches for certain types of proteins may require larger database sizes.

Sequence sorting is an important consideration in the application of alignment algorithms. Our aim is to return a continuous list biologically relevant database hits, defined here as sequences within the SCOP superfamily. For a sequence alignment list returned, the SRH counts the number of consecutive hits within the superfamily as determined by SCOP hierarchies. If the continuous sequence is broken by more than 5 non-superfamily hits, the SRH stops evaluating the hits returned. The SRH score is the number of continuous results is divided by the family size, which weights families similarly. The best case is an algorithm will return all of the members of a superfamily contiguously, scoring a 1.

Our observation is that BLAST may find all the homologous members of a superfamily, but they will be interdispersed with non-members. The non-members may be sequences that have high similarities of biologically uninteresting functions (for example, very common protein motifs for a larger class than the superfamily). Alternativley, they may be remotely homologous sequences that may not have been categorized in the superfamily. These are the most biologically interesting sequences, though hardest to find. To increase the specificity of the remote homology detection, we conduct a Smith-Waterman alignment of all the BLAST returns. By our metric (fig 3a,b), this at least moves more of the homologous sequences to the top of the list and at best will include remote homologues that BLAST sorted much lower.

Our investigation was an initial inquiry to using Smith-Waterman alignments within the BLAST framework. The FASTA package actually does something very similar, in a time scale similar to BLAST. It finds 'seeds' of high sequence similarity, then generates sequences of high homology probability and computes Smith-Waterman alignment on the relevant subsections (Pearson 2000, Durbin 1998). Algorithmically, our procedure is different from the FASTA because the subregions of sequence for alignment by Smith-Waterman are not determined heuristically. The advantage is we do not bias the alignments toward the original 'seed' sequences of high identity. This may allow alternative alignments of biological relevance in some cases and should be verified by eBLAST comparison to FASTA.

Literature Cited

Alberts, Bruce et al. *Molecular Biology of the Cell: Third Edition*. Garland Publishing, NY, 1994.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.*Nucleic Acids Res.* 1997;25:3389-3402. http://ncbi.nlm.nih.gov/BLAST

Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* 2000 Jan 1;28(1):254-6.

Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C., SCOP: a structural classification of proteins database. *Nucleic Acids Res* 2000 Jan 1;28(1):257-9

Durbin, R. Eddy, S., Krogh, A., Mitchinson, G. *Biological Sequence Analysis.* Cambridge University Press, Cambridge, UK, 1998.

Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol*. 2000;132:185-219. http://alpha10.bioch.virginia.edu/fasta/

Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. *J Comput Biol.* 1999 Feb-Apr;7(1-2):95-114.

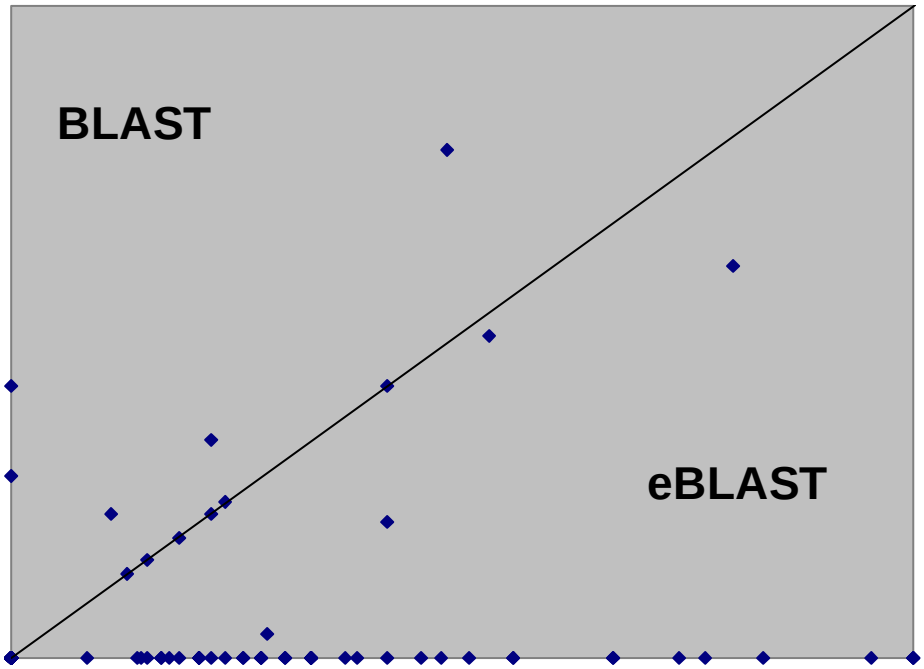

Figure 3a: Performance of BLAST and eBLAST relative to Smith-Waterman by the SRH metric. Each point in the BLAST region is a search where BLASt was closer to Smith-Waterman score than eBLAST and vice versa. A point on the diagonal is a tie.

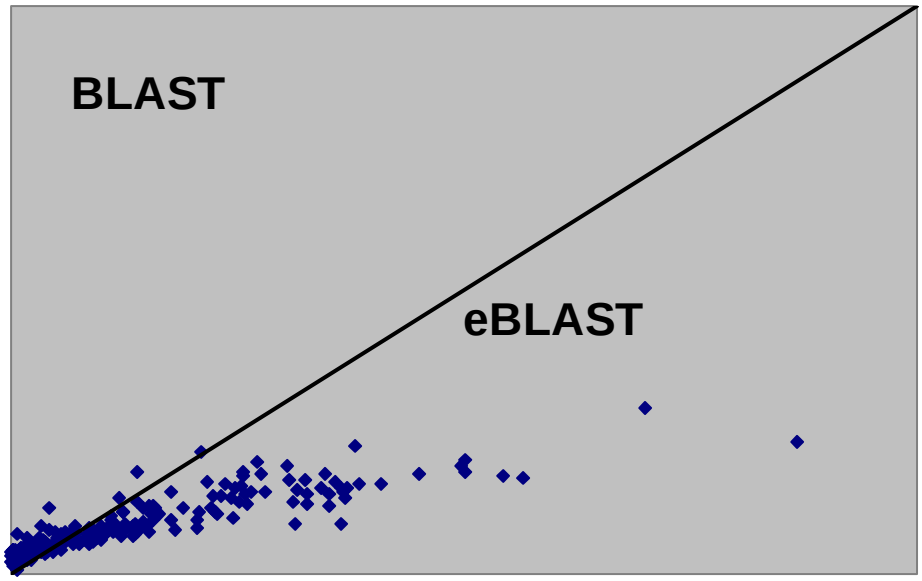

Figure 3b: Performance of BLAST and eBLAST relative to Smith-Waterman by the SRH metric. Each point in the BLAST region is a search where BLASt was closer to Smith-Waterman score than eBLAST and vice versa. A point on the diagonal is a tie.