Short Communication

Alternatively Spliced Human Genes by Exon Skipping – A Database (ASHESdb)

Meena Kishore Sakharkara,a,*, Bagavathi S. Perumala, Yun Ping Limb, Lee Pern Cherna,
Yiting Yuanda and Pandjassarame Kangueanea

aNanyang Centre for Supercomputing and Visualization, Nanyang Technological University, Singapore 639798
bBioInformatics Institute, Matrix, 30 BioPolis Street, Singapore

Edited by E. Wingender; received 10 November 2004; revised and accepted 29 November 2004; published 8 December 2004

ABSTRACT: Alternative splicing of mRNA allows many gene products with different functions to be produced from a single coding sequence. Exon skipping is the most commonly known alternative splicing mechanism. A comprehensive database of alternative splicing by exon skipping is made available for the human genome data. 1,229 human genes are identified to exhibit alternative splicing by exon skipping.

Availability: http://sege.ntu.edu.sg/wester/ashes/

KEYWORDS: Alternative splicing, eukaryotic gene structure, exon skipping, multiple splice variants, protein isoforms, protein diversity, skipping patterns, genome wide splicing, full length cDNA sequences, tissue specificity, computational analysis, database

INTRODUCTION

Alternative splicing is the major contributor to protein diversity in human [Smith et al., 2000]. Some genes can generate as many as thousand protein isoforms by alternative splicing [Brett et al., 2000; Liang et al., 2000]. The mechanism of alternative splicing in normal and diseased states is perplexing. Differential joining of exons during alternative splicing is important in detecting genetic disorders [Philips and Cooper, 2000]. Alternative splicing is reported to regulate the sub-cellular localization of divalent metal transporter 1 isoforms [Tabuchi et al., 2002] and the NMDA R1 receptor gene [Grabowski and Black, 2001]. Therefore, a comprehensive knowledge on alternative splicing (mechanism and combinatorial protein diversity) is critical in efficient gene discovery and target validation. Alternative splicing can change the mRNA product in several ways. At its simplest level, an exon can be removed (exon skip), lengthened or shortened (alternative 5′ or 3′ splicing). In recent years several databases have been developed to study alternative splicing. ASDB contains information about protein products of alternatively spliced genes by selecting all Swiss-Prot entries containing the words ‘alternative splicing’

*Corresponding author. Tel.: +65 6790 5836; Fax: +65 6791 1859; E-mail: mmeena@ntu.edu.sg.

Electronic publication can be found in In Silico Biol. 5, 0021 <http://www.bioinfo.de/isb/2004/05/0021/>, 8 December 2004.

1386-6338/05/$17.00 © 2005 – IOS Press and Bioinformation Systems e.V. and the authors. All rights reserved
SpliceDB [Burset et al., 2001], STACKDB [Christoffels et al., 2001], TAP [Kan et al., 2001], and ASAP [Lee et al., 2003] use expressed sequence tags (ESTs) to identify splice variants. However, HASDB [Modrek et al., 2001] and EASED [Pospisil et al., 2004] use mRNA/EST data, while PALSdb uses Unigene cluster and EST for studying alternative splicing [Huang et al., 2002]. ProSplicer uses alignment of proteins, mRNA sequences and ESTs against human genomic DNA sequences to reveal splice variants of genes [Huang et al., 2003]. ASD is a database of computationally delineated alternative splice events as seen in alignments of EST/cDNA sequences with genome sequences, and a database of alternatively spliced exons collected from literature [Thanaraj et al., 2004].

However, identification of splice variants remains tricky and arduous mainly due to large intervening sequences and lack of tissue specific cDNA sequence data. As can be seen majority of currently known splice variants are identified using EST and EST coverage in the protein coding sequence of many genes is still inadequate to predict splicing to a large extent [Modrek and Lee, 2002]. Moreover, there are limitations in accuracy resulting from the single-pass sequencing that has been used to identify ESTs. In this report, we describe a database for alternatively spliced (exon skipping) human genes identified strictly using full-length cDNA sequences. This novel approach makes the detection of splice variants more reliable and accurate. This circumvents the greatest challenges in using EST databases to understand alternative splicing and thereby facilitates the task of comprehending the relationships of these short EST sequences to each other and to other genes.

The database integrates a variety of data for each gene ranging from gene map, gene structure, splice variants and tissue information. Information on mouse orthologs showing exon-skipping patterns is also provided. This database can be used to study the impact of alternative splicing on protein function and could be a useful resource to researchers who have found a new cDNA or human gene and wish to find additional information.

DATABASE CONSTRUCTION AND IMPLEMENTATION


**Human splice variants – Dataset-1**

The human genome chromosomal fragment sequences were subjected to BLASTN against the cDNA data set at an E-value cut-off of $10^{-100}$ using the standard $+1/-3$ scoring scheme [Altschul et al., 1990]. The human genes with two or more cDNA sequence hits were analyzed by mapping the positions of the exons onto the genomic DNA. Genes with two or more cDNA hits with different exon combinations are characterized as exon skipping cases. Excluded (skipped) exons are denoted as ‘0’ and included (non-skipped) exons are denoted as ‘1’. Each entry in the database is hyperlinked to the corresponding GenBank entry at NCBI and to a local file illustrating the putative exon skipping patterns. The identified variants are further classified based on the tissues in which they are expressed.

**Mouse splice variants – Dataset-2**

Similar data was obtained for mouse genome.
Table 1
Distribution of various exon skipping patterns

<table>
<thead>
<tr>
<th>Organism</th>
<th>Total # of genes</th>
<th>Total # of splice variants</th>
<th>Skipping pattern in variants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0*0 (splice variants/gene)</td>
</tr>
<tr>
<td>Human</td>
<td>1229</td>
<td>9049</td>
<td>5347/906</td>
</tr>
<tr>
<td>Mouse</td>
<td>1197</td>
<td>4794</td>
<td>2475/827</td>
</tr>
</tbody>
</table>

Terminal exon skipping – (0*0), (0*1) and (1*0); internal exon skipping – (1*1).

Dataset-3

A BLASTN was performed for dataset-1 against dataset-2 and orthologs were identified. The data are made available through a WWW interface. The database is housed in a POSTGRES.

Interface

A user-friendly query system is provided for efficient search of the data. Thus ASHESdb can be searched with gene name, chromosome number and exon skipping pattern. The results are summarized in tabular and graphical forms with appropriate links for better understanding and visualization of the gene features. In addition, a BLAST [Altschul et al., 1990] query page is also provided for searching an input sequence against the ASHESdb. This allows for checking of homology against the genes showing alternative splicing by exon skipping in the human genome.

Data

Out of a total of 22,948 annotated genes in the human genome, 1,229 genes are identified to exhibit alternative splicing by exon skipping in 9049 splice variants identified using full-length cDNA as a novel approach. Similar analysis on mouse identified 1197 genes to exhibit alternative splicing by exon skipping in 4794 splice variants. The exon-skipping patterns are classified into two groups namely: (1) terminal skipping defined as exclusion of a 5′ and/or 3′ – terminal exon ((0*0), (0*1), (1*0)) and (2) internal skipping defined as exclusion of alternative internal exons – (1*1) (‘*’ represents wild card and can have a value of ‘0’ for excluded and ‘1’ for included exons). In this notation ‘0’ represents excluded exon and ‘1’ represents included exon. We observed that the total number of genes with terminal skipping patterns are 1941 for human and 1835 for mouse with 8835 and 4571 terminal skipping splice variants in human and mouse respectively. The number of genes involved in terminal skipping pattern was maximum for the (0*0) type both in human and mouse (Table 1). These results support the earlier finding that the amount of alternative splicing is comparable between eukaryotes [Brett et al., 2002]. The bias in terminal skipping patterns hints hints at the usage of different promoters (or transcription start sites) and different poly(A) sites.

It was observed that out of 9049 alternatively spliced variants identified 5925 are involved in genetic diseases like carcinoma, melanoma and neuroblastoma etc (data available online). These data hint that splicing signals are subtle and complex and the flexibility of splicing can sometimes be a liability, and several genetic diseases result from changes that cause exon skipping.

The list of genes with information on their exon-skipping patterns and gene structure is available online.
CAVEATS

One must take note of the fact that due to reduced efficiency of the reverse transcription reaction and the lack of an efficient technique for selecting only full-length cDNA, the full length cDNA collection in MGC is not complete for all of the protein diversity in the human genome. This could be one of the factors towards the low detection of true (internal) exon-skipping events in the human genome by this methodology. The lack of full length cDNA matching the genomic data at the specified cutoff could be another factor contributing towards lower detection of exon skipping events. Perhaps as the full length cDNA data increases by large scale sequencing and thorough prediction of gene structures, it may become more feasible to detect almost all exon skipping events in the human genome.

It must also be noted that the traditional gene finding algorithms treat the translation start site as the 5′ boundary of the gene and there are currently no computational tools to predict the non coding first exons or non coding portions of the first exon except where the true full-length mRNA sequences are available [Davuluri et al., 2001]. As this analysis is strictly based on CDS feature in genome data, it does not take into account the first exon and is biased towards internal coding exons of the gene.

Nonetheless, this dataset provides a unique and reliable collection of genes exhibiting alternative splicing by exon skipping in the human genome.

ACKNOWLEDGEMENTS

This manuscript is based on the project supported by A*STAR – BMRC grant #03/1/22/19/242.

REFERENCES


