Title

An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs.

Journal

Nucleic acids research, 34(10)

ISSN

0305-1048

Authors

Xing, Yi
Yu, Tianwei
Wu, Ying Nian
et al.

Publication Date

2006

DOI

10.1093/nar/gkl396

Peer reviewed

3150–3160 *Nucleic Acids Research, 2006, Vol. 34, No. 10*
doi:10.1093/nar/gkl396

An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs

Yi Xing[1,*], Tianwei Yu[2,3], Ying Nian Wu[2], Meenakshi Roy[1], Joseph Kim[1] and Christopher Lee[1,*]

[1]Molecular Biology Institute, Center for Computational Biology, Department of Chemistry and Biochemistry, University of California, Los Angeles, USA, [2]Department of Statistics, University of California, Los Angeles, USA and [3]Dental Research Institute, School of Dentistry, University of California, Los Angeles, USA

Received January 29, 2006; Revised April 13, 2006; Accepted May 10, 2006

ABSTRACT

Reconstructing full-length transcript isoforms from sequence fragments (such as ESTs) is a major interest and challenge for bioinformatic analysis of pre-mRNA alternative splicing. This problem has been formulated as finding traversals across the splice graph, which is a directed acyclic graph (DAG) representation of gene structure and alternative splicing. In this manuscript we introduce a probabilistic formulation of the isoform reconstruction problem, and provide an expectation-maximization (EM) algorithm for its maximum likelihood solution. Using a series of simulated data and expressed sequences from real human genes, we demonstrate that our EM algorithm can correctly handle various situations of fragmentation and coupling in the input data. Our work establishes a general probabilistic framework for splice graph-based reconstructions of full-length isoforms.

INTRODUCTION

Alternative splicing is a widespread mechanism of gene regulation in higher eukaryotes (1–3). It refers to the production of different mRNA transcripts from a single gene through alternative combinations of exons or alternative selections of splice sites (4). Among multi-exon genes in the human genome, it is estimated that as many as 74% are alternatively spliced (2). Alternative splicing can impact important functional regions of the proteins, such as protein interaction domains (5) and structural elements (6,7), leading to multiple protein products with distinct functions. It can be regulated in a tissue-specific (8) or developmental-dependent (9) manner. Aberrant alternative splicing is a major cause for many human diseases (10).

There are two types of experimental data for detecting alternative splicing: full-length sequences and sequence fragments. Sequencing of cDNAs can reveal the complete set of splicing events of full-length transcripts, leading to discoveries of splice variants (11). By contrast, most high-throughput genomic technologies [such as shotgun sequencing of ESTs (12) and oligonucleotide microarrays (13)] produce information about sequence fragments. The vast majority of alternative splicing events in the human and other eukaryotic genomes are discovered using fragmentary sequence data, such as ESTs (1) and microarray probe signals (2). ESTs are short fragments of full-length mRNA sequences (12). To date over 6.5 million human ESTs have been deposited in the UniGene database (compared to ~200 000 full-length mRNAs, see the statistics at http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=9606), which are extremely useful for discovering mRNA diversity, such as alternative splicing. However, as sequence fragments, ESTs only provide partial information on the complete gene structure (e.g. how 2 exons are combined). It is difficult to infer the likely functional impact of a splice variant, without the knowledge of its full-length transcript and protein product (14). In fact, over 80% of alternative splicing events in the human transcriptome are detected from EST sequences, with no corresponding full-length transcript or protein sequences available in sequence databases, such as GenBank (15). This problem becomes even more complicated in splicing microarrays, which return highly fragmentary information from probes targeting specific exons or exon–exon junctions (2,16,17). For these reasons, the computational reconstruction of full-length transcript isoforms from sequence fragments becomes an important problem and challenge for bioinformatic analyses of alternative splicing (18).

*To whom correspondence should be addressed. Tel: +1 310 825 7374; Fax: +1 310 206 7286. Email: leec@mbi.ucla.edu
*Correspondence may also be addressed to Yi Xing. Email: yxing@ucla.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

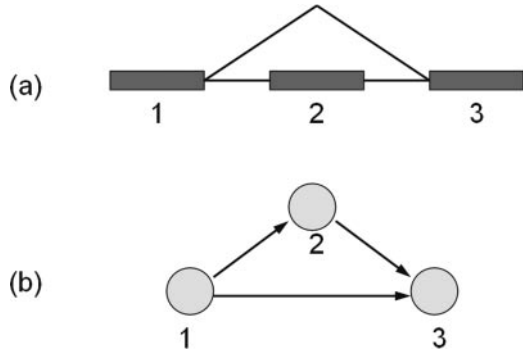

Figure 1. Splice graph of a three-exon alternatively spliced gene. (**a**) Gene structure for a three-exon gene. The second exon is a cassette exon. (**b**) The splice graph representation of the gene structure. The exon skipping event is represented by a directed edge from node 1 to node 3.

Over the years, there has been a large body of work on this isoform reconstruction problem (15,19–28). The centerpiece of these studies is a graph representation of gene structure and alternative splicing, often referred to as the 'splice graph'. In a pioneering study, Heber and colleagues introduced the concept of splice graph (23), which represented exons as nodes and splicing events as directed edges (see Figure 1 for a simple illustration). Different types of alternative splicing events, such as exon skipping, alternative donor/acceptor splice sites, mutually exclusive exon usage, intron retention, can be easily represented using splice graphs (e.g. see the representation of exon skipping events in Figure 1). The problem of isoform reconstruction can be formulated as a splice graph traversal problem (15,23,29). Due to alternative splicing, multiple traversals exist, corresponding to multiple isoforms of the gene (see Figure 1). Some methods enumerate all possible traversals across the splice graph (23,24). For genes with multiple alternatively spliced regions, such methods potentially generate a large number of random exon combinations. Several other methods use certain rules to produce a minimal set of traversals sufficient to explain all the input data (15,20–22). In some cases the results of isoform reconstruction are dependent on the order of sequence observations in the input data (15,18).

Despite the vital contributions by these studies, several aspects of isoform reconstruction need to be improved. First, many genes have multiple alternatively spliced regions and complex patterns of alternative splicing (30,31). Multiple alternative splicing events in a single gene can be regulated in a highly coordinated fashion as shown experimentally in fibronectin, and other genes (32). One classic example is the alternative splicing of CD44. Only ~20 isoforms have been observed for CD44, despite the possibility of ~1000 random combinations of cassette exons (33,34). It is estimated that coupling of alternative splicing exists in at least 25% of alternatively spliced genes (32). In principle a splice graph traversal algorithm based on dynamic programming [such as the Heaviest Bundling algorithm of ours (15)], which makes local choices for finding the optimal traversal, cannot guarantee the correct treatment of coupled edges in the splice graph. Second, not all traversals across the splice graph are equally likely. It is important for the isoform reconstruction algorithm to reflect the strength of evidence for a particular splice graph traversal. An essential question is how to weigh evidence from sequence fragments, which might be consistent with a large number of traversals of the splice graph. These issues point to the need for an explicitly probabilistic approach to the isoform reconstruction problem.

In this manuscript, we introduce a probabilistic formulation of the isoform reconstruction problem, and provide a solution based on the maximum likelihood principle. EM algorithm (35) has been used in many areas of computational biology, such as haplotype inference (36), evolutionary selection pressure (e.g. Ka/Ks) estimation (37), predictions of domain–domain interactions (38) and motif detection (39), etc. We describe an EM algorithm to estimate the probability for each traversal across the splice graph, which maximizes the total likelihood of the observed input data. A large body of work has used expressed sequences to reconstruct full-length isoforms (15,19–28), or to quantify alternative splicing on the exon level (8,40–42). Consistent with these previous studies, here we use a series of simulated sequence observations and expressed sequences of real human genes to demonstrate our probabilistic reconstruction of full-length isoforms from splice graphs. We want to emphasize that our method is not limited to the analyses of EST data. A variety of high-throughput genomic technologies [such as mass spectrometry (43), microarray (2), massively parallel signature sequencing (MPSS) (44), etc] produce fragments that can be used to detect alternative splicing. The goal of this manuscript is to establish a general probabilistic framework for splice graph-based reconstruction of full-length isoforms.

MATERIALS AND METHODS

Enumeration of putative isoforms from the splice graph

A splice graph is a directed acyclic graph (DAG), whose nodes represent exons and edges represent splicing events (23) (see Figure 1). We enumerated all possible traversals across the splice graph using a breadth-first-search (BFS) algorithm. These traversals corresponded to putative isoforms of a gene. We compared each sequence observation with each putative isoform to derive their consistency relationship. A sequence observation was defined as being consistent with a putative isoform if it was fully contained in the putative isoform. We constructed an indicator matrix, which recorded the consistency relationship of all sequence observations with all putative isoforms (see details below).

Probabilistic formulation and EM algorithm

Multinomial model with uncommitted categorization. Suppose there are K possible isoforms for a gene. Let's denote them by $I_1, I_2, \ldots, I_K$. For each sequence observation, the probability that it is generated by isoform I_k is p_k, where $k = 1, \ldots, K$ and $p_1 + p_2 + \ldots + p_K = 1$. This probability model is called multinomial model.

Suppose we observe N sequence observations. Let's denote them by $O_1, O_2, \ldots, O_N$. We can use an $N \times K$ indicator matrix $\mathbf{Z} = (z_{i,k})_{i=1,\ldots,N,\ k=1,\ldots,K}$ to record the categorizations of these sequence observations. Specifically, if the ith sequence observation is generated by isoform I_k, then $z_{i,k} = 1$; otherwise, $z_{i,k} = 0$.

The probabilities (p_k, $k = 1, \ldots, K$) can be estimated by isoform proportions. Specifically, we count the number of

sequence observations that are generated by isoform k, i.e. $n_k = \sum_{i=1}^{N} z_{i,k}$, then the estimated probability $\hat{p}_k = n_k/N$.

The complication is that most sequence observations are not of full-length, so that they are consistent with more than one isoform. Therefore, the indicator matrix $\mathbf{Z}$ is not fully observed. What is observed is another indicator matrix $\mathbf{Y} = (y_{i,k})_{i=1,\ldots,N;k=1,\ldots,K}$, where $y_{i,k} = 1$ if the ith sequence observation is consistent with isoform I_k, and $y_{i,k} = 0$ otherwise. Unlike matrix $\mathbf{Z}$, which has one and only one 1 in each row, the matrix $\mathbf{Y}$ has one or more 1's in each row. If $y_{i,k} = 0$, then $z_{i,k}$ must be 0, but if $y_{i,k} = 1$, then $z_{i,k}$ may or may not be 1. We call $\mathbf{Y}$ the uncommitted categorization, and $\mathbf{Z}$ the underlying committed categorization.

We denote $\theta = (p_1, \ldots, p_K)$. The log likelihood function of the multinomial model with uncommitted categorization is

$$l(\theta \mid \mathbf{Y}) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} y_{i,k} p_k\right).$$

The maximum likelihood estimates (MLE) of θ, $\hat{\theta} = \arg\max_{\theta} l(\theta \mid \mathbf{Y})$, cannot be obtained in closed-form.

EM: soft categorization and fractional counts. The EM algorithm (35) can be used to compute the maximum likelihood estimates (MLE) of the category probabilities $\theta = (p_k, k = 1, \ldots, K)$ from the observed data $\mathbf{Y}$. The EM algorithm is an iterative algorithm. In describing the algorithm, we add subscript $\theta^{(t)}$ to the relevant quantities. For instance, $\theta_{(t)}$ is the parameter value computed after t th iteration. The algorithm starts from an initial guess $\theta^{(0)} = (p_k^{(0)}, k = 1, \ldots, K)$. For instance, $p_k^{(0)} = 1/K$. Each iteration is a mapping from $\theta^{(t)}$ to $\theta^{(t+1)}$, which is accomplished via the following two steps:

E-step:

$$z_{i,k}^{(t+1)} = E[z_{i,k} \mid Y_i, \theta^{(t)}] = \Pr(z_{i,k} = 1 \mid Y_i, \theta^{(t)})$$
$$= \frac{y_{i,k} p_k^{(t)}}{\sum_{k=1}^{K} y_{i,k} p_k^{(t)}}, \forall i, k.$$

M-step: Let $n_k^{(t+1)} = \sum_{i=1}^{N} z_{i,k}^{(t+1)}, \quad \forall k,$

$$p_k^{(t+1)} = \frac{n_k^{(t+1)}}{N}, \quad \forall k.$$

Intuitively, the E-step splits the ith sequence observation into different isoforms. Each isoform k with $y_{i,k} = 1$ gets a fraction of O_i in proportion to $p_k^{(t)}$, and this fraction is $z_{i,k}^{(t+1)}$. We call this soft categorization. The M-step updates the probability of each isoform by counting the sequence observations that are categorized into this isoform. Because of soft categorization, we have to sum up the fractional indicators as if they were 0/1 indicators.

We can run this algorithm until $\mid \theta^{(t+1)} - \theta^{(t)} \mid = \sum_{k=1}^{K} \mid p_k^{(t+1)} - p_k^{(t)} \mid < \varepsilon$, where ε is a pre-specified stopping criterion. Throughout this manuscript, we used an ε of 10^{-6}.

The algorithm is justified in Appendix 1 of the online supplements. The standard errors of the parameter estimates can be obtained after EM converges. See Appendix 2 of the online supplements.

Table 1. The simulation study to test the robustness of the EM algorithm

Isoform	Probability in set a	Probability in set b	Probability in set c	Probability in set d
1-2-3-4-5-6-7-8-9-10	0.25	0.5	0	0.95
1-3-4-5-6-7-8-9-10	0.25	0	0.5	0
1-2-3-4-5-6-7-8-10	0.25	0	0.5	0
1-3-4-5-6-7-8-10	0.25	0.5	0	0.05

Simulation study

We performed a simulation study to examine the robustness of our EM algorithm against sampling biases in the input data. EST generation is subject to two major sources of variation: (i) isoform sampling based on isoform abundance and (ii) random fragmentation of isoforms. These variations translate into the variation in the MLE estimates of the probabilities. We simulated a gene with 10 exons. The terminal exons were 250 bp in length. All the internal exons were 150 bp. Exon 2 and 9 were alternatively spliced cassette exons. Four isoforms can be generated from this gene: the isoform with all exons, the isoform without exon 2, the isoform without exon 9, and the isoform without exon 2 and 9 (see Table 1).

We set certain fixed probabilities for these isoforms (see Table 1). We generated a simulated expressed sequence using the following three-step procedure: (i) randomly sample the four isoforms, to generate a full-length mRNA, (ii) randomly sample the empirical distribution of the length of ESTs (taken from human UniGene data, see Supplementary Data), to decide the length of the simulated expressed sequence and (iii) randomly truncate the simulated mRNA according to the length obtained from the previous step, to make the simulated expressed sequence. If the length was longer than the mRNA, the whole mRNA was the simulated expressed sequence.

After gathering N ($N = 10$, 25, 50, 100 and 250) simulated expressed sequences, we ran the EM algorithm to obtain the estimates of probabilities. The deviation of the estimated probabilities from the truth was measured by the total variation distance, which is $\frac{1}{2}\sum_{i=1}^{4} \mid \hat{p}_i - p_i \mid$. For each N, we repeated the process 100 times and inspected the distributions of the total variation distance.

Probabilistic isoform reconstruction using human expressed sequence data

For a human gene, we aligned all of its mRNA and EST sequences to its genomic sequence using POA (45), and calculated the exons and splicing events from the multiple sequence alignment (15), using June 2003 download of UniGene data and human genome sequences. We constructed the splice graph based on observed splicing events among its exons in the expressed sequence data. Details of the splice graph construction were described in Xing *et al*. (15). We enumerated all possible traversals across the splice graph as the total set of putative isoforms, using a BFS algorithm. We used the EM algorithm described above to estimate the probability of each isoform.

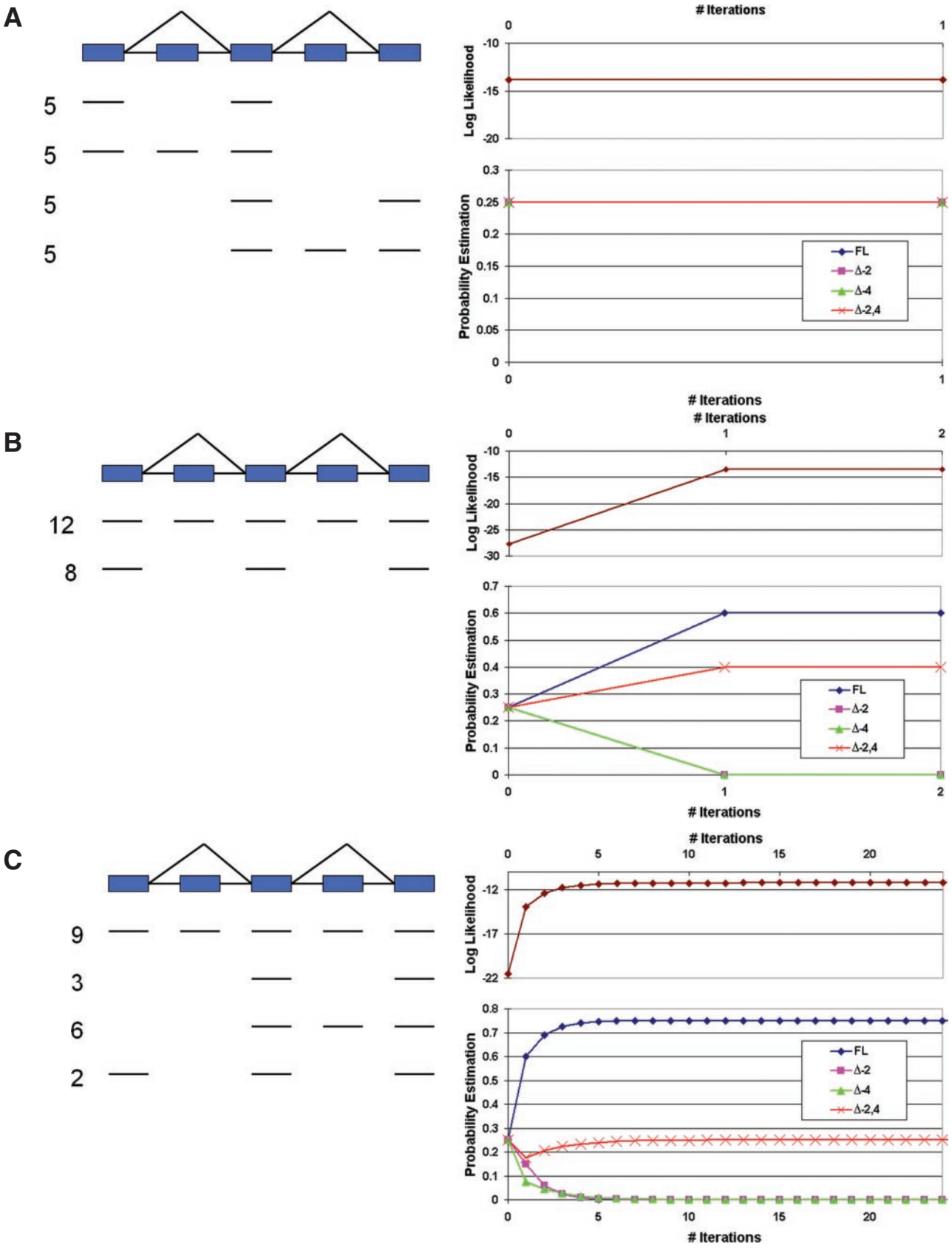

Figure 2. Probabilistic isoform reconstruction for a simulated gene. Left panel: sequence observations; right panel: results of probabilistic isoform reconstruction. The upper graph indicates the overall log likelihood; the lower graph shows the estimated probabilities for individual isoforms until convergence. **A**–**D** represents different situations of sequence observations.

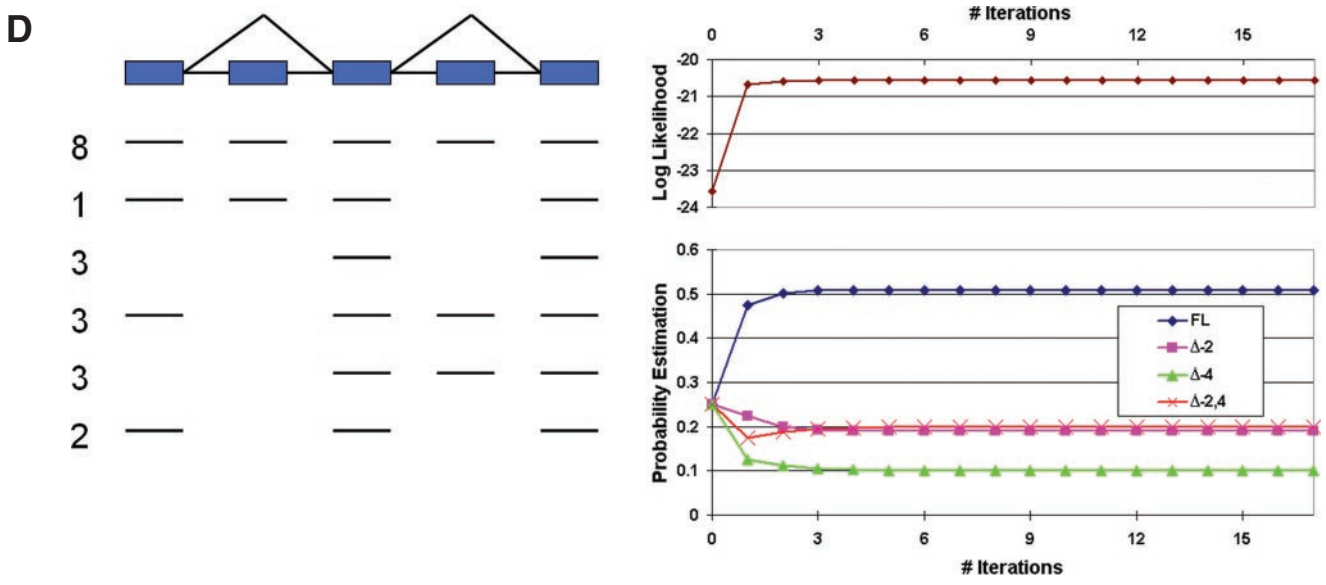

Figure 2. Continued

RESULTS

Probabilistic isoform reconstruction for a simulated gene

To test and illustrate the behavior of the EM algorithm, we designed a simulated gene with 5 exons. Exon 2 and 4 were cassette exons that could be skipped entirely. An exhaustive traversal of this splice graph produced four isoforms for this gene: the one that contained all 5 exons (FL), the one that lacked exon 2 (Δ−2), the one that lacked exon 4 (Δ−4) and the one that lacked both exon 2 and 4 (Δ−2, 4). Of course, the result of our probabilistic isoform reconstruction should be dependent on the actual sequence observations for this gene.

We designed four different sets of sequence observations, and ran the EM algorithm on each of them to infer isoform probabilities. Each set contained 20 sequences, and represented distinct situations of sequence observations. Details of these dataset are illustrated in Figure 2 (left panels).

In sequence set 1, each sequence indicated the inclusion or skipping of one cassette exon (exon 2 or exon 4). No sequence covered both alternatively spliced regions. Five sequences contained exon 1 and 3, indicating the skipping of exon 2; five sequences contained exon 1, 2 and 3, indicating the inclusion of exon 2. These ten sequences gave no information about the inclusion or skipping of exon 4. Similarly, five sequences indicated the inclusion and another five sequences indicated the skipping of exon 4, but carried no information on the status of exon 2. In summary, this sequence dataset represented a situation where all the input sequences were highly fragmented. There was no indication on how cassette exon 2 and 4 were combined in the final transcripts. Our EM algorithm converged at equal probabilities for all possible isoforms (see Figure 2A, right panel).

Sequence set 2 represented a completely opposite situation. Every sequence in set 2 was a full-length sequence. Twelve sequences contained all 5 exons, and eight sequences contained only exon 1, 3 and 5. Obviously, although we could produce four different traversals across this splice graph, only two traversals were observed in our input data and were necessary to explain all the 20 sequences. In fact, the skipping events of exon 2 and exon 4 were always coupled in the input data. Initiating from equal probabilities (0.25) for all isoforms, our EM algorithm converged at 0.6 for the FL and 0.4 for the Δ−2,4 isoform, and 0 for the Δ−2 and Δ−4 isoform (see Figure 2B). The result of our EM algorithm was consistent with the proposed strategy of producing a minimal set of isoforms to explain all the sequence observations (15,21,22).

Sequence set 3 contained both full-length sequences and sequence fragments (see Figure 2C, left panel). It was easy to notice that six sequence fragments (which contained exon 3, 4 and 5) were consistent with both FL and Δ−2 isoforms. During the iterations, the categorization of these six sequences to the FL and Δ−2 isoform was gradually assigned to the FL isoform, due to additional sequence evidence (nine sequences) that only supported the FL isoform. The algorithm converged at 0.75 for the FL and 0 for the Δ−2 isoform. A similar situation occurred during the probability estimation for the other two isoforms: Δ−2,4 received a probability of 0.25 while Δ−4 received a probability of 0 (see Figure 2C, right panel).

In sequence set 4, all four isoforms were observed in the input data. Our EM algorithm converged at non-zero probabilities for all isoforms (see Figure 2D).

The results of our probabilistic isoform reconstruction on this simulated gene indicate that our EM algorithm can correctly handle various situations of fragmentation and coupling in the sequence data. It produces the most likely probability estimate for each isoform by considering all the sequence observations simultaneously. Unlike our previous method (15), this probabilistic reconstruction is completely independent of the order of input sequence observations.

Simulation study to test the robustness of the EM algorithm

Sampling biases in the input data can affect the maximum likelihood estimates of the isoform probabilities. In particular, when the number of sequence observations for a gene is relatively small, the MLE might have considerable deviations from the true isoform probabilities. To assess the impact of sampling biases, we performed a simulation study on a simple gene model with 10 exons. We used four sets of probability distributions of isoforms (see Table 1), and then simulated the process of random EST generation (see Materials and Methods). For each probability distribution we randomly generated 10, 25, 50, 100 and 250 sequences (see Materials and Methods).

We compared our MLE estimates to the true probabilities to calculate the total variation distance (see the definition in Materials and Methods). When the total number of sequence observations was only 10, there was a considerable difference between the estimated isoform probabilities and the true probabilities. By contrast, with 250 sequence observations, these two sets of probabilities were fairly close (see Figure 3A). An additional simulation that randomly initiated the true probability distribution of all isoforms produced a similar result (see Figure 3B). As expected, our simulation study indicates that the accuracy of our estimates is dependent on the number of sequence observations for a gene.

Isoform reconstruction using expressed sequence data for HLA-DMB and TPM1

Figure 4A shows the gene structure for a well-studied gene *HLA-DMB*. *HLA-DMB* plays an important role in antigen presentation and the activation of the humoral immune response, by facilitating the loading of class II MHC molecules with exogenous peptide antigens (46–48). This process occurs in early lysosomal compartments. *HLA-DMB* has 6 exons. Exon 4 encodes a hydrophobic transmembrane (TM) domain, and exon 5 encodes a lysosomal targeting (LT) signal. Our EST-genome alignment revealed alternative splice forms that skipped exon 4 or exon 5 of *HLA-DMB*, or both. Therefore in addition to the FL form of *HLA-DMB*, expressed sequence data indicate three additional isoforms that lack the TM domain (Δ-TM), or the LT signal (Δ-LT), or both (Δ-TM and LT). By excluding likely EST artifacts from unspliced genomic DNA (1), we used 97 cDNA/EST sequences as the input data for our probabilistic isoform reconstruction. Our calculation shows that the FL *HLA-DMB* is the predominant isoform for this gene, with an estimated probability of 0.732. The three shorter *HLA-DMB* isoforms are present at a much lower level (see Figure 4C). The probability estimates of these *HLA-DMB* isoforms agree well with our RT–PCR analysis of *HLA-DMB* isoforms (Figure 4D; see Appendix 3 of the online supplements for details on RT–PCR analyses and sequencing of *HLA-DMB*).

We also reconstructed isoforms for Tropomyosin 1 (*TPM1*). *TPM1* has multiple regions with complicated patterns of alternative splicing. Even after we excluded likely EST artifacts, the remaining sequence observations still gave rise to a splice graph with 16 possible traversals. Previous studies of *TPM1* alternative splicing revealed two major alternative splicing events: the mutually exclusive usage of exon 6a and 6b; the alternative use of exon 9/10 or exon 11 as the 3′ terminal exons (see Figure 5A). There is evidence in the expressed sequence data that alternative splicing of these two regions is not independent (see Figure 5B). In fact, quantitative PCR analyses of *TPM1* alternative splicing showed that the inclusion of exon 6b together with exon 9/10 was the predominant isoform in muscle (49). On the other hand, the inclusion of exon 6a and exon 11 was predominant in non-muscle tissues (49). Although such coupling information was present in the input sequence observations, our previous isoform generation method failed to capture such a signal for coupled alternative splicing events (15), demonstrating the limitation of dynamic programming in recognizing coupled edges in the splice graph. By contrast, our EM algorithm converged at five isoforms with an inferred probability of at least 0.05. The two isoforms that received the highest probability estimates corresponded to the muscle-specific and non-muscle isoforms of *TPM1* (see Figure 5C). This analysis demonstrates that our EM algorithm provides a global, probabilistic solution that can deal with coupling events in the splice graph.

Isoform reconstruction for genes on human chromosome 22

To assess the computational feasibility of our algorithm, we performed isoform reconstruction for 186 alternatively spliced genes from human chromosome 22. We recorded the computation time on a PC (AMD Athlon 1500+ with 320MB of RAM) for each gene. The result is plotted in Figure 6. The average CPU time is 6.7 seconds. The maximal CPU time is 298.1 seconds for Hs.26593 (HDAC10), which has 768 possible traversals across its splice graph. This analysis demonstrates that our algorithm is computationally feasible.

One important and difficult aspect of isoform reconstruction is to assess the confidence in different traversals of the splice graph, by integrating full-length sequences and sequence fragments. A widely used approach in non-probabilistic isoform reconstruction method is to evaluate isoforms based on numbers of consistent sequence evidence. For each of the 186 UniGene clusters on chromosome 22, we scored its putative isoforms using (i) probability estimates from our EM algorithm; (ii) numbers of consistent sequence evidence. We classified all putative isoforms into two categories: those with mRNA evidence, and those with only EST evidence. The first category was treated as goldstandard high-confidence isoforms, since the presence of mRNA evidence was widely used as a criterion for real isoforms, as opposed to rare spliceosomal errors or EST artifacts (1). For UniGene clusters on chromosome 22 with both mRNAs and ESTs, we summed the scores for both categories of putative isoforms separately and calculated their ratio. Using probability estimates as the scores, the overall ratio (isoforms with mRNA evidence versus isoforms with only EST evidence) is 4.5. By contrast, scoring by numbers of consistent sequence evidence yielded an overall ratio of 1.1. This analysis indicates that a higher probability out of our EM algorithm is correlated with a higher confidence in the putative isoforms. Isoform evaluation by EM has a more significant tendency of favoring

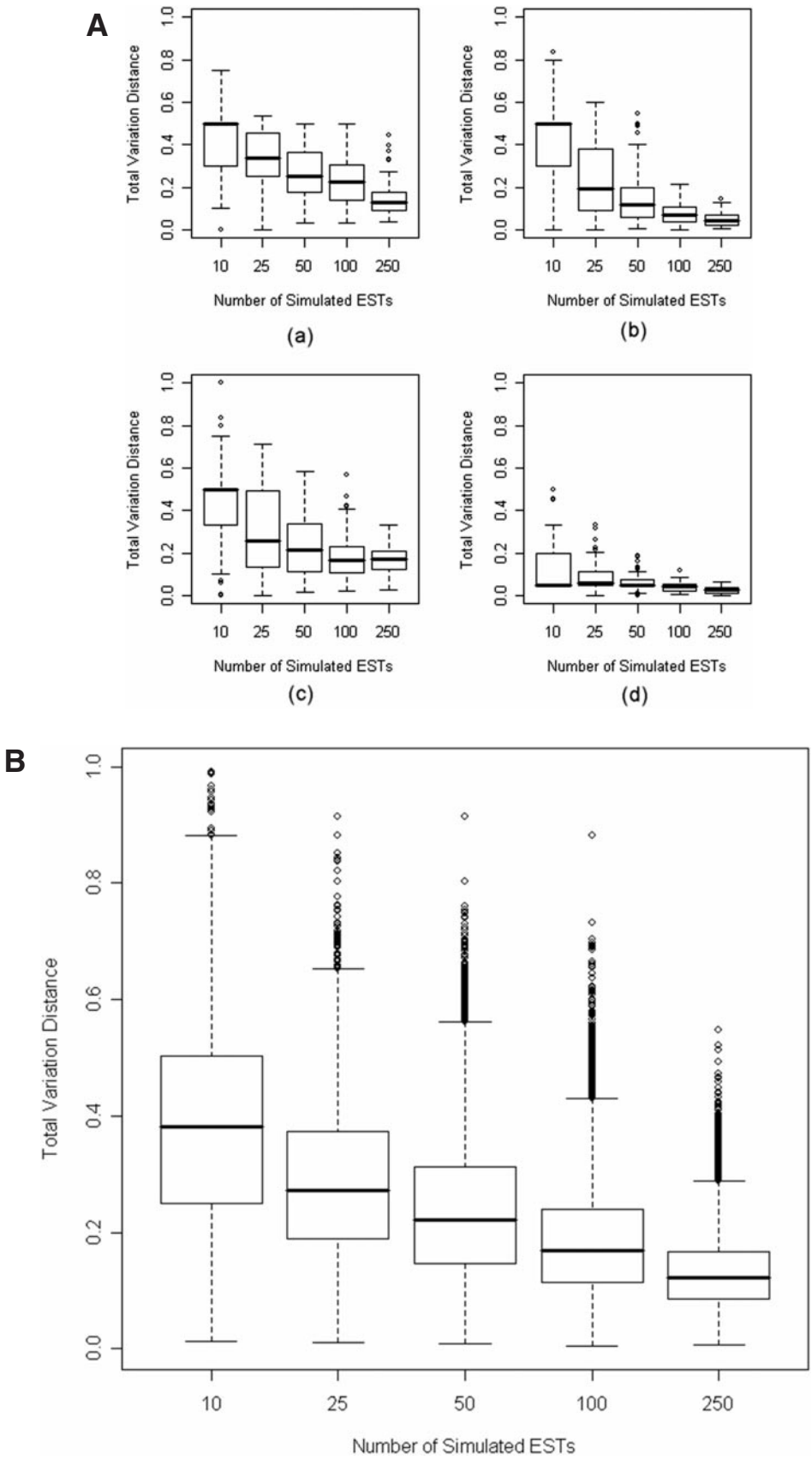

Figure 3. Simulation study to test the robustness of the EM algorithm. X-axis: the total number of sequence observations being simulated; Y-axis: the total variation distance between the true probabilities and the estimated probabilities (see Materials and Methods). (**A**) Simulation studies using fixed probabilities of four isoforms. The probabilities are listed in Table 1. (**B**) A simulation study using randomized probabilities of four isoforms.

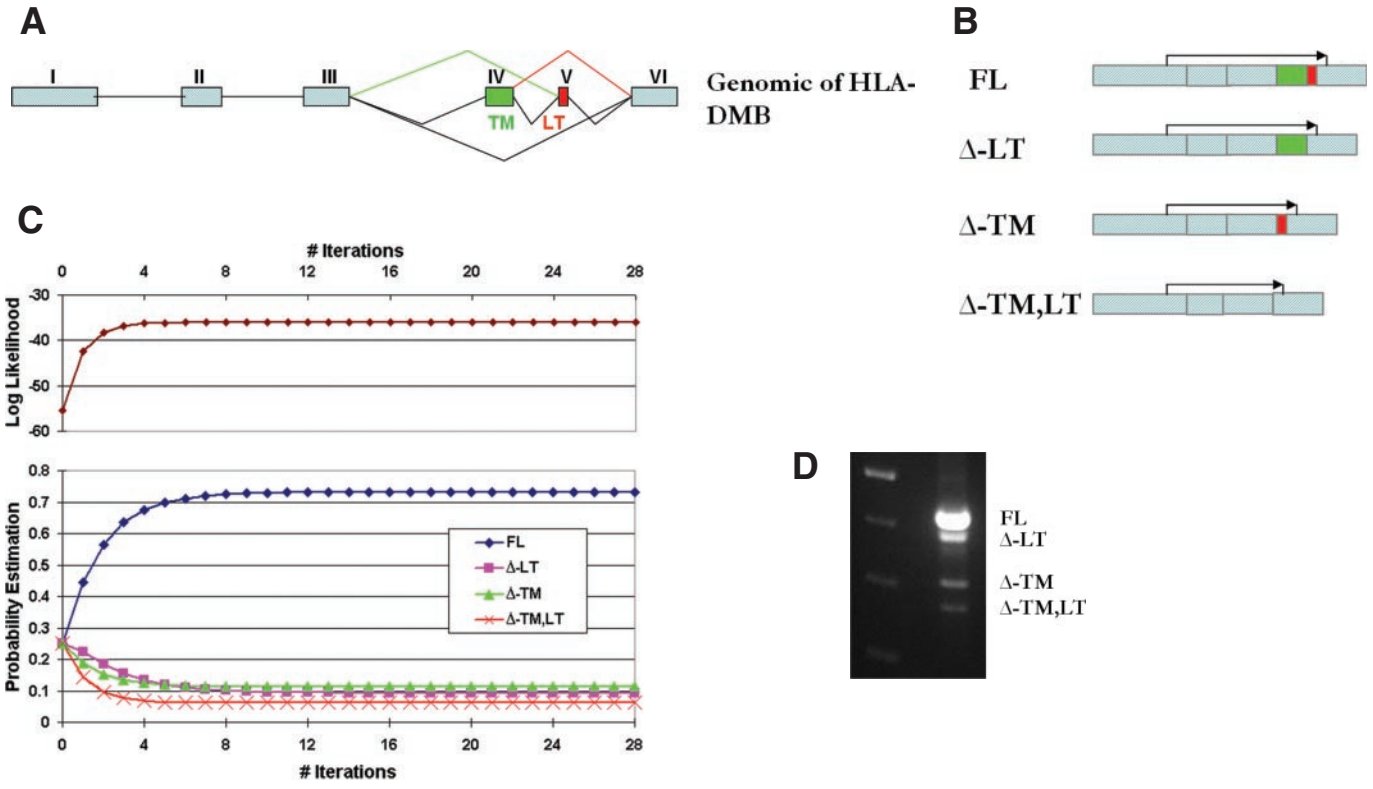

Figure 4. Probabilistic isoform reconstruction for HLA-DMB. (**A**) Gene structure and alternative splicing of HLA-DMB. Exon 4 encodes the TM domain. Exon 5 encodes the LT signal. (**B**) Four putative isoforms of HLA-DMB. (**C**) Probabilistic isoform reconstruction of HLA-DMB. The upper graph indicates the overall log likelihood; the lower graph shows the estimated probabilities for individual isoforms until convergence. The FL form has the highest estimated probability, followed by Δ-LT, Δ-TM and Δ-TM,LT. (**D**) RT–PCR analysis of HLA-DMB isoforms in pooled human tissues (see Appendix 3 of the online supplements for details of the experiment). Left lane: marker; Right lane: HLA-DMB.

high-confidence isoforms, compared to the non-probabilistic method. In addition, EM algorithm does a much better job in distinguishing high-scoring isoforms from low-scoring isoforms (data not shown).

DISCUSSION

Our study is an extension to a large body of previous work which reconstruct full-length isoforms from the splice graph (15,19–26). An essential question in every isoform reconstruction method is how to integrate evidence from full-length sequences and sequence fragments. This question is complicated by the combinatorial nature of alternative splicing, i.e. a sequence fragment can be consistent evidence for a large number of isoforms. Some previous studies enumerate all possible isoforms (23,24). Several others use certain rules to produce a minimal set of isoforms sufficient to explain all the input data (15,20–22). The resulting minimal set of isoforms can be input-order dependent (15,18). There is a genuine need for a global consideration of all the observational data simultaneously, and a probabilistic approach to measure evidence for every possible traversal across the splice graph. In this manuscript, we introduce a probabilistic formulation of the isoform reconstruction problem, and describe an EM algorithm for its maximum likelihood solution. Through the analyses of simulated data and sequence data of real human genes, we demonstrate that our algorithm provide a robust, probabilistic solution for various situations of fragmentation and coupling in the input data.

Our probabilistic framework provides the basis for several areas of extensions. One extension is to account for errors or redundancies in the sequence observations. For example, EST data contain various types of artifacts. Sorek and colleagues (50,51) proposed a method to detect EST libraries with high levels of artifacts. It has also been suggested that due to issues related to the normalization of cDNA libraries (52), multiple ESTs from a single library should be regarded as weaker evidence for an alternative splice form than the same number of ESTs from different libraries. We can use an additional weighting parameter to reflect our confidence about an individual EST sequence. This weighting parameter can be easily incorporated in the EM algorithm for maximum likelihood estimation. Specifically, we can write the log likelihood as the weighted sum of the individual log terms, and this leads to an EM algorithm where the sum of $z_{i,k}$ becomes the weight sum of $z_{i,k}$. Of course, we can also be more formal by adding an extra layer of randomization in the model to reflect the uncertainty in the sequence observations, and then develop the EM algorithm for maximum likelihood estimation.

A second extension is to detect full-length isoforms that are specifically enriched in certain conditions (e.g. tissues,

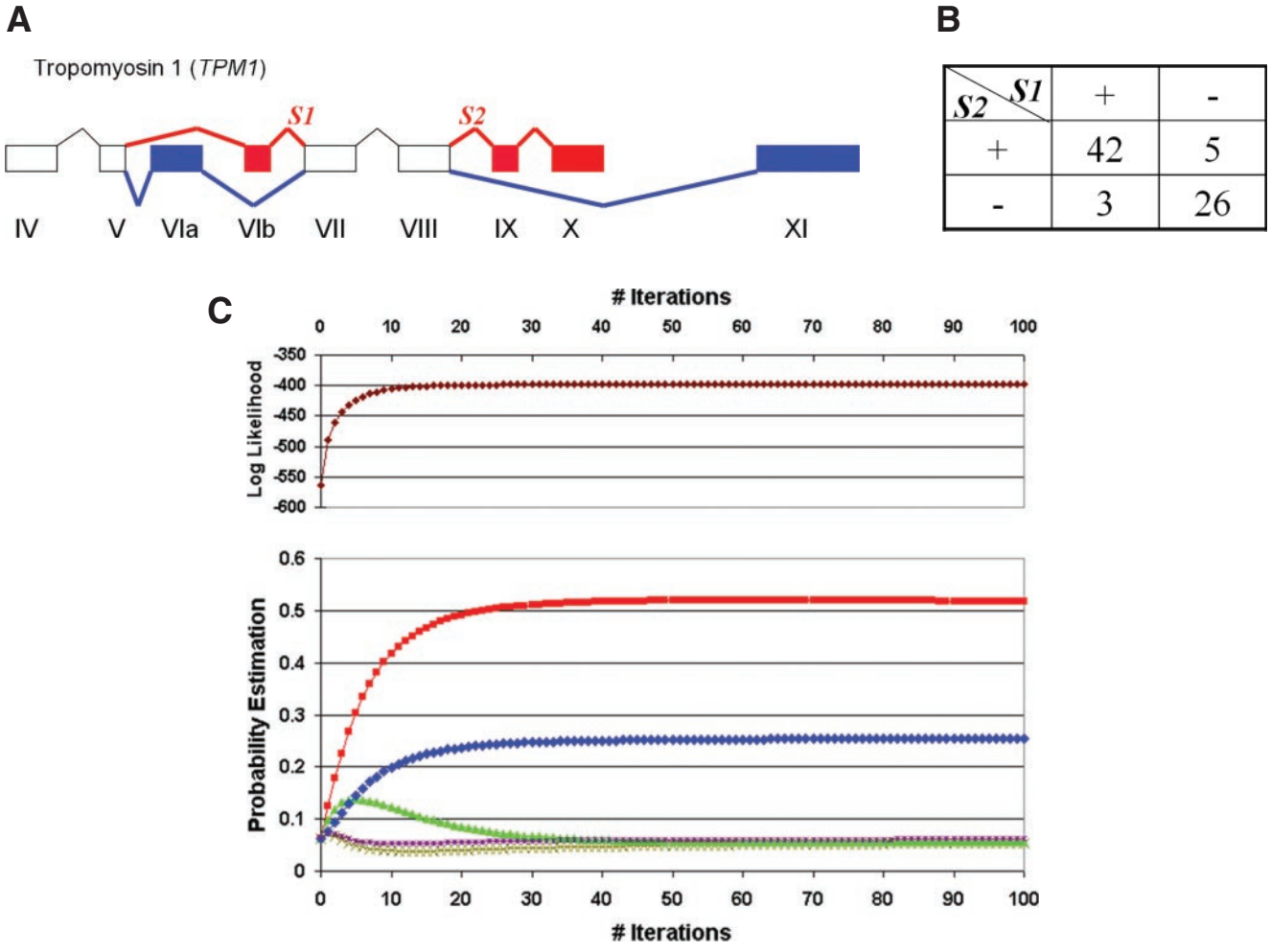

Figure 5. Probabilistic isoform reconstruction for TPM1. (**A**) The gene structure of TPM1 from exon 4 to exon 11. (**B**) EST evidence indicates coupled alternative splicing events in TPM1. (**C**) Probabilistic isoform reconstruction of TPM1. The upper graph indicates the overall log likelihood; the lower graph shows the estimated probabilities for individual isoforms until convergence. Only isoforms with >0.05 probability are shown in the graph.

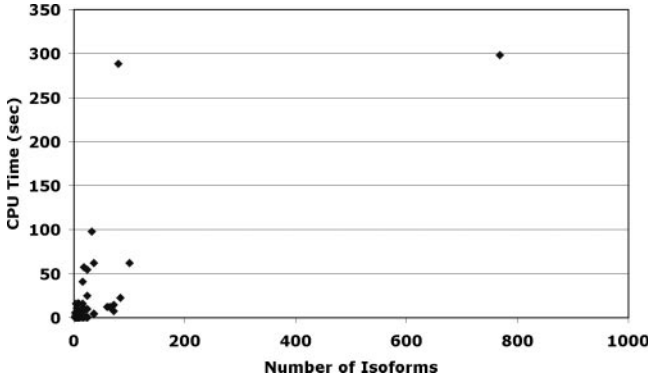

Figure 6. CPU time of probabilistic isoform reconstruction for 186 genes on human chromosome 22. X-axis: numbers of putative isoforms. Y-axis: CPU time on a PC (AMD Athlon 1500+ with 320MB of RAM).

developmental states, diseases) by considering the origin of sequence observations. Several studies used expressed sequences to detect tissue-specific and cancer-specific exons (8,52–54). However, information of specificity in sequence fragments cannot be directly translated into knowledge about full-length tissue-specific and cancer-specific isoforms. In our current probabilistic framework, the EM algorithm can run separately on different sets of sequence observations for a given gene (e.g. one set from cancerous libraries and another set from normal libraries). The likelihood ratio test (LRT) can be used to assess whether the abundance of a particular isoform varies significantly across different sets of sequence observations.

One of the most exciting future developments is to incorporate splicing microarray data. Recently, microarray technology has been developed for high-throughput analyses of alternative splicing (2,16). By designing probes targeting specific exons or exon–exon junctions, microarray allows a rapid quantitative analysis of alternative splicing of thousands of exons on a single array. However, the signal from microarray experiments is even more fragmented compared to expressed sequences. There will be a growing need for an isoform reconstruction method that integrates sequence data (e.g. expressed sequences) and microarray probe signals (55). We will describe our splice graph-based analysis of splicing microarray data in a separate manuscript.

ONLINE SUPPLEMENTS

Online Supplementary Data, including the R code for the EM calculation and simulation studies, can be accessed at http://bioinfo.mbi.ucla.edu/yxing/isoform/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Stott Parker and Namshin Kim for comments on this manuscript. This work was supported by NIH Grant U54-RR021813, a Teacher-Scholar award to C.J.L. from the Dreyfus Foundation, a DOE grant DE-FC02-02ER63421, and a Dissertation Year Fellowship to Y.X. from UCLA. Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health, Bethesda, Maryland, USA.

Conflict of interest statement. None declared.

REFERENCES

1. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet*., **30**, 13–19.
2. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
3. Lareau,L.F., Green,R.E., Bhatnagar,R.S. and Brenner,S.E. (2004) The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol*., **14**, 273–282.
4. Gravely,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*., **17**, 100–107.
5. Resch,A., Xing,Y., Modrek,B., Gorlick,M., Riley,R. and Lee,C. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. *J. Proteome Res*., **3**, 76–83.
6. Garcia,J., Gerber,S.H., Sugita,S., Sudhof,T.C. and Rizo,J. (2004) A conformational switch in the Piccolo C2A domain regulated by alternative splicing. *Nature Struct. Mol. Biol*., **11**, 45–53.
7. Wen,F., Li,F., Xia,H., Lu,X., Zhang,X. and Li,Y. (2004) The impact of very short alternative splicing on protein structures and functions in the human genome. *Trends Genet*., **20**, 232–236.
8. Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*., **30**, 3754–3766.
9. Xu,X., Yang,D., Ding,J.H., Wang,W., Chu,P.H., Dalton,N.D., Wang,H.Y., Bermingham,J.R.,Jr, Ye,Z., Liu,F. *et al*. (2005) ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle. *Cell*, **120**, 59–72.
10. Garcia-Blanco,M.A., Baraniak,A.P. and Lasda,E.L. (2004) Alternative splicing in disease and therapy. *Nat. Biotechnol*., **22**, 535–546.
11. Kochiwa,H., Suzuki,R., Washio,T., Saito,R., Bono,H., Carninci,P., Okazaki,Y., Miki,R., Hayashizaki,Y. and Tomita,M. (2002) Inferring alternative splicing patterns in mouse from a full-length cDNA library and microarray data. *Genome Res*., **12**, 1286–1293.
12. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST--database for 'expressed sequence tags'. *Nature Genet*., **4**, 332–333.
13. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
14. Boue,S., Vingron,M., Kriventseva,E. and Koch,I. (2002) Theoretical analysis of alternative splice forms using computational methods. *Bioinformatics*, **18**, S65–S73.
15. Xing,Y., Resch,A. and Lee,C. (2004) The Multiassembly Problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res*., **14**, 426–441.
16. Pan,Q., Shai,O., Misquitta,C., Zhang,W., Saltzman,A.L., Mohammad,N., Babak,T., Siu,H., Hughes,T.R., Morris,Q.D. *et al*. (2004) Revealing global regulatory features of Mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell*, **16**, 929–941.
17. Wang,H., Hubbell,E., Hu,J.S., Mei,G., Cline,M., Lu,G., Clark,T., Siani-Rose,M.A., Ares,M., Kulp,D.C. *et al*. (2003) Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics*, **19**, i315–i322.
18. Lee,C. and Wang,Q. (2005) Bioinformatics analysis of alternative splicing. *Brief Bioinform*., **6**, 23–33.
19. Sharov,A.A., Dudekula,D.B. and Ko,M.S. (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. *Genome Res*., **15**, 748–754.
20. Kim,P., Kim,N., Lee,Y., Kim,B., Shin,Y. and Lee,S. (2005) ECgene: genome annotation for alternative splicing. *Nucleic Acids Res*., **33**, D75–D79.
21. Eyras,E., Caccamo,M., Curwen,V. and Clamp,M. (2004) ESTGenes: alternative splicing from ESTs in Ensembl. *Genome Res*., **14**, 976–987.
22. Florea,L., Di Francesco,V., Miller,J., Turner,R., Yao,A., Harris,M., Walenz,B., Mobarry,C., Merkulov,G.V., Charlab,R. *et al*. (2005) Gene and alternative splicing annotation with AIR. *Genome Res*., **15**, 54–66.
23. Heber,S., Alekseyev,M., Sze,S.H., Tang,H. and Pevzner,P.A. (2002) Splicing graphs and EST assembly problem. *Bioinformatics*, **18**, S181–S188.
24. Leipzig,J., Pevzner,P. and Heber,S. (2004) The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res*., **32**, 3977–3983.
25. Lee,B.T., Tan,T.W. and Ranganathan,S. (2004) DEDB: a database of *Drosophila melanogaster* exons in splicing graph form. *BMC Bioinformatics*, **5**, 189.
26. Neverov,A.D., Artamonova,II, Nurtdinov,R.N., Frishman,D., Gelfand,M.S. and Mironov,A.A. (2005) Alternative splicing and protein function. *BMC Bioinformatics,*, **6**, 266.
27. Chang,H.C., Yu,P.S., Huang,T.W., Lin,Y.L. and Hsu,F.R. (2004) The Application of Alternative Splicing Graphs in Quantitative Analysis of Alternative Splicing Form from EST Database. *Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04)*. p. 293.
28. Haas,B.J., Delcher,A.L., Mount,S.M., Wortman,J.R., Smith,R.K.,Jr, Hannick,L.I., Maiti,R., Ronning,C.M., Rusch,D.B., Town,C.D. *et al*. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*., **31**, 5654–5666.
29. Malde,K., Coward,E. and Jonassen,I. (2005) A graph based algorithm for generating EST consensus sequences. *Bioinformatics*, **21**, 1371–1375.
30. Roberts,G.C. and Smith,C.W. (2002) Alternative splicing: combinatorial output from the genome. *Curr. Opin. Chem. Biol*., **6**, 375–383.
31. Smith,C.W. and Valcarcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci*., **25**, 381–388.
32. Fededa,J.P., Petrillo,E., Gelfand,M.S., Neverov,A.D., Kadener,S., Nogues,G., Pelisch,F., Baralle,F.E., Muro,A.F. and Kornblihtt,A.R. (2005) A polar mechanism coordinates different regions of alternative splicing within a single gene. *Mol. Cell*, **19**, 393–404.
33. Zhu,J., Shendure,J., Mitra,R.D. and Church,G.M. (2003) Single molecule profiling of alternative pre-mRNA splicing. *Science*, **301**, 836–838.
34. Bell,M.V., Cowper,A.E., Lefranc,M.P., Bell,J.I. and Screaton,G.R. (1998) Influence of intron length on alternative splicing of CD44. *Mol. Cell Biol*., **18**, 5930–5941.
35. Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Stat. Soc. B*, **39**, 1–38.
36. Excoffier,L. and Slatkin,M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol*., **12**, 921–927.
37. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*., **13**, 555–556.
38. Deng,M., Mehta,S., Sun,F. and Chen,T. (2002) Inferring domain-domain interactions from protein–protein interactions. *Genome Res*., **12**, 1540–1548.

39. Moses,A.M., Chiang,D.Y. and Eisen,M.B. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp. Biocomput.*, 324–335.
40. Hide,W.A., Babenko,V.N., van Heusden,P.A., Seoighe,C. and Kelso,J.F. (2001) The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res.*, **11**, 1848–1853.
41. Modrek,B. and Lee,C. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased rate of exon creation/loss. *Nature Genet.*, **34**, 177–180.
42. Yeo,G., Holste,D., Kreiman,G. and Burge,C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol.*, **5**, R74.
43. McCullough,R.M., Cantor,C.R. and Ding,C. (2005) High-throughput alternative splicing quantification by primer extension and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Nucleic Acids Res.*, **33**, e99.
44. Meyers,B.C., Vu,T.H., Tej,S.S., Ghazal,H., Matvienko,M., Agrawal,V., Ning,J. and Haudenschild,C.D. (2004) Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. *Nat. Biotechnol.*, **22**, 1006–1011.
45. Lee,C., Grasso,C. and Sharlow,M. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.
46. Kelly,A.P., Monaco,J.J., Cho,S.G. and Trowsdale,J. (1991) A new human HLA class II-related locus, DM. *Nature*, **353**, 571–573.
47. Sanderson,F., Kleijmeer,M.J., Kelly,A.P., Verwoerd,D., Tulp,A., Neefjes,J., Geueze,H.J. and Trowsdale,J. (1994) Accumulation of HLA-DM, a regulator of antigen presentation, in MHC class II compartments. *Science*, **266**, 1566–1569.
48. Weber,D.A., Dao,C.T., Jun,J., Wigal,J.L. and Jensen,P.E. (2001) Transmembrane domain-mediated colocalization of HLA-DM and HLA-DR is required for optimal HLA-DM catalytic activity. *J. Immunol.*, **167**, 5167–5174.
49. Le,K., Mitsouras,K., Roy,M., Wang,Q., Xu,Q., Nelson,S.F. and Lee,C. (2004) Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res.*, **32**, e180.
50. Sorek,R., Basechess,O. and Safer,H.M. (2003) Expressed sequence tags: clean before using. Correspondence re: Z. Wang *et al.*, computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. Cancer Res., **63**: 655–657, 2003. *Cancer Res.*, **63**, 6996 author reply 6996–6997.
51. Sorek,R. and Safer,H.M. (2003) A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.*, **31**, 1067–1074.
52. Gupta,S., Zink,D., Korn,B., Vingron,M. and Haas,S.A. (2004) Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics*, **5**, 72.
53. Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.*, **31**, 5635–5643.
54. Wang,Z., Lo,H.S., Yang,H., Gere,S., Hu,Y., Buetow,K.H. and Lee,M.P. (2003) Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res.*, **63**, 655–657.
55. Lee,C. and Roy,M. (2004) Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol.*, **5**, 231.