EstSearch: A Tool for More Informative EST Alignments

Vera Asodi vera@compugen.co.il
Alex Diber alex@compugen.co.il
Raveh Gill-More raveh@compugen.co.il
Hershel Safer hersh@compugen.co.il
Mor Amitai mor@compugen.co.il
Compugen Ltd., Pinchas Rosen 72, Tel Aviv 69512, Israel

Keywords: pairwise sequence alignment, expressed sequence tag

1 Introduction

Differences between homologous expressed DNA sequences arise in multiple ways. Evolution causes codon substitution, insertion, and deletion; sequencing errors cause the same changes, but for nucleotides. Existing EST database search tools do not differentiate between these sources of variation.

This paper introduces EstSearch, a tool that fills this gap. Instead of comparing sequences nucleotide-by-nucleotide, EstSearch progresses codon-by-codon. It shifts frames on the fly and recognizes the correct orientation of aligned ESTs.

EST database search sensitivity was compared for EstSearch and four popular DNA-to-DNA alignment algorithms. TBlastX discriminated most effectively at moderate distances. EstSearch was most effective for more distant homologues, by a margin that increased with the distance.

A C implementation of EstSearch and the datasets described in this paper are freely available by anonymous FTP to ftp.compugen.co.il in the directory /pub/research/estsearch.

2 A New Comparison Model

The EstSearch model, illustrated below, captures both evolutionary changes and sequencing errors when aligning expressed DNA sequences. Nodes represent codon alignment columns. An arc directed into node M indicates that the next codon in X is aligned with the next codon in Y (they may or may not be identical). An arc into node I represents insertion of a codon in X, relative to Y; an arc into node D, deletion of a codon from X.

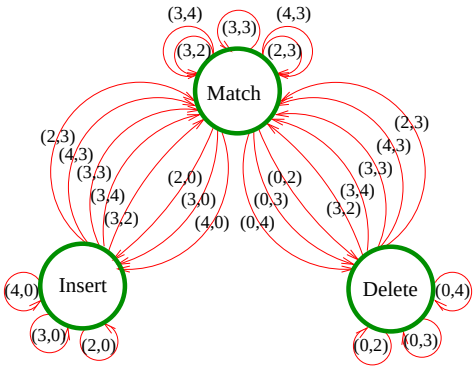

Figure 1: EstSearch codon comparison model.

Arcs represent the sequence of codon alignments as well as nucleotide insertions and deletions (indels). An arc's label indicates the numbers of X and Y nucleotides used. The most common arc

is $M(3,3)$, which means that full codons from X and Y are aligned. Other arcs represent indels. For example, arc $M(2,3)$ is used when the X codon suffers a nucleotide deletion. The alignment advances only two nucleotide positions in X, but advances all three nucleotide positions in Y.

EstSearch performs four alignments for each sequence pair, using the original and reverse complement of each sequence. Tools that progress nucleotide-by-nucleotide only benefit from two of these alignments. Since amino acids produced by reverse complements differ, however, EstSearch yields more informative results by examining all four cases. Generally one alignment gives a much higher score than the other three, so EstSearch also finds the correct orientation of the ESTs.

3 Implementation

The key innovation in EstSearch is the comparison model; the software uses standard dynamic programming methods to find alignments. Parameters that determine scores for alignment operations are similar to those in other alignment tools, modified to reflect a focus on codon-to-codon comparisons. Parameters can be set by the user, but the default values work well for a wide range of data.

Codon substitution matrix: Each entry represents the probability of aligning two codons. Frequencies found from good protein alignments were modified to reflect the possibility of sequencing errors.

Codon gap open and gap extension: A large set of good protein alignments was used to estimate the probabilities of initiating and extending gaps in alignments of homologous proteins.

Nucleotide deletion and insertion: Observed insertion probabilities were modified to account for the number of identical nucleotides prior to the insertion; the model considers runs of length zero through five. It also includes the probability of inserting an unspecified character ("N").

Computing the scores: The score for a transition accounts for the from and to states, the arc used, and the number of nucleotide indels.

4 Tool Comparison

Search sensitivities of EstSearch, SSearch, FastaA, BlastN, and TBlastX were compared using an approach similar to that in [1]. Proteins of known relationship were identified; two proteins were considered homologous if they have the same values for class, fold, superfamily, and family in the SCOP database [2]. The test was focused on the "twilight zone" [4] by using a subset of SCOP entries that exhibit at most 40% similarity to each other. Entries from the dbEST database corresponding to these proteins were identified; this resulted in a database of 267 ESTs with known relationships.

The five tools were used to align every pair of ESTs and each tool's output was sorted by score (p-scores for BlastN and TBlastX; e-scores for the others). Self-comparisons were omitted. All algorithms include about ten non-homologous pairs among the ˜30 highest-scoring alignments. For the next ˜70 pairs, TBlastX includes the fewest non-homologous pairs. After this, EstSearch outperforms TBlastX and the other algorithms, by a margin that increases with the evolutionary distance. Equivalence numbers [3] yield the same conclusion; EstSearch results were 30% better than those from TBlastX.

References

[1] Brenner, S.E., Chothia, C. and Hubbard, T.J., Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, *Proc. Natl. Acad. Sci. U.S.A.*, 95(11):6073–6078, 1998.

[2] Hubbard, T. J. P., *et al.*, SCOP: a Structural Classification of Proteins database. *Nucl. Acids Res.* 27:254–6, 1999.

[3] Pearson, W. R., Comparison of methods for searching protein sequence databases. *Prot. Sci.* 4:1145–60, 1995.

[4] Rost, B., Twilight zone of protein sequence alignments, *Protein Eng.*, 12(2):85–94, 1999.