D80–D85 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 4 November 2010*
doi:10.1093/nar/gkq1073

ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing

Pier L. Martelli[1], Mattia D'Antonio[2], Paola Bonizzoni[3], Tiziana Castrignanò[2], Anna M. D'Erchia[4], Paolo D'Onorio De Meo[2], Piero Fariselli[1], Michele Finelli[5], Flavio Licciulli[6], Marina Mangiulli[4], Flavio Mignone[7], Giulio Pavesi[8], Ernesto Picardi[3], Raffaella Rizzi[3], Ivan Rossi[5], Alessio Valletti[4], Andrea Zauli[5], Federico Zambelli[8], Rita Casadio[1,*] and Graziano Pesole[4,9,*]

[1]Biocomputing Group, University of Bologna, Bologna 40126, [2]Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca (CASPUR), Rome 00185, [3]DISCo, University of Milan-Bicocca, Milan, 20135, [4]Dipartimento di Biochimica e Biologia Molecolare, University of Bari, Bari 70126, [5]BioDec srl, Casalecchio di Reno, Bologna 40033, [6]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Bari 70126, [7]Dipartimento di Chimica Strutturale e Stereochimica Inorganica, University of Milan, [8]Dipartimento di Scienze Biomolecolari e Biotecnologie, University of Milan, Milan 20133 and [9]Istituto Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari 70125, Italy

Received August 9, 2010; Revised and Accepted October 14, 2010

ABSTRACT

Alternative splicing is emerging as a major mechanism for the expansion of the transcriptome and proteome diversity, particularly in human and other vertebrates. However, the proportion of alternative transcripts and proteins actually endowed with functional activity is currently highly debated. We present here a new release of ASPicDB which now provides a unique annotation resource of human protein variants generated by alternative splicing. A total of 256 939 protein variants from 17 191 multi-exon genes have been extensively annotated through state of the art machine learning tools providing information of the protein type (globular and transmembrane), localization, presence of PFAM domains, signal peptides, GPI-anchor propeptides, transmembrane and coiled-coil segments. Furthermore, full-length variants can be now specifically selected based on the annotation of CAGE-tags and polyA signal and/or polyA sites, marking transcription initiation and termination sites, respectively. The retrieval can be carried out at gene, transcript, exon, protein or splice site level allowing the selection of data sets fulfilling one or more features settled by the user. The retrieval interface also enables the selection of protein variants showing specific differences in the annotated features. ASPicDB is available at http://www.caspur.it/ASPicDB/.

INTRODUCTION

Alternative splicing is a well characterized mechanism which, coupled with alternative initiation and termination of transcription (1), may expand the transcriptome and proteome complexity in human and other organisms by over one order of magnitude with respect to the number of annotated genes (2,3). In particular, it is now widely demonstrated that virtually all multi-exon genes may generate multiple transcripts and protein variants (3,4) and that the splicing process is tightly regulated in different physiological conditions, tissues or developmental stages (5). Furthermore, alterations of the splicing process can be observed in several genetic diseases and in cancer (6–10).

The huge amount of EST sequences (11) together with the relevant reference genome sequence has been used to carry out an extensive analysis of alternative splicing in

*To whom correspondence should be addressed. Tel: +39 080 5443588; Fax: +39 080 5443317; Email: graziano.pesole@biologia.uniba.it
Correspondence may also be addressed to Rita Casadio. Tel: +39 0512094005; Fax: +39 051242576; Email: casadio@alma.unibo.it

The authors wish it to be known that in their opinion, the first two authors are the joint First Authors.

human through the ASPIC algorithm (12–14). The alternative splicing pattern of human multi-exon genes, determined by ASPIC, has been collected in ASPicDB, a database resource which presents some unique features with respect to other similar databases (15). The ASPIC algorithm implements an optimization strategy that, performing a multiple alignment of all available transcript data (including full-length cDNA and EST sequences) to the relevant genome sequence, detects the set of introns that minimizes the number of splicing sites. It also generates through a directed-acyclic graph combinatorial procedure the minimal set of non-mergeable transcript isoforms compatible with the detected splicing events (14). The reliability of splicing isoforms detected by ASPIC has been recently established through a comparative assessment (16).

The advent of massive transcriptome sequence data generated by RNA-Seq (17) is steadily increasing the number of validated splicing sites and isoforms in human and other organisms thus suggesting that a fraction of alternative splicing events are the result of background noise in the splicing process (18) which generates non-functional isoforms expressed at low level. Therefore, extensive research efforts are required to distinguish functional species-specific variants from non-functional ones originated from neutral drift in the splicing process, as well as to asses the biological role of functional isoforms.

The annotation of the protein variants predicted with ASPIC is an essential step for exploring the functional and structural diversity of the proteins originating from the same gene by means of alternative splicing and therefore for unraveling the complex physiological effects of alternative splicing events (19). Indeed, currently available databases, such as ASD (20), ASAP II (21), ASTALAVISTA (22) and H-DBAS (23), mostly collect information on alternative transcripts at the mRNA level, without considering the effect of alternative splicing on the protein structure and function. The ProSAS (24) database contains structural information as derived from comparative modeling procedure, but due to the limitations of the modeling techniques, only $\sim$15% of the human transcripts are endowed with a reliable protein structure prediction.

ASPICdb aims at filling the gap of structural and functional annotation of protein splicing variants, by adopting a set of analysis and prediction tools that do not rely only on annotation transfer by sequence similarity. It provides a thorough computational annotation of predicted human protein variants including PFAM domains (25), N-terminal signal peptides, GPI-anchor propeptides, transmembrane domains, subcellular localization and other features, also reporting the relevant crosslinks to UniprotKB/Swissprot (26) and PDB databases (27). A comprehensive annotation of the domain architecture and other structural features could also be extremely useful to critically assess the reliability of the functional classification provided the GO System (25), which still neglects much of the relevant information for alternative splicing products.

In addition, in consideration of the fragmented nature of the available transcript data, the new version of ASPicDB include the annotation of CAGE tags (28) in order to identify truly transcription initiation sites and discriminate between full-length isoforms using alternative transcription initiations and $5'$-partial transcripts for which a full-length CDS and the encoded protein cannot be reliably predicted.

ANNOTATION PIPELINE OF HUMAN PROTEIN VARIANTS

The computational pipeline implemented for supplementing the ASPicDB protein sequences with functional and structural annotations is represented in Figure 1 and integrates several state-of-the-art tools for similarity search and for machine-learning based prediction of protein features starting from residue sequence.

For each one of the 256 939 protein variants coming from 17 191 human genes, a first layer of annotation consists in the retrieval of similar sequences from the two major repositories containing well-characterized proteins, namely: (i) the UniProtKB/SwissProt data base (26) (rel. 2010_07, June 2010), that contains 547 011 protein sequences with curated annotations, including 517 802 principal entries and 29 209 splicing variants (UniProt Consotium, 2010); (ii) the Protein Data Bank (rel. July 2010), that contains resolved three dimensional structures for 50 171 different protein sequences (29).

Similarity searches were performed with BLAST (30) setting the *E*-value threshold to 10^{-3}.

A second layer of annotation is obtained by mapping the structural and functional domains collected in the

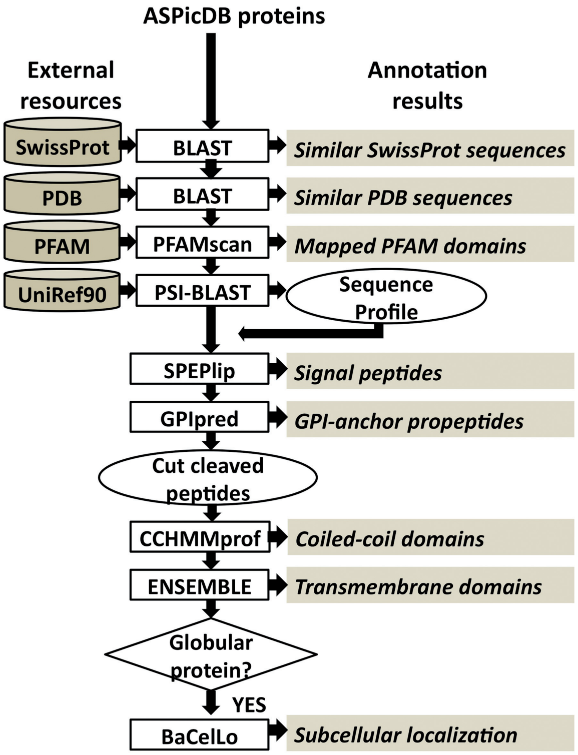

Figure 1. Pipeline for the annotation of alternative transcripts.

PFAM-A database (rel. 24.0, October 2009) that contains curated multiple sequence alignments based on hidden Markov models (HMM) for 8691 families, 2985 domains, 162 repeats and 74 motifs (25). The PFAM models were mapped on the ASPicDB protein sequences by means of the pfam_scan.pl program (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/), based on HMMER3.0 (31).

The third layer of annotation results from the integration of several predictors based on machine learning tools, such as neural networks, hidden Markov models, support vector machines and conditional random fields. Since most of the methods take advantage of the evolutionary information encoded in sequence profiles, we compiled them starting from the similar sequences retrieved with two PSI-BLAST iterations (setting the *E*-value threshold to 10^{-3}) from the UniRef90 data set consisting of 6 955 504 sequences (July 2010). The first predicted features are the presence of N-terminal signal peptide and of C-terminal GPI-anchor propeptides, with SPEPlip (32) and PredGPI (33), respectively. Both the methods are among the best available predictors, scoring with accuracy as high as 95% the former and 88% the latter. When present, the signal peptide and the propeptide are cleaved from the protein sequence. The presence of coiled-coil domains is predicted with CCHMM-PROF that is able to locate coiled-coil segments in protein sequences with 80% accuracy (34). α-Helical transmembrane domains are then predicted with ENSEMBLE (35), that discriminates transmembrane from globular proteins with false positive and false negative rates both equal to 3%. The same tool is adopted for predicting the number and the position of transmembrane segments along the sequence, with an accuracy of 90% on the protein base. The subcellular localization of globular proteins is predicted with BaCelLo (36), which discriminates four localizations in animals (secretory pathway, cytoplasm, nucleus and mitochondrion) with 74% accuracy.

ASPicDB CONTENT AND ANNOTATION OF PROTEIN VARIANTS

Table 1 reports some statistics on the data contained in the current version of ASPicDB (version 2.0, August 2010) which refers only to human multi-exon genes annotated in NCBI Entrez Gene (37) with at least one RefSeq transcript (38) and the relevant Unigene cluster (39) collecting all available gene-specific cDNA and EST sequences.

In the current version of ASPicDB some more features are available including the annotation of the CAGE tags (28) which define truly transcription initiation sites and a comprehensive protein annotation. A total of 12 789 394 CAGE tags have been mapped thus supporting constitutive or alternative transcription start sites. To each transcript variant a 'unique identifier' (16) has been associated in order to make possible the unambiguous comparison with alternative transcripts collected in other databases.

All alternative proteins collected in ASPicDB have been compared with UniprotKB/SwissProt (26) and PDB (29) databases. The results of similarity searches are reported in Table 2. Only 17% of the ASPicDB protein sequences are identical to proteins deposited in UniProtKB/SwissProt database. However, 94% of the sequences share significant similarity with proteins annotated in the same database, prompting the possibility of a reliable annotation transfer. Moreover, 54% of ASPicDB sequences are similar to proteins deposited in the PDB suggesting that their structures can be modeled, at least partially.

A considerable amount of PFAM models map on the ASPicDB sequences (Table 2). On the overall, 71% of sequences match with at least one model. This result is in agreement with the reported sequence coverage on the human proteome of the current PFAM release, which is equal to 72.5% (25). It is worth noticing that, although all the models map with an *E*-value < 10^{-5}, only 20% of the matches are complete (that is, involve the whole model). A note of caution is necessary when inferring features from partial matches and the actual extent of the match has to be evaluated for each instance.

Table 3 summarizes the results of the annotation process performed with machine learning based predictors. Two percent of proteins were not predicted since they are shorter than 50 residues, 16% of proteins are predicted as transmembrane and 82% are predicted as globular. Among the globular proteins, 12% are predicted as secreted, 35% as cytoplasmic, 27% as globular and 8% as mitochondrial. Signal peptides and GPI-anchor propeptides are predicted in the 12 and 0.7% of the

Table 1. Statistics of the ASPicDB content (v2.0, August 2010)

	ASPicDB v2.0
Genes	17 191
Transcripts	319 092
Proteins	256 939
Exons	390 886
Splicing sites	351 345
U2	302 164
U12	1712
Splicing events	233 717

The number of splicing sites belonging to the U2 or U12 class and of splicing events is also reported.

Table 2. Annotation of human variants upon similarity and PFAM searches

Sequence repository	No of proteins[a]	No of genes[a]
UniProtKB/SwissProt, %		
E-value < 10^{-3}, %	239 814 (93)	17 054 (99)
Identical, %	42 601 (17)	13 043 (76)
PDB		
E-value < 10^{-3}, %	137 528 (54)	11 062 (64)
Identical, %	1079 (0.4)	316 (2)
PFAM		
All matches, *E*-value < 10^{-5}, %	183 483 (71)	14 205 (83)
Complete matches, *E*-value < 10^{-5}, %	46 630 (18)	5621 (33)

[a]The percentages are computed with respect to 256 939 protein variants and 17 191 genes.

sequences, respectively. Coiled-coil domains are predicted in 1.3% of the proteins. At the gene level, 30 and 92% of genes encode for transmembrane and globular proteins, respectively. Since the sum exceed 100%, it follows that 22% of the genes encode for both globular and transmembrane variants. The same consideration holds for the other annotations as reported in Table 4. The amount of genes predicted to encode for proteins with different subcellular localization achieves 56%. This is partially explained by the fact that BaCelLo scores with an accuracy equal to 74%, which is the lowest among the methods included in the pipeline. Indeed the discrimination between the 'cytoplasmic' and the 'nuclear' classes is still a difficult task for all subcellular localization predictors (40). When the two classes are merged together, the BaCelLo accuracy increases up to 91%, but the rate of genes encoding for proteins with different localizations is still as high as 44%, suggesting that localization diversity is inherent in the ASPicDB protein variants. The structure of PFAM annotations is also highly variable: 38% of genes encode for variants matching with different number and/or type of PFAM models. Altogether, results listed in Table 4 suggest that alternative transcripts can encode for proteins endowed with different structural and functional features. ASPicDB provides a unique resource reporting the annotation of alternative splicing variants at the protein level and an interface enabling the discovery of such differences.

Table 3. Machine learning-based prediction of the human proteins deposited ASPicDB

Annotation	No. of proteins[a]	No. of genes[a]
Type		
Globular, %	210 608 (82)	15 513 (90)
Transmembrane, %	41 561 (16)	5439 (32)
Localization (globular proteins)		
Secretory pathway, %	31 917 (12)	7348 (43)
Cytoplasm, %	90 046 (35)	10 327 (60)
Nucleus, %	69 167 (27)	8183 (48)
Mitochondrion, %	19 478 (8)	4698 (27)
Domains		
Signal peptide, %	30 508 (12)	5153 (30)
GPI-anchor propeptide, %	1673 (0.7)	629 (4)
Coiled-coil segments, %	3423 (1.3)	497 (2.8)

[a] The percentages are computed with respect to 256 939 protein variants and 17 191 genes.

Table 4. Differences among alternative proteins encoded by the same human gene

Annotation	No. of genes[a], %
Type (globular/transmembrane)	3817 (22)
Subcellular localization (globular proteins)	9593 (56)
Presence of signal peptide	3939 (23)
Presence of GPI-anchor propeptide	591 (3.4)
Presence of coiled-coil domains	464 (2.7)
Number of transmembrane helices	2140 (12)
PFAM models (all matches)	6575 (38)

[a] The percentages are computed with respect to 17 191 genes.

ASPicDB RETRIEVAL INTERFACE

ASPicDB can be accessed though simple or advanced query forms. The simple query form allows the user to obtain the splicing pattern of one or more genes selected according to several criteria (e.g. HGNC name, RefSeq or Unigene accession IDs, etc.). The advanced query form allows the user to search for (i) genes, (ii) transcripts; (iii) exons; (iv) splicing sites; and (v) proteins, fulfilling different criteria (e.g. exons in a given length range, etc.). Depending on the choice separate query forms appear. The 'gene', 'transcript' and 'splicing sites' query forms have been described previously (15) whereas the 'exon' and 'protein' query forms are novel features of this version of ASPicDB. The exon query form allows the user to select exons in a given length range, belonging to a specific type (initial, internal or teminal), flanked by specific splicing sites or associated to one or more Affimetrix ExonArray probeset IDs.

The 'protein' query form allows the retrieval of transcripts encoding proteins isoforms of a specific class (e.g. globular or transmembrane), subcellular localization (e.g. mitochondrion, nucleus, secretory, cytoplasm) or containing one or more features, including occurrence and number of PFAM or transmembrane domains, GPI-anchor propeptides, signal peptides. Finally, it is also possible to retrieve genes encoding for alternative proteins that show differences in the above mentioned features.

ASPicDB OUTPUT

After a simple or advanced query has been submitted the output for each selected gene is shown which is organized in eight panels.

(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 other popular prediction programs such as ASAP2 (21), ASD (20) and ACEVIEW (41) as well as to ASPIC results for orthologous genes in other species.
(2) *Gene structure* view provides a schematic graphical view of the gene structure including all predicted exons/introns.
(3) *Predicted transcripts* show a graphical representation of the assembled transcripts with predicted annotations of 5′-UTR, CDS and 3′-UTR, CAGE tag mapping, Premature Termination Codons (PTC) and polyA sites.
(4) *Transcript table* lists the details of all predicted alternative transcripts including their length, number of exons and presence of a protein coding sequence. The 'variant type' column lists all the alternative splicing events using a RefSeq mRNA as the reference transcript. The transcript signature is also reported which consists in a unique ID for alternatively spliced variants generated according to (16).
(5) *Predicted proteins* show a graphical representation of the encoded proteins with matching domains (Figure 2). For each mapped domain the sequence

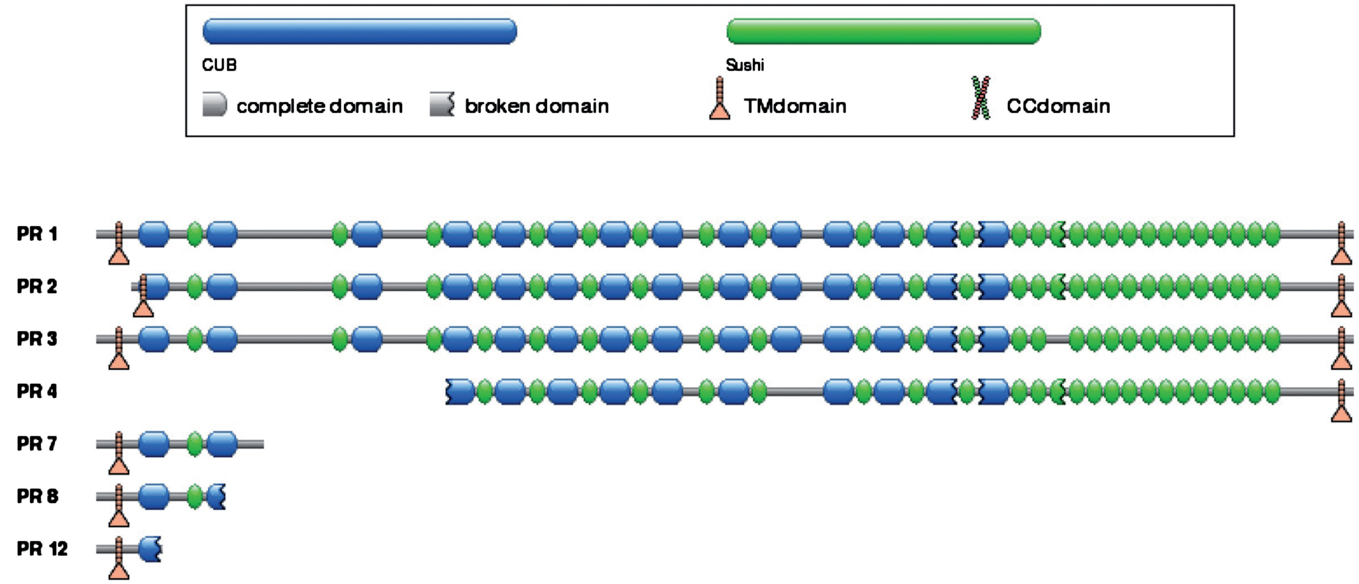

Figure 2. 'Predicted proteins' panel for gene CSMD3 (CUB and Sushi multiple domains 3). The gene is predicted to encode for 12 transcripts and 7 different protein sequences. Variants labeled as PR1, PR2 and PR3 are identical to the isoforms reported in the CSMD3_HUMAN entry of SwissProt/UniprotKB. Two more variants are reported in that file, although lacking of experimental annotations. Several repetitions of Sushi and CUB domains are predicted with PFAM (25) and represented with symbols indicating whether the model is completely or partially mapped to the sequence. The two transmembrane helices are predicted with ENSEMBLE (35).

coordinates are reported and different symbols indicate whether the mapping involves the complete domain or only a part of it.

(6) *Protein table* lists the predicted features of the alternative proteins that include: (i) the best hits obtained from the similarity searches against the UniProtKB/SwissProt and PDB databases, along with the identity value and coverage of the alignment with respect to both the query and the subject sequence lengths; (ii) the features predicted by the pipeline based on machine-learning tools.

(7) *Predicted splice sites* shows the multiple alignment between the genomic sequence and the expressed sequences (i.e. mRNAs and ESTs) near the boundaries (splice sites) of all predicted introns.

(8) *Intron table* lists all predicted introns and their relevant features; All results can be also downloaded by the user in textual format following the 'gene transfer format' (GTF) (see the Gene Information panel).

After a query at the gene, transcript, exon, protein or splice site level has been completed, the user can also download specific sets of sequences in FASTA format for further analyses, e.g. genes, transcripts, exons, proteins, 5′-UTRs, coding sequences, 3′-UTRs, introns as well as sequence regions surrounding splice site boundaries.

FUTURE PERSPECTIVES

ASPicDB is an ongoing project and we plan to further develop it in the next releases. In particular we plan to add specific annotations on splicing regulatory elements and their interacting RNA-binding proteins located both in exonic and intronic regions. We also plan to update alternative splicing prediction by using the huge amount of RNA-Seq data which are now being produced by next generation sequencing, possibly annotating splicing events as constitutive or tissue-specific. Furthermore, literature-screened splicing patterns related to diseases will be annotated as they represent potential molecular biomarkers and possible targets for therapy. Finally, the inclusion in the database of data related to other organisms will certainly favor a better understanding of the alternative splicing process through comparative analyses.

FUNDING

Ministero dell'Istruzione, dell'Università e della Ricerca: Fondo Italiano Ricerca di Base: 'Laboratorio Internazionale di Bioinformatica' (LIBI); Laboratorio di Bioinformatica per la Biodiversità Molecolare (MBLAB) and Telethon (project GGP01658). Funding for open access charge: Ministero dell'Università e della Ricerca: Fondo Italiano Ricerca di Base: 'Laboratorio Internazionale di Bioinformatica' (LIBI).

Conflict of interest statement. None declared.

REFERENCES

1. Frith,M.C., Valen,E., Krogh,A., Hayashizaki,Y., Carninci,P. and Sandelin,A. (2008) A code for transcription initiation in mammalian genomes. *Genome Res.*, **18**, 1–12.
2. Matlin,A.J., Clark,F. and Smith,C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, **6**, 386–398.

3. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
4. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
5. Barash,Y., Calarco,J.A., Gao,W., Pan,Q., Wang,X., Shai,O., Blencowe,B.J. and Frey,B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.
6. Faustino,N.A. and Cooper,T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev.*, **17**, 419–437.
7. Wang,G.S. and Cooper,T.A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, **8**, 749–761.
8. Pettigrew,C.A. and Brown,M.A. (2008) Pre-mRNA splicing aberrations and cancer. *Front. Biosci.*, **13**, 1090–1105.
9. Srebrow,A. and Kornblihtt,A.R. (2006) The connection between splicing and cancer. *J. Cell Sci.*, **119**, 2635–2641.
10. Venables,J.P. (2004) Aberrant and alternative splicing in cancer. *Cancer Res.*, **64**, 7647–7654.
11. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST–database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
12. Bonizzoni,P., Rizzi,R. and Pesole,G. (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.
13. Castrignano,T., Rizzi,R., Talamo,I.G., De Meo,P.D., Anselmo,A., Bonizzoni,P. and Pesole,G. (2006) ASPIC: a web resource for alternative splicing prediction and transcript isoforms characterization. *Nucleic Acids Res.*, **34**, W440–W443.
14. Bonizzoni,P., Mauri,G., Pesole,G., Picardi,E., Pirola,Y. and Rizzi,R. (2009) Detecting alternative gene structures from spliced ESTs: a computational approach. *J. Comput. Biol.*, **16**, 43–66.
15. Castrignano,T., D'Antonio,M., Anselmo,A., Carrabino,D., D'Onorio De Meo,A., D'Erchia,A.M., Licciulli,F., Mangiulli,M., Mignone,F., Pavesi,G. *et al.* (2008) ASPicDB: a database resource for alternative splicing analysis. *Bioinformatics*, **24**, 1300–1304.
16. Riva,A. and Pesole,G. (2009) A unique, consistent identifier for alternatively spliced transcript variants. *PLoS ONE*, **4**, e7631.
17. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
18. Melamud,E. and Moult,J. (2009) Stochastic noise in splicing machinery. *Nucleic Acids Res.*, **37**, 4873–4886.
19. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.J., Yeats,C., Olason,P.L., Albrecht,M., Hegyi,H., Giorgetti,A. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.
20. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46–D55.
21. Kim,N., Alekseyenko,A.V., Roy,M. and Lee,C. (2007) The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res.*, **35**, D93–D98.
22. Foissac,S. and Sammeth,M. (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.*, **35**, W297–W299.
23. Takeda,J., Suzuki,Y., Sakate,R., Sato,Y., Gojobori,T., Imanishi,T. and Sugano,S. (2010) H-DBAS: human-transcriptome database for alternative splicing: update 2010. *Nucleic Acids Res.*, **38**, D86–D90.
24. Birzele,F., Kuffner,R., Meier,F., Oefinger,F., Potthast,C. and Zimmer,R. (2008) ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.*, **36**, D63–D68.
25. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
26. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.*, **406**, 89–112.
27. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
28. Kodzius,R., Kojima,M., Nishiyori,H., Nakamura,M., Fukuda,S., Tagami,M., Sasaki,D., Imamura,K., Kai,C., Harbers,M. *et al.* (2006) CAGE: cap analysis of gene expression. *Nat. Methods*, **3**, 211–222.
29. Dutta,S., Burkhardt,K., Swaminathan,G.J., Kosada,T., Henrick,K., Nakamura,H. and Berman,H.M. (2008) Data deposition and annotation at the worldwide protein data bank. *Methods Mol. Biol.*, **426**, 81–101.
30. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
31. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
32. Fariselli,P., Finocchiaro,G. and Casadio,R. (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, **19**, 2498–2499.
33. Pierleoni,A., Martelli,P.L. and Casadio,R. (2008) PredGPI: a GPI-anchor predictor. *BMC Bioinformatics*, **9**, 392.
34. Bartoli,L., Fariselli,P., Krogh,A. and Casadio,R. (2009) CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information. *Bioinformatics*, **25**, 2757–2763.
35. Martelli,P.L., Fariselli,P. and Casadio,R. (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics*, **19(Suppl. 1)**, i205–i211.
36. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.
37. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–D16.
38. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
39. Zhuo,D., Zhao,W.D., Wright,F.A., Yang,H.Y., Wang,J.P., Sears,R., Baer,T., Kwon,D.H., Gordon,D., Gibbs,S. *et al.* (2001) Assembly, annotation, and integration of UNIGENE clusters into the human genome draft. *Genome Res.*, **11**, 904–918.
40. Casadio,R., Martelli,P.L. and Pierleoni,A. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief. Funct. Genomic Proteomic*, **7**, 63–73.
41. Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7(Suppl. 1)**, S12, 1–14.