Genome wide identification and classification of alternative splicing based on EST data

S. Gupta1,∗, D. Zink2, B. Korn2, M. Vingron1 and S. A. Haas1

1Computational Molecular Biology Max Planck Institute for Molecular Genetics Ihnestrasse 73, D-14195 Berlin, Germany and 2German Resource Center for Genome Research INF 580, 69120 Heidelberg, Germany

Received on October 10, 2003; revised on March 16, 2004; accepted on April 20, 2004
Advance Access publication April 29, 2004

ABSTRACT

Motivation: Alternative splicing is currently seen to explain the vast disparity between the number of predicted genes in the human genome and the highly diverse proteome. The mapping of expressed sequences tag (EST) consensus sequences derived from the GeneNest database onto the genome provides an efficient way of predicting exon–intron boundaries, gene structure and alternative splicing events. However, the alternative splicing events are obscured by a large number of putatively artificial exon boundaries arising due to genomic contamination or alignment errors. The current work describes a methodology to associate quality values to the predicted exon–intron boundaries. High quality exon–intron boundaries are used to predict constitutive and alternative splicing ranked by confidence values, aiming to facilitate large-scale analysis of alternative splicing and splicing in general.

Results: Applying the current methodology, constitutive splicing is observed in 33 270 EST clusters, out of which 45% are alternatively spliced. The classification derived from the computed confidence values for 17 of these splice events frequently correlate (15/17) with RT–PCR experiments performed for 40 different tissue samples. As an application of the confidence measure, an evaluation of distribution of alternative splicing revealed that majority of variants correspond to the coding regions of the genes. However, still a significant fraction maps to non-coding regions, thereby indicating a functional relevance of alternative splicing in untranslated regions.

Availability: The predicted alternative splice variants are visualized in the SpliceNest database at http://splicenest.molgen.mpg.de

Contact: gupta@molgen.mpg.de

Supplementary information: For supplementary data, please refer Bioinformatics Online.

1 INTRODUCTION

The concept of alternative splicing was discovered with the identification of mosaic structure of genes in adenovirus (Sambrook, 1977). The terms ‘intron’ and ‘exon’ were coined in 1978, along with the formulation of a hypothesis for the phenomenon of alternative splicing (Gilbert, 1978).

For a long time, alternative splicing was thought to be an exception. However, over the last few years, it became apparent that alternative splicing is more the rule than the exception (Mironov et al., 1999; Kan et al., 2001). Alternative splicing contributes significantly to the wealth of gene products. One of the genes, which has been widely examined and known to produce extraordinary molecular diversity, is the Drosophila homologue of human Down syndrome cell adhesion molecule (Dscam). Alternative splicing in this gene can potentially generate more than 38 000 Dscam isoforms, which may contribute to the specificity of neuronal connectivity (Schmucker et al., 2000).

Despite the enormous potential to generate innumerable splice isoforms, only a small fraction of the variants have been observed in nature. This implies a complicated regulation mechanism that guides the expression of certain splice isoforms instead of a random combinatorial output (Smith and Valcarcel, 2000). Several attempts have been made to reveal the mechanism of regulation of alternative splicing (Fury et al., 2002) leading to the identification of conserved sequences, which are associated with exon skipping (Miriami et al., 2003). Both computational and experimental studies are underway to understand the mechanism and regulation of alternative splicing (Muro et al., 2003; Munch et al., 2003).

With such a complicated machinery to regulate splicing, it is evident that any failure or error is likely to result in disorders (Cooper and Mattox, 1997; Caceras and Kornbluth, 2002). It has been estimated that splicing defects account for ~15% of disease-causing mutations in humans, most of which are often attributed to point mutations in the 5′ and 3′ splice sites (Krawczak et al., 1992). However, some diseases (e.g. Alzheimer’s disease) are attributed to non-optimal regulation of splicing (Selkoe, 2001). Unusual splice isoforms were also described in various cancers and tumours (Hsieh et al., 2003; Parareda et al., 2003). Being a bottleneck for several medical
applications, deciphering the mechanism and regulation of constitute and alternative splicing becomes critical.

Currently, the large-scale identification and analysis of splice events is mainly based on the evaluation of the alignment of expressed sequence tag (EST) sequences related to a specific gene with the appropriate genomic sequence (Kan et al., 2002; Modrek and Lee, 2002). In contrast to the mapping of raw EST data, we base our strategy on EST consensus sequences derived from GeneNest (Haas et al., 2000). These consensus sequences reconstruct parts of transcripts containing a lower number of sequencing errors than the original EST sequences. Therefore, when different consensus sequences are mapped to the genome, in order to reveal splice junctions (SpliceNest: Coward et al., 2002), the consensus sequences produce fewer splicing artefacts. Nevertheless, the resulting data will still contain wrongly predicted isoforms, because of genomic contamination for example. Therefore, this emphasizes the need for a strategy to detect real alternative splice events.

Consequently, we describe a methodology to identify and rank potential alternative splicing events by computing confidence values for the splice events. Such confidence-based prediction of constitutive and alternative splice sites ameliorate the SpliceNest database. The computed confidence values are evaluated experimentally based on 17 arbitrarily selected alternative splice events on chromosomes 21, 22 and X, predicted using an earlier version of SpliceNest without the quality computation module. As an application, the predicted splice variants are applied to analyse the frequency of alternative splicing in coding and non-coding regions of known genes.

2 METHODS

2.1 Quality computation for exon–intron boundaries and exons

Our quality computation methodology is based on the GeneNest consensus sequences (Human UniGene Build 161) and the subsequent mapping of these consensus sequences to the genome (SpliceNest, April 2003 freeze of human genome). The parameters used for the computation of the quality of exon–intron boundaries are grouped into four main categories.

Alignment quality: Alignment quality is computed based on the sim4 (Florea et al., 1998) alignment score, reflecting local similarity, as well as on the number of ESTs covering the relevant region.

Signal information: The presence of a splice signal is defined either by the presence of the consensus (GT–AG) splice signal or by the presence of a non-consensus splice signal with perfect alignment in vicinity (±10 bases) of the splice junction. The latter case refers to the observation that while non-consensus splicing events are a known feature of splicing (Clark and Thanaraj, 2002), these are also often caused by misalignments.

Common boundaries: Common boundaries observed among related exons of different transcripts provides further evidence for a splice junction, and is therefore encoded as another parameter for the computation of quality values of splice junctions.

Neighbourhood parameters: An overall splicing direction of all transcripts of the gene is computed and an inconsistency in the direction of a splice signal is interpreted as the absence of splice signals. This information is grouped along with the alignment score and splice signal information of the adjacent exons to determine the neighbourhood quality.

All these parameters are heuristically weighted based on manual inspection of the behaviour of certain randomly chosen test clusters, describing various splicing patterns and intuitive data artefacts arising either due to clustering errors (like antisense transcripts clustered together in Hs177559) or due to increased variation observed in terminal exons (as in Hs17118, Fig. 3). Since splice signals are well-studied markers for exon junctions (Thanaraj and Clark, 2001; Weir and Rice, 2004), the parameter encoding splice site information is weighted 50% while the remaining three parameters share equal weights (Fig. 1). This relative weighting of different parameters is implemented using the fuzzy logic (Zadeh, 1965) based system (TILLShell Pro, Togai Infralogic Ltd), which encodes the relative weightage of the parameters using a set of defined rules and outputs the quality of the relevant exon–intron boundary. The rules are defined such that to achieve the defined threshold of 55%, either signal information must be optimal or at least two other parameters must contribute positively to the quality value. Subsequently, the quality of relevant exon–intron boundaries are used to compute confidence values for all observed alternative splicing events.

2.2 Experimental validation

A set of 17 putative alternative splice events on chromosomes 21, 22 and X was arbitrarily selected using a previous version of SpliceNest data not employing the quality computation methodology. For these splice isoforms, PCR primers were generated on either side of alternative splice events using the primer design software GenomePRIDE (Haas et al., 2003). The computed primers were then used for RT–PCR experiments on 40 different tissue samples (see supplementary data for a list of tissues). These primers were subsequently mapped to the updated version of SpliceNest (which includes the quality computation module), thereby formulating a basis for validation of the methodology.

2.3 Classification of alternative splicing

Based on the pattern of exon boundaries described by alternative transcripts, the splice variants are classified into four types (Fig. 2).

Skipped exons: This type of alternative splicing includes those events in which one of the transcripts contained an exon
Quality-based prediction of alternative splicing

Fig. 1. The flow chart of confidence computation.

Fig. 2. Classification of alternative splicing. (1) Skipped exon event in which one of the transcripts has an additional exon. (2) An intron in a transcript is the part of an exon in another transcript. (3) Alternative donor or acceptor sites. (4) An intron in a transcript is the part of an exon in another transcript.

that is skipped in another transcript. This class contains cas-sette exons and mutually exclusive exons as described recently (Roberts and Smith, 2002).

Multiple skipped exons: A multiple skipped exon event refers to an event in which several skipped exons are observed in a row.

Alternative donor/acceptor sites: This category includes instances of alternative splicing in which either the 5' or the 3' splice site is different between the related exons of different transcripts. These events include competing 5'/3' sites and multiple promoter/polyA (Roberts and Smith, 2002).

Retained introns: An event in which an exon in one of the transcripts connects two adjacent exons in another transcript is called a retained intron event.

2.4 Alternative splicing in coding/non-coding regions

To evaluate the distribution of alternative splicing events in the coding and non-coding regions of the transcripts, a BLAST search against the SWISSPROT protein database was performed for all clusters that are predicted to be alternatively spliced. The following criteria were applied to create a dataset suitable for analysing the distribution of alternative splicing in known genes.

(1) Consider only the transcripts of genes for which a gene symbol is annotated in the Unigene data.

(2) Alternative splicing prediction quality > 70%, which enriches for transcripts with alternative regions covered by at least three ESTs and exon boundaries characterized by reliable splice signals.

(3) Blast similarity score > 100.

(4) Percentage identity > 95%.

(5) Consider only skipped exon events.

The alternative exons were mapped to the 5'-untranslated region (5'-UTR), 3'-UTR and coding regions of the genes.

3 RESULTS

Figure 3 shows a schematic alignment of EST consensus sequences with the genome sequence (SpliceNest). Alternative regions are highlighted (i.e. quality > 55%) if
Fig. 3. Visualization of alternative splicing in SpliceNest (Hs17118). The uppermost and lowermost horizontal bars represent the genomic sequence, which is greyed out in the regions representing repeats. The alignments in the middle represent different contigs/transcripts of a cluster/gene. These contigs split up into exons (thick bars) and introns (lines) with the presence or absence of both splice signals GT–AG (arrows on the intron line). A vertical line at the end of a bar represents the end of a consensus sequence. The colour of the exon represents percentage alignment with the genomic sequence (as defined by the top colour bar). The green vertical bars label significant differences as alternative splicing, while the potential data artefacts like genomic contamination towards the ends of consensus sequences (e.g. the leftmost exon of contig Hs17118.9) are not labeled.

some of the contributing exon boundaries are shared by other transcripts with at least two ESTs covering the alternative region, and if the splice signals exist for alternatively spliced exons. Potential alternative regions below the quality cut-off are not interpreted as splice isoforms. Applying this strategy for analysing human genes represented by GeneNest EST clusters (total 108 094 clusters), constitutive splicing is observed in about one-third of these genes (33 270), out of which 45% are alternatively spliced. The remaining two-thirds of the clusters represent singleton clusters, clusters that are not mapped to the genome and clusters that do not contain any reliable splice site some of which might reflect single exon genes like the human melanocortin 4-receptor gene (Brocke et al., 2002).

Out of the 17 alternative splice events analysed by RT–PCR (Table 1), 11 were correctly attributed as true variants. In addition, four variants were correctly assigned confidence values below the threshold. However, the remaining two predicted splice variants, with confidence values above the threshold, were not observed in the experiments.

The frequency of alternative splicing events grouped by different types revealed skipped-exon (58%) and alternative donor/acceptor (26%) as the most common types of alternative splicing. In contrast, multiple skipped-exon events are observed in 11% of the cases while retained introns are observed in a small percentage of variants (5%).

The comparison of predicted genes with the proteins in the SWISSPROT database mapped 7974 spliced genes to the protein sequences, out of which 2926 exons corresponding to 1730 genes are alternatively spliced. Among these exons, a large fraction (62%) map to the coding regions of the genes. Nevertheless, 14% of the exons map to the 5’ UTR while 7% map to the 3’ UTR. The remaining 17% of the exons overlap the translated and the un-translated parts of genes, putatively representing alternative start and alternative stop of translation.

4 DISCUSSION

The knowledge of alternative isoforms is the basis for understanding regulation of alternative splicing and its role in tumours and diseases. The EST data provide an opportunity to detect alternative splicing, but with the inherent problem of error-prone sequences leading to a high rate of false-positively predicted variants. However, our computation of quality values forms the basis to define a confidence measure for the predicted splice variants and to differentiate them from the misaligned transcripts. Such a confidence measure provides flexibility in selecting a set of splice variants for experimental or theoretical analysis. However, the variants with confidence values below the threshold might include some of the atypically expressed variants (e.g. variants with non-consensus splice signals, lowly expressed variants), which lack the redundant sequence information needed to remedy the otherwise low sequence quality of ESTs.

The predicted splice variants are classified into different types of alternative splicing events. The splicing events, most reliably recognized, are skipped exons since these events
are nicely defined by splice signals on both sides of the alternative exon. A subset of the skipped exon events, the multiple skipped exons are usually covered by only few ESTs and are therefore more likely to reflect leakage of the splicing machinery or data artefacts. Alternatively, this unusual splicing behavior might represent splice variants in a different biological context like in tumours (Wang et al., 2003). In contrast to the skipped exons, retained introns are represented by a single splice signal, thereby reducing the confidence in these boundaries. Additionally, the frequent genomic contamination in some cDNA libraries (Sorek and Safer, 2003), will produce a splicing pattern resembling a retained intron. Consequently, the retained intron events are the least reliable type of alternative splicing. Nevertheless, retained introns with high EST coverage and optimal alignment are still detected although with a lower confidence than skipped exon events with similar alignment quality.

The predicted splice isoforms were subsequently applied to study the distribution of alternative splicing events in coding and non-coding regions of known genes. The coding regions contain the largest fraction (62%) of alternative variants, which are likely to result in functionally divergent proteins (Yan et al., 1996). This large fraction of alternative splicing may partially reflect the current experimental focus on the analysis of protein sequences, thereby increasing the chance of detecting splice variants in coding regions, whereas splice variant detection in the non-coding region is mainly based on ESTs. Furthermore, due to the experimental protocol of EST generation the 3′-UTR is usually covered by a larger number of ESTs than regions further upstream, thus increasing the chance of detecting splice variants. Therefore, in 5′-UTR the EST coverage is often much lower, thereby complicating the detection of splice variants in this region. Nevertheless, the fraction of alternative splice variants in the 5′-UTR is two times higher than in the 3′-UTR implying a higher functional importance of splicing in the 5′-UTR as compared with 3′-UTR. Assuming that alternative splice variants in the non-coding regions of the gene are not just related to nonsense mediated mRNA decay (NMD, Lewis et al., 2003), these variants could, for instance, represent translational control mechanisms as summarized by Wilkie et al., (2003) and Kuersten and Goodwin (2003). In case of 5′ UTR, alternative splicing may affect exons carrying an upstream ORF that influences the expression of the downstream ORF (Jin et al., 2003). Additionally, the 5′-UTR variants along with a fraction of variants overlapping 5′-UTR and coding region may represent alternative promoters (Itani et al., 2003; Delaloy et al., 2003; Brown et al., 1999). In summary, the confidence measure for constitutive and alternative splicing provides a basis for large-scale experimental and theoretical analysis of various phenomena related to splicing. As an example, we showed that alternative splicing in a non-coding region indicates a possible role of splicing in post-transcriptional regulation, thereby illustrating an additional theme for analysis of gene regulation. We plan to adapt our tools to tissue specific and disease-related splicing which may provide more insight into the regulation of splicing, implying several diagnostic and therapeutic applications.

### Table 1. Validated alternative splice events

<table>
<thead>
<tr>
<th>Old cluster</th>
<th>Version</th>
<th>New cluster</th>
<th>Chromosome</th>
<th>Prediction</th>
<th>RT–PCR</th>
<th>Validation (quality)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs8811</td>
<td>Sep. ’99</td>
<td>Hs10267</td>
<td>22</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (88%)</td>
</tr>
<tr>
<td>Hs75527</td>
<td>Sep. ’99</td>
<td>Hs75527</td>
<td>22</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (86%)</td>
</tr>
<tr>
<td>Hs110447</td>
<td>Sep. ’99</td>
<td>Hs92260</td>
<td>22</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (89%)</td>
</tr>
<tr>
<td>Hs155156</td>
<td>Sep. ’99</td>
<td>Hs351478</td>
<td>22</td>
<td>No</td>
<td>One variant verified</td>
<td>True negative</td>
</tr>
<tr>
<td>Hs129829</td>
<td>Sep. ’99</td>
<td>Hs129829</td>
<td>21</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (86%)</td>
</tr>
<tr>
<td>Hs181188</td>
<td>Sep. ’99</td>
<td>Hs235887</td>
<td>21</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (91%)</td>
</tr>
<tr>
<td>Hs181581</td>
<td>Sep. ’99</td>
<td>Hs181581</td>
<td>21</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (84%)</td>
</tr>
<tr>
<td>Hs184045</td>
<td>Sep. ’99</td>
<td>Hs198308</td>
<td>21</td>
<td>No</td>
<td>One variant verified</td>
<td>True negative</td>
</tr>
<tr>
<td>Hs198694</td>
<td>Sep. ’99</td>
<td>Hs457939</td>
<td>21</td>
<td>No</td>
<td>One variant verified</td>
<td>True negative</td>
</tr>
<tr>
<td>Hs25854</td>
<td>Sep. ’99</td>
<td>Hs350208</td>
<td>21</td>
<td>Yes</td>
<td>One variant verified</td>
<td>False Positive (89%)—foetal brain.</td>
</tr>
<tr>
<td>Hs93931</td>
<td>Sep. ’99</td>
<td>Hs43931</td>
<td>21</td>
<td>Yes</td>
<td>One variant verified</td>
<td>False Positive (89%)—normalized ESTs</td>
</tr>
<tr>
<td>Hs38668</td>
<td>Sep. ’99</td>
<td>Hs38668</td>
<td>21</td>
<td>No</td>
<td>One variant verified</td>
<td>True negative</td>
</tr>
<tr>
<td>Hs66493</td>
<td>Sep. ’99</td>
<td>Hs408790</td>
<td>21</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (86%)</td>
</tr>
<tr>
<td>Hs75238</td>
<td>Sep. ’99</td>
<td>Hs75238</td>
<td>21</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (84%)</td>
</tr>
<tr>
<td>Hs86958</td>
<td>Sep. ’99</td>
<td>Hs86958</td>
<td>21</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (87%)</td>
</tr>
<tr>
<td>Hs821</td>
<td>Mar. ’01</td>
<td>Hs821</td>
<td>X</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive—non-consensus splicing</td>
</tr>
<tr>
<td>Hs1757</td>
<td>Mar. ’01</td>
<td>Hs1757</td>
<td>X</td>
<td>Yes</td>
<td>Both variants verified</td>
<td>True positive (89%)</td>
</tr>
</tbody>
</table>

The list shows the cluster ID from a previous version of SpliceNest database not employing the quality criterion with positive/negative results in the RT–PCR experiments. 15/17 clusters are now correctly annotated using the quality criterion, with the relevant quality values in parenthesis. The two mispredictions might be a result of low expression and/or developmental stage specific splicing (foetal brain specific).
REFERENCES


