Software

ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data

Hsien-Da Huang*, Jorng-Tzong Horng*†, Chau-Chin Lee‡ and Baw-Jhiune Liu‡

Addresses: *Department of Computer Science and Information Engineering, †Department of Life Science, National Central University, Jung-Li City 320, Taiwan. ‡Department of Computer Science and Engineering, Yuan-Ze University, Jung-Li City 320, Taiwan.

Correspondence: Jorng-Tzong Horng. E-mail: horng@db.csie.ncu.edu.tw

Published: 14 March 2003

Genome ***Biology*** 2003, **4**:R29

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/4/R29

Received: 4 September 2002
Revised: 14 November 2002
Accepted: 24 January 2003

Abstract

ProSplicer is a database of putative alternative splicing information derived from the alignment of proteins, mRNA sequences and expressed sequence tags (ESTs) against human genomic DNA sequences. Proteins, mRNA and ESTs provide valuable evidence that can reveal splice variants of genes. The alternative splicing information in the database can help users investigate the alternative splicing and tissue-specific expression of genes.

Rationale

Alternative splicing is a widely occurring and important mechanism for controlling the expression of cellular and viral genes. It changes the effects of a gene in different tissues and developmental states by generating distinct mRNA isoforms composed of different selections of exons, which produce variant proteins. This phenomenon is widespread in the human genome and it was commonly believed that alternative splicing existed in only about 30 to 40% of all genes [1,2].

Because the number of sequences - that is, proteins, mRNAs, and ESTs - in the databases is increasing exponentially, it is possible to decipher alternative splicing forms by computational alignment methods such as BLAST [3] and SIM4 [4]. mRNA and EST sequences, as well as protein sequences that are theoretically translated into nucleotides in six reading frames, provide gene-expression evidence revealing alternative splicing events when aligned against genomic sequences. The alignment tools BLAST and SIM4 can both align mRNA and EST sequences against the reference genomic sequence. However, only SIM4 provides boundary information that can be used to postulate the splicing sites (donor sites and acceptor sites) when considering alternative splicing issues.

In ProSplicer, complete genomic sequences of known and novel human genes are used to investigate alternative splicing variants. ProSplicer provides alternative splicing information for known and novel genes by aligning three major types of expressed gene evidence, that is protein, mRNA and EST sequences. The alternative splicing forms predicted by considering protein, mRNA, and EST sequences together are more complete than just the exons predicted by EST sequences. The tissue-specific expression information provided by mRNA and EST sequences is also helpful in revealing alternative splicing forms favored in different tissues, for example by exon skipping. The database provides keyword search for retrieving and searching the contents and a graphical user interface displays the alternative splice forms. The alternative splice sites predicted by protein, mRNA or EST sequences individually are also given.

Results

As shown in Table 1, ProSplicer uses material on 21,786 genes from ENSEMBL [5], a total of 2,311,460 sequences including protein, mRNA, and EST sequences, to investigate local sequence similarities that can reveal alternative splicing variants. The number of exon candidates generated by alignment tools is shown in Table 1: 442,077, 395,619, and 12,361,685 exon candidates are predicted by aligning protein sequences, mRNA sequences and EST sequences, respectively, against the genomic sequences. ProSplicer also takes mouse protein sequences into account to reveal the cross-species comparison of the alternative splicing variants of a gene. That is, the mouse protein sequences are aligned to human genomic sequences and matching blocks are generated to be the exon candidates.

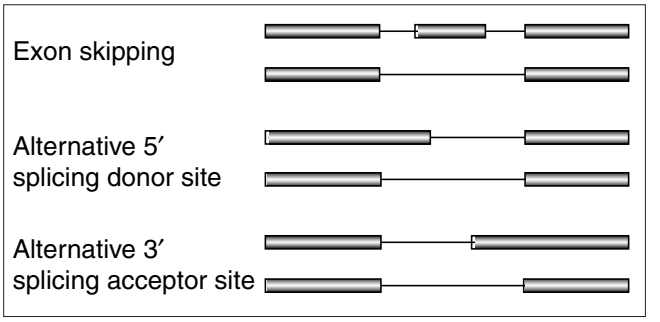

Figure 1
Comparison of pairs of transcripts from the same gene showing three types of alternative splicing events. The shaded bars indicate the exon candidates; the thin lines indicate the intron regions.

Query interfaces

In ProSplicer, all the related evidence sequences, that is, mRNA, EST and protein sequences that are maintained in the database, are pre-aligned to the genomic gene sequences. All alternative splicing variants revealed by the alignment after the filtering phase are also stored. Both text and graphical information are provided in ProSplicer, as well as the query interfaces provided via the web.

By considering the alternative splicing forms of a gene provided in ProSplicer, an exon might be left out or selected after comparison to other protein, mRNA or EST sequences. The three major types of alternative splicing events, including exon skipping, alternative 5′ splicing donor sites, and alternative 3′ splicing acceptor sites [6] are included in the database. Figure 1 shows an example of three types of alternative splicing events. The three types of alternative splicing forms can be shown directly in the graphical user interface in ProSplicer, as shown in Figure 2.

Search tools

ProSplicer provides several keyword search criteria, such as Ensembl gene identification numbers, gene symbols or names, protein id, and UniGene id. Users can submit a gene symbol as keyword and the database returns the query result containing the keyword. All the gene-related information, including supporting evidence, that is mRNA, EST and protein sequences, are also provided in the interface. Protein id and UniGene id can also be submitted by the user and the query result returns the genes that are supported by the query protein sequences or UniGene clusters.

Table 1

Numbers of sequences and exon candidates in ProSplicer

Sequence type	Number of sequences		Number of exon candidates	
Protein	44,184	26,115 (human) 18,069 (mouse)	442,077	279,656 (human) 162,421 (mouse)
mRNA	20,577		395,619	
EST	2,246,699		12,361,685	
Total	2,311,460		13,199,381	

Gene information

ProSplicer provides related reference links to other biological databases and sequences related to the genes selected. The related annotations and reference database links of a gene include Ensembl id numbers, gene symbols, genomic locations and gene descriptions. As shown in Figure 3, the available reference links include GO (Gene Ontology data) [7], HUGO (providing access to the list of currently approved human gene symbols) [8], GeneCard [9] (integrating human genes, their products and their involvement in diseases), LocusLink [10] (organizing information around genes to generate a central hub for accessing gene-specific information), RefSeq [11] (providing reference sequence standards for genomes, transcripts and proteins) and OMIM [12].

Graphical splicing view

The splicing view consists of two parts - 'overview' and 'detailed view'. The overview interface provides a graphical view of the selected gene's location on the chromosome. Figure 4 shows the detailed view in ProSplicer. There are two graphical components in the detailed view. The first is an adjustment bar to scale and move the viewer along the chromosome. The second shows the alignment result of mRNA, EST and protein sequence against the gene genomic sequences to reveal the alternative splicing variants. The graphical interface provides the following functions.

Jumping to specific region. You can jump to a user-specified region of the genomic sequence (see A in Figure 4) where all the related sequences and alignment result are also shown in the detailed view.

Scaling the view. You can scale the view to 1/8, 1/4, 1/2, 2, 4 or 8 times its current window size (see B in Figure 4).

Figure 2
An example ProSplicer analysis showing how the three types of alternative splicing event will be displayed.

Ensembl gene ID	ENSG00000067191	SpliceView
Gene Symbol	CACNB1	
Genomic Location	Chromosome: **17** Chromosome strand:**0** From **39366305** to **39394500**	
Gene Description	DIHYDROPYRIDINE-SENSITIVE L-TYPE, CALCIUM CHANNEL BETA-1M SUBUNIT(BETA-1 ISOFORM C) (BETA-1A).	
UniGene Cluster	Hs.635	
Referenced Database Links	GO	GO:0005624 GO:0006811 GO:0006832 GO:0006936 GO:0005245 GO:0005891 GO:0006816 GO:0005624 GO:0006811 GO:0006832 GO:0006936 GO:0005245 GO:0005891 GO:0006816
	HUGO	CACNB1
	GeneCard	CACNB1
	LocusLink	782
	RefSeq	NM_000723
	OMIM	CACNB1
Supported Sequences	Protein (11 entries)	O15331 CCBA_RAT CCBC_HUMAN CCBB_HUMAN CCBA_HUMAN Q96NZ4 Q96NZ5 Q9C085 Q9EPT9 Q9Y340 Q9Y341
	mRNA (4 entries)	M92303 NM_000726 NM_000724 U07139
	EST (42 entries)	SpliceView
Other Alternative Splicing Datebases	PALS db:	CACNB1
	SpliceNest:	Hs.635

Figure 3
Gene information and links in ProSplicer.

Figure 4
The 'detailed view' graphical interface in ProSplicer.

Moving the view. You can move to the left or right of the current view (see C in Figure 4).

Also shown in Figure 4 is the main graphical view of the alternative splicing view. This comprises basic gene information: gene id (D), gene symbol (E), and gene description (F). The items provided in the splicing view are: the quality of alignment (G), with the degree of similarity between matching blocks, that is, exon candidates, represented by different colors; the length of the selected gene region (H); and sequence identification (I) - each 'Sequence ID' of a nucleotide or protein sequence is hyperlinked to SWISS-PROT, GenBank and dbEST. When you click on an exon candidate (J), a new browsing window opens showing the alignment flat file. The different color fills of the exon blocks refer to the alignment quality as set out in G. When you click on an intron block, a new browsing window showing the alignment flat file opens. The display also includes tissue information, with different tissues represented in different colors, and species information on the source organism of the protein or mRNA sequences.

A comparison of existing alternative splicing databases and tools

Several alternative splicing databases, such as AsMamDB [13], ASDB [14] and SpliceDB [15], are constructed on the basis of genes annotated containing the keywords 'alternative splicing'. AsMamDB contains information about alternative splicing in several mammals. SpliceNest [16], SpliceDB, AsMmDB, and HASDB [17] map clustered ESTs onto human genomic DNA to compute gene structures and splice variants. PALS db [6] takes the longest mRNA sequence in each UniGene [18] cluster as the reference sequence, which is aligned with ESTs and mRNA sequences in the same cluster to predict alternative splicing sites. The BLAT server [19] is a BLAST-like alignment tool that aligns an input nucleotide sequence to human genomic sequences, mRNA, EST and protein sequences. BLAT builds an index of

the database and then scans linearly through the query sequence for local alignments. It then stitches them together into a larger alignment. Finally, BLAT revisits small internal exons possibly missed at the first stage and, where feasible, adjusts large gap boundaries that have canonical splice sites. BLAT is more accurate and 500 times faster than popular existing tools for mRNA/DNA alignments. BLAT is very effective for doing alignments between mRNA and genomic DNA from the same species, and can reveal splicing variants from the alignment result. ProSplicer pre-aligns known and novel gene sequences to the available mRNA, EST and protein sequences. ProSplicer is useful when the user wants to find alternative splicing variants by inputing a gene. We briefly summarize the difference between ProSplicer and BLAT as follows.

First, researchers can input gene names in ProSplicer as opposed to nucleotide sequences in the query stage. Second, the methods of alignment and filtering of sequences are most likely to be very different. We describe our method more fully in the Materials and method section. Third, in ProSplicer, links to various databases and functional information on particular genes (OMIM, RefSeq, GO, HUGO, and so on) are provided.

A comparison of several alternative splicing databases and tools is given in Table 2. The column 'Referenced sequence' indicates genomic sequences, or the longest mRNA sequence in UniGene clusters when used. The 'Types of sequence supported' column shows the materials, including proteins, mRNA or EST sequences, which are used to analyze and then investigate alternative splicing forms of genes. The alignment tool used in each approach is also shown. Whether the alternative splicing criterion for inclusion of genes has been determined through literature search or not is also given in Table 2.

The future of ProSplicer

Novel genes can also be analyzed and integrated into the ProSplicer database. The ProSplicer database does not search by input raw nucleotide sequence because of the limitations of computational power. To handle the database search successfully, the query nucleotide sequences need to be aligned to a BLAST database of nucleotide, EST and protein sequences. We would like to support the query feature in the future by improving the computational power.

Materials and methods

The genomic sequences and gene annotation information are obtained from ENSEMBL [5] (Release 28 May, 2002) and contains 21,786 genes, including known and novel genes. The related mRNA sequences and EST sequences of the genes are retrieved from UniGene [18] (Release 147), and contain 96,105 human gene clusters. The EST sequences are from dbEST (Release 22 February, 2002), and there are about 3,991,208 human EST. The protein sequences are obtained from the SWISS-PROT and TrEMBL (Release 20 March, 2002) [20]. SWISS-PROT (Release 40) contains 101,602 entries in total and 29,751 human entries. From TrEMBL 25,972 rodent entries are taken into account.

Table 2

A comparison of different alternative splicing databases and tools

	Referenced sequence	Types of sequence supported			Statistics	Alignment tool	Literature search	Organisms
		Protein	mRNA	ESTs				
SpliceNest	Genomic sequence		Yes	Yes	90,000 EST clusters	REPuter [23]		Human and *Arabidopsis*
PALS db (Release 2)	Message RNA		Yes	Yes	19,936 (human) 16,615 (mouse) UniGene clusters	BLAST		Human and mouse
SpliceDB	Genomic sequence		Yes	Yes	43,337 splice site pairs	BLAST	Yes	Mammalian
AsMmDB (version 1.0)	Genomic sequence		Yes	Yes	1,563 alternative splicing genes	FASTA	Yes	Human, mouse and rat
ASDB (version 2.1)		Yes			1,922 proteins		Yes	
HASDB	Genomic sequence		Yes	Yes	6,201 UniGene clusters	BLAST		Human
BLAT server	User-submitted sequence	Yes	Yes	Yes		BLAT		Human and mouse
ProSplicer	Genomic sequence	Yes	Yes	Yes	21,786 genes	BLAST (protein) SIM4 (mRNA and ESTs)		Human

The predictive approach to alternative splicing consists of three main phases and the system flow is shown in Figure 5. The three phases are the preprocessing phase, the alignment phase and the filtering phase. In the preprocessing phase, the gene genomic, EST, mRNA and protein sequences, which are stored in different databases, are collected, converted and integrated into a single database, namely GeneInfo. All the sequences are maintained and are prepared for analysis in the alignment phase. Here, the protein sequences are aligned to the gene genomic sequences by TBLASTN [3], and the mRNA and EST sequences are aligned by the alignment tool SIM4 [4]. Exon candidates are generated by both alignment tools. In the filtering phase, we filter the noise of the exon candidates, and connect the exon candidates as reasonable transcript forms of each EST, mRNA and protein sequence by considering the sequential order of matching blocks. Finally, the exon candidates of the alternative splicing forms are provided in the database of ProSplicer.

The preprocessing phase

In this phase, all the sequences and gene information are first collected and maintained. The data files are obtained from dbEST [21], UniGene [18], SWISS-PROT [20] and TrEMBL [20] databases. These files are then parsed and imported into a MySQL database. Ensembl [5] provides database files of MySQL, which we import directly into the database under construction. In the human genome data in Ensembl, each chromosome is represented as assembled fragments of nucleotide sequences. For each known gene and novel gene, Ensembl provides its location and direction in one or multiple chromosome. The genomic sequence of a gene is extracted from the assembled chromosome sequences.

The alignment phase

In ProSplicer, alternative splicing events are revealed by aligning the EST, mRNA and protein sequences against the gene genomic sequences. When aligning the nucleotide sequences, that is, mRNA and EST sequences, both types of nucleotide sequence are viewed as query sequences which are aligned to the gene genomic sequences using the alignment tool SIM4 [4]. As the EST sequences are partial sequences of gene transcripts, the alignment result may show partial matching regions when they are compared to the genomic coding sequences. The mRNA sequences, which in general are cloned and sequenced from mature mRNA, provide more complete alignment results than do EST sequences to reveal the alternative splice forms.

Protein sequences are aligned to genomic sequence in all six possible reading frames by the TBLASTN module of BLAST.

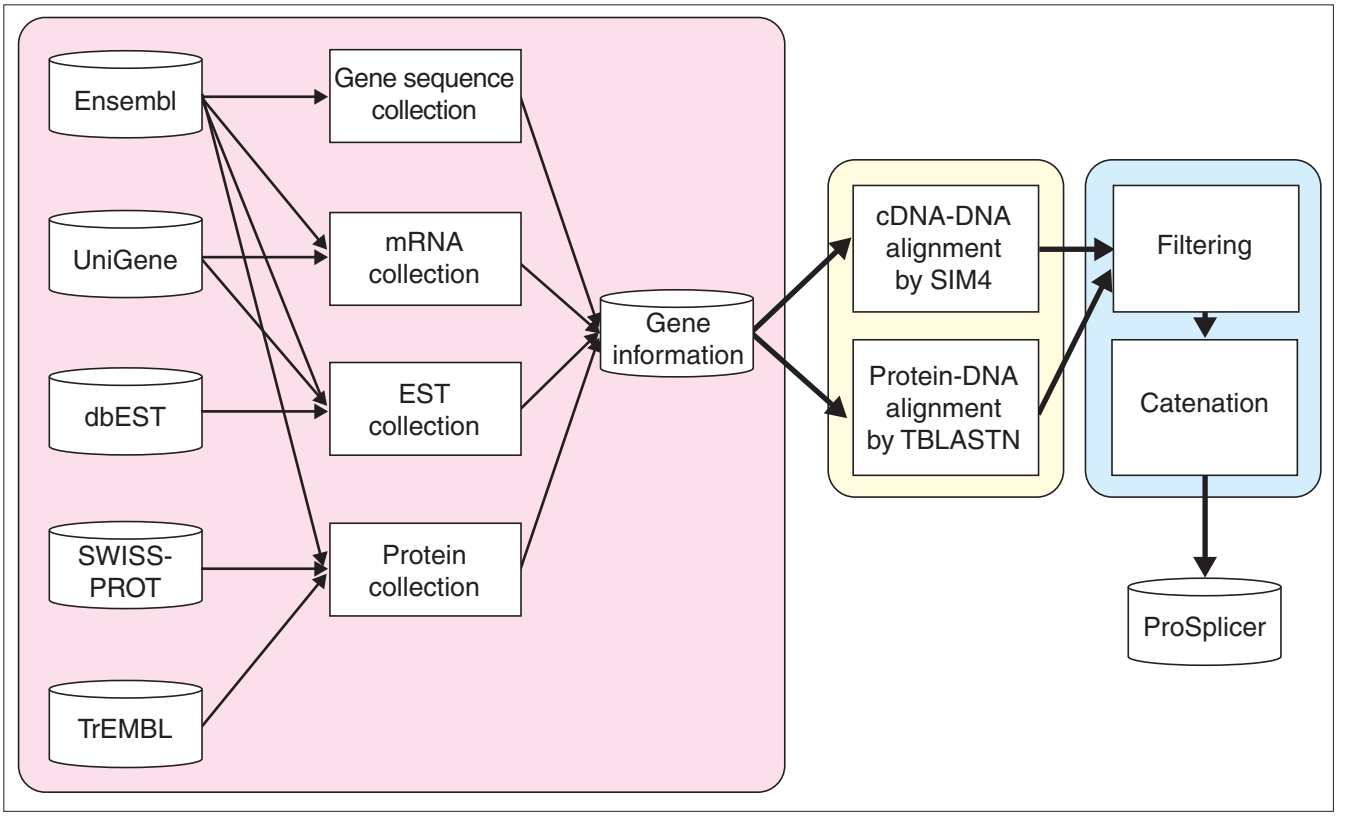

Figure 5
System flow of ProSplicer.

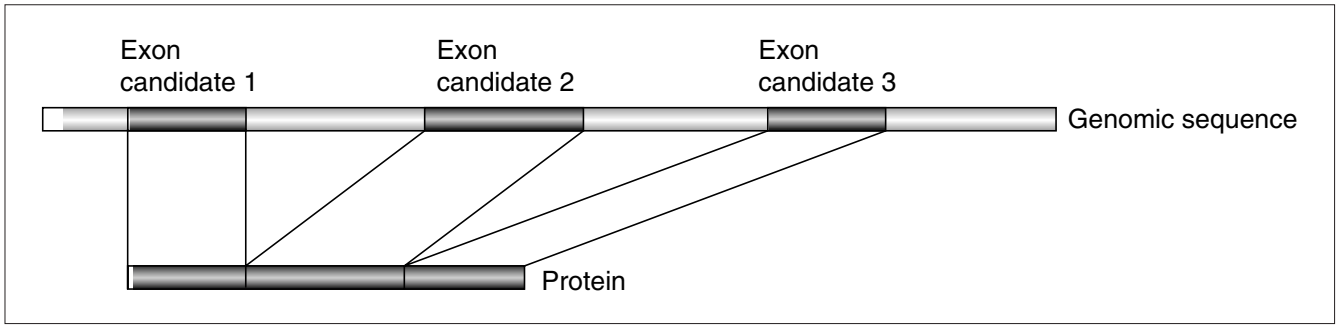

Figure 6
The concept of alignment between protein and genomic sequences to identify candidate exons.

The protein-to-DNA alignment is done with the scoring matrix PAM30 and the parameters are extending gap cost 2, opening gap cost 8, and expectation value limitation 0.001. An example of a protein and genomic sequence alignment is shown in Figure 6.

The filtering phase

Aligning the mRNA, EST and protein sequences to the genomic sequences gives matching blocks which can be considered as exon candidates. However, there is some noise within the alignment result. The major sources of noise are small match blocks, repeat sequences and matching blocks on the complementary strand. The exon candidates generated by aligning the mRNA and EST sequences are eliminated if they are less than 15 base-pairs (bp) long, and those generated by protein comparisons if their length is less than 10 bp. Exon candidates are eliminated if the sequences are matched on the complementary strands of the gene genomic sequences. The remaining exon candidates generated from a single mRNA, EST or protein sequence are assembled into a complete splicing form by considering the sequential order of the matching blocks on the corresponding genomic sequence. To generate the splicing variants for each evidential sequence, we connect those exon candidates to reconstruct a splicing form by reference to the genomic sequence. This is done by comparing the position of the start and end sites of the matching blocks generated by aligning the EST, mRNA and protein sequences. The matching blocks in each evidential sequence are connected as a splicing form.

Availability

ProSplicer is now available at [22].

References

1. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes.** *Genome Res* 1999, **9:**1288-1293.
2. Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.** *FEBS Lett* 2000, **474:**83-86.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
4. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8:**967-974.
5. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, *et al.*: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30:**38-41.
6. Huang YH, Chen YT, Lai JJ, Yang ST, Yang UC: **PALS db: putative alternative splicing database.** *Nucleic Acids Res* 2002, **30:**186-190.
7. The Gene Ontology Consortium: **Creating the Gene Ontology resource: design and implementation.** *Genome Res* 2001, **11:**1425-1433.
8. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain HM: **The HUGO Gene Nomenclature Committee (HGNC). Nomenclature Recommendations.** *Hum Genet* 2001, **109:**678-680.
9. **GeneCards** [http://bioinfo.weizmann.ac.il/cards/index.html]
10. **LocusLink** [http://www.ncbi.nlm.nih.gov/LocusLink/index.html]
11. **RefSeq** [http://www.ncbi.nih.gov/RefSeq/]
12. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30:**52-55.
13. Ji H, Zhou Q, Wen F, Xia H, Lu X, Li Y: **AsMamDB: an alternative splice database of mammals.** *Nucleic Acids Res* 2001, **29:**260-263.
14. Dubchak I, Brudno M, Gelfand MS, Zorn M, Dralyuk I: **ASDB: database of alternatively spliced genes.** *Nucleic Acids Res* 2000, **28:**296-297.
15. Burset M, Seledtsov IA, Solovyev VV: **SpliceDB: database of canonical and non-canonical mammalian splice sites.** *Nucleic Acids Res* 2001, **29:**255-259.
16. Coward E, Haas SA, Vingron M: **SpliceNest: visualization of gene structure and alternative splicing based on EST clusters.** *Trends Genet* 2002, **18:**53-55.
17. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** *Nucleic Acids Res* 2001, **29:**2850-2859.
18. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, *et al.*: **A gene map of the human genome.** *Science* 1996, **274:**540-546.
19. Kent WJ: **BLAT - the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
20. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45-48.
21. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
22. **ProSplicer** [http://bioinfo.csie.ncu.edu.tw/ProSplicer]
23. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R: **REPuter: the manifold applications of repeat analysis on a genomic scale.** *Nucleic Acids Res* 2001, **29:**4633-4642.