D80–D85 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki129

SpliceInfo: an information repository for mRNA alternative splicing in human genome

Hsien-Da Huang, Jorng-Tzong Horng[1,2,*], Feng-Mao Lin[1], Yu-Chung Chang[3] and Chen-Chia Huang[1]

Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan, [1]Department of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan, [2]Department of Life Science, National Central University, Chung-Li 320, Taiwan and [3]Department of Biotechnology, Ming Chuan University, Taoyuan 333, Taiwan

Received August 15, 2004; Revised and Accepted October 25, 2004

ABSTRACT

We have developed an information repository named SpliceInfo to collect the occurrences of the four major alternative-splicing (AS) modes in human genome; these include exon skipping, 5′-alternative splicing, 3′-alternative splicing and intron retention. The dataset is derived by comparing the nucleotide and protein sequences available for a given gene for evidence of AS. Additional features such as the tissue specificity of the mRNA, the protein domain contained by exons, the GC-ratio of exons, the repeats contained within the exons, and the Gene Ontology are annotated computationally for each exonic region that is alternatively spliced. Motivated by a previous investigation of AS-related motifs such as exonic splicing enhancer and exonic splicing silencer, this resource also provides a means of identifying motifs candidates and this should help to identify potential regulatory mechanisms within a particular exonic sequence set and its two flanking intronic sequence sets. This is carried out using motif discovery tools to identify motif candidates related to alternative splicing regulation and together with a secondary structure prediction tool, will help in the identification of the structural properties of such regulatory motifs. The integrated resource is now available on http://SpliceInfo.mbc.NCTU.edu.tw/.

INTRODUCTION

Alternative splicing is a major mechanism for controlling the expression of cellular and viral genes and is a widely occurring phenomenon. It affects how a gene acts in different tissues and under developmental states, by generating distinct mRNA isoforms that are composed of different selections of the exons, which result in variant proteins. This phenomenon occurs extensively in the human genome, and alternative splicing is commonly believed to occur in ∼30–40% of all genes.

Alternative splicing modes have been categorized into several types to reveal alternative splicing mechanisms. As depicted in Figure 1, the four major modes of mRNA alternative splicing are: (i) exon skipping; (ii) 3′-alternative splicing; (iii) 5′-alternative splicing; and (iv) intron retention. These are all defined in the SpliceInfo resource. The occurrence of the four alternative-splicing modes is identified by comparing a pair of nucleotide sequences (DNA with mRNA) or DNA sequence with protein sequence for evidence of alternative splicing in a given gene. In complex pre-mRNA systems, more than one alternative splicing mode may be found for a given pair of sequences.

As the number of the expressed nucleotide sequences and proteins has rapidly increased, it has now become possible to identify alternative splicing isoforms by computational methods. Alternative splicing has been reported to be detectable when using such evidence as expressed sequence tags (ESTs), the mRNA sequences and the protein sequences. This is because the mature mRNA embeds the alternative splicing information. ProSplicer (1) incorporates protein sequences, mRNA and ESTs as the gene expressed evidence from which to derive the exon–intron boundary of splicing within the pre-mRNA using computational alignment methods such as BLAST and SIM4. ASDB (2) is an alternative exon database, which has been constructed based on collected experimental data for alternatively spliced exons. Thanaraj *et al*. (3) developed ASD, which extended the implementation of ASDB. ASD comprises a database of computationally delineated alternative splice events as seen by the alignments of EST–cDNA sequences within genome sequences and a database of alternatively spliced exons collected from the literature. AsMamDB (4) is an alternative splicing database

*To whom correspondence should be addressed. Tel: +886 3 4227151 (Ext. 34307); Fax: +886 3 4222681; Email: horng@db.csie.ncu.edu.tw

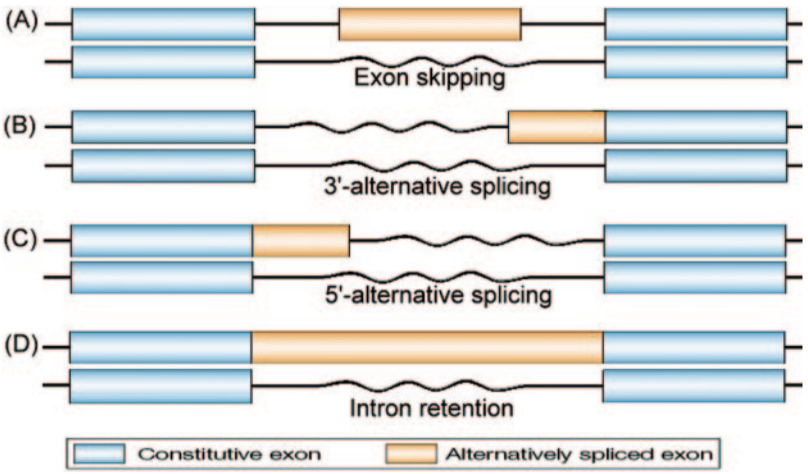

Figure 1. Four major alternative splicing modes defined in the SpliceInfo resource. In each case, the constitutively spliced exons are indicated in blue and the alternatively spliced exons are shown in orange. In case (**A**), the exon in the middle is alternative spliced into the first isoform and spliced out of the second isoform. In cases (**B**) and (**C**), the different selection of the splicing sites causes alternative isoforms. Especially in the last case (**D**), the first sequence without splicing in alternative pathway gives rise to intron retention.

that facilitates the systematic study of alternatively spliced genes in mammals. AsMamDB contains 1563 alternatively spliced genes found in human, mouse and rat, and each is associated with a cluster of nucleotide sequences. Alternative splicing patterns are represented by multiple alignments of the various gene transcripts and by graphs of their topological structures. HASDB (5) has identified 6201 alternative splice relationships in human genes based on a genome-wide analysis of expressed sequence tags (ESTs). HASDB mapped expressed sequences onto the draft human genome sequence and only accepted splices that obeyed rules for the standard splice site consensus.

Previous research has reported that exonic splicing enhancer (ESE) is a binding site of serine/arginine-rich protein (SR proteins). SR proteins belong to a family of conserved splicing factors and were first implicated in splicing when it was discovered that they are components of the spliceosome (6). SR proteins are bound to ESEs and can promote exon definition by directly recruiting the splicing machinery (6). The ability to identify the motifs shared by a set of exonic sequences that are alternatively spliced may provide targets for the investigation of alternative splicing regulatory mechanisms. Miriami *et al.* (7) performed a computer analysis of 54 sequences that have been documented as undergoing exon skipping. They identified two intronic motifs in the upstream and the downstream regions of the skipped exons. In their study, exon skipping was suggested to be controlled by sequences in the adjacent introns. They found that one motif is greatly enriched in pyrimidines (mostly C residues), and the other motif is greatly enriched in purines (mostly G residues). These two motifs differ from the known *cis*-elements at the 5′- and 3′-splice sites. They are complementary, and their relative positional order is always conserved. Numerous cancers and inherited diseases in humans are associated with mutations that cause unnatural exon skipping (8). Mutations located outside of the traditional splice sites, either in the exon or in the flanking intron sequences, have also been reported to be associated with exon skipping and diseases.

ESEfinder (9) is a web resource for identifying exonic splicing enhancers, and has been employed to develop methods of identifying putative ESEs that are related to the human SR proteins SF2/ASF, SC35, SRp40 and SRp55 and also to predict what type of exonic mutations can occur in these elements.

Previous research reveals that RNA motifs are abundant in exonic or intronic sequences associated with the regulatory mechanisms of alternative splicing. To facilitate the investigation of the regulatory mechanisms involved in alternative splicing, an integrated resource is crucial and there is a need to collect the occurrences of the major alternative splicing modes, as well as to annotate other features, such as tissue specificity of mRNA, protein domain contained by exons, GC-ratio of exons, repeats contained by exons and the gene ontology of the exonic regions plus their flanking intronic regions. This work has developed an integrated resource that collects the occurrences of various alternative splicing modes as defined in Figure 1 and provides a tool for the further analysis of the identified motif sequences involved in alternative splicing regulation. Various features provided by SpliceInfo allow the selection of a range of functions to retrieve the database contents in an efficient manner and the selection of a variety of graphical interfaces to effectively represent the analyzed results.

DATABASE STATISTICS

We have identified 14 618 known genes that have available at least two separate evidence sequences, either mRNA or protein. Each pair of the evidence sequences is compared to obtain the occurrences of alternative splicing modes as defined in Figure 1 and the data thus obtained are given in Table 1. The results show that there are 203 645 occurrences of alternative splicing contained within 6309 genes. The number of genes where there is an occurrence of ‘exon skipping’, ‘5′-alternative splicing’, ‘3′-alternative splicing’ and ‘intron retention’ are 3747, 1622, 2925 and 1776, respectively. Table 2 gives the distribution of genes containing occurrences of a particular alternative splicing mode from SpliceInfo.

Table 1. The SpliceInfo data statistics for the four major alternative splicing (AS) modes

Amount	AS modes Exon skipping	5′-alternative splicing	3′-alternative splicing	Intron retention	Total
Occurrences	149 560	25 918	14 851	13 316	203 645
Genes containing the occurrences	3747	1622	2925	1776	6309* (at least one AS-mode)
Average occurrences per gene	39.91	15.9	5.07	7.50	32.28

The first row gives the number of occurrences of each alternative splicing mode; the second row gives the number of genes containing the occurrence of each AS-mode; the last row shows the average number of occurrence per gene. The symbol '*' indicates that there are 6309 genes that contain at least one occurrences of any modes of alternative splicing. This is not the total number of genes that make up each alternative splicing mode.

Table 2. The distribution of genes containing the occurrence of a particular AS mode in SpliceInfo

AS modes Exon skipping	5′-alternative splicing	3′-alternative splicing	Intron retention	Number of genes containing the occurrences of particular AS-modes
v				1818
		v		944
			v	619
v		v		573
v	v	v		357
	v			311
		v	v	291
v	v			290
	v	v		240
v			v	223
v		v	v	219
v	v	v	v	205
	v	v	v	96
v	v		v	62
	v		v	61

The 'v' annotated in the first four columns means that the genes contain that particular mode of alternative splicing. For example, 1818 genes contain the occurrences of only one mode, namely, 'exon skipping'. While in the sixth row, 357 genes contain three modes, namely, 'exon skipping', '5′-alternative splicing' and '3-alternative splicing'.

MATERIALS AND METHODS

The data flow of the SpliceInfo is briefly depicted in Figure 2.

Deriving the occurrences of alternative splicing modes

The occurrences of the alternative splicing modes are first derived computationally from several data sources, such as ProSplicer release 3 (1) and Ensembl version 19_34 (10). The pairing of a DNA sequence with an mRNA sequence or a protein sequence from the same gene allows a comparison that can derive the alternative splicing pathway. In case of nucleotide–nucleotide comparison, the exonic–intronic boundaries of the two evidence sequences with respect to the same gene are compared. Each differential region between the two sequential exon sets is used to determine the type of alternative splicing mode involved. Similarly, in the case of nucleotide–protein comparison, because the nucleotide evidence and the protein evidence are aligned within the same gene sequence, the differential regions between the two pieces of information can also be determined directly. At least 10 bp for each occurrence within the exonic regions undergoing alternative splicing are derived and stored in the database.

Annotating sequence features in the exonic regions that are alternatively spliced

The sequence features within the exonic sequences, which are alternatively spliced, are annotated. The protein domain is another interesting aspect of splicing (11). Liu and Altman (11) carried out a large-scale study of protein domain distribution in the context of alternative splicing and this revealed various facts about protein domains, namely, that they are disproportionately distributed, and many are on different sequences after alternative splicing. The protein domain data were obtained from InterPro (12), which is an integrated resource of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. Each reading frame of the exonic nucleotide sequences is translated into proteins and then the protein sequence is scanned for protein domain instances. All the hits within the exonic regions are stored in the SpliceInfo database. The data for gene functions and tissue-specificity were obtained from Gene Ontology (GO) (13), which is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO collaborators are developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. Other sequences features such as repeats like SINE, LINE, STR, etc., and the GC-ratio were obtained from the Ensembl database (10) for the genomic regions of the considered exonic or intronic sequences.

Selecting exonic sequences by alternative splicing modes and other features

Various features selecting filters are provided in the resource and these allow users to select the exonic sequences based on specific constraints such as tissue specificity of the mRNA, the protein domains contained by exons, the GC-ratio of exons, the repeats contained by exons and the GO. For instance, a set of exonic sequences and two sets of flanking intronic sequences can be constructed in the case of the 'exon skipping' mode.

Discovering motifs related to alternative splicing regulation

For further analyses of the motif sequences related to the regulatory mechanism of alternative splicing, DNA/RNA

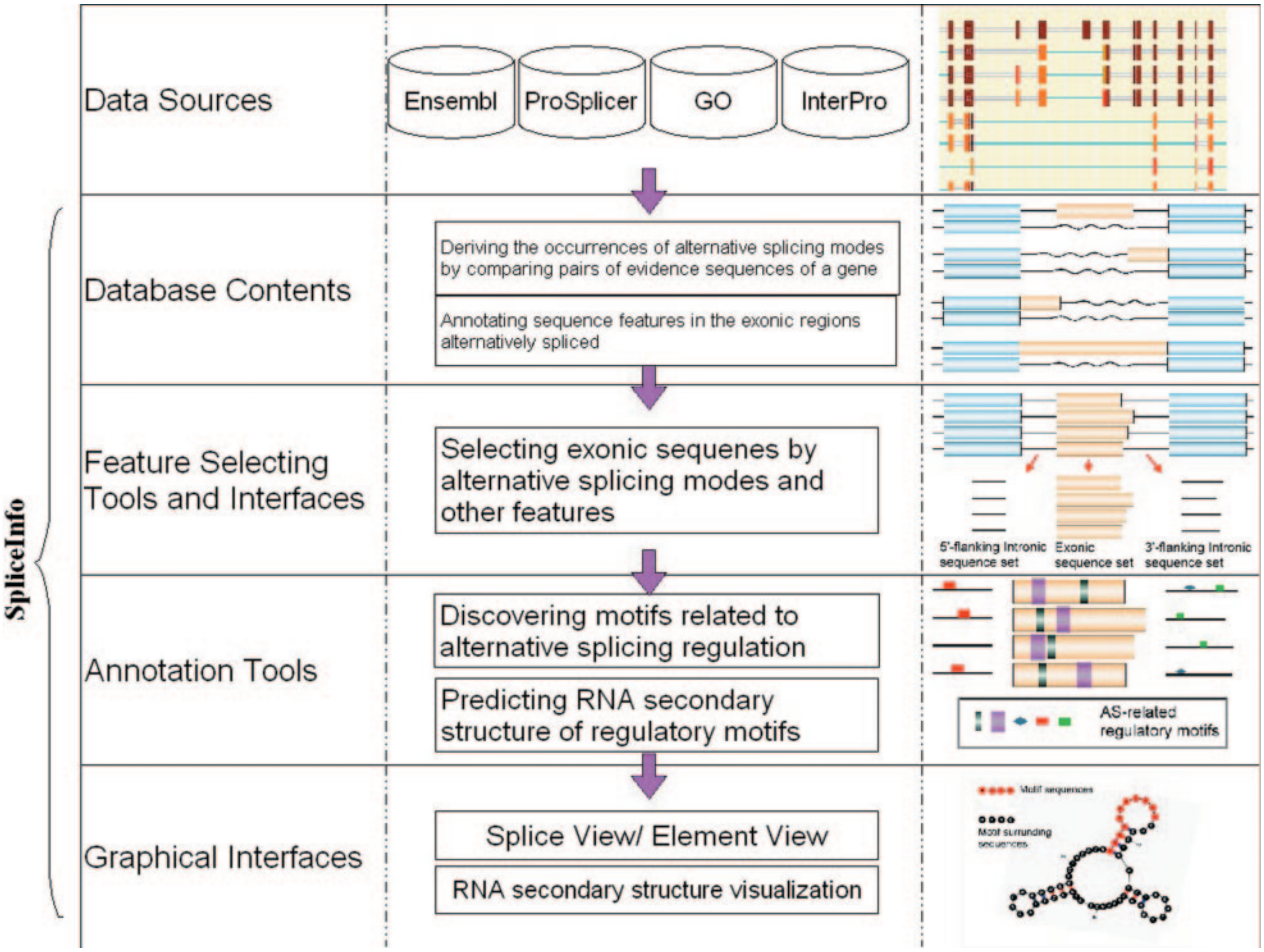
SpliceInfo
Data Sources
Ensembl
ProSplicer
GO
InterPro
Database Contents
Deriving the occurrences of alternative splicing modes by comparing pairs of evidence sequences of a gene
Annotating sequence features in the exonic regions alternatively spliced
Feature Selecting Tools and Interfaces
Selecting exonic sequenes by alternative splicing modes and other features
5'-flanking Intronic sequence set
Exonic sequence set
3'-flanking Intronic sequence set
Annotation Tools
Discovering motifs related to alternative splicing regulation
Predicting RNA secondary structure of regulatory motifs
AS-related regulatory motifs
Graphical Interfaces
Splice View/ Element View
RNA secondary structure visualization
Motif sequences
Motif surrunding sequences

Figure 2. Data flow of the SpliceInfo resource. In the right-bottom subfigure, the nucleotides indicated by the red circles constitute the motif, itself, and the nucleotides indicated by the black circles are the motif surrounding sequences.

motif discovery tools, such as MEME (14) and Gibbs sampler (15), and the RNA secondary structure prediction tool, Mfold (16), are integrated into SpliceInfo. In order to identify the binding motifs in a group of intronic or exonic sequences, we integrated two popular regulatory motif prediction programs, namely, the Gibbs sampler and MEME to identify DNA motifs that are potentially regulatory motifs involved in the regulation of alternative splicing. Based on the result of the DNA motif discovery, consensus patterns or position matrices are obtained and these are stored as motifs in the form of a consensus pattern including both motif sequences occurring in the intronic or exonic regions. When motifs are found, a sequence logo (17) is created for each motif. Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment. In general, a sequence logo provides a richer and more precise description of a binding site, than would a consensus sequence.

Predicting RNA secondary structure of regulatory motifs

SpliceInfo also provides the RNA secondary structure of the predicted motifs. Users can build the secondary structure on the web interface by specifying different parameters. Mfold (16), which is a tool for predicting the RNA secondary structure, is used to predict the secondary structure of the regulatory motif. It will predict the optimal secondary structure as well as some suboptimal secondary structures and the best are displayed by the web interface while the others are discarded.

Interfaces

To facilitate the data access and further analyses, several query forms and graphical interfaces have been designed and implemented in the resource. A feature filtering form allows users to specify particular constraints when selecting the sequence regions. For instance, a user can specify the alternative splicing mode that should occur for a particular group of genes and exonic sequences that interest the user. After the user has specified the query options and the form has been submitted, the sequence regions that meet the query constraints are extracted from the database of the SpliceInfo resource. The resource facilitates users in the tailoring of 5′-flanking sequence and 3′-flanking sequences for each selected region.

The SpliceInfo integrates two motif discovery tools and facilitates the detection of motifs in the sequence region set established constructed by user. Two motif discovery tools, MEME (14) and Gibbs sampler (15), are provided by the web

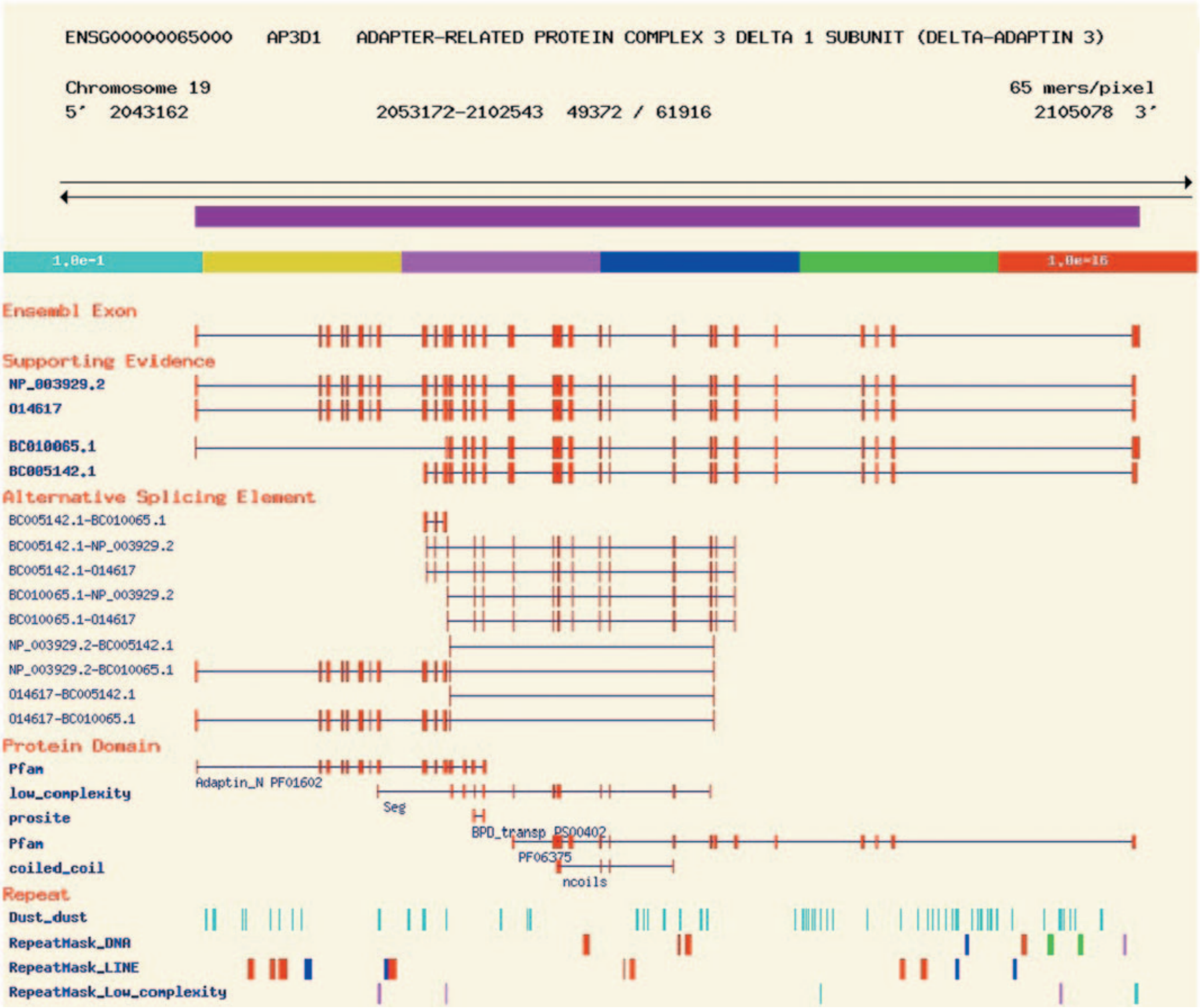

Figure 3. Gene view in SpliceInfo.

interface. Users can apply the tools separately to each sequence set. The sequence logo (17) is employed to present the content of the motif. Accordingly, a user can easily understand the composition of the nucleotide motif sequence and the nucleotide percentage of the motif. Figure 3 presents one graphical view of examples of such motifs and other sequence features in exonic and intronic regions for all the evidence sequences of a gene. Users can zoom in/out, shift the window left/right and select particular locations to check regulatory motifs or features related to alternative splicing in an easy manner.

CONCLUSIONS

We present here an information repository that collects the occurrences of alternative splicing in the human genome; the resulting dataset is derived from a pair of evidence sequences for each gene. Annotated features are also provided for the characterization of the alternative splicing events. Further, analyzing tool supports the identification of motifs related to the regulatory mechanism of alternative splicing both at the primary level and at the structural level. Further development of the resource will result in the inclusion of (i) a collection of the occurrence data of alternative splicing modes in other species such as mouse, rat, yeast and fly; (ii) support for more complex alternative splicing modes such as 'mutually exclusive exons'; and (iii) the incorporation of a control sequence set that will aid the motif discovery process by assessing the significance of the predicted motifs.

ACKNOWLEDGEMENTS

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 93-2213-E-008-024 and Contract No. NSC-93-2213-E-009-075.

REFERENCES

1. Huang,H.D., Horng,J.T., Lee,C.C. and Liu,B.J. (2003) ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data. *Genome Biol.*, **4**, R29.
2. Dralyuk,I., Brudno,M., Gelfand,M.S., Zorn,M. and Dubchak,I. (2000) ASDB: database of alternatively spliced genes. *Nucleic Acids Res.*, **28**, 296–297.
3. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res.*, **32**, D64–D69.
4. Ji,H., Zhou,Q., Wen,F., Xia,H., Lu,X. and Li,Y. (2001) AsMamDB: an alternative splice database of mammals. *Nucleic Acids Res.*, **29**, 260–263.
5. Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.
6. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.*, **3**, 285–298.
7. Miriami,E., Margalit,H. and Sperling,R. (2003) Conserved sequence elements associated with exon skipping. *Nucleic Acids Res.*, **31**, 1974–1983.
8. Murakami,T., Sakane,F., Imai,S., Houkin,K. and Kanoh,H. (2003) Identification and characterization of two splice variants of human diacylglycerol kinase eta. *J. Biol. Chem.*, **278**, 34364–34372.
9. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
10. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
11. Liu,S. and Altman,R.B. (2003) Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic Acids Res.*, **31**, 4828–4835.
12. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
13. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
14. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
15. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
16. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
17. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a Sequence Logo Generator. *Genome Res.*, **14**, 1188–1190.