First publ. in: Phyloinformatics (2004), 1:6

EverEST – A phylogenomic EST database approach

*Dirk Steinke, Walter Salzburger & Axel Meyer**

*Lehrstuhl für Zoologie und Evolutionsbiologie
Department of Biology, University of Konstanz,
D-78457 Konstanz*

** Axel.Meyer@uni-konstanz.de*

Received: 27 October 2004 - Accepted: 20 December 2004

Abstract

The use of local similarity search algorithms (e.g. BLAST) against public databases is often the preferred method of choice for screening short expressed sequence tags (EST). The need for automation of processing, similarity analyses, and annotation of ESTs is highlighted in cases when several BLAST results should be analysed in a combined approach.
EverEST is a WINDOWS program that automizes database management and phylogenetic analyses of ESTs. Together with its interactive visualization tools and a variety of built-in data, EverEST effectively integrates data mining, annotation, and phylogenetic evaluation into one application. EverEST is constructed to maximize evolutionary relevant information that is contained in large amounts of DNA data while attempting to minimize computation time. The database package will be freely available at: http://www.evolutionsbiologie.uni-konstanz.de/~dirk/software.htm.

Introduction

Large-scale sequencing of partial cDNA clones as expressed sequence tags (ESTs) and similarity searches of these against public DNA and protein sequence databases is becoming a more widely-used method for gaining information on gene content and genomic complexity for many phyla (Hedges & Kumar 2002). The use of local similarity search algorithms against the public databases is often the method of choice for screening short EST sequences, because they determine scores for only those regions conserved between sequences (Altschul et al. 1990). Furthermore, local similarities of translated EST sequences to the public protein databases can often be detected even when similarities to the public DNA databases appear coincidental (Gish & States 1993). Several EST projects (e.g. Whitfield et al. 2002, Fizames et al. 2004, Renn et al. 2004) have used similarity searches for positive identification of known genes and the determination of putative functions of others. The most commonly used method for this is to run the BLAST similarity search programs for each EST. Hits from EST similarity searches are the basis for the inference of probable biological functions and also the evolutionary history, i.e., homology, whereas a lack of hits at least suggests the possibility of the discovery of a novel gene or convergence of genes.
The need for automation of processing, similarity analysis, and annotation of ESTs is highlighted in cases when several BLAST

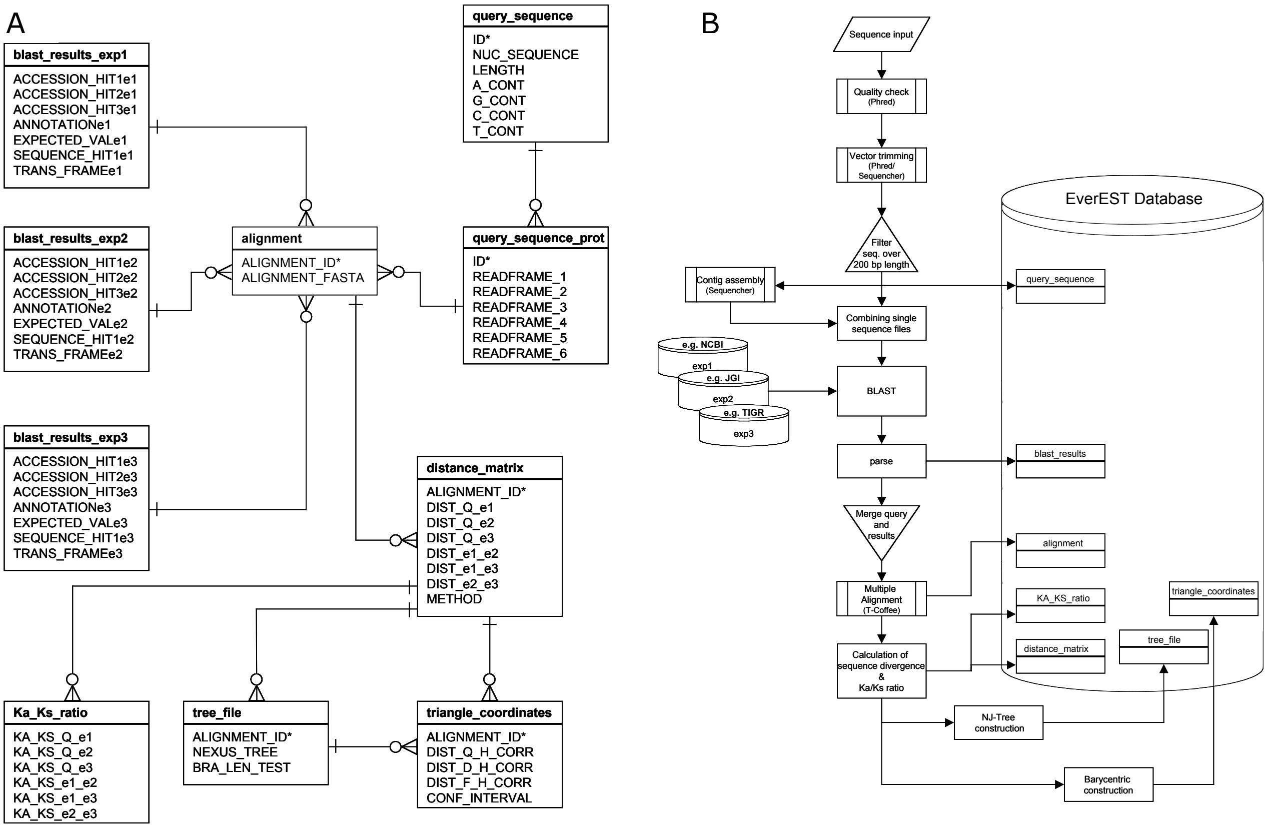

Figure 1a: Entity relationship diagram describing the association between the query sequence, the BLAST results and the phylogenetic analyses results. Primary keys are indicated with an asterisk.
Figure 1b: Typical process for pre-processing and analysing an EST sequence and for integrating the query and BLAST resources in the EverEST database.

results are analysed in a combined approach.
Here, we describe a database software we term EverEST, for processing simultaneous database searches using the BLAST algorithm against three databases to identify the best hits for any given EST sequence. In a further step EST sequences are associated with BLAST results and phylogenetic analyses in a relational database using its own specific database management system.

The program

A flow-chart for our phylogenomic EST database is shown in figure 1a. Most records in this flow-chart involve a ‘one-to-one relationship’ with the exception of the distance_matrix table, where a ‘one-to-many relationship’ exists. The distance_matrix table consists of all genetic distances between sequences in one entry of the alignment table and the used distance method. Figure 1a depicts that the blast_results are joined to the query_sequence tables through the alignment table. This table associates the distance_matrix table and through this the tree_file and triangle_coordinates tables.

The tree_file table consists of the tree representation in the Newick tree format and the result of a branch length test (Takezaki et al. 1995). The branch length test is a test of rate difference for each sequence under the tree root from the average rate of all sequences. This enables a linearized graphic representation of a group of ORF alignments in a ternary plot for which all related data are stored in the triangle_coordinates table. The primary key is the query sequence ID associating every single EST in the database with BLAST results, related alignments, distance matrices, NJ trees, ternary coordinates and Ka/Ks ratios.
Figure 1b depicts the processes involved in populating the tables in figure 1a. In case of

newly generated EST data, preprocessing includes base-calling, filtering of low-quality sequences, identification of sequence features, and vector trimming. Automatic base-calling and quality and vector trimming maybe performed with PHRED (Ewing et al., 1998). The input source for EverEST are preprocessed high quality ESTs of a length of more than 200 bp that are screened by local BLAST searches with an expected value threshold of e.g. < 1 x 10-15 against several databases. These should be chosen based on completeness and taxonomic relatedness to the source of the cDNA. The BLAST interface of EverEST is command line driven following the NCBI syntax for standalone BLAST and needs fastA input files (Pearson & Lipman 1988). The results are parsed into the blast_results tables which contain GenBank Accession Numbers of the three best hits, the sequence of the best hit and its annotation as well as related e-Values and information concerning translation frames. The query sequence and all possible three best hits of every single search are translated into amino acid code, combined by a Visual Basic routine, and aligned using the T-Coffee algorithm (Notredame et al. 2000) which is a component of the EverEST package. T-Coffee is a progressive multiple alignment program which considers information from all of the sequences during each step, not just those being aligned at that particular stage. The alignment is stored into the alignment table in the fastA file format. Following the alignment, sequence divergences for every pair in the alignment is estimated as the observed proportion of amino acid sites at which the two sequences to be compared are different. All alignment positions with gaps were excluded previously (complete deletion). This option is generally desirable because different regions of DNA or amino acid sequences often evolve under different evolutionary forces. The user can choose between two methods for estimating evolutionary distances: Poisson-correction distance and Gamma distance (Dayhoff, 1978). The distances are used to construct a neighbor-joining tree and a ternary graphic representation as depicted in Figure 2. The ratio of the number of nonsynonymous substitutions per nonsynonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) is calculated by a subroutine to evaluate the selective forces acting on the protein (McDonald and Kreitman, 1991).

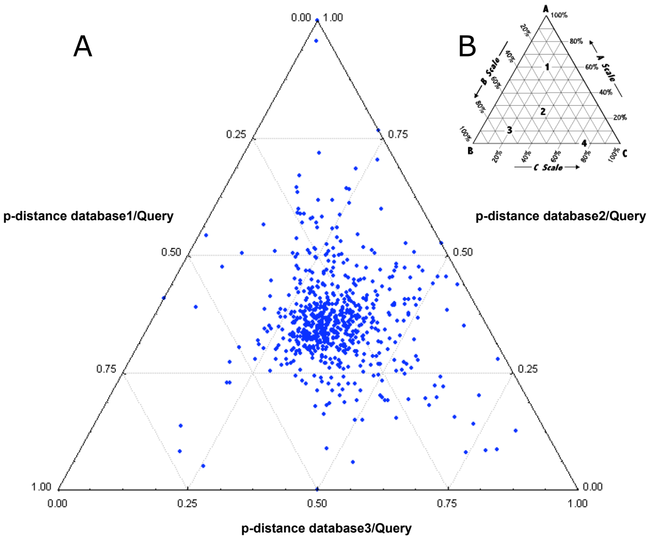

Figure 2a: Example for a ternary diagram of p-distances between query and BLAST results of 3 database searches respectively.

Figure 2b: Examples how the ternary diagram should be read. Note the numbers 1 - 4 on the diagram. The composition for each of these points is: 1 (60%/20%/20%), 2 (25%/40%/35%), 3 (10%/70%/20%) and 4 (0.0%/25%/75%). Please note that the ternary diagram is read counter clockwise.

The EverEST database was recently implemented on WINDOWS XP and 2000. It requires a preinstalled standalone BLAST version with at least the two components blastall and formatdb. The parse routine is written in C++ and has been successfully tested on several variants of the BLAST output files e.g. BLASTN, BLASTP or TBLASTX results. The other utilities supporting the database (e.g. divergence calculation, NJ-Tree construction and barycentric construction) were developed in Visual Basic. To visualize the results, several support features are available, e.g. output of trees in Newick format or the output of distance matrices in tab-delimited text format. The ternary graphic window allows choosing single ORF groups to obtain all related information in the database by a simple mouse click. Configurability was one of the main design issues in the development of EverEST. The usage is not limited to the query sequences and databases.

This flow-chart presented here (Figure 1b) is constructed to maximize evolutionary relevant information that is contained in large amounts of DNA data while attempting to minimize computation time. This system is geared towards as much automation as possible, while maintaining, organizing and flagging all results for individual inspection. Moreover, increased automation reduces the number of error-prone steps. By reducing errors in the data processing and manipulation, the quality of data submitted to the public databases should be increased. The system we have described here is functional, and in use on a regular basis, however, it is evolving. The areas that we continue to address include the implementation of the data preprocessing, enhancements in putative function assignment, and the automatic GenBank submission after successful annotation. For those ESTs that have insufficient hits, methods for automatically re-executing similarity searches periodically as the public databases get updated will be developed.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. 1990. Basic alignment search tool, *J. Mol. Biol.* 215: 403-410.

Dayhoff, M.O. 1978. Survey of new data and computer methods of analysis. In M. O. Dayhoff, ed., *Atlas of Protein Sequence and Structure*, vol. 5, supp. 3, pp. 29, National Biomedical Research Foundation, Silver Springs, Maryland.

Ewing, B.G., Hiller, L., Wendl, M.C. & Green, P. 1998. Basecalling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res.* 8: 175-185.

Fizames, C., Munos, S., Cazettes, C., Nacry, P., Boucherez, J., Gaymard, F., Piquemal, D., Delorme, V., Commes, T., Doumas, P., Cooke, R., Marti, J., Sentenac, H. & Gojon, A. 2004. The Arabidopsis root transcriptome by serial analysis of gene expression. Gene identification using the genome sequence. *Plant Physiol*. 134: 67-80

Gish, W. & States, D.J. 1993. Identification of protein coding regions by database similarity search, Nature Genet. 3: 266-272.

Hedges, SB. & Kumar, S. 2002. Genomics. Vertebrate genomes compared. *Science 297*: 1283-1285.

McDonald, J.H. & Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila, *Nature 251*: 652-654.

Notredame, C., Higgins, D.G. & Heringa, J. 2000. T-Coffee : A novel method for fast and accurate multiple sequence alignment, *J. Mol. Bio.* 302 : 205-217.

Pearson, W.R. & Lipman, D.J. 1988 Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85: 2444-2448.

Renn S.C., Aubin-Horth N. & Hofmann H.A. 2004. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray, *BMC Genomics* 5: 42.

Takezaki, N., Rzhetsky, A. & Nei, M. 1995. Phylogenetic test of the molecular clock and linearized trees, *Mol. Biol. Evol.* 12: 823-833.

Whitfield, C.W., Band, M.R., Bonaldo, M.F., Kumar, C.G., Liu, L., Pardinas, J.R., Robertson, H.M., Soares, M.B. & Robinson, G.E. 2003 Annotated expressed sequence tags and cDNA microarrays for studies of brain and behaviour in the honey bee, *Genome Res.* 12: 555-566.