Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Web Database**

AVATAR: A database for genome-wide alternative splicing event detection using large scale ESTs and mRNAs

Fang Rong Hsu[1*], Hwan-You Chang[2], Yaw-Lin Lin[3], Yin-Te Tsai[3], Hui-Ling Peng[4], Ying Tsong Chen[4], Chia Yang Cheng[5], Min Yao Shih[6], Chia-Hung Liu[6], Chin-Feng Chen[7]

[1]Bioinformatics Research Center, Department of Information Engineering and Computer Science, Feng Chia University, Taiwan; [2]Department of Life Science, National Tsing-Hua University, Hsin-chu, Taiwan; [3]Department of Computer Science and Information Management, Providence University; [4]Department of Biological Science and Technology, National Chiao-Tung University; [5]Department of Computer Science, National Tsing-Hua University, Hsin-chu, Taiwan; [6]Department of Bioinformatics, Taichung Healthcare & Management University; [7]Department of Information Engineering, Taichung Healthcare & Management University; No. 100 Wenhwa Rd., Seatwen, Taichung, Taiwan 40724, R.O.C;

Fang Rong Hsu* - Email: frhsu@fcu.edu.tw; * Corresponding author

received April 15, 2005; revised April 20, 2005; accepted April 20, 2005; published online April 22, 2005

Abstract:

In the past years, identification of alternative splicing (AS) variants has been gaining momentum. We developed AVATAR, a database for documenting AS using 5,469,433 human EST sequences and 26,159 human mRNA sequences. AVATAR contains 12000 alternative splicing sites identified by mapping ESTs and mRNAs with the whole human genome sequence. AVATAR also contains AS information for 6 eukaryotes. We mapped EST alignment information into a graph model where exons and introns are represented with vertices and edges, respectively. AVATAR can be queried using, (1) gene names, (2) number of identified AS events in a gene, (3) minimal number of ESTs supporting a splicing site, etc. as search parameters. The system provides visualized AS information for queried genes.

Availability: The database is available for free at http://avatar.iecs.fcu.edu.tw/

Keywords: alternative splicing; EST; sequence alignment; mRNA; protein diversity; database; human; eukaryotes

Background:

Alternative splicing (AS) is an important mechanism for functional diversity in eukaryotic cells. AS allow processing of one pre-mRNA into different transcripts in a cell type. This results in protein diversity with each protein having distinct function. **[1, 2, 3]** To address this problem we used EST (short, single pass cDNA sequences generated from randomly selected library clones produced in a high throughput manner from different tissues, individuals and conditions) and mRNA sequences to detect AS variants. The detected variants (using 5,469,433 EST and 26,159 mRNA sequences) were stored in a database called AVATAR.

Although, AS databases are available in the public domain, not many contain AS information for multiple eukaryotes (a comparison summarized in AVATAR web site). Therefore, it is important to document AS information for multiple eukaryotes. Hence, we developed AVATAR containing AS information for six eukaryotes. Here, we describe AVATAR development, its content and utility.

Methodology:

Dataset used:

The dbEST database (Jan 16, 2004) at NCBI contains nearly 5.4 million human EST sequences and this dataset is used in the current analysis. **[4]** The human genome sequences (CONTIG build 3.4) in Genbank format is obtained from NCBI. **[5]** Gene information and mRNA sequence were downloaded from the NCBI RefSeq project.

Identification of AS: The identification of AS in AVATAR is performed in three steps (described below) and is illustrated in Figure 1.

Step 1: Alignment of EST and mRNA with the genome sequence: EST sequences were aligned to the whole genome sequence using Mugup. **[6]** Mugup is a sequence alignment program developed in Windows platform. This procedure identified splice sites in the ESTs (Figure 1 panel A and B). The matched regions and gaps correspond to exons and introns, respectively. EST and mRNA alignments with scores greater than 94% were used for further analysis.

Step 2: Clustering EST and mRNA: EST and mRNA were clustered according to their location in the genome (Figure 1 panel C). EST and mRNA with overlapping regions were then assembled together.

Step 3: Detection of AS sites: The mapping of EST alignment with genome sequence to intron positions helps to identify skipped exons and included exons.

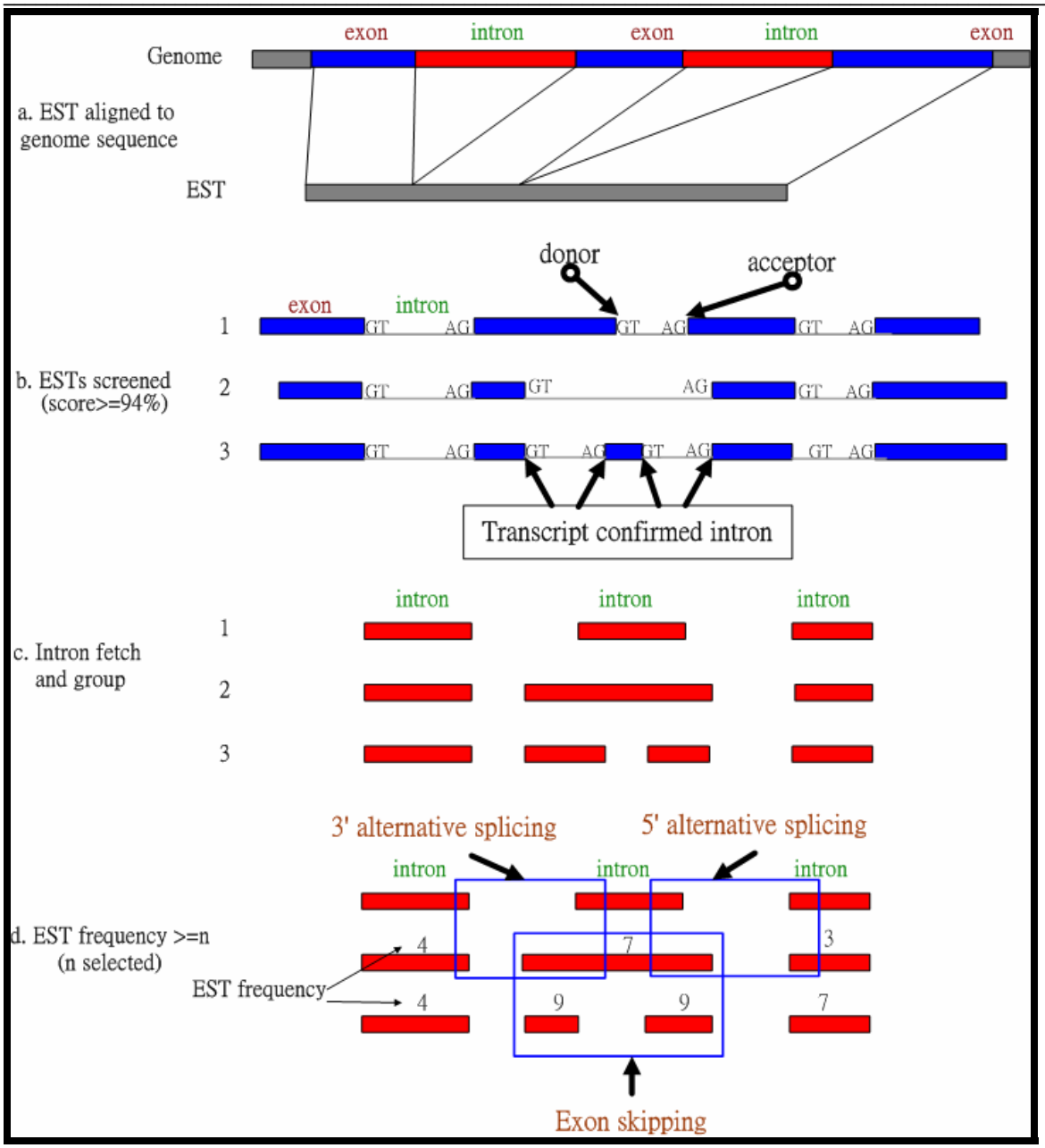

Figure 1: Process of EST alignment and screening. (a) Search the genomic location for each EST. (b) Screening ESTs scores with greater than 94%. (c) Grouping intron by splicing site matched within 3 bp. (d) Detection of AS sites.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Web Database**

Database statistics:

Organism	Exon skipping	3' AS	5' AS	Total
Homo sapiens	5800	3227	3213	12240
Mus musculus	2772	1504	1488	5764
Rattus norvegicus	158	145	162	465
Drosophila melanogaster	8	100	106	214
Caenorhabditis elegans	7	50	63	120
Arabidopsis thaliana	2	59	76	137

Searching AVATAR:
AVATAR can be queried using keywords. The keywords include accession number, gene name, gene isoform, gene location, cytogenetic locations, chromosome number and number of AS events. The database search produces AS visuals for queried gene.

Utility to the Biological Community:
AVATAR is a collection of AS information for 6 eukaryotic organisms. The database can be queried simultaneously for 6 organisms. It can also be searched using gene names and desired number of AS events. EST sequences are error prone resulting in the detection of aberrant transcripts. Frequency of EST alignment at a specific site provides improved detection in AVATAR.

Caveats:
AS information on paralogous genes in eukaryotic genomes are not included in AVATAR due to the difficulty in identifying their corresponding chromosomal locations using EST sequences.

Future Developments:
New EST sequences are generated in laboratories every day. Hence, it is a time consuming to keep AS databases updated due to the growth of genome and mRNA sequences. Hence, we are in the process of developing a computer agent which can update AVATAR automatically. We also plan to include tumor specific AS data.

Acknowledgement:
This work was supported by National Science Council under Grant, NSC92-3312-B-468-001.

References:
[1] R. E. Breitbart, *et al.*, *Annu. Rev. Biochem.*, 56:467 (1987)
[2] P. J. Grabowski, *et al.*, *Progress Neurobiol.*, 65:289 (2001) [PubMed: 11473790]
[3] B. R. Graveley, *Trends Genet.*, 17:100 (2001) [PubMed: 11173120]
[4] ftp.ncbi.nih.gov/repository/dbEST/gzipped/dbEST.reports.date.no.gz
[5] ftp.ncbi.nih.gov/genbank/genomes/
[6] http://avatar.iecs.fcu.edu.tw/mugup

Edited by M.K. Sakharkar
Citation: Hsu *et al.*, Bioinformation 1(1): 16-18 (2005)