On the Comparison of Sets of Alternative Transcripts

Aïda Ouangraoua[1], Krister M. Swenson[2], and Anne Bergeron[3]

[1] INRIA Lille, LIFL, Université Lille 1, Villeneuve d'Ascq, France
[2] Université de Montréal and McGill University, Canada
[3] Lacim, Université du Québec à Montréal, Montréal, Canada

Abstract. Alternative splicing is pervasive among complex eukaryote species. For some genes shared by numerous species, dozens of alternative transcripts are already annotated in databases. Most recent studies compare and catalog alternate splicing events within or across species, but there is an urgent need to be able to compare sets of whole transcripts both manually and automatically.

In this paper, we propose a general framework to compare sets of transcripts that are transcribed from orthologous loci of several species. The model is based on the construction of a common reference sequence, and on annotations that allow the reconstruction of ancestral sequences, the identification of conserved events, and the inference of gains and losses of donor/acceptors sites, exons, introns and transcripts.

Our representation of sets of transcripts is straightforward, and readable by both humans and computers. On the other hand, the model has a precise, formal specification that insures its coherence, consistency and scalability. We give several examples, among them a comparison of 24 Smox gene transcripts across five species.

1 Introduction

One of the most intriguing and powerful discoveries of the post-genomic era is the revelation of the extent of alternative splicing in eukaryote genomes, where a single gene sequence can produce a multitude of transcripts [3,5]. The "one gene, one protein" dogma of the last century has not merely been shaken, it has been shattered into pieces, and these pieces tell a story in which genome sequences acquire new function not only by mutation, but by being processed differently.

The main inspiration for this work is the recent paper by [12] that describes the variety of splicing events in vertebrates: the authors carefully annotated and validated hundreds of transcripts from over three hundred genes in human, mouse and other genomes, yielding dozens of conserved or species-specific splicing events. The results are given as combined statistics by species or group of species, and cataloged as one of 68 different kinds of splicing events.

However, beyond recognizing that two transcripts are conserved between species, or that a specific alternative splicing event is conserved, there is no formal setting for the comparison of two or more sets of transcripts that are

L. Bleris et al. (Eds.): ISBRA 2012, LNBI 7292, pp. 201–212, 2012.

transcribed from homologous genes of different genomes. The most widely used approach is to resort to comparing all pairs of transcripts within a set, or between two sets (see [19] for a review).

There are several hurdles on the way to a good representation. The first comes from the fact that, when alternate transcripts were scarce, much of the focus was directed towards the representation of alternative splicing events: splicing graphs [8] or pictograms [1] are adequate but do not scale easily to genes that can have dozens of transcripts, or to comparison between multiple species. Other representation techniques, such as bit matrices and codes (see [13,16] and references therein), proposed for the identification and the categorization of alternative splicing events are often more appropriate for computers than for human beings. A second problem is the identification of the features to compare. The splicing machinery is entangled with a myriad of bits and pieces that can vary within and between species: transcripts, coding sequences, exons, introns, splicing donor and acceptor sites, start and stop codons, untranslated regions of arbitrary lengths, frame shifts, etc. Ideally, a model would capture as much as is known about transcripts, including the underlying sequences. In that direction, the goal of the Exalign method [14,18] is to integrate the exon-intron structure of transcripts with gene comparison, in order to find "splicing orthology" for pairs of transcripts. What about orthologous sets of transcripts?

Here we propose a switch from the paradigm of comparing single transcripts between species, to comparing all transcripts with respect to a common reference sequence — derived from a multiple alignment when several species are considered — on which splicing events are represented in a consistent manner. We show that this yields very flexible tools to compare sets of transcripts, that can incorporate the various mechanisms that drive transcript evolution.

The paper is organized as follows. Section 2 presents the model in the single species context. This elementary model is developed in Sections 3 and 4 to include coding sequences and multiple species comparison. Section 6 contains the formal details of the common reference sequence construction.

2 Basic Representations of Transcripts

In this section, we give an intuitive presentation of the basic techniques for representing transcripts. It is self-contained, and the definitions should give the reader a basic understanding of the model. Section 6 presents the formal model, with a level of detail that is not necessary for most situations.

We call an RNA molecule that has made it to a sequencing machine a *transcript*. The sequence for a transcript usually matches the genomic sequence, or *locus*, from which it was transcribed in a piecewise way: parts of the genomic sequence are spliced out, called *introns*, and parts are retained, the *exons*. Different transcripts from the same genomic region may exhibit different combinations of introns and exons in what is called *alternative splicing*, as illustrated in Figure 1 (A) for the human Smox gene.

In order to represent a variety of transcripts in a simple way, we first define blocks of consecutive exons that have no internal variation:

Definition 1. *Given a set of transcripts from the same genomic sequence, an* exon block *is a maximal sequence of adjacent exons or exon fragments that always appear together in the set of transcripts.*

An exon block may contain introns, for example block B of Figure 1 contains two introns. These introns are inconsequential to our comparisons, since they always start and end at the same position of the genomic sequence in all transcripts of the set: they are said to be *identically spliced*. Maximal intron fragments that are common to all transcripts but not identically spliced form *intron blocks*, such as blocks C, G and I of Figure 1.

Labeling the segments delimited by exon and intron blocks on the genome sequence, from left to right, gives the *reference sequence* $\mathcal{R}$ corresponding to a set of transcripts. Each segment of the reference sequence is simply called a *block*. For example, the reference sequence for the Smox gene of Figure 1 is ABCDEFGHIJ.

A *splicing event* is the removal of a consecutive set of blocks in the reference sequence $\mathcal{R}$. Transcripts can be represented by sequences of block labels, such as AB.D.F.J for transcript H001 in Figure 1 (B), where the dots between blocks indicate the position of the splicing events that created the transcript.

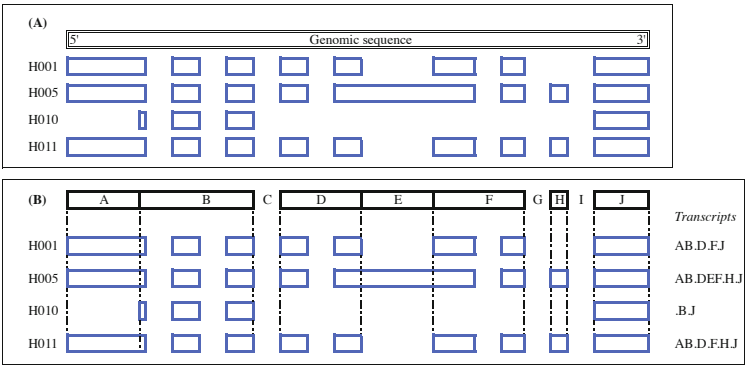

Fig. 1. Four transcripts of the human gene Smox. (A) The transcripts mapped to the genome sequence. Exons are depicted as blue boxes. Transcripts are named with a four symbol code that corresponds to the first letter of the species name, followed by the last three symbols of their transcript identifier in the Ensembl database (see Appendix 1). (B) Blocks of exons or exons fragments that always appear together in the transcripts are depicted by black boxes. Labeling the segments from A to J gives the reference sequence. The four transcripts are represented by the sequences AB.D.F.J, AB.DEF.H.J, .B.J and AB.D.F.H.J.

Introns that are removed by a splicing event are characterized by a *donor site* at their 5' end, known as the exon-intron junction, and by an *acceptor site* at their 3' end, known as the intron-exon junction. The existence of donor or acceptor sites is an attribute of a specific locus or gene sequence. Thus, given a set of transcripts that are transcribed from the same locus, it is natural to annotate a reference sequence with their donor/acceptor sites. We use the following convention:

Definition 2. *In a reference sequence $\mathcal{R}$, we denote by '<b' the fact that the beginning of block b is a donor site, and by 'b>' the fact that the end of block b is an acceptor site.*

For example, the annotated reference sequence for the Smox gene of Figure 1 is:

A>B<C>D<E>F<G>H<I>J

Since splicing events are consecutive along a transcript, it is also possible to use these annotations to indicate the splicing events that constructed transcripts. For example the four transcripts of Figure 1 may be represented by:

H001: AB<C>D<E>F<GHI>J
H005: AB<C>DEF<G>H<I>J
H010: A>B<CDEFGHI>J
H011: AB<C>D<E>F<G>H<I>J

In the representation of individual transcripts, it is understood that, for internal splicing events, each '<' symbol is paired with the next '>' symbol to create a splicing event. Since donor and acceptor annotations are sets, all the usual set operations can be applied to these annotated sequences. For example, the common annotations for the four transcripts of the human Smox gene is:

AB<CDEFGHI>J

In the next sections, we will see that the main advantages of this representation is that it can be used immediately to represent sets of coding transcripts, and to compare sets of transcripts from different species.

3 Integrating Coding Information

Many transcripts will eventually be translated into proteins. The formalism of the preceding section can be adapted to indicate coding exons by adding block separators before each start codon, and after each stop codon. We associate to each block its *coding sequence*: for blocks that contain introns, the coding sequence is the concatenation of its exons or exon fragments; for blocks that contain only intron fragments, the coding sequence is empty. A *coding block* is a block whose coding sequence is not empty.

Figure 2 shows an example using four transcripts of the human Crem gene. Shaded rectangles represent translated regions.

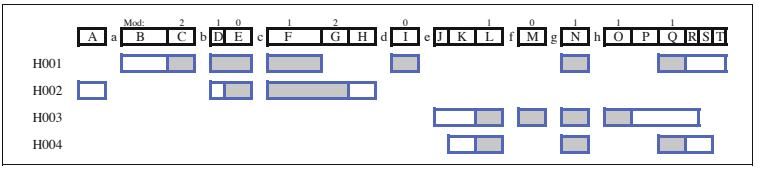

Fig. 2. Four transcripts of the human gene Crem. Shaded rectangles represent translated regions.

The length of a block coding sequence is not necessarily a multiple of three, although the total length of all coding sequences of a transcript must be. When a coding block is skipped, this event may introduce a *frameshift* if the length of the skipped block is not a multiple of three. In order to keep track of the possibility of frameshift, we associate to each block a Mod value which is the remainder of the length of its coding sequence divided by 3. When studying transcripts from a single species, these values can help determine alternative early stop codons, as illustrated by H003 compared to H001 and H004 in Figure 2. However, we will see in Section 6 that these values also have an impact on block construction.

Trimming the blocks that precede the first start codon (in the set), and those that succeed the last stop codon (in the set) does not significantly alter the information content. In the following table, the untranslated blocks of the transcripts of Figure 2 have been painted in blue, and the second column give the shortened versions of the transcripts, where the blocks preceding block C and following block Q have been trimmed.

H001	.BC.DE.F.I.N.QRST	C.DE.F.I.N.Q
H002	A.DE.FGH.	.DE.FGH.
H003	.JKL.M.N.OPQR.	.JKL.M.N.OPQ
H004	.KL.N.QRS.	.KL.N.Q

Early start codons, and late stop codons, are still identifiable in the shortened version, and the untranslated regions that are coding in alternate transcripts are visible. Note that skipping the untranslated region before the start codons is often done in practice due to the variability of the transcription initiation site [5], and splicing events that occur in untranslated regions after a stop codon often lead to a condition known as *nonsense mediated decay* that prevents the translation of the transcript [11].

4 Multi-species Comparisons

Here we show how to adapt the representations of the last two sections to sets of transcripts that come from orthologous loci in multiple species.

Our goal is to construct a common reference sequence using a multiple alignment, and cut it into blocks. The examples of this section are straightforward

applications of the representation. Exceptions and limiting cases are treated in Section 6.

Given an alignment of two or more genes and a transcript t from one of these genes, it is always possible to color the exons of t in the corresponding gene sequence. For example, Figure 3 shows an alignment of the human and mouse Ensa gene, human transcript H004 has been colored in blue for the human sequence, and mouse transcript M201 has been colored in red. In the single species context, the blocks of the reference sequence were segments of the genomic sequence, here the blocks are defined as intervals of the columns of the alignment. Block junctions correspond to positions where changes of color occur. The common

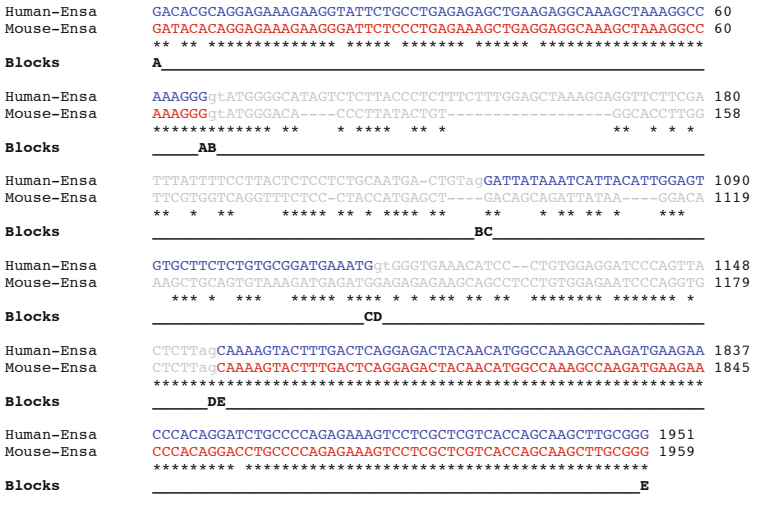

Fig. 3. An alignment of the partial human and mouse Ensa gene, where the exons of human transcript H004 have been colored in blue, and the exons of mouse transcript M201 in red. The blocks of the reference sequence are given below the alignment.

reference sequence for the human and mouse Ensa gene would be the sequence ABCDE as defined by Figure 3. Although block C is an exon in the human transcript, preceded by the canonical 'ag' acceptor site at its 5' junction, and followed by the canonical 'gt' donor site at its 3' junction, these features do not exist in the mouse sequence. However, both transcripts can be represented with the formalism of the preceding section, human as A<B>C<D>E and mouse as A<BCD>E.

This approach can be applied to genes that have a variety of transcripts in many different species. Figure 4 shows 24 transcripts of the Smox gene: 4 from the human gene, 1 from the chimpanzee gene, 2 from the orangutan gene, 10 from the mouse gene and 7 from the rat gene. We chose transcripts that were annotated in the CCDS database [15] for human and mouse, along with transcripts that were annotated as protein coding in the Ensembl database [7].

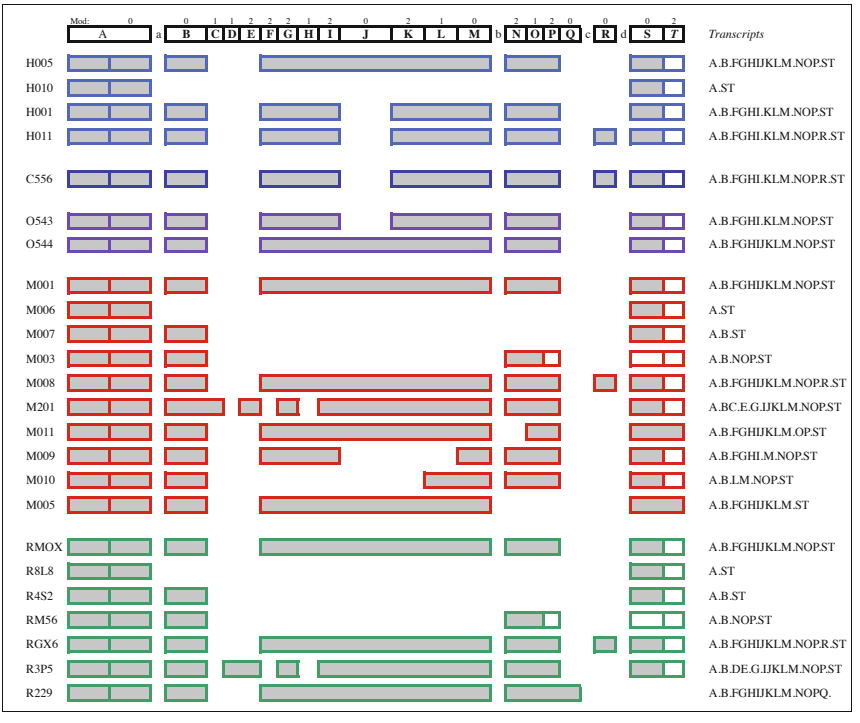

Fig. 4. An alignment of the Smox gene, where the exons of 4 human transcripts have been colored in blue, 1 chimpanzee transcript in dark blue, 2 orangutan transcripts in indigo, 10 mouse transcripts in red, and 7 rat transcripts in green

In the common reference sequence, there are 19 blocks that contain at least one exonic sequence, labeled from 'A' to 'T', and 4 blocks that are common intron fragments, labeled from 'a' to 'd'.

Transcripts that have the same representation are *splicing orthologs* [14]. For example, in Figure 4, the following sets of transcripts are sets of orthologs: {H005, O544, M001, RMOX}, {H010, M006, R8L8}, {H011, C556}, {H001, O543}, {M007, R4S2}, {M003, RM56} and {M008, RGX6}. In order to say more about the relations between these sets, we may compare the reference sequences for each species, constructed using the common reference:

Human: A<a>B<CDE>FGHI<J>KLM<b>NOP<Qc>R<d>ST
Chimp: A<a>B<CDE>FGHI<J>KLM<b>NOP<Qc>R<d>ST
Orang: A<a>B<CDE>FGHI<J>KLM<b>NOP<QcRd>ST
Mouse: A<a>B<C<D>E><F>G<H>I<JK>L>M<b>N>OP<Qc>R<d>ST
Rat: A<a>B<C>DE><F>G<H>IJKLM<b>NOP<Q<c>R<d>ST

Annotations of the human and chimpanzee are equal, even though only one transcript was observed in the chimpanzee. Annotations of the orangutan are contained in the human-chimpanzee annotations, since exon >R< has not – yet? – been

observed in the orangutan. The annotations that are common to all five species reference sequences are: <a>, <C, E>, <b>, <Q, and d>. The common annotations are captured by the sequence:

$$A<a>B<CDE>FGHIJKLM<b>NOP<QcRd>ST$$

Annotations can also be used in the context of ancestral reconstruction. Each donor or acceptor site is either present or absent in a locus reference sequence. Given a phylogenetic tree for a set of loci, it is possible to apply a Fitch-like algorithm [6] to each site in order to determine the ancestral states yielding the minimum number of gains/losses of sites. Figure 5 gives an example of such a reconstruction, using the five reference sequences of the Smox gene without their common annotations. Annotations unambiguously assigned to the ancestor of the five species are <J, c> and <d. The rodent ancestor has <F> and <H>, and the primate ancestor has J>.

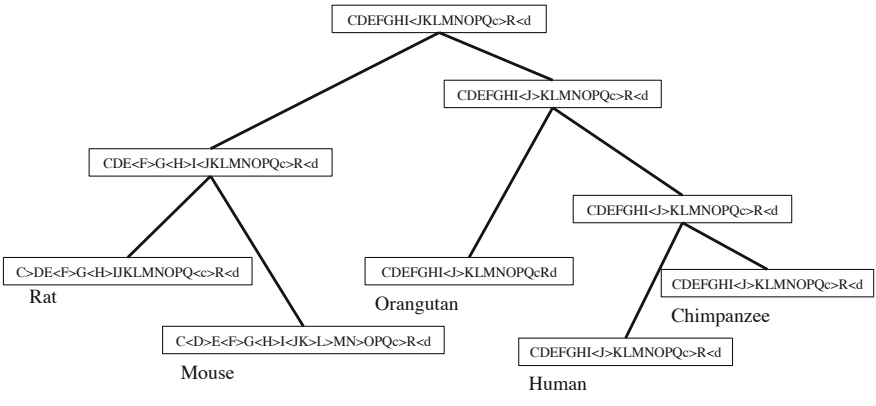

Fig. 5. Ancestor reconstruction. Donor/acceptor sites are assigned to each node with a Fitch-like algorithm to minimize the number of gains/losses of sites.

5 Comparison with Other Formalisms

Most formalisms for the representation of splicing events were developed for the comparative study of transcripts of a single gene [1,8], or the comparison of single events in the same gene or between orthologuous genes [16,12]. In order to be able to assess splicing orthology between two sets of transcripts, as proposed in [18], it is necessary to be able to represent, with the same formalism, exon and intron structure of single transcripts, common and diverging structures of transcripts from the same gene, and common and diverging structures of sets of transcripts from different species.

In the initial phase of this project, we explored some generalizations of splicing graphs [8] and their variants [1,2]. The simplest solution was to construct the graphs using the common blocks of the different species, and compare sets of

transcripts using graph comparisons. Assuming that these tasks are simple, there remains a major problem in the fact that two sets of transcripts that have no common elements may still yield the same graph [9]: it is thus necessary to keep track of individual transcripts, together with more general representations that capture their similarities.

The approach of Mudge et al. [12] is adapted to the comparison of splicing events among several species, but focus on the alternative aspect of events: two species may share a transcript ABCDE, but if one has the alternative event A<B>CDE, and the other event ABC<D>E, a local approach would conclude that there is no conserved splicing event. In this case, we think that the fact that the two species have a common transcript should witness a certain degree of similarity. The Bubbles formalism [17] has been proposed to describe the splicing events of a set of transcripts, but can only be applied to the set of common transcripts of a group of species.

Zambelli et al. [18] introduce the concept of "iso-orthology" between transcripts whose less stringent class corresponds roughly to equality of sequences in our proposed representation. However, iso-orthology cannot capture more general similarities between sets of transcripts. For example, in Figure 4, the sets of transcripts of the human, chimpanzee and orangutan are remarkably similar, despite the fact that there are no common iso-orthologs in the three species; the mouse and rat transcripts have five iso-orthologs, but they also each have two transcripts (M201 and R3P5) that are not iso-orthologs, but share a quite distinct conserved feature, exon G, that seems to be unique to the rodents. On the other hand, the similarity, or dissimilarity, of the reference sequences proposed in Section 4 would capture these relations.

6 Formalities of Block Construction

Blocks are constructed using transcript annotations and a multiple alignment. For simplicity, we assume that the transcripts of a species are all transcribed from a contiguous locus, or *gene*, and that the annotations give the position of each exon with respect to the coordinates of the locus.

A multiple alignment of all genes under consideration is obtained using standard software – we used ClustalW [10] for the examples presented in this paper. The quality of the comparison will clearly be influenced by the quality of the alignment: genes that share some highly homologous exons will be easier to compare, even if some exons might be absent in some species.

Let n be the number of columns of the alignment. A transcript t from gene g is represented by a list of n values, 0,1, or '-'. Position i has value 1 if and only if the nucleotide at column i in gene g belongs to an exon of transcript t, or column i belongs to a gap flanked by two nucleotides of the same exon of transcript t. Position i has value 0 if and only if the nucleotide at column i in gene g belongs to an intron of transcript t, or column i belongs to a gap flanked by two nucleotides of the same intron of transcript t. Otherwise the value at position i is the gap character '-'. This yields the *transcript matrix*, which is reminiscent of the bit matrices used in [13].

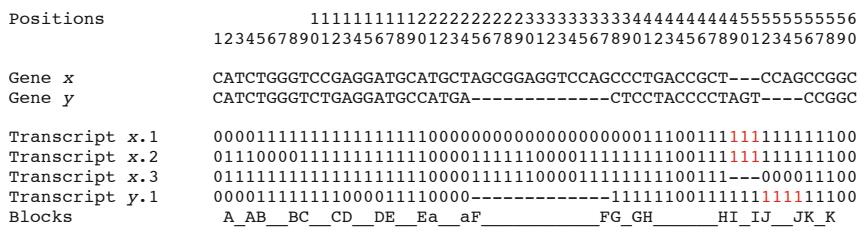

Fig. 6. An example of block construction. Values in red correspond to gaps that are within exons.

For example, Figure 6 shows an alignment of two genes, with transcripts $x.1$, $x.2$ and $x.3$ from Gene x, and transcript $y.1$ from Gene y.

A *positive* column in the transcript matrix is a column that contains at least one value 1. Blocks are defined as maximal intervals $[i..j]$ such that 1) column i and j of the transcript matrix are positive, and all the positive columns in the interval are equal, or 2) maximal intervals between two such intervals. For a gene, the *sequence* of a block $[i..j]$ is the subsequence of nucleotides that corresponds to positive columns between i and j. When all the Mod values of the non-empty sequences of a block are equal, then the Mod value of the block can be clearly assigned. Since the block structure should reflect ortholog exons or exon parts, different Mod values require further investigation especially for coding exons.

In the above example, block F is defined by positions [25..37] of the alignment. Of these 13 positions, only 9 positions have a positive column, thus the sequence of block F for Gene x is 'AGCGGACAG', and for Gene y, the empty sequence. For blocks that do not contain any positive column, all the associated sequences are empty. The sequence $\mathcal{R}$ of all blocks of a multiple alignment is called the *reference sequence*. A consecutive sequence of blocks is called an *exon* if the sequence is the longest possible that exactly contains the nucleotide sequence of an exon of one of the transcripts. The sequence of blocks between two exons of the same transcript is called an *intron*. Some 'real' introns may not be representable as sequences of blocks if they are always identically spliced between the same exons. An example is the intron contained in block H in Figure 6.

The reference sequence of the above example is ABCDEaFGHIJK. The four transcripts are represented as follows: $x.1$ as BCDEHIJK, $x.2$ as ACDEFGHIJK, $x.3$ as ABCDEFGHIK, $y.1$ as BCEFGHIJK. Note that the first of these transcripts does not have block F, since it is included in one of its introns. The last has block F, but the corresponding sequence is empty. The second exon of transcript $x.3$ is represented by the sequence FGHI, even though the sequence of block I is empty for Gene x. In this model, we assume that the multiple alignment includes all genes under consideration, even if some genes do not have observed transcripts. Adding new transcripts may split existing blocks.

Proposition 1. *The nucleotide sequence of a transcript from a gene s is the concatenation of the nucleotide sequences of the corresponding blocks $b_1 \dots b_k$ for gene s.*

As in Section 2, a splicing event removes consecutive blocks from the reference sequence. Any such event defines a donor and/or an acceptor site:

Definition 3. *Block $b = [i..j]$ contains a* donor site*, denoted '$<b$', if one of the transcripts has an intron starting at position i of the alignment. It contains an* acceptor site*, denoted '$b>$', if one of the transcripts has an intron ending at position j of the alignment.*

7 Conclusion

We have described a succinct and readable representation for transcripts of alternatively spliced genes. Our representation is amenable to the comparison of sets of transcripts for a single gene, or to sets of transcripts corresponding to orthologous genes across multiple species. To our knowledge, this is the first such representation in the literature; existing studies consider single transcripts, or splicing events across multiple species in isolation. The utility of the representation was demonstrated first on the set of transcripts from the human Smox gene and then on the coding transcripts of the human Crem gene. The Smox gene was then considered in a phylogenetic context where ancestral sets of transcripts were inferred via maximum parsimony. Beyond ancestral inference, we expect that this representation will lead to new tools for phylogeny reconstruction [see [4], for example], transcript discovery, and homologous gene discovery.

References

1. Bollina, D., Lee, B.T., Tan, T.W., Ranganathan, S.: ASGS: an alternative splicing graph web service. Nucleic Acids Res. 34, W444–W447 (2006)
2. Bonizzoni, P., Mauri, G., Pesole, G., Picardi, E., Pirola, Y., Rizzi, R.: Detecting alternative gene structures from spliced ESTs: a computational approach. J. Comput. Biol. 16, 43–66 (2009)
3. Carninci, P., Kasukawa, T., Katayama, S., et al.: The transcriptional landscape of the mammalian genome. Science 309, 1559–1563 (2005)
4. Christinat, Y., Moret, B.M.E.: Inferring transcript phylogenies. In: Proc. of IEEE International Conference on Bioinformatics and Biomedecine, pp. 208–215 (2011)
5. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. Nature 447, 799–816 (2007)
6. Fitch, W.M.: Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20(4), 406–416 (1971)
7. Flicek, P., Amode, M.R., Barrell, D., et al.: Ensembl 2011. Nucleic Acids Res. 39, D800–D806 (2011)
8. Heber, S., Alekseyev, M., Sze, S.H., Tang, H., Pevzner, P.A.: Splicing graphs and EST assembly problem. Bioinformatics 18(suppl.1), S181–S188 (2002)

9. Lacroix, V., Sammeth, M., Guigo, R., Bergeron, A.: Exact Transcriptome Reconstruction from Short Sequence Reads. In: Crandall, K.A., Lagergren, J. (eds.) WABI 2008. LNCS (LNBI), vol. 5251, pp. 50–63. Springer, Heidelberg (2008)
10. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, A., Wilm, I.M., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.: Clustal w and clustal x version 2.0. Bioinformatics 23, 2947–2948 (2007)
11. Mendell, J.T., Sharifi, N.A., Meyers, J.L., Martinez-Murillo, F., Dietz, H.C.: Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nature Genetics 36, 1073–1078 (2004)
12. Mudge, J.M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald, C., Reymond, A., Guigo, R., Hubbard, T., Harrow, J.: The origins, evolution and functional potential of alternative splicing in vertebrates. Molecular Biology and Evolution 28, 2949–2959 (2011)
13. Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M., Gotoh, O.: Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. Bioinformatics 22(10), 1211–1216 (2006)
14. Pavesi, G., Zambelli, F., Caggese, C., Pesole, G.: Exalign: a new method for comparative analysis of exon-intron gene structures. Nucleic Acids Res. 36, e47 (2008)
15. Pruitt, K.D., Harrow, J., Harte, R.A., et al.: The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 19, 1316–1323 (2009)
16. Sammeth, M., Foissac, S., Guigo, R.: A general definition and nomenclature for alternative splicing events. PLoS Computational Biology 8, e1000147 (2008)
17. Sammeth, M., Valiente, G., Guigo, R.: Bubbles: Alternative Splicing Events of Arbitrary Dimension in Splicing Graphs. In: Vingron, M., Wong, L. (eds.) RECOMB 2008. LNCS (LNBI), vol. 4955, pp. 372–395. Springer, Heidelberg (2008)
18. Zambelli, F., Pavesi, G., Gissi, C., Horner, D.S., Pesole, G.: Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC Genomics 11, 534 (2010)
19. Zavolan, M., van Nimwegen, E.: The types and prevalence of alternative splice forms. Curr. Opin. Struct. Biol. 16, 362–367 (2006)