Systematic Reconstruction of Splicing Regulatory Modules by Integrating Many RNA-Seq Datasets

Chao Dai [1,2,*] and Wenyuan Li [2,*] and Juan Liu [1] and Xianghong Jasmine Zhou [2,†]
1 School of Computer, Wuhan University, Wuhan 430072, PR China
2 Molecular and Computational Biology, Department of Biological Sciences
University of Southern California, Los Angeles, CA 90089, USA

Abstract—**Alternative splicing is a ubiquitous gene regulatory mechanism that dramatically increases the complexity of the proteome. In this paper we study *splicing module*, which we define as a set of cassette exons co-regulated by the same splicing factors. We have designed a tensor-based approach to identify co-splicing clusters that appear *frequently* across multiple conditions, thus very likely to represent splicing modules – a unit in the splicing regulatory network. In particular, we model each RNA-seq dataset as a co-splicing network, where the nodes represent exons and the edges are weighted by the correlations between exon inclusion rate profiles. We apply our tensor-based method to the 19 co-splicing networks derived from RNA-seq datasets and identify an atlas of frequent co-splicing clusters. We demonstrate that these identified clusters represent splicing modules by validating against four biological knowledge databases. The likelihood that a frequent co-splicing cluster is biologically meaningful increases with its recurrence across multiple datasets, highlighting the importance of the integrative approach. We also demonstrate that the co-splicing clusters reveal novel functional groups which cannot be identified by co-expression clusters, and that the same exons can dynamically participate in different pathways depending on different conditions and different other exons that are co-spliced.**

I. Introduction

Alternative splicing provides an important means for generating proteomic diversity. Recent estimates indicate that nearly 95% of human multi-exon genes are alternatively spliced [1]. The mechanism for regulating alternative splicing is still poorly understood, and its complexity attributes to the combinatorial regulation of many factors, e.g. splicing factors, cis-acting elements, and RNA secondary structure [2], [3]. A fundamental task of alternative splicing research is to decipher splicing code and understand the mechanism of how an exon is alternatively spliced in tissue-specific manner.

A central concept in transcription regulation is the *transcription module*, defined as a set of genes that are co-regulated by the same transcription factor(s). Analogously, such coordinated regulation also occurs at the splicing level [4], [5], [6]. For example, the splicing factor *Nova* regulates exon splicing of a set of genes that shape the synapse [6]. However, the study of such coordinated splicing regulation has thus far been limited to individual cases [5], [6], [7], [8], [9]. In this paper, we define a *splicing module* as a set of exons that are regulated by the same splicing factors. The exons in a splicing module can belong to different genes, but they exhibit correlated splicing patterns (in terms of being included or excluded in their respective transcripts) across different conditions, thus form an exon co-splicing cluster.

The recent development of RNA-seq technology provides a revolutionary tool to study alternative splicing. From each RNA-seq dataset, we can derive not only the expression levels of genes, but also those of exons and transcripts (i.e., splicing isoforms). Given an RNA-seq dataset containing a set of samples, we can calculate the inclusion rate of each exon [1] in every sample, as the ratio between its expression level and that of the host gene. A recent study provided a nice example of studying splicing regulatory relationships using a network of exon-exon, exon-gene, and gene-gene links [10]. Here, we construct from each RNA-seq dataset a *weighted co-splicing network* where the nodes represent exons and the edge weights are correlations between the inclusion rates of two exons across all samples in the dataset. While directly comparing the inclusion rates for the same exon in different datasets could be biased by platforms and protocols, the *correlations* between inclusion rates for a given exon pair are comparable across datasets. From a series of RNA-seq datasets, we can therefore derive a series of co-splicing networks, which can be subjected to comparative network analysis and provide an effective way to integrate a large number of RNA-seq experiments conducted in different laboratories and using different technology platforms.

A heavy subgraph in a weighted co-splicing network represents a set of exons that are highly correlated in their inclusion rate profiles; i.e., they are co-spliced. A set of exons which *frequently* form a heavy subgraph in multiple datasets are likely to be regulated by the same splicing factors, and thus form a splicing module. We call such patterns *frequent co-splicing clusters*. Due to the enhanced signal to noise separation, frequent clusters are more robust and are more likely to be regulated by the same splicing factors (thus more likely to represent splicing modules) than those heavy subgraphs derived from a single dataset. In our previous research [11], we showed that the likelihood for a gene co-expression cluster to be a transcription module increases significantly with the recurrence of clusters in multiple datasets. A similar principle applies to splicing modules.

[*] Equally contributed joint first authors.

[†] To whom correspondence should be addressed: xjzhou@usc.edu.

[1] In this study we only consider cassette exons, which are common in alternative splicing events. Henceforth, the term "exon" always means "cassette exon."

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

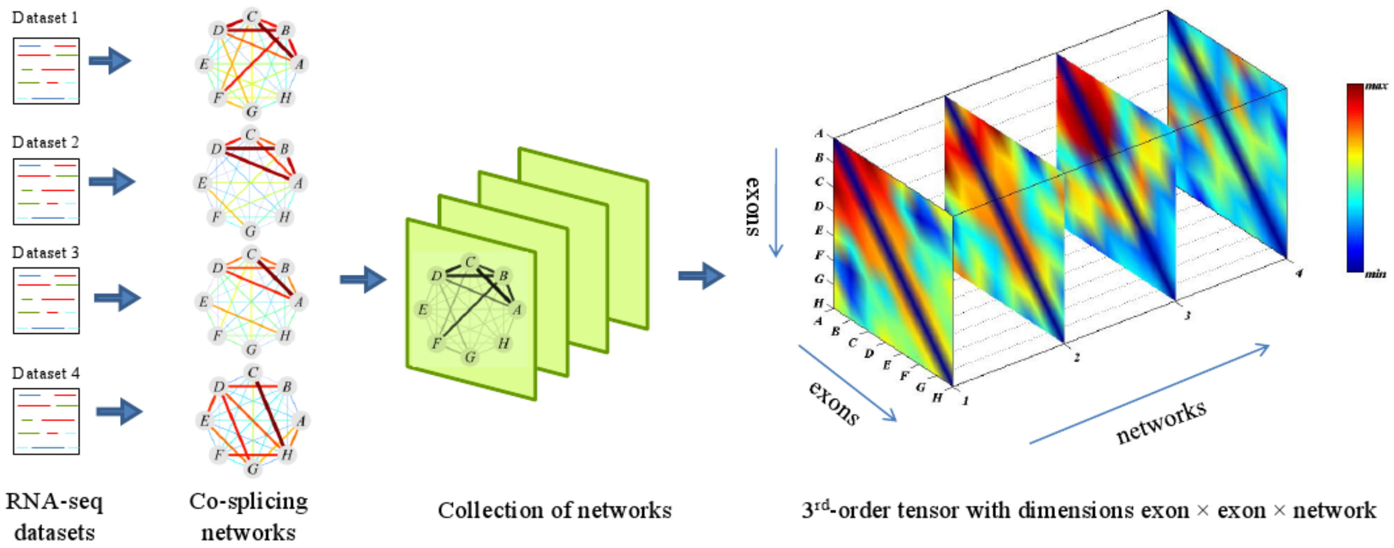

Fig. 1. A collection of co-splicing networks can be "stacked" into a third-order tensor such that each slice represents the adjacency matrix of one network. The weights of edges in the co-splicing networks and their corresponding entries in the tensor are color-coded according to the scale to the right of the figure. After reordering the tensor by the exon and network membership vectors, a frequent co-splicing cluster (colored in red) emerges in the top-left corner. It is composed of exons A, B, C, D which are heavily interconnected in networks $1, 2, 3$.

In this paper, we adopt our recently developed tensor-based approach to find the heavy subgraph that frequently occur in multiple weighted networks [12]. Our goal here is to identify co-spliced exon clusters that frequently occur across multiple weighted co-splicing networks. A co-splicing network of n nodes (exons) can be represented as an $n \times n$ adjacency matrix A, where element a_{ij} is the weight of the edge between nodes i and j. This weight represents the correlation between the two exons' inclusion rate profiles. Given m co-splicing networks with the same n nodes but different edge weights, we can represent the whole system as a 3rd-order tensor (or 3-dimensional array) of size $n \times n \times m$. An element a_{ijk} of the tensor is the weight of the edge between nodes i and j in the k^{th} network (Fig. 1). A co-splicing cluster appears as a heavy subgraph in the co-splicing network, which in turn corresponds to a heavy region in the adjacency matrix. A *frequent* co-splicing cluster is one that appears in multiple datasets, and appears as a heavy region of the tensor (Fig. 1). Thus, the problem of identifying frequent co-splicing clusters can intuitively be formulated as the problem of identifying heavy subtensors in a tensor. By representing networks and formulating the problem in this tensor form, we gain access to a wealth of established optimization methods for multidimensional arrays. Reformulating a discrete graph discovery problem as a continuous optimization problem is a long-standing tradition in graph theory. There are many successful examples, such as using a Hopfield neural network to solve the traveling salesman problem [13] and applying the Motzkin-Straus theorem to the clique-finding problem [14]. Moreover, when a graph-based pattern mining problem is transformed into a continuous optimization problem, it becomes easy to incorporate constraints representing prior knowledge. Finally, advanced continuous optimization techniques require very few *ad hoc* parameters, in contrast with most heuristic graph combinatorial algorithms.

We applied our tensor algorithm to 19 weighted exon co-splicing networks derived from human RNA-seq datasets. We identified an atlas of frequent co-splicing clusters and validated them against four biological knowledge bases: Gene Ontology annotations, RNA-binding motif database, 191 ENCODE genome-wide ChIP-seq profiles, and protein complex database. We demonstrate that the likelihood for an exon cluster to be biologically meaningful increases with its recurrence across multiple datasets, highlighting the benefit of the integrative approach. Moreover, we show that co-splicing clusters can reveal novel functional groups that cannot be identified by co-expression clusters. Finally, we show that the same exons can dynamically participate in different pathways, depending on different conditions and different other exons that are co-spliced.

II. Methods

Given an RNA-seq dataset, we construct a co-splicing network where nodes represent exons and edges are weighted by the correlation between two exon inclusion rate profiles. Given m co-splicing networks with the same n nodes but different edge weights, we can represent the whole system as a 3rd-order tensor $\mathcal{A} = (a_{ijk})_{n\times n\times m}$. A *frequent co-splicing cluster* (FSC) in the tensor $\mathcal{A}$ can be defined by two membership vectors: (i) the *exon membership vector* $\mathbf{x} = (x_1, \ldots, x_n)^T$, where $x_i = 1$ if exon i belongs to the cluster and $x_i = 0$ otherwise; and (ii) the *network membership vector* $\mathbf{y} = (y_1, \ldots, y_m)^T$, where $y_j = 1$ if the exons of the cluster are heavily interconnected in network j and $y_j = 0$ otherwise. The summed weight of all edges in the FSC is

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

$$H_{\mathcal{A}}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} a_{ijk}x_i x_j y_k. \qquad (1)$$

Note that only the weights of edges a_{ijk} with $x_i = x_j = y_k = 1$ are counted in $H_{\mathcal{A}}$. Thus, $H_{\mathcal{A}}(\mathbf{x},\mathbf{y})$ measures the "heaviness" of the FSC defined by $\mathbf{x}$ and $\mathbf{y}$. The problem of discovering a frequent co-splicing cluster can be formulated as a discrete combinatorial optimization problem: *among all patterns of fixed size (K_1 member exons and K_2 member networks), we look for the heaviest.* This is also an integer programming problem: find the binary membership vectors $\mathbf{x}$ and $\mathbf{y}$ that jointly maximize $H_{\mathcal{A}}$ under the constraints $\sum_{i=1}^{n} x_i = K_1$ and $\sum_{j=1}^{m} y_j = K_2$. However, there are several major drawbacks to this discrete formulation. The first is *parameter dependence*, meaning that the size parameters K_1 and K_2 are hard for users to provide and control. The second is *high computational complexity*; the task is proved to be NP-hard (see Supplementary Material [2]) and therefore not solvable in a reasonable time even for small datasets. Therefore, the discrete optimization problem is infeasible for an integrative analysis of many massive networks. Instead, we solve a continuous optimization problem with the same objective by relaxing integer constraints to continuous constraints. That is, we look for non-negative real vectors $\mathbf{x}$ and $\mathbf{y}$ that jointly maximize $H_{\mathcal{A}}$. This optimization problem is formally expressed as follows:

$$\begin{array}{ll} \max_{\mathbf{x}\in\mathbb{R}_+^n,\mathbf{y}\in\mathbb{R}_+^m} & H_{\mathcal{A}}(\mathbf{x},\mathbf{y}) \\ \text{subject to } f(\mathbf{x}) = 1 & \text{and} \quad g(\mathbf{y}) = 1 \end{array} \qquad (2)$$

where $\mathbb{R}_+$ is a non-negative real space, and $f(\mathbf{x})$ and $g(\mathbf{y})$ are vector norms. After solving Eq. (2), users can easily identify the top-ranking networks (after sorting the tensor by $\mathbf{y}$) and top-ranking exons (after sorting each network by $\mathbf{x}$) contributing to the objective function. After rearranging the networks in this manner, the FSC with the largest heaviness occupies a corner of the 3D tensor. We can then mask all edges in the heaviest FSC with zeros, and optimize Eq. (2) again to search for the next FSC.

The choice of vector norms in Eq. (2) has a significant impact on the outcome of the optimization. A vector norm defined as $\|\mathbf{x}\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$, where $p > 0$, is also called an "L_p-vector norm". In general, the closer p is to zero, the sparser the solution favored by the L_p-norm; that is, fewer components of the optimized vectors are significantly different from zero [15]. In contrast, as p increases, the solution favored by the L_p-norm grows smoother; in the extreme case $p \to \infty$, the elements of the optimized vector are approximately equal to each other. For more details on these vector norms, refer to the Supplementary Material. Our ideal membership vector selects *a small number of exons ("sparse") whose values are close to each other in magnitude ("smooth"), while the rest of exons are close to zero*. Our past research [12] has shown that this goal can be achieved using the mixed norm $L_{0,\infty}(\mathbf{x}) = \alpha\|\mathbf{x}\|_0 + (1-\alpha)\|\mathbf{x}\|_\infty$ $(0 < \alpha < 1)$ for $f(\mathbf{x})$. The norm L_0 favors sparsity while the norm L_∞ encourages smoothness in the non-zero components of $\mathbf{x}$. In practice, we approximate $L_{0,\infty}(\mathbf{x})$ with another mixed norm: $L_{p,2}(\mathbf{x}) = \alpha\|\mathbf{x}\|_p + (1-\alpha)\|\mathbf{x}\|_2$, where $p < 1$. Our criteria for the network membership vector are similar. We want the exon cluster to appear in as many networks as possible, so the network membership values should be non-zero and close to each other. This is the typical outcome of optimization using the L_∞ norm. In practice, we approximate L_∞ with $L_q(\mathbf{y})$, where $q > 1$ for $g(\mathbf{y})$. Therefore, the vector norms $f(\mathbf{x})$ and $g(\mathbf{y})$ are fully specified as follows,

$$f(\mathbf{x}) = \alpha\|\mathbf{x}\|_p + (1-\alpha)\|\mathbf{x}\|_2 \quad \text{and} \quad g(\mathbf{y}) = \|\mathbf{y}\|_q \qquad (3)$$

We performed simulations to determine suitable values for the parameters p, α, and q, applying our tensor method to collections of random weighted networks. We randomly placed FSCs of varying size, recurrence, and heaviness in a subset of the random networks. We then tried different combinations of p, α, and q, and adopted the combination ($p = 0.8$, $\alpha = 0.2$, and $q = 10$) that led to the discovery of the most FSCs. More details on these simulations are provided in the Supplementary Material.

Since the vector norm $f(\mathbf{x})$ is non-convex, our tensor method requires an optimization protocol that can deal with non-convex constraints. The quality of the optimum discovered for a non-convex problem depends heavily on the numerical procedure. Standard numerical techniques such as gradient descent converge to a local minimum of the solution space, and different procedures often find different local minima. Thus, it is important to find a theoretically justified numerical procedure. We use an advanced framework known as multi-stage convex relaxation, which has good numerical properties for non-convex optimization problems [15]. In this framework, concave duality is used to construct a sequence of convex relaxations that give increasingly accurate approximations to the original non-convex problem. We approximate the sparse constraint function $f(\mathbf{x})$ by the convex function $\widetilde{f}_{\mathbf{v}}(\mathbf{x}) = \mathbf{v}^T h(\mathbf{x}) - f_h^*(\mathbf{v})$, where $h(\mathbf{x})$ is a specific convex function $h(x) = x^2$ and $f_h^*(\mathbf{v})$ is the concave dual of the function $\overline{f}_h(\mathbf{v})$ (defined as $f(\mathbf{v}) = \overline{f}_h(h(\mathbf{v}))$). The vector $\mathbf{v}$ contains coefficients that will be automatically generated during the optimization process. After each optimization, the new coefficient vector $\mathbf{v}$ yields a convex function $\widetilde{f}_{\mathbf{v}}(\mathbf{x})$ that more closely approximates the original non-convex function $f(\mathbf{x})$. Details of our tensor-based optimization method can be found in the Supplementary Material.

Once the membership vectors (i.e., the solution of Eq. (2)) have been found by optimization, the frequent co-splicing clusters can be intuitively obtained by including those exons and networks with large membership values. However, any given solution can result in multiple overlapping patterns whose "heaviness" is greater than a specified threshold. Here, *heaviness* is defined as the average weight of all edges in the pattern. To identify the most representative pattern, we first

[2] Supplementary Material is available at http://zhoulab.usc.edu/cosplicing

rank exons and networks in decreasing order of their membership values in $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. Then we extract two representative patterns that satisfy the heaviness threshold: the pattern that occurs in the most networks while having at least the minimum number of top-ranking exons (e.g., 5), and the pattern with the largest number of top-ranking exons while appearing in at least the minimum number of top-ranking networks (e.g., 4). Both patterns are included as co-splicing clusters in our results. After discovering a pattern, we can mask its edges in those networks where it occurs (replacing those elements of the tensor with zeroes) and optimize Eq. (2) again to search for the next frequent co-splicing cluster.

III. Results

We identified 19 human RNA-seq datasets from the NCBI Sequence Read Archive [3], each with at least six samples providing transcriptome profiling under multiple experimental conditions, such as diverse tissues or various diseases. For each dataset, we used the Bowtie [16] tool to map short reads to the *hg18* reference genome, setting the program options to report only the optimal alignment and discard those reads that map equally well to multiple positions. Next, we applied the transcript assembly tool Cufflinks [17] to estimate the expressions for all transcripts with known ensemble transcription annotations. We calculated the inclusion rate of each exon, as the ratio between its expression (the sum of RPKM [4] over all transcripts that cover the exon) and the host gene's expression (the sum of RPKM over all transcripts of the gene). It is worth noting that in RNA-seq experiments, a gene expression with low RPKM is usually not precisely estimated because the number of reads mapped to the gene is quite small. In order to work with reasonably accurate estimates of exon inclusion rates, as pointed out by [19], we calculated inclusion rates only for those genes whose expressions are above 70th percentile across at least 2/3 of the samples. This criterion resulted in inclusion rate profiles for 5422 exons covering 3343 genes. Based on these profiles, we constructed an exon co-splicing network from each RNA-seq dataset by using Pearson's correlation between exons' inclusion rate profiles.

We applied our method to 19 RNA-seq datasets generated under various experimental conditions. Adopting the empirical criteria of "heaviness" $\geqslant 0.4$ and cluster size $\geqslant 5$ exons, we identified 2334 co-splicing clusters with recurrences $\geqslant 4$, 1064 co-splicing clusters with recurrences $\geqslant 5$, and 442 co-splicing clusters with recurrences $\geqslant 6$.

A. Frequent co-splicing clusters are likely to represent functional modules, splicing modules, transcriptional modules, and protein complexes

To assess the biological significance of the identified patterns, we evaluate the extent to which these exon clusters represent functional modules, splicing modules, transcriptional regulatory modules, and protein complexes.

a) Functional analysis: We evaluated the functional homogeneity of the host genes in an exon cluster using Gene Ontology (GO) annotations. To ensure the specificity of GO terms, we filtered out general GO terms associated with > 500 genes. If the host genes of exons in a cluster are statistically enriched in a GO term with p-value$<$1E-4 (based on the hypergeometric test), we declare the exon cluster to be functionally homogeneous. We found that 14.9% of clusters appearing in $\geqslant$6 datasets are functionally homogenous, compared to only 5.4% of randomly generated clusters with the same sizes. functionally homogenous clusters cover a wide range of post-transcriptional associated GO terms, such as "RNA splicing", "ribonucleoprotein binding", "heterogeneous nuclear ribonucleoprotein complex", "negative regulation of protein catabolic process", and "cellular protein localization". When we perform the same analysis for clusters with lower recurrences (4 or 5 datasets), it is clear that functional homogeneity is more likely among more frequent clusters (Fig. 2A). These results confirm the benefits of the integrative approach in improving the quality of detected patterns.

b) Splicing regulatory analysis: By construction, the exons in our identified co-splicing clusters have highly correlated inclusion rate profiles across different experimental conditions. Clusters meeting this criterion are likely to consist of exons co-regulated by the same splicing factors. It has been shown that splicing factors can affect alternative splicing by interacting with cis-acting elements in a position-dependent manner [20]. We collected the binding motifs of 33 RNA-binding proteins from the RBPDB database (version 1.2.1 released on 25/01/2011) [21]. These proteins include known and potential splicing factors. To identify possible splicing factors associated with a co-splicing cluster, for each exon of a co-splicing cluster, we retrieved the internal exon region and its 100bp flanking intron region to check whether those regions contain one or more of the exact motifs of those 33 RNA-binding proteins. If the exons of a cluster are highly enriched in the targets of an RNA-binding protein, then this protein could serve as the common splicing regulator of the cluster. In this case, we consider the cluster to be "splicing homogenous". At the p-value$<$0.01 level (based on the hypergeometric test), 12.2% clusters with $\geqslant$5 exons and $\geqslant$6 recurrences are splicing homogenous. Performing the same analysis for less frequent clusters, we find that as the recurrence increases, so does the percentage of splicing homogenous modules (Fig. 2B). The four most frequently enriched RNA-binding proteins are *RBM4*, *YTHDC1*, *YBX1* and *SFRS1*. *RBM4* is known to be involved in diverse cellular processes including alternative splicing of pre-mRNA, translation, and RNA silencing [22]. *YTHDC1* has been shown to modulate alternative splice site selection in a concentration-dependent manner [23], and its malfunction is associated with a number of diseases[24], [25]. The RNA splicing mediated by *YBX1* is inhibited by TLS/CHOP in human myxoid liposarcoma cells [26]. *SF2/SFRS1* promotes alternative exon inclusion, and prevents inappropriate exon skipping in natural alternatively spliced pre-mRNAs [27].

We found that some splicing factors tend to co-bind to the

[3] http://www.ncbi.nlm.nih.gov/sra

[4] RPKM stands for "Reads Per Kilobase of exon model per Million mapped reads", as defined in [18].

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

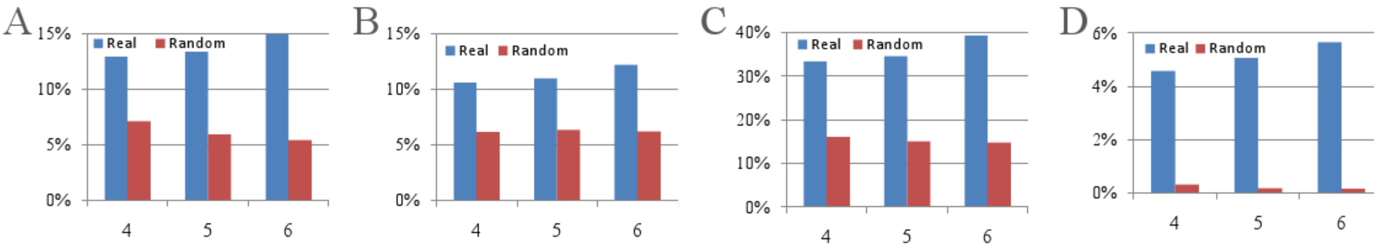

Fig. 2. Evaluation of the functional, transcriptional, splicing, and protein complex homogeneity of co-splicing clusters with different recurrences. Four types of databases are used: (A) Gene Ontology for functional enrichment, (B) RBPDB database for splicing enrichment, (C) ENCODE database for transcriptional and epigenetic enrichment, and (D) CORUM database for protein complex enrichment. The x-axis is recurrence and y-axis is enrichment rate.

cis-regulatory regions of exons in a co-splicing cluster, suggesting the combinatorial regulation of those splicing factors. *SFRS1* and *RBM4* are simultaneously enriched in 12 clusters, whose major functions (by GO term enrichment) are related to transcriptional regulation, such as "transcription factor binding" (p-value=1.22E-3), "transcription repressor activity" (p-value=2.99E-3), and "positive regulation of gene-specific transcription" (p-value=4.46E-3). *KHDRBS3*, *Fox-1* and *YBX2*, are simultaneously enriched in 6 clusters. These clusters are associated with post-transcriptional regulation, for example, "proteolysis involved in cellular protein catabolic process" (p-value=8.31E-4), "ubiquitin-dependent protein catabolic process" (p-value=4.85E-4), and "post-transcriptional regulation of gene expression" (p-value=4.53E-3). Our results suggest that combinatorial splicing regulation can occur in both co-transcriptional and post-transcriptional processes.

c) Transcriptional and epigenomic analysis: To evaluate how co-splicing is affected by transcriptional regulation, we used 191 ChIP-seq profiles generated by the Encyclopedia of DNA Elements (ENCODE) consortium [28]. This dataset includes the genome-wide bindings of 40 transcription factors (TF), 9 histone modification marks, and 3 other markers (DNase, FAIRE, and DNA methylation) on 25 different cell lines. For a detailed description of the signal extraction procedure, see the Supplementary Material. If the host genes of an exon cluster are highly enriched in the targets of any regulatory factor, we consider the cluster to be "transcription homogenous". At the significance level p-value < 0.01, 39.4% clusters with recurrences $\geqslant$6 are transcription homogenous, compared to only 14.8% of randomly generated clusters with the same sizes. As expected, the percentage of transcription homogenous modules increases with recurrence (Fig. 2C). This result suggests a strong association between transcription and splicing for a significant number of genes. The 5 most frequently enriched regulatory factors are *JUN*, *H3K9me1*, *STAT2*, *H4K20me1* and *H3K36me3*. *JUN* and *STAT2* are transcriptional factors regulating a wide range of biological processes, while the roles of *H3K9me1* and *H4K20me1* in transcription or splicing are not yet clear. Of particular interest is *H3K36me3*, a histone modification mark closely related to alternative splicing. The causal effect of *H3K36me3* on alternative splicing was recently discovered: increasing *H3K36me3* reduces the inclusion of PTB-dependent exons in *FGFR2*, *TPM2*, *TPM1* and *PKM2* mRNA [29]. Furthermore, the mechanism is recruitment of PTB to *H3K36me3*-modified chromatin through protein *MRG15* [29]. *H3K36me3* is also potentially linked to transcription elongation by RNA polymerase II [30], which may be a regulator for transcription-coupled alternative splicing. Based on the kinetic model, the rate of transcription elongation influences the inclusion of alternative exons by affecting whether the splicing machinery is recruited sufficiently quickly for spliceosome assembly and splicing to occur [3].

d) Protein complex analysis: We evaluate the extent to which host genes of our identified exon clusters are protein complexes by using the Comprehensive Resource of Mammalian protein complexes database (CORUM, September 2009 version) [31]. At the significance level p-value < 0.01, 5.7% of co-splicing clusters with recurrences $\geqslant$6 are enriched in genes belonging to a protein complex, versus only 0.16% of randomly generated clusters with the same sizes. The percentage of clusters enriched in protein complexes increases with the cluster recurrence (Fig. 2D). The 4 most frequently enriched protein complexes are "large *Drosha* complex", "C complex spliceosome", "*TNFα/NF-κB* signaling complex", "*PABPC1-HSPA8-HNRPD-EIF4G1* complex". At least 1/3 of subunits in the highest enriched complex "large *Drosha* complex" contain proteins associated with splicing function, especially heterogeneous nuclear ribonucleoproteins such as *HNRNPH1*, *HNRNPM*, *HNRNPU*, *HNRNPUL1* and *HNRNPDL* [31].

B. Co-splicing clusters reveal novel functions that are not identified by co-expression clusters

Studies have shown that genes that are co-regulated transcriptionally do not necessarily overlap with those that are co-spliced [32]. Therefore, the identification of co-splicing clusters can reveal functionally related genes that could not be discovered from transcription analysis. In order to identify novel functions associated with co-splicing but not co-expression, we complement the above analysis by constructing a gene co-expression network from each RNA-seq dataset. The nodes of these networks represent genes, and the edges are weighted by Pearson's correlation between two gene expression profiles. We then apply our tensor-based pattern mining algorithm to identify frequent co-expression clusters in the 19 co-expression networks. The same functional enrichment analysis described above for co-splicing clusters was performed on the resulting co-expression clusters. We found that 97.7% of co-splicing clusters with recurrences $\geqslant 4$ have low expression correlations (average correlations $\leqslant 0.4$). Therefore, many of the functions associated with post-transcriptional regulation are enriched in co-splicing clusters but not in co-expression clusters. These functions include "maintenance of protein location", "regulation of protein catabolic process", "cytoplasmic sequestering of protein", "regulation of intracellular protein transport", "regulation of ubiquitin-protein ligase activity", "ribonucleoprotein complex assembly", "RNA splicing, via transesterification reactions", and "RNA export from nucleus".

For example, one co-splicing cluster has seven host genes: *HNRNPC*, *HERPUD1*, *RALY*, *MAPKAP1*, *PUM1*, *PKM2*, and *DCAF11*. This cluster cannot be found from co-expression data, for the expression profiles of the host genes have low correlations. However, this set of host genes is enriched with several splicing associated functions including "spliceosomal complex" (p-value=1.15E-3) and "RNA splicing" (p-value=4.76E-3). Out of the seven host genes, *RALY* and *PUM1* encode RNA-binding proteins, and *HNRNPC* encodes heterogenous nuclear ribonucleoproteins C1/C2. This co-splicing cluster reveals a cascade splicing effect: the co-spliced genes encode RNA-binding proteins or splicing factors, which can participate in the splicing of downstream genes. Clearly, co-splicing clusters can provide complementary information on functionally related gene groups in addition to co-transcription clusters. In particular, co-splicing clusters can grant new insights into functions associated with post-transcriptional regulation.

C. Exons can dynamically participate in different pathways upon different co-splicing mechanisms

Alternatively skipping or including a cassette exon can change the functions of a protein by deleting or inserting a protein domain. In other words, protein isoforms alternatively spliced from the same gene may participate in different pathways. In our results, we observed that 45.3%/35.7%/26.1% of exons are members of at least two clusters (recurrence$\geqslant$4/5/6) with different functions. For example, exon8 of the gene *Rela* appears in three co-splicing clusters, which are enriched with the following distinct functions respectively: "regulation of NF-κB cascade" (p-value=2.71E-6), "negative regulation of protein catabolic process" (p-value=2.04E-5), and "kinase binding" (p-value=8.99E-5). *Rela* encodes the transcription factor *p65*, which is an important subunit of the *NF-κB* complex that affects several hundred genes by *NF-κB* signaling. Recent research has identified several alternative splice variants of *Rela*, e.g. *p65△*, *p65△2* and *p65△3*. In fact, *p65△* arises by the use of an alternative splice site located 30 nucleotides into exon8, and *p65△3* was identified as a splice variant lacking exon7 and exon8 [33]. These facts are consistent with our finding that exon8 is dynamically included in multiple co-splicing clusters. As another example, exon7 of *PRMT5* appears in two co-splicing clusters, which are enriched with two distinct splicing functions, "ribonucleoprotein complex assembly" (p-value=9.75E-4) and "ribonucleoprotein binding" (p-value=5.66E-5). This is consistent with recent genome-wide studies that *PRMT5* contributes to the regulation of many pre-messenger-RNA splicing events in various ways [34]. These examples demonstrate that exons can contribute to different functionalities of proteins depending on different splicing regulatory mechanisms.

IV. Conclusion

Splicing code is determined by a combination of many factors, such as cis-acting elements and trans-acting factors. If some exons share the same splicing code, they may form a splicing module: a unit in the splicing regulatory network. Therefore, identifying co-splicing clusters first and then investigating their cis-acting elements and associated trans-acting factors can serve as an important step to decipher the splicing code. Our tensor-based approach can identify co-spliced exon clusters that frequently appear in multiple RNA-seq datasets. The exons in a frequent co-splicing cluster can belong to different genes, but are very likely to be co-regulated by the same splicing factors, thus forming a splicing module. We demonstrated that the identified clusters represent meaningful biological modules, i.e. functional modules, splicing modules, transcriptional modules, and protein complexes, by validating against four biological knowledge databases. In all four types of enrichment results, the likelihood that a co-splicing cluster is biologically meaningful increases with its recurrence. This consistent behavior highlights the importance of the integrative approach. We also showed that the co-splicing clusters can reveal novel functional related genes that cannot be identified by co-expression clusters, and that the same exons can dynamically participate in different pathways depending on different conditions and different other exons that are co-spliced. The *NCBI Sequence Reader Achieve* database currently stores 6293 RNA-seq profiles, and this number is expected to dramatically increase in the near future. We expect to apply our approach to the rapidly accumulating RNA-seq data of multiple organisms, and to identify a large number of splicing modules and their associated phenotype conditions. This analysis can serve as a first step towards the reconstruction of tissue- and disease-specific splicing regulatory networks.

Acknowledgment

The work presented in this paper was supported by National Institutes of Health Grants R01GM074163 and U54CA112952, and NSF Grant 0747475.

References

[1] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, Nov. 2008.

[2] A. J. Matlin, F. Clark, and C. W. J. Smith, "Understanding alternative splicing: towards a cellular code," *Nat Rev Mol Cell Biol*, vol. 6, no. 5, pp. 386–398, May 2005.

[3] M. Chen and J. L. Manley, "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches," *Nat Rev Mol Cell Biol*, vol. 10, no. 11, pp. 741–754, Nov. 2009.

[4] R. Nagoshi, M. McKeown, K. Burtis, J. Belote, and B. Baker, "The control of alternative splicing at genes regulating sexual differentiation in D. melanogaster," *Cell*, vol. 53, no. 2, pp. 229–236, 1988.

[5] M. Hedley and T. Maniatis, “Sex-specific splicing and polyadenylation of dsx pre-mRNA requires a sequence that binds specifically to tra-2 protein in vitro,” *Cell*, vol. 65, no. 4, pp. 579–586, 1991.

[6] A. Jernej Ule, J. Ule, J. Alan Williams, H. Melissa Cline, C. Tyson Clark, B. Matteo Ruggiu, J. David Kane, and R. John Blume, “Nova regulates brain-specific splicing to shape the synapse,” *Nature Genetics*, vol. 37, no. 8, pp. 844–852, 2005.

[7] M. Hentze and L. Kuhn, “Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 16, pp. 8175–82, 1996.

[8] C. Zhang, Z. Zhang, J. Castle, S. Sun, J. Johnson, A. Krainer, and M. Zhang, “Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2,” *Genes & Development*, vol. 22, no. 18, pp. 2550–2563, 2008.

[9] M. Moore, Q. Wang, C. Kennedy, and P. Silver, “An alternative splicing network links cell-cycle control to apoptosis,” *Cell*, vol. 142, no. 4, pp. 625–636, 2010.

[10] L. Chen and S. Zheng, “Studying alternative splicing regulatory networks through partial correlation analysis,” *Genome biology*, vol. 10, no. 1, p. R3, 2009.

[11] X. Yan, M. Mehan, Y. Huang, M. Waterman, P. Yu, and X. Zhou, “A graph-based approach to systematically reconstruct human transcriptional regulatory modules,” *Bioinformatics*, vol. 23, no. 13, p. i577, 2007.

[12] W. Li, C. Liu, T. Zhang, H. Li, M. Waterman, and X. Zhou, “Integrative analysis of many weighted co-expression networks using tensor computation,” *PLoS Computational Biology*, vol. 7, no. 5, p. in press, 2011.

[13] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proc Natl Acad Sci USA.*, vol. 79, no. 8, pp. 2554–2558.

[14] T. S. Motzkin and E. G. Straus, “Maxima for graphs and a new proof of a theorem of Turán,” *Canadian Journal of Mathematics*, vol. 17, no. 4, pp. 533–540, 1965.

[15] T. Zhang, “Analysis of multi-stage convex relaxation for sparse regularization,” *J Mach Learn Res*, vol. 11, pp. 1081–1107, 2010.

[16] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” *Genome Biology*, vol. 10, no. 3, p. R25, 2009, PMID: 19261174.

[17] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation,” *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, May 2010, PMID: 20436464.

[18] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[19] H. Jiang and W. H. Wong, “Statistical inferences for isoform expression in RNA-Seq,” *Bioinformatics*, vol. 25, no. 8, pp. 1026–1032, Apr. 2009, PMID: 19244387.

[20] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell, “HITS-CLIP yields genome-wide insights into brain alternative RNA processing,” *Nature*, vol. 456, no. 7221, pp. 464–469, Nov. 2008. [Online]. Available: http://dx.doi.org/10.1038/nature07488

[21] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes, “RBPDB: a database of RNA-binding specificities,” vol. 39, no. Database issue, pp. D301–D308, Jan. 2011.

[22] M. A. Markus, B. Heinrich, O. Raitskin, D. J. Adams, H. Mangs, C. Goy, M. Ladomery, R. Sperling, S. Stamm, and B. J. Morris, “WT1 interacts with the splicing protein RBM4 and regulates its ability to modulate alternative splicing in vivo,” *Experimental Cell Research*, vol. 312, no. 17, pp. 3379–3388, Oct. 2006, PMID: 16934801.

[23] A. M. Hartmann, O. Nayler, F. W. Schwaiger, A. Obermeier, and S. Stamm, “The interaction and colocalization of sam68 with the splicing-associated factor YT521-B in nuclear dots is regulated by the src family kinase p59(fyn),” *Molecular Biology of the Cell*, vol. 10, no. 11, pp. 3909–3926, Nov. 1999, PMID: 10564280.

[24] B. Zhang, A. zur Hausen, M. Orlowska-Volk, M. J "ager, H. Bettendorf, S. Stamm, M. Hirschfeld, O. Yiqin, X. Tong, G. Gitsch *et al.*, “Alternative splicing-related factor yt521: An independent prognostic factor in endometrial cancer,” *International Journal of Gynecological Cancer*, vol. 20, no. 4, p. 492, 2010.

[25] F. Wilkinson, J. Holaska, Z. Zhang, A. Sharma, S. Manilal, I. Holt, S. Stamm, K. Wilson, and G. Morris, “Emerin interacts in vitro with the splicing-associated factor, yt521-b,” *European Journal of Biochemistry*, vol. 270, no. 11, pp. 2459–2466, 2003.

[26] T. B. Rapp, L. Yang, r. Conrad, Ernest U, N. Mandahl, and H. A. Chansky, “RNA splicing mediated by YB-1 is inhibited by TLS/CHOP in human myxoid liposarcoma cells,” *Journal of Orthopaedic Research: Official Publication of the Orthopaedic Research Society*, vol. 20, no. 4, pp. 723–729, Jul. 2002, PMID: 12168660.

[27] A. Mayeda, D. M. Helfman, and A. R. Krainer, “Modulation of exon skipping and inclusion by heterogeneous nuclear ribonucleoprotein a1 and pre-mRNA splicing factor SF2/ASF,” *Molecular and Cellular Biology*, vol. 13, no. 5, pp. 2993–3001, May 1993, PMID: 8474457.

[28] D. J. Thomas, K. R. Rosenbloom, H. Clawson, A. S. Hinrichs, H. Trumbower, B. J. Raney, D. Karolchik, G. P. Barber, R. A. Harte, J. Hillman-Jackson, R. M. Kuhn, B. L. Rhead, K. E. Smith, A. Thakkapallayil, A. S. Zweig, D. Haussler, and W. J. Kent, “The ENCODE project at UC santa cruz,” *Nucleic Acids Res*, vol. 35, no. Database issue, pp. D663–D667, 2007, PMID: 17166863. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17166863

[29] R. F. Luco, Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith, and T. Misteli, “Regulation of alternative splicing by histone modifications,” *Science (New York, N.Y.)*, vol. 327, no. 5968, pp. 996–1000, Feb. 2010, PMID: 20133523. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20133523

[30] R. J. Sims III and D. Reinberg, “Processing the H3K36me3 signature,” *Nat Genet*, vol. 41, no. 3, pp. 270–271, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1038/ng0309-270

[31] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H. Mewes, “CORUM: the comprehensive resource of mammalian protein complexes–2009,” *Nucl. Acids Res.*, vol. 38, no. Database issue, pp. D497–D501, Jan. 2010, PMID: 19884131.

[32] Q. Pan, O. Shai, C. Misquitta, W. Zhang, A. L. Saltzman, N. Mohammad, T. Babak, H. Siu, T. R. Hughes, Q. D. Morris, B. J. Frey, and B. J. Blencowe, “Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform,” *Molecular Cell*, vol. 16, no. 6, pp. 929–941, Dec. 2004, PMID: 15610736.

[33] J. R. Leeman and T. D. Gilmore, “Alternative splicing in the NF-kappaB signaling pathway,” *Gene*, vol. 423, no. 2, pp. 97–107, Nov. 2008.

[34] S. E. Sanchez, E. Petrillo, E. J. Beckwith, X. Zhang, M. L. Rugnone, C. E. Hernando, J. C. Cuevas, M. A. Godoy Herz, A. Depetris-Chauvin, C. G. Simpson, J. W. S. Brown, P. D. Cerdan, J. O. Borevitz, P. Mas, M. F. Ceriani, A. R. Kornblihtt, and M. J. Yanovsky, “A methyl transferase links the circadian clock to the regulation of alternative splicing,” *Nature*, vol. 468, no. 7320, pp. 112–116, Nov. 2010.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE