Expressed Sequence Tags and Gene Prediction

Neeta Maitre
Department of Computer Science and Engineering
G. H. Raisoni College of Engineering, Nagpur, Maharashtra, India

Dr.M.M.Kshirsagar
Department of Computer Technology
Yeshwantrao Chavan College of Engineeirng, Nagpur, Maharashtra, India

Abstract. Gene prediction is of prime importance as far as genetics is concerned. This leads to application of computational algorithms to the biological data. Biological sequences majorly considers genomics and proteomics. DNA sequences are thus processed in order to locate genes and then study of their functions is initiated. Expressed sequence Tags may be used to identify gene transcripts, and are instrumental in gene discovery and in gene-sequence determination. Agriculture is the field where genetics majorly plays a vital role. Crops and their suitability, resistance to pest and diseases are of prime importance in agriculture. This research paper talks about the validation of use of EST in gene prediction and is analyzed for the cotton, Gossypium Raimondii dataset.

Keywords – Expressed sequence tags (EST), gene prediction, Gossypium Raimondii, Bioinformatics

I. INTRODUCTION

The genome sequence is an organism's blueprint: the set of instructions dictating its biological traits.[1] In eukaryotes, a gene is a combination of coding segments (exons) that are interrupted by non-coding segments (introns).This makes computational gene prediction in eukaryotes even more difficult. Prokaryotes (e.g. bacteria) don't have introns - their genes are contiguous.[2]

cDNA or complementary DNA is reversely transcribed from mRNA using reverse transcriptase enzyme. cDNA is widely used in cloning of genes in eukaryotes.
Expressed sequence tags (EST) are the short subsequences of cDNA (complimentary DNA) which contain exon region and thus mostly represent expressed genes. The length of ESTs is limited to 200-500 bps which makes them easy to handle and process.

The ESTs can be generated by following steps [3]:

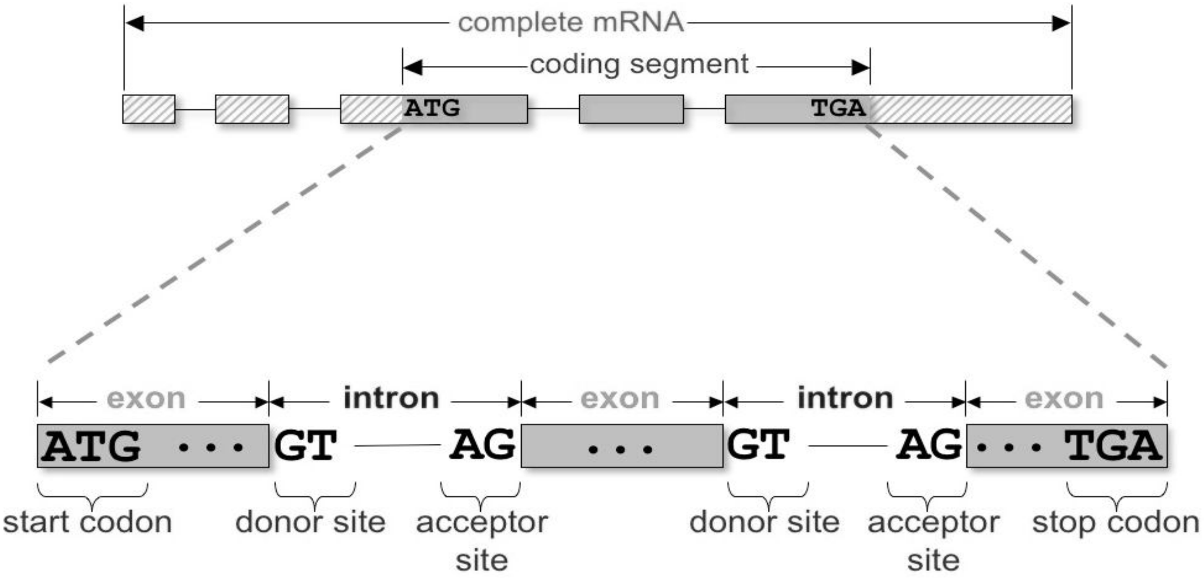
Transcription of Genomic DNA: Genomic DNA is first transcribed to generate Nascent mRNA followed by splicing of synthesize perfect mRNA.
Reverse transcription of mRNA: mRNA can also be directly isolated from the species by using different kits (e.g. RNAgent Promega). mRNA synthesized undergoes reverse transcription to form cDNA library.
Generation of ESTs: From the cDNA library 5' or 3'-ESTs are generated by cDNA end sequencing. 5' EST is formed from a region of transcript which forms protein whereas the ending portion of cDNA forms 3'EST. Assembly and organization of ESTs: The constructed ESTs can then be assembled separately in multimember sequence assembly, Bridged sequence assembly and small clusters on the basis of size of ESTs.

The rest of the paper is organized as follows. Proposed combinational technique and gene prediction techniques are explained in section II. Experimental results are presented in section III. Concluding remarks are given in section IV.

II. PROPOSED TECHNIQUE

A. Gene Prediction Techniques –

Gene prediction in a given biological sequence can be done in two ways. One way is, ab initio technique which calls for locating gene in a given sequence by the existence of gene signals, which include start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites, and polyadenylation (poly-A) sites. Finding Open Reading Frames (ORF) is the most popular one. ORFs are the frames on the DNA strand where the evidence of the gene is found this is necessitated by the presence of start and stop codons. The start codon is ATG and the stop codons could be TAA, TAG, TGA . The portion of nucleotide between start and stop codon is considered to be potential sequence encoding genes. This ultimately governs the functional proteins.



From “Advancing the State of the Art in Computational Gene Prediction”, by William H. Majoros, Uwe

Figure 1. Ab initio based gene prediction using Open Reading Frames (ORF)

Figure 2.

B. Combinational Approach –

For validating the usage of Expressed sequence Tags, we proposed a technique which uses a combination of both, ab initio and the direct method. Direct method is applied by performing a local alignment search by using Basic Local Alignment Search Tool (BLAST) and the generated results are subject to the ORF Finder.

BLAST is a collection of sequence alignment programs from National Center for Biotechnology Information (NCBI) that uses the same heuristic approach to identify the best local alignments between the input query and sequences in the target database/sequences. BLAST uses statistical methods to compare a DNA or protein input sequence (a.k.a. “query sequence”) to a database of sequences (“subject sequences”) and return those sequences that have a significant level of similarity to the query sequence[6]. The BLAST algorithm calculates similarity scores for local alignments (i.e., the most similar regions between 2 sequences) between the query sequence and subject sequences using specific scoring matrices, and returns a table of the best matches (“hits”) from the database. The hit table includes several useful pieces of information, including the similarity score, query coverage (percent of the query sequence that overlaps the subject sequence), Expect-value and max identity (percent similarity between the query and subject sequences over the length of the coverage area).

III. Experiment and Result

Gossypium Raimondii is a Diploid cotton species share a common chromosome number (n = 13). A case study of Gossypium Raimondii is done in order to validate ESTs through evidence based approach of ORFs. The ESTs of G. Raimondii are BLAST against Gossypium Raimondii genome (taxid:29730). Genome size varies from 880 Mb to 2500 Mb.

BLAST search is performed for individual ESTs and are searched for local alignment against its genome. The results are noted for total number of BLAST hits. The separate results are also maintained for high degree of identity (>95%) and a considerable query cover and Expect value.

The ESTs then are further processed by ORF finder in order to get reading frames for locating genes. 96 % results on the dataset under consideration shown existence of ORFs corresponding to BLAST hits. This ultimately refers to the evidence of potential gene locations. In remaining 4% , only 61% have shown existence of ORFs. In 39% cases , ESTs that don’t have a match and also show that there are no ORFs. Thus, in totality 97% of the results validated the use of ESTs in gene prediction. The graph shows actual data analyzed along with the values of total BLAST hits and total ORFs found for the given EST.

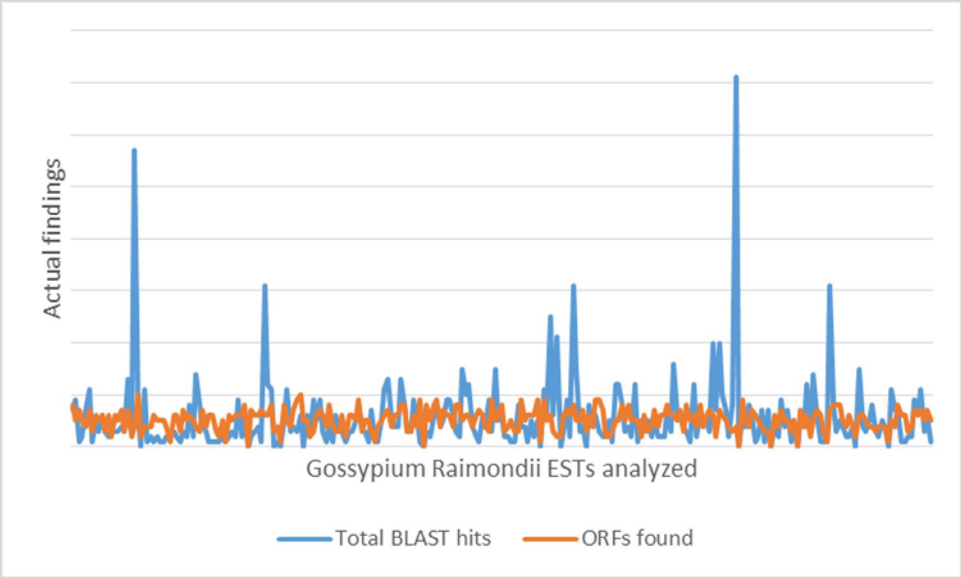


Figure 3. Analysis of Gossypium Raimondii ESTs

IV.CONCLUSION

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. Agriculture being the main occupation in India must be enhanced with the advanced techniques of bioinformatics. Cotton is one of the most economically important crop plants worldwide. This research paper summarizes work done to relate basic concepts of bioinformatics and its applications to the genomic sequence analysis of cotton. Thus, validation approach for ESTs with the help of ORF computation is found to be crucial and instrumental in extensive usage of ESTs in gene prediction.

REFERENCE

[1] Jennifer Harrow, Alinda Nagy et al. “Identifying protein-coding genes in genomic sequences”, *Review in* Genome Biology 2009, 10:201 (doi:10.1186/gb-2009-10-1-201) .

[2] Sergei L Kosakovsky Pond " Computational Gene Prediction," CSE/BIMM/BENG 181 MAY 24, 2011

[3] Dr. Vikash Kumar Dubey, “Expressed Sequence Tags” lecture notes in IIT Guwahati

[4] William R. Pearson, " An Introduction to Sequence Similarity (“Homology”) Searching ", Curr Protoc Bioinformatics. 2013 June

[5] Jin Xiong, "Essential Bioinformatics", Texas A&M University, Cabridge University Press

[6] Todd Osmundson," A brief tutorial on BLAST ", ESPM 150/290, 2011.

[7] NCBI Handbook: BLAST. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook&part=ch16#A614

[8] NCBI: The Statistics of Sequence Similarity Scores. http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html