EXPRESSED SEQUENCE TAGS (EST)

Liu Peigang and Yang Qian*

Department of Life Science and Engineering, Harbin Insistitute of Technology,
Harbin 150001, P.R.China

ABSTRACT

The focus on genome research has shifted from structural genomics to functional annotation. A rapidly growing area of functional research is the generation of expressed sequence tags (EST) in which a large number of randomly selected cDNA clones are partially sequenced. The generation of EST has proven to be a rapid and economical approach to the identification and characterization of expressed genes. Currently expressed sequence tag projects have accumulated over 15 million un-annotated files from more than 5318 cDNA libraries deposited in the public domain. Therefore, in this review we summarize the advantages of EST research , strategies for EST study, weaknesses of EST analysis and finally conclude with our work in EST research.

Keywords: EST, genome, bioinformatics

1. INTRODUCTION

Expressed sequence tags (ESTs) are nucleotide sequences generated from the ends of randomly selected cDNA clones. The remarkable expansion of EST efforts in the past ten years [1-3] has undoubtedly led to a revolutionary change in the strategies used by molecular geneticists for identifying and cloning novel genes. The first EST project was begun in 1991. More than 15,879,985 entries (release 030703) generated from different tissues and species are now stored in the public EST database (dbEST) [4], maintained at the National Center for Biotechnology Information(NCBI). Human and mouse sequences form the majority of the data held in this collection, with 5,037,058 and 3,679,027 entries, respectively (as of march 7,2003) [5]. For most of the clones, ESTs have been generated at both ends and the

*Corresponding author. .Tel:0451-86412952 E-mail: yangq@hope.hit.edu.cn

corresponding sequence traces can be easily retrieved for quality checking [6]; furthermore, it is possible to obtain in a few days most of the EST cDNA clones, such as the ones generated by the IMAGE consortium and the Institute for Genomic Research (TIGR) [7], from international distributors. EST clones represent useful molecular tools for gene characterization experiments, expression studies and expression of recombinant proteins. In most cases, a single EST clone (average insert approximately 1.5 kb) does not span the entire coding region of a gene. However, before starting any experimental procedure aimed at the isolation of the remaining part of the transcript, it must be pointed out that several bioinformatic tools exist that allow one to obtain full-length transcript information and refined chromosomal mapping assignment.

2. STRATEGIES FOR EST STUDY

EST study requires the following: formation of cDNA Library, Nucleotide Sequencing and Data Management and Bioinformatics Analysis. A brief description of the procedures as carried out in the author's laboratory is given below.

cDNA Library

Total RNA was isolated from tissues or cellsby standard methods. Poly(A)+ RNA was selected using a commercially available poly(A)+ RNA purification kit (such as Pharmacia). cDNA was synthesized using a cDNA synthesis kit and was cloned into vectors using unphosphorylated adaptors. The plasmid library was plated on 15-cm Luria Bertani agar plates with ampicillin. Individual colonies were propagated and saved at -80°C until further use.

Nucleotide Sequencing

The template DNAs for the sequencing reaction were prepared by an alkaline lysis method. Using a Perkin-Elmer 9600 thermal cycler and an ABI 373A sequencer (Applied Biosystems), the 5' ends of the cDNA clones were sequenced according to the thermal cycling protocol with a *Taq* Dye Primer Cycle Sequencing Kit (Applied Biosystems).

Data Management and Bioinformatics Analysis

FACTURA software (Perkin–Elmer) and BLAST analysis were used to remove vector sequence from the ESTs and to identify ''trash'' sequences, defined as sequences from bacterial DNA, sequences from primer polymers, sequences containing 3% of ambiguous bases (N), or sequences 100 bp long. All sequence data were preserved on record tape. These sequences were searched against GenBank and dbEST databases for homology comparison by using BLAST and FASTA in the Genetics Computer Group program package. EST sequences were considered part of known genes if they shared at least 95% homology over at least 100 bp of DNA sequence on BLAST search. These ESTs were divided into some categories according to the functions of corresponding genes as proposed in the literature [8].

3. APPLICATIONS OF ESTs

The EST resources have been used for novel gene identification and positional cloning, as well as genetic and physical mapping, radiation hybrid, molecular marker and micro-array development.

Novel Genes Identification

Expressed sequence tags (ESTs) have proven to be a powerful and rapid approach to identifying new genes that are preferentially expressed in certain tissues or cell types. Analysis of gene expression has moved rapidly from classical or single of a few genes toward genome wide studies on multiple genes. One of the major goals in any genome study is to identify the full set of genes in the targeted genome . Among various tools designed for gene identification, physical isolation of the genes through analyzing mRNA is the most reliable one . Expressed sequence tags (ESTs) are used most often for that purpose . ESTs represent 200–500 nucleotide gene signatures that provide information not derivable from pattern recognition techniques alone. In principle, similarity to an EST is a highly reliable indicator that a sequence is associated with a gene or pseudogene, because cloning of non-mRNAs into the cDNA libraries from which ESTs are derived is presumed to be rare. In practice, this assumption remains to be tested. The cDNA clones associated with most human ESTs are also publicly available, so that identification of an EST often provides rapid access to a laboratory reagent useful for further characterizing a potential gene of interest. The performance of various EST projects [9]in the past decade has resulted in the accumulation of nearly 4 million ESTs from the human genome and large numbers of ESTs from genomes of other species .

Genetic and Physical Mapping

ESTs help to quickly identify functions of expressed genes and to understand the complexity of gene expression. ESTs have also served as molecular genetic markers in genomic mapping. An EST is simply a segment of a sequence from a cDNA clone that corresponds to an mRNA. ESTs longer than 150bp were found to be the most useful for physical and genetic mapping as sequence-tagged sites (STSs), which are becoming standard markers for the mapping of the human genome. These short sequences from physically mapped clones represent uniquely identified map positions. ESTs can serve the same purpose as the random genomic DNA STSs and provide the additional feature of pointing directly to an expressed gene. Sequencing ESTs in different organs, tissues, or cells of the human body is complimentary to the genomic DNA sequencing in the human genome project . During the past years, tens of thousands of EST sequences were mapped on chromosomes by using RH to create sequence tagged sites, which serve both as the scaffold of a physical map and as gene labels in a transcription map [10]. Currently, systematic screening of transcription units builds up the link between structural and the functional genomics.

Other information associated with ESTs, such as the library of origin, can supply useful information about expression patterns within an organism, or conservation of structure across species. ESTs are also particularly valuable reagents for detecting alternative splicing and polymorphisms and for locating the 3' ends of genes, which can be used to distinguish related genes from each other.

4. WEAKNESSES AND BOTTLENECKS

EST procedures must deal with several obstacles leading to invalid or breakdown results, notably including sequencing errors, short sequenced bases, the alternate spicing, the problem of abundance and the tools for analysis. Large-scale annotation work is a relatively novel task and, by definition, a difficult and risky process. Essentially based on prediction tools, computer technology and biological knowledge always in evolution, the annotation of a EST sequence is never perfect and always inevitably incomplete. Classical errors and limits inherent to the annotations will be discussed here.

1. EST sequencing is far more accurate for the reasons of base alternative, insertion or deletion. This procedure is a enzyme-mediated course on base of PCR, which leads to poor sequencing accuracy in contrast to genome sequencing. Moreover, only partial fragments (150-400bp) were sequenced which greatly relied on the information of some known genes and tools for analysis.
2. The mRNA expression patterns of genes at different stages of the development and differentiation of distinct tissues and cell types can differ greatly. This is essentially determined by a precise regulation of gene transcriptional expression in a time-and-space-dependent manner. As result of this, gene expression is greatly unbalanced and produce different adundance, some housekeeping genes occupy almost 50% of sequenced fragments so as to waste time and money. Some strategies, such as normalization or substraction library , can deal with this problem in some degree.
3. The multinational and, therefore, fragmented organization of EST annotation allows scientists to follow the daily huge sequence production, but sets the problem of the heterogeneous character of the results and provides annotations that are difficult to compare and to exploit by automatic routines of sequencing. Moreover, the nomenclature of genes, proteins and their function is also a source of ambiguities. There is a clear lack of controlled vocabulary both in the literature and the databases. This problem is linked to sequence redundancy in the databases, which can contain several times the same genes under different names. The resulting loss in time for the search and the annotation is very serious [11].
4. Prediction softwares evolve very rapidly but it may take some time to recognize which is the most efficient. The ideal case is when the different prediction softwares are in agreement with each other and with the detected conserved regions [12].

5. OUR ACCOMPLISHMENTS

As a lab working on biocontrol fungi , we have successfully generated 1381 ESTs of *Chaetomium globosum*, some of which play a key role in biocontrol. We are trying to achieve full length characterization of some cDNAs related to multidrug resistant proteins (asscess number BP099630).

6. CONCLUSION

EST procedure is a relatively efficient, simple, less expensive but labor intensive method to identify new genes. With the increase of huge ESTs in database, we can deal with genes of interest by even more accurate softwares. It is a bridge to connect structural genome to functional one.

REFERENCES

[1] International Human Genome Sequencing Consortium, *Nature* (London), **409**, 2001, 860–921.

[2] J. Kawai, A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, Y. Ishii, T. Arakawa, A. Hara, Y. Fukunishi and H. Konno, *Nature* (London), **409**, 2001, 685–690.

[3] M.D. Adams, A.R. Kerlavage, C. Fields, and J.C. Venter, *Nat. Genet.*, **4**, 1993, 256–267.

[4] M.D. Adams, J.M. Kelley, J.D. Gocayne and M. Dubnick, (1991). Single-run partial sequencing of randomly selected cDNA clones is now a widely used tool in genome research

[5] H. Xiao, C.R. Merril, A. Wu, 6. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McCombie and J.C. Venter, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1991, 1651-1656.

[6] M.S. Boguski, T.M.J. Lowe and C.M. Tolstoshew, dbEST-database for expressed sequence tags. *Nature Genet.*, **4**, 1993, 332-333.
(Sasaki et al., 1994Sasaki T, Song J, Koga-Ban Y, Matsui E, Fang F, Higo H, Nagasaki)

[7] M. Hori, M. Miya, E. Murayama-Kayano, T. Takiguchi, A. Taka-suga, T. Niki, K. Ishimaru, H. Ikeda, Y. Yamamoto, Y. Mukai, I. Ohta, N. Miyadera, I. Havukkala and Y. Minobe, Toward cataloguing a11 rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. *Plant J.*, *6*, 1994, 615-24.

[8] M.H. Polymeropoulos, H. Xiao, C.R. Merril, A.B. Wu, R.F. Olde and R.F. Moreno, Complementary DNA sequencing: .Expressed sequence tags and human genome project. *Science* **252**, 1991, 1651–1656.

[9] M.D. Adams, M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields and J.C. Venter, Sequence identification of 2375 human brain genes.

Nature, **355**, 1992, 632–634.

[10] M. D. Adams, A. R. Kerlavage, R. D. Fleischmann, R. A. Fuldner, C. J. Bult, N. H. Lee, E. F. Kirkness, K. G. Weinstock, J. D. Gocayne and O. White, Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature,* **377,** 1995, 3–17.

[11] S.J. Wheelan and M.S. Boguski, Late-night thoughts on the sequence annotation problem, *Genome Res*., **8**, 1998, 168–169.

[12] N. Pavy, S. Rombauts, P. Déhais, C. Mathé, D.V. Ramana, P. Leroy and P. Rouzé, Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences, *Bioinformatics*, **15**, 1999, 887–899.