EST alignments suggest that [secret number]% of Arabidopsis thaliana genes are alternatively spliced

Dan Morris
Stanford University Robotics Lab
Computer Science Department
Stanford, CA 94305-9010
dmorris@cs.stanford.edu

THIS DOCUMENT'S LIMITED USEFULNESS…

This document contains cool pictures with no useful data; this work was done as part of a corporate internship that used proprietary data, so we couldn't release or publish sequence information or real numbers. It's purely a description of what I worked on, with no constructive scientific content. APPENDIX A contains a number of (unlabeled) images representing examples of alternative splicing. The extremely informal APPENDIX B describes how we arrived at some of our filtering procedures, using public-domain examples.

ABSTRACT

Recent estimates of the frequency of alternative splicing in human genes, in conjunction with recent estimates of the number of genes in the human genome, have confirmed that alternative splicing plays a critical role in human genetic complexity. However, despite this recent interest in alternative splicing as a fundamental genetic mechanism, and despite numerous experimentally confirmed examples of alternative splicing in plants, no genome-wide studies have been performed to estimate the frequency of alternative splicing in any plant species. Here we present an estimate of the number of alternatively spliced genes in Arabidopsis thaliana, based on genomic alignments of approximately [very large number] EST sequences. After strict filtering to eliminate false-positives, we estimate that the total number of alternatively spliced genes in Arabidopsis is [top secret number], or about [top secret number] of the total gene number.

Keywords
Arabidopsis, EST's, useless documents

1. INTRODUCTION

Alternative splicing (AS) – the processing of individual pre-mRNA's into multiple mature mRNA forms – is used extensively by animal cells to modulate gene function. The high frequency of AS in many animal species allows the effective number of protein products available to an organism to be significantly higher than the number of genes in the genome. As a result, AS plays a key role in animal function and development, and is known to be involved in numerous physiological processes, such as apoptosis (Boise et al, 1993) and leukocyte differentiation (Eckhart et al, 2001). AS has also been implicated in numerous human disease states (Citron et al, 2002; Qi and Byers, 1998).

With the recent availability of human genetic information, it has become possible to study this important phenomenon on a genome-wide scale. Several approaches have been taken to estimate the number of alternatively spliced genes and/or splice variants in the human genome: Brett et al (2000) aligned EST sequences against mRNA sequences to find overlapping alignments, and estimated that 38% of human mRNA's contain possible splice forms; Modrek et al (2001) aligned EST sequences against the draft human genome sequence, and estimated that the proportion of alternatively spliced human genes is 42% or higher.

Alternative splicing does not appear to be nearly as common in plants, but several examples of AS in plants have been confirmed biochemically (e.g. Werneke et al, 1989; Golovkin and Reddy, 1996). In some cases, plant AS has been shown to be regulated in a functionally important manner – for example, Mano et al (2000) demonstrated light-regulated alternative splicing of a leaf protein. Thus it seems that AS may be an important mechanism in plant cells as well as in animal cells, although it is almost certainly not as common.

With the recent sequencing of the five-chromosome, 125-megabase Arabidopsis thaliana genome (Arabidopsis Genome Initiative, 2000), it became possible to perform a genome-wide analysis of alternative splicing in Arabidopsis. We have used this genomic information, in conjunction with the public Arabidopsis expressed sequence tag (EST) database and our own in-house EST database, to computationally estimate the frequency of AS in Arabidopsis using a method that is similar to those applied to the human genome. We estimate that over [secret number] Arabidopsis genes (about [secret number]% of the total gene number) are alternatively spliced, and nearly all of the genes we have identified represent novel splice variants.

2. RESULTS

By aligning ESTs against the Arabidopsis genomic sequence, we identified [secret number] genes for which EST alignments appear to demonstrate multiple splice variants. A strict filtering process (discussed below) was used to remove false-positives that may have resulted from sequencing error, alignment error, gene duplication, etc. Thus we assign a high confidence level to each of the alternative splicing events that we include in the final count.

We classified each of the AS events into exon-skipping events and exon-extension events (Figure 1), and the exon-extension events were further classified by the length of the "extension": exons that differed in length between splice variants by a multiple of 3 base pairs (thus maintaining the reading frame of the gene) were separated from those events that did not demonstrate a 3n-bp extension. The results are not displayed in Table 1.

Table 1 also includes (in your imagination) the total number of genes for which EST alignments displayed "intron-retention", which are not included in our total count of alternatively spliced genes. Although cases have been reported in which one mature splice variant contains a complete "intron" that is spliced out of another RNA (Golovkin and Reddy, 1996), our methodology does not allow us to distinguish these cases from erroneous inclusion of unspliced pre-RNA in our EST databases.

The total number of exon-extension and exon-skipping events not reported in Table 1 is [secret number]. This figure represents the total number of exons that are alternatively spliced. These [secret number] alternatively spliced exons are distributed over [secret number] genes, so many of the genes that we suggest are alternatively spliced actually demonstrate alternative splicing at more than one site. Note that this figure does not allow us to compute the total number of splice variants for any of the genes that demonstrate multiple sites for alternative splicing. We use un-assembled ESTs for our detection of alternative splicing, so we cannot determine whether alternative splicing that is observed at multiple sites within a gene occurs independently at each site. Thus we cannot assess the number of splice variants per gene using this method.

After identifying alternatively spliced genes and categorizing them as discussed, we used the public database of confirmed and predicted Arabidopsis coding sequences to analyze the fraction of alternative splicing events that occur in coding and non-coding regions. The results are not displayed in Table 2.

[Secret number] percent of the predicted alternative splicing events could not be associated with a predicted coding sequence from the public database, and could thus not be classified as coding, 3'-UTR, or 5'-UTR. These may represent non-translated genes, false negative predictions in the public database, or a neighboring gene with an extremely long UTR, which might cause an AS event to appear to be between genes.

3. MATERIALS AND METHODS

3.1 Data Sources

The complete genome of Arabidopsis thaliana is made publicly available by the Arabidopsis Genome Initiative, and can be downloaded from the AGI website:

ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/Sequences/

Approximately [some fraction] of our EST's were collected from dbEST:

http://www.ncbi.nlm.nih.gov/dbEST/

The remainder were sequenced in-house.

3.2 Alternative Splicing Analysis

All [very large number] ESTs were assigned preliminary alignments against the Arabidopsis genome using BLAST (Altschul et al, 1990) with a threshold of E < 10-30. Using software written in Perl and C++, we then extracted the genomic regions corresponding to the top BLAST hit for each EST, including a 2-kilobase flanking sequence on either side of the hit region. The sim4 alignment program (Florea et al, 1998) was used to align each EST to the corresponding extracted genomic segment, producing a high-quality genomic alignment for each EST. At this stage, we threw out any ESTs that aligned to multiple genomic regions with greater than 95% identity, to avoid errors caused by alignments to pseudogenes or duplicated genes.

Initial candidate AS events were identified using a program written in C++, which sorted all introns and exons generated by sim4 by genomic coordinates and generated a list of cases in which an intron or exon in one EST aligned with an intron-exon boundary in another EST (Figure 1 demonstrates that this is the basic criteria for the types of alternative splicing we are investigating). In subsequent discussion, an "AS event" will represent two ESTs that contain this type of overlapping alignment.

Note that we did not classify the ends of ESTs as intron-exon boundaries; only internal intron-exon boundaries were considered in our analysis, since EST ends are especially prone to sequencing error and are difficult to classify according to our AS categories. Similarly, we only considered ESTs that contain at least one intron.

After identification of preliminary candidate events, we subjected each candidate AS event to a series of filters to remove false positives, which generally resulted from alignment or sequencing error. We removed any events for which any exons on either EST aligned with less than 95% identity, any events for which an intron adjacent to the event was longer than 2000 bp, any events for which an exon adjacent to the event was shorter than 20bp, and any events for which the identified intron-exon boundary was within 50bp of the end of either EST. These filters removed the most common types of false positives. We also removed any events that represent exon-extensions where an extension of the same size occurs at the opposite end of the relevant intron, to avoid errors caused by a shifting of bases from one end of an intron to the other in the alignment process.

Redundant events were removed, and each of the remaining AS events was classified into exon-extension, exon-skipping, or intron-retention.

In order to label events as coding, 3'-UTR, or 5'-UTR, we also aligned the AGI gene predictions to our genomic sequence (also using BLAST and sim4). Each AS event was associated with the AGI predicted coding sequence with which it most closely aligned.

4. DISCUSSION

The data presented here represent an estimate of the frequency of alternative splicing in Arabidopsis thaliana, which we take to be about [secret number] of the total number of genes in the genome (25,498) (Arabidopsis Genome Initiative, 2000). As expected, this is significantly lower than the corresponding estimates made in studies of the human genome. But it represents a significant fraction of genes in Arabidopsis, and suggests that alternative splicing may also be a significant genetic mechanism in plant species.

The locations of AS events within genes are fairly consistent with previous distributions established for the human genome. Our

data indicate that blah%, blah%, and blah% of AS events are found in coding regions, 5'-UTR, and 3'-UTR respectively, and previous results indicate that 74%, 22%, and 4% of human AS events are located in coding regions, 5'-UTR, and 3'-UTR respectively (Modrek et al, 2001).

Given our strict filtering criteria, we expect that nearly all of our proposed alternative splicing events actually represent alternative splicing. However, due to incomplete and non-uniform EST coverage, and due to incomplete and non-uniform sampling of Arabidopsis tissue types, we may have significantly underestimated the fraction of genes that actually contain multiple splice variants. It is very difficult to overcome this limitation without continued sequencing of large numbers of Arabidopsis ESTs.

The same limitations prevent us from determining whether specific splice variants tend to be localized to particular tissue types, since in most cases we do not have large numbers of ESTs representing each splice variant. We did examine this relationship using our current EST set, but found no specific biases among tissue types. That is, the distribution of alternatively spliced ESTs among tissue types was consistent with the distribution of ESTs in our database: blah% leaf, blah% flowering bud, blah% stem, blah% opened flower, blah% seed, blah% root.

We also examined the relationship between alternatively splicing and the nucleotide sequence at intron-exon boundaries, examining the possibility that non-GT/AG splice sites might be linked to alternative splicing, but again found no statistically significant biases. To avoid biases introduced by sim4 (which trims its alignments to consensus splice site sequences), we performed this analysis using BLAST alignments that displayed 100% identity. The distribution of nucleotides around splice sites was consistent with published standards, and no bias toward non-standard splice site sequences was observed in alternatively spliced ESTs.

REFERENCES

[1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410

[2] Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815

[3] Boise LH, Gonzalez-Garcia M, Postema CE, Ding L, Lindsten T, Turka LA, Mao X, Nunez G, Thompson CB (1993) bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell 74(4): 597-608

[4] Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett 474(1): 83-6

[5] Citron BA, Suo Z, SantaCruz K, Davies PJ, Oin F, Festoff BW (2002) Protein crosslinking, tissue transglutaminase, alternative splicing and neurodegeneration. Neurochem Int 40(1): 69-78

[6] Eckhart L, Henry M, Santos-Beneit AM, Schmitz I, Krueger A, Fischer H, Bach J, Ban J, Kirchhoff S, Krammer PH, Mollinedo F, Tschachler E (2001) Alternative Splicing of Caspase-8 mRNA during Differentiation of Human Leukocytes. Biochem Biophys Res Commun 289(4): 777-81

[7] Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8(9): 967-74.

[8] Golovkin M, Reddy AS (1996) Structure and expression of a plant U1 snRNP 70K gene: alternative splicing of U1 snRNP 70K pre-mRNAs produces two different transcripts. Plant Cell 8(8): 1421-35.

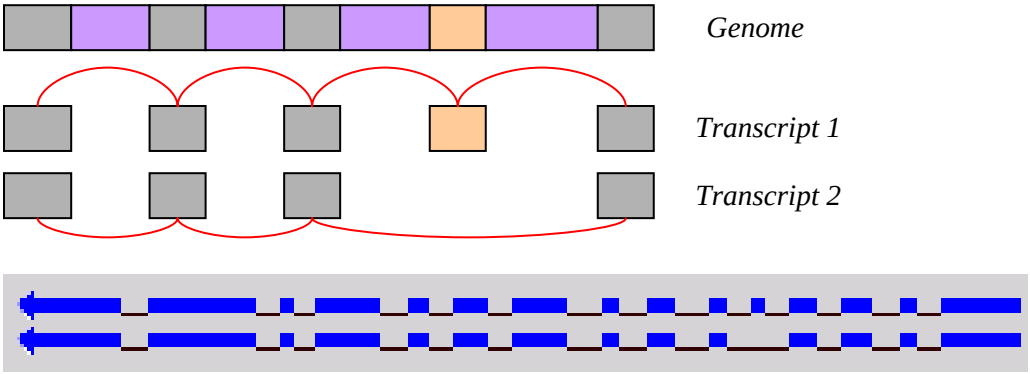
[9] International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921

[10] Klamt B, Koziell A, Poulat F, Wieacker P, Scambler P, Berta P, Gessler M (1998) Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of WT1 +/-KTS splice isoforms. Hum Mol Genet 7(4): 709-14

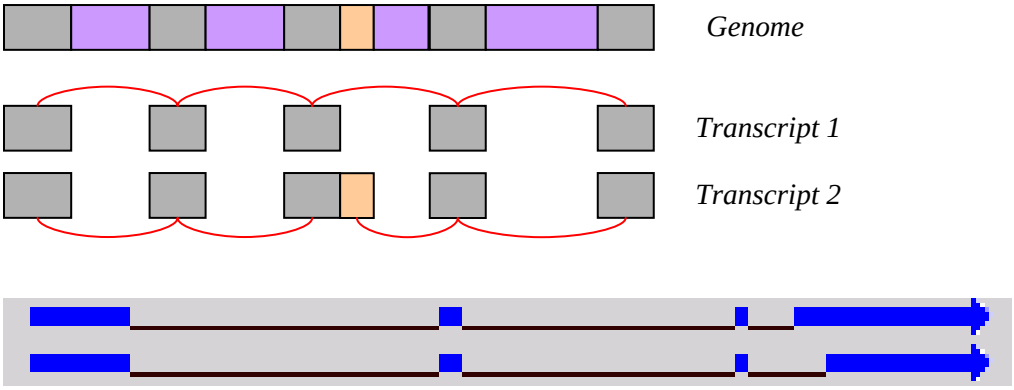
[11] Mano S, Hayashi M, Nishimura M (2000) A leaf-peroxisomal protein, hydroxypyruvate reductase, is produced by light-regulated alternative splicing. Cell Biochem Biophys 32:147-54

[12] Modrek B, Resch A, Grasso C, Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 29(13): 2850-9

Category I : Exon-skipping



Category II : Exon-extension



Category III : Intron-retention

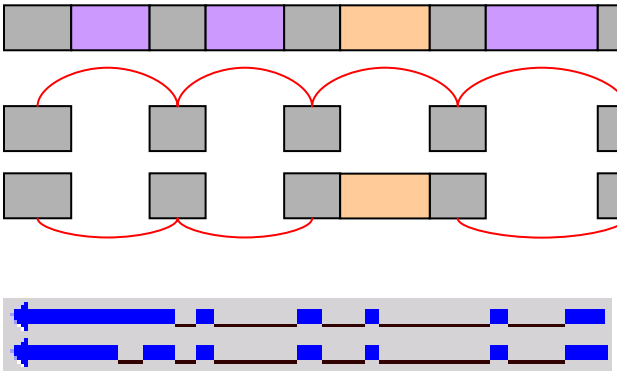
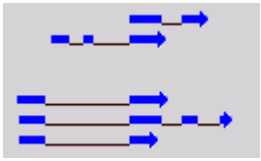
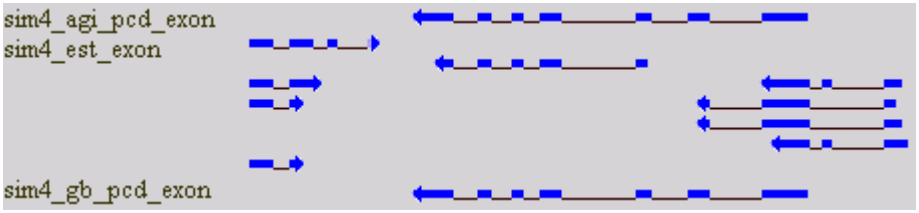


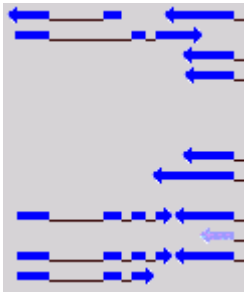
FIGURE 1, *Alternative splicing event categories*

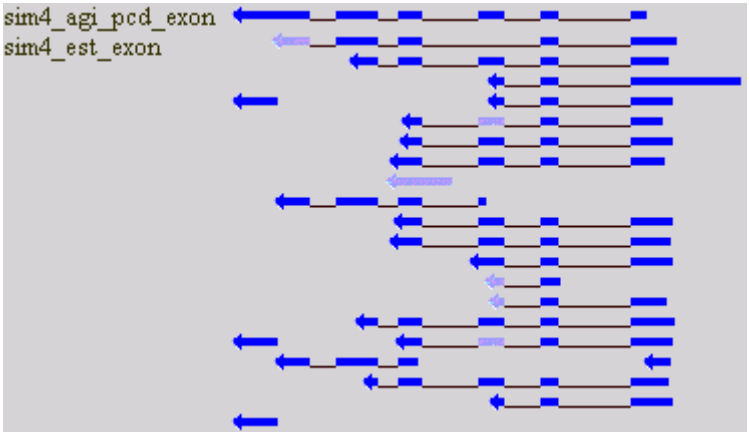
For each of the three categories into which we placed AS events, we display a schematic representation of the region around the alternative splice site, along with an actual example from our genomic EST alignments. The third category, intron-retention, includes alignments that may represent actual splice variants but more often represent pre-mRNA that was included in our EST database. Events in this category are not included in any of our reported statistics (which isn't interesting to you, since we don't report them here anyway!).

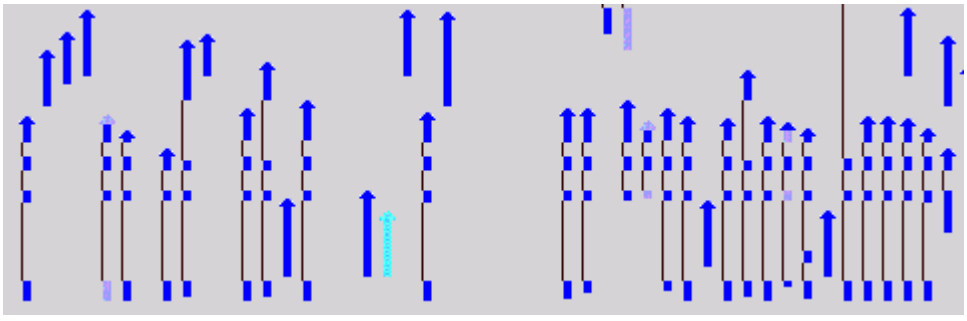
APPENDIX A: EXAMPLES OF IDENTIFIED ALTERNATIVE SPLICE VARIANTS (WITHOUT USEFUL LABELS)

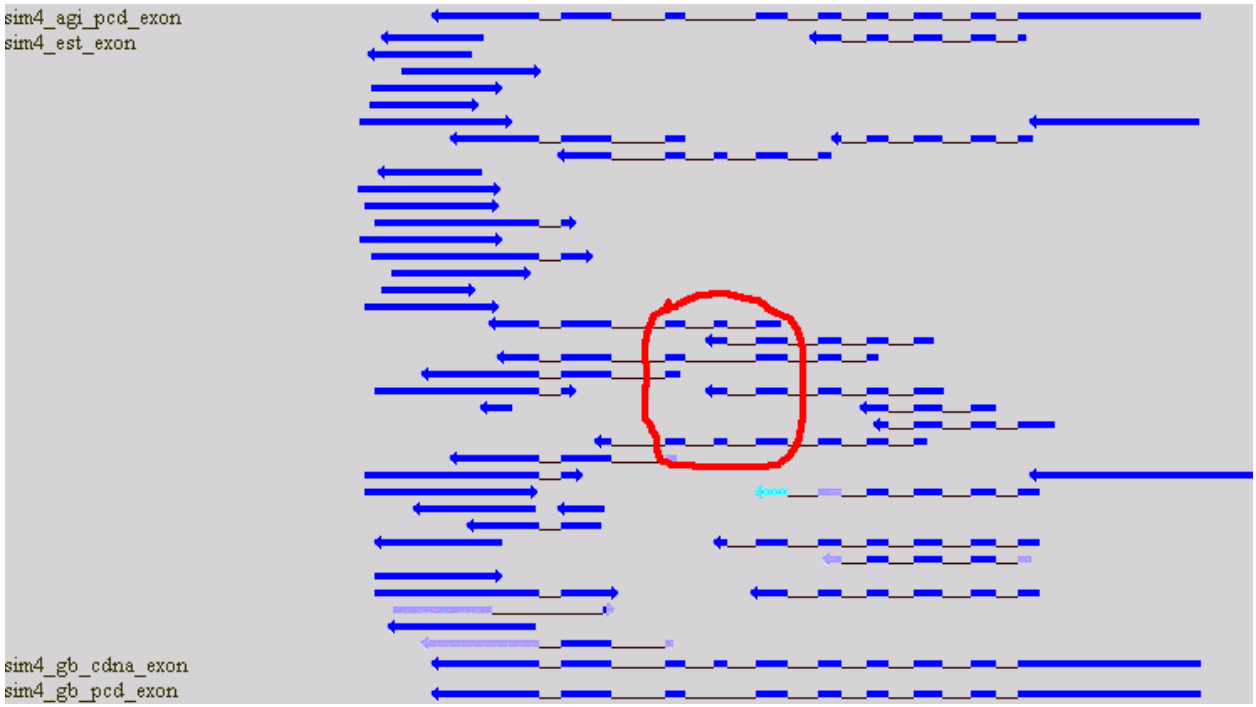


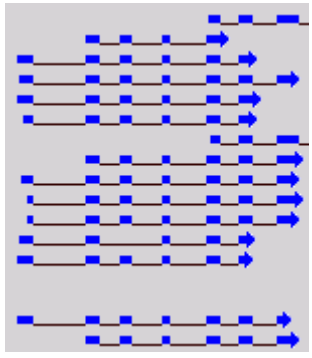


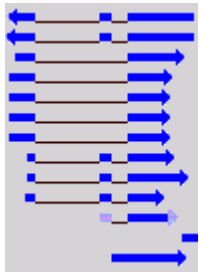


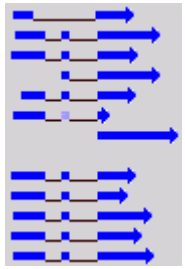


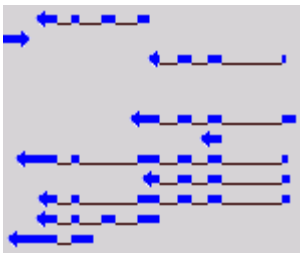


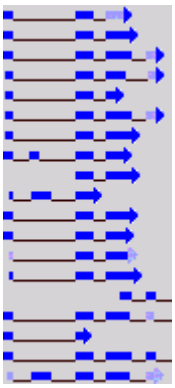

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_cdna_exon
sim4_gb_pcd_exon

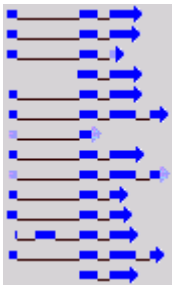


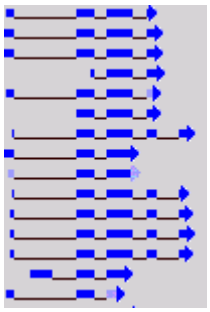


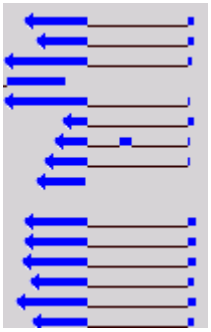


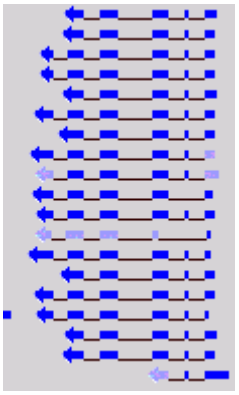


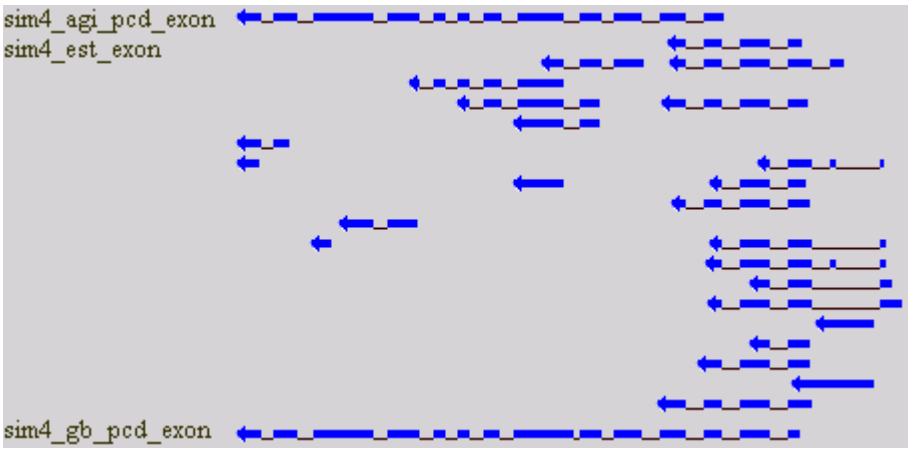


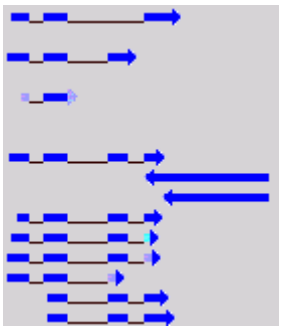


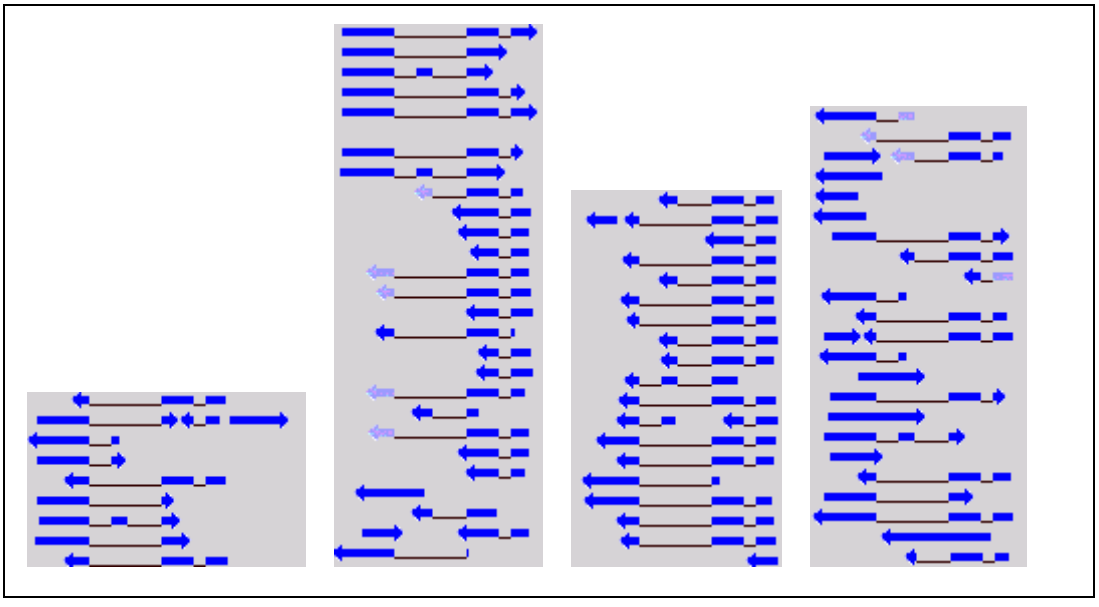


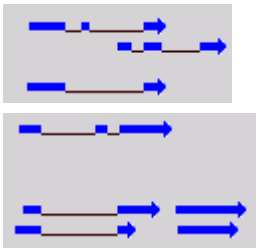


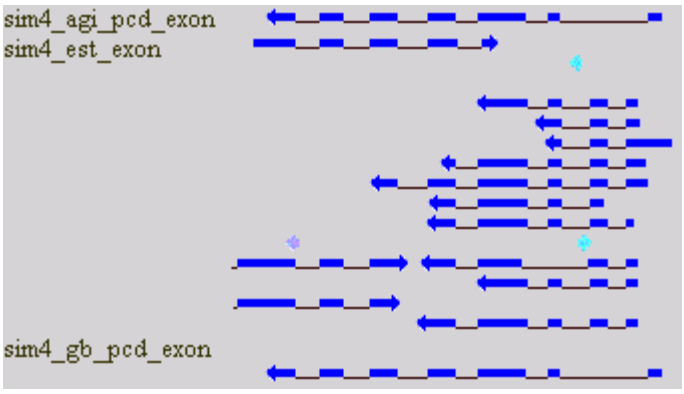


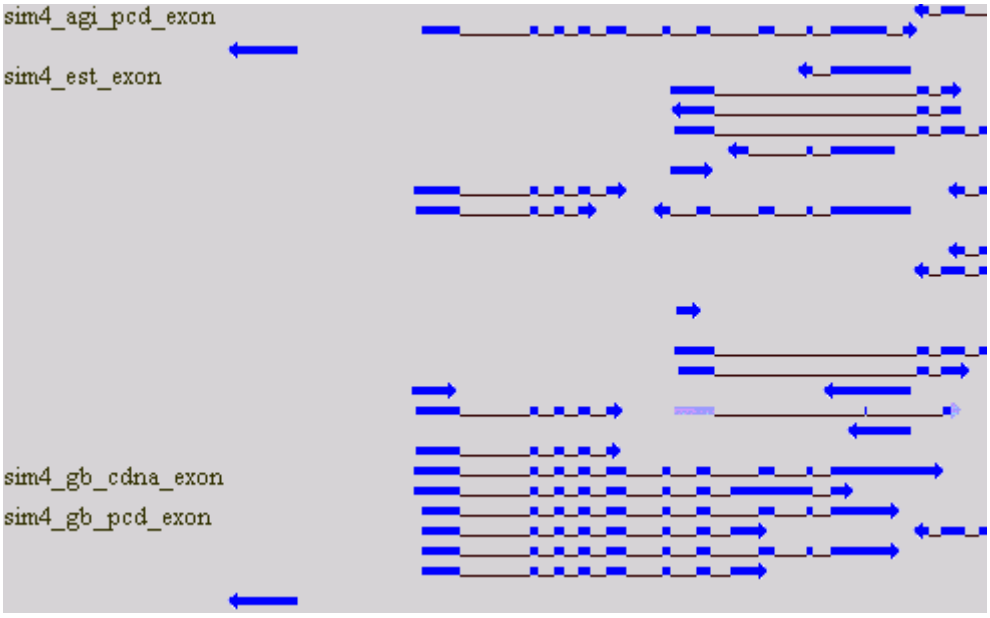

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

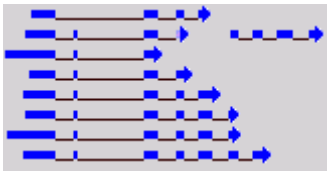


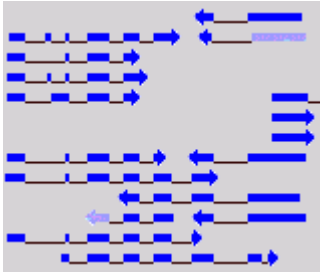


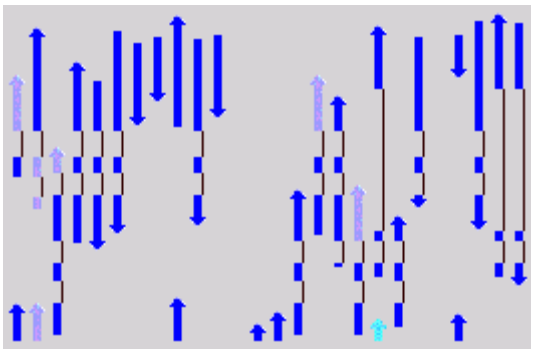


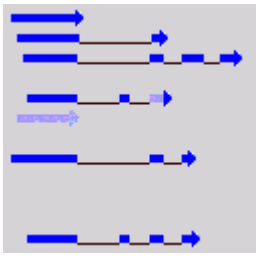

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

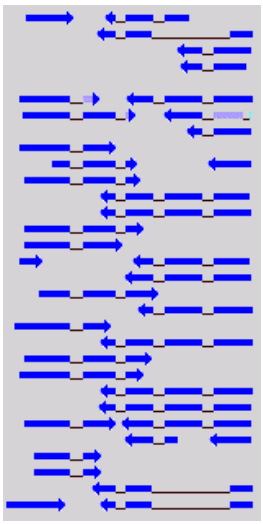

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_cdna_exon
sim4_gb_pcd_exon

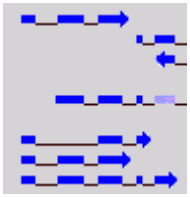


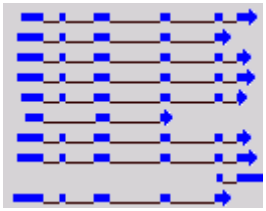


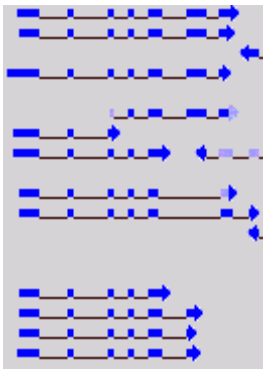


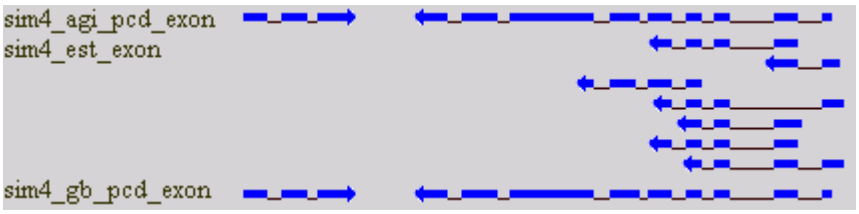


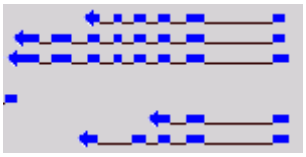


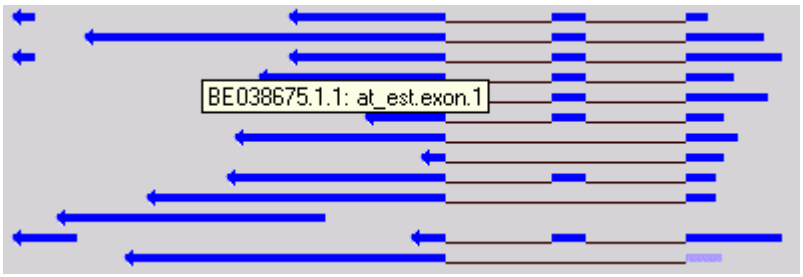


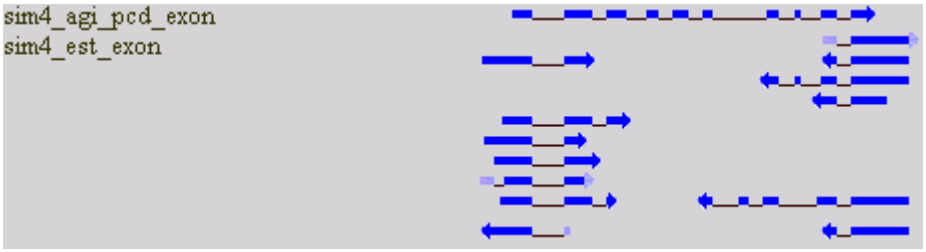


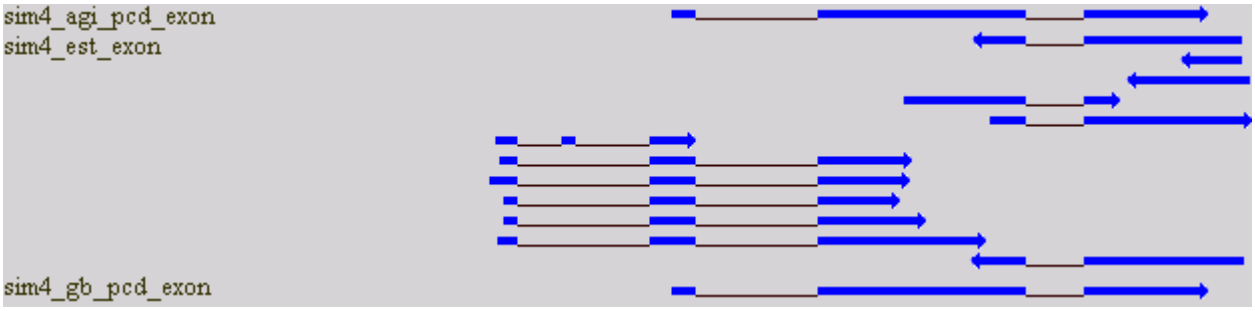


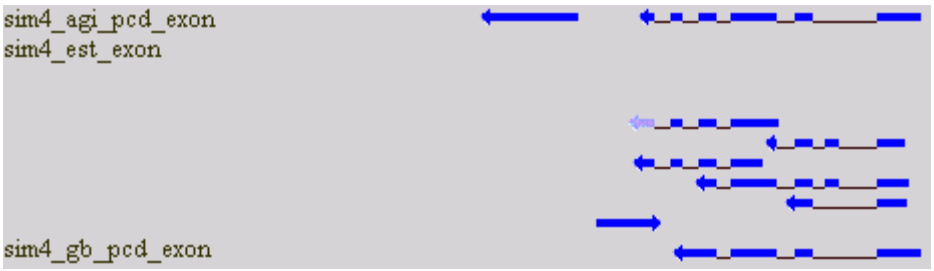

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

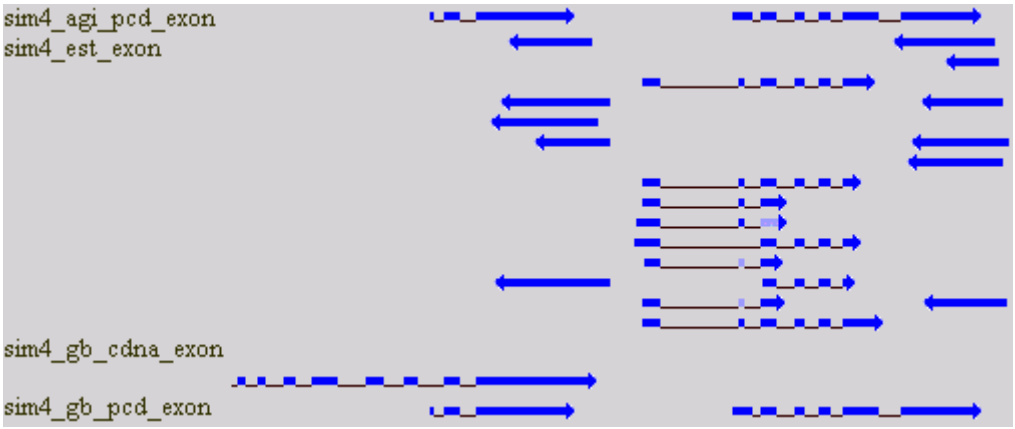


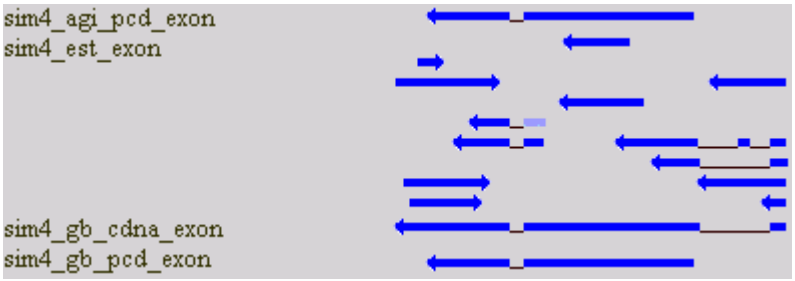

BE038675.1.1: at_est.exon.1

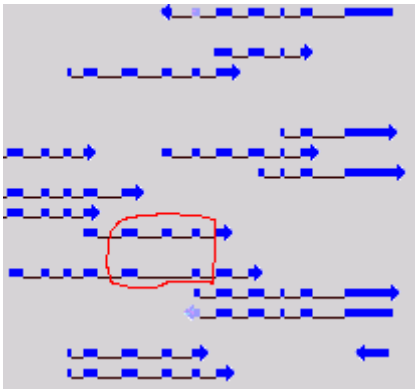

sim4_agi_pcd_exon
sim4_est_exon

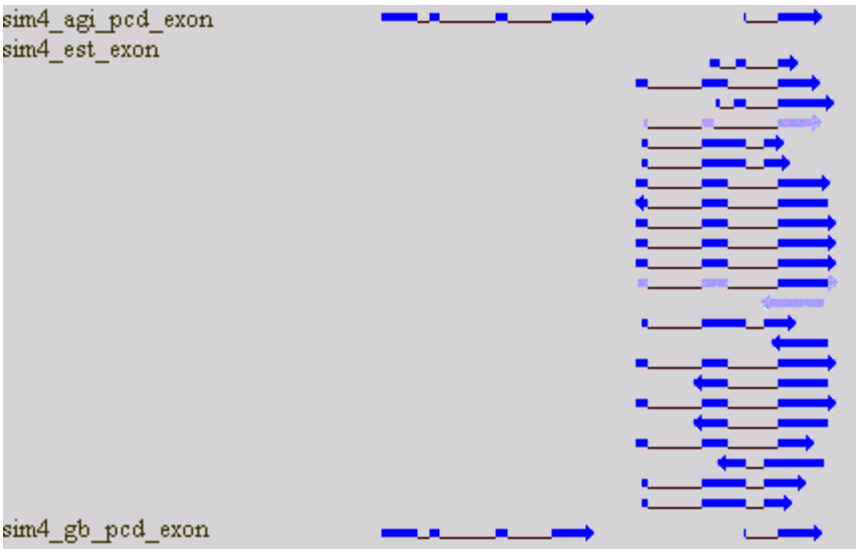

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

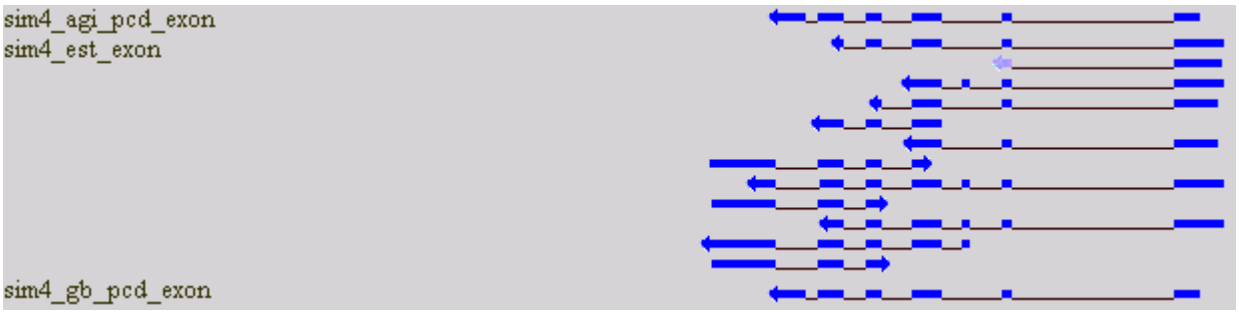

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

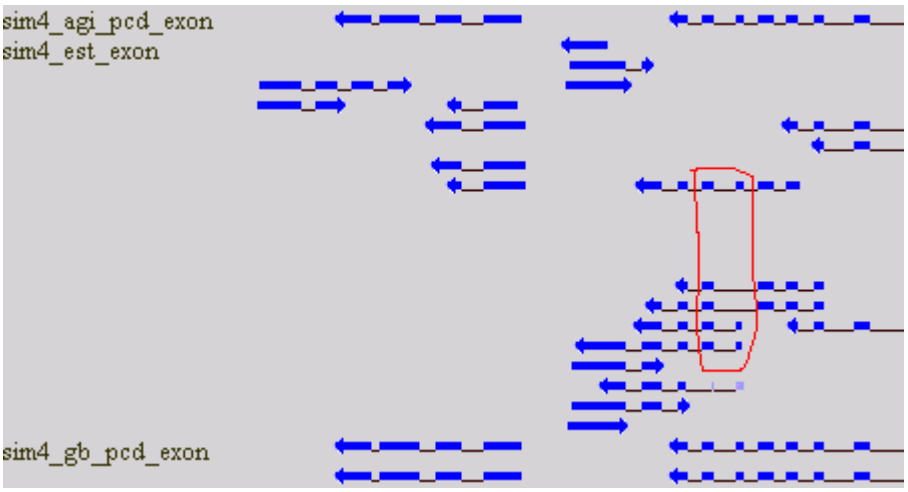

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_cdna_exon
sim4_gb_pcd_exon

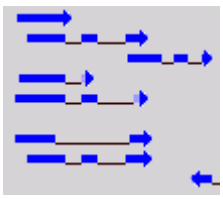

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_cdna_exon
sim4_gb_pcd_exon

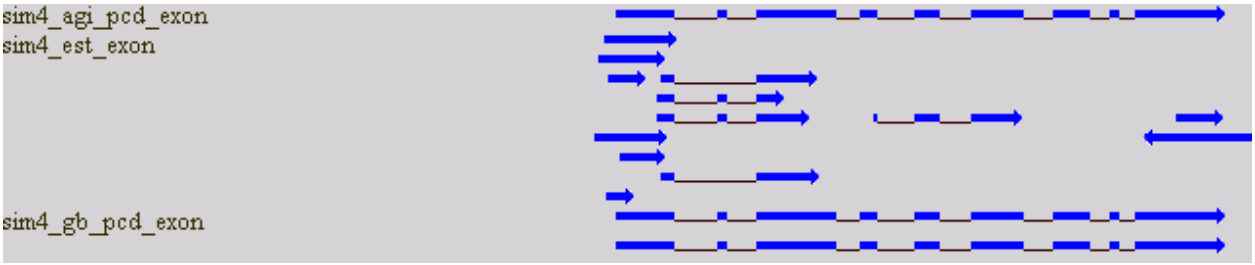


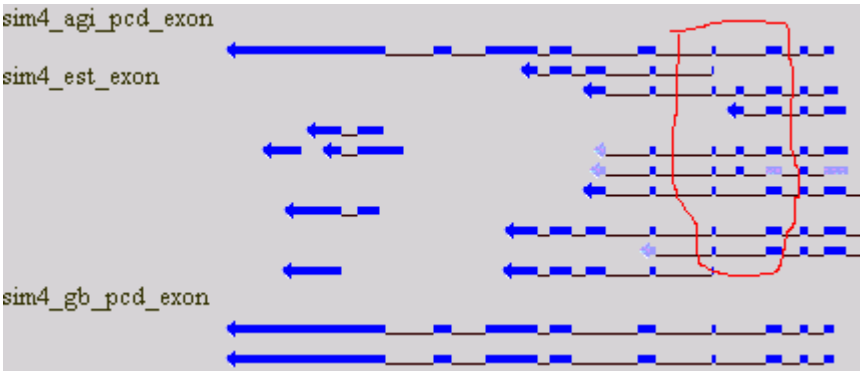

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

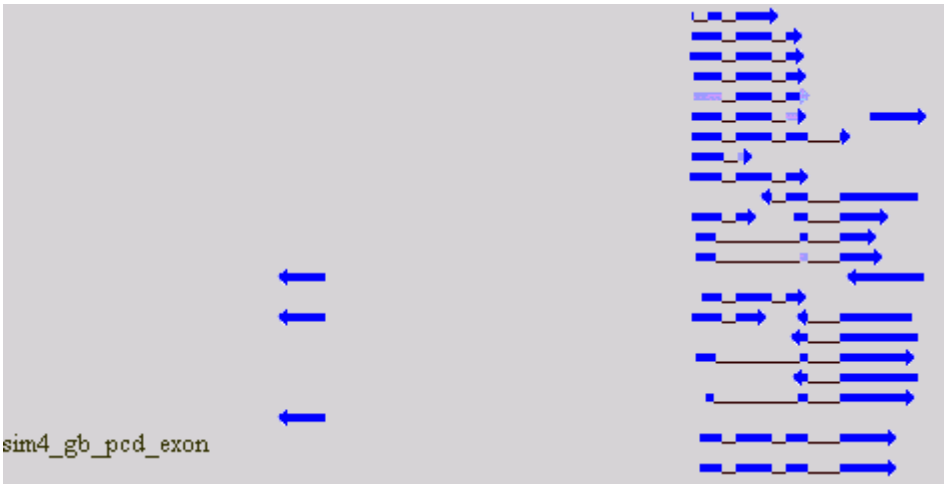

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

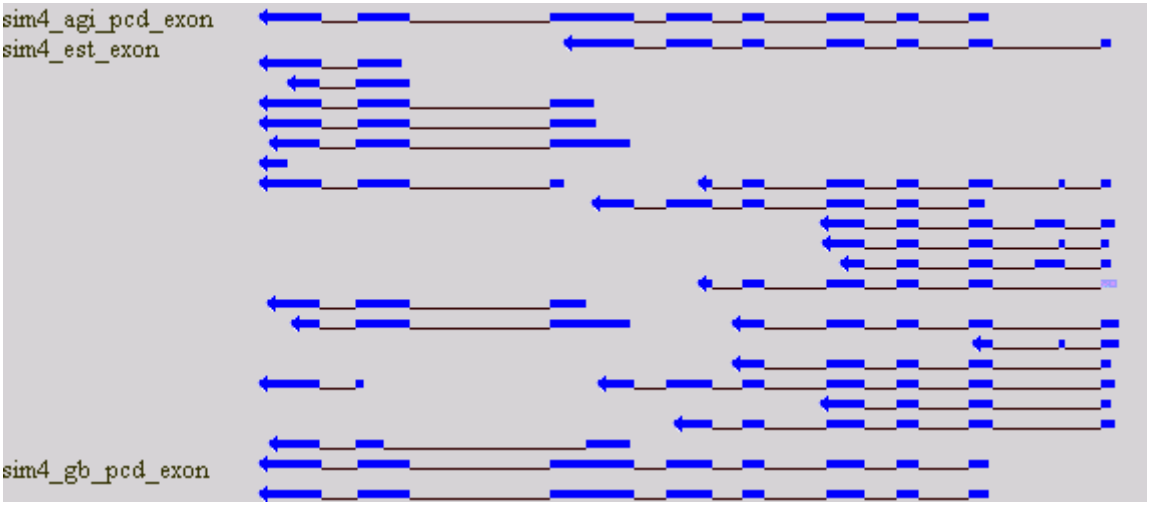

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

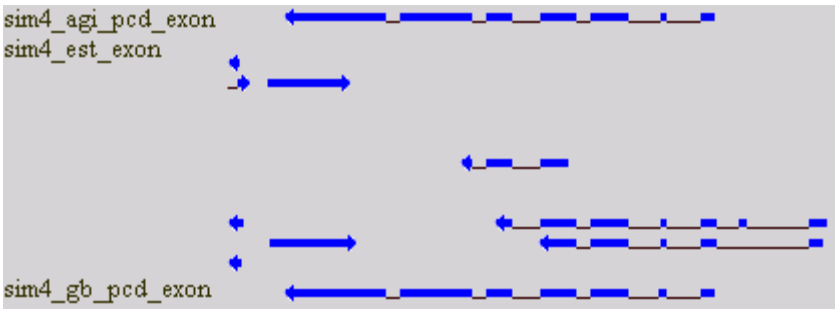


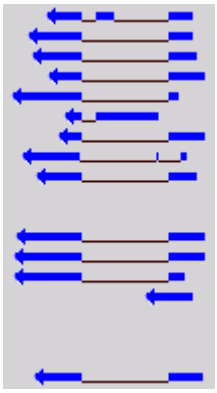

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

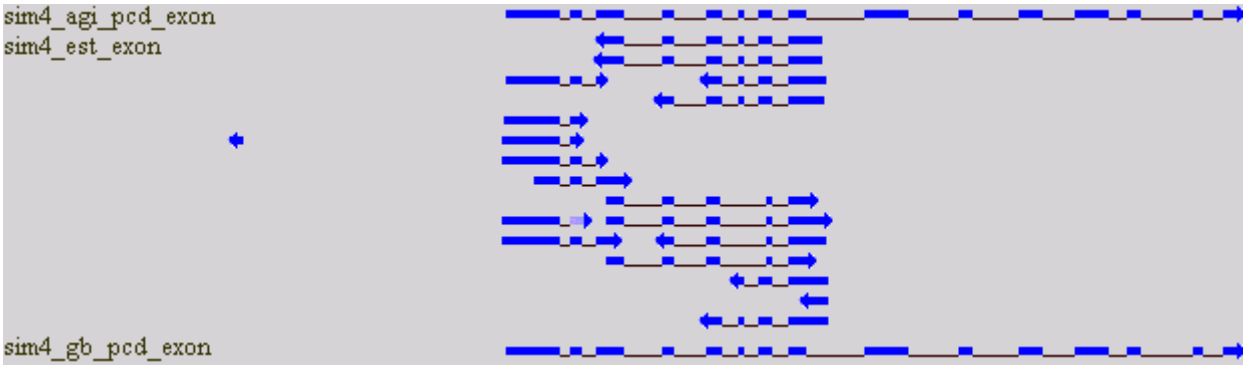

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon

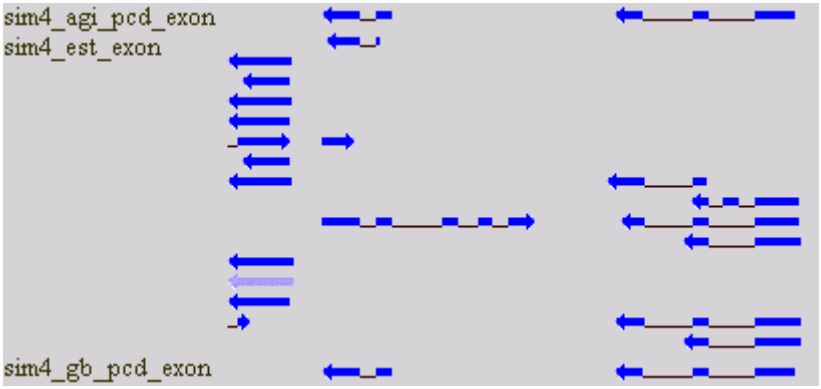

sim4_gb_pcd_exon

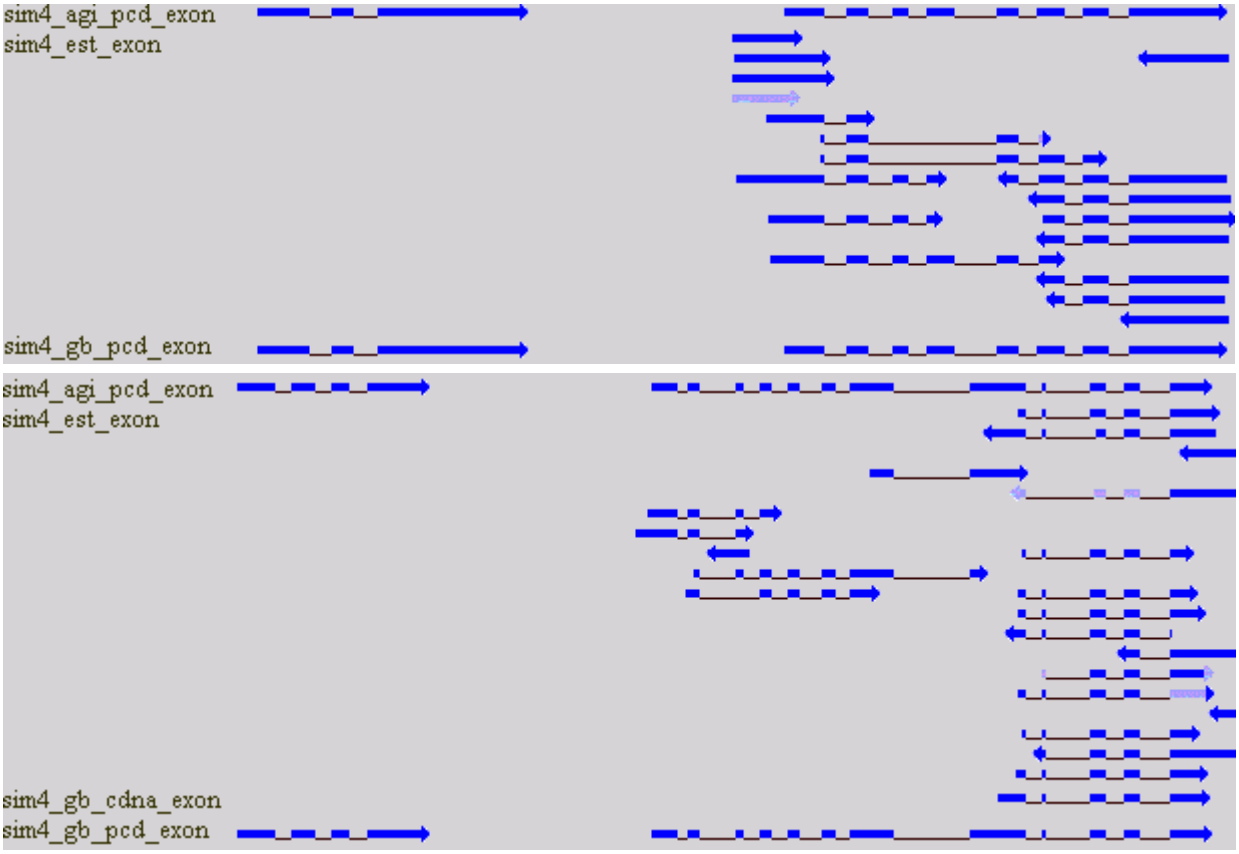

sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon


sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon




sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon


sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon


sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_pcd_exon
sim4_agi_pcd_exon
sim4_est_exon
sim4_gb_cdna_exon
sim4_gb_pcd_exon

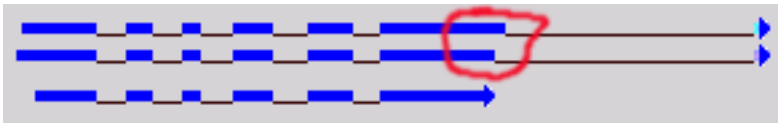
APPENDIX B: GENBANK (PUBLIC-DOMAIN) CDNA ALIGNMENTS WITH ARABIDOPSIS CHROMOSOME 4 (PUBLIC-DOMAIN): ALTERNATIVE SPLICING FALSE-POSITIVE ANALYSIS

This extremely informal appendix describes how we arrived at some of our filtering procedures, using public-domain examples.

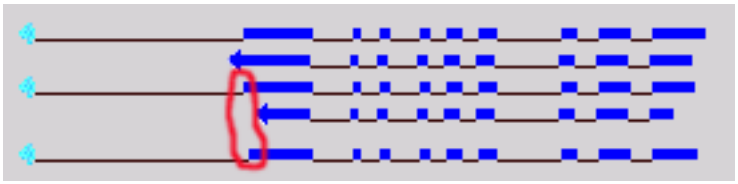
I. This is the most common false-positive case: a small number of bases are aligned with a terminal exon in some cDNA's and with an internal exon in others. Not surprisingly, sim4's authors report that the algorithm's biggest shortcoming is in aligning small terminal fragments, since splice site signals are misleading or not available in these cases. This sort of problem can probably be filtered out by:

- De-prioritizing or ignoring boundaries that are adjacent to low-quality exons
- Giving higher priority to boundaries that differ by more significant amounts (currently I report any boundaries that differ by more than 6bp, as we discussed)

Exon : C4 AF134129.1 5 5374131 5374510 bound : AF326900.1 5374477

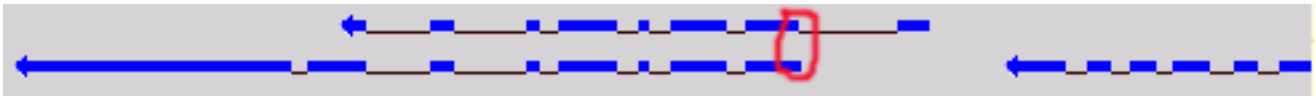


Exon : C4 AF098072.1 1 10734665 10735086 bound : AJ004881.1 10734691

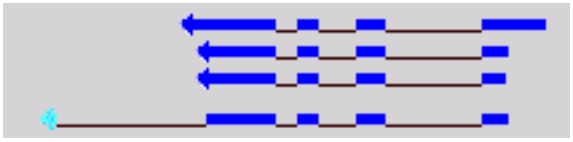


Exon : C4 Y13175.1 2 7466636 7467133 bound : AF043535.1 7467123

Exon : C4 AJ296155.1 7 11288523 11288775 bound : AF012658.1 11288767

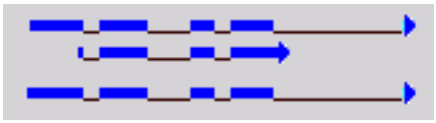


Exon : C4 AF325019.2 0 12880743 12881118 T bound : Z14990.1 12880810



Exon : C4 U62746.1 6 15639342 15639655 bound : AF031427.1 15639608

Exon : C4 D85193.1 3 15674744 15675029 bound : U78297.1 15674999

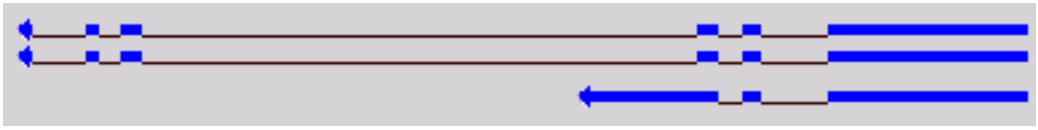


II. Two exons from two ESTs aligned perfectly, but one was very low-quality, so a very small gap was introduced right in the middle. This can be filtered out by sorting the output to give priority to boundaries that are adjacent to higher-quality sequences, and by de-prioritizing or ignoring very small introns.

Exon : C4 X69377.1 3 6986782 6987106 bound : AJ002892.1 6986867

III. This one is less obvious… I'm not actually sure how to explain the observed results here. The alignments were good, so maybe it's a possibility, but it's not the typical pattern that we're looking for. It's more likely that perhaps the third case should actually have aligned to a paralogous gene somewhere else. I would not propose that any specific algorithmic steps should be taken to avoid reporting and examining cases like this one.

Exon : C4 Z82991.1 0 8178099 8178581 T bound : Z82989.1 8178503



IV. As far as I know, simply skipping an intron is not part of the typical pattern of alternative splicing; the following cases are probably attributable to contamination by unspliced mRNA? I should be able to filter these out by ignoring or de-prioritizing cases in which a single exon overlaps an entire small intron.

It seems very odd that 2 of 3 sequences in the first case and 4 of 5 sequences in the third case would contain unspliced RNA in the same exon. Maybe the cDNA's that do show splicing at that site should actually have aligned to another genomic region, and we're just seeing two paralogous genes, one with a large deletion relative to the other? That would be odd, since these alignments represent the top BLAST hit for each cDNA, and it seems like a gap that size would lower the alignment score. But that still seems to be the best explanation.

Exon : C4 AF117063.1 4 8957147 8958005 bound : AJ002295.1 8957363

Exon : C4 U53865.1 1 10621910 10623221 bound : U56922.1 10622125

Exon : C4 X63157.1 3 11732322 11732591 bound : D31713.1 11732443

V. The remaining cases are the "miscellaneous" cases; the sort that are probably not incorrect according to any parameters that we will use to find alternative splicing events, but are clearly not cases of alternative splicing.

Exon : C4 AF325021.1 0 11212087 11212551 T bound : AF216387.1 11212438

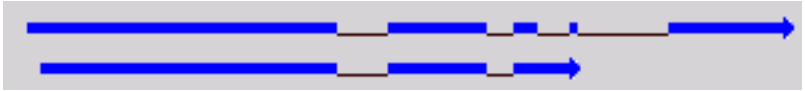
It's hard to believe that the bottom cDNA is for real, but I think it's also very difficult to rule this one out algorithmically. Something like this will probably have to remain a candidate until it can be manually investigated; I imagine that we'll find the lower cDNA to be a relatively low-quality BLAST hit.

Perhaps I should give lower priority to events arising from single-exon transcripts, even if they align well. This would be no big deal with the cDNA's, most of which have six or seven exons, but it might be a problem for the EST data set, where many of the transcripts don't cross an intron-exon boundary.

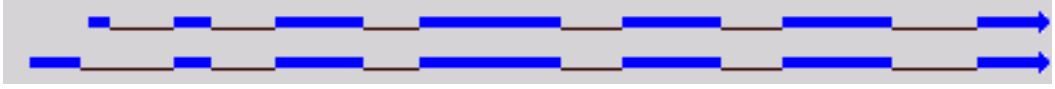
Intron : C4 AB049628.1 1 12017069 12018635 bound : AF062867.1 12017807

I'm actually surprised that there aren't more cases like this; this appears to be an inexplicable bad alignment. It clearly doesn't look much like alternative splicing. This is different from the first set of cases above in that although the intron in question is adjacent to a low-quality exon, the *boundaries* that caused this case to be flagged are boundaries with high-quality exons. Perhaps I should also give a low priority to cases in which a very long intron is the range that caused the event to be flagged.

Exon : C4 AF334210.1 2 15262333 15262500 bound : AF093546.1 15262405



Intron : C4 U70425.1 1 15804608 15804749 bound : AF034387.1 15804652



Exon : C4 AB031047.1 7 15923528 15923677 bound : AF322228.1 15923557