Genome-Wide Detection of Alternative Splicing in Expressed Sequences Using Partial Order Multiple Sequence Alignment Graphs

C. Grasso, B. Modrek, Y. Xing, and C. Lee

Pacific Symposium on Biocomputing 9:29-41(2004)

GENOME-WIDE DETECTION OF ALTERNATIVE SPLICING IN EXPRESSED SEQUENCES USING PARTIAL ORDER MULTIPLE SEQUENCE ALIGNMENT GRAPHS

C. GRASSO, B. MODREK, Y. XING, C. LEE

Department of Chemistry and Biochemistry,
University of California, 611 Charles E. Young Drive East,
Los Angeles, CA 90095-1570, USA
E-mail: leec@mbi.ucla.edu

We present a method for high-throughput alternative splicing detection in expressed sequence data. This method effectively copes with many of the problems inherent in making inferences about splicing and alternative splicing on the basis of EST sequences, which in addition to being fragmentary and full of sequencing errors, may also be chimeric, mis-oriented, or contaminated with genomic sequence. Our method, which relies both on the Partial Order Alignment (POA) program for constructing multiple sequence alignments, and its Heaviest Bundling function for generating consensus sequences, accounts for the real complexity of expressed sequence data by building and analyzing a single multiple sequence alignment containing all of the expressed sequences in a particular cluster aligned to genomic sequence. We illustrate application of this method to human UniGene Cluster Hs.1162, which contains expressed sequences from the human HLA-DMB gene. We have used this method to generate databases, published elsewhere, of splices and alternative splicing relationships for the human, mouse and rat genomes. We present statistics from these calculations, as well as the CPU time for running our method on expressed sequence clusters of varying size, to verify that it truly scales to complete genomes.

1 Introduction

Alternative splicing describes the process by which multiple exons can be spliced together to produce different mRNA isoforms, encoding structurally and functionally distinct protein products.[1,2] Recent studies have indicated that the mechanism of alternative splicing not only plays a large role in expanding the repertoire of gene function during the lifetime of an organism, but also facilitates the evolution of novel functions in alternatively spliced exons, which are less subject to the effects of natural selection.[3] Alternative splicing appears increasingly to make an important contribution to the complexity of the higher eukaryotes, by multiplying the number of gene products possible from the baseline number of genes. This issue has received much attention since the human genome (once estimated to contain up to 120,000 genes[4]) was reported to contain only ~32,000 genes.[5,6] Large-scale expressed sequence tag (EST) and mRNA sequencing has made possible multiple bioinformatics studies of alternative splicing.[5,7-11] In contrast with previous expectations that alternative splicing plays a relatively minor role in functional regulation (affecting perhaps 5 – 15% of genes), these EST-based studies have reported that alternative splicing is ubiquitous, observed in 40 – 60% of human genes.

While these results have aroused increasing interest in alternative splicing, there are many unanswered questions for the next phase of research. First of all, these studies were very different in their detailed methodology and results. For example, these methods divide into two very different camps. Some methods directly compare expressed sequences (ESTs and mRNA) to each other to identify divergent forms (insertions and deletions), which are interpreted as alternative splicing.[8,12] Other methods compare the expressed sequences individually to the genomic sequence to identify divergent patterns of exon inclusion.[7,10] These two approaches, which we will refer to as "EST Comparison" and "Genomic Mapping", cause very different patterns of false positive and false negative errors, and neither approach is by itself ideal. Second, EST-based alternative splice detection faces many fundamental technical challenges, concerning the experimental data, bioinformatics methods, and biological interpretation.[13] Thus, it is now essential to assess the key technical factors that determine the reliability of such alternative splicing analyses.[14] In this paper we present a detailed examination of the technical problems we have encountered in undertaking high-throughput analyses of alternative splicing over the last four years, and the specific solutions we have developed for these problems, in seeking to minimize both false positive and false negative errors.

2 Methods

2.1 Overview

In theory, detection of alternative splicing is straightforward: comparison of expressed sequences from a given gene can identify insertions and deletions that indicate alternative exon usage. In practice, however, this apparently simple task is complicated by serious technical problems that can produce artifacts resembling alternative splicing. The subtlety of these challenges is well illustrated by the question of whether to use EST Comparison vs. Genomic Mapping. As we will show in our analysis below, EST Comparison is vulnerable to a wide variety of problems (paralog mixing and genomic contamination, to name a few) that cause false positive errors (alternative splice predictions that are not reliable). However, this does not necessarily mean that Genomic Mapping is preferable. As we will show, Genomic Mapping not only raises many problems of computational load but also of accuracy, including significant false negatives. Thus, we have concluded that neither method is adequate by itself, and our approach combines *both* methods in an unusual hybrid approach.

A flow chart detailing our alternative splicing analysis is shown in Figure 1. Our analysis takes as input a single UniGene EST cluster[15] that contains both mRNA and EST sequences from a particular organism along with the organism's complete genome sequence. Our analysis produces as output a mapping of the cluster onto genomic sequence, a multiple sequence alignment of the set of expressed sequences aligned both to each other and to the genomic sequence, a set of detected splices stored as pairs of indices in the genomic sequence, and a set of alternative splicing relationship stored as pairs of indices of splice sites.

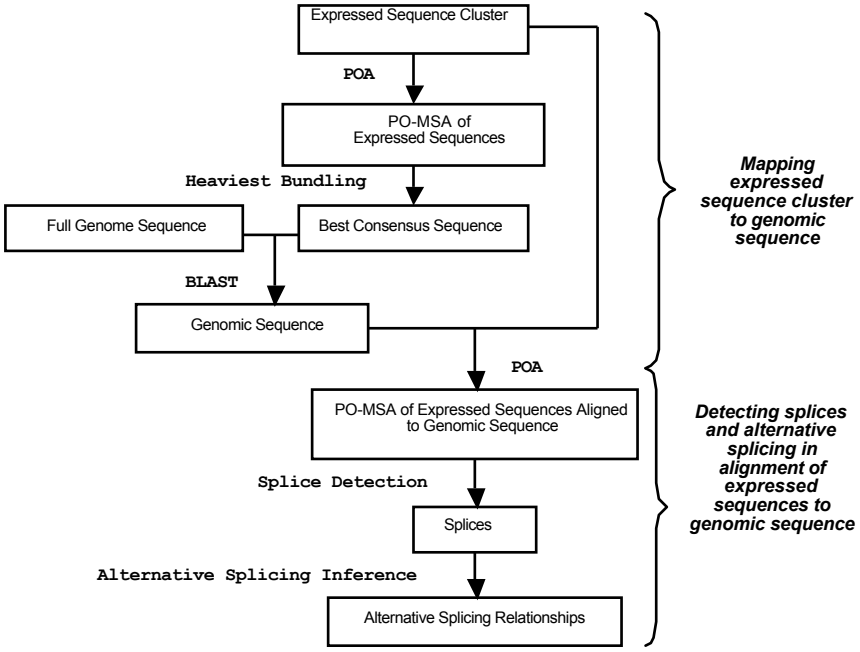

Figure 1. Flow-chart depicting our alternative splicing detection method.
Nodes are labeled with the input/output of each step in the method. Edges are labeled with the process undertaken at each step in the method.

2.2 Mapping the cluster of expressed sequences to genomic sequence

Extensive analysis of EST alignments has demonstrated that they are a valuable source of polymorphism identification, including SNPs and alternative splicing.[7,16] However, we have found that such analysis is very vulnerable to artifacts in both the experimental methods and the bioinformatics interpretation.[17] Since alternative splicing is identified in these alignments as large insertions and deletions, any artifact that gives rise to such differences in ESTs will cause false positives that can be difficult to screen out.[13] We have identified a number of such causes of artifacts. First, *genomic contamination* (EST library clones derived from genomic DNA rather than mRNA) and *incomplete mRNA processing* (clones derived from mRNA molecules whose splicing has not been completed) will produce the appearance of large insertions, due to retention of some intron sequences. Second, *paralog contamination* (mixing of ESTs derived from different, paralogous genes as a single EST cluster) can also produce the artifactual appearance of alternative splicing, which actually reflects differences between paralogous genes. Third, the EST data

are frequently massive and complex. For example, a UniGene cluster for a single gene can contain up to 5000 ESTs, far too large for most multiple sequence alignment programs to compare.

Genomic mapping provides an obvious solution to many of these problems, by permitting easy recognition of genomic contamination / intron retention, and verification of which gene a given EST is from.[7,10,13] When the complete genome sequence is available, it enables one to check definitively for the presence of possible paralogs, and to require that each EST match perfectly to its target gene (allowing for sequencing error) as a condition for inclusion in our calculation.[11]

On the other hand, attempting to map EST sequences directly to the entire genome itself poses serious problems. Because ESTs are short single-pass fragments and full of sequencing errors, BLASTing them individually against the genome sequence is both computationally expensive (e.g. for the human genome, 4 million ESTs vs. 3 billion bases of genomic sequence) and error prone, leading to a high false negative rate for splicing and alternative splicing detection. Matching a short, error-filled EST fragment against short genomic exons (150 nt on average, but can be as short as 10 nt) separated by large introns (from 1 kb up to >20kb) is very challenging, and both standard search programs (such as BLAST[18]) and multiple sequence alignment programs (such as CLUSTAL[19]) cannot guarantee reliable results.

To solve these problems, our method constructs a multiple sequence alignment (MSA) for the cluster of ESTs, extracts one or more "consensus" sequences that represent the aligned ESTs, and maps these consensus sequences to the genome using BLAST. The BLAST mapping step is straightforward, and has been described in detail.[11] However, the MSA and consensus construction steps pose significant new challenges. First, the large number of expressed sequences that must be aligned (up to 5000 ESTs in a single UniGene cluster) exceed the time and memory limitations of most MSA software.[12] To solve this problem, we use Partial Order Alignment (POA), whose time and memory requirements grow linearly with the number and length of EST sequences to be aligned.[20] POA can align 5,000 EST sequences in approximately 4 h on an inexpensive Pentium II PC. More importantly, POA generates the EST alignment as a graph structure that is able to represent both regions of match and regions of divergence: in regions where an EST matches other ESTs, it follows their path in the alignment graph; in regions where it diverges, it produces a new branch in the alignment graph. The alignment graph can accurately represent any level of complexity in the input sequence data: while a simple dataset of EST fragments of a single mRNA isoform would produce a single, linear path, a set containing a mix of ESTs from paralogous genes, genomic contaminants, or chimeric sequences would result in a branched alignment structure that reflects this complexity.

Moreover, this approach provides a natural, robust way for dealing with this complexity so that it does not cause artifacts in alternative splice detection. Specifically, we generate consensus sequence(s) by analyzing the Partial Order Multiple Sequence Alignment (PO-MSA) using the paralog separation algorithm of the Heaviest Bundling function of the POA program.[17] This method finds multiple

consensus paths through the PO-MSA graph, and then associates with each consensus path all of the expressed sequences which follow that path (with an allowance for sequencing error).[17] By separating ESTs that show signs of substantial divergence from the majority, POA's consensus generation is insulated from artifacts due to paralog mixing, genomic contamination, etc.

Ordinarily, Genomic Mapping confronts the twin difficulties of poor sensitivity and enormous inefficiency due to the high levels of sequencing error and redundancy in ESTs. Mapping individual ESTs is both harder (due to their short size and poor sequence quality) and very time consuming. We resolve both these issues by using the consensus sequences obtained from Partial Order Alignment. This both converts the EST data to reliable, assembled consensus sequence (greatly increasing sensitivity and robustness), and drastically reduces the number of search steps that must be performed. For large EST clusters (>100 ESTs) we have found this reduces the number of BLAST searches by 20 to 100-fold. In order to map the UniGene cluster to genomic sequence, we select the consensus sequence to which the majority of the expressed sequences have been bundled, since it most closely approximates a full-length mRNA transcript. The remaining consensus sequences, to which have been bundled the paralogous ESTs, chimeric ESTs, and mis-oriented ESTs that are not 90% identical to the majority consensus sequence, summarize the experimental and bioinformatics artifacts in the data.

To assess the value of using POA and Heaviest Bundling to cope with the complexity of the UniGene expressed sequence data, we have constructed the PO-MSAs of 80,000 Human UniGene clusters using POA, and run the Heaviest Bundling function to extract the minimum number of linear consensus sequences required to describe the aligned EST sequences to at least 90% identity. The number of consensus sequences generated by Heaviest Bundling is a useful measure of the degree of complexity of the data. For all Human UniGene clusters containing at least 10 ESTs, we counted the number of consensus sequences generated by Heaviest Bundling. Remarkably, a single consensus sequence was generated for only 16% of the Human UniGene clusters; two or three consensus sequences were generated for 41% of the clusters; four to ten consensus sequences were generated for 43% of the clusters. These data suggest that the large insertions and deletions in multiple sequence alignments of expressed sequence clusters, which result from experimental and bioinformatics errors, are not a minor phenomenon in the UniGene data, but are instead the norm. Their prevalence in the UniGene data necessitates the application of POA and Heaviest Bundling to the problem of mapping a UniGene cluster to genomic sequence.

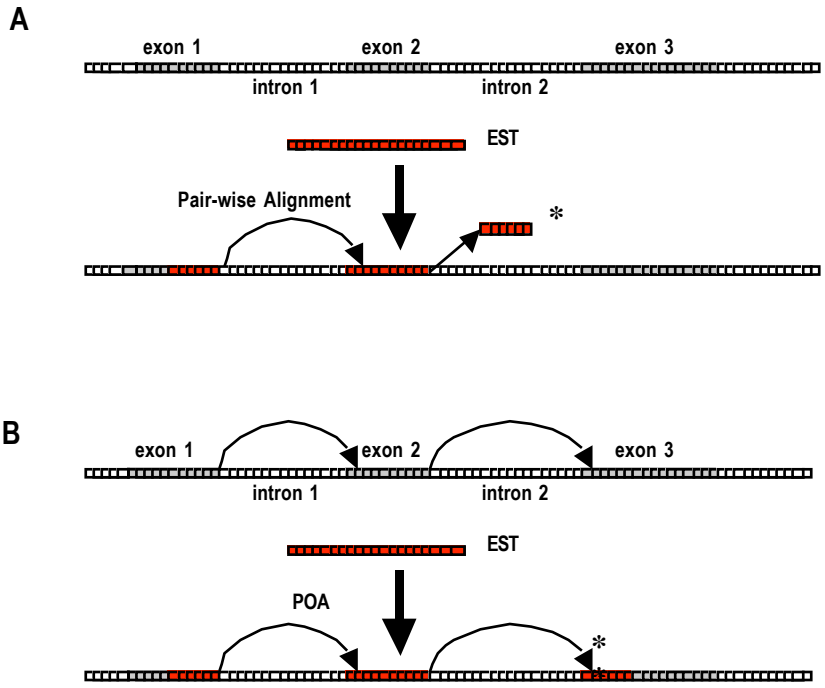

Figure 2. POA facilitates accurate alignment of EST fragments to genomic sequence.
In this figure, all alignments are represented as PO-MSAs, regardless of the manner in which they were constructed. The nodes in the PO-MSA are represented as squares and directed edges are shown only at branch points; nodes containing genomic sequence are white with grey nodes indicating exons, while nodes containing the EST sequence are always colored red. In A, the EST fragment cannot pay the gap penalty in order to align its last six nucleotides to exon 3, instead the six nucleotides are not aligned to genomic sequence at all (*) and so they do not provide evidence for the splicing of intron 2. In B, the EST fragment aligns to the PO-MSA containing multiple ESTs and mRNAs aligned to genomic sequence along the edge connecting exons 2 and 3. In this case, aligning the six nucleotides to exon 3 (**) does not require the payment of a large gap penalty and so the EST provides evidence for the splicing of intron 2.

2.3 Aligning expressed sequences to genomic sequence and to each other

Once a genomic location for an EST cluster has been identified, the method must next compare each EST to the genomic sequence to identify alternative exon usage. Once again, this apparently simple task is undercut by many technical difficulties. Whereas gene mapping only requires finding the right genomic region, reliable splice detection requires an exact, robust alignment of each EST to the genomic sequence. This is much harder to ensure. Whereas EST Comparison based methods rely on multiple sequence alignment, Genomic Mapping based methods

rely on pairwise alignment, i.e. aligning each individual EST to the genomic sequence. While pairwise alignments between full-length mRNA and genomic sequence are likely to be reliable, pair-wise alignments between EST fragments and genomic sequence are much more difficult to construct accurately, because ESTs are short, randomly fragmented, and full of sequencing errors. Figure 2A shows a pair-wise sequence alignment between an EST fragment and genomic sequence. The six nucleotides at the end of the EST fragment, which should align to the third exon in the genomic sequence, fail to do so because the score for perfectly matching them to genomic sequence is insufficient to compensate for the large gap penalty required to accommodate the large intron between exons 2 and 3. Instead, these six residues do not align to genomic sequence at all. Any attempt to detect splices on the basis of the resulting pair-wise alignment alone would fail to identify the splice that removes the second intron, resulting in a false negative.

Partial Order Alignment provides a systematic solution to this problem. As long as the PO-MSA contains at least one EST aligned across the gap, aligning a new EST can follow this path without any gap penalty. In this case, even the short EST fragment will align correctly across the gap from exon 2 to exon 3 (see Figure 2B).

The key difference here is that POA provides a hybrid method between conventional EST Comparison and Genomic Mapping: each EST is aligned not only against the genomic sequence, but also against the set of all previous ESTs at the same time, to identify the best scoring alignment path. In practice, we align full-length mRNA sequences to genomic sequence first, and then align EST sequences to the growing PO-MSA in order of decreasing length. This ensures that the evidence for splices, for which any sequence observation is able to pay the gap penalty, may be augmented by fragmentary sequence observations. In this way, our method is able to not only accurately align all EST fragments to genomic sequence, but also to combine the evidence for splicing from multiple ESTs.

This is valuable not only to rescue many EST splice observations that would normally be lost, but also to detect when several ESTs show a similar divergence from the genomic sequence (for example, indicating that they may actually be derived from a paralogous gene). These ESTs would be aligned to each other as a distinct path in the alignment, branching away from the genomic sequence. This information is then used to filter the set of ESTs that are retained for analyzing splicing. The detailed retention criteria have been previously described.[11]

2.4 Splicing and alternative splicing detection in PO-MSAs

Figure 3B shows the PO-MSA of all of the expressed sequences in human UniGene cluster Hs.1162 aligned to genomic sequence. Once the PO-MSA is constructed detecting splices amounts to finding large deletions in expressed sequences relative to genomic sequence. These deletions manifest themselves as directed edges in the

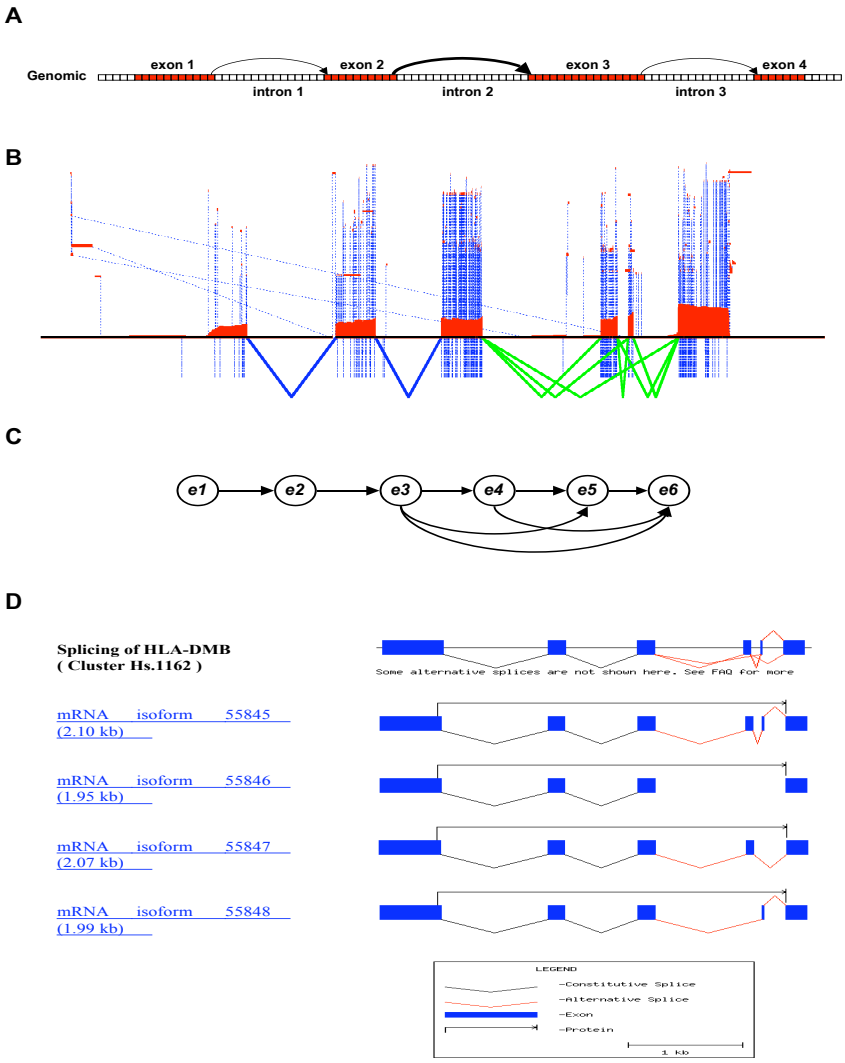

Figure 3. **Splicing and alternative splicing detection in a human expressed sequence cluster.** Figure A shows the PO-MSA of a single expressed sequence aligned to genomic sequence. The nodes in the PO-MSA are represented as squares and the directed edges are shown only at branch points; nodes containing genomic sequence are white, while nodes containing the expressed sequence are red. Figure B shows the PO-MSA graph, constructed by POA, of the ESTs and mRNAs in UniGene cluster Hs.1162 aligned to genomic sequence. In this visualization, rendered by the Partial Order Alignment Visualizer (POAVIZ),[16] nodes are red dots, directed edges are dotted blue lines, aligned nodes are adjacent to each other and genomic sequence is shown as a black line. Nodes and edges corresponding to insertions relative to genomic sequence are shown above the genomic sequence, and

appear as blue spikes topped with red dots, while the edges corresponding to deletions relative to genomic sequence are shown below, and appear solely as blue spikes. The number of expressed sequences aligned to a particular position in the genomic sequence is reflected in the thickness of red dots along the black line representing the genomic sequence so that the regions of genomic sequence corresponding to exons appear as red rectangles above the black line. The edges corresponding to splices are in bold, while edges corresponding to alternative splices are shown in green. Figure C shows the splicing graph constructed from the PO-MSA shown in B. Nodes correspond to exons, while edges correspond to splices. Figure D shows the graphical representation, provided by the ASAP database,[14] of the splicing and alternative splicing of the HLA-DMB gene inferred from the expressed sequences in human UniGene cluster Hs.1162 using our method. The four mRNA isoforms inferred from this data are shown as well.

alignment graph between a node containing genomic nucleotide *i* and a node containing genomic nucleotide *i*+*x*, where *x* is greater than 10 (Figure 3A). While one might be tempted to find these edges by depth first or breadth first search, the easiest way to find all of the large deletions in the expressed sequences relative to genomic sequence is to extract from the PO-MSA the set of pair-wise sequence alignments between each of them and genomic sequence and then analyze them directly. In order to verify each splice once it has been detected, we check not only the number of expressed sequences indicating the existence of the splice, but also whether the splice has valid intronic splice site sequences (GT / AG). Next, verified splices, which are stored as pairs of indices in genomic sequences, are compared to each other in order to infer alternative splice relationships. If the genomic sequence delimited by two of pairs of splices overlap, we identify the pairs as having an alternative splice relationship (Figure 4).

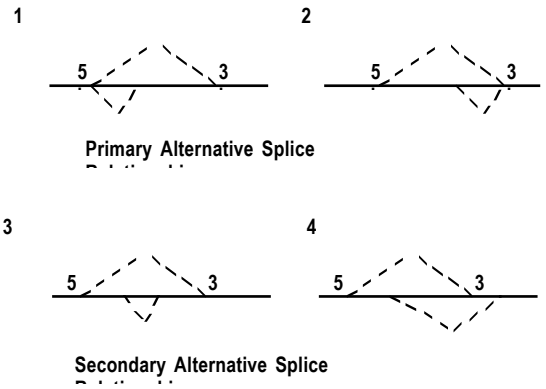

Figure 4: Filtering of alternative splicing by splice pair overlap relationships.
In all four diagrams, genomic sequence is shown as a straight black line, and splices are shown as dotted black lines. The 5' splice sites and 3' splice sites are labeled for the splice depicted above genomic sequence. The ASAP database reports alternative splice relationship types 1 and 2, which produce alternative 3' splicing, alternative 5' splicing, and exon skipping.

In order to filter out alternative splice relationships that are the result of genomic contamination, we identify only those alternative splicing relationships between mutually exclusive splices, i.e. pairs of splices whose 5' splice sites or 3' splice sites are the same, as valid. These valid alternative splice relationships are the basis on which we make inferences about alternative 5' splicing, alternative 3' splicing, and exon skipping. Figure 3B shows the eight splices and eight valid alternative splicing relationships detected in human UniGene cluster Hs.1162 using this method.

3 Results

We have applied our method to genome-wide detection of splicing and alternative splicing in the human, mouse, and rat genomes. This procedure is fully automated, can be applied to any genome, and its computation time scales linearly with the amount of EST data (Figure 5).

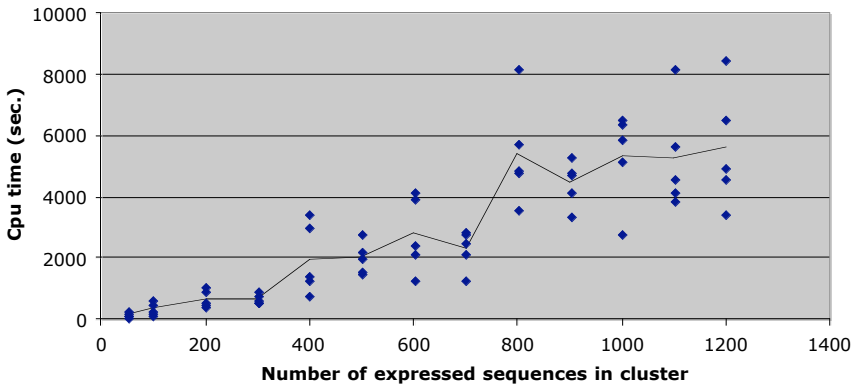

Figure 5: Total computation time as a function of increasing EST data.
For each number of sequences in the range of 50 to 1200, the cpu time was computed for five sequence clusters containing roughly the same number of expresses sequences. The black line plots the average of these five cpu times versus the number of sequences. These calculations were performed in an 1.4 GHz AMD Athlon running Linux.

Our method mapped 17,656 multi-exon genes to exact locations in the human genome (January 2003 data), 14,556 multi-exon genes in mouse, and 8,342 multi-exon genes in rat. In the human data, it detected over 35,000 alternative splicing relationships, more than doubling the number of predicted gene products versus the number expected from the estimated 32,000 human genes without alternative splicing. We detected a total of 115,518 splices and 35,433 alternative splice relationships, of these 30,891 were novel and 12,615 were novel and were

supported by multiple expressed sequences. Using the January 2002 mouse UniGene data we detected 91,225 splices and 12,528 alternative splice relationships, of these 11,687 were novel and 4,090 were novel and were supported by multiple expressed sequence observations. Using the January 2002 rat UniGene data we detected 31,177 and 1,143 alternative splice relationships, of these 11,687 were novel and 4,090 were novel and were supported by multiple expressed sequence observations.

Table 1: Total alternative splice detection in three genome-wide analyses.

Total	Human 1/02	Mouse 1/02	Rat 1/02	Human 1/03
Clusters	9610	8504	6158	11106
Clusters with a consensus sequence	9604	8387	5666	11092
Clusters mapped to genome	6801	5411	3958	6457
Splices detected	133369/ 1817	91225/ 1455	31177/834	115518/1765
Alternative splice relationships	30793/799	12528/489	1143/68	35433/783
Alternative splice relationships with multiple evidence	14656/520	4931/248	468/27	17157/530
Novel alternative splice relationships	26504/739	11687/469	919/581	30891/731
Novel alternative splice relationships with multiple evidence	10367/409	4090/217	244/16	12615/431

N.B. ratios are the number of splices of a particular type divided by the total number of clusters in which they occur.

4 Discussion and conclusions

We have presented a method for genome-wide detection of splicing and alternative splicing using expressed sequence data. We have demonstrated that this method can be run on a genome-wide scale, both by running it on the full human, mouse and rat genomes, and by assessing the cpu time required to run it on clusters with number of sequences ranging from 50 to 1200. We have also provided evidence that the partial order alignment algorithms are useful for coping with the true complexity of expressed sequence data, screening out experimental and bioinformatics artifacts in EST data that might cause spurious alternative splices. In addition, we have argued for the value of POA for simultaneously aligning expressed sequences to each other and to genomic sequence in order to effectively cope with EST fragmentation, which contributes to the loss of evidence for splicing and alternative splicing when short ESTs cannot be accurately aligned to genomic sequence.

While we have briefly explained the process by which we detect splicing and alternative splicing in the PO-MSA of all of the expressed sequences in a cluster

aligned to genomic sequence, we have not discussed the benefits of this approach. One of the major advantages of the PO-MSA representation is that its structure, which reflects exons, introns, and splices, can be easily abstracted as a splicing graph[21] (see Figure 3C). We have been able to exploit this feature in order to design algorithms for inferring full-length mRNAs isoforms either from the PO-MSA of the expressed sequences directly,[17] or from the splicing graph inferred from the PO-MSA of the expressed sequences aligned to genomic sequence.[22] The other major advantage of the PO-MSA representation is that it stores all of the evidence for a particular splice or alternative splice in a single data structure. This could be useful for calculating statistics measuring the evidence for a particular splice or alternative splice relationship from multiple EST and mRNA observations. By applying such methods, we would be able to associate lod scores with all of the splices and alternative splicing relationships in our datasets. These lod scores would be very useful for molecular biologists as they determine the direction of their expensive and time consuming experimental work in the area of alternative splicing.

Acknowledgments

C.G. was supported by a DOE Computational Science Graduate Fellowship; B.M. by NSF IGERT #DGE-9987641, and C.L. by NIMH / NINDS Grant #MH65166.

References

1. W. Gilbert, "Why genes in pieces?" *Nature* **271**, 501 (1978)
2. T. Maniatis and B. Tanis, "Alternative pre-mRNA splicing and proteome expansion in metazoans." *Nature* **418**, 236-243. (2002)
3. B. Modrek and C. Lee, "Alternative splicing in the human, mouse and rat genomes is associated with an increased rate of exon creation / loss." *Nature Genet.* **34**, 177-180 (2003)
4. F. Liang, et al., "Gene Index analysis of the human genome estimates approximately 120,000 genes." *Nature Genet.* **25**, 239-240 (2000)
5. I. H. G. S. Consortium., "Initial sequencing and analysis of the human genome." *Nature* **409**, 860-921 (2001)
6. D. Brett, et al., "Alternative splicing and genome complexity." *Nature Genet.* **30**, 29-30 (2002)
7. A. A. Mironov, J. W. Fickett and M. S. Gelfand, "Frequent alternative splicing of human genes." *Genome Res.* **9**, 1288-1293 (1999)
8. D. Brett, et al., "EST comparison indicates 38% of human mRNAs contain possible alternative splice forms." *FEBS Letters* **474**, 83-86 (2000)
9. L. Croft, et al., "ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome." *Nature Genet.* **24**, 340-1 (2000)

10. Z. Kan, E. C. Rouchka, W. R. Gish and D. J. States, "Gene structure prediction and alternative splicing analysis using genomically aligned ESTs." *Genome Res.* **11**, 889-900 (2001)
11. B. Modrek, A. Resch, C. Grasso and C. Lee, "Genome-wide analysis of alternative splicing using human expressed sequence data." *Nucleic Acids Res.* **29**, 2850-9 (2001)
12. J. Burke, H. Wang, W. Hide and D. B. Davison, "Alternative gene form discovery and candidate gene selection from gene indexing projects." *Genome Res.* **8**, 276-290 (1998)
13. B. Modrek and C. Lee, "A genomic view of alternative splicing." *Nature Genet.* **30**, 13-9 (2002)
14. Z. Kan, D. States and W. Gish, "Selecting for Functional Alternative Splices in ESTs" *Genome Res.* **12**, 1837-45 (2002)
15. G. Schuler, "Pieces of the puzzle: expressed sequence tags and the catalog of human genes." *J. Mol. Med.* **75**, 694-698 (1997)
16. K. Irizarry, et al., "Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences." *Nature Genet.* **26**, 233-236 (2000)
17. C. Lee, "Generating consensus sequences from partial order multiple sequence alignment graphs." *Bioinformatics* **19**, 999-1008 (2003)
18. S. F. Altschul, et al., "Basic local alignment search tool." *J. Mol. Biol.* **215**, 403-410 (1990)
19. J. D. Thompson, D. G. Higgins and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Res.* **22**, 4673-80 (1994)
20. C. Lee, C. Grasso and M. Sharlow, "Multiple sequence alignment using partial order graphs." *Bioinformatics* **18**, 452-464 (2002)
21. S. Heber, et al., "Splicing graphs and EST assembly problem." *Bioinformatics* **18 Suppl. 1**, S181-8 (2002)
22. Y. Xing, A. Resch and C. Lee, "The Multiassembly Problem: reconstructing multiple transcript isoforms from EST fragment mixtures," submitted.