D46–D55 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj031

ASD: a bioinformatics resource on alternative splicing

Stefan Stamm[1], Jean-Jack Riethoven, Vincent Le Texier, Chellappa Gopalakrishnan, Vasudev Kumanduri, Yesheng Tang[1], Nuno L. Barbosa-Morais[2] and Thangavel Alphonse Thanaraj*

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, [1]University of Erlangen, Institute for Biochemistry, Fahrstrasse 17, 91054 Erlangen, Germany and [2]Faculty of Medicine, Institute of Molecular Medicine, University of Lisbon, 1649-028 Lisbon, Portugal

Received August 23, 2005; Revised and Accepted September 22, 2005

ABSTRACT

Alternative splicing is an important regulatory mechanism of mammalian gene expression. The alternative splicing database (ASD) consortium is systematically collecting and annotating data on alternative splicing. We present the continuation and upgrade of the ASD [T. A. Thanaraj, S. Stamm, F. Clark, J. J. Riethoven, V. Le Texier, J. Muilu (2004) *Nucleic Acids Res.* 32, D64–D69] that consists of computationally and manually generated data. Its largest parts are AltSplice, a value-added database of computationally delineated alternative splicing events. Its data include alternatively spliced introns/exons, events, isoform splicing patterns and isoform peptide sequences. AltSplice data are generated by examining gene-transcript alignments. The data are annotated for various biological features including splicing signals, expression states, (SNP)-mediated splicing and cross-species conservation. AEdb forms the manually curated component of ASD. It is a literature-based data set containing sequence and properties of alternatively spliced exons, functional enumeration of observed splicing events, characterization of observed splicing regulatory elements, and a collection of experimentally clarified minigene constructs. ASD includes a workbench, which is an analysis tool that enables users to carry out splicing related analysis such as characterization of introns for various splicing signals, identification of splicing regulatory elements on a given RNA sequence, prediction of putative exons and prediction of putative translation start codons. The different ASD modules are integrated and can be accessed through user-friendly interfaces and visualization tools. ASD data has been integrated with Ensembl genome annotation project as a Distributed Annotation System (DAS) resource and can be viewed on Ensembl genome browser. The ASD resource is presented at (http://www.ebi.ac.uk/asd).

INTRODUCTION

Alternative pre-mRNA splicing is emerging as one of the most important mechanisms to control eukaryotic gene expression. Recent array data indicate that as much as 76% of genes generate alternatively spliced products (1). Alternative splicing regulates numerous aspects of protein function, such as binding properties, intracellular localization, enzymatic activity, stability and post-translational modifications. Reports in literature indicate that protein isoforms generated by alternative splicing show in most cases only subtle differences. However, in some cases alternative splicing can lead to large functional differences, e.g. by generating dominant negative isoforms (2). Finally, 10–15% of genes could be switched off due to the coupling between nonsense-mediated decay and alternative splicing (3). This indicates that alternative splicing controls both transcript composition and abundance.

Despite intense research, the mechanisms leading to splice site selection are not fully understood. Currently, it is not possible to accurately predict alternative exons from genomic sequences. It is not at all possible to predict the tissue or developmental expression profile of an alternative exon. The major obstacle for an accurate prediction is the lack of conservation in the regulatory sequences of the pre-mRNA that can only be described by consensus sequences or

*To whom correspondence should be addressed. Tel: +44 1223 494650; Fax: +44 1223 494468; Email: thanaraj@ebi.ac.uk

expectation matrices. However, *in vivo*, alternative exons are recognized and regulated with high fidelity, because numerous proteins bind to pre-mRNA and help in exon recognition. Due to the combination of multiple weak protein–protein and protein–RNA interactions, alternative exons can be faithfully recognized (4,5). The importance of proper splicing site recognition is underlined by an increasing number of diseases that are caused or associated with the selection of wrong splicing sites (6,7).

Alternative splicing events have been compiled previously in different databases [reviewed in (8)]. Here, we present the continuation and upgrade of the alternative splicing database (ASD) [the previous version was reported earlier in (9)] as a bioinformatics resource that integrates data on alternative splicing, derived from computational as well as literature based approaches, and bioinformatics analysis tools.

ASD

The different component data sets of ASD

Data generated from computational approaches and data reported in the literature are the two major sources for databases on alternative splicing. The major advantage of data from computational approaches, such as EST comparison, is the large size of the data sets. However, these data sets lack biological information about the alternative splicing events. In contrast, data sets derived from the literature contain biologically relevant information, but these data sets are smaller. In order to combine these two approaches, we carried out two activities namely, (i) we developed a computational pipeline (AltSplice) that generates genome-wide value-added data on alternative exons, and (ii) we developed a procedure (AEdb) that manually collects data on alternative exons from literature. In addition, data on motifs, functions and minigenes described in the literature were collected in databases. We then built an integrated database (Figure 1) from these heterogeneous data resources. The integrated database is named ASD for which we developed query interfaces that are flexible enough to handle the heterogeneity. A single-query bar provides a quick access to all of ASD data, allowing retrieval of data using keyword search or sequence comparison searches. In addition, each data set can be queried using a data set-specific interface. Finally, all data sets can be downloaded as flat file distributions.

The statistics on different data sets of the ASD are listed in Table 1.

AltSplice. AltSplice data are generated by an automated computational pipeline. The basal data includes transcript-confirmed introns/exons, alternative splicing events and isoform splicing patterns. The basal data are generated through computational comparison of EST/mRNA alignments with genomic sequences [see (9,10) for details on the computational

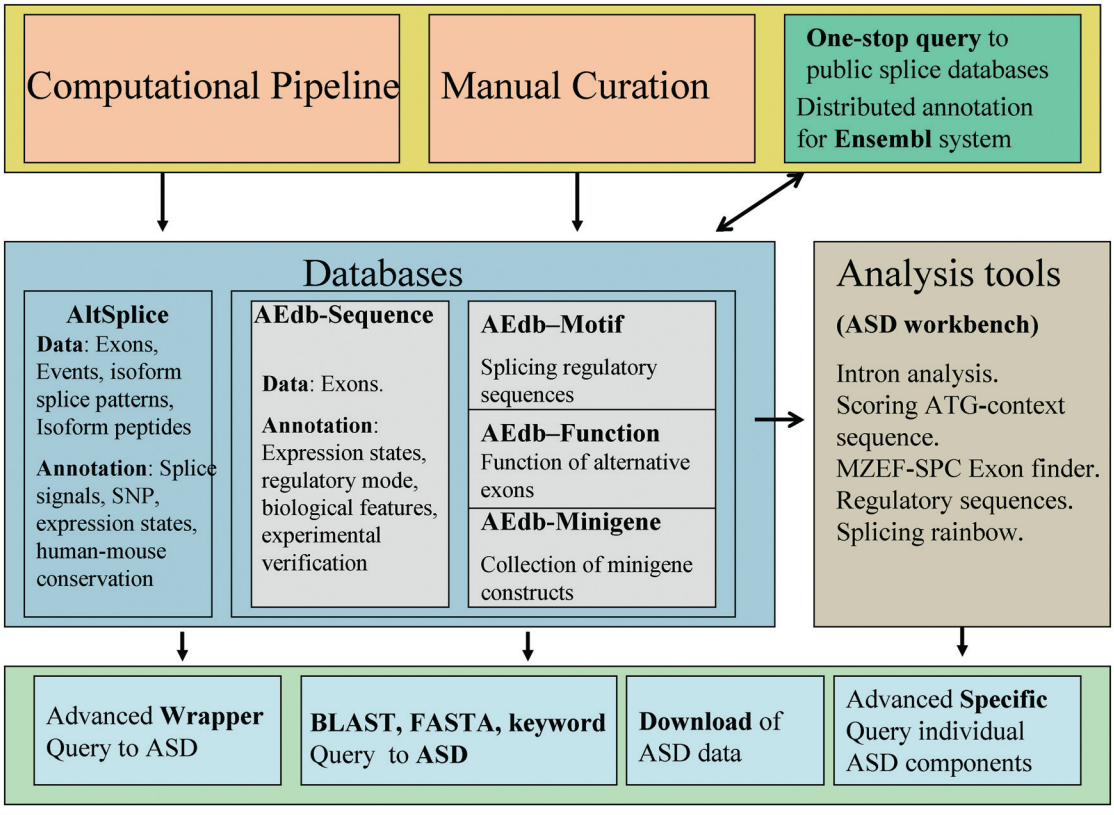


Figure 1. Structure of ASD. We used computational pipelines and manual curation (top, pink) to create the modular databases of ASD (middle, blue). The individual databases are integrated, cross-linked and are available through a variety of interface tools (bottom, sky blue). Currently the databases are the computer generated Altsplice and the manually-curated AEdb-Sequence, AEdb-Motif, AEdb-Function and AEdb-Minigene databases. The ASD data are integrated with Ensembl genome annotation system and is visible from Ensembl genome browser; Publicly-available databases on alternative splicing are accessible from ASD interfaces (top right, dark green) The databases are connected to analysis tools that are collected in the ASD workbench (middle right, grey).

Table 1. Statistics on ASD data

AltSplice data statistics (Release 2 of AltSplice-Human and AltSplice-Mouse)	
Genes with alternative splicing	9929 (61%) of 16 293 human genes
	8211 (50%) of 16 391 (mouse) genes
Alternative splicing events per gene	2.6 (human); 2.2 (mouse)
Alternative splicing patterns per gene	3.9 (human); 3.7 (mouse)
Distribution of the observed event types	Cassette exon (52%)
	Alternate acceptor or donor (27%)
	Intron retention (17%)
	Mutually exclusive (4%)
Flanking exons often undergo extension or truncation	26% of Cassette exon events
	23% of Retained introns
	19% of Mutually exclusive exons
Human-mouse conservation	Orthologous gene pairs = 5176
	Number of conserved events = 1177
	Number of conserved exons = 28 276
Entries with peptide sequence annotation for at least 1 splicing pattern	6848 (human); 4366 (mouse)
Entries with peptide sequence annotation for ⩾2 splicing patterns	2896 (human); 977 (mouse)
Entries for which variant peptide data are available in UniProt	2083 (human); 896 (mouse)
Entries presenting data for ⩾2 isoform peptide sequences (by either of the above two sources)	3994 (human); 1573 (mouse)
Entries associated with AEdb-Sequence database	367 (human); 118 (mouse)
AEdb-Sequence data statistics-2255 entries	
Distribution	Number of entries
Organism distribution	Human (1283); mouse (413); rat (232); drosophila (100); others (227)
Event type distribution	Cassette exon (1281)
	Alternative acceptor or donor (395)
	Intron retention (154)
	Mutually exclusive exons (130)
	Alternative 3′ exon by polyA variant (71)
Regulation associated with disease	295
Regulation associated with development	282
Regulation associated with tissue type	312
Regulation causing frameshift	151
Regulation introducing stop codons	260
Alternative exon being noncoding exon	222
Entries associated with AltSplice	1198 (human and mouse entries)
AEdb-Function data statistics-354 entries	
Functional role	Number of entries
Modulation of protein interaction	136
Internal structural change	119
Novel carboxyl terminus	87
Novel amino terminus	38
Association with disease	81
Intracellular location	76
Enzymatic activity	64
Channel activity	54
Others	37
AEdb-Motif data statistics-255 entries	
Type of regulator sequence	Number of entries
Exon enhancer	97
Exon silencer	44
Intron enhancer	56
Intron silencer	37
AEdb-Minigene data statistics—82 entries	
Distribution	Number of entries
Organism distribution	Human (46); mouse (17); rat (15); others (9)
Splicing mechanism distribution	Cassette exon [single exon, 45; multiple casette exons (3); incremental combinatorial exons (2)]; Alternative acceptor or donor sites (17); Mutually exclusive exons (13); Intron retention (2)
Reported tissue specificity	55
Known regulatory factors	32
Deduced enhancer and silencer sequences	97
Hyperlinks to AEdb-Sequence database	78 (to 105 AEdb-Sequence entries)

methods]. The data are annotated for biological features (such as splicing site characteristics, expression state of the isoforms, allele usage at SNP positions, conservation of intron/exon-events across species and peptide isoforms) by various computational modules that are part of AltSplice pipeline. AltSplice data (Table 1) indicates that up to 61% of human genes (and 50% of mouse genes) undergo alternative splicing. The available transcriptome data indicates that in an average 3.9 isoform splicing patterns can be expressed from a single human gene. Cassette exon events outnumber the other event

types and one in every four cassette exon events is accompanied by extension/truncation of the flanking exons. The number of human-mouse orthologous gene pairs (with data on alternative splicing) present in AltSplice is around 5200, and such a large data set is valuable for studies on evolution of alternative splicing. Finally, AltSplice presents data on isoform peptide sequences for around 4000 human genes, and such a large data set of variant peptides is valuable for studies aimed at deciphering splicing mediated functional and structural changes in proteins.

AEdb-Sequence. AEdb-Sequence is a literature based, manually curated database of alternative exons. We used 'alternative splicing' as a keyword to search PubMed bibliography data and collected information on the following features from the resultant research articles: organism, splicing mechanism, tissue-specificity, regulation during development stages, disease association, regulatory features of the exon and the sequence of the alternatively spliced exon as well as its flanking constitutive exons. It is seen that more than half the number of AEdb-Sequence entries are from human (Table 1). As is in the case of AltSplice data, cassette exon events outnumber other event types. The data set reports splicing events that are specific to cellular states, such as tissue type, development stage and disease state. Roughly 10% of the entries report events that introduce premature stop codons and this data set can serve the studies on nonsense mediated decay of transcripts. Finally, 10% of the reported exons are from non-coding regions of the genes.

AEdb-Function. The function database is a literature based, manually curated database of known functions of the alternative exons. Functional differences between the protein isoforms generated by alternative splicing are enumerated from the literature and are organized into 11 well-defined categories, such as 'Modulation of protein interaction' or 'Internal structural change' (Table 1). An analysis of the function of alternative exons based on this data set has been published previously (2).

AEdb-Motif. Alternative splicing site selection is partially regulated by weak binding of proteins to highly degenerate regulatory sequences. As a first attempt to understand the combinatorial control behind this regulation, we collected splicing regulatory motifs described in literature and expanded upon the previous collections of intronic regulatory sequences (11), exonic regulatory sequences (12,13) and disease-causing mutations (6). The collection reports 153 enhancer sequences and 81 silencer sequences (Table 1). The entries are annotated with value-added information, such as the experimental technique used, the nucleotide sequence of the motif, mutations that are studied and the protein that binds at the motif.

AEdb-Minigenes. A minigene is a genomic fragment that includes the alternative exon and the surrounding introns as well as the flanking constitutively spliced exons. Constructs derived by cloning the insert in an eukaryotic expression vector are increasingly used to study alternative splicing (14,15). We compiled all minigenes described in the literature. The splicing patterns and deduced regulatory sequences are represented in a graphic format. The minigene collection includes 82 entries for which a total of 97 regulatory sequences are ascribed. The reported minigene constructs representing cassette exon events outnumber those for other event types (Table 1). The minigene entries are linked to appropriate entries in AEdb-Sequence data set, which allows the user to quickly identify experimentally useful minigenes by searching the database.

DATA INTEGRATION

Integration of data across the different data sets of ASD

Extensive integration has been made between AltSplice and AEdb-Sequence. Alternative exons and events that are common in AltSplice and AEdb-Sequence are identified and are annotated in both the databases. This allows associating the manually-collected annotations to the computationally generated data in AltSplice. Related entries among AEdb-Sequence, -Function and -Minigenes are identified and are annotated. Table 1 shows that from the 1700 AEdb-Sequence entries, as much as 1200 are associated with AltSplice entries; and as much as 78 of 82 entries from AEdb-Minigenes data set are associated with AEdb-Sequence data set.

Integration with other resources

Ensembl (16) and UniProt (17) are among the most important resources on sequence data. Both include significant data relating to alternative splicing, e.g. Ensembl reports alternative transcripts while UniProt reports curated data on isoform peptide sequences. AltSplice uses Ensembl genes as the starting gene set for deriving splicing patterns. Therefore, the AltSplice data are intrinsically associated with Ensembl annotation of alternate transcripts and related information. Data on peptide variants collected in UniProt are integrated with AltSplice and they complement the set of AltSplice-derived peptide isoform data.

ACCESS TO DATABASES

ASD interfaces

ASD contains heterogeneous data sets—AltSplice is created through computational analysis of gene-transcript alignments, and AEdb is created by manual curation of literature data. Furthermore, there are differences in the extent of annotation and in the adopted vocabularies. We therefore generated interfaces that are flexible enough to handle this heterogeneity and that allow an easy retrieval of information. Different layers of interfaces are available and they provide either a single-point access to all the data sets or advanced searches of individual data sets.

Single-query bar and wrapper interfaces. Both these interfaces provide a quick access to all of ASD data. The single-query bar accepts commonly-used search terms, such as keywords, gene symbols (or their synonyms) and database cross-references. The wrapper interface queries all of ASD data against a given search term; these terms include the above-mentioned commonly used terms and splicing event type. Queries can be selectively restricted to specific sets of gene entries, such as set of human-mouse orthologous gene pairs or set of gene

ALTSPLICE Human : 2 matche(s) [first 10 matches shown]:

ID	Gene symbols	Protein description	AEdb association	Mouse Orthologous Entry
ENSG00000136527	SFRS10 , SRFS10 , TRA2-BETA , HTRA2-BETA , TRA2B	SPLICING FACTOR, ARGININE/SERINE-RICH 10 (TRANSFORMER 2 HOMOLOG, DROSOPHILA) [HOMO SAPIENS]. [SOURCE:REFSEQ;ACC:NM_004593]	2290 , 2289 , 1493 , 1097 , 1492 , 1098	
ENSG00000164548	TRA2A , HSU53209	TRANSFORMER-2 PROTEIN HOMOLOG (TRA-2 ALPHA). [SOURCE:UNIPROT/SWISSPROT;ACC:Q13595]		ENSMUSG00000029617

AEDB-SEQUENCE database : 7 matche(s) [first 10 matches shown]:

ID	Gene symbols	AEdb annotation	Altsplice association	Organism
1097	SILG41 , SIG41 , SFRS10 , TRA2B	TRA2-BETA, EXON III	ENSG00000136527	Homo sapiens, human
1098	SILG41 , SIG41 , SFRS10 , TRA2B	TRA2-BETA1, EXON II	ENSG00000136527	Homo sapiens, human

AEDB-FUNCTION database : 1 matche(s) [first 10 matches shown]:

ID	Gene symbols	AEdb annotation	Organism
467		TRA2BETA(SFRS10) , EXON2	Homo sapiens, human

AEDB-MOTIF database : 5 matche(s) [first 10 matches shown]:

ID	AEdb annotation	Organism
91	FTDP-17 TAU EXON 10	Homo sapiens, human
130	DMD, PSEUDO EXON FROM INTRON	Homo sapiens, human
261	HNRNPG BINDING MOTIF	Homo sapiens, human

Figure 2. Result page of query to all of ASD data. The ASD was queried using the wrapper interface with the term 'tra2a* | tra2b*' and this resulted in the retrieval of data entries from AltSplice-Human (two entries), AEdb-Sequence (seven entries), AEdb-Function (one entry) and AEdb-Motif (five entries). This figure illustrates the integration among the different data sets of ASD—(i) the AltSplice-Human entries are seen associated with entries from AEdb-Sequence, and from AltSplice-Mouse; (ii) the AEdb-Sequence entries are seen associated with entries from AltSplice-Human. These associations are hyperlinked.

entries for which data on isoform peptide sequences is available or integrated set of AltSplice-AEdb entries.

Advanced search query interfaces. The individual data sets differ in terms of the type of data and annotation—e.g. AltSplice reports splicing events, splicing patterns and introns/exons while AEdb-Motif reports splicing regulatory sequences. Thus specialized query interfaces have been built for individual data sets.

Interface for AltSplice. Genes can be queried by chromosomal location, gene names and synonyms, protein keywords and database cross-references [such as EMBL (18) and UniProt accession nos (17), HUGO gene symbols (19), Gene Ontology identifiers (20) and protein identifiers], types of splicing events and allele specificity at SNP positions. Browsers allowing selection of eVOC standard vocabularies for EST library annotation (21) and GO classifiers facilitate querying through expression states and through protein function/process/location. Queries can be selectively restricted to specific sets of gene entries, such as to the set of human-mouse orthologous gene pairs, or to the set of gene entries for which data on isoform peptide sequences is available, or to the integrated set of AltSplice-AEdb entries. A particularly useful query for experimentalists is 'Library Subtraction Tool', which let users retrieve gene entries with splicing patterns that are differentially expressed in different cell states.

Interface for AEdb-Sequence. The data can be queried by gene names and synonyms, database cross-references, type of splicing events and type of regulatory roles, such as introducing premature termination codons or frameshifts. Further, the data can be queried for disease association and developmental specificity.

Interface for AEdb-Function. The data can be queried by gene names, protein keywords and database cross-references. Further, queries based on the functional enumeration of the isoform peptide sequence can be raised by selecting from a predefined list of functional categories (for the list of functional categories see Table 1).

Interface for AEdb-Motif. The interface allows free-text search. The search items include gene names, sequence of

(a)

Altsplice-Human : Entry ENSG00000136527

GENE INFORMATION :	
Gene symbols :	SFRS10 , SRFS10 , TRA2B , TRA2-BETA , HTRA2-BETA
Protein description :	splicing factor, arginine/serine-rich 10 (transformer 2 homolog, Drosophila) [Homo sapiens]. [Source:RefSeq;Acc:NM_004593]
Gene sequence :	Viewthe gene sequence

EVIDENCES :			
AEdb associations :	o Exon level	o Event level	1097 , 1098 , 1492 , 1493 , 2289 , 2290
Uniprot peptide isoforms :	Viewall the peptide isoform sequences from UniProt		
Ensembl transcript structures :	Ensembl predicts mutiple transcript structures for this gene		

Cassette Exon Event

CASSETTE EXON(s)	EVENT TYPE	AEDB ASSOCIATION	CONSERVATION
14280..14413	CCE-EB-3P SCE	1097 , 1493 , 2290	
21518..21577	SCE		

Intron Isoform Event

INTRON ISOFORMS	EVENT TYPE	AEDB ASSOCIATION	CONSERVATION
3190..15387|3190..14279	Alternative donor		

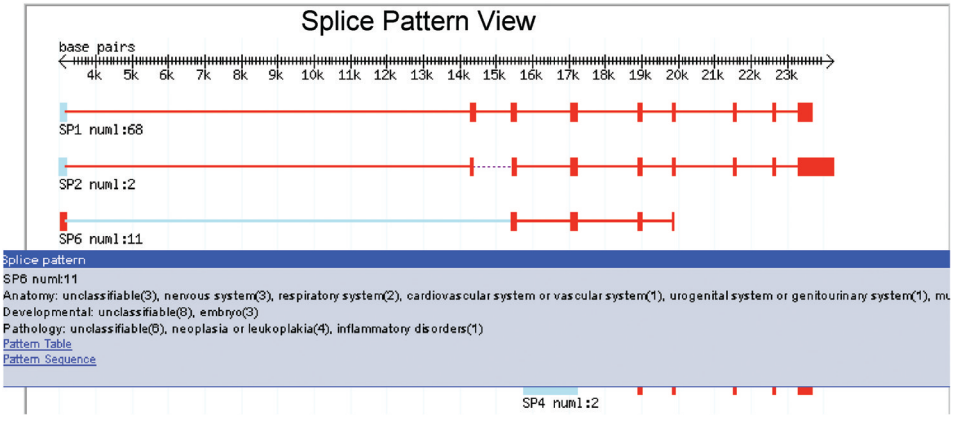
Intron Retention Event

RETAINED INTRON(s)	EVENT TYPE	AEDB ASSOCIATION	CONSERVATION
17219..18887	SIR		

(b)

Splice Pattern Table

PATTERN SEQUENCE	PEPTIDE SEQUENCE	STRUCTURE	CONFIRMING EST/mRNA's	CLONE LIBRARIES	IDENTIFIED SNP's
1	3154..23299 (288 aa)	~3011..3189, 14280..14413, 15388..15550, 17030..17218, 18888..19003, 19827..19910, 21518..21577, 22576..22649, 23289..~23686	176	68	48
2	3154..23299 (252 aa)	~2984..3189, 14280..~14360, ~15433..15550, 17030..17218, 18888..19003, 19827..19910, 21518..21577, 22576..22649, 23289..~24265	2	2	45
3		~14279..14413, 15388..15550, 17030..17218, 18888..19003, 19827..19910, 22576..~22651	4	4	24



(c)

GENE INFORMATION :	
Organism :	Homo sapiens, human
Gene symbols :	
AEdb gene annotation :	tra2beta(sfrs10)
AEdb exon annotation :	exon2
EMBL/Genbank/DDBJ/RefSeq Accession numbers :	U87836 , NM_004593
Ensembl IDs:	ENSG00000136527

BIBLIOGRAPHY INFORMATION :	
Publication :	Stoilov P, Daoud R, Nayler O, Stamm S.(2004) "Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA." Hum Mol Genet.13(5):509-24.
Comments :	It introduces a frameshift and inclusion of this exon generates a mRNA named Tra2beta4 which is not translated.
PubMed IDs :	14709600

FUNCTION INFORMATION :	
Description :	exon2 is skipped in Tra2-beta1. Tra2-beta1 binds to four enhancers present in exon 2, which activates its inclusion. Inclusion of exon 2 generates mRNAs that are not translated into proteins.
Keywords :	alternative splicing,tra2beta,sfrs10,rs domain,clk2,nmd
Roles :	Modulation of protein interaction Transcription factor property

Figure 3. Display of data on alternative splicing of human tra2-beta gene as seen in AltSplice and AEdb-Function data sets. (**a**) This figure presents a display of data sections on Gene Information, Evidences and Splicing events as seen in AltSplice. Gene information section provides hyperlinks to a page listing the gene entry from HUGO Gene Nomenclature database and to a page which lists the sequence of the gene. The evidences section provides hyperlinks to the associated entries from AEdb-Sequence, to pages that list variant peptide sequences for the gene from UniProt or to pages that list the Ensembl transcript sequences for the gene. The events section lists all the splicing events that AltSplice has identified for the gene. Column 1 lists the gene coordinates of alternatively spliced exons/introns. Column 2 indicates whether the event involves modifications in the flanking exons as well; entries are hyperlinked to pages listing detailed information on the event. Column 3 indicates the identifier of the associated entry from AEdb-Sequence (if any) and the entry is hyperlinked. Column 4 indicates the identifier of the orthologous gene (if any) and the coordinates of the exon orthologous to the one presented in column 1; the entry is hyperlinked to the orthologous gene entry. (**b**) This figure presents the textual and graphical display of observed splicing patterns for tra2-beta gene as seen in AltSplice data. Splice Pattern Table: Entry in column 1 is hyperlinked to a page listing the sequence of the splicing pattern. Entry in column 2 gives the coding start and end positions on the gene and the length of the translated peptide sequence and is hyperlinked to a page listing the peptide sequence. Entry in column 3 lists the structure of the splicing pattern as a string of exons (exon boundaries are presented in gene coordinates). Entry in column 4 is hyperlinked to pages listing detailed information on the confirming transcript sequences. Entry in column 5 is hyperlinked to pages listing expression states. Entry in column 6 is hyperlinked to pages listing allele specificity of the splicing pattern. Splice Pattern View: Exons are indicated by boxes and introns by lines. Exons/introns that are variants are indicated in blue color. Browsing the cursor over various elements of the pattern displays pop-up's giving detailed information on the elements. The displayed pop-up in this example shows information on the expression state of Splicing Pattern 6. (**c**) This figure presents data on the functional changes due to alternative splicing in tra2-beta as seen in AEdb-Function data set. The data are organised into three sections namely, gene information, bibliography information and functional information. This figure illustrates the wealth of knowledge captured from literature.

the regulatory motifs and type of regulatory sequence (enhancer or silencer).

BLAST and FASTA searches to ASD. The nucleotide and peptide sequences from ASD can be searched through both BLAST (WU-BLAST2) (http://blast.wustl.edu) and FASTA utilities (22). The objective of these search programs is to identify sequence similarities between novel sequences and alternatively spliced sequences collected in ASD. BLAST reports regions of high similarity. FASTA can be very useful to identify long regions of low similarity between highly diverged sequences. Further, FASTA is helpful when the query sequence is short, since BLAST usually fails to report results for short query sequences.

One-stop query system to access publicly available databases on alternative splicing. Several databases on alternative splicing are publicly available. To facilitate extraction of all known information about splicing of a gene, we generated a single interface that queries various databases simultaneously. Presently, seven alternative splicing databases namely, ASD, ASG, PALS, SpliceInfo, MAASE and HASDB (23–27) are made available from this interface. The interface accepts typical search terms (such as keywords, gene names and cross-references) and queries all these databases. The results

that are obtained from the individual database servers are presented with hyperlinks to the individual databases.

Example of data search and of data content

Figures 2 and 3 illustrate the results of searching the database for all entries of tra2-alpha and tra2-beta, two important splicing regulators. Querying the ASD using wrapper query interface for 'tra2a* | tra2b*' as keyword produces an output page (Figure 2) that lists entries from the different component data sets. As can be seen from the figure, related data entries across the different data sets are hyperlinked to one another. Figure 3a and b illustrate the presentation of some of the data items from AltSplice for tra2-beta; Figure 3a shows sections on gene information, on evidences for alternative splicing of tra2-beta, and on the observed splicing events in AltSplice. Figure 3b shows splicing patterns presented in textual form (as Splice Pattern Table) and in graphical form (as Splice Pattern view). Individual data items in the display page are hyperlinked to pages that list detailed information. For example, the splicing pattern entry number in the table is linked to a page that lets the user to perform multiple alignments on the sequences of all the observed isoform splicing patterns or peptides. Figure 3c shows the display page of tra2-beta entry from AEdb-Function data set; it illustrates the wealth of information that is captured from published literature.

ASD WORKBENCH

The workbench provides a set of online tools that enable users to carry out analysis of pre-mRNA sequences. It includes tools for intron analysis, scoring ATG-context sequence, finding exons and identifying splicing regulatory sequences. These tools are accessed either through a single wrapper interface or through interfaces that are specialised for individual tools.

Intron analysis

The tool examines intron sequences (as provided by the user) for putative branch point (BP) sites and polypyrimidine tracts (PPT). It further calculates the strength of the donor and acceptor sites. The methods are described elsewhere (10). The user has a choice of weight matrices for donor and acceptor sites tailored for different intron types, such as U2-type GT-AG and GC-AG and U12-type GT-AG and AT-AC.

Scoring ATG-context sequence

This tool examines each occurrence of ATG in a given transcript sequence for its ability to act as translation start codon. Each ATG is scored for Kozak's ATG-context sequence (28) using a weight matrix that we built from experimentally confirmed translation initiation sites. The sequence of translated peptide from each occurrence of ATG is presented along with the ATG-context score. FASTA/BLAST searches against UniProt sequence data can be launched for each of the translated peptide sequence.

MZEF-SPC exon finder

This tool identifies potential exons in a given nucleotide sequence. It integrates Michael Zhang's Exon Finder (29) and Thanaraj's SpliceProximalCheck (30). MZEF identifies putative exons using quadratic discriminant analysis. SPC is a decision tree implementation of splicing signals that differentiate genuine human splicing sites from the proximal false sites and thus specialises in validating the predicted exon boundary for exactness.

Detection of short regulatory sequences

Exons are regulated by short, degenerative sequences that bind to interacting splicing factors and proteins. These sequences are collected in the AEdb-Motif database. The Regulatory Sequence tool uses these motifs to examine a given nucleotide sequence for their presence. Users have a choice to specify the extent of allowed mismatches. The identified motifs are hyperlinked to the corresponding entries in AEdb-Motif database (See Figure 4 for illustration of identified motifs in tra2-beta gene). The splicing rainbow is a visualizations tool that colour-codes presence of different regulatory motifs in a user-supplied sequence.

SUMMARY OF UPDATES AND ENHANCEMENTS

The current release of ASD includes a large number of improvements over that reported earlier (9). AltSplice (the production pipeline) supersedes AltExtron (the research and development pipeline) in functionalities, data content, integrations and data presentations. As a result, AltExtron has become redundant and is not maintained any further; however for archival purposes, the earlier versions of AltExtron data are still presented in ASD web pages. Examples of enhancements in AltSplice data content include data on evidences for alternative splicing and isoform peptide sequences; those in integrations include related data from UniProt and Ensembl; those in query tools include differential library expression profiler; and those in presentation of data include a complete redesign of both the query and data (textual and graphical) results pages and multiple alignment view of isoform splice pattern and peptide sequences. In addition to the AEdb-Sequence data set presented in (9), further AEdb data sets (namely AEdb-Function, AEdb-Motif and AEdb-Minigenes) are presented. The ASD workbench is a new addition that complements the data content. The ASD sequence data are now available for search through BLAST and FASA tools. The current version of ASD provides three levels of search facilities, namely single-query bar, wrapper and data set specific advanced search query page. Further, a one-stop query system that enables users to query many publicly-available databases (including all the data sets of ASD) on alternative splicing is now provided. Integrations with splice variant data from UniProt and Ensembl enhance the value of ASD resources; AltSplice data are presented on Ensembl genome annotation browsers as DAS (Distributed Annotation Server) tracks. Since the last report, the ASD pipeline has matured to high production standards. The pipeline and the database are now handled by EBI database-production team which is committed to making regular data releases, to expanding the repertoire of organisms for which the data are made available, and to make seamless integration and hyperlinks both among the different data sets of ASD and to various other external databases.

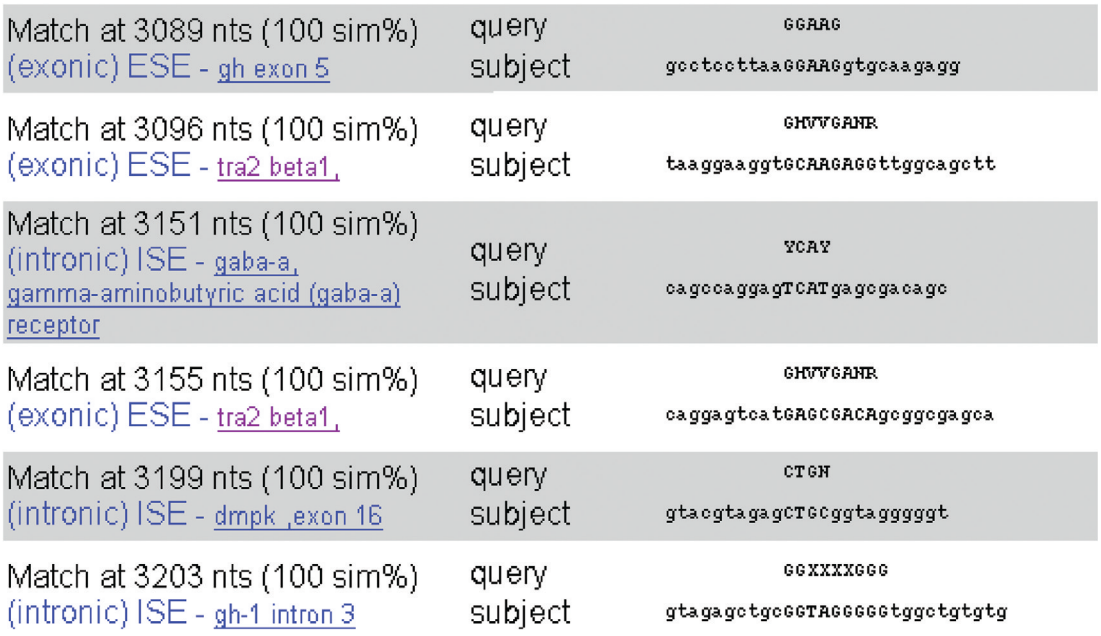


Figure 4. Example output from workbench tool that detects splicing regulatory sequences. The figure displays a portion of the page that reports splicing regulatory sequences in tra2-beta gene. The names of identified motifs are hyperlinked to appropriate entries in AEdb-Motif database. Matches against tra2-beta regulatory sequences are also seen. It is known that tra2-beta1 auto regulates its protein concentration by influencing alternative splicing of its pre-mRNA (31).

CONCLUSIONS

We present here ASD, a bioinformatics resource for alternative splicing. The individual components of the resource are (i) AltSplice—a value-added data set generated by our AltSplice, (ii) AEdb—a manually curated data set of alternatively spliced exons and their properties, and (iii) ASD Workbench—a collection of tools that carry our various analyses on pre-mRNA sequences. These individual components are integrated to one another and to related data from UniProt and Ensembl. The resource also provides a one-stop query system that accesses various other publicly available databases on alternative splicing. The integrated resource is available to the community (from http://www.ebi.ac.uk/asd) through user-friendly interfaces. The future releases will contain data that are being generated by other members of the splicing community, e.g. array expression data on alternative splicing of genes.

General enquiries on the ASD database can be mailed at asd-ebi@ebi.ac.uk.

ACKNOWLEDGEMENTS

T.A.T. and S.S. thank European Commission for the ASD grant (QLRT-CT-2001-02062). Involvement of Francis Clark and Juha Muilu in the earlier stages of the project is acknowledged. NLB-M acknowledges Portugal Foundation for Science and Technology for financial support (Fellowship SFRH/BD/2914/2000). Juan Valcarcel is acknowledged for mentoring NLB-M in developing the Splicing Rainbow tool. Funding to pay the Open Access publication charges for this article was provided by the ASD grant from European Commission.

Conflict of interest statement. None declared.

REFERENCES

1. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
2. Stamm,S., Ben-Ari,S., Rafalska,I., Tang,Y., Zhang,Z., Toiber,D., Thanaraj,T.A. and Soreq,H. (2005) Function of alternative splicing. *Gene*, **344**, 1–20.
3. Lewis,B.P., Green,R.E. and Brenner,S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.
4. Smith,C.W. and Valcarcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci*., **25**, 381–388.
5. Caceres,J.F. and Kornblihtt,A.R. (2002) Alternative splicing regulation: multiple control mechanisms and involvement in human diseases. *Trends Genet*., **18**, 186–193.
6. Stoilov,P., Meshorer,E., Gencheva,M., Glick,D., Soreq,H. and Stamm,S. (2002) Defects in pre-mRNA processing as causes and predisposition to diseases. *DNA Cell Biol*., **21**, 803–818.
7. Faustino,N.A. and Cooper,T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev*., **17**, 419–437.
8. Thanaraj,T.A. and Stamm,S. (2003) Prediction and statistical analysis of alternatively spliced exons. *Prog. Mol. Subcell. Biol*., **31**, 1–31.
9. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res*., **32**, D64–D69.
10. Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet*., **11**, 1–14, 451–464.
11. Ladd,A.N. and Cooper,T.A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol*., **3**, 8.1–8.16.
12. Zheng,Z.M. (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J. Biomed. Sci*., **11**, 278–294.
13. Bourgeois,C.F., Lejeune,F. and Stevenin,J. (2004) Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. *Prog. Nucleic Acid Res. Mol. Biol*., **78**, 37–88.
14. Tang,Y., Novoyatleva,T., Benderska,N., Kishore,S., Thanaraj,T.A. and Stamm,S. (2005) Analysis of alternative splicing *in vivo* using minigenes.

In Westhof,E., Bindereif,A., Schön,A. and Hartmann,K. (eds), *Handbook of RNA Biochemistry*. Wiley-VCH, Verlag, Weinheim., Vol. 2, pp. 755–782.
15. Stoss,O., Stoilov,P., Hartmann,A.M., Nayler,O. and Stamm,S. (1999) The *in-vivo* minigene approach to analyse tissue-specific splicing. *Brain Res. Protoc.*, **4**, 383–394.
16. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
17. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The universal protein resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
18. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **33**, D29–D33.
19. Wain,H.M., Lush,M., Ducluzeau,F. and Povey,S. (2002) Genew: the human gene nomenclature database. *Nucleic Acids Res.*, **30**, 169–171.
20. Ashburner,M., Ball,C.A., Blake,J.A., Butler,H., Cherry,J.M., Corradi,J., Dolinski,K., Janan,T., Eppig,J.T., Harris,M. *et al.* (2001) Creating the Gene Ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
21. Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien-Kruger,S., Smedley,D., Otgaar,D., Greyling,G., Jongeneel,V., McCarthy,M.I. *et al.* (2003) eVOC: a controlled vocabulary for gene expression data. *Genome Res.*, **13**, 1222–1230.
22. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
23. Leipzig,J., Pevzner,P. and Heber,S. (2004) The alternative splicing gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res.*, **32**, 3977–3983.
24. Huang,Y.H., Chen,Y.T., Lai,J.J., Yang,S.T. and Yang,U.C. (2002) PALS db: putative alternative splicing database. *Nucleic Acids Res.*, **30**, 186–190.
25. Huang,H.D., Horng,J.T., Lin,F.M., Chang,Y.C. and Huang,C.C. (2005) SpliceInfo: an information repository for mRNA alternative splicing in human genome. *Nucleic Acids Res.*, **33**, D80–85.
26. Zheng,C.L., Nair,T.M., Gribskov,M., Kwon,Y.S., Li,H.R. and Fu,X.D. (2004) A database designed to computationally aid an experimental approach to alternative splicing. *Pac. Symp. Biocomput.*, 78–88.
27. Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.
28. Kozak,M. (1984) Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.*, **12**, 857–872.
29. Zhang,M.Q. (2003) Using MZEF to find internal coding exons. In Baxevanis,A.D. and Davison,D.B. (eds), *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., New York, Vol. 1, pp. 4.2.1–18.
30. Thanaraj,T.A. and Robinson,A. (2000) Prediction of exact boundaries of exons. *Brief Bioinform.*, **1**, 343–356.
31. Stoilov,P., Daoud,R., Nayler,O. and Stamm,S. (2004) Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. *Hum. Mol. Genet.*, **13**, 509–524.