EXPRESSED SEQUENCE TAG: PAVE THE WAY TO NEW HORIZONS IN GENETIC RESEARCH

Awanish Kumar

*Assistant Professor, Department of Biotechnology,
National Institute of Technology, Raipur, India
Email: {awanik.bt@nitrr.ac.in}*

Abstract

Expressed Sequence Tag (EST) is a small portion of a gene that can be used to identify unknown genes and to map their positions on chromosome. In other words it is fragments of mRNA sequences derived from cDNA libraries through single sequencing reactions executed on randomly selected clones. Million of ESTs have been generated from different species. It is used as low cost alternatives for discovery of genes. From a functional point of view, they allow the determination of the expression profiles of genes in different conditions/status of any particular tissue and in this way it helps in the identification of genes. They have also found as vast application in different fields such as phylogenetics, transcript profiling and proteomics. ESTs also contribute to comparative genetics and it can help to decipher gene function by comparison between species, even genetically variant. Thus combining functional, localization and sequence data, ESTs contribute to an integrated approach to a genome. It is highly helpful in genetic research and made gene discovery very simple. Today, researchers are using ESTs to study the human and other organism's genome. In this way, it paves the way to new horizons in genetic research. This review provides a brief introduction, process of synthesis and search of EST, which endorse it as a tool for mapping and discovery of gene(s).

Key Words : EST, Gene expression, Genomics, Disease.

1. Introduction

Genome biologists are working meticulously to sequence and assemble the genomes of various organisms for a number of imperative reasons. Although important goal of any sequencing project may be to obtain sequence and identify of a complete set of genes. Its ultimate objective is to gain an understanding of when, where, how, and why a gene is turned on for expression [1]. Once we begin to understand how and where a gene is expressed under normal conditions, we can then study what occurs in an altered state, especially in case of disease. To accomplish the latter objective, however, researchers must identify and study a protein(s), coded for by a particular gene. We understand that finding a gene those codes for a protein(s) is not a simple job [2]. Traditionally, researchers would start their search by defining a biological problem and developing a strategy for finding genes. Search of the scientific literature provided various clues about how to carry on for identification of gene(s). Time to time researchers has published data that has established a link between a particular protein and a disease of interest. Researchers would then work to isolate that protein, establish their function, and locate that particular gene that coded for the protein. Alternatively, research could conduct what is referred to as linkage studies to determine the chromosomal location of a gene. Once the chromosomal location is

deciphered, genetic engineers would use biochemical methods to isolate the gene and its corresponding protein. Either way, these methods took a great deal in terms of time and yielded the description and localization of only a small percentage of the genes found in the genome of an organism. The development of such technology used to locate and fully describe a gene is known as Expressed Sequence Tags (ESTs). It provides researchers with an inexpensive and quick route for discovering new genes, for obtaining data on gene expression/regulation, and for constructing genome maps more easily.

2. Generation/Synthesis of EST

ESTs are small portions of DNA sequence (200-500 nt in length) that are generated from either one or both ends of an expressed gene after sequencing process [3]. The idea is to sequence small pieces of DNA that represent genes expressed in definite cells, tissues, or organs from different organisms and use these tags to fish a gene out of a portion of chromosomal DNA by matching with sequence. The challenge associated with identifying genes from genomic sequences varies among organisms and is dependent upon size of genome as well as presence/absence of introns in a particular genome. Introns are non-coding or the intervening DNA sequences that interrupts the protein coding sequence of a gene. Gene identification

is very difficult in eukaryotes, because most of their genome is composed of introns intermingled with a relative few DNA coding sequences i.e., exons. These genes are expressed into proteins via a complex process known as transcription followed by translation. Each gene (DNA) must be converted, or transcribed, into messenger RNA (transcription). This mRNA serves as a template for protein synthesis (translation). Broadly, protein synthesis is the process where DNA codes for the production of amino acids that ultimately gives proteins. This process is divided into two parts: transcription and translation (Figure 1). During the process of transcription, one strand of a DNA double helix is used as a template by mRNA polymerase to synthesize mRNA. In this step, mRNA passes through various phases, including one called splicing, where the non-coding sequences are removed from transcript. mRNA directs the synthesis of the protein by inserting amino acids, one by one, as dictated by the DNA and represented by the mRNA in translation. Interestingly, mRNAs in a cell do not contain sequences from the regions between genes, nor from the non-coding introns that are present within many genes [4]. Therefore, isolation of mRNA is a key to finding expressed genes in the vast span of the complex genome like human.

Fig. 1 : Schematic representation of protein synthesis process (adopted from <http://www.ncbi.nlm.nih.gov/>)

mRNA is very unstable outside the cell and it degrades quickly. This is the key problem associated with researcher during in vitro experiment. Therefore, researchers convert mRNA into complementary DNA (cDNA) with the help of an enzymes reverse transcriptase. cDNA production is the reverse of the usual process of DNA replication. It is produced by using mRNA as a template rather than DNA strand. Unlike genomic DNA, cDNA contains only expressed DNA sequences i.e., exons. cDNA is a much more stable compound, because it is generated from mRNA in which the intronic part have been removed. cDNA represents only expressed DNA sequence which is helpful in EST generation.

3. CDNAs to ESTs

Researchers then sequence a few hundred nucleotides from either end of the molecule to create two different kinds of ESTs, once an expressed gene has been isolated in terms of cDNA [5]. 5' EST is produced by sequencing the beginning portion of the cDNA. 5' EST is acquired from the portion of a transcript that typically codes for a protein. This region tends to be conserved across the species and do not change much within a family of gene. It is a group of closely related genes whose product is similar protein. Sequencing of the last portion of the cDNA molecule is known as 3' EST. Because these ESTs are produced from the 3' end of a mRNA, therefore; they are likely to fall within untranslated regions (UTRs; a part of a gene that is not translated into protein) or non-coding region. It leans to exhibit less cross species conservation than do coding sequences of DNA/mRNA. ESTs are generated by sequencing cDNA with the help of forward and reverse sequencing primer, which itself is synthesized from the mRNA molecules during cellular process [6]. mRNA does not contain sequences from the regions between genes, nor from the non-coding (introns) that are present within in many parts of the genome as an interspersed part. An overview of how EST is produced from cDNA is shown in figure 2 that depicts the finding of EST from cDNA of particular gene with the help of primers.

Fig. 2 : An overview of how ESTs are generated from cDNA (adopted from <http://www.ncbi.nlm.nih.gov/>)

4. Primers for the amplification of DNA

Primer is a small and single strand of DNA bases that enables DNA to be replicated into another. It principally jump starts the process of DNA replication. Primers are used mainly in Polymerase Chain Reaction (PCR). PCR is processes which reproduce DNA to make millions of copies that means PCR amplify the DNA fragment into millions of copies [7]. A set of primer is required in PCR which is known as forward primer and reverse primer. Each primer is single stranded DNA and is designed to match a specific piece of template of DNA. The specificity arises from the fact that each DNA base can only pair with one other DNA base i.e., adenine (A) pairs only with thymine (T) in DNA and uracil (U) in RNA, whereas guanine (G) pairs only with cytosine (C). The primer must bind to the right piece of DNA and the bases must match in order for copies of DNA to be made. If the matching of base pair occurs, then DNA polymerase enzyme can bind and amplify the DNA fragment [8]. DNA polymerase adds new base pair and makes another copy of the DNA. If the primer does not match the DNA template, then the DNA polymerase can not bind and no copies of DNA will be made. To get exactly the right order of A, T, G, and C, we can order primers from companies that can cord together the sequence that we want or we can synthesize primer also for PCR with the help of primer synthesizer. In PCR technology, the gene sequence of the organism is known at the DNA level so that primers can be designed on that basis. A piece of the

DNA can be opened/unfolded by heat, and after cooling of the unfolded DNA, two primers (forward and reverse) will attach with specific regions of the single stranded DNA. A heat stable DNA polymerase will complete the complementary strand of DNA in the presence of nucleotides from the 3' end and the 5' end. Thus, in one PCR cycle, one molecule of DNA will become two, and after another cycle, two will become four and so on. If process continues, one molecule of DNA can be amplified to 106 in few hours [9]. The genetic sequence of the DNA must be known, and the appropriate primers must be developed for successful use for this powerful tool.

5. ESTs: a tool for mapping and discovery of gene

To find his destination a person need a map while driving a car, in the same way researchers need genome maps to search new gene. Genome map helps them to navigate through the billions of nucleotides to search a particular gene. These nucleotides are the make up the genome of an organism. It must include reliable landmarks or markers for a map to make navigational sense [10]. Sequence Tagged Site (STS) is currently the most powerful mapping technique. It has been used to generate many genome maps [11]. An STS is a short DNA sequence that is easily recognizable and occurs only once in the genome. The 3' ESTs serve as a common source of STSs, because of their likelihood of being unique to a particular species and provide the additional feature of pointing directly to an expressed gene.

ESTs are prevailing tools in the hunt for known genes because they greatly reduce the time required to locate a gene. ESTs represent a copy of interesting part of a genome, which is expressed in cell/tissue. They have proven themselves again and again as powerful tools in the pursue for genes involved in heredity and diseases. ESTs also have a number of practical advantages in that their sequences and can be generated rapidly and inexpensively. Only one sequencing experiment is needed per each cDNA generated. They do not have to be checked for sequencing errors because mistakes do not prevent identification of the gene from which the EST was derived. Using ESTs, scientists have rapidly isolated some of the genes involved in a particular disease. Using this approach for finding a gene responsible for disease, researchers first use observable biological clues to identify ESTs that may correspond to gene candidate for a disease. Researchers then scrutinize the DNA of disease patients for mutations in one or more of these candidate genes to confirm gene

identity. Scientists have already isolated genes involved in Alzheimer's disease, colon cancer, and many other diseases using this method.

6. Search and access of ESTs

They must be organized in a searchable database that also provides access to genome data for ESTs to be easily accessed and useful as gene discovery tools. Because of their speed (with which they may be generated), utility, and the low cost associated with this technology, many individual researchers as well as large genome sequencing centers have been generating thousands of ESTs for research and public use. Once an EST was generated, researchers submit their tags to GenBank (a NIH sequence database operated by NCBI, USA) [12]. It became difficult to identify a sequence that had already been deposited in the database due to the rapid submission of so many ESTs. It was becoming increasingly apparent to NCBI investigators that if ESTs were to be easily accessed and useful as gene discovery tools, ESTs are needed to be organized in a searchable database that also provided access to other genome data. Therefore, scientists at NCBI developed a new database designed to serve as a collection point for ESTs in 1992, which is known as dbEST [13]. Once an EST that was submitted to GenBank, it is screened and annotated and then it was deposited in dbEST.

Scientists at NCBI create dbEST to store, organize, and provide access to the researcher and public. EST data is accumulated and continues to grow daily in dbEST. A researcher can access not only data on human ESTs but information on ESTs from over 300 other organisms as well using dbEST [13]. NCBI scientists annotate the EST record with any known information whenever possible. For example, if an EST matches a DNA sequence that codes for a known gene with a known function, that gene's name and function then they are placed from the EST record. Annotating EST records allows public scientists to use dbEST as an avenue for the discovery of new genes. Any interested person can conduct sequence similarity searches against dbEST by using a database search tool, such as NCBI's BLAST [14]. Because a gene is expressed in terms of mRNA, ESTs ultimately derived from this mRNA many times and may be redundant. That is why there may be many identical, or similar, copies of the same EST. Such redundancy and overlap means, when someone searches dbEST for a particular EST, they may retrieve a long list of tags, many of which may represent the same gene. It is time consuming to search all of these identical ESTs. NCBI investigators have developed the UniGene database to resolve the

redundancy and overlap problem [15]. Automatically UniGene partitions GenBank sequences into a non redundant set of gene oriented clusters.

7. Discussion

It is widely recognized that the generation of ESTs constitutes a competent strategy to identify genes as a result of technological advancement. It is important to acknowledge that there are several limitations associated with the EST approach despite its enormous advantages. First, it is very tricky and difficult to isolate mRNA from some cell types and tissues. This results in a paucity of data on certain genes that may only be found in a particular tissues or cell types. Second is that important gene regulatory sequences may be found within an intron. ESTs are small segments of cDNA free from introns, which is generated from mRNA. Introns been removed totally from EST. But valuable information may be lost by focusing only on cDNA sequencing and this is the limitations of EST. Despite these limitations, ESTs continue to be invaluable in characterizing the human genome, as well as the genomes of other organisms. They have enabled the mapping of many genes to chromosomal sites and have also assisted in the discovery of many new genes and thus pave the way to new horizons in the present and future genetic research.

Acknowledgment

Author is thankful to National Institute of Technology, Raipur (CG), India for providing facility and resources for this work.

References

- [1] Inaba K, Padma P, Satouh Y, Shinl T, Kohara Y, Satoh N, Satou Y. 2002. EST analysis of gene expression in testis of the ascidian *Ciona intestinalis*. *Mol Reprod Dev* 62: (4) 431.
- [2] Bailey J, Manoil C. 2002. Genome-wide internal tagging of bacterial exported proteins. *Nat Biotechnol* 20: (8) 839.
- [3] Van der Hoeven R, Ronning C, Giovannoni J, Martin G, Tanksley S. 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* 14: (7) 1441.
- [4] Nagaraj SH, Gasser RB, Ranganathan S (2007). "A hitchhiker's guide to expressed sequence tag (EST) analysis". *Brief Bioinformatics* 8 (1): 6–21.

- [5] Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. (1991) Complementary DNA sequencing: expressed sequence tags and the human genome project. *Science* 252, 1651-1656
- [6] Parkinson, J., Guiliano, D. and Blaxter, M. (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinf* 3, 31.
- [7] Bartlett, J. M. S.; Stirling, D. (2003). "A Short History of the Polymerase Chain Reaction". *PCR Protocols. Methods in Molecular Biology* 226 (2nd ed.). pp. 3–6.
- [8] Lawyer FC, Stoffel S, RK, et al. (1989) Isolation, characterization, and expression in *Escherichia coli* of the DNA polymerase from *Thermus aquaticus*. *J Biol Chem*, 264: 6427–6437.
- [9] Kainz P. (2000) The PCR plateau phase- towards an understanding of its limitations. *Biochem Biophys Acta*, 1494: 23–27.
- [10] Khowaja, F. S., Norton, G. J. Courtois, B. and Price, A. H. 2009. Improved resolution in the position of drought-related QTLs in a single mapping population of rice by meta-analysis. *BMC Genomics.*, 10: 276-290.
- [11] <http://www.ncbi.nlm.nih.gov/probe/docs/techsts/>
- [12] <http://www.ncbi.nlm.nih.gov/>
- [13] Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) dbEST - database for "expressed sequence tags". *Nat Genet* 4, 332-333.
- [14] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* 215, 403-410. 27.
- [15] <http://www.ncbi.nlm.nih.gov/unigene>