W440–W443 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl324

ASPIC: a web resource for alternative splicing prediction and transcript isoforms characterization

Tiziana Castrignanò, Raffaella Rizzi[1], Ivano Giuseppe Talamo, Paolo D'Onorio De Meo, Anna Anselmo[2], Paola Bonizzoni[1] and Graziano Pesole[3,*]

Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Universita' e Ricerca, CASPUR, Rome, Italy, [1]DISCo, University of Milan Bicocca, via Bicocca degli Arcimboldi, 8, Milan, 20135, Italy, [2]Dipartimento di Scienze Biomolecolari e Biotecnologie, University of Milan, via Celoria 26, Milan 20133, Italy and [3]Dipartimento di Biochimica e Biologia Molecolare, University of Bari, via Orabona, 4, Bari 70126, Italy

Received February 18, 2006; Revised March 8, 2006; Accepted April 13, 2006

ABSTRACT

Alternative splicing (AS) is now emerging as a major mechanism contributing to the expansion of the transcriptome and proteome complexity of multicellular organisms. The fact that a single gene locus may give rise to multiple mRNAs and protein isoforms, showing both major and subtle structural variations, is an exceptionally versatile tool in the optimization of the coding capacity of the eukaryotic genome. The huge and continuously increasing number of genome and transcript sequences provides an essential information source for the computational detection of genes AS pattern. However, much of this information is not optimally or comprehensively used in gene annotation by current genome annotation pipelines. We present here a web resource implementing the ASPIC algorithm which we developed previously for the investigation of AS of user submitted genes, based on comparative analysis of available transcript and genome data from a variety of species. The ASPIC web resource provides graphical and tabular views of the splicing patterns of all full-length mRNA isoforms compatible with the detected splice sites of genes under investigation as well as relevant structural and functional annotation. The ASPIC web resource—available at http://www.caspur.it/ASPIC/—is dynamically interconnected with the Ensembl and Unigene databases and also implements an upload facility.

INTRODUCTION

Alternative splicing (AS) is increasingly emerging as a major mechanism in the expansion of transcript and protein complexity in eukaryotes. Indeed, recent experimental studies directed towards the characterization of human and mouse transcriptomes have revealed that AS is a widespread phenomenon affecting >60% (a constantly increasing estimate) of human genes (1). These discoveries imply that current microarray- or SAGE-based methods are not fully adequate for determining the cell specific expression profile of genes, since they do not take into account all of the possible alternative transcripts and thus can only provide a partial and incomplete estimate of the actual expression level of given gene isoform. Equally important is the capacity to annotate different possible transcripts generated by AS. In particular, transcripts generated by AS may differ both in the untranslated region (UTR) and in coding regions (CDS). It is also possible that certain transcribed isoforms may completely lack coding capacity but be involved in regulatory activities (2). A profound appreciation of the impact of AS at the level of protein structure and protein interactions also represents a challenge in the understanding of the functional impact of AS in cell metabolism. Recent descriptions of the functional implications of AS in tissue-specificity (3), different biological processes (4) and tumor development (5) has generated an explosion of interest and activity in this field particularly with respect to the development of suitable computational methods for AS prediction. Such methods are generally based on large-scale comparisons of transcript [mostly expressed sequence tags (ESTs)] and genomic sequences. A number of AS databases, such as ASD (6), ASAP (7) and ECgene (8), have been developed but these are often limited to a limited number of organisms, particularly human and mouse.

Computational methods for AS prediction can be subdivided into three groups: methods based on the comparison of expressed sequences to each other (9); methods based on the progressive alignment of expressed sequences to the genomic sequence (10); and methods that combine the previous two approaches thus avoiding their specific limitations (11,12).

*To whom correspondence should be addressed. Tel: +39 080 5443588; Fax: +39 080 5443317; Email: graziano.pesole@biologia.uniba.it

We have developed a new method for the prediction of splice sites and transcript isoforms. Our approach adopts an optimization procedure that considers multiple alignments of ESTs to the genomic sequence, minimizing the number of splice site predictions and of transcript isoforms. This method, implemented in the ASPIC software, has been shown to outperform other similar tools both in term of sensitivity and selectivity (11).

In this manuscript we present the ASPIC web resource that allows the user to determine the splicing pattern of a user submitted gene and the relevant transcript and protein products. The ASPIC methodology is applicable to a variety of species. Input data can be retrieved from the Ensembl and Unigene databases—to which the web resource is dynamically interconnected—or directly provided by the user.

METHODS

ASPIC adopts an optimization procedure that minimizes the set of splice sites compatible with the multiple alignments of all transcript data against the genomic sequence (11). This approach overcomes the limitations of methods that (erroneously) assume independence of single transcript-genome alignments.

The alignment between genome and the transcript sequences is carried out by a specifically designed aligner that produces a factorization of all expressed sequences into high-quality alignments to the genomic sequence (exons or factors) and then finds the solution that minimizes the corresponding factors in the genomic sequence. As the occurrence of repeated sequences in the genomic sequences may induce an over-factorization a backtracking procedure is carried out for concatenating wrongly split exons.

A maximum parsimony criterion is also used for the final assessment of intron–exon boundaries whereby computed EST factors (candidate exons) are merged whenever they differ at only a few positions—likely because of sequencing errors.

Furthermore, ASPIC implements specific algorithmic strategies to improve the quality of splice locations. More precisely, it applies an algorithm based on dynamic programming (DP), producing for regions close to splice sites, an alignment between the ESTs and the genomic sequence with a large gap of cost zero (the intron) and the minimum number of mismatches and insertions/deletions. Alternative alignments of the same quality (identity %) are differentially weighted by the DP procedure according to a scoring system using position frequency matrices of donor and acceptor splice sites (13).

Following the determination of EST factors and their corresponding genomic factors, the *TransView* module of ASPIC generates the minimum set of non-mergeable transcripts supported by experimental evidence (provided by the previously determined genome-transcripts alignments). Briefly, the algorithm builds an assembly graph of EST factorizations constructed by representing partial order relationships among spliced ESTs. More precisely, nodes of the graphs are spliced ESTs which are connected by edges if they overlap, i.e. they share at least one splice site. Given transcripts t_1 consisting of genomic exons $a_1, a_2, \ldots, a_n$ and $t_2 = b_1, b_2, \ldots, b_m$, then t_2 overlaps t_1, iff b_1 is a suffix of a_j for some $1 \leq j \leq n$, b_k is equal to a_{j+k-1} for $k = 2 \ldots n - j$, a_n is a prefix of b_{n-j+1}, $n - j + 1 \leq m$ and a_n is not a terminal exon [i.e. does not present a poly(A) tail].

The *Transview* algorithm then generates the alternative full-length transcripts by exploring all distinct plausible and non-redundant paths of the assembly graph. To increase the accuracy of transcript assembly, only spliced ESTs with high-quality splice sites (i.e. supported by more than two ESTs or showing perfect identity with the genomic sequence and canonical splices) are included in the graph.

cDNA/EST sequences with poly(A) tails are used to infer poly(A) cleavage sites (CS) in the assembled transcripts. The computation pipeline adopted in ASPIC is similar to that reported in Ref. (14) requiring at least eight or more consecutive As after the predicted CS at the 3′ end of the transcript. To exclude internal priming artifacts the genomic sequence from −10 to +10 with respect to the predicted CS should not contain more than six continuous As or more than seven As in a 10 nt window. Poly(A) CSs located within a 24 nt window are considered to be generated from heterogeneous cleavage of mRNA and clustered together. To further support the occurrence of a poly(A) site the occurrence of a canonical polyadenylation site (AAUAAA or AUUAAA) is sought from 40 nt upstream of the CS.

THE ASPIC WEB RESOURCE

Input

The ASPIC input consists of the genomic sequence corresponding to a specific gene (and possibly some flanking sequences) and a collection of related transcribed sequences. The web interface allows the user to paste or upload both the genomic and transcript sequences, or more conveniently may automatically get all the relevant sequences (once the organism and the gene under investigation have been defined) from Ensembl and Unigene databases. Genomic sequences can be automatically retrieved by providing chromosomal ranges, or typing the Ensembl gene ID (Hugo ID is also allowed for human genes), If a gene ID is provided the transcribed sequences collected in the corresponding Unigene cluster are automatically extracted. In addition to the Unigene sequences, the user can input additional transcribed sequences by the paste and/or upload facility.

Output

ASPIC outputs provide both graphical and tabular views of the splicing patterns of the gene under investigation. The nucleotide sequences of the inferred full-length isoforms as well as their structural and functional annotation are also produced. Two graphical views are generated: (i) the gene view and (ii) the transcript view. The gene view (Figure 1a) shows a full snapshot of the splicing pattern of a gene showing all detected exons and introns. The transcript view (Figure 1b) shows the overall exon–intron scheme of the assembled full-length transcripts, where the 5′-UTR, CDS, 3′-UTR and poly(A) site are annotated. Two tabular views are generated: (i) the intron table and (ii) the transcript table. The intron table (Figure 2) reports the relative and

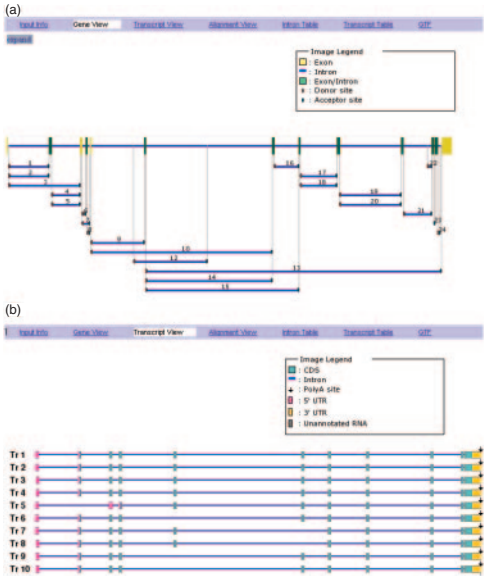

Figure 1. (**a**) Snapshot of the gene view for the human HNRPR gene showing the gene structure and detected introns numbered progressively. Constitutive and alternative exons are shown in yellow and green respectively. (**b**) Sample transcript view showing the inferred structure of assembled alternative transcripts starting from the reference transcript and reporting the annotation of the 5′-UTR, CDS, 3′-UTR and poly(A) tail.

Intron	Relative Start	Relative End	Absolute Start	Absolute End	Length	#ESTs	Donor & Acceptor	Mismatch (15 bp upstream donor)	Mismatch (15 bp downstream acceptor)
1	101	3290	1:23412820	1:23416009	3190	68	GT-AG	0.00	0.00
2	101	3293	1:23412817	1:23416009	3193	82	GT-AG	0.16	0.41
3	101	5702	1:23410408	1:23416009	5602	304	GT-AG	0.15	0.39
4	3460	5702	1:23410408	1:23412650	2243	161	GT-AG	0.17	0.54
5	3464	5715	1:23410395	1:23412646	2252	1	AA-TA	13.33	20.00
6	5822	6157	1:23409953	1:23410288	336	1	GT-AG	0.00	0.00
7	5822	6449	1:23409661	1:23410288	628	484	GT-AG	0.14	0.18
8	6249	6449	1:23409661	1:23409861	201	2	GT-AG	0.00	0.00
9	6558	10679	1:23405431	1:23409552	4122	479	GT-AG	0.38	0.40
10	6558	20578	1:23395532	1:23409552	14021	17	GT-AG	0.39	0.78

Figure 2. Sample intron table for the human HNRPR gene showing the relative and absolute coordinates of each detected intron, their lengths, the number of supporting ESTs, the donor and acceptor sites and the alignment quality (overall mismatch percentage) near to intron boundaries.

Transcript	Exons	L (nt)	CDS	CCDS start/stop	Prot L (aa)	Variant_type
TR1	11	2681	110..2011		633	Reference TR
TR2	11	2690	110..2020	yes/yes	636	A3E (I7, -9 nt), CDS;
TR3	11	2683	112..2013	yes/yes	633	A3E (I1, -3 nt), 5utr;
TR4	11	2692	112..2022	yes/yes	636	A3E (I7, -9 nt), CDS; A3E (I1, -3 nt), 5utr;
TR5	10	2524	247..1854	no/yes	535	A3E (I7, -9 nt), CDS; skip(E2), uCDS;
TR6	10	2567	110..1897	yes/yes	595	skip(E5), CDS;
TR7	10	2504	110..1834	yes/yes	574	skip(E6), CDS;
TR8	10	2506	112..1836	yes/yes	574	A3E (I1, -3 nt), 5utr; skip(E6), CDS;
TR9	10	2569	112..1899	yes/yes	595	A3E (I1, -3 nt), 5utr; skip(E5), CDS;
TR10	10	2578	112..1908	yes/yes	598	A3E (I7, -9 nt), CDS; A3E (I1, -3 nt), 5utr; skip(E5), CDS;

Figure 3. Transcript table for the first 10 alternative transcripts of the human HNRPR gene, showing: the transcript ID, number of exons, length, CDS annotation, occurrence of the CCDS start/stop, inferred protein length and variant type. The variant type provides information on the type of splicing events and their gene (E, exon; I, intron) and mRNA locations (5utr, CDS or 3utr).

absolute coordinates of each detected intron as well as the number of supporting ESTs and the alignment quality near to the intron boundaries. The transcript table (Figure 3) shows the general structural features of all alternative full-length transcripts—such as the length, the number of exons, the putative location of the CDS, the length of the putative encoded protein and the transcript variant type. The variant type column reports—for each full-length transcript—the type of splicing event (e.g. alternative 5′ or 3′ end, exon skipping), the affected exon or intron as well as its location in the coding and/or UTRs of the transcript. Splicing variants are labeled in comparison with a reference transcript given by the longest inferred transcript containing a CDS corresponding to the one annotated in the CCDS database (http://www.ncbi.nlm.nih.gov/CCDS/) for human or by the longest transcript with the longest open reading frame (ORF) in other species. The CDS in alternative full-length transcripts not containing the CCDS start and stop codons is determined as the longest ORF if longer than 100 codons. Finally, a full textual output of ASPIC analysis can be downloaded whereby exon, intron and transcript coordinates and sequences can be downloaded in GTF format.

System

PHP scripts (http://www.php.net/) have been developed for job submission; they launch a java background program (that manages the whole run) and eventually plot dynamic web results. The architecture of the software system can be divided into two main parts: a pre-processing phase where the java program queries two web services (WS) built on Simple Object Access Protocol open standard and a core processing phase where C programs run to detect splicing sites. The two WSs are called depending on input parameter selection:

(i) GeneInfo (http://t.caspur.it:8080/axis/webservices/GeneInfo.jws?wsdl) accepts a gene identifier, gives information on all the other available identifiers (Hugo name, Alias, Ensembl, Unigene) and downloads the ESTs cluster;
(ii) EnsjWS (http://t.caspur.it:8080/axis/webservices/EnsjWS.jws?wsdl) uses the ensembl java API to download the genomic sequence; it accepts an organism and a gene identifier or a chromosomal range, and returns the genomic sequence.

A PHP part of the web interface—related to the pre-processing phase—has been developed with a service-oriented approach to enable users to download a huge amount of genomic sequence and related transcripts.

The Aspic web tool has been implemented on a 4-processor server (HP DL585). The web tool accessible at http://www.caspur.it/ASPIC is implemented on a Linux server (SUSE SLESS 9) running the apache web server version 2.0 (www.apache.org).

CONCLUSIONS AND PERSPECTIVES

AS has been shown to be a key mechanism for the optimization of the information encoded by a single gene. A single gene may generate a large number of different transcripts and proteins through a combinatorial assortment of alternative exons and introns. Thus, many transcripts differing in 5′- and 3′-UTRs and in the coding region may be generated from a single gene. Such transcripts may be subjected to different posttranscriptional regulatory pathways and may encode several proteins with different functional and structural features—such as stability, intracellular localization or binding properties (4). Consequently, AS offers an exceptionally versatile way to fine tune gene expression according to the specific cell type and physiological status.

The ASPIC web resource provides biologists with a powerful tool for detecting the transcriptional profile of a specific gene using all the currently available genome and transcript data from a variety of species. Results thus obtained may contribute many new functional insights into known and novel genes and may suitably direct or focus further experimental studies.

ACKNOWLEDGEMENTS

We thank Gabriele Ravanelli for some technical support and David Horner for valuable comments on the manuscript. This work was supported by Fondo Italiano Ricerca di Base (FIRB) projects 'Bioinformatica per la Genomica e la Proteomica' and 'Laboratorio Italiano di Bioinformatica', EU STREP project TRANSCODE and by AIRC. Funding to pay the Open Access publication charges for this article was provided by FIRB project 'Laboratorio Italiano di Bioinformatica', Italy.

Conflict of interest statement. None declared.

REFERENCES

1. Matlin,A.J., Clark,F. and Smith,C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nature Rev. Mol. Cell Biol.*, **6**, 386–398.
2. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature*, **420**, 563–573.
3. Yeo,G., Holste,D., Kreiman,G. and Burge,C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol.*, **5**, R74.
4. 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.
5. Venables,J.P. (2004) Aberrant and alternative splicing in cancer. *Cancer Res.*, **64**, 7647–7654.
6. 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.
7. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res.*, **31**, 101–105.
8. Kim,P., Kim,N., Lee,Y., Kim,B., Shin,Y. and Lee,S. (2005) ECgene: genome annotation for alternative splicing. *Nucleic Acids Res.*, **33**, D75–D79.
9. Brett,D., Hanke,J., Lehmann,G., Haase,S., Delbruck,S., Krueger,S., Reich,J. and Bork,P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, 83–86.
10. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
11. 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.
12. Grasso,C., Modrek,B., Xing,Y. and Lee,C. (2004) Genome-wide detection of alternative splicing in expressed sequences using partial order multiple sequence alignment graphs. *Pac. Symp. Biocomput.*, 29-41.
13. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.
14. Zhang,H., Hu,J., Recce,M. and Tian,B. (2005) PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.*, **33**, D116–D120.