

Detection of Novel Splice Forms in Human and Mouse Using Cross-Species Approach

Z. Kan, J. Castle, J. M. Johnson, and N. F. Tsinoremas

Pacific Symposium on Biocomputing 9:42-53(2004)

DETECTION OF NOVEL SPLICE FORMS IN HUMAN AND MOUSE USING CROSS-SPECIES APPROACH

Z. KAN, J. CASTLE, J. M. JOHNSON, N. F. TSINOREMAS

Rosetta Inpharmatics, 12040 115th Ave. N.E.

Kirkland, WA 98034

E-mail: zhengyan_kan@merck.com

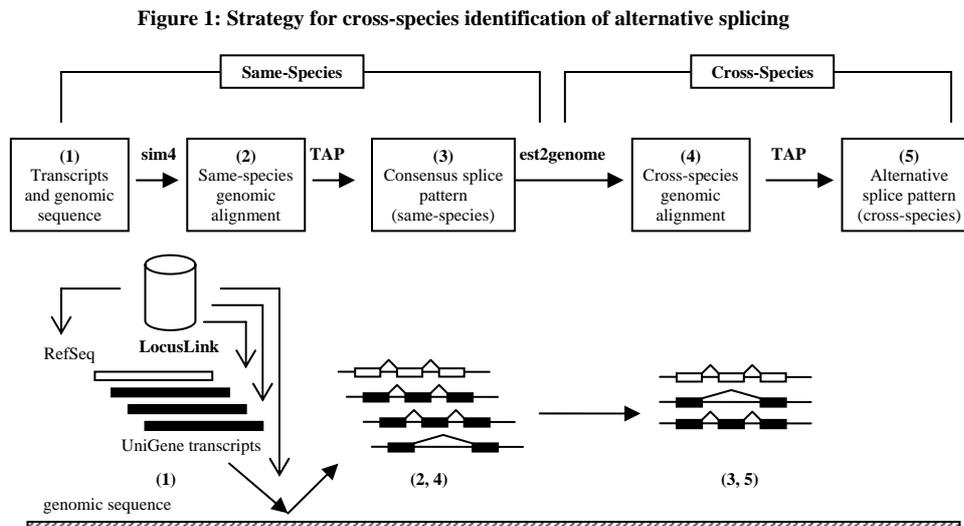
Millions of transcript sequences have become available for characterizing the transcriptome of human and mouse. Transcript databases have been extensively mined for extracting alternative splicing information within the same species; but they also represent a potentially valuable resource for the discovery of alternative splice variants in another species. In this study, we have performed analysis of alternative splicing patterns for 7,475 pairs of human and mouse genes. We found that cross-species transcript analysis could accomplish the same level of sensitivity in detecting constitutive splice patterns as EST resource from the same species. In contrast, identifying alternative splice patterns in human genes, mouse transcripts achieved only 50% of the sensitivity of human EST and 70% of the sensitivity of human mRNA. While identifying alternative splice patterns in mouse genes, human transcripts are 38% more sensitive than mouse mRNA, and reach 60% of the sensitivity of mouse EST. Furthermore, using the cross-species approach, we predicted novel alternative splice patterns for 42% of human genes and 51% of mouse genes. Splice site motif analysis suggests that the majority of predicted novel splice patterns are expressed in human. EST-based frequency analysis shows that novel splice patterns are expressed at lower frequency than alternative splice patterns present in the transcript data from both species, possibly explaining why they remain undetected in the transcript data of the same species.

1 Introduction

Alternative splicing is an important mechanism for regulating gene functions [7] and has been implicated in many human diseases [8]. Genome-wide EST analyses have found evidence of alternative splicing for the majority of human genes [9] and are being used for mining novel splice forms in human genes of therapeutic interest [14]. In addition to human transcript databases, mouse transcripts represent a potentially valuable resource for discovering alternative splice variants of human genes. There currently exist more than 3 million mouse ESTs, and 100,000 mouse mRNAs in the public domain. Novel splice variants of human genes may be predicted by mining the mouse transcript data. In addition, classifying individual human splice variants as conserved across species or as human specific is important for evolutionary analysis and functional investigation of alternate splice forms [5, 10, 12, 15]. However, evolutionary divergence between human and mouse poses a new and considerable challenge to alternative splicing analysis. Cross-species alignment data is noisier than same-species alignment due to divergence at the sequence level that results in a higher error rate in delineating splice patterns. Recent studies also indicate that alternative splicing could be less well-conserved from human to mouse than constitutive splicing, although no clear agreement emerges on how conserved alternative splicing is [5, 10, 12, 15].

This study is focused on detecting and delineating alternative splice patterns using transcript sequences from a different species origin. We employed a bidirectional strategy for the parallel identification of splice variants for 7,475 orthologous pairs of human and mouse genes. A simple method was developed to screen errors in cross-species alignment by requiring splice junction consistency. We found that mouse transcripts could be used to predict 21% of known alternative splice patterns in human genes, and human transcripts could be used to predict 27% of known alternative splice patterns in mouse genes. In addition, potentially novel alternative splicing patterns were identified for 42% of human genes and 51% of mouse genes using the cross-species approach. Splice site motif analysis was introduced to assess the authenticity of a novel splice site. The methods developed in this work are applicable to future cross-species studies of splicing. This study also demonstrates that cross-species analysis would significantly enrich our knowledge of alternative splicing in human genes, and to an even larger extent in mouse genes.

2 Methods



2.1 *Strategy for cross-species identification of alternative splice patterns*

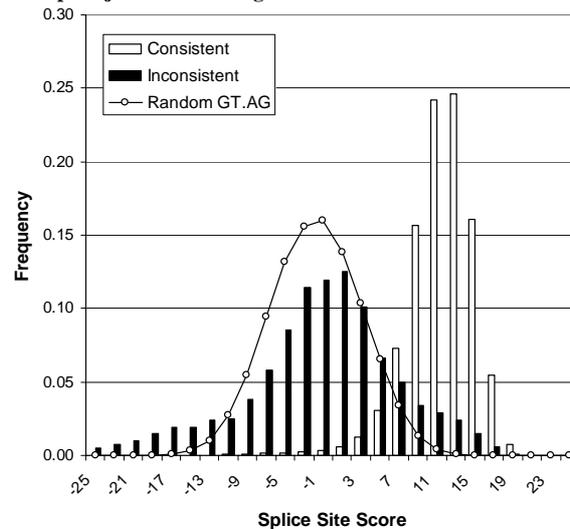
We employed a bidirectional strategy that enables cross-species identification of alternative splice patterns for human and mouse genes in parallel (Fig. 1). In the first phase, a program called TAP [4, 5] identified consensus splice patterns, including both constitutive and alternative patterns, for two genome-wide collections of human and mouse genes based on same-species alignments, obtained by aligning transcript sequences to the genome of the same species. We used both EST and mRNA sequences in GenBank. In the second phase, cross-species alignments are generated by aligning mouse consensus sequences to the human genome and vice versa. Cross-species alignments are then used to identify alternative splice patterns using TAP. The first three steps accomplish the detection of alternative splice patterns in all human and mouse genes using transcripts of the same species origin. (1) For each gene in the LocusLink database [13], we obtain the following data, a RefSeq sequence, sequence of the corresponding genomic region and UniGene cluster [16]. (2) Transcripts in the UniGene cluster, including EST and Genbank mRNA sequences, are aligned to the genomic sequence using *sim4* [3]. (3) Genomic alignments are processed by TAP and clustered into consensus splice patterns, each representing a distinct splice form. Consensus splice patterns that are mutually exclusive to the reference gene structure are identified as alternative splice patterns. The next phase is cross-species analysis. (4) Consensus sequences from one species are aligned to the orthologous genomic templates using *est2genome* [11]. Two genes are “orthologous” if they are reciprocal best matches as annotated in the Homologene database [16]. (5) TAP analysis is performed to identify alternative splice patterns from “cross-species” transcript alignment data. Step (5) is similar to step (3) as cross-species alignment is treated the same as same-species alignment. One minor modification involves reducing the percent identity requirement for screening poor alignment from 92% to 70%. A refinement procedure described below is used to screen errors in cross-species alignment.

2.2 *Splice junction consistency check*

Orthologous human and mouse transcripts exhibit a wide range of sequence homology from 70% to 95% [8]. Due to sequence divergence, cross-species alignment is more error-prone than same-species alignment in term of accuracy for inferring splice patterns. Furthermore, false splice patterns resulting from alignment errors would be mistaken as alternative splice patterns since they are distinct from the reference gene structure. To address this issue, we developed a refinement procedure for examining the consistency of splice junction inference by comparing alignments of the same sequence to different genomes, one from the same species and one from another species. A transcript sequence is aligned to the genome of origin and aligned to the genome from a different species. Each alignment indicates a splice pattern, a series of intron/exon boundaries, on the genome and a set of splice junctions on the transcript sequence. A splice junction from the cross-species

alignment is “consistent” if it is located at the same position as a splice junction from the same-species alignment. If no matching junction can be found for a splice junction, it is classified as “inconsistent”.

Figure 2: Inconsistent splice junctions are alignment errors



Shown here is a clear distinction between consistent and inconsistent splice junctions in term of splice site score, the sum of donor motif score and acceptor motif score. The score distribution of inconsistent splice junctions is similar to that of randomly selected splice junctions containing the canonical GT.AG motifs, indicating that they are artifacts of the alignment program, which only looks for the canonical motifs.

The splice site sequences of all putative splices from cross-species alignments are scored using a weight matrix method taking into account the contexts surrounding the donor and acceptor sites as well as the canonical GT.AG motifs [2, 6]. The donor motif sequence is extracted from an 11-nt window (-2, +8) flanking the donor splice site on the genomic sequence, and the acceptor motif sequence is a 20-nt window (-2, +18) flanking the acceptor splice site. Log odds scores are calculated for individual motif sequences using two weight matrices, one derived from known splice sites and one from background genomic sequences. Figure 2 shows that consistent splices receive much higher scores than inconsistent ones, and the score distribution for inconsistent splices closely resembles that of randomly selected splice sites containing the canonical GT.AG motifs. In addition, inconsistent splices are rarely “reproducible”, also identified using transcript resources of the same species. Less than 8% of inconsistent splices (649/8129) from the mouse-human alignments are reproducible in the human transcripts, whereas 91% (56,678/62246) of consistent splices are reproducible. Based on above evidence, we decide to filter out inconsistent splices. In total, 12% of all splices and

60% of alternative splices from mouse-human alignments that are not reproducible were inconsistent splice junctions. These numbers suggest that alignment error compounded with a lack of sequence conservation could cause a dramatic drop in the accuracy of alternative splicing prediction using the cross-species approach.

2.3 Frequency analysis of splicing event

In the EST-based frequency analysis [5], alternative splice patterns are treated as mutually exclusive outcomes of a stochastic process. The biological frequency of a splicing event, represented by a splice, can be estimated from the frequency of observations in EST sequences. Z-score stands for the likelihood that the biological frequency f of a splicing event is greater than the expected frequency p , set to 10% in this study. The following formula was used to calculate the z-score.

$$z = \frac{\frac{k}{n} - p \pm \frac{0.5}{n}}{\sqrt{\frac{p(1-p)}{n}}}$$

k : the number of ESTs showing a particular splice

$(n - k)$: the number of ESTs showing mutually exclusive splices

The binomial probability $P(f \geq p | n, k)$ that an outcome occurs k or more times in n trials with an expected frequency of p is calculated. If $n * p < 5$, Poisson approximation to the binomial probability is used when $n \geq 200$; the exact binomial probability is calculated when $n < 200$. Probability is converted to z-score using the standard error function.

2.4 Sequence data resources

The December 2002 version of the LocusLink database was used for linking genes, RefSeq sequences, genomic contig mapping and UniGene clusters. For each gene, a single RefSeq sequence is used as the reference sequence. For 95% of loci in LocusLink, there was only a single RefSeq recorded for a gene. If one locus is linked to multiple RefSeq sequences, the RefSeq with the earliest accession number is chosen. Gene loci without a RefSeq sequence are not included in the study. Each gene is linked to a UniGene cluster consisting of both EST and GenBank mRNA sequences. Human and mouse transcript sequences are derived from UniGene build 154. Genomic sequences were retrieved from NCBI contig databases [16, 17] updated as of December 2002. A genomic template sequence for each gene is extracted with five kbs of extension at both ends according to genomic contig locations specified in LocusLink. Orthologous pairing between human and mouse genes require a reciprocal best match relationship according to annotation in the Homologene database.

3 Results

In this study, we used a dataset of 7,454 orthologous pairs of human and mouse genes based on annotations in Homologene [16]. Alternative splice patterns are defined based on mutually exclusive relationship with the reference gene structure of the RefSeq sequence chosen to represent a gene. Each gene has four resources of splice pattern information: EST, mRNA from the same species, EST and mRNA from the other species. Splice patterns identified from different resources are characterized and compared on the basis of individual splice. A "splice" refers to a pair of donor/acceptor splice sites flanking a putative intron on the genomic sequence. Splices are classified under several categories. One category is the source of inference, such as human to human EST alignment, or mouse to human mRNA alignment. Another category defines the alternative splicing relationship. A splice is labeled as "RefSeq" if it is found in the RefSeq gene structure or "alternative" if it is mutually exclusive to a RefSeq splice. RefSeq splices are likely to be constitutive splice patterns although it is not necessarily true in a minority of cases.

3.1 *Detection of Known Constitutive and Alternative Splice Patterns*

Cross-species transcript alignment data and same-species data were compared at the level of individual splice. For each splice present in the RefSeq gene structures, we examine if the exact same splice is present in the splice patterns derived from different transcript resources. We found that human ESTs can identify 84% of RefSeq splices, whereas mouse EST and mRNA combined can identify 82% of them, indicating that mouse transcripts are very informative about constitutive splicing in human genes. A similar trend is observed for detecting constitutive splicing of mouse genes using human transcripts (Table 1A).

We sought to determine how well the cross-species approach predicts splice variants. A test set consisting of 8,786 known alternative splices in human genes was derived from human mRNA. As shown in Table 1B, when compared with splice patterns from human EST data, only 40% of the known alternative splices could be identified. Lower sensitivity for detecting alternative splices indicates the difficulty of "capturing" splice variants that are often expressed at low levels or under specific conditions. Mouse transcripts, including both mRNA and EST sequences, identified 21% of the known alternative splices, more than 50% of the sensitivity of human EST. Human transcripts could identify 27% of known alternative splice patterns in mouse genes, about 60% of the detection power of mouse EST. Greater sensitivity is expected for human to mouse alignment because of greater sequence coverage. It is also worth noting that mRNA seems to be equally powerful as EST for detecting alternative splicing across species (Table 1B).

Table 1: Alternative Splicing Statistics

Species	RefSeq Splices, Total	Alignment Evidence	Splices Identified	Sensitivity
Human	65,925 by human RefSeq	Human EST	55,660	84%
		Mouse EST/mRNA	53,806	82%
		Mouse EST	45,349	69%
		Mouse mRNA	51,822	79%
Mouse	64,608 by mouse RefSeq	Mouse EST	52,854	82%
		Human EST/mRNA	53,153	82%
		Human EST	45,836	71%
		Human mRNA	48,023	74%

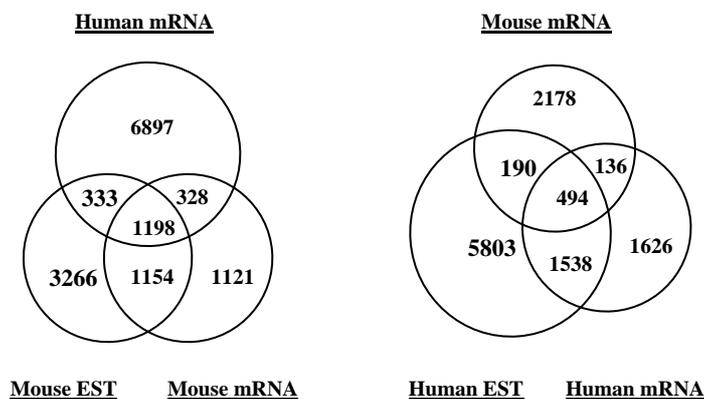
(A)

Species	Alternative Splices, Total	Alignment Evidence	Splices Identified	Sensitivity
Human	8,786 by human mRNA	Human EST	3,517	40%
		Mouse EST/mRNA	1,871	21%
		Mouse EST	1,540	18%
		Mouse mRNA	1,535	18%
	21,099 by human EST	Human mRNA	3,495	17%
		Mouse EST/mRNA	2,449	12%
		Mouse EST	2,149	10%
		Mouse mRNA	1,702	8%
Mouse	2,998 by mouse mRNA	Mouse EST	1,353	45%
		Human EST/mRNA	820	27%
		Human EST	684	23%
		Human mRNA	630	21%
	10,162 by mouse EST	Mouse mRNA	1,353	13%
		Human EST/mRNA	1,784	18%
		Human EST	1,557	15%
		Human mRNA	1,094	11%

(B)

Table (A) shows that cross-species analysis can detect constitutive splice patterns almost as effectively as same-species analysis. “Species” indicates the species origin of the gene under consideration. “RefSeq Splices” include all splices from 7,475 RefSeq gene structures “Alignment Evidence” refers to the type of transcript sequence data that is used for identifying RefSeq and alternative splice patterns. A splice in one data resource is “identified” if both splice sites are exactly matched with a splice inferred using a different resource. “Sensitivity” stands for the fraction of the total splices that are identified using one type of alignment evidence. Table (B) compares the detection power for identifying alternative splice patterns between cross-species resource and same-species resource. Known alternative splices are taken from same-species mRNA alignments or from same-species EST alignments.

Figure 3: Cross-species identification of known alternative splices



Known alternative splices are represented by alternative splices identified from same-species mRNA alignment data. Shown on the left is a Venn diagram showing the overlaps between known alternative splices in human and alternative splices identified using two types of cross-species alignment evidence, EST and mRNA. Shown on the right is the same type of Venn diagram for mouse.

3.2 Characterization of Novel Alternative Splice Patterns

Figure 3 shows that the majority of alternative splices predicted through cross-species analysis are novel, meaning that no match can be found among the transcript alignments in the same species. These splices are all junction consistent, matching a splice junction in the same-species alignment. Mouse EST and mRNA predicted novel alternative splices for 42% (3157) of human genes, whereas human transcripts predicted novel splices for 57% (4250) of mouse genes (Table 2).

Predicted novel splice patterns are further characterized by frequency analysis. Based on the frequency of observing a particular splice in mouse EST sequences, a z-score is calculated for each splice, representing the chance that the real frequency of a splice pattern is greater than the expected frequency, set to 10% in this study. The greater the z-score, the more likely that the splice variant giving rise to the said splice pattern accounts for more than 10% of all splice variants originating from the same gene [5]. Interestingly, novel alternative splices exhibit a clear separation from alternative splices that are reproduced in same-species analysis (Fig. 4). This observation points to low frequency, whether due to low expression level or rare expression pattern, as one contributing factor to the absence of these splice patterns in the human transcript data. Human and mouse transcripts can be thought of as two repeat samples of a set of splice patterns. While high-frequency patterns are likely to have been detected using human transcripts alone, rare patterns may be identified in one sample but are missing from another sample. Even though the coverage of

human transcripts appears to be more comprehensive than mouse transcripts, cross-species transcript analysis can still uncover novel splice patterns.

A novel splice pattern derived from a mouse transcript is not necessarily expressed by human genes. Nonetheless, sequence motifs associated with putative splice sites delineated by cross-species alignment are from the human genome. Each sequence motif could be evaluated for the likelihood that it is a “real” splice site rather than randomly selected using the splice site motif score. Within the set of novel splice patterns, we further identified novel splice sites, required to be at least 10 bases apart from any known splice sites. From mouse-human data, we found 1,135 novel donor sites, and 60% (676) of them receive motif score > 3 . There are 1,420 novel acceptor sites and 53% (759) receive motif score > 2 . These score cutoffs are selected to maximally discriminate real splice sites from the background, randomly selected sequences containing the GT.AG motif. This is a strong indication that many predicted novel alternative splice patterns are likely to be real as they correspond to biological motifs. Work is currently underway to validate these predictions using RT-PCR.

4 Discussion

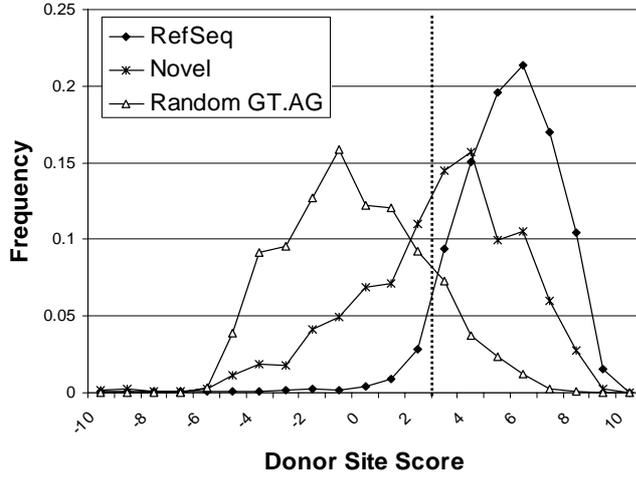
A bidirectional strategy that precedes cross-species analysis with same-species analysis is used to identify alternative splice patterns for both human and mouse genes (Fig. 1). This strategy helps to resolve several problems in transcript-based alternative splicing analysis in the context of cross-species analysis. (1) Artifacts. EST sequences are single sequencing reads often poor in quality and sometimes derived from chimeric cDNA clones. (2) Paralogs. Sequences of closely related paralogous genes are hard to differentiate from each other. (3) Redundancy. In the transcript database, many sequences, EST in particular, exhibit the same splice patterns and are therefore redundant for the purpose of discovering splice variants.

Table 2: Novel alternative splice patterns

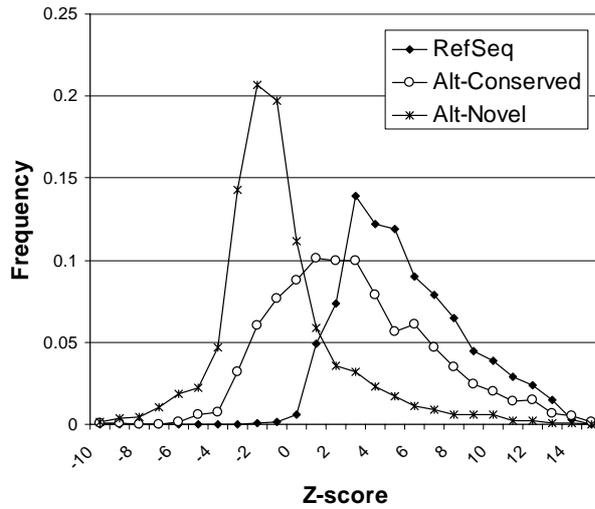
Alignment	Resource	Splices	Genes
Mouse Transcript to Human Genome	EST	4,420	2,677
	mRNA	2,275	1,508
	EST or mRNA	5,568	3,157
	EST and mRNA	1,154	863
Human Transcript to Mouse Genome	EST	7,341	3,403
	mRNA	3,164	1,756
	EST or mRNA	9,060	3,831
	EST and mRNA	1,538	1,022

“Splices” refers to the number of novel alternative splices predicted by cross-species analysis. “EST or mRNA” is the union of two resources, and “EST and mRNA” is the intersection.

Figure 4: Characterization of novel alternative splices



(A)



(B)

(A) The majority of novel donor splice sites are likely to be real, as indicated by the clear separation of score distribution from the randomly selected sequences containing GT.AG motif. The cutoff score of 3 (dashed line) is selected based on the maximum separation between the random set and RefSeq donor sites. (B) Novel alternative splices are expressed at lower frequencies than alternative splices reproducible in human transcripts. Z-scores based on frequency information in mouse ESTs (see methods) are calculated for three classes of splices derived from the Mouse-Human EST alignments. "RefSeq" splices are found in the human RefSeq gene structures. "Alt-Conserved" stands for alternative splices that are also identified in either human mRNA or EST sequences. "ALT-Novel" stands for junction-consistent splices not identified in any human transcript.

While directly aligning mouse sequence to the human genome, it is difficult to tell if a poor alignment is due to evolutionary divergence or other issues such as artifacts or paralogs. By filtering transcript sequences that are not aligned to the genome with near perfect identity in the same-species phase, we can effectively eliminate poor quality sequences, chimeric clones and paralogs. In addition, TAP analysis in the first phase clusters redundant transcript sequences into consensus splice patterns. This procedure substantially reduces the computational cost of performing cross-species alignment, which is often the bottleneck in data analysis on a genomic scale. For example, there are 780,797 human ESTs mapped to 7,454 genes. Only 46,944 consensus splice patterns were aligned to the mouse genome, resulting in a 17-fold reduction in computational cost.

In this study, we have performed genome-wide alternative splicing analysis for both mouse and human. We characterized the transcript resources in term of detecting known patterns of constitutive and alternative splicing across species. Furthermore, we have predicted novel splice forms for 42% of human genes and 51% of mouse genes through cross-species analysis. Work is underway to experimentally validate these predictions. While bioinformatics analysis has predicted many splice variants in human genes, the vast majority of which are poorly characterized, conserved splice variants may constitute an important subset as they remained unchanged across 75 million years of evolutionary drift. Having the mouse counterparts also offers many opportunities for comparative studies that would help elucidate the function and regulation of alternative splicing in the mammalian system.

5 Acknowledgements

We sincerely thank Chen Ronghua for his help with databases.

References

1. Caceres, J. F., and Kornblihtt, A. R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet* *18*, 186-193.
2. Clark, F., and Thanaraj, T. A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet* *11*, 451-464.
3. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* *8*, 967-974.

4. Kan, Z., Eric, R., Warren, G., and David, S. (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res* *11*, 889-900.
5. Kan, Z., States, D., and Gish, W. (2002). Selecting for functional alternative splices. *Genome Res* *12*, 1837-1845.
6. Lim, L. P., and Burge, C. B. (2001). A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* *98*, 11193-11198.
7. Lopez, A. J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet* *32*, 279-305.
8. Makalowski, W., and Boguski, M. S. (1998). Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci U S A* *95*, 9407-9412.
9. Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. *Nat Genet* *30*, 13-19.
10. Modrek, B., and Lee, C. J. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet* *34*, 177-180.
11. Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci* *13*, 477-478.
12. Nurtdinov, R. N., Artamonova, I., Mironov, A. A., and Gelfand, M. S. (2003). Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum Mol Genet* *12*, 1313-1320.
13. Pruitt, K. D., and Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* *29*, 137-140.
14. Sorek, R., and Amitai, M. (2001). Piecing together the significance of splicing. *Nat Biotechnol* *19*, 196.
15. Thanaraj, T. A., Clark, F., and Muilu, J. (2003). Conservation of human alternative splice events in mouse. *Nucleic Acids Res* *31*, 2544-2552.
16. Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* *420*, 520-562.
17. Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A., and Wagner, L. (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Res* *31*, 28-33.