

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

Hsien-Da Huang¹, Jorng-Tzong Horng^{2, 3, *}, Feng-Mao Lin², Yu-Chung Chang⁴, and
Chen-Chia Huang²

¹*Department of Biological Science and Technology, Institute of Bioinformatics
National Chiao Tung University, Hsin-Chu, Taiwan*

²*Department of Computer Science and Information Engineering, National Central University, Taiwan,*

³*Department of Life Science, National Central University, Taiwan,*

⁴*Department of Biotechnology, Ming Chuan University, Taiwan*

** Author for correspondence: e-mail: horng@db.csie.ncu.edu.tw
Phone: +886-3-4227151 Ext. 34519*

ABSTRACT

As an integrated database, SpliceInfo identifies automatically the conserved sequences in selected exon/intron regions of a gene group. Several modes of mRNA alternative splicing, such as exon skipping, alternative 5'-splicing sites, alternative 3'-splicing sites and mutually exclusive exons are computationally derived and extracted. Finally, for each type of alternative splicing, the flanking intronic sequences are collected and then exploited by motif discovery tools. The tissue-specific information and gene functionalities that correspond to the selected regions are also considered. The database provides a means of investigating alternative splicing and can be used for identifying alternative splicing - related motifs, such as the exonic splicing enhancer (ESE), the exonic splicing silencer (ESS) and other intronic splicing motifs. The integrated resource is now available on <http://140.113.239.236/SpliceInfo/>.

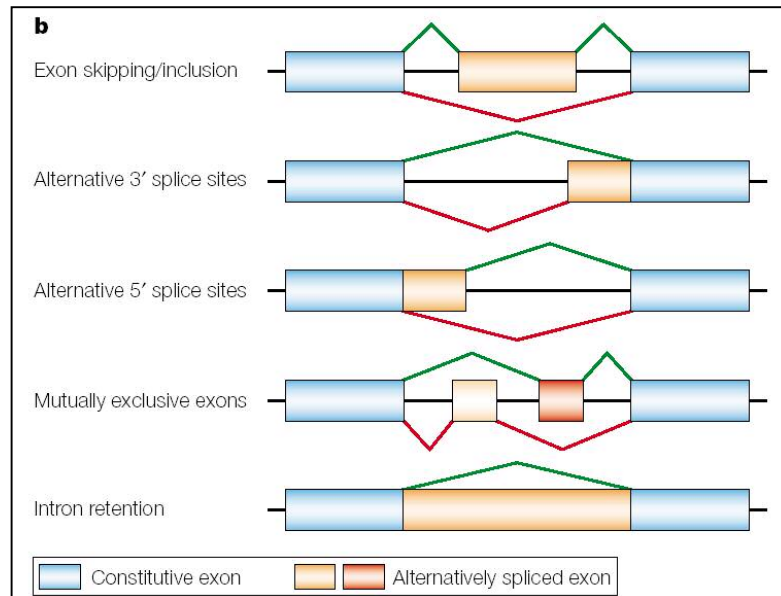
INTRODUCTION

Alternative splicing is a major mechanism for controlling the expression of cellular and viral genes. It is an extensively occurring phenomenon. It affects how a gene acts in different tissues and developmental states, by generating distinct mRNA isoforms that are composed of different selections of exons and produce variant proteins. This phenomenon occurs extensively in the human genome, and alternative splicing is commonly believed to occur in only approximately 30% ~ 40% of all genes. Alternative splicing can generate variant isoforms of mRNA and has implicated in many processes, such as sex determination, apoptosis and acoustic

tuning in the ear.

The alternative splicing modes are categorized into several types to reveal alternative splicing mechanisms. As depicted in Fig. 1 (1), five classical alternative splicing types exist; these are exon skipping, alternative 3' splice sites, alternative 5' splice sites and mutually exclusive exons and intron retention.

Figure 1. Classical splicing signals and modes of alternative splicing (1).



Exonic Splicing Enhancer (ESE) is a binding site of Serine/Arginine-rich protein (SR proteins). SR proteins belong to a conserved splicing factors family and are firstly implicated in splicing when it was discovered that they are components of the spliceosome (1). SR proteins that are bound to ESEs can promote exon definition by directly recruiting the splicing machinery (1).

Numerous cancers and inherited diseases in humans are associated with mutations that cause unnatural exon skipping. Generally, the mutations affect the splice sites; for example, a point mutation at the 5' splice site of exon 7 of Wilm's tumor (2) suppressor gene causes unnatural skipping of exon 7 and generates a truncated protein that is associated with Wilm's tumor. 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.

Today, the number of discovered sequences is increasing exponentially and alternative splicing sites can be predicted by computational methods. Alternative splicing has been recently reported to be detected in expressed sequence tag sequence. EST sequences embed alternative splicing information. The protein sequence

translated from mRNA also embeds alternative splicing information.

Previous research reveals that RNA motifs are conserved in the exonic or intronic sequences that are associated with the mechanisms of the alternative splicing. An integrated system that facilitates the prediction of conserved sequence elements associated with a particular type of alternative splicing is crucial in deciphering the mechanisms of alternative splicing.

This work develops an integrated approach to provide various modes of alternative splicing information and automatically identify alternative splicing (AS)-related motif sequences in the human genome. The aim here is to derive and extract the information about alternative splicing from the gene expression evidences, such as the covering exon skipping, the 5' - alternative splicing, the 3' - alternative splicing and other alternative splicing modes. Various filtering functions provide means to query the SpliceInfo database. Also, tissue-specific information and gene functionalities are also considered. Several analyzing tools such DNA/RNA motif discovery tools and RNA secondary structure prediction are provided and integrated and can be further applied to the selected modes of alternative splicing.

RELATED WORKS

In previously work by the authors' group (3), ProSplicer was used to extract information of the alternative splicing using computational alignment methods such as BLAST and SIM4. The nucleotide sequences, mRNA and EST, provide evidence of gene expression to reveal the alternative splicing modes of the gene; the protein sequences are theoretically translated into nucleotides in six reading frames and are aligned against the genomic sequences.

InterPro (4) is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.

The Gene Ontology (GO) project (5) 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.

In order to identify the binding motifs in a group of intron/exon regions, we integrate three popular regulatory motif prediction programs, which are Gibbs sampler and MEME to discover DNA motifs, which are potentially regulatory motifs. The Gibbs sampler (Charles Lawrence's Gibbs Motif Sampler (Version 1.01.009) was used (6), with the option 'motif sampler'. 100 different 'seeds' or starting points were used, a maximum of

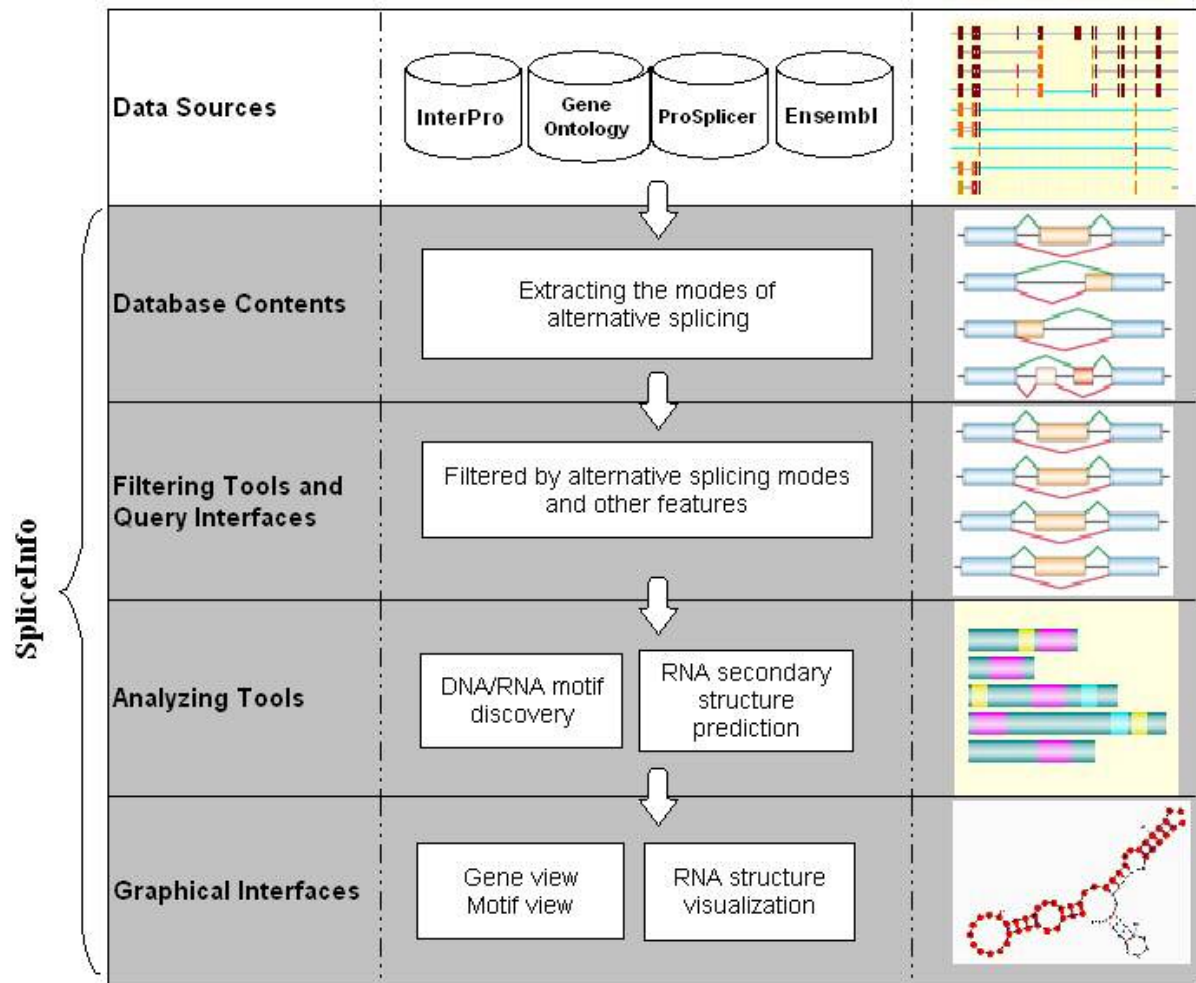
2000 iterations were performed for each run, and the highest scoring result was reported. The MEME algorithm uses an expectation maximization algorithm for finding patterns in input sequences. MEME version 2.2 (7) was run over the MEME web-server. The top scoring result was reported. As the result of the DNA motif discovery methods all result in consensus patterns or position matrices, we store the motifs as the format of the consensus pattern including the motif sequences occurring in the intron/exon regions.

Mfold (8) (9) is a tool for predicting the secondary structure of RNA and DNA, mainly by using thermodynamic methods. The core algorithm predicts a minimum free energy as well as minimum free energies for foldings that must contain any particular base pair. Base pairs within this free energy increment are chosen either automatically or else by the user. Then foldings that contain the chosen base pair are computed.

WebLogo (10) is a web-based application designed to make the generation of sequence logos (11) as easy and painless as possible. Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by Tom Schneider and Mike Stephens. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

Miriam *et al.* performed a computer analysis of 54 sequences, documented as undergoing exon skipping, and identified two motifs in both the upstream and the downstream introns of the skipped exons (12). Exon skipping has been 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. ESEfinder (13) is a web resource for identifying exonic splicing enhancers, and has been employed to develop methods of identifying putative ESEs, which correspond to the human SR proteins SF2/ASF, SC35, SRp40 and SRp55, also want to predict what type of exonic mutations occur in these elements.

Figure 2. System flow of the SpliceInfo system.



METHOD

The system flow of the SpliceInfo system is briefly depicted in Fig. 2. The modes of mRNA alternative splicing are firstly defined. The database contents of the modes of mRNA alternative splicing are extracted from several data sources, such as ProSplicer (3) and Ensembl (14). The annotated features such as protein domains, gene functions and tissue-specificities are obtained from InterPro (4) and Gene Ontology (5). Various filtering tools are provided in the system and allow users to query the genes under particular constraints, such as the alternative splicing modes, protein domain containing, gene functions, repetitive features and tissue specificities. A set of genes and the selected sequence regions under the same alternative-splicing consideration is constructed.

DNA/RNA motif discovery tools, such as MEME (15), and RNA secondary structure prediction tool, such as Mfold (9), are integrated. It facilitates users to execute the tool to their constructed sequence set on the web. Furthermore, several graphical interfaces are designed and implemented in the system and will be described in

tissue-specificity of a protein was extracted from the UniProt database (16). A connection is made between the expressed tissue of ESTs and tissue-specificity. The simplest way - collecting all ESTs sequence expressed on the chosen tissue – us used. The authors hope to be able to use more statistical information on tissue-specificity in the future.

The protein domain is another interesting aspect of slicing (17). Liu et al. studies in a large-scale the protein domain distribution in the context of alternative splicing and revealed various facts about protein domains; they are disproportionately distributed, and many are on different sequences after alternative splicing. Alternative splicing can control protein function and related protein domains, so protein domain and protein family information from InterPro (4), and protein product descriptions from Gene Ontology (5) were used herein. Sequence elements were linked to the protein features, which were used to prove that a dataset with a single feature, can yield some conserved sequence elements, and be involved in splicing to produce a particular function of protein.

Not only they determined that the domains are disproportionately distributed, but they also suggested that alternative splicing is related to RNA structure. Spliceosome interacts with the RNA sequence during splicing. Some information about RNA structure is sought, and more information on the RNA secondary structure is provided herein.

DNA/RNA Motif Discovery Tools

The web system can use several motif finding tools, MEME (15), Gibbs Sampler (6) and AlignACE (18), to analyze the sequence of regions and identify conserved sequence elements, the motif, in the selected exon/intron regions. Each tool is separately applied to two or three sequence regions. Numerous motifs are found using these tools. The authors hope that the motifs found using motif-finding tools are specific motifs with specific functions, conserved order motifs, complementary motifs with specific structures and other types of motifs). Motifs with specific function are those that always interact or are expressed under some particular conditions. Conserved order motifs are those that are always shown in a conserved order and perhaps co-expressed under some conditions.

When motifs were found, a sequence logo was created for each motif. Accordingly, other data are associated with each motif. These data are PSSM, FASTA, relative position and statistics. Then, the user can view the sequence logo and all other information using the motif display part of the web system.

RNA Secondary Structure Prediction

This work also provides the secondary structure of the motifs, after an Mfold (9) was applied using some commonly used parameters. A motif consists of several conserved sequences, so the first sequence with the lowest e-value is used to present the RNA secondary structure of this motif. Additionally, a user can rebuild the secondary structure on the web system by specifying different parameters. Mfold, which is a tool for predicting the RNA secondary structure, is used to predict the secondary structure of the motif discovered in the motif discovery phase. It will predict the optimal secondary structure and some suboptimal secondary structures, of which the best is displayed in the web system while the others are left behind. After a picture of the secondary structure has been established, plt2gif is used to annotate the secondary structure. Red and black are used to annotate the relative position of the motif occurred: red indicates the position of the motif and black indicates the position of motif surrounding sequences.

QUERY INTERFACE

As presented in Fig. 4, a filtering form allows users to specify particular constraints to select the sequence regions. For instance, a user can specify which alternative mode of splicing is considered. Additionally, several filtering functions are provided in the form given in Fig. 4, including GC ratio, expressed tissues, gene functions and repeat features.

Figure 4. Form of filtering functions in web interface.

Gene	Ensembl ID Group	<input type="text"/>
	Gene Name Group	<input type="text"/>
	Splicing Type	<input type="text"/>
	With Spliced-Out Protein Domain	<input type="checkbox"/>
Exon / Intron	Exon	<input type="checkbox"/> Use whole exon regions
	Intron	<input type="checkbox"/> Use whole intron regions
Element (Selected Region)	Alternative Splicing Type	<input type="text"/>
	Cover Repeat Feature	<input type="text"/>
	GC Ratio	>= <input type="text"/>
	Element Minimum Length	>= 9 <input type="text"/>
	Tissue which expressed in Expressed Sequence Tags (ESTs)	<input type="text"/> Pick Tissue
Protein	Tissue Specificity of Human Protein	<input type="text"/> Pick Tissue
	Covers whole or partial protein domain	<input type="checkbox"/>
	Gene Ontology Aspect	<input type="text"/>
	Gene Ontology Term	<input type="text"/>
		<input type="button" value="List all filtered data"/> <input type="button" value="View Selected Elements"/>

After users have specified query options and the form has been submitted, the sequence regions that meet the query constraints are extracted from the database in the SpliceInfo system. The system helps users to tailor a 5'-flanking sequence and 3'-flanking sequences of each selected region. Using the form shown in Fig. 5, users can specify the length of both flanks of the selected region. For instance, if a user constructs a sequence set which consists of exons skipped in "exon skipping" mode, then the two flanking regions are the 5'-flanking intronic sequence and the 3'-flanking intronic sequence. All three sequences - the selected regions, the 5'-flanking regions and the 3'-flanking regions - are further analyzed.

Figure 5. Sequence tailoring form for selected region and flanking regions.

Motif Discovery			
Sequence Type	Length	FASTA	Run?
5' upstream	500	<pre>>ASMotif_1_1 50491033 50491119 -1 ENSG00000185104 500 AGTAAGTGACAGTCAAATTTTCAGTTTATATCCTCCTTCTCTTTGGCCACTTTCAGTCTAATTGGCTTTGATAATAAG >ASMotif_2_1 1215439 1215884 -1 ENSG00000175756 500 GGTGAGTGGCCCCCTGCCTGGCCCTGGGAGGGCAAAGTGTGAGAACAGTTTCTTTGCCACGAATTACTGGCGGTCCC >ASMotif_3_1 50431712 50431807 -1 ENSG00000185104 500 AGTAAGCCTGTTTGGCCAGCTTTTCTTTGAACAGAAAGGCCCATCCCCACCTACCTGGCCACTTCCCTTTTTTTTATCTG >ASMotif_4_9 91421212 91421322 1 ENSG00000165238 500 ACACCCCGCATACAGCTTCTCCCGGGCTGGGACAGTGGAAAGGGTGCCTTGCCATCACCTCCCTGCACACTTGTCTTGC</pre>	<input type="checkbox"/>
selected region		<pre>>ASMotif_1_1 50491033 50491119 -1 ENSG00000185104 87 GGTAAAGAGAAATGTGTATGACCTTACAAGTATCCCGTTCCGCCACCAATTATGGGAGGGCTGGCCAACTTCTGTCTACAG >ASMotif_2_1 1215439 1215884 -1 ENSG00000175756 446 GGCGCCCGCCCGCTTGGCCCGTCTCTGGAGTGTCTGGGAGCCGGGTCTGGGGCCCTTTACAGCACATCGCCGGCCGG >ASMotif_3_1 50431712 50431807 -1 ENSG00000185104 96 GATGTGTCTTGTGTAATCAGGGCTCTTATCCCTGCCATCGACTTACAGTGGGAAGAGATCTTACCTGCACAGACCC >ASMotif_4_9 91421212 91421322 1 ENSG00000165238 111 TGACCGCACCTCGAGCAGGAGTGGGATGCCACGTCTGCCCCAGCCCGCCCGCTCTGTCCACACGGTCAATCCCGGA</pre>	<input checked="" type="checkbox"/>
3' downstream	300	<pre>>ASMotif_1_1 50491033 50491119 -1 ENSG00000185104 300 TTAGTGAACTGGGTGCCTGAAGTTTCTTGAAGTAAAAGAAATGAAGTACAGCCAGGAGTTGAAAAACAAGAGGGGAGGATA >ASMotif_2_1 1215439 1215884 -1 ENSG00000175756 300 GGCGCCTGACTTCCCAGCTGTTGAGGGCCGTTCTTGGGAGGTAGGAAGCCCGGGGGGGGATGGGAGGATGCACAC >ASMotif_3_1 50431712 50431807 -1 ENSG00000185104 300 ACGCACATTGAGGCAGTGGGGGATTGGGGGCAAGTATGTATGATTTGGCTTATTTGGGCTCATACCACTTCTCAGAAC >ASMotif_4_9 91421212 91421322 1 ENSG00000165238 300 GTACTGCCCTCTCCACCTCCCAACCCCAACCTGTGGGCTGCCTTGGGTCAGTACAGAGGCTCTGCCACGCTCCAGTG</pre>	<input type="checkbox"/>

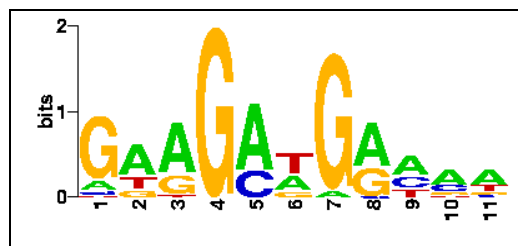
The SpliceInfo system integrates numerous motif discovery tools and facilitates the detection of motifs in the sequence region set established constructed by users. Three motif discovery tools - MEME (15), AlignACE (18) and Gibbs sampler (6) - are provided in the web interface. Users can apply the tools separately to each sequence set. As shown in Fig. 6, before the tool is used, a user can specify some parameters that specify each motif discovery tool.

Figure 6. Interface of the motif discovery tools.

Motif Discovery			
Program Name	Parameter Name	Parameter Value	Run?
MEME	Distribution among the sequences	<input type="radio"/> One per sequence <input checked="" type="radio"/> Zero or one per sequence <input type="radio"/> Any number of repetitions	<input type="checkbox"/> Original HTML <input type="button" value="MEME"/>
	Maximum number of motifs to find	3	
	Optimum number of sites [Optional]	<input type="text"/> Minimum sites (≥ 2) <input type="text"/> Maximum sites (≤ 300)	
	Optimum width of each motif	<input type="text"/> Minimum width (≥ 2) <input type="text"/> Maximum width (≤ 300)	
Gibbs sampler	No. of different motifs (patterns):	5	<input type="button" value="Gibbs sampler"/>
	Max sites per seq: (recursive sampler)	5	
	Motif Width(s):*	10,10,8,10,10	
	Est. total sites for each motif type:	18,12,13,10,10	
AlignACE	Number of columns to align	10	<input type="button" value="AlignACE"/>
	Number of sites to expect	10	
	Fractional background GC content	0.38	

As depicted in Fig. 7, the sequence logo (10) is employed to present the content of the motif. Accordingly, a user can easily understand the composition of the nucleotide and the nucleotide percentage of the motif.

Figure 7. Sequence logo of a conserved sequence elements.



As shown in Fig. 8, a user can easily understand the energy of the input sequence, which consists of the motif sequence and two flanking sequences. The energy dot plot tells the user some information about the energy of the input sequence and the user can easily determine the possible structure at particular positions. An RNA secondary structure is predicted with the color-coded annotation, which differentiates the motif and the two flanking sequences. The nucleotides indicated by the red circles constitute the motif, itself, and the nucleotides indicated by the black circles are the motif surrounding sequences.

Figure 8. RNA secondary structure of an alternative splicing motif.

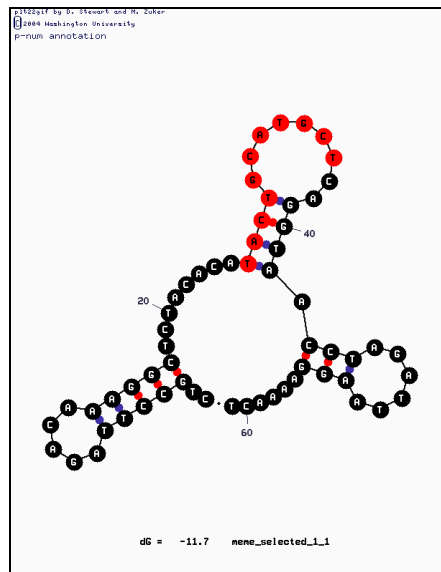
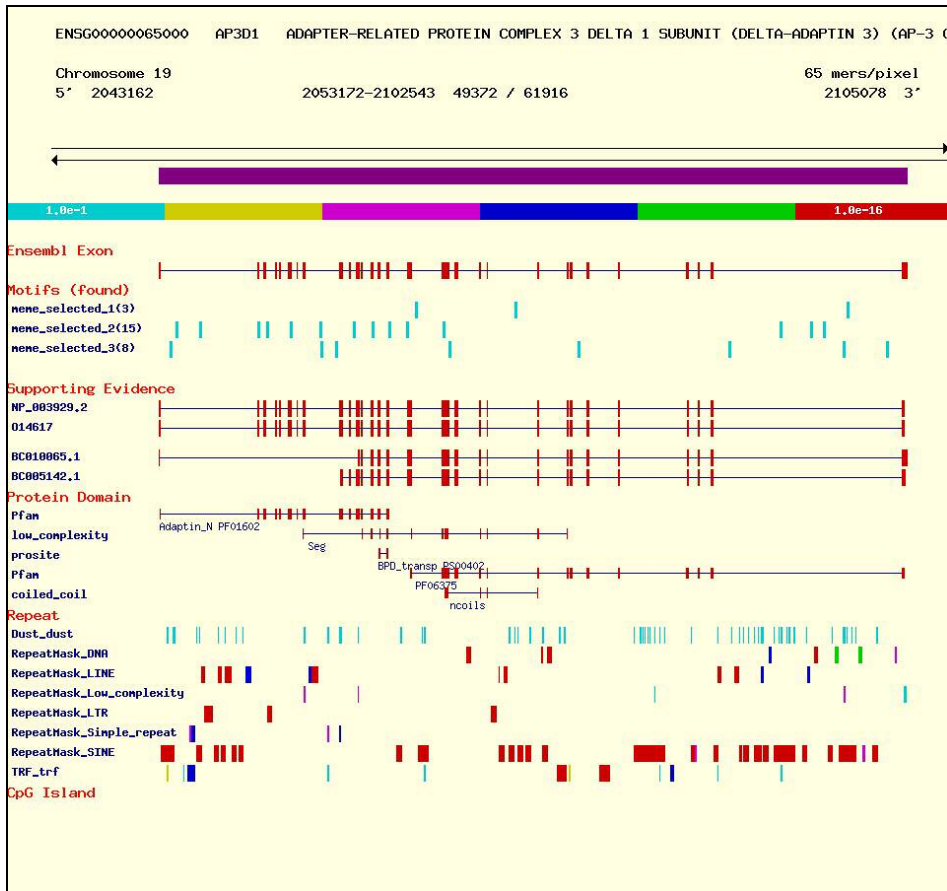


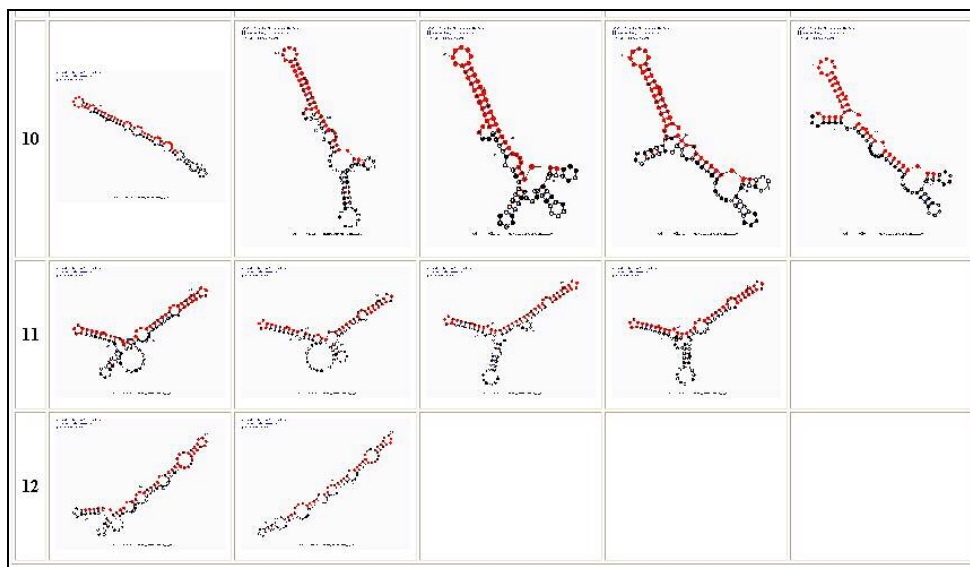
Figure 9 present is one of the graphical views to show the found motifs in exonic and intronic regions in all the evidence sequences of a particular gene. The user can zoom in/out, shift the window left/right, and present specific position on the gene or chromosome orientation. The user can easily view more widely or more deeply the sequence, and check information on the motifs.

Figure 9. Gene view of the motifs.



The RNA secondary structures of all instances of a alternative splicing-related motif are predicted. All the structures are provided at the same time in the same web page, as depicted in Fig. 10, to allow users compare all the structures of the motif.

Figure 10. Sub-optimal RNA secondary structures.

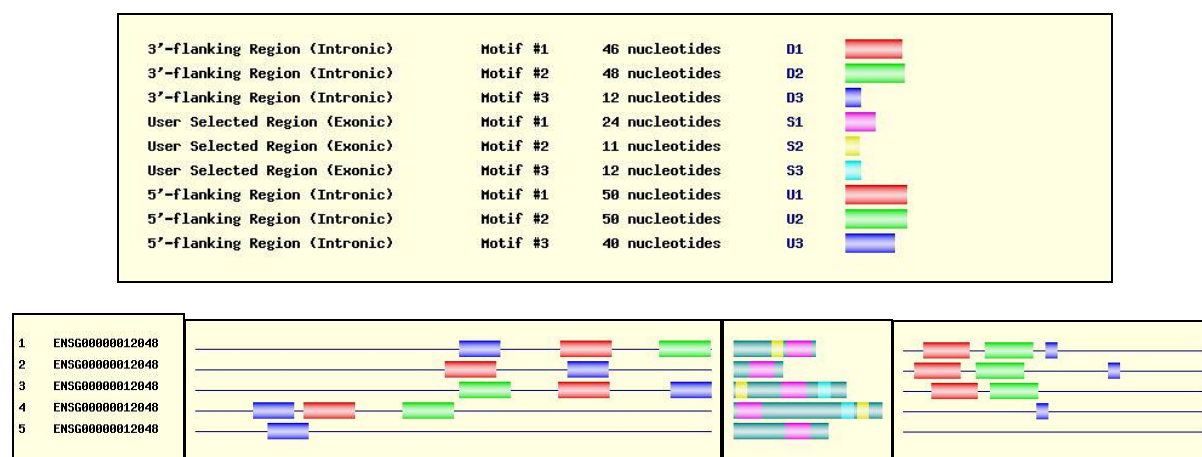


BIOLOGICAL APPLICATIONS

An earlier investigation found that the full-length BRCA1, the $\Delta(9,10)$ (NM_007302), the $\Delta(11q)$ (NM_007305) and the $\Delta(9,10,11q)$ (NM_007304) variants are expressed in a variety of tissues, under various conditions. They are therefore called predominant splice variants.

Five alternative spliced elements, annotated to be skipped exons as a selected region were collected and the motif discovery process was applied. The aim was to find some conserved sequence elements that regulate exon skipping, such as one underlying mechanism of exon skipping on BRCA1 or another gene. Figure 11 reveals some results concerning motifs on skipped exons of BRCA1.

Figure 11. Motifs in regions of exon skipping in BRCA1.



A Variety of Genes Which Has ESEs

Other investigations have predicted ESEs on several genes (ACF, BRCA1, BRCA2, FBN1, IGF1, PDHA1, SMN1, SMN2, TNFRSF5 and CFTR), which have also been verified, as given in Table 1. The web system can be used to find these ESEs and validate them.

Table 1. A variety of genes which has ESEs.

Gene Name	Ensembl Gene ID	Exon #	Strand	Chr	Exons	
					Start	End
ACF (NM_014576)	ENSG00000148584	12	-1	10	37418	37552
BRCA1	ENSG00000012048	17, 18, 19	-1	17	52944, 55475, 59822	52989, 55580, 59961
BRCA2	ENSG00000139618	17, 19	1	13	47044, 54923	47217, 55078
FBN1	ENSG00000166147	51	-1	15	99055	99180
IGF1	ENSG00000017427	5	-1	12	73389	73545
PDHA1	ENSG00000131828	7, 8	1	X	11417, 11754	11572, 11825
SMN1	ENSG00000172062,	7	1	5	20963	21073
SMN2	178958, 179850	7	-1	5	13079	13150
TNFRSF5	ENSG00000101017	3,5	1	20	3967, 4860	4092, 4953
CFTR	ENSG00000001626	9, 12	1	7	62054, 107777	62146, 107871

The case studies above demonstrate that that the proposed system can discover conserved sequence elements within genomic sequences, and these motifs represent the functional site of user-selected regions or two flanking regions.

DISCUSSIONS

The system has a few limitations of input data size, execution time and waiting time. Basically, the system is an online web system, so some congenital restrictions apply. Limited by computational power, the system cannot accept large datasets that consists of very long sequences, because the waiting and executing times are too long. Numerous web users cannot wait for the program to respond after several hours. Accordingly, the aim is to reduce waiting time to maybe under an hour. The authors hope to be able to apply a more computationally powerful server to enable users to obtain results in minutes.

ESEfinder (13) uses four SR proteins as materials that are stored as matrices. It accepts sequences in FASTA format. The proposed system can predict conserved sequence elements within a set of sequence regions and the motifs associated with the particular alternative splicing signals are found. It provides “sequence logos” for DNA motifs, and predicts the RNA secondary structure for a motif and its flanking sequences.

More attention should be paid to the improvements of the developed system, such as large dataset support and response time reduction. The system can be extended in several ways. The first is to filter alternative splicing modes. More filtering functions to yield more specific datasets and identify more specific motifs, can be provided to present specific genomic sequence characteristics and splicing mechanisms.

The RNA secondary structures of a motif and its flanking sequences can be further investigated. A motif predicted from a set of sequence regions demonstrates the conservation of the sites in the primary level. The RNA secondary structures of all the instances of a motif can be predicted. However, an RNA secondary structural comparison method should be developed to asses the structural conservation.

REFERENCES

1. Cartegni, L., Chew, S.L. and Krainer, A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing *Nat Rev Genet*, **3**, 285-298.
2. 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.
3. 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.
4. 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.
 5. 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 Database issue**, D258-261.
 6. 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.
 7. 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.
 8. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure *J Mol Biol*, **288**, 911-940.
 9. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction *Nucleic Acids Res*, **31**, 3406-3415.
 10. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: A Sequence Logo Generator *Genome Res*, **14**, 1188-1190.
 11. Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences *Nucleic Acids Res*, **18**, 6097-6100.
 12. Miriami, E., Margalit, H. and Sperling, R. (2003) Conserved sequence elements associated with exon skipping *Nucleic Acids Res*, **31**, 1974-1983.
 13. 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.
 14. 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.
 15. Bailey, T.L. and Elkan, C. (1994).
 16. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. *et al.* (2004) UniProt: the Universal Protein knowledgebase *Nucleic Acids Res*, **32 Database issue**, D115-119.
 17. 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.
 18. Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000), *J Mol Biol*, Vol. 296, pp. 1205-1214.