Expressed sequence tag

From Wikipedia, the free encyclopedia

An **expressed sequence tag** or **EST** is a short sub-sequence of a cDNA sequence.[1] They may be used to identify gene transcripts, and are instrumental in gene discovery and gene sequence determination.[2] The identification of ESTs has proceeded rapidly, with approximately 65.9 million ESTs now available in public databases (e.g. GenBank 18 June 2010, all species).

An EST results from one-shot sequencing of a cloned mRNA (i.e. several hundred base pairs of sequence starting from an end of a cDNA). The cDNAs used for EST generation are typically individual clones from a cDNA library. The resulting sequence is a relatively low quality fragment whose length is limited by current technology to approximately 500 to 800 nucleotides. Because these clones consist of DNA that is complementary to mRNA, the ESTs represent portions of expressed genes. They may be represented in databases as either cDNA/mRNA sequence or as the reverse complement of the mRNA, the template strand.

ESTs can be mapped to specific chromosome locations using physical mapping techniques, such as radiation hybrid mapping, Happy mapping, or FISH. Alternatively, if the genome of the organism that originated the EST has been sequenced, one can align the EST sequence to that genome using a computer.

The current understanding of the human set of genes (as of 2006) includes the existence of thousands of genes based solely on EST evidence. In this respect, ESTs have become a tool to refine the predicted transcripts for those genes, which leads to the prediction of their protein products and ultimately their function. Moreover, the situation in which those ESTs are obtained (tissue, organ, disease state - e.g. cancer) gives information on the conditions in which the corresponding gene is acting. ESTs contain enough information to permit the design of precise probes for DNA microarrays that then can be used to determine the gene expression.

Some authors use the term "EST" to describe genes for which little or no further information exists besides the tag.[3]

The significance of ESTs, their properties, methods to analyze EST dataset and their applications in different areas of biology have been reviewed by Nagaraj et al. (2007).[4]

Contents

Sources of data and annotations

dbEST

dbEST is a division of Genbank established in 1992. As for GenBank, data in dbEST is directly submitted by laboratories worldwide and is not curated.

EST contigs

Because of the way ESTs are sequenced, many distinct expressed sequence tags are often partial sequences that correspond to the same mRNA of an organism. In an effort to reduce the number of expressed sequence tags for downstream gene discovery analyses, several groups assembled expressed sequence tags into EST contigs. Example of resources that provide EST contigs include:

- TIGR gene indices [5]
- Unigene [6]
- STACK [7]

Constructing EST contigs is not trivial and may yield artifacts (contigs that contain two distinct gene products). When the complete genome sequence of an organism is available and transcripts are annotated, it is possible to bypass contig assembly and directly match transcripts with ESTs. This approach is used in the TissueInfo system (see below) and makes it easy to link annotations in the genomic database to tissue information provided by EST data.

Tissue information

High-throughput analyses of ESTs often encounter similar data management challenges. A first challenge is that tissue provenance of EST libraries is described in plain English in dbEST.[8] This makes it difficult to write programs that can non ambiguously determine that two EST libraries were sequenced from the same tissue. Similarly, disease conditions for the tissue are not annotated in a computationally friendly manner. For instance, cancer origin of a library is often mixed with the tissue name (e.g., the tissue name "glioblastoma" indicates that the EST library was sequenced from brain tissue and the disease condition is cancer).[9] With the notable exception of cancer, the disease condition is often not recorded in dbEST entries. The TissueInfo project was started in 2000 to help with these challenges. The project provides curated data (updated daily) to disambiguate tissue origin and disease state (cancer/non cancer), offers a tissue ontology that links tissues and organs by "is part of" relationships (i.e., formalizes knowledge that hypothalamus is part of brain, and that brain is part of the central nervous system) and distributes open-source software for linking transcript annotations from sequenced genomes to tissue expression profiles calculated with data in dbEST.[10]

See also

- gene expression
- complementary DNA (cDNA)
- IMAGE cDNA clones
- Whole genome sequencing (WGS)

References

1. ^ ESTs Factsheet (http://www.ncbi.nlm.nih.gov /About/primer/est.html) . National Center for Biotechnology Information.
2. ^ Adams MD, Kelley JM, Gocayne JD, *et al.* (Jun 1991). "Complementary DNA sequencing: expressed sequence tags and human genome project" (http://www.sciencemag.org /cgi/pmidlookup?view=long&pmid=2047873) .

Science **252** (5013): 1651–6. doi:10.1126/science.2047873 (http://dx.doi.org/10.1126%2Fscience.2047873) . PMID 2047873 (http://www.ncbi.nlm.nih.gov/pubmed/2047873) . http://www.sciencemag.org/cgi/pmidlookup?view=long&pmid=2047873 .

3. ^ dbEST (http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html)
4. ^ Nagaraj SH, Gasser RB, Ranganathan S (Jan 2007). "A hitchhiker's guide to expressed sequence tag (EST) analysis" (http://bib.oxfordjournals.org/cgi/content/short/8/1/6?rss=1) . *Brief. Bioinformatics* **8** (1): 6–21. doi:10.1093/bib/bbl015 (http://dx.doi.org/10.1093%2Fbib%2Fbbl015) . PMID 16772268 (http://www.ncbi.nlm.nih.gov/pubmed/16772268) . http://bib.oxfordjournals.org/cgi/content/short/8/1/6?rss=1 .
5. ^ Lee Y, Tsai J, Sunkara S, *et al.* (Jan 2005). "The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes". *Nucleic Acids Res.* **33** (Database issue): D71–4. doi:10.1093/nar/gki064 (http://dx.doi.org/10.1093%2Fnar%2Fgki064) . PMC 540018 (http://www.pubmedcentral.gov/articlerender.fcgi?tool=pmcentrez&artid=540018) . PMID 15608288 (http://www.ncbi.nlm.nih.gov/pubmed/15608288) .
6. ^ Stanton JA, Macgregor AB, Green DP (2003). "Identifying tissue-enriched gene expression in mouse tissues using the NIH UniGene database". *Appl Bioinformatics* **2** (3 Suppl): S65–73. PMID 15130819 (http://www.ncbi.nlm.nih.gov/pubmed/15130819) .
7. ^ Christoffels A, van Gelder A, Greyling G, Miller R, Hide T, Hide W (Jan 2001). "STACK: Sequence Tag Alignment and Consensus Knowledgebase" (http://nar.oxfordjournals.org/cgi/pmidlookup?view=long&pmid=11125101) . *Nucleic Acids Res.* **29** (1): 234–8. doi:10.1093/nar/29.1.234 (http://dx.doi.org/10.1093%2Fnar%2F29.1.234) . PMC 29830 (http://www.pubmedcentral.gov/articlerender.fcgi?tool=pmcentrez&artid=29830) . PMID 11125101 (http://www.ncbi.nlm.nih.gov/pubmed/11125101) . http://nar.oxfordjournals.org/cgi/pmidlookup?view=long&pmid=11125101 .
8. ^ Skrabanek L, Campagne F (Nov 2001). "TissueInfo: high-throughput identification of tissue expression profiles and specificity" (http://nar.oxfordjournals.org/cgi/pmidlookup?view=long&pmid=11691939) . *Nucleic Acids Res.* **29** (21): E102–2. doi:10.1093/nar/29.21.e102 (http://dx.doi.org/10.1093%2Fnar%2F29.21.e102) . PMC 60201 (http://www.pubmedcentral.gov/articlerender.fcgi?tool=pmcentrez&artid=60201) . PMID 11691939 (http://www.ncbi.nlm.nih.gov/pubmed/11691939) . http://nar.oxfordjournals.org/cgi/pmidlookup?view=long&pmid=11691939 .
9. ^ Campagne F, Skrabanek L (2006). "Mining expressed sequence tags identifies cancer markers of clinical interest". *BMC Bioinformatics* **7**: 481. doi:10.1186/1471-2105-7-481 (http://dx.doi.org/10.1186%2F1471-2105-7-481) . PMC 1635568 (http://www.pubmedcentral.gov/articlerender.fcgi?tool=pmcentrez&artid=1635568) . PMID 17078886 (http://www.ncbi.nlm.nih.gov/pubmed/17078886) .
10. ^ :institute for computational biomedicine::TissueInfo (http://icb.med.cornell.edu/crt/tissueinfo/)

External links

- ESTs Factsheet (http://www.ncbi.nlm.nih.gov/About/primer/est.html) from NCBI, a good and easy to read introduction to ESTs.
- The NCBI Handbook, Part 3, Chapter 21 (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.section.858) has a very nice overview.
- ECLAT (http://mips.gsf.de/proj/est/) a server for the classification of ESTs from mixed EST pools (from fungus infected plants) using codon usage.
- The current number of EST sequences in the (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html) GenBank division dbEST.
- Web Resources for EST data and analysis (http://biolinfo.org/EST/)
- http://icb.med.cornell.edu/crt/tissueinfo/ TissueInfo project: Curated EST tissue provenance, tissue ontology, open-source software.
- http://www.estinformatics.org/ Web resource contains contains all publicly available ESTs which has been processed through various cleaning steps where contaminating DNA e.g. vector, E coli and short sequences (<100bp) removed.