Nucleic Acids Research, 2003, Vol. 31, No. 1 **101–105**
DOI: 10.1093/nar/gkg029

ASAP: the Alternative Splicing Annotation Project

Christopher Lee*, Levan Atanelov, Barmak Modrek and Yi Xing

Molecular Biology Institute, Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095-1570, USA

Received August 14, 2002; Revised and Accepted October 2, 2002

ABSTRACT

Recently, genomics analyses have demonstrated that alternative splicing is widespread in mammalian genomes (30–60% of genes reported to have multiple isoforms), and may be one of their most important mechanisms of functional regulation. However, by comparison with other genomics data such as genome annotation, SNPs, or gene expression, there exists relatively little database infrastructure for the study of alternative splicing. We have constructed an online database ASAP (the Alternative Splicing Annotation Project) for biologists to access and mine the enormous wealth of alternative splicing information coming from genomics and proteomics. ASAP is based on genome-wide analyses of alternative splicing in human (30 793 alternative splice relationships found) from detailed alignment of expressed sequences onto the genomic sequence. ASAP provides precise gene exon–intron structure, alternative splicing, tissue specificity of alternative splice forms, and protein isoform sequences resulting from alternative splicing. Moreover, it can help biologists design probe sequences for distinguishing specific mRNA isoforms. ASAP is intended to be a community resource for collaborative annotation of alternative splice forms, their regulation, and biological functions. The URL for ASAP is http://www.bioinformatics.ucla.edu/ASAP.

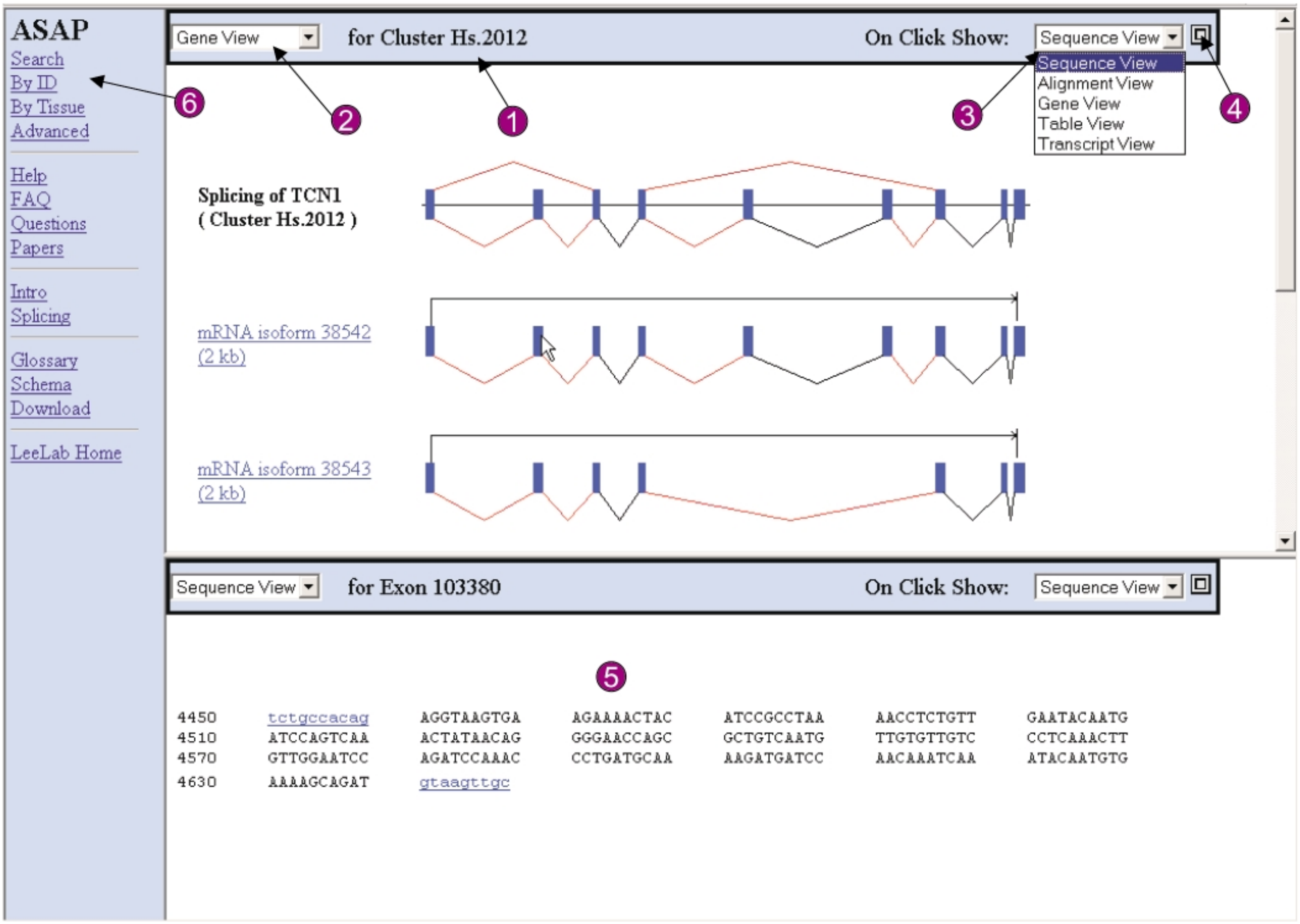
INTRODUCTION

Alternative splicing has recently emerged as a major mechanism for expanding and regulating the repertoire of gene functions (1). Previously considered to be an unusual event (estimated to occur in 5% of genes) (2), it has recently been identified in 30–60% of mammalian genes by large-scale genomics studies (3–8). If alternative splicing plays a large and important role in gene function regulation, many new questions will need to be answered (9). What is the total set of mRNA isoforms in different tissues for human, mouse and other organisms? How are these alternative splice forms regulated? Can the corresponding protein isoforms be detected and distinguished? What infrastructure is needed for high-throughput detection and quantitation of mRNA and protein isoforms, [e.g. using microarrays (10,11) and mass-spectrometry]? Most importantly, how can we establish a unified community mechanism for functional annotation of alternative splice forms?

These questions require major new database infrastructure for supporting the alternative splicing research community. There is an enormous gap between the relatively small amount of detailed biochemical studies of alternative splicing (compared with the enormous amount of work on transcriptional regulation), and the recent rapid growth in alternative splicing data generated by high-throughput genomics studies. For example, tissue specificities for only a small number of alternatively spliced genes (about 50) are listed in current alternative splicing databases (12,13), and mechanisms of alternative splicing regulation have been studied in only a small number of genes (14). By contrast, a recent genomics study identified 667–2873 tissue-specific alternative splice relationships in human genes (15). These high-throughput data could be useful to many biologists, leading to a rapid expansion in alternative splicing research, if there were a community annotation database similar to those that have been successful for gene annotation [e.g. GenBank (16)] and genetic polymorphism [e.g. dbSNP (17)].

What is needed for an alternative splicing community annotation database? First, it must bridge the gap between the enormous amount of high-throughput data and scientists who focus on specific genes or mechanistic studies. It must provide the genomics results in a form that is readily searchable and intelligible at a glance. Second, it should allow biologists to evaluate the detailed experimental evidence for each alternative splice form, and provide them with tools for probe design for distinguishing specific isoforms in their own experiments. Third, it should show the specific effects of alternative splicing on full-length protein isoform sequences, providing biochemists with useful functional predictions (such as detecting removal of a protein functional domain or localization signal). Fourth, it should provide information on regulation and tissue specificity of these isoforms, which can be useful both for mechanistic studies of splicing regulation, and for suggesting biological function. In the long run, it must enable researchers to deposit their own annotations of alternative splice forms (such as results from probing for specific isoforms in different tissues, or functional assays of different isoforms' effects). The biological functions of alternative splice forms can only be determined by

*To whom correspondence should be addressed. Tel: +1 3108257374; Fax: +1 3102670248; Email: leec@mbi.ucla.edu



```
4450    tctgccacag   AGGTAAGTGA   AGAAAACTAC   ATCCGCCTAA   AACCTCTGTT   GAATACAATG
4510    ATCCAGTCAA   ACTATAACAG   GGGAACCAGC   GCTGTCAATG   TTGTGTTGTC   CCTCAAACTT
4570    GTTGGAATCC   AGATCCAAAC   CCTGATGCAA   AAGATGATCC   AACAAATCAA   ATACAATGTG
4630    AAAAGCAGAT   gtaagttgc
```

Figure 1. Screenshot of ASAP's gene view for transcobalamin I, and sequence view for its second exon. Each view has a title bar (tag #1); left menu (tag #2) that switches between various views for the current object; right menu (tag #3) that controls what view is shown when the user clicks on a feature; and maximize/split button (tag #4) that toggles between maximizing the view to fill the whole browser window, or splitting it into two views so the user can click on a feature in the upper view and see its detailed results in the lower view (tag #5). New searches, help, and additional information are available from the navigation bar (tag #6).

experimental studies from the entire research community. Currently, in the absence of such a community annotation database, this information is not being captured in a usable form, and there is no single database where one can find answers to these questions. Instead, one's only recourse is to search and read the original literature, which is practical only for small numbers of genes.

Here, we describe the Alternative Splicing Annotation Project database (ASAP, http://www.bioinformatics.ucla.edu/ASAP), an outgrowth of our previous Human Alternative Splicing Database [HASDB (8)]. ASAP is the largest human alternative splicing database available, and is expanding to other genomes. It provides information on gene structure, alternative splicing, tissue specificity, and protein isoforms.

RESULTS

Viewing alternative splicing data in ASAP

ASAP can be searched by several different criteria such as gene name, keywords, ID [UniGene (18), GenBank (16), etc.], alternative splice type, confidence level, and tissue specificity (Fig. 1, see tag #6). It provides both graphical and table views of the alternative splicing results. The Graphical views show the 'big picture' of alternative splicing in a gene, in a form that is easily comprehensible at a glance. For example, the Gene view (Fig. 1) shows the detailed gene structure of exon and introns, with observed splicing patterns and inferred full-length isoforms. Alternative splice relationships are highlighted in red. Each feature of these Graphical views (e.g. splices, exons, isoforms) can be clicked on to 'drill down' to more specific data. As another example, the Transcript view shows the mapping of exons and the translated protein sequence for each isoform, allowing the user to easily see the relationship between gene structure, alternative splicing, and protein isoforms. These Graphical views are particularly useful for biologists who want to check a gene of interest for novel alternative splice forms and their impact. For users who wish to probe into the most detailed aspects of these data, the Table Views provide full empirical evidence for all alternative splice forms, and links to other data sources [NCBI Entrez/Genbank (16), UniGene (18), OMIM (19), GeneCards (20), etc.].

Table 1. ASAP data statistics (August 2002)

	All genes			Genes with mRNA		
	Number	Clusters		Number	Clusters	
Total unigene clusters		96 109			20 817	
Mapped to draft genome		68 032	71%		17 552	84%
Detected splices	133 369	18 173	27%	121 172	12 537	71%
Isoforms	19 384	14 015	77%	14 038	9 117	73%
Alternative splice relationships	30 793	7 991	44%	28 947	7 143	57%
Tissue specific alternative splice	2 873	1 572	20%	2 745	1 496	21%
High confidence	667	454		656	446	

Tabulation of the number of splices and number of distinct UniGene clusters in which they were observed from the total dataset (all), and clusters containing partial and/or full-length mRNAs (genes with mRNA). Percentages are given for the fraction of gene clusters successfully mapped by our procedure (8) to the January 2002 draft human genome sequence (mapped to draft genome); the fraction of mapped gene clusters observed to contain at least one splice (detected splices); the fraction of gene clusters containing isoforms, out of the total observed to contain at least one splice (isoforms); the fraction of gene clusters containing alternative splices, out of the total observed to contain at least one splice (alternative splicing relationships); the fraction of gene clusters containing tissue specific alternative splices, out of the total observed to contain alternative splices (tissue specific alternative splice); and the subset of tissue specific alternative splices that were observed with high confidence (high confidence).

One major emphasis of ASAP is on the detailed experimental evidence for each splice form. Possible alternative splices in expressed sequences are first validated by mapping to a specific genomic location by strict alignment criteria, and careful verification of the implied splice sites in the genomic sequence (8). Second, ASAP restricts alternative splicing to a very specific pattern: two observed splices that match exactly at one splice site, but differ at their opposite sites. This excludes many kinds of possible artifacts such as incomplete mRNA processing (8). This definition includes alternative splicing types such as alternative 3′, alternative 5′, exon skip and mutually exclusive exons, but not other events such as intron retention. We are extending ASAP to report additional types of alternative transcript forms. ASAP's Alignment view shows the detailed splice sites and evidence of mRNA–EST-genomic sequence alignment for each splice. ASAP's Table Views enable easy browsing and drill-down for the statistics of this evidence. ASAP's gene structures have been extensively validated by comparison with NCBI Assembly (8), and its detection of splice form tissue-specificity has been validated by comparison with independent literature (15). ASAP currently restricts its Graphical Views to alternative splicing events that appear to produce full-length protein products as judged by several criteria. This filters out some alternative splicing events that may be experimental artifacts, and which are of less immediate interest to most biologists. To view all detected alternative splices regardless of their apparent impact on the protein product, users can query the Table Views. We also provide an advanced search where users can query by alternative splice type or confidence level.

By clicking the maximize/split button (Fig. 1, tag #4) the user can choose to see one view (maximized mode) or two views (split mode) simultaneously on the same page. The split mode can be used to see detailed information in one view while keeping the 'big picture' context in the other view (Fig. 1). Menus on the title bar (at the top of each view) navigate through the various views. Its left menu (tag #2) toggles between different views for the current object, while the right menu (tag #3) controls what view will be shown when the user clicks on specific objects in the window. For example, if the right menu is set to 'Sequence View', clicking on an exon displays its sequence (tag #5).

ASAP data statistics

ASAP was launched in August 2002, and currently provides alternative splicing data for human genes. Genome-wide analyses of several other organisms (mouse, *Arabidopsis*, etc.) are ongoing. In general, ASAP will seek to include data for all genomes as they become available. ASAP's current dataset has 18 173 genes with more than one exon, with a total of 30 793 total alternative splice relationships in 7991 human genes (Table 1). These results are about four-fold larger than our results from one year before (8), demonstrating the continuing rapid growth in the raw genomics data. Approximately, half of these alternative splices are supported by multiple EST or mRNA observations. Moreover, 86% are detectible only in EST sequences (as opposed to human-curated mRNA sequences) and thus are likely to be novel.

Tissue specificity of alternative splicing

ASAP also provides results from a genome-wide analysis of tissue specificity of human alternative splice forms (15). A total of 667 tissue-specific alternative splice relationships were found at high confidence (Table 1). These tissue specificity data were detected as individual splices preferred (major form) in one tissue, and not preferred (i.e. absent, or the minor form) in the set of all other tissues. The largest categories of tissue-specific splice forms are brain, testis, skin, and lymph. For each tissue-specific splice form, the user can request the detailed evidence that is regulated in a tissue-specific manner. This includes a tissue-specificity (TS) score, robustness score [see (15) for details], counts of EST observations of this splice and other splice forms in the tissue of interest versus the pool of other tissues. The user can click to see the specific cDNA clone and library preparation information for each EST. In ASAP, tissue-specificity information is divided into two user-selectable categories: a high-confidence group designed to avoid false positives (but potentially missing many real tissue-specific forms), and a larger low-confidence group that indicates many more potential tissue-specific splices. When a sample of the high-confidence set was checked against independent literature, a high fraction (8/10) of the tissue-specificities were confirmed by independent data (15).

Protein analysis

Biologists want to go from experimental detection of an alternative splice, to its effect on full-length protein sequence. ASAP makes this connection easy, by providing protein isoform sequences for each splice form. Since ASAP uses EST alignment to genomic sequence for alternative splice detection, this furnishes exact sequence information for the sequence changes it causes, in many cases suggesting a clearly interpretable functional effect. Each protein translation is generated from the human genomic sequence, using an exon–intron structure determined from EST and mRNA transcript data. By contrast, northern blot, PCR-based, or microarray hybridization-based methods do not directly read out the sequence of a novel form as sequencing does, and instead would require other means to infer its sequence. Knowing the exact sequence change that an alternative splice produces is the difference between simply having a new 'band on a gel', versus being able to apply the full resources of sequence analysis and available literature to interpreting its likely functional impact.

Experimental splice detection resources

The purpose of our alternative splicing database is to help researchers identify interesting isoforms for further experimental study. To facilitate design of new experiments, ASAP provides exon sequences that can be utilized for generation of probes. It also provides predicted mRNA and protein isoform sequences for PCR-based detection strategies, and lists of tissues in which each splice has been observed. While many further experimental approaches (such as protein mass spectrometry) are needed for studying alternative splicing and its functional importance, these features may help many researchers make a direct step from finding novel isoforms in ASAP to investigating them rapidly in the lab. We hope to add further experimental linkages and utilities to ASAP over time.

Future directions

The long-term purpose of ASAP is to provide a community annotation resource, enabling biologists to deposit their own data and functional annotations for alternative splice forms. To this end, a number of features will be added to ASAP. The database will be expanded to include alternative splicing information for other organisms (mouse, *Arabidopsis*, *Drosophila*, etc.). ASAP will provide predictions of specific isoform sizes for detection by northern or western blot along with comprehensive analysis of the impact of alternative splicing on protein functional domains. Once a researcher identifies a particular alternative splicing relationship, ASAP will generate a set of oligo probe designs to quantitate these isoforms. Researchers will then be allowed to deposit their own annotations on experimental testing of these isoforms and their functional effects. Long term, we hope that ASAP can become a community repository for functional annotation, based on open participation, oversight, and curation by members of the research community.

CONCLUSION

ASAP can be useful to biologists in several ways. Our comparisons with published literature suggest that up to 78% of our tissue specificity findings are novel (15). These data can furnish biologists with many new functional insights into well-studied genes (by identifying a novel tissue-specific splice form), which can be of great interest for further experimental study. Our data can also provide interesting functional suggestions for unknown genes, since observation of tissue-specificity (combined with other information such as homology) may itself suggest fruitful directions for research. Finally, researchers who study regulation of splicing can benefit from this large, searchable database of tissue-specific alternative splicing spanning many distinct tissue types.

ACKNOWLEDGEMENTS

We wish to thank D. Black, A. Resch, Q. Xu and K. Ke for their helpful discussions and comments on this work. This material is based upon work supported by the National Science Foundation under Grant No. 0082964. C.L. was supported by Department of Energy grant DEFG0387ER60615. B.M. is a predoctoral trainee supported by NSF IGERT Award #DGE-9987641.

REFERENCES

1. Gravely,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
2. Sharp,P.A. (1994) Split genes and RNA splicing. *Cell*, **77**, 805–815.
3. Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
4. 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.
5. Croft,L., Schandorff,S., Clark,F., Burrage,K., Arctander,P. and Mattick,J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genet.*, **24**, 340–341.
6. International Human Genome Sequencing Consortium, (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
7. Kan,Z., Rouchka,E.C., Gish,W.R. and States,D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **11**, 889–900.
8. Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide analysis of alternative splicing using human expressed sequence data. *Nucleic Acids Res.*, **29**, 2850–2859.
9. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
10. Hu,G.K., Madore,S.J., Moldover,B., Jatkoe,T., Balaban,D., Thomas,J. and Wang,Y. (2001) Predicting splice variant from DNA chip expression data. *Genome Res.*, **11**, 1237–1245.
11. Yeakley,J.M., Fan,J.B., Doucet,D., Luo,L., Wickham,E., Ye,Z., Chee,M.S. and Fu,X.D. (2002) Profiling alternative splicing on fiber-optic arrays. *Nat. Biotechnol.*, **20**, 353–358.
12. Stamm,S., Zhu,J., Nakai,K., Stoilov,P., Stoss,O. and Zhang,M.Q. (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol.*, **19**, 739–756.
13. Brudno,M., Gelfand,M.S., Splengler,S., Zorn,M., Dubchak,I. and Conboy,J.G. (2001) Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Res.*, **29**, 2338–2348.

14. Maniatis,T. and Tanis,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
15. Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754–3766.
16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
17. Sherry,S.T., Ward,M. and Sirotkin,K. (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, **9**, 677–679.
18. Schuler,G. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
19. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick, V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
20. Rebhan,M., Chalifa-Caspi,V., Priluski,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.