Databases and ontologies

ASPicDB: A database resource for alternative splicing analysis

T. Castrignanò1, M. D’Antonio1, A. Anselmo2, D. Carrabino1, A. D’Onorio De Meo1, A. M. D’Erchia3, F. Licciulli4, M. Mangiulli3, F. Mignone2, G. Pavesi2, E. Picardi3, A. Riva3,5, R. Rizzi6, P. Bonizzoni6 and G. Pesole3,4,*

1Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca, CASPUR, Rome, 2University of Milan, Dipartimento di Scienze Biomolecolari e Biotecnologie, via Celoria 26, Milan 20133, 3University of Bari, Dipartimento di Biochimica e Biologia Molecolare, via Orabona, 4, Bari 70126, 4Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, via Amendola 122/D, Bari 70126, Italy, 5Department of Molecular Genetics and Microbiology, University of Florida, PO Box 103610, Gainesville, FL 32610-3610, USA and 6DISCo, University of Milan Bicocca, via Bicocca degli Arcimboldi, 8, Milan, 20135, Italy

Received on November 30, 2007; revised on February 26, 2008; accepted on March 28, 2008

Advance Access publication April 3, 2008
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Alternative splicing has recently emerged as a key mechanism responsible for the expansion of transcriptome and proteome complexity in human and other organisms. Although several online resources devoted to alternative splicing analysis are available they may suffer from limitations related both to the computational methodologies adopted and to the extent of the annotations they provide that prevent the full exploitation of the available data. Furthermore, current resources provide limited query and download facilities.

Results: ASPicDB is a database designed to provide access to reliable annotations of the alternative splicing pattern of human genes and to the functional annotation of predicted splicing isoforms. Splice-site detection and full-length transcript modeling have been carried out by a genome-wide application of the ASPic algorithm, based on the multiple alignments of gene-related transcripts (typically a Unigene cluster) to the genomic sequence, a strategy that greatly improves prediction accuracy compared to methods based on independent and progressive alignments. Enhanced query and download facilities for annotations and sequences allow users to select and extract specific sets of data related to genes, transcripts and introns fulfilling a combination of user-defined criteria. Several tabular and graphical views of the results are presented, providing a comprehensive assessment of the functional implication of alternative splicing in the gene set under investigation.

ASPicDB, which is regularly updated on a monthly basis, also includes information on tissue-specific splicing patterns of normal and cancer cells, based on available EST sequences and their library source annotation.

Availability: www.caspur.it/ASPicDB

Contact: graziano.pesole@biologia.uniba.it

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Alternative splicing and alternative initiation/termination of transcription have recently emerged as the major mechanisms responsible for the expansion of the transcriptome and proteome complexity in human and other organisms (Brett et al., 2002; Kopelman et al., 2005; Talavera et al., 2007). Although the functional potential of splicing variants has not been widely studied so far, several examples are known where they are involved in the creation of novel or specialized functions (Black, 2003; Blencowe, 2006; Lopez, 1998). In fact, recent experimental studies aimed at the characterization of human and mouse transcriptomes have revealed a remarkable heterogeneity in transcription initiation, and have also shown that alternative splicing is a widespread phenomenon affecting more than 60% of human genes (Matlin et al., 2005; Tress et al., 2007), an estimate constantly increasing from the 35% of the first genome-wide assessment (Mironov et al., 1999).

Splicing regulation, resulting from a combination of cis-elements and trans-acting factors, is thus a key mechanism to tune gene expression to a variety of conditions, while its dysfunction may often be at the basis of the onset of genetic diseases and cancer (Garcia-Blanco et al., 2004).

Most of the methods currently available for the investigation and prediction of gene splicing patterns are based on independent and progressive alignments of transcript data (mostly ESTs) to a genomic sequence. Since these methods present some limitations mostly due to the sequence errors frequently occurring in ESTs and to the repetitive structure of the genome sequence, we previously developed a novel methodology implemented in the ASPic algorithm and software that take into account these issues [see Methods and (Bonizzoni et al., 2005; Castrignano et al., 2006)].

Here we present ASPicDB (version 1.2, January 2008), a database resource for alternative splicing analysis that collects and provides access to the results of a genome-wide analysis of human genes carried out by the ASPic software. ASPicDB is regularly updated on a monthly basis with the latest releases of NCBI Entrez gene and Unigene databases.
A large variety of tools and databases is currently available for genome-wide investigation and prediction of alternative splicing in human and other organisms: ASD (Stamm et al., 2006), ECgene (Lee et al., 2007), ASAP2 (Kim et al., 2007), Hollywood (Holste et al., 2006), AceView (Thierry-Mieg and Thierry-Mieg, 2006). However, the data they present are often significantly discordant due to differences in the input data, in the algorithm adopted for splice site prediction and transcript assembly [see (Bonizzoni et al., 2006) for further discussion] as well as on the level of stringency adopted to predict a splice site (e.g. occurrence of canonical donor/acceptor, number of supporting ESTs, quality of the alignment at the splice boundaries). ASPicDB combines the high-quality predictions produced by the ASPic algorithm with a powerful user-interface that offers several unique features. Enhanced retrieval tools allow for the definition of composite queries for the selection of subsets of genes, transcripts or introns related to specific features, for example introns belonging to the U2 or U12 class, or found in a specific mRNA localization (e.g. 5'UTR), or supported by a minimum number of ESTs, and so on. Furthermore, the database also includes information on tissue specific splicing in normal and cancer cells based on available expressed sequence tags and their source library annotations. Finally, ASPicDB provides download facilities for sequence extraction (e.g. sequence regions surrounding splice site boundaries featuring specific conditions) that could be used to assist bioinformatics analyses aimed at the detection of splicing regulatory elements.

2 METHODS

2.1 Data resources

ASPicDB is a relational database populated with the results obtained by the genome-wide application of the ASPic program, whose input consists of the genomic sequence corresponding to a specific gene and the collection of related expressed sequences, typically the expressed sequences contained in the relevant Unigene cluster. Only human genes for which a RefSeq NM curated transcript and a Unigene cluster were available were included in the database.

Genomic and expressed sequences used for the prediction were dynamically extracted from an ad hoc developed MySQL genomic platform, that includes data from the latest version of the human genome assembly (NCBI36) and coordinates resulting from BLAT mapping of RefSeq and Unigene entries updated on January 2008.

The genomic coordinates of each input gene were determined as the leftmost and rightmost mapping position of the relevant Unigene entries that also included the RefSeq sequences. In the case that more than one Unigene cluster was associated with a specific gene we selected the one containing the relevant gene-related RefSeq entries.

In order to investigate the differential expression of genes and alternative splicing isoforms in normal versus cancer tissues we annotated Unigene ESTs based on the standard eVOC ontology (Kelso et al., 2003). The eVOC database was also used as the source of expression pattern information. To avoid biases or inconsistencies in the statistical analysis of the expression pattern we excluded from the analysis the ESTs from normalized libraries and considered only tissues represented by more than 40000 ESTs.

2.2 Computational method for alternative splicing prediction

ASPic implements an algorithm for splice site prediction that performs a multiple EST sequence comparison and alignment against the genomic sequence (Bonizzoni et al., 2005). The algorithm uses an optimization procedure to minimize the number of detected splice sites thus reducing the number of false splice predictions, a common problem with other methods as we previously showed in (Bonizzoni et al., 2006).

ASPic overcomes the limitations of programs based on independent pairwise alignment of ESTs against the genome that tend to predict artefactual splice sites (Bonizzoni et al., 2006) suggested by alternative individual EST alignments, mainly, because of the repetitive structure of genomic sequences and of sequencing errors frequently occurring in ESTs. Furthermore, a dynamic programming procedure is adopted for further refinement of the alignment at exon–intron boundaries, trying to reconcile, where possible, non-canonical splice sites to canonical ones (Bonizzoni et al., 2005), and taking into account the scoring matrices for donor, acceptor and branch sites of U2 and U12 introns derived from (Sheth et al., 2006). However, all 256 possible splice site pairs are acceptable and used in transcript assembly under the condition they are supported by at least two mRNA/ESTs and no mismatches are observed in a 15-bp long sequence upstream and downstream from the intron. The main result of the strategy of aligning split ESTs to the genome is a more reliable prediction of introns. Then, a suitable algorithm based on a directed acyclic graph (DAG) has been then designed to assemble spliced ESTs and predicted introns into a minimum set of non-mergeable transcripts which are also annotated with respect to the location of the coding sequence and to the presence of premature stop codon (Maquat, 2004), polyA signal and polyA tail (Zhang et al., 2005). The annotation of splicing variants is done using a RefSeq mRNA as reference (Pruitt et al., 2007).

3 RESULTS

3.1 Database content

Table 1 reports some statistics on the data contained in the current version of ASPicDB (version 1.2, January 2008) which currently contains splicing predictions for 18 442 human genes. We estimated that over 91% of multi-exon genes may generate alternative isoforms and that each gene—on average—may generate about 12 different transcripts and 11 different proteins, most of them translated in frame with the RefSeq annotated protein. The resulting 10% of ‘untranslated’ isoforms includes those transcripts for which a reliable open reading frame (ORF) could not be annotated automatically (see the online documentation for the criteria used for ORF annotation).

ASPicDB also contains information about cancer versus normal tissue specificity for 17 tissue types at both gene and

Table 1. Statistics for ASPicDB (version 1.2, January 2008)

<table>
<thead>
<tr>
<th>Statistics</th>
<th>ASPicDB version 1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes (Refseq transcripts)</td>
<td>18 442 (27 056)</td>
</tr>
<tr>
<td>% Alternatively spliced multi-exon genes</td>
<td>91.1</td>
</tr>
<tr>
<td>Predicted transcripts</td>
<td>229 123</td>
</tr>
<tr>
<td>Predicted proteins</td>
<td>207 850</td>
</tr>
<tr>
<td>% in frame Proteins</td>
<td>89.7</td>
</tr>
<tr>
<td>Independent splicing events</td>
<td>185 446</td>
</tr>
<tr>
<td>U2 introns*</td>
<td>292 740 (226 600)</td>
</tr>
<tr>
<td>U12 introns*</td>
<td>1793 (917)</td>
</tr>
<tr>
<td>Other introns*</td>
<td>54 047 (9478)</td>
</tr>
</tbody>
</table>

\*supported by ≥2 ESTs.
splice site level (see an example in Supplementary Fig. 2): breast, cardiovascular system (e.g. heart), central nervous system (e.g. brain), gastrointestinal tract (e.g. colon, stomach), dermal system (e.g. skin), endocrine system (e.g. parathyroid, thyroid, pancreatic islet), hematological system (e.g. blood, bone marrow), liver and biliary system, musculoskeletal system (e.g. muscle), pancreas, peripheral nervous system (e.g. eye), ovary, placenta, uterus, prostate, respiratory system (e.g. lung), stomach, testis, urinary system (e.g. kidney, bladder).

3.2 Database mining
The database can be accessed through simple or advanced query interfaces. The simple query form allows the user to obtain the ASPic output for one or more genes selected according to one of their HGNC (e.g. NSMCE1), Unigene (e.g. Hs.481720), RefSeq (e.g. NM_152713), Entrez (e.g. 3248) or MIM (e.g. 202300) IDs (upper panel), or according to a keyword term (e.g. survivin), or to their associated Gene Ontology (GO) IDs (e.g. 0015248) or textual terms (e.g. ‘calcium ion transport’) belonging to the ‘biological process’, ‘molecular function’ or ‘cellular component’ categories (lower panel).

The advanced query form allows the user to search for: (1) genes; (2) transcripts or (3) splice sites, fulfilling different criteria (e.g. type of splicing event, type of donor/acceptor splice site, etc.). Depending on this choice three separate query forms appear.

The ‘Gene’ retrieval form has been designed to select genes fulfilling specific criteria (Fig. 1A) or showing a specific expression pattern in normal/tumor cells belonging to a given tissue (Fig. 1B).

Thus, users can select genes that are: (1) exclusively expressed in tumor cells (TT); (2) significantly more expressed in tumor cells (T); (3) equally expressed in tumor and normal cells (E); (4) significantly more expressed in normal cells (N) and (5) exclusively expressed in normal cells (NN). The classification is based on a simple $\chi^2$ statistic obtained from the comparison of observed and expected number of ESTs in normal and tumor tissues, calculated taking into account the relevant library sizes (for more information see the online Help documentation).

A transcript search can be performed in a similar fashion, with the additional possibility of selecting transcripts potentially targeted to nonsense mediated decay (NMD) for the presence of a premature termination codon (PTC) defined according to (Maquat, 2004) or with a polyA tail and a polyA signal, defined according to (Zhang et al., 2005).

Finally, the splice site retrieval form allows users to retrieve splice sites (introns) fulfilling one or more criteria, like for example the type of donor/acceptor site, the intron class [U2, U12 or unclassifiable, according to Sheth et al., (2006)], the number of supporting ESTs (defining their reliability degree), and their specific overrepresentation in normal versus tumor cell of a specific tissue, based on the number and the source of supporting ESTs following (Wang et al., 2003).

3.3 Results output
After a ‘Gene’, ‘Transcript’ or ‘Splice site’ query has been completed a Result Table is shown (Supplementary Fig. 1) listing the genes (Supplementary Fig. 1A), transcripts (Supplementary Fig. 1B) or introns (Supplelmentary Fig. 1C) that satisfy the selected criteria. The table contains general information about the query results and provide hyperlinks to access the ASPic results and the expression pattern output at the gene (Supplementary Fig. 2A) or splice site (Supplementary Fig. 2B) level.

A more detailed description of the Aspic output can be found in the online documentation where some example screenshots are also shown.

Briefly, the output is organized in six sections:

1. **Gene Information** reports a summary of the genomic and transcript data used by ASPIC to generate the prediction, downloadable by the user, and links to the results of other popular prediction programs [e.g. ASAP2 (Kim et al., 2007), ASTD (Stamm et al., 2006), AceView (Thierry-Mieg and Thierry-Mieg, 2006)].

2. **Gene Structure View** provides a schematic graphical view of the gene structure including all predicted exons/introns, using different colors for RefSeq or novel exon/introns.

3. **Predicted Transcripts** shows a graphical representation of the assembled transcripts with predicted annotations of 5’UTR, CDS and 3’UTR, PTC and polyA sites.

Fig. 1. ASPic DB gene advanced query form. The form is split into two parts (A and B) to search for genes fulfilling one or more criteria (A) or showing a specific expression profile in normal/tumor cells of a given tissue (B).
(4) **Predicted Splice Sites** shows the multiple sequence alignment between the genomic sequence and the expressed sequences (i.e. mRNAs and ESTs) near the boundaries (splice sites) of all predicted introns. Also included are the donor and acceptor predictive scores, the intron type (U2 or U12), and information on the score and position of the branching site, if available.

(5) **Intron Table** lists all predicted introns and their relevant features, including their chromosomal location, length, class and donor/acceptor sequences.

(6) **Transcript Table** lists the details of all predicted alternative transcripts including their length, number of exons, and presence of a predicted coding sequence. The ‘variant type’ columns lists all the alternative splicing events using a RelSeq mRNA as the reference transcript.

All results can also be downloaded by the user in the ‘gene transfer format’ (GTF) (see the Gene Information panel). The user can also download specific sets of sequences in FASTA format for further analyses, e.g. genes, transcripts, proteins, 5'UTRs, coding sequences, 3'UTRs, introns as well as sequence regions surrounding splice site boundaries.

To show the use of ASPicDB on a simple example, we retrieved all splice sites with a non-canonical AG donor and a canonical AG acceptor, supported by at least four ESTs. This search, which is easily performed using the ‘Advanced Search’ on ‘splice sites’, resulted in a total of 26 splice sites from 25 genes. We then downloaded the 22-bp long sequences, ranging from different genes. This way, for example, users may query for all introns with non-canonical splice sites supported by more than 10 ESTs, and belonging to the U2 or U12 class, optionally restricting the output to those located on a specific chromosome or chromosome region or chromosome. Other specific features of ASPicDB are the possibility to query genes narrowing the search to the ones containing alternative splicing events specifically affecting the 5'UTR, CDS or 3'UTR, or presenting a specific type of splicing event (e.g. exon skip, alternative donor or acceptor, and so on), or producing a given number (or range) of alternative transcripts/proteins, as well as any combination of the above criteria.

The possibility to query genes or splice sites showing a specific expression pattern in normal versus tumor cells in one or more tissues is another key feature of the database. Supplementary Figure 2A shows the expression pattern observed for the tp53 gene in 11 different tissues. Notably, a statistically significant cancer-specific expression of this gene is observed in urinary system, prostate, kidney and brain. The normal versus tumor expression pattern is also generated at the splice site level (Supplementary Fig. 2B) as previously described. Indeed, if a gene is exclusively expressed in tumor cells in a given tissue type it is meaningless to mine for tumor-specific isoforms. Such isoforms should instead be characterized by the occurrence of tumor specific introns (or vice versa normal) in genes expressed in both the normal and tumor status. Supplementary Figure 2B shows that splicing of intron #35 (numbering according to the ASPidDB annotation) in the gene MTO1 is remarkably cancer-specific in endothrme system, breast and muscle. Consequently, all transcript isoforms carrying this specific splice site are likely to be tumor-specific as well.

Finally, the download facilities add another unique feature to ASPicDB. In particular, the possibility to extract specific sequence regions surrounding boundaries of splice sites fulfilling specific criteria (e.g. a given assortment of donor and acceptor sites or a specific expression pattern) may greatly aid customized sequence analyses aimed at the identification of splicing regulatory elements (see also the example in Fig. 2).
5 FUTURE DIRECTIONS

ASPicDB is an ongoing project and we plan to further develop it in the next releases. The annotation of predicted isoforms will be further enriched by including information on specific regulatory elements in alternative mRNA untranslated regions and the functional features of the predicted protein isoforms (e.g. occurrence of Pfam domains, signal peptides, transmembrane helices, etc.).

We also plan to extend the database to other organisms for which the genome sequence and a suitable amount of expressed sequences are available (e.g. mouse, rat, zebrafish, etc.) and to add facilities to perform comparative analysis of alternative splicing of homologous genes.

ACKNOWLEDGEMENTS

This work was supported by FIRB (MIUR), EU STREP project ‘TRANSPOSE’, AIRC, and Telethon.

Conflicts of Interest: none declared.

REFERENCES

Bonizzoni, P. et al. (2005) ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences. BMC Bioinformatics, 6, 244.

Hsu, F.R. et al. (2005) AVATAR: A database for genome-wide alternative splicing event detection using large scale ESTs and mRNAs. Bioinformation, 1, 16–18.