EST clustering

Swiss Institute of Bioinformatics (SIB)

26-30 November 2000

Expressed sequence tags (ESTs)

ESTs represent the most extensive available survey of the transcribed portion of the genome.

ESTs are single-pass reads from the 5' and/or 3' end of cDNA clones.

ESTs represent partial sequences of cDNA clones (~ 300 bp).

High-volume and high-throughput data production.

ESTs are used extensively and are indispensable for gene discovery and genomic mapping.

ESTs are often associated with tissues.

There are 9,372,718 of EST entries in GeneBank (October 26, 2001):

- 3,859,807 entries of human ESTs;
- 2,328,188 entries of mouse ESTs;
- ...

Expressed sequence tags (ESTs)

ESTs are difficult to use effectively:

- partial gene sequences;
- high error rates ($\sim 1/100$) because of the sequence single-pass;
- not a defined protein product;
- not curate in a highly annotated form;
- high redundancy in the data.

The value of ESTs is greatly enhanced if they are used to construct high-fidelity set of non-redundant transcripts:

- fewer sequences to analyze;
- solving redundancy can help to correct errors;
- longer sequences;
- better annotated;
- easier association to mRNAs and proteins;
- detection of splice variants;
- they can be used for extensive functional annotation;
- facilitates gene expression studies.

EST clustering

The goal of the clustering process is to incorporate overlapping ESTs which tag the same transcript of the same gene in a single cluster.

Once clustering is completed, one or more consensus assemblies for each individual cluster can be produced.

EST clustering and assembling presents a number of distinct computational problems:

- can be extremely time consuming due to the intrinsic need for all pairs of ESTs to be tested for overlap;
- ESTs derive from a wide variety of sources representing the polymorphism in the original samples;
- high sequencing errors;
- high rate of insertions and deletions;
- contaminations by vector and linker sequences;

$\Rightarrow$ the degree of identity in overlapping sequences from the same gene will be lower than in genomic projects.

EST clustering

Patterns of overlapping sequences caused by alternative splicing will be different from the ones observed in the genomics shotgun projects.

A big number of EST sequences lack the base-calling quality values or the associated chromatograms.

$\Rightarrow$ Although the clustering and assembling of ESTs is a problem similar to clustering and assembling of genome shotgun data, software and parameters have to be adapted to solve the ESTs specific problems.

Different clustering/assembly procedures have been proposed with associated resulting database, also called gene indices:

- UniGene (http://www.ncbi.nlm.nih.gov/UniGene)
- TIGR Gene Indices (http://www.tigr.org/tdb/tgi.shtml)
- STACK (http://www.sanbi.ac.za/Dbases.html)
- trEST (ftp://ftp.isrec.isb-sib.ch/pub/databases/trest)

Pipeline for EST clustering

The steps for EST clustering:

- Obtain EST sequences of interest (sequencing project, dbEST, ...) and/or chromatograms.
- Process chromatograms for low-quality regions (Phred).
- Mask repeats (pairwise comparison programs, RepBase).
- Mask/delete contaminants (pairwise comparison programs, mitochondrial DNA, ribosomal DNA, ...).
- Clustering (time expensive):
  - ▷ "Strict" cluster: pairwise alignment programs are used to compare each sequence of the dataset. Only sequences sharing good homology of the same regions are clustered together.
  - ▷ "Loose" cluster: accept sequences in the cluster if they have a region of homology with at least one sequence of the cluster.
- Assembling of the clusters to generate consensus sequences and contigs.
  - ▷ This step is performed by programs like Phrap, CAP3, ...
  - ▷ Each cluster can produce ≥ 1 contigs, representing chimeras, splice variants, sequencing errors, ...
- Some post-processing can be done.

UniGene

UniGene characteristics:

- contains clusters of sequences deriving from GeneBank;
- each cluster represents a unique gene as well as tissue types where the gene is expressed and map locations;
- ESTs are included;
- no attempts to produce contigs or consensus sequences;
- all splice variants for a gene are put into the same cluster.

UniGene uses pairwise sequence comparison at various levels of stringency to group related sequences, placing closely related and alternatively spliced transcripts into clusters.

UniGene data can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene.

UniGene

UniGene build procedure:

- Screen for contaminants, repeats, and low-complexity regions in GeneBank:
  - Low-complexity are detected using *Dust*;
  - Contaminants (vector, linker, bacterial, mitochondrial, ribosomal sequences) are detected using pairwise alignment programs;
  - Repeat masking of repeated regions (RepeatMasker).
- Clustering procedure, which results in clusters called anchored clusters:
  - Build clusters of genes and mRNAs.
  - Add ESTs to previous clusters (megablast).
  - ESTs that join two clusters of genes/mRNAs are discarded.
  - Any resulting cluster without a polyadenilation signal or two 3' ESTs is discarded.
- Ensures 5' and 3' ESTs from the same clone belongs to the same cluster.
- ESTs that have not been clustered, are reprocessed with lower level of stringency. ESTs added during this step are called guest members.
- Clusters of size 1 are compared against the rest of the clusters with a lower level of stringency and merged with the cluster containing the most similar sequence.

TIGR Gene Indices

TIGR Gene Indices uses assembly algorithms, rather than clustering, to produce tentative consensus (TC) sequences that represent the underlying mRNA transcripts.

The TIGR Gene Indices building method tightly groups highly related sequences and discard under-represented, divergent, or noisy sequences.

TIGR Gene Indices characteristics:

- separate closely related genes into distinct consensus sequences;
- separate splice variants into separate clusters;
- low level of contamination.

TC sequences can be used for genome annotation, genome mapping, and identification of orthologs/paralogs genes.

References:

- Quackenbush *et al.* (2000) *Nucleic Acid Research*,**28**, 141-145.
- Quackenbush *et al.* (2001) *Nucleic Acid Research*,**29**, 159-164.

TIGR Gene Indices

Construction process of the TIGR Gene Indexes (identical process for each gene index):

- EST sequences recovered form dbEST (http://www.ncbi.nlm.nih.gov/dbEST);
- Sequences are trimmed to remove:
  - vectors
  - polyA/T tails
  - adaptor sequences
  - bacterial sequences
  - mitochondrial sequences
  - ribosomal sequences
  - low quality sequences

TIGR Gene Indices

- Get expressed transcripts (ETs) from EGAD (http://www.tigr.org/tdb/egad/egad.shtml):
  - EGAD (Expressed Gene Anatomy Database) is based on mRNA and CDS (coding sequences) from GeneBank.
- Cleaned ESTs and ETs are compared using FLAST (a rapid pairwise comparison program). Sequences are grouped in the same cluster if both conditions are true:
  - they share $\geq 95\%$ identity over 40 nt or longer regions
  - < 20 bases of mismatch
- Each cluster is assembled using CAP3 assembling program to produce tentative consensus (TC) sequences.
  - CAP3 can generate multiple consensus sequences for each cluster
  - CAP3 rejects chimeric, low-quality and non-overlapping sequences.
- Builded TCs are loaded in the TIGR Gene Indices database and annotated.

STACK

Based on "loose" clustering, followed by strict assembly and analysis tools to identify, characterize, view and isolate sequence divergence.

The "loose" clustering approach, d2_cluster, is not based on alignments, but performs comparisons via non-contextual assessment of the composition and multiplicity of words within each sequence.

STACK produces longer consensus sequences than TIGR Gene Indices.

STACK concentrates on human data.

References:

- Miller *et al.* (1999) *Genome Research*,**9**, 1143-1155.
- Christoffels *et al.* (2001) *Nucleic Acid Research*,**29**, 234-238.

STACK

The STACK procedure:

- Sub-partitioning.
  - Select human ESTs from GeneBank;
  - Sequences are grouped in tissue-based categories.
- Masking using cross-match.
  - Parameters: minmatch=12 minscore=20;
  - Human repeat sequences (RepBase);
  - Vector sequences (ftp://ncbi.nlm.nih.gov/blast/db/vector.Z);
  - Ribosomal and mitochondrial DNA.
- "Loose" clustering (d2_cluster).
  - Parameters: word_size=6 similarity_cutoff=0.96 minimum_sequence_size=50 window_size=100;
  - The algorithm looks for a window of size 100 bases having at least 96% identity.
  - Clusters highly related sequences;
  - Clusters also sequences related by rearrangements or alternative splicing.

STACK

- Assembly (Phrap).
  - Parameters: vector_bound=0 trim_score=150 forcelevel=0 oenalty=-2 gap_init=-4 gep_ext=-3 ins_gap_ext=-3 del_gap_ext=-3 maxgap=30;
  - STACK don't use quality information available from chromatograms;
  - The lack of trace information is largely compensated by the redundancy of the ESTs data;
  - Multiple contigs are generated within clusters, corresponding to divergent groups.
- Alignment analysis (CRAW).
  - Parameters: sig=0.5 window_size=100 ignore_first=50;
  - Generates consensus sequence with maximized length;
  - Partitioning of sub-ensembles;
  - Annotate polymorphic regions and alternative splicing
- Linking.
  - Link cDNA clone ID to a cluster if two transcripts correspond to the same clone;
  - Merge contigs linked to the same cDNA clone.

trEST

trEST is an attempt to produce contigs from clusters of ESTs and to translate them into proteins.

trEST uses UniGene clusters and clusters produced from in-house software.

To assemble clusters trEST uses Phrap and CAP3 algorithms.

Contigs produced by the assembling step are translated into protein sequences using the ESTscan program, which corrects most of the frame-shift errors and predicts transcripts with a position error of few amino acids.

sim4: mapping expressed sequences to genomic sequences

sim4 is an algorithm that maps ESTs, cDNAs, mRNAs to genomic sequences.

sim4 algorithm finds matching blocks reppresenting the "exon cores".

The algorithm used by sim4 is similar to the blast algorithm:

- Determine high-scoring segment pairs (HSPs).
  - ▷ High scoring gap-free regions.
  - ▷ Setects exact matches of length 12.
  - ▷ Extend matches in both directions with a score of 1 for a match and -5 for a mismatch until no increase of the score.
- Select HSPs tha could represent a gene.
  - ▷ Use dynamic programming algorithm to find a chain of HSPs with the following constrains:
    1. Their starting position are in increasing order.
    2. The diagonals of consecutive HSPs are nearly the same ("exon cores") or differ enough to be a plausible intron.

sim4: mapping expressed sequences to genomic sequences

- Find exon boundaries.
  - If "exon cores" overlap, the ends are trimmed to find boundary sequences (GT..AG or CT..AC).
  - If "exon cores" don't overlap, they are extended using a "greedy" method. Then the ends are trimmed to find bounary sequences.
  - If this last step fails, the region between two adjacent exon cores is searched for HSPs at a reduced stringency.
- Determine alignments.
  - Found exons with anchored boundaries are realigned by a method to align very similar DNA sequences (*Chao et al., 1997*).

sim4 souces are available from http://globin.cse.psu.edu/globin/html/software.html.