Tutorial section

‘We are gathered here today’
– EST cluster databases

Expressed sequence tags (ESTs) are short, first pass sequences (approximately 500 bp) from cDNA clones and Adams et al. first published the method by which they were obtained over a decade ago.1 The motivation for sequencing ESTs is that they represent those parts of the genome that are transcribed, an important subset of the total genome and an aid to gene identification in eukaryotes. They were fast and cost effective and have since been instrumental in further research over a wide variety of organisms. ESTs were deposited in the dbEST database,2 created and maintained by the NCBI.

As with many other methods, accuracy was sacrificed for speed, and an error rate of approximately 2 per cent proved a hindrance to genome investigation. In addition, there were obviously many more ESTs than actual genes in the genome as sequences were taken from almost any location on the clone. In an effort to provide a more useful research tool, UniGene3 was initiated in 1995 by the NCBI. EST data was taken from dbEST and clustered according to the genetic information represented in each sequence tag. Thus non-redundant information was removed and researchers could focus on a particular area of the genome and possible splice variants in that region without wading through thousands of irrelevant pieces of data. It also solved many annotation errors and allowed easier association of ESTs with corresponding mRNA and protein sequences. Unigene is currently updated approximately once per month to coincide with additional data submissions.

Since UniGene was established there have been other bioinformatics projects designed to cluster EST information and disseminate it to the scientific population. STACK4 is an initiative from the South African National Bioinformatics Institute (SANBI) and was conceived in 1998 to cluster ESTs in tissue source categories. It is updated approximately twice per year. The Institute for Genome Research (TIGR) created Gene Indices,5 which cluster EST and transcript information for individual organisms. Updates to relevant Indices occur at quarterly intervals.

Prior to clustering, ESTs are extracted from dbEST and preprocessed to remove common contaminations such as repeat and vector sequences that may be artefacts of the experimental process. Low-complexity regions may also confuse the clustering process. Masking the sequences using databases of repeats and vectors removes these contaminants. Where contamination is found, regions are converted into Ns (DNA) or Xs (protein) which are then ignored by further clustering processes. Poly A tails may also be trimmed.

Each EST must have following information that is generally available in the header of each tag:

- sequence ID;
- clone ID from which the EST has been generated;
- location in respect of the poly A (3’ or 5’);
- organism;
- tissue and/or conditions;
- EST sequence.
EST clustering is not based on statistical cluster analysis, but rather classifies clusters into relational groups of matching sequences (index classes). Similarity (also known as distance) is measured between EST sequences and reduced to a binary value – either the sequences match or they do not. If the sequences match, they are accepted into the cluster, and if they do not, they are rejected. This distance is often measured using pairwise alignment algorithms. BLAST\textsuperscript{6} is one of the more common methods and can be tuned to align EST sequences. At NCBI, for example, a word size of 11 is used for UniGene initial clustering. STACK does not use a pairwise alignment algorithm, and instead relies on a d2 cluster algorithm\textsuperscript{7} that is based on a weighted word comparison. TIGR uses its own Assembler\textsuperscript{8} method to produce TC (tentative consensus) sequences.

There are two main types of clustering method. \textit{Loose clustering} uses a lower quality of sequence in the initial phases and repeats the clustering alignment (match or mismatch) several times. This allows for greater coverage of expressed data and a higher likelihood of isoform (different members of a gene family) inclusion. The main disadvantage of this method is the increased risk of including paralogous sequences in a cluster. Assembly of such clusters results in a longer consensus sequence. \textit{Stringent clustering} is completed using a single pass cluster grouping, but the sequence data used is of a higher quality and has undergone a more rigorous preprocessing phase. Owing to this initial stage, stringent cluster methods result in a lower coverage of expressed gene data, and reduced isoform inclusion. Assembly of such clusters results in a shorter consensus sequence.

Clustering itself is either supervised, or unsupervised. \textit{Supervised} clusters are classified with respect to seed sequences. These seeds are often reference sequences consisting of full-length mRNA, although exon constructs from genomic sequences, or EST cluster consensi may also be used. Reference seeds may be taken from Refseq as well as other sources. In \textit{unsupervised} clustering, ESTs are classified with no external sequence reference point.

Table 1 compares the three EST databases.

Through annotation and clustering of EST sequences, researchers can make use of a vast array of data that was previously less accessible, not only to research simple gene expression, but also to study alternative splicing as a source of protein manufacture and gene regulation. The Alternative Splice Database\textsuperscript{12} and Ensembl Genome Browser EST map\textsuperscript{13}

\textbf{Table 1: Comparison of the three EST databases}

<table>
<thead>
<tr>
<th>Database</th>
<th>Input format</th>
<th>Output format</th>
<th>Preprocessing</th>
<th>Clustering method</th>
<th>Contig assembly</th>
<th>Alignment processing</th>
<th>Cluster joining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigene</td>
<td>FASTA</td>
<td>FASTA</td>
<td>Repeats; vector; low complexity; clean data (\geq 100) bp accepted</td>
<td>Loose; supervised</td>
<td>None</td>
<td>None</td>
<td>No</td>
</tr>
<tr>
<td>STACK</td>
<td>FASTA</td>
<td>FASTA, GDE</td>
<td>Human repeats; vector; ribosomal and mitochondrial DNA; clean data (\geq 50) bp accepted</td>
<td>Loose; unsupervised</td>
<td>PHRAP</td>
<td>CRAW\textsuperscript{9}</td>
<td>Yes\textsuperscript{*}</td>
</tr>
<tr>
<td>TIGR</td>
<td>FASTA</td>
<td>FASTA</td>
<td>Vector; PolyA/T tails; adaptor sequences; bacterial sequences; sequences from EGAD\textsuperscript{10} used</td>
<td>Stringent; supervised</td>
<td>CAP3\textsuperscript{11}</td>
<td>None</td>
<td>No</td>
</tr>
</tbody>
</table>

\textsuperscript{*} Clusters are joined using common clone IDs. If two ESTs in one cluster are from clones X and Y and three ESTs in a separate cluster are from clones A, B and X, then the clusters may be joined as the gene represented by both clusters.
are just two examples of an array of tools and repositories designed to utilise EST information in order to assist the researcher with the ever-growing wealth of information on isoform gene expression.

Lisa Mullan
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton,
Cambridge CB10 1SD, UK
E-mail: lisa@ebi.ac.uk

References


10. URL: http://www.tigr.org/tdb/egad/egad.shtml

