International Journal of Basic and Applied Biology
p-ISSN: 2394-5820, e-ISSN: 2349-2539, Volume 3, Issue 1; January-March, 2016, pp. 22-24

http://www.krishisanskriti.org/Publication.html

Expressed Sequence Tags and Gene Prediction in Cotton Genome

Neeta Maitre[1] and Manali Kshirsagar[2]

[1,2]G.H.Raisoni College of Engineering, Nagpur
E-mail: [1]neeta.maitre@gmail.com, [2]manali_kshirsagar@yahoo.com

Abstract—*Cotton is a highly commercial value crop. It is a major fiber crop of global importance and found in more than 70 countries worldwide. Expressed sequence tags are short sequences of mRNA and can serve as the low-cost alternative for gene prediction. Gene prediction relates majorly to the finding of Open reading frames (ORFs) with the help of introns and exons. Expressed sequence tags thus can help in local and limited search for ORFs.*
Cotton genome is a eukaryotic genome and majorly found in 4 breeds for cultivated cottons. Gene prediction in a given genomic sequence can be addressed by aligning it with available ESTs. As ESTs are shorter sequences than DNA sequences, the method can be employed in a faster way than any other DNA sequence mapper. ESTs being short snapshot of cDNA , challenges are observed in cases of false positive intergenic region and false negative intergenic region.

1. INTRODUCTION

Bioinformtics is the blend of biology, information sciences and mathematics. It plays an important role in genomic databases and its analysis. Agriculture field is no longer away from the tools and techniques used in bioinformatics to enhance crop quality.

India has the largest cotton growing area in the world with contribution of about 16% of the global cotton production. Most commercially cultivated cotton is derived from two species, *G.hirsutum* (Upland cotton, 90% of world plantings) and *G.barbadense* (Pima, or Long-staple cotton). Two other species, *G.arboreum* and *G.herbaceum*, are indigenous to Asia and Africa and are popularly referred as desi cottons in India.[1].

Expressed Sequence Tag, is a tiny portion of an entire gene, a fragment of a cDNA clone that has been sequenced. The process by which ESTs are manufactured requires the construction of an mRNA library.

2. EST GENERATION

As per the diagram (Fig. 1), it is observed that a typical EST sequence is only a very short copy of the mRNA itself and is highly error prone, especially at the ends. The overall sequence quality is usually significantly better in the middle. Vector and repeat sequences either in the end or rarely in the middle are excised during EST pre-processing.

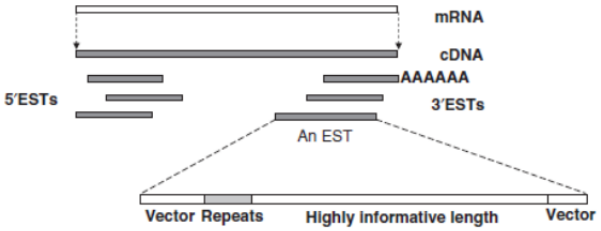

Fig. 1: Characteristics of EST sequence[2]

3. GENERIC ANALYSIS OF ESTS

Steps involved in generic analysis of ESTs are:

- EST preprocessing- It is aimed at reduction in overall noise in EST data. Low quality, singleton, very short ESTs are discarded. Repeat masking is also done in this phase.
- EST clustering – overlapping ESTs from the same transcript of a single gene are collected and are framed as a single cluster to reduce redundancy
- EST assembly – Assembler programs like CAP3 are used to assemble ESTs and these critically evaluate clustering used for the same.
- Database similarity searches – NCBI provides a set of tools for database similarity searches.
- Translation of ESTs – This involves identification of open reading frames from consensus EST sequences to enhance gene prediction.

4. OPEN READING FRAMES

An open reading frame is a portion of a DNA molecule that, when translated into amino acids, contains no stop codons. The genetic code reads DNA sequences in groups of three base pairs, which means that a double-stranded DNA molecule can read in any of six possible reading frames--three in the forward direction and three in the reverse. A long open reading frame is likely part of a gene. An ORF is a sequence

of DNA that starts with start codon “ATG” (not always) and ends with any of the three termination codons (TAA, TAG, TGA).

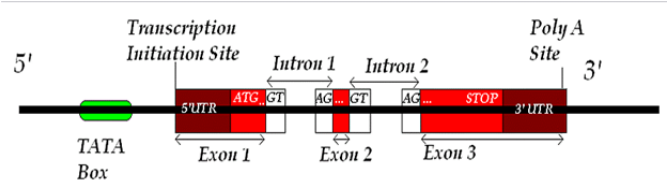

Fig. 2: Open reading frame

For example, the following sequence of DNA can be read in six reading frames. Three in the forward and three in the reverse direction. The three reading frames in the forward direction are shown with the translated amino acids below each DNA sequence. Frame **1** starts with the "a", Frame **2** with the "t" and Frame **3** with the "g". Stop codons are indicated by an "*" in the protein sequence. The longest ORF is in Frame **1**.

5' 3'

atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa

1 atg ccc aag ctg aat agc gta gag ggg

M P K L N S V E G

ttt tca tca ttt gag gac gat gta taa
F S S F E D D V *

2 tgc cca agc tga ata gcg tag agg ggt

C P S * I A * R G

ttt cat cat ttg agg acg atg tat
F H H L R T M Y

3 gcc caa gct gaa tag cgt aga ggg gtt

A Q A E * R R G V

ttc atc att tga gga cga tgt ata
F I I * G R C I

ORF may contain introns as well while the Coding Sequence (CDS) is the actual region of DNA that is translated to form proteins.

Accuracy of gene prediction also depends on accuracy of exons. Exons can be divided into three basic types:

- Initial exons - ORFs delimited by a start site and a 5'(donor) site
- Internal exons – ORFs delimited by a 3' (acceptor) site and a 5' (donor)site.
- Terminal exons – ORFs delimited by a 3'(acceptor) site and a stop codon.

5. GENE PREDICTION METHODS

ESTs offer a rapid and inexpensive route to gene discovery, reveal expression and regulation data [3], highlight gene sequence diversity and splicing [4].

- Searching by signal: analysis of sequence signals potentially involved in gene specification
- Searching by content : the analysis of regions showing compositional bias that has been correlated with coding regions
- Homology- based gene prediction : comparing sequence of interest against known coding sequences
- Comparative gene prediction: comparing sequences of interest against anonymous genomic sequences.

Predicting location of genes in genomic sequences through a combination of one or more of the above approaches.

6. STEPS FOR COMPUTATIONAL GENE PREDICTION IN COTTON GENOME

- Identifying and scoring suitable splice sites and start and stop signals along the query sequence
- Predicting candidate exons , as deduced through the detection of these signals
- Scoring these exons as a function of signals used to detect exons and on coding statistics computed on the putative exon sequence itself
- Assembling a subset of these exons candidates into a predicted gene structure

7. DISCUSSIONS

Intergenic region is a stretch of DNA sequences located between genes. They are the subsets of non-coding DNA. Taking into account a set of known genes and gene prediction methods, results can be obtained for true positives (TP), false positives (FP), true negatives

(TN), and false negatives (FN) for N number of pairs.

To calculate

– Specificity = TP/(TP+FN)

– Sensitivity = TN/(TN+FP)

– Correlation coefficient = [(TP)(TN)-(FP)(FN)]/

SQRT[(TN+FP)(TP+FP)(TP+FN)(TN+FN)]

8. CONCLUSION

Cotton is an immensely important crop for the sustainable economy of the country and livelihood of the Indian farming community. Advanced techniques applied to genomic information of this crop will help in getting good quality fabric. Thus EST data sets have been utilized to complement genome sequencing and ultimately gene prediction. Both functional and evolutionary information can be inferred from customized queries and alignments.

REFERENCES

[1] Shivashankar H. Nagaraj, Robin B. Gasser and Shoba Ranganathan , “A hitchhiker’s guide to expressed sequence tag (EST) analysis”,BRIEFINGS IN BIOINFORMATICS . VOL 8. NO 1. 6 May 23, 2006

[2] “Biology of cotton” , Department of Biotechnology, Ministry of Science and technology, Government of India

[3] Vasmatzis, G., M. Essand, U. Brinkmann, B. Lee, and I. Pastan. 1998. Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. Proc. Natl. Acad. Sci. USA 95(1):300-304 .

[4] Wolfberg, T.G. and D. Landsman. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Research 25(8):1626-1632.

[5] Mohammadreza Ghodsi, Bo Liu and Mihai Pop, “DNACLUST: accurate and efficient clustering of phylogenetic marker genes” , BMC Bioinformatics

[6] Weizhong Li, Limin Fu, Beifang Niu, SitaoWu and JohnWooley, “Ultrafast clustering algorithms for metagenomic sequence analysis” , May2012

[7] Book: Bioinformatics- a practical guide to the analysis of Genes and Proteins by Andreas Baxevanis, B.F. Francis Ouellette