Algorithm to Identify Frequent Coupled Modules from Two-Layered Network Series: Application to Study Transcription and Splicing Coupling

WENYUAN LI,^{1,*} CHAO DAI,^{1,*} CHUN-CHI LIU,² and XIANGHONG JASMINE ZHOU¹

ABSTRACT

Current network analysis methods all focus on one or multiple networks of the same type. However, cells are organized by multi-layer networks (e.g., transcriptional regulatory networks, splicing regulatory networks, protein-protein interaction networks), which interact and influence each other. Elucidating the coupling mechanisms among those different types of networks is essential in understanding the functions and mechanisms of cellular activities. In this article, we developed the first computational method for pattern mining across many two-layered graphs, with the two layers representing different types yet coupled biological networks. We formulated the problem of identifying frequent coupled clusters between the two layers of networks into a tensor-based computation problem, and proposed an efficient solution to solve the problem. We applied the method to 38 two-layered co-transcription and co-splicing networks, derived from 38 RNA-seq datasets. With the identified atlas of coupled transcription-splicing modules, we explored to what extent, for which cellular functions, and by what mechanisms transcription-splicing coupling takes place.

Key word: algorithms.

1. INTRODUCTION

THE RECENT DEVELOPMENT OF HIGH-THROUGHPUT TECHNOLOGIES provides numerous opportunities to systematically characterize diverse biological networks. Network Biology is an emerging field (Barabasi and Oltvai, 2004). Thus far, most computational methods have focused on the analysis of one or more of the same type of biological networks, e.g. protein-protein interaction networks, transcriptional regulatory networks, or metabolic networks, each representing a single layer of organization in the complex cellular system. In reality, these multiple levels of organization interact and influence each other, harboring sophisticated coupling mechanisms that are essential in maintaining the function and robustness of cells. However, thus far, no computational methods exist to analyze the coupling between different types of biological networks. Yet, more and more experimental studies tend to simultaneously profiling biological

¹Program in Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California.

²Institute of Genomics and Bioinformatics, National Chung Hsing University, Taiwan.

*These two authors contributed equally to this article.

systems at multiple levels. For example, the Cancer Genome Atlas (TCGA) project is generating multi-dimensional maps of key genomic changes (e.g., SNP, DNA methylation, gene expression, and microRNA expression) for the same set of tumor samples (Cancer Genome Atlas Research Network, 2008). The NCI60 project has profiled 60 human cancer cell lines in terms of drug responses (Weinstein et al., 1997; Bussey et al., 2006; Scherf et al., 2000), gene expression (Staunton et al., 2001), protein expression (Shankavaram et al., 2007), and miRNA expression (Blower et al., 2007). The most abundant multi-level profiling data, in fact, is RNA-seq data, since every RNA-seq profile provides the information at both the splicing level (the quantification of exon abundances) and the transcription level (the quantification of gene expressions). How to utilize such multi-dimensional data to study the coupling between different layers of cellular networks is a new challenge in computational biology. In this article, we will develop a novel algorithm to mine coupled network modules that occur frequently across a series of two-layered networks. We will use the rapidly accumulating RNA-seq data as the testing system to study the coupling between co-expression and co-splicing networks.

Transcription and splicing regulate gene expression differently, yet in a coordinated manner (Rosonina and Blencowe, 2002; Fong and Zhou, 2001; Kornblihtt et al., 2004). Recent reports have shown that splicing couples with transcription (Rosonina and Blencowe, 2002; Kornblihtt et al., 2004; Pandit et al., 2008), with all three splicing reactions (capping, splicing and cleavage/polyadenylation) occurring in intimate association with the elongating RNA polymerase II (Céaceres and Kornblihtt, 2002). Multiple splicing regulators are known to be linked to the transcription machinery via protein-protein interactions (Fong and Zhou, 2001; Lin et al., 2008; Pandit et al., 2008; Brès et al., 2005). Whereas most previous studies have focused on how transcription impacts splicing, Lin et al. (2008) reported the first evidence that splicing factors also affect transcription. Although mechanism study of such coupling has emerged as an active research area (Perales and Bentley, 2009; Muñoz et al., 2009; Mapendano et al., 2010; Barboric et al., 2009), the function, scope, and mechanism of the coupling has not yet been well studied. Thus far, it is not clear to what extent, by what mechanisms, and for which cellular functions, transcription-splicing coupling takes place.

From a RNA-seq dataset, we can obtain not only the expression levels of genes to construct a gene co-expression network, but also those of exons from which an exon co-splicing network can be constructed. In an exon co-splicing network, the nodes represent exons and the edge weights represent correlations between the inclusion rates of two exons across all samples in the dataset. These two kinds of networks are “coupled,” in the sense that each gene in the gene network contains several exons in the exon network. To study coupling mechanisms, in this paper, we propose to identify coupled transcription-splicing modules in a series of paired gene co-expression and exon co-splicing networks, with each pair being derived from a RNA dataset. The concept of our approach is illustrated in Figure 1. A set of co-expressed genes (i.e., heavily interconnected in the gene co-expression networks) is likely to be co-regulated by the same transcription factor and thus may represent a *transcription module*. Similarly, a set of co-spliced exons (i.e., heavily interconnected in the exon co-splicing networks) is likely to be co-spliced by the same splicing factor and thus may represent a *splicing module*. When genes of a transcription module have their exons or at least a subset of their exons to form a splicing module, it is likely to implicate the transcription and splicing coupling is taking place, and thus forms a so-called “*coupled module*.” In particular, if the coupled cluster (a co-expressed gene cluster coupled with a co-spliced exon cluster) recurrently appears across multiple paired gene and exon networks, then the enhanced signal to noise ratio would point to a higher likelihood for this *frequent coupled cluster* (FCC) to represent a *coupled module* than would those coupled clusters derived from a single dataset. We previously showed (Yan et al., 2007; Li et al., 2011) that the likelihood that a gene co-expression cluster is a transcription module increases significantly with the recurrence of clusters in multiple datasets. In this article, we hypothesize and validate that a similar principle holds for coupled modules: i.e., the likelihood that a coupled gene-exon cluster represents a coupled transcription-splicing module increases with its recurrence frequency.

To identify FCCs from a large collection of many (say > 30) paired edge-weighted co-expression and co-splicing networks, we propose a computational method based on the tensor model. Generally speaking, a tensor is a multi-dimensional array and a matrix is a 2^{nd} -order tensor. Given L RNA-seq datasets, a collection of L gene co-expression networks with the same N gene nodes but different topologies can be represented as a 3^{rd} -order tensor $\mathcal{G} = (g_{ijk})_{N \times N \times L}$ (Fig. 1). Each element g_{ijk} is the weight of the edge between genes i and j calculated from the k^{th} RNA-seq dataset. Correspondingly, a collection of L exon co-splicing networks with the same M exon nodes can be modeled as the tensor $\mathcal{E} = (e_{ijk})_{M \times M \times L}$. Representing a set of networks as a 3^{rd} -order tensor brings the following advantages: (1)

FIG. 1. Illustration of the tensor model for collections of networks. From each RNA-seq dataset, we can build a gene co-expression network and an exon co-splicing network. Because all of the gene networks share the same set of genes, the collection of gene co-expression networks can be “stacked” into a third-order tensor $\mathcal{G}$, such that each slice represents the adjacency matrix of one network. The same scenario applies to the exon co-splicing networks, which form a third-order tensor $\mathcal{E}$. Weights of the edges in each network and their corresponding entries in the tensor are color-coded according to the scale at the right of the figure. The relationships of the gene set and the exon set in the two tensors can be described by a binary relation matrix $\mathbf{R}$, in which $r_{ij} = 1$ when the i^{th} gene contains the j^{th} exon; otherwise $r_{ij} = 0$.

This model provides access to a wealth of numerical methods—particularly continuous optimization methods. In fact, reformulating discrete problems as continuous optimization problems is a long-standing tradition in graph theory. There are many successful examples, such as using a Hopfield neural network for the traveling salesman problem (Hopfield) and applying the Motzkin–Straus theorem to solve the clique-finding problem (Motzkin and Straus, 1965). (2) Advanced continuous optimization techniques require very few *ad hoc* parameters, in contrast with many heuristic graph algorithms. Both unweighted and weighted networks can be equally modeled as tensor models leading to the same tensor-based computational method; while many existing graph methods on unweighted networks cannot be easily adapted to weighted networks. (3) By transforming a graph pattern mining problem into a continuous optimization problem, it becomes easy to incorporate constraints representing prior knowledge.

As shown in Figure 2, an FCC can be described with a tensor model as follows: its gene cluster and exon cluster intuitively correspond to a heavy region of the tensor $\mathcal{G}$ and $\mathcal{E}$, respectively, which can be called as the heavy subtensor. Thus, the FCC can be found by reordering the tensors $\mathcal{G}$ and $\mathcal{E}$ simultaneously such that the heaviest subtensors move toward the top-left corner, while their constituent genes and exons keep “belong-to” relationships. The subtensor in the top-left corner can then be expanded outwards from the left-top corner until the FCC reaches its optimal size.

This is the first study to identify coupled modules from two-layer networks systematically. We applied our tensor-based method to 38 paired weighted co-expression and co-splicing networks derived from human RNA-seq data, and identified an atlas of *frequent coupled clusters*. We show that these clusters are highly likely to represent functional, transcriptional, and splicing modules. The likelihood for an FCC to be biologically meaningful increases significantly with its recurrence and heaviness. A large proportion of the identified coupled modules are involved in post-transcriptional processing, and most modules are centered on the information

FIG. 2. Illustration of the *frequent coupled cluster* (FCC). In a collection of three paired gene co-expression and exon co-splicing networks, a subset of genes $\{1,2,3\}$ are heavily interconnected and their exons $\{A,B,C,D\}$ are also heavily interconnected. These subsets form an FCC that represents the coupled transcription-splicing module. The gene and exon clusters intuitively corresponds to the heavy subtensors in $\mathcal{G}$ and $\mathcal{E}$.

transfer pathway from DNAs to proteins. We also show that an important mechanism of transcription-splicing coupling is mediated by protein-protein interactions between transcription and splicing factors.

2. METHODS

From the Sequence Read Archive (SRA) of NCBI,¹ we selected all human RNA-seq datasets, each of which contains at least six samples (the minimum for robust correlation estimation). This results in a total of 38 datasets. For each dataset, we used the Tophat (Trapnell et al., 2009) tool to map short reads to the *hg18* reference genome and applied the transcript assembly tool Cufflinks (Trapnell et al., 2010) to estimate expressions for all transcripts with known UCSC annotations (Fujita et al., 2011). We calculated the inclusion rate of each exon² in every sample as the ratio between its expression (i.e., sum of FPKM³ over all transcripts that cover the exon) and the expression of the host gene (i.e., sum of FPKM over all transcripts of the gene). Then for each dataset, we built a *weighted gene co-expression network*, in which nodes represent genes and edges are weighted by expression correlations between two genes, and a *weighted exon co-splicing network*, in which nodes represent exons and edge weights represent correlations between the inclusion rates of two exons. We then performed network normalization to make the edge weights comparable across datasets and used non-uniform edge sampling in all networks for fast computation. These procedures are detailed in Appendices A and B.

2.1. Problem formulation

A naive approach of identifying an FCC is to first identify gene heavy subgraphs which frequently occur across multiple co-expression networks, then to identify a subset of those member genes whose exons (or subset of exons) form heavy subgraphs frequently occurring in the pair-matched co-splicing networks. However, because of the fact that *a subgraph of a heavy subgraph may not be heavy*, we cannot simultaneously guarantee the heaviness of such derived gene and exon subgraphs; therefore, the naive approach is not applicable. In this article, we propose a novel computational method to identify FCCs.

Given L RNA-seq datasets from which we construct L gene co-expression networks (with the same N genes but different topologies) and their paired L exon co-splicing networks (with the same M exons but different topologies), the whole system can be represented as two 3rd-order tensors, $\mathcal{G} = (g_{ijk})_{N \times N \times L}$, $\mathcal{E} = (e_{ijk})_{M \times M \times L}$ and a binary relation matrix between N genes and M exons, $R = (r_{ij})_{N \times M}$. Each element g_{ijk} (e_{ijk}) in the tensor is the non-negative weight of the edge between genes (exons) i and j in the k^{th} gene co-expression (exon co-splicing) network. Note that $g_{iik} = 0$ ($e_{iik} = 0$) and $g_{ijk} = g_{jik}$ ($e_{ijk} = e_{jik}$) for any i, j, k , because all networks are assumed to be undirected and without self-loops. As each exon belongs to only one gene but a gene may have at least one exons, the relation matrix R has special characteristics: (1) $r_{ij} = 1$ when gene i contains exon j , $r_{ij} = 0$ otherwise; and (2) each column vector has only one non-zero and the rest are zero. Therefore, R is very sparse, with exactly M non-zeros.

An FCC is defined as follows: “a set of genes G that are frequently co-expressed in a set of datasets D (forming a heavy subgraph in multiple gene networks), and a set of their exons E that are co-spliced in the same set of datasets (forming a heavy subgraph in multiple exon networks).” Figure 2 gives an example. We formulate the problem of identifying an FCC as follows,

Definition of an FCC

An FCC consists of a set of genes G , a set of exons E and a set of datasets D that satisfy the following two criteria:

Heavy subgraph criterion: Genes of G are heavily connected to each other in each dataset of D (i.e., active dataset); and exons of E are heavily connected to each other in each of the same set of datasets.

Relation criterion: Each gene of G contains ≥ 1 exons of E , whereas each exon of E is contained by ≤ 1 genes of G .

¹www.ncbi.nlm.nih.gov/sra.

²Throughout this study, we only consider cassette exons, which are common in alternative splicing events. Henceforth, the term “exon” always means “cassette exon.”

³FPKM stands for “Fragments Per Kilobase of exon per Million fragments mapped,” as defined in Trapnell et al. (2010).

Any FCC can be described by three membership vectors: (i) the *gene membership vector* $\mathbf{x} = (x_1, \dots, x_N)^T$, where $x_i = 1$ if gene i belongs to the gene set G of the cluster and $x_i = 0$ otherwise; (ii) the *exon membership vector* $\mathbf{y} = (y_1, \dots, y_M)^T$, where $y_j = 1$ if exon j belongs to the exon set E of the cluster and $y_j = 0$ otherwise; and (iii) the *dataset membership vector* $\mathbf{w} = (w_1, \dots, w_L)^T$, where $w_k = 1$ if the cluster appears in the dataset k of D (we term such a dataset as an “active dataset” for the cluster) and $w_k = 0$ otherwise.

Using these three membership vectors, the “heavy subgraph criterion” defined above can be formulated by maximizing the “heaviness” functions of the gene and exon subgraphs in the active datasets D . The “heaviness” functions of gene subgraph $H_G(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^L g_{ijk} x_i x_j w_k$ is the summed weight of all edges of the gene subgraph in its active datasets. The “heaviness” functions of exon subgraph $H_E(\mathbf{y}, \mathbf{w}) = \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^L e_{ijk} x_i x_j w_k$ is the summed weight of all edges of the exon subgraph in its active datasets D . Only the weights of edges $x_i = x_j = 1$ ($y_i = y_j = 1$) are counted in $H_G(H_E)$. Therefore, $H_G(\mathbf{x}, \mathbf{w})$ measures the “heaviness” of a gene subgraph G , defined by the gene membership vector $\mathbf{x}$ in the set of active dataset D defined by the dataset membership vector $\mathbf{w}$. This is similarly applied to $H_E(\mathbf{y})$.

The relation criterion can be formulated by using the idea of the linear assignment problem formulation in operation science (Burkard et al., 2009). Let the relation variables $Z = (z_{ij})_{N \times M}$ indicate the matching between genes and exons, where $z_{ij} = 1$ if gene i contains exon j and both belong to the FCC and 0 otherwise. Then relation criterion can be formulated by maximizing the objective function $O_R(Z) = \sum_{r_{ij} \neq 0} z_{ij} r_{ij} x_i y_j$ with the constraint $\sum_{j=1}^M z_{ij} \leq K_2$ for all $i = 1, \dots, N$, where K_2 is the number of exons in the cluster). Therefore, maximizing $O_R(Z)$ becomes a process of finding genes and exons that have one-to-many relationships between each other and simultaneously have large associated weights x_i and y_j , because the special characteristic of R already guarantees that each exon belongs to only ONE gene.

With the above formulations of each individual criterion, discovering an FCC can be formulated as a discrete combinatorial optimization problem: *among all FCCs of fixed size (K_1 member genes, K_2 member exons, and K_3 member datasets), we look for the pattern that can arrive at the maximum of the combined objective function: $O(\mathbf{x}, \mathbf{y}, \mathbf{w}, Z) = H_G(\mathbf{x}, \mathbf{w}) + \lambda H_E(\mathbf{y}, \mathbf{w}) + \mu O_R(\mathbf{x}, \mathbf{y}, Z)$, where $\lambda, \mu > 0$ are constant weights of individual criteria. This is an integer programming problem of looking for the binary membership vectors $\mathbf{x}, \mathbf{y}, \mathbf{w}$ and the binary matrix Z that jointly maximize the combined objective function under the constraints $\sum_{i=1}^N x_i = K_1$, $\sum_{j=1}^M y_j = K_2$, and $\sum_{k=1}^L w_k = K_3$. However, there are several major drawbacks to this discrete formulation. The first is *parameter dependence*: as with K -heaviest/densest subgraph problems, the size parameters K_1, K_2 , and K_3 are difficult for users to provide and control. The second is *high computational complexity*: the task is NP-hard, as indicated in Lemma 1 (see Appendix C for its proof), and therefore is not solvable in reasonable time even for small datasets. As our own interest is pattern mining in a large set of massive networks, the discrete optimization problem is infeasible.*

Lemma 1. *The problem of maximizing $O(\mathbf{x}, \mathbf{y}, \mathbf{w}, Z) = H_G(\mathbf{x}, \mathbf{w}) + \lambda H_E(\mathbf{y}, \mathbf{w}) + \mu O_R(\mathbf{x}, \mathbf{y}, Z)$ with the constraints $\sum_{i=1}^N x_i = K_1$, $\sum_{j=1}^M y_j = K_2$, $\sum_{k=1}^L w_k = K_3$ and $\sum_{j=1}^M z_{ij} \leq K_2$ for all $i = 1, \dots, N$ is NP-hard.*

To address these two drawbacks, we instead solved a continuous optimization problem with the same objective by relaxing integer constraints to continuous constraints. That is, we looked for non-negative real vectors $\mathbf{x}, \mathbf{y}, \mathbf{w}$ and the non-negative real sparse matrix Z that jointly maximize $O(\mathbf{x}, \mathbf{y}, \mathbf{w}, Z)$. This optimization problem is formally expressed as follows:

$$\begin{aligned} & \max_{\mathbf{x}, \mathbf{y}, \mathbf{w}, Z \in \mathbb{R}^+} O(\mathbf{x}, \mathbf{y}, \mathbf{w}, Z) = H_G(\mathbf{x}, \mathbf{w}) + \lambda H_E(\mathbf{y}, \mathbf{w}) + \mu O_R(\mathbf{x}, \mathbf{y}, Z) \\ & \text{subject to } \begin{cases} \text{Constraint I: } \|\mathbf{x}\|_f = 1, & 1 \leq f < 2 \\ \text{Constraint II: } \|\mathbf{y}\|_g = 1, & 1 \leq g < 2 \\ \text{Constraint III: } \|\mathbf{w}\|_h = 1, & h \geq 2 \\ \text{Constraint IV: } \sum_{j=1}^M z_{ij} \leq 1, & i = 1, 2, \dots, N \end{cases} \end{aligned} \quad (1)$$

where $\mathbb{R}^+$ is a non-negative real space of vectors or matrix, and $\|\mathbf{x}\|_p = (\sum_i |x_i|^p)^{1/p}$ is the L_p norm of the vector $\mathbf{x}$. The $L_p = 1$ ($1 \leq p < 2$) norm constraint of genes and exons (Constraints I and II above) is a well-known convex sparse coding scheme (Li et al., 2011; Ding et al., 2006) that can lead to the sparse solution where only a few elements of the gene (exon) membership vector $\mathbf{x}$ ($\mathbf{y}$) are significantly different from zero and therefore their corresponding genes (exons) can be selected into the module. Therefore it is

well suited to our purpose of selecting only a small number of genes (exons) into the module. On the contrary, the $L_h = 1$ ($h \geq 2$) norm constraints of datasets (Constraint III) gives a “smooth” solution where the elements of the optimized vector $\mathbf{w}$ are approximately equal and therefore lead to the discovered coupled cluster occurring in as many datasets as possible. This problem definition and formulation can be viewed as the generalization of the recurrent heavy subgraph (RHS) problem defined in Li et al. (2011). However, requiring mapping relationships between gene subgraph and exon subgraph by introducing a relation criterion to the problem makes it much more complicated than the RHS problem. Such nature of simultaneously analyzing two different tensors makes it distinct from the problem defined in the single tensor.

Eq. (1) defines a tensor-based optimization formulation for the problem of identifying FCCs. By solving Eq. (1), users can easily identify the top-ranking datasets (after sorting elements of $\mathbf{w}$ in non-increasing order) and top-ranking genes/exons (after sorting elements $\mathbf{x}/\mathbf{y}$ in non-increasing order) contributing to the objective function. After rearranging the tensors in this manner, the optimum FCC occupies the corner of the 3D tensor $\mathcal{G}$ and $\mathcal{E}$. We then mask the edges contained in this cluster in both gene and exon networks with zeros and optimize Eq. (1) again to search for the next module. The constant weights λ and μ in the objective function provide us the opportunity to control the importance of criteria and data. To fully exploit such capability, we collect all FCCs discovered in all combinations⁴ of different λ and μ , then remove duplicates. Details of the optimization procedure and pattern extraction process will be presented in the next section. Therefore, our problem formulation requires only three parameters f , g and h which can be fixed through simulation study.

We performed simulation studies to determine suitable values for the parameters f , g , and h by applying our tensor method to collections of random weighted networks. In subsets of these networks, we randomly placed FCC of varying size, occurrence, and heaviness. We then tested different combinations of f , g , and h , and adopted the combination ($f = g = 1$, $h = 10$) that led to the discovery of the most FCCs. More details on these simulations are provided in Appendix D.

2.2. Optimization algorithm and pattern extraction

We derived an iterative algorithm to maximize the objective function with the constraints defined in Eq. (1). This algorithm repeatedly applies the four updating steps (shown below) until the objective function converges to a fixed point, yielding the solution to Eq. (1). Each step of this procedure individually updates $\mathbf{x}$ ($\mathbf{y}$ or $\mathbf{w}$ or Z) while fixing the other three variables.

ALGORITHM 1 FOR SOLVING EQ. (1)

$$\begin{aligned}
 \text{Step 1 : } x_i &\leftarrow \left(\frac{x_i \sum_{j,k} g_{ijk} x_j w_k + \mu [(Z \cdot R) \mathbf{y}]_i}{\sum_{i,j,k} g_{ijk} x_i x_j w_k + \mu \mathbf{x}^T (Z \cdot R) \mathbf{y}} \right)^{1/f} & i = 1, \dots, N \\
 \text{Step 2 : } y_j &\leftarrow \left(\frac{y_j \frac{\lambda \sum_{i,k} e_{ijk} y_i w_k + \mu [(Z \cdot R)^T \mathbf{x}]_j}{\lambda \sum_{i,j,k} e_{ijk} y_i y_j w_k + \mu \mathbf{x}^T (Z \cdot R) \mathbf{y}}}{\lambda \sum_{i,j,k} e_{ijk} y_i y_j w_k + \mu \mathbf{x}^T (Z \cdot R) \mathbf{y}} \right)^{1/g} & j = 1, \dots, M \\
 \text{Step 3 : } w_k &\leftarrow \left(\frac{w_k \frac{\sum_{i,j} g_{ijk} x_i x_j + \lambda \sum_{i,j} e_{ijk} y_i y_j}{\sum_{i,j,k} g_{ijk} x_i x_j w_k + \lambda \sum_{i,j,k} e_{ijk} y_i y_j w_k}}{\sum_{i,j,k} g_{ijk} x_i x_j w_k + \lambda \sum_{i,j,k} e_{ijk} y_i y_j w_k} \right)^{1/h} & k = 1, \dots, L \\
 \text{Step 4 : } z_{ij} &\leftarrow z_{ij} \frac{[R \cdot (\mathbf{xy}^T)]_{ij}}{[R \cdot (\mathbf{xy}^T)]_{i*}^T Z_{i*}} & \text{for all } i, j \text{ whose } [R]_{ij} \neq 0
 \end{aligned}$$

Note: $[\mathbf{x}]_i$ is the i^{th} element of the vector $\mathbf{x}$. $[X]_{ij}$ is the element of the matrix X at the i^{th} row and the j^{th} column, $[X]_{i*}$ denotes the i^{th} row vector of the matrix X , and $Z = X \cdot Y$ is an element-wise matrix multiplication where $z_{ij} = x_{ij} y_{ij}$.

⁴In practice, we used the following combinations: λ is any of the six predefined values $\{0.01, 0.05, 0.1, 1, 10, 50\}$, and μ is any of the five predefined values $\{10, 20, 50, 100, 500\}$.

The following lemmas guarantee the correctness and convergence of this algorithm (see Appendices E and F for their proofs).

Lemma 2 (Correctness). *The fixed point of the updating steps in Algorithm 1 satisfies the Karush-Kuhn-Tucker (KKT) condition of Eq.(1).*

Lemma 3 (Convergence). *The objective function $O(\mathbf{x}, \mathbf{y}, \mathbf{w}, Z)$ is nondecreasing under the updating rules in Algorithm 1. The objective function is invariant under these updates if and only if $\mathbf{x}, \mathbf{y}, \mathbf{w}$, and Z are at a stationary point of the objective function.*

As the relation matrix R is extremely sparse (with only M non-zeros), all element-wise matrix multiplication operations involving R and Z in Algorithm 1 could be implemented efficiently by only treating z_{ij} whose corresponding $r_{ij} \neq 0$ and simply assigning the rest $z_{ij} = 0$ whose corresponding $r_{ij} = 0$. Thus, the complexity of Algorithm 1 is linear with respect to the total number of edges in two tensors $\mathcal{G}$ and $\mathcal{E}$.

According to the definition of three aforementioned membership vectors $\mathbf{x}, \mathbf{y}$, and $\mathbf{w}$, FCCs can be intuitively obtained by including those top-ranking genes, exons, and datasets with large membership values (after sorting $\mathbf{x}, \mathbf{y}$ and $\mathbf{w}$ in descending order, respectively). In this article, we empirically define an FCC extracted from these top-ranking genes, exons and datasets as a triplet (G', E', D') , i.e., a set of ≥ 5 genes G' , a set of ≥ 5 exons E' and a set of ≥ 2 active datasets D' satisfying: (1) each gene subgraph formed by G' and exon subgraph formed by E' in each active dataset of D' has an average edge weight (or so-called “heaviness”) $\geq$ a predefined threshold (i.e., ≥ 0.4) and (2) the fraction of host genes of E' that are included in G' , so-called $coverage_{gene}(G', E') = \frac{|G' \cap hostgene(E')|}{|G'|} \geq$ a predefined threshold (i.e., ≥ 0.7) and the fraction of exons of G' which are included in E' , so-called $coverage_{exon}(G', E') = \frac{|E' \cap exon(G')|}{|E'|} \geq$ a predefined threshold (i.e., ≥ 0.7). The first criterion gives a certain degree of how heavy and frequent the cluster should be, and the second criterion gives a measure of what degree of “coupling” should exist between a gene set and an exon set. Two examples of coupled gene set and exon set with $coverage_{gene} > 0.6$ and $coverage_{exon} > 0.6$ are shown in Figure 3. Based on this definition, we can generate one or more FCCs that are highly overlapping with each other from combinations of different sizes of top-ranking genes, exons, and datasets which satisfy above two criteria. Therefore, we call a group of FCCs derived from the same membership vectors $\mathbf{x}, \mathbf{y}$, and $\mathbf{w}$ as a family of FCCs.

3. RESULTS

After we applied our method to 38 paired gene co-expression networks and exon co-splicing networks derived from RNA-seq datasets, we identified 8,667 FCC families containing the total 43,580 FCCs. Each FCC contains ≥ 5 member genes, ≥ 5 member exons, appears in ≥ 2 RNA-seq datasets, has a “heaviness” ≥ 0.4 , “ $coverage_{gene} \geq 0.7$ ” and “ $coverage_{exon} \geq 0.7$.” The average gene/exon size of these patterns are 12.61/12.64 and the average recurrence is 2.04. To assess the statistical significance of the identified FCCs, we applied our method to 38 paired random networks (each of which is generated from one of the 38 paired weighted networks by the edge randomization method⁵) to identify FCCs with the same properties of size, heaviness, and coupling coverage. We repeated this process 10 times, and each

FIG. 3. Illustration of the coupling degree between a gene set and an exon set. Blue edges between genes and exons indicate that there are relationships between them. Red edges within genes or exons indicate that they are co-expressed or co-spliced. (A) Perfect coupling. (B) Coupling with $coverage_{gene} > 0.6$ and $coverage_{exon} > 0.6$.

FIG. 4. FCCs are highly likely to represent functional modules. The enrichment fold ratio, calculated by dividing the percentage of functionally homogenous FCC families by the percentage of functionally homogenous random patterns, increases significantly with the recurrence and heaviness of the FCCs. (A) The minimum number of active datasets in which FCCs recur is varied while fixing the minimum heaviness as 0.4. (B) The minimum heaviness of FCCs is varied while fixing the minimum number of active datasets as 2.

time only 0 ~ 3 FCC families (average 0.9 families) were identified. This is an extremely low value compared to 8,667 FCC families discovered in real data that indicates the significance of our results (detail is provided in Appendix G).

To assess the biological significance of the identified FCCs, we evaluate the extent to which these FCCs represent functional modules, protein complexes, transcriptional, and splicing modules. Since FCCs within the same family are highly overlapping, we treat a family of FCCs as a unit in these following analyses. For example, we denote a family of FCC to be functionally homogenous if any of its containing FCCs is significantly enriched in genes of the same functional category.

3.1. Frequent coupled clusters are likely to represent functional modules and protein complexes

We evaluated the functional homogeneity of the merged genes and host genes of exons in a *frequent coupling clusters* using specific Gene Ontology (GO) terms (Ashburner et al., 2000). To ensure the specificity of GO terms, we filtered out those general terms associated with >300 genes. If the gene set of any FCCs within a family is significantly enriched in a GO term with a q -value <0.05 (the q -value is the hypergeometric p -value after a False Discovery Rate multiple testing correction), we declare this family as functionally homogeneous. According to this definition, 50.4% of the FCC families was functionally homogenous. In an ensemble of randomly generated FCC families with the same size distribution as our patterns, only 5.7% of the families was functionally homogenous. This enrichment fold ratio of 8.8 between real and random patterns demonstrates the strong biological relevance of the identified patterns. Furthermore, the enrichment fold ratio increases dramatically with increasing heaviness and recurrence of FCCs (Fig. 4). For example, when the FCCs are required to recur in at least 4 datasets, their enrichment fold ratio compared to random patterns increases to 48.0, highlighting the benefit of pursuing integrative analysis of *multiple* RNA-seq datasets. When the FCCs are required to be heavier than 0.6, their enrichment fold ratio raises to 65.0.

Functional modules represented by FCCs cover a large number of categories related to post-transcriptional processing, such as “*regulation of mRNA stability*,” “*posttranscriptional regulation of gene expression*,” “*RNA polyadenylation*,” “*RNA splicing*,” “*mRNA processing*,” “*RNA transport*,” “*establishment of RNA localization*,” “*mRNA 3'-end processing*,” “*nucleic acid transport*,” and “*RNA transport*.” This is not surprising, since recent experimental studies revealed the existence of factories in which individual machines for the transcription, post-transcription processing, and mRNA transport are functionally coupled (Reed, 2003). During expression of protein-coding genes, pre-mRNAs are transcribed in the nucleus and undergo several processing steps, including capping, splicing, 3'-end processing and polyadenylation, prior to be transported to the cytoplasm for translation. The coupling between those individual steps is believed to enable the proofreading and streamlining of the entire process of gene expression *in vivo* (Reed, 2003). Other

⁵Given a real weighted network, the random weighted network is generated by a random redistribution of the actual weights on the random unweighted network with the same number of edges which can be generated by the widely used degree-preserving randomization procedure (Maslov and Sneppen, 2002). Detail is provided in Appendix G.

FIG. 5. FCCs are highly likely to represent protein complexes. The enrichment fold ratio increases significantly with the recurrence and heaviness of the FCCs. (A) The minimum number of active datasets in which FCCs recur is varied while fixing the minimum heaviness as 0.4. (B) The minimum heaviness of FCCs is varied while fixing the minimum number of active datasets as 2.

functional categories over-represented by FCCs include “nucleosome organization,” linking to the recent evidences that the transcription elongation, an important step in transcription-splicing coupling, can be regulated by nucleosome (Lindstrom et al., 2003).

Similar results were achieved by using the Comprehensive Resource of Mammalian protein complexes (CORUM) database (Ruepp et al., 2010) to assess the association between FCCs and known protein complexes. The merged sets of genes and exon-host-genes of 78.6% of FCC families are significantly enriched in genes from the same complexes with a q -value < 0.05 , compared to only 5.05% of randomly generated patterns with the same size distribution. The enrichment fold ratio dramatically increases with increasing heaviness and recurrence of the FCCs (Fig. 5).

Protein complexes represented by FCCs generally fall into the following categories: (1) *Splicing regulation*, e.g. *SNW1* complex, *CDC5L* complex, C complex spliceosome; (2) *Transcription regulation*, e.g., *SNW1* complex, *TLE1* complex, *H2AX* complex; (3) *microRNA processing*, e.g., Large Drosha complex, *DGCR8* multiprotein complex; (4) *DNA or chromatin processing*, e.g., *DNA-PK-Ku-eIF2-NF90-NF45* complex, Emerin complex; (5) *Translation*, e.g., *eif* complex, *Nop56p*-associated pre-rRNA complex; and (6) *Protein folding*, e.g., *CCT* micro-complex. Interestingly, the functions of all of those complexes are exclusively centered on the central-dogma-process, namely, the sequential information transfer from DNAs to proteins. Likely, all functions under coupled transcription and splicing regulation are those that need strict and precise regulation to ensure efficient and error-free information processing. Some of those complexes have been experimentally confirmed to couple splicing and transcription. For example, the *SNW1* complex contains components of the spliceosome, and it is known to couple vitamin D receptor-mediated transcription and RNA splicing (Zhang et al., 2003). Since research on such coupling is still in its infancy, we expect that more protein complexes will be shown to link those individual regulatory steps into a coherent and smooth production line.

3.2. Frequent coupled clusters are likely to represent transcription and splicing modules

Because the gene set of an FCC are strongly co-expressed in multiple datasets generated under different conditions, it is likely to represent a transcription module. To assess this possibility, we used the 191 ChIP-seq profiles generated by the Encyclopedia of DNA Elements (ENCODE) consortium (Thomas et al., 2007). This dataset includes the genome-wide binding of 40 transcription factors (TF), 9 histone modification marks, and 3 other marks (DNase, FAIRE, and DNA methylation) on 25 different cell lines. For a detailed description of the signal extraction procedure, see Appendix H. These data provide the potential targets of regulatory factors that may or may not be active under a specific condition. However, if the gene set of an FCC is found to be highly enriched in the targets for a regulatory factor, then this factor is likely to actively regulate genes in this FCC, and we declare the family containing this FCC to be “transcriptionally homogenous.” According to this definition, 61.2% of FCC families are transcriptionally homogenous with an enrichment q -value < 0.05 , compared to 8.9% of the randomly generated patterns with the same size distribution. As shown in Figure 6, with the increasing heaviness and recurrence of FCCs, the enrichment fold ratio between FCCs and random patterns dramatically increases. The five most frequently enriched regulators are *YY1*, *E2F4*, *c-Myc*, *MAX*, and *TAF*, all of which have been implicated in cancer pathogenesis or progression (Gordon et al., 2005; Nevins, 2001; Little et al., 1983; Nair and Burley, 2003; Suzuki et al., 2010).

FIG. 6. FCCs are highly likely to represent transcription modules. The enrichment fold ratio increases significantly with the recurrence and heaviness of the FCCs. **(A)** The minimum number of active datasets in which FCCs recur is varied while fixing the minimum heaviness as 0.4. **(B)** The minimum heaviness of FCCs is varied while fixing the minimum number of active datasets as 2.

Since the exon set of an FCC has highly correlated inclusion rate profiles across different experimental conditions, they are likely to be co-regulated by the same splicing factors. We collected the experimental RNA target motifs (2,220 RNA binding sites) of 62 splicing factors from the SpliceAid2 database (Piva et al., 2011). For each exon in an FCC, we retrieved the internal exon region and its 100bp flanking intron region which are enriched in the motifs of those 62 splicing proteins by performing BLAST search (E-score <0.05). If the exon set of any FCC within a family is significantly (q -value <0.05) enriched in the targets of a splicing factor, we consider this family to be “splicing homogenous.” Although the collection of known splicing motifs is very limited, we still observed that 8.9% of the FCC families are splicing homogenous, compared to 5.1% of randomly generated patterns with the same size distribution. Again, as shown in Figure 7, with the increasing heaviness and recurrence of the FCCs, the enrichment fold ratio increases as well. The five most frequently enriched splicing regulators are *PSF*, *hnRNP-D*, *hnRNP-C1*, *SLM-2*, and *HuB*.

3.3. Exploring the mechanisms of the transcription and splicing coupling

Recently, individual case studies have suggested that the strong connection between transcription and splicing might be due to (a) the coupled recruitment of transcription and splicing factors and (b) the kinetic elongation control by RNA *Pol II* (Kornblihtt et al., 2004). However, global analyses of these hypotheses have not been conducted. Our catalogues of coupled transcription-splicing modules provide a unique opportunity for us to systematically assess the mechanism (a). In particular, we evaluate the hypothesis that the functionally coupled recruitment of both types of factors could be mediated by the direct or indirect protein-protein interactions between them. This hypothesis is supported by the following biological evidences: (1) the presence of an intron or simply a 5' splice site immediately downstream from a promoter, i.e., the short distance between the splicing and transcription factors, greatly enhances transcription, both in mammalian and yeast genes (Kornblihtt et al., 2004); and (2) human spliceosome contains at least 30 proteins with known or putative roles involved in transcription, presenting direct interactions between the transcription and splicing (Zhou et al., 2002; Rappsilber et al., 2002).

We used protein-protein interaction (PPI) repository BioGRID (Stark et al., 2006) to examine the PPIs between enriched transcription and splicing factors in coupled module families. We consider not only the

FIG. 7. FCCs are likely to represent splicing modules. The enrichment fold ratio increases significantly with the recurrence and heaviness of the FCCs. **(A)** The minimum number of active datasets in which FCCs recur is varied while fixing the minimum heaviness as 0.4. **(B)** The minimum heaviness of FCCs is varied while fixing the minimum number of active datasets as 2.

direct PPIs between two factors, but also the indirect interactions through a mediator protein (one-hop interaction), because experiments showed that transcription and splicing factors could be recruited by the same other proteins (Nayler et al., 1998; Blencowe et al., 1999). In order to have a broad coverage of transcription factors, we used 109 human transcription factors in ENCODE (Thomas et al., 2007) and JASPAR database (Sandelin et al., 2004), where 10,278 DNA binding motifs are downloaded from Tabach et al. (2007), with which we performed BLAST search (E-score < 0.05) in 1000bp regions of the transcription start sites of all genes, and then used hyper-geometry test (q -value < 0.05) to identify enriched transcription factors for genes in an FCC. The putative splicing factors for each FCC were identified based on the description in Section 3.2. The total number of PPIs between transcription and splicing factors within the same FCC families (including direct and one-hop interactions) is 105, compared to only 14.8 in random families with the same size of gene and exon sets, with the fold ratio 7.1. This evidence supports that PPI mediated association can be an important mechanism of transcription-splicing coupling.

An interesting example is the FCC family associated with the splicing regulator *PSF* and the transcription regulator *TAT-SF1*. *PSF* is a part of the human spliceosome and essential in RNA splicing (Peng et al., 2006; Garcia-Jurado et al., 2011). *TAT-SF1* functions as a general transcription factor playing a role in the process of transcriptional elongation (Zhou and Sharp, 1996). Based on the kinetic model that transcription rate affects splicing outcomes during transcription elongation (Kornblihtt et al., 2004), previous studies suggested that *TAT-SF1* couple with *PSF* for co-transcriptional splicing. In fact, the spliceosome has been shown to be associated with the transcription elongation complexes (Kameoka et al., 2004), and *TAT-SF1* was also characterized to be part of the spliceosome (Rappsilber et al., 2002). We found that *TAT-SF1* and *PSF* both interact with the protein *WBP4*, a general spliceosomal protein that may play a role in cross-intron bridging of *U1* and *U2 snRNPs* (Bedford et al., 1998). Thus, we hypothesize that the elongation factor *TAT-SF1* may functionally couple with *PSF* through the mediator *WBP4* in spliceosome.

As another example, the co-spliced exons of an FCC family is enriched in the RNA binding motif of the splicing factor *hnRNP D*, and the co-expressed genes of the same FCC family is over-represented by the DNA binding motif of the transcription factor *TDP43*. *hnRNP D* belongs to heterogeneous nuclear ribonucleoprotein family, which generally suppress splicing through binding to exonic splicing silencer (Matlin et al., 2005). *TDP43* was originally identified as a transcriptional repressor that binds to chromosomally integrated *TAR* DNA and represses *HIV-1* transcription (Ou et al., 1995). Interestingly, *TDP43* was later shown to also have splicing regulatory functions, wherein the recruitment of *TDP43* inhibits exon recognition (Buratti and Baralle, 2001). We suspect that *TDP43* may couple with *hnRNP D* through *hnRNP A1*, because there is a direct protein-protein interaction between *hnRNP A1* and *hnRNP D/TDP43* (Stark et al., 2006), whereas *TDP43-hnRNP A1* interaction is suggested to be mediated by C-terminal region of *TDP43* (Ayala et al., 2010). Furthermore, experiments have shown that *TDP43*, *hnRNP D* and *hnRNP A1* are all shuttled between nucleus and cytoplasm (Loflin et al., 1999; Ayala et al., 2008). Thus, it is very possible that *TDP43* and *hnRNP* couple the transcription-splicing process via *hnRNP A1*.

4. CONCLUSION

We have developed a novel tensor-based approach to identify frequent coupled modules in many two-layered networks. This is the first computational method for pattern mining across multi-layered network series. Advancement of genomic technologies resulted in a rapid accumulation of multi-dimensional genomic data that simultaneously profile molecular activities at a different level (e.g., gene expression, DNA methylation, histone modifications, protein abundances) of the same biological samples. These data provides unique opportunities and challenges to study the complex coupling mechanisms among multiple layers of biological networks. Our work addresses this challenge timely. By using an elegant tensor-based problem formulation, our approach can be easily extended to mine K -layered ($K > 2$) network series and can conveniently integrate additional pattern constraints.

We applied our method to study the coupling between transcription and splicing, an area that recently attracted much attention of experimental biologists but thus far focused only on individual case studies. Our comprehensive catalogue of coupled transcription-splicing modules provides a unique resource to globally examine the scope, functions, and mechanisms of the coupling. Although the current collection of transcription and splicing factor binding information is very limited, we were able to draw meaningful inferences. For example, the strongly coupled transcription-splicing modules mainly participate in processes that need strict

and precise regulation, e.g. the sequential information transfer from DNAs to proteins. We also show that protein-protein interactions couple transcription and splicing factors in many identified modules. With the continued growth of RNA-seq data repositories and regulatory information, we expect that our method, using only public data, can provide many more interesting hypotheses for further targeted experimental studies.

5. APPENDIX

A. RNA-seq datasets selection and processing, and network construction

From NCBI's Sequence Read Archive (SRA), we selected all human RNA-seq datasets, each of which contains at least six samples (the minimum for robust correlation estimation). This results in a total of 38 datasets. For each dataset, we used the Tophat (Trapnell et al., 2009) 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 (Trapnell et al., 2010) to estimate expressions for all transcripts with known UCSC transcription annotations (Fujita et al., 2011). We calculated the inclusion rate of each exon in every sample, as the ratio between its expression (i.e., sum of FPKM over all transcripts that cover the exon) and the expression of the host gene (i.e., sum of FPKM over all transcripts of the gene). It is worth noting that in RNA-seq experiments, a gene expression with low FPKM 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 Jiang and Wong (2009), we calculated inclusion rates only for those exons whose host genes' expressions are above 80th percentile across at least six samples. Throughout this study, we only considered the genes containing cassette exons whose inclusion rate profiles met the above criterion. This resulted in 12,287 exons covering 8,196 genes. The 38 datasets that met these criteria on January 30, 2011 were used for the analysis described herein.

For each RNA-seq dataset containing a set of samples, a *weighted gene co-expression network* can be constructed by representing genes as nodes and edges weights as correlations between the expression profiles of two genes. The same procedure is applied to build a *weighted exon co-splicing network* for the same RNA-seq dataset, in which nodes represent exons and edge weights represent correlations between the inclusion rates of two exons. To determine the weights, we first compute the correlation between two genes (exons) as the leave-one-out Pearson correlation coefficient estimate (Zhou et al., 2005). The resulting correlation estimate is conservative and sensitive to similarities in the patterns, yet robust to single experimental outliers. To make the correlation estimates comparable across datasets, we then applied Fisher's z transform (Anderson, 2003). Given a correlation estimate r , Fisher's transformation score is calculated as $z = \frac{\sqrt{n-3}}{2} \ln\left(\frac{1+r}{1-r}\right)$. Note that the sample size n may be different for different datasets and even for different gene (exon) pairs due to missing values. Practically, we observed the distributions of z -scores may still vary from dataset to dataset, we standardized the z -scores to enforce zero mean and unit variance in each dataset by following the normalization procedure introduced in Xu et al. (2008) and Li et al. (2011). Then, the "normalized" correlations r' are obtained by inverting the z -score with the same virtual sample size $n' = 10$: $r' = \frac{\exp(\frac{2}{\sqrt{n'-3}}z) - 1}{\exp(\frac{2}{\sqrt{n'-3}}z) + 1}$. Finally, the absolute value of r' is used as the edge weight of networks.

B. Non-uniform sampling for fast computation

Even though our optimization method is efficient, its computation time can still be long for large sets of networks with many edges. In such cases, edge sampling can provide an efficient approximation to many graph problems (Tsay et al., 1999; Achlioptas and McSherry, 2007). From the perspective of matrix or tensor computation, such sampling methods can also be viewed as matrix/tensor sparsification (Arora et al., 2006). As FCCs predominately contain edges with large weights, we designed a non-uniform sampling method that preferentially selects edges with large weights. Specifically, given a tensor $\mathcal{A}$ which could be $\mathcal{G}$ (or $\mathcal{E}$), each edge a_{ijk} is sampled with probability p_{ijk} :

$$p_{ijk} = \begin{cases} 1, & \text{if } a_{ijk} \geq \tilde{a} \\ p(\frac{a_{ijk}}{\tilde{a}})^b, & \text{if } a_{ijk} < \tilde{a} \end{cases} \quad (2)$$

where $\tilde{a} \in (0, 1)$, $b \in [1, \infty)$ and $p \in (0, \tilde{a}^b]$ are constants that control the number of sampled edges. Note that Eq. (2) *always* samples edges with weights $\geq \tilde{a}$. It selects an edge of weight $a_{ijk} < \tilde{a}$ with probability

p_{ijk} proportional to the b^{th} power of the weight. We choose $\tilde{a}=0.4$, $b=3$, and $p=0.1$ as a reasonable tradeoff between computational efficiency and the quality of the sampled tensor, meanwhile satisfying the conditions of Theorem 4, i.e., $p\tilde{a}^{b-1}$ and $b \geq 1$.

To correct the bias caused by this sampling method, the weight of each edge is corrected by its relative probability: $\hat{a}_{ijk} = a_{ijk}/p_{ijk}$. The expected weight of the sampled network, $E(\hat{a}_{ijk})$, is therefore equal to the weight of the original network. However, in practice, when the adjusted edge weight $\hat{a}_{ijk} > \tilde{a}$ (but the original edge weight $a_{ijk} < \tilde{a}$), we enforced it to be $\hat{a}_{ijk} = \tilde{a}$ to avoid too large edge weights. The overall edge sampling procedure adopts the simple random-sampling based single-pass sparsification procedure introduced in Arora et al. (2006). Details of the sampling procedure is given below. This single-pass sampling procedure's time complexity is $O(N^2L)$. It is obvious that the sparsification procedure in Arora et al. (2006) is a special case of our sampling procedure when all entries of $\mathcal{A}$ are non-negative and $p=1$, $\tilde{a} = \frac{\epsilon}{N+N+L}$, $b=1$. After edge sampling, the procedure described above will use the corrected tensor $\hat{\mathcal{A}} = (\hat{a}_{ijk})_{N \times N \times L}$ instead of the original tensor $\mathcal{A}$.

Edge sampling procedure of the tensor $A = (a_{ijk})_{N \times N \times L}$
Procedure SAMPLING($\mathcal{A}, p, \tilde{a}, b$)
for each $i \in [1, N], j \in [1, N], k \in [1, L]$ **do**
 if $a_{ijk} \geq \tilde{a}$ **then**
 $\hat{a}_{ijk} = a_{ijk}$
 else
 $\hat{a}_{ijk} = \begin{cases} \min(\frac{a_{ijk}}{p_{ijk}}, \tilde{a}), & \text{with probability } p_{ijk} = p \left(\frac{a_{ijk}}{\tilde{a}} \right)^b \\ 0 & \text{with probability } 1 - p_{ijk} \end{cases}$
return $\hat{\mathcal{A}} = (\hat{a}_{ijk})_{N \times N \times L}$

Based on the well-known Chernoff-Hoeffding bounds (Motwani and Raghavan, 1995), we gave the bound of the non-zero entries in the corrected tensor $\hat{\mathcal{A}}$ after sampling in Lemma 4. Therefore, the computational complexity of the proposed algorithm on the tensors $\hat{\mathcal{G}}$ and $\hat{\mathcal{E}}$ after sampling is linear to the number of the non-zero entries of $\hat{\mathcal{G}}$ and $\hat{\mathcal{E}}$.

Lemma 4. Given $p\tilde{a}^{b-1}$, $b \geq 1$ and $0 \leq a_{ijk} \leq 1$ (for any i, j, k), with probability at least $1 - \exp(-\Omega(S'))$, the tensor $\hat{\mathcal{A}}$ contains at most $O(\frac{S}{\tilde{a}})$ non-zero entries, where $S = \sum_{i,j,k} a_{ijk}$ and $S' = \frac{p}{\tilde{a}^b} \sum_{i,j,k} a_{ijk}^b$.

Proof. This proof is similar to Lemma 1 in Arora et al. (2006). Since $S = \sum_{i,j,k} a_{ijk}$, the number of a_{ijk} that are no less than $\tilde{a}$ is at most $\frac{S}{\tilde{a}}$; otherwise, the sum of all entries $\geq \tilde{a}$ would be greater than S . Now consider all non-zero a_{ijk} entries that are smaller than $\tilde{a}$. ■

The Chernoff bound (Motwani and Raghavan, 1995) asserts that if $X_1, X_2, \dots, X_N$ are indicator random variables and $X = \sum_i X_i$ with $\mathbb{E}[X] = \mu$, then for any $\delta > 0$

$$\Pr[X > (1+\delta)\mu] < \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\mu \quad (3)$$

In our case, we set up indicator random variables X_{ijk} which are 0 or 1 depending on whether $\tilde{a} = 0$ or not. Then $X = \sum_{i,j,k} X_{ijk}$ is the number of non-zero entries of $\hat{\mathcal{A}}$, and

$$\mu = \mathbb{E}[X] = \sum_{i,j,k} p_{ijk} = \frac{p}{\tilde{a}^b} \sum_{i,j,k} a_{ijk}^b = S' \quad (4)$$

Since $p \leq \tilde{a}^{b-1}$ and $a_{ijk}^b \leq a_{ijk}$ (because $b \geq 1$ and $0 \leq a_{ijk} \leq 1$), we have $\frac{p}{\tilde{a}^b} \sum_{i,j,k} a_{ijk}^b \leq \frac{1}{\tilde{a}} \sum_{i,j,k} a_{ijk}$. Then, we can get

$$\mu = \frac{p}{\tilde{a}^b} \sum_{i,j,k} a_{ijk}^b \leq \frac{1}{\tilde{a}} \sum_{i,j,k} a_{ijk} = \frac{S}{\tilde{a}} \quad (5)$$

Then we have

$$\Pr \left[X(1+\delta) \frac{S}{a} \right] \geq \Pr[X \leq (1+\delta)\mu] \geq 1 - \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\mu \quad (6)$$

We use the Chernoff bound with $(1 + \delta) = e$ and arrive at the following

$$\Pr \left[Xe \frac{S}{a} \right] \geq 1 - \exp(-\mu) = 1 - \exp(-S') \quad (7)$$

So the claim holds. ■

C. Proof of Lemma 1

Proof. We can reduce the well-known NP-complete K -clique problem⁶ (i.e., *is there a clique with K nodes in a graph?*) to this problem and therefore prove its NP-hardness.

Let $G(V, E)$ be an undirected unweighted graph. We can copy this graph L times to generate a graph set consisting of the L graphs $\mathcal{G} = \{G_1(V, E), \dots, G_m(V, E)\}$ in which all graphs have the same vertex set V and edge set E . Then we consider our problem with $\lambda = \mu = 0$ in the set of L graphs $\mathcal{G}$. In other words, our problem with $\lambda = \mu = 0$ can be rewritten as follows,

$$\begin{aligned} \max \quad & H_G(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^L g_{ijk} x_i x_j w_k \\ \text{subject to} \quad & \begin{cases} \sum_{i=1}^N x_i = K_1 \\ \sum_{i=1}^L w_j = K_3 \\ x_i \in \{0, 1\} & \text{for any } 1i \leq N \\ w_j \in \{0, 1\} & \text{for any } 1j \leq L \end{cases} \end{aligned} \quad (8)$$

The solution vectors $\mathbf{x}^*$ and $\mathbf{w}^*$ of this problem defines a module having K_1 nodes whose corresponding $x_i = 1$ and K_3 networks whose corresponding $w_j = 1$. Therefore, this module can be called as “recurrent dense subgraph” because its K_1 nodes are most densely interconnected in the K_3 networks. Then the question of “*is there a clique with K vertices in G ?*” can be answered by solving above defined problem, because we can easily claim:

- If the “recurrent dense subgraph” with K nodes and K_3 networks found in the graph set $\mathcal{G}$ is a clique with K vertices recurring in the K_3 graphs, then there exist a clique with K vertices in G . This claim is obvious and straightforward.
- If the “recurrent dense subgraph” with K nodes and K_3 networks found in the graph set $\mathcal{G}$ is not a clique with K vertices recurring in the K_3 graphs, then a clique with K vertices does not exist in G . This claim can be proved by contradiction: supposing there exists a clique with K vertices in G , since all graphs G_i in the graph set $\mathcal{G}$ were copied from G , this K -clique in G must also exist at least K_3 graphs of the graph set $\mathcal{G}$. Among all “recurrent dense subgraphs” with K nodes and K_3 networks in m unweighted graphs, the K -cliques recurring in K_3 graphs must be the “recurrent dense subgraph” having the largest total sum of edge weights. This K -clique recurring in K_3 graphs must be the solution of the above defined problem. So it contradicts with the statement “the recurrent dense subgraph with K nodes and K_3 networks found is not a clique with K vertices recurring in the K_2 graphs.”

Since the K -clique problem is NP-complete, the above defined problem is NP-hard and therefore our problem is NP-hard. ■

D. Simulation study

We generated 72 sets of random weighted paired gene and exon networks where each set contains 100 genes, 200 exons, 40 gene networks, and 40 exon networks, and all their edges weights follow the

⁶The NP-completeness proof of the K -clique problem can be found in textbooks introducing “algorithms and complexity” and lectures’ exercises (e.g., <http://people.bath.ac.uk/masn/v/Teaching/AAlg10Sol.pdf>).

uniform distribution in the range $[0, 0.3]$. For each set of paired networks, a random FCC pattern with K_1 member genes and K_2 member exons, and K_3 member networks is generated by making their edges weights follow the uniform distribution in the range $[\theta, 1]$. Here, K_1 is any of the three predefined values $\{5, 10, 20\}$, K_2 is any of the three predefined values $\{5, 10, 20\}$ (and simultaneously satisfying $K_2 \geq K_1$), K_3 is any of the three predefined values $\{5, 10, 20\}$, and θ is any of the four predefined values $\{0.6, 0.7, 0.8, 0.9\}$. Therefore, there are a total of $\binom{3}{2} \times 3 \times 4 = 72$ FCC patterns generated in all 72 sets of networks. Each FCC pattern is then placed into a set of networks, by randomly selecting $K_1/K_2/K_3$ genes/exons/networks and replacing edges weights with the corresponding FCC pattern's edges weights. These simulated networks and patterns have taken into account of various factors that may affect the performance. We can evaluate the performance by counting the number of these predefined patterns found or hit by the method with each parameter combination.

We performed our tensor method with different values of the parameters f and g on each set of networks and obtained the result as shown in the following box. As explained in this table, $f = 1$ and $g = 1$ is one of the best choices and is more computationally efficient than other values. For the value of h , in practice, any value ≥ 2 can be used. Therefore, we choose $h = 10$.

$f = g =$	1	1.1	1.2	1.3	1.4	1.5
hit rate =	100%	100%	97.2%	91.6 %	91.6	%90.2%

Hits table of performing our tensor method with different values of f and g and α on all 72 sets of paired networks, where f uses the same value of g , and $\text{hit} = \frac{\text{number of predefined patterns obtained / hit by tensor method}}{\text{number of all predefined patterns}}$. In this table, we chose the $f = g = 1$ whose hitting values are always $\geq$ the rest values of f and g .

E. Proof of Lemma 2

Proof. The Lagrangian function incorporating constraints is formed as follows,

$$\mathcal{L} = H_G(\mathbf{x}, \mathbf{w}) + \lambda H_E(\mathbf{y}, \mathbf{w}) + \mu O_R(\mathbf{x}, \mathbf{y}, Z) - \tau_1 (\|\mathbf{x}\|_f^f - 1) - \tau_2 (\|\mathbf{y}\|_g^g - 1) - \tau_3 (\|\mathbf{w}\|_h^h - 1) - \sum_{i=1}^N \psi_i \left(\sum_{j=1}^M z_{ij} - 1 \right) \quad (9)$$

where τ_1 , τ_2 , τ_3 and $\psi_i (i=1, \dots, N)$ are Lagrangian multipliers for enforcing constraints. According to the Karush-Kuhn-Tucker (KKT) conditions, the partial derivatives of $\mathcal{L}$ with respect to x_i , y_j , w_k , and z_{ij} should be equal to zero: $\frac{\partial \mathcal{L}}{\partial x_i} = 0$, $\frac{\partial \mathcal{L}}{\partial y_j} = 0$, $\frac{\partial \mathcal{L}}{\partial w_k} = 0$ and $\frac{\partial \mathcal{L}}{\partial z_{ij}} = 0$. The non-negativeness of $\mathbf{x}$, $\mathbf{y}$, $\mathbf{w}$, and Z is enforced by the complementarity condition, $\frac{\partial \mathcal{L}}{\partial x_i} x_i = 0$, $\frac{\partial \mathcal{L}}{\partial y_j} y_j = 0$, $\frac{\partial \mathcal{L}}{\partial w_k} w_k = 0$, and $\frac{\partial \mathcal{L}}{\partial z_{ij}} z_{ij} = 0$. This leads to,

$$\left(\sum_{j,k} g_{ijk} x_j w_k + \mu [(Z \cdot R) \mathbf{y}]_i - \tau_1 f x_i^{f-1} \right) x_i = 0 \quad (10)$$

$$\left(\lambda \sum_{i,k} e_{ijk} y_i w_k + \mu [(Z \cdot R)^T \mathbf{x}]_j - \tau_2 g y_j^{g-1} \right) y_j = 0 \quad (11)$$

$$\left(\sum_{i,j} g_{ijk} x_i x_j + \lambda \sum_{i,j} e_{ijk} y_i y_j - \tau_3 h w_k^{h-1} \right) w_k = 0 \quad (12)$$

$$(\mu r_{ij} x_i y_j - \psi_i) z_{ij} = 0 \quad (13)$$

Summing over index i in Eq. (10), over index j in Eq. (11), over index k in Eq. (12), over index j in Eq. (13) for each index i , we obtain the values of Lagrangian multipliers:

$$\tau_1 = \frac{\sum_{i,j,k} g_{ijk} x_i x_j w_k + \mu \mathbf{x}^T (Z \cdot R) \mathbf{y}}{f} \quad (14)$$

$$\tau_2 = \frac{\lambda \sum_{i,j,k} e_{ijk} y_i y_j w_k + \mu \mathbf{x}^T (Z \cdot R) \mathbf{y}}{g} \quad (15)$$

$$\tau_3 = \frac{\sum_{i,j,k} g_{ijk} x_i x_j w_k + \lambda \sum_{i,j,k} e_{ijk} y_i y_j w_k}{h} \quad (16)$$

$$\psi_i = \mu [R \cdot (\mathbf{x} \mathbf{y}^T)]_{i*}^T Z_{i*} \quad i = 1, \dots, N \quad (17)$$

Note that Eq. (17) is obtained by enforcing the constraint $\sum_{j=1}^M z_{ij} = 1$ for all i . This doesn't impact the correctness of the algorithm because it can be proved that the constraint $\sum_{j=1}^M z_{ij} = 1$ does not hurt the optimization when using the constraint $\sum_{j=1}^M z_{ij} \leq 1$ when $\mathbf{x}$ and $\mathbf{y}$ are enforced by sparsity constraints.

Therefore, it is obvious to verify that the fixed point of the updating steps in Algorithm 1 satisfies Eq. (10,11,12,13). So it is proved.

F. Proof of Lemma 3

Proof. We use the auxiliary function approach to prove the theorem. $G(\mathbf{x}, \mathbf{x}')$ is an auxiliary function of the function $O(\mathbf{x})$ if

$$G(\mathbf{x}, \mathbf{x}') O(\mathbf{x}) \quad \text{and} \quad G(\mathbf{x}, \mathbf{x}) = O(\mathbf{x})$$

We have the following lemma for the auxiliary function.

Lemma 5. If $G(\mathbf{x})$ is an auxiliary function of $O(\mathbf{x})$, then $O(\mathbf{x})$ is nondecreasing under the updates $\mathbf{x}^{(t+1)} = \arg \max_{\mathbf{x}} G(\mathbf{x}, \mathbf{x}^{(t)})$, because of the fact $O(\mathbf{x}^{(t)}) = G(\mathbf{x}^{(t)}, \mathbf{x}^{(t)}) \leq G(\mathbf{x}^{(t+1)}, \mathbf{x}^{(t)}) \leq O(\mathbf{x}^{(t+1)})$.

To design a suitable auxiliary function of our Lagrangian function, we used the inequality of $x \geq 1 + \log(x)$ for any positive x which was introduced by Ding et al. (2006) for proving convergence. So the auxiliary function of our Lagrangian function in Eq. (9) is introduced as follows,

$$\begin{aligned} G(\mathbf{x}, \mathbf{y}, \mathbf{w}, Z, \mathbf{x}', \mathbf{y}', \mathbf{w}', Z') = & \sum_{i,j,k} g_{ijk} x'_i x'_j w'_k \left(1 + \log \frac{x_i x_j w_k}{x'_i x'_j w'_k} \right) + \lambda \sum_{i,j,k} e_{ijk} y'_i y'_j w'_k \left(1 + \log \frac{y_i y_j w_k}{y'_i y'_j w'_k} \right) \\ & + \mu \sum_{r_{ij} \neq 0} z'_{ij} r_{ij} x'_i y'_j \log \frac{z_{ij} x_i y_j}{z'_{ij} x'_i y'_j} - \tau_1 \left(\sum_i x_i^f - 1 \right) - \tau_2 \left(\sum_j y_j^g - 1 \right) \\ & - \tau_3 \left(\sum_k w_k^h - 1 \right) - \sum_{i=1}^N \psi_i \left(\sum_{j=1}^M z_{ij} - 1 \right) \end{aligned} \quad (18)$$

where $\mathbf{x}'$ denotes $\mathbf{x}^{(t)}$. The maximum for Eq. (18) is obtained by setting its gradient to zero,

$$\frac{\partial G}{\partial x_i} = \frac{x'_i \sum_{j,k} g_{ijk} y'_j w'_k}{x_i} + \mu \frac{x'_i \sum_j r_{ij} z'_{ij} y'_j}{x_i} - \tau_1 f x_i^{f-1} = 0 \quad (19)$$

$$\frac{\partial G}{\partial y_j} = \lambda \frac{y'_j \sum_{i,k} e_{ijk} x'_i w'_k}{y_j} + \mu \frac{y'_j \sum_i r_{ij} z'_{ij} x'_i}{y_j} - \tau_2 g y_j^{g-1} = 0 \quad (20)$$

$$\frac{\partial G}{\partial w_k} = \frac{w'_k \sum_{i,j} g_{ijk} x'_i y'_j}{w_k} + \lambda \frac{w'_k \sum_{i,j} e_{ijk} x'_i y'_j}{w_k} - \tau_3 h w_k^{h-1} = 0 \quad (21)$$

$$\frac{\partial G}{\partial z_{ij}} = \mu \frac{r_{ij} z'_{ij} x'_i y'_j}{z_{ij}} - \psi_i = 0 \quad (22)$$

The above equations can be re-written as

$$x_i = \left(x'_i \frac{\sum_{j,k} g_{ijk} x'_j w'_k + \mu [(Z' \cdot R) \mathbf{y}']_i}{\tau_1 f} \right)^{1/f} \quad (23)$$

$$y_j = \left(y'_j \frac{\lambda \sum_{i,k} e_{ijk} y'_i w'_k + \mu [(Z' \cdot R)^T \mathbf{x}']_j}{\tau_2 g} \right)^{1/g} \quad (24)$$

$$w_k = \left(w'_k \frac{\sum_{i,j} g_{ijk} x'_i y'_j + \lambda \sum_{i,j} e_{ijk} y'_i y'_j}{\tau_3 h} \right)^{1/h} \quad (25)$$

$$z_{ij} = z'_{ij} \frac{[R \cdot (\mathbf{x}' \mathbf{y}'^T)]_{ij}}{\psi_i \frac{1}{\mu}} \quad (26)$$

Since above variables must satisfy the constraints in Eq. (1), we can tune $\tau_1, \tau_2, \tau_3, \psi_i$ like Eqs. (14–17) so that they meet the constraints. Therefore, we can recover the updating rules in Algorithm 1, by letting $\mathbf{x}^{(r+1)} = \mathbf{x}, \mathbf{y}^{(r+1)} = \mathbf{y}, \mathbf{w}^{(r+1)} = \mathbf{w}, Z^{(r+1)} = Z$, and $\mathbf{x}^{(r)} = \mathbf{x}', \mathbf{y}^{(r)} = \mathbf{y}', \mathbf{w}^{(r)} = \mathbf{w}', Z^{(r)} = Z'$.

To ensure the solution gives a maxima, we show that the second order derivatives (Hessian matrix) of the auxiliary function are negative definite. We have

$$\begin{aligned} \frac{\partial^2 G}{\partial x_i \partial x_i} &= - \left[\frac{x'_i \sum_{j,k} g_{ijk} y'_j w'_k}{x_i^2} + \mu \frac{x'_i \sum_j r_{ij} z'_{ij} y'_j}{x_i^2} + \tau_1 f(f-1) x_i^{f-2} \right] \\ \frac{\partial^2 G}{\partial y_j \partial y_j} &= - \left[\lambda \frac{y'_j \sum_{i,k} e_{ijk} x'_i w'_k}{y_j^2} + \mu \frac{y'_j \sum_i r_{ij} z'_{ij} x'_i}{y_j^2} + \tau_2 g(g-1) y_j^{g-2} \right] \\ \frac{\partial^2 G}{\partial w_k \partial w_k} &= - \left[\frac{w'_k \sum_{i,j} g_{ijk} x'_i y'_j}{w_k^2} + \lambda \frac{w'_k \sum_{i,j} e_{ijk} x'_i y'_j}{w_k^2} + \tau_3 h(h-1) w_k^{h-2} \right] \\ \frac{\partial^2 G}{\partial z_{ij} \partial z_{ij}} &= - \left[\mu \frac{r_{ij} z'_{ij} x'_i y'_j}{z_{ij}^2} \right] \\ \frac{\partial^2 G}{\partial x_i \partial y_j} &= \frac{\partial^2 G}{\partial x_i \partial w_k} = \frac{\partial^2 G}{\partial x_i \partial z_{ij}} = \frac{\partial^2 G}{\partial y_j \partial w_k} = \frac{\partial^2 G}{\partial y_j \partial z_{ij}} = \frac{\partial^2 G}{\partial w_k \partial z_{ij}} = 0 \end{aligned}$$

Because $f, g, h \geq 1$, this Hessian matrix is a diagonal matrix with negative quantities on the diagonals. Therefore $G(\mathbf{x}, \mathbf{y}, \mathbf{w}, Z, \mathbf{x}', \mathbf{y}', \mathbf{w}', Z')$ is a concave function in $\mathbf{x}, \mathbf{y}, \mathbf{w}$, and Z and has a unique global maximum. This completes the proof of the theorem. ■

G. Random network generation and comparison with real networks

To further evaluate the significance of FCCs discovered in real networks, we applied the proposed method on random networks to obtain FCCs. The random network is generated by randomizing a given network $G(V, E, W)$ (where V is the node set, E is the edge set and W is the sequence of the edges' weights corresponding to edges in E) with the following steps.

Given a real weighted network, the corresponding random weighted network is generated by a random redistribution of the actual weights on the randomly generated unweighted graph. In other words, for a real weighted network, we firstly generated a random unweighted network by using the widely used degree-preserving randomization procedure (Maslov and Sneppen, 2002) (the MATLAB code is provided at www.cmth.bnl.gov/~maslov/matlab.htm). Then the weights collected from the real weighted network were randomly distributed among the edges in the random unweighted network. This procedure is similar to our previous work (Li et al., 2011, 2007) and is formally presented in the following box.

Random weighted network generation procedure

Procedure GENERATE-RANDOM-NETWORK($G(V, E, W)$)

Generate random unweighted network: apply the degree-preserving randomization procedure (Maslov and Sneppen, 2002) on the edge set E to obtain a randomized edge set E' whose nodes' degree distribution is the same as that of E .

Randomly assign weights to edges of the random unweighted network: randomly permute the edge weights of the sequence W , and assign the values of the new randomized sequence W' to each edge of the edge set E' .

Return the random weighted network $G' (V, E', W')$

For our case, we applied this randomization procedure to each weighted network in paired tensors $\mathcal{G}$ and $\mathcal{E}$. The resulted random tensors is denoted as $\mathcal{G}'$ and $\mathcal{E}'$. We generated 10 sets of random paired tensors $\mathcal{G}'$ and $\mathcal{E}'$ from the same real tensors $\mathcal{G}$ and $\mathcal{E}$, and applied the proposed FCCs discovery method on each of them to discover the FCCs with ≥ 5 genes, ≥ 5 exons, ≥ 2 datasets, “heaviness” ≥ 0.4 , “ $coverage_{gene} \geq 0.7$,” and “ $coverage_{exon} \geq 0.7$.” Among 10 sets of random paired tensors, only 0 ~ 3 FCC families in each set of data (average 0.9 families) were identified. In contrast, 8,667 FCC families were identified in the real tensors. This comparison indicates the significance of FCCs identified in the tensors of real weighted paired networks.

H. Signal extraction procedure of ENCODE data

Recently, in the ENCODE production phase (September 2007 to the present), there are 191 ENCODE genome-wide tables for ChIP-seq. To perform enrichment analysis, we integrated these ChIP-seq samples in Genome-wide ENCODE data including chromatin modification, TF binding, Methylation, and chromatin accessibility. UCSC database provide the peak tables for these ChIP-Seq data.

However, different data types have different criterion for the significance. We used both top N and threshold condition to select peaks as follows:

1. If the table is Methylation-Seq data, select the peaks score > 0.6 ;
2. Else if the peaks have p-value, select the peaks with p-value $< 1E-4$ and top 20K;
3. Otherwise, select the peaks with signal value > 1 and top 20K.

Then for each table, we selected the target genes whose transcriptional start site is nearby the peaks within 5000 bases.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (grant R01GM074163) and the National Science Foundation (Grant 0747475).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Achlioptas, D., and McSherry, F. 2007. Fast computation of low-rank matrix approximations. *J ACM* 54 (article No. 9).
- Anderson, T.W. 2003. *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley-Interscience, Hoboken, NJ.
- Arora, S., Hazan, E., and Kale, S. 2006. A fast random sampling algorithm for sparsifying matrices, 272–279. In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. Springer-Verlag, Berlin.
- Ashburner, M., Ball, C., Blake, J., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.
- Ayala, Y., Laura De Conti, S., Vázquez, A., et al. 2010. TDP-43 regulates its mRNA levels through a negative feedback loop. *EMBO J.* 30, 277–288.
- Ayala, Y., Zago, P., D’Ambrogio, A., et al. 2008. Structural determinants of the cellular localization and shuttling of TDP-43. *J. Cell Sci.* 121, 3778.
- Barabasi, A., and Oltvai, Z. 2004. Network biology: understanding the cell’s functional organization. *Nat. Rev. Genet.* 5, 101–113.
- Barboric, M., Lenasi, T., Chen, H., et al. 2009. 7SK snRNP/P-TEFb couples transcription elongation with alternative splicing and is essential for vertebrate development. *Proc. Natl. Acad. Sci. USA* 106, 7798–7803.

- Bedford, M., Reed, R., and Leder, P. 1998. WW domain-mediated interactions reveal a spliceosome-associated protein that binds a third class of proline-rich motif: the proline glycine and methionine-rich motif. *Proc. Natl. Acad. Sci. USA* 95, 10602.
- Blencowe, B., Bowman, J., McCracken, S., et al. 1999. SR-related proteins and the processing of messenger RNA precursors. *Biochem. Cell Biol.* 77, 277–291.
- Blower, P.E., Verducci, J.S., Lin, S., et al. 2007. MicroRNA expression profiles for the NCI-60 cancer cell panel. *Mol. Cancer Therap.* 6, 1483–1491.
- Brès, V., Gomes, N., Pickle, L., et al. 2005. A human splicing factor, SKIP, associates with P-TEFb and enhances transcription elongation by HIV-1 Tat. *Genes Dev.* 19, 1211.
- Buratti, E., and Baralle, F. 2001. Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR exon 9. *J. Biol. Chem.* 276, 36337.
- Burkard, R., Dell’Amico, M., and Martello, S. 2009. *Assignment Problems*. SIAM, Philadelphia.
- Bussey, K.J., Chin, K., Lababidi, S., et al. 2006. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol. Cancer Therap.* 5, 853–867.
- Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068.
- Céacères, J.F., and Kornblihtt, A.R. 2002. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* 18, 186–193.
- Ding, C., Zhang, Y., Li, T., et al. 2006. Biclustering protein complex interactions with a biclique finding algorithm. *Proc. 6th Int. Conf. Data Mining*, 178–187.
- Fong, Y., and Zhou, Q. 2001. Stimulatory effect of splicing factors on transcriptional elongation. *Nature* 414, 929–933.
- Fujita, P.A., Rhead, B., Zweig, A.S., et al. 2011. The UCSC genome browser database: update 2011. *Nucleic Acids Res.* 39, D876–D882.
- Garcia-Jurado, G., Llanes, D., Moreno, A., et al. 2011. Monoclonal antibody that recognizes a domain on heterogeneous nuclear ribonucleoprotein K and PTB-associated splicing factor. *Hybridoma* 30, 53–59.
- Gordon, S., Akopyan, G., Garban, H., et al. 2005. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene* 25, 1125–1142.
- Hopfield, J.J., Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* 79, 2554–2558.
- Jiang, H., and Wong, W.H. 2009. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25, 1026–1032.
- Kameoka, S., Duque, P., and Konarska, M. 2004. p54nrb associates with the 5′ splice site within large transcription/splicing complexes. *EMBO J.* 23, 1782–1791.
- Kornblihtt, A., de la Mata, M., Fededa, J., et al. 2004. Multiple links between transcription and splicing. *RNA* 10, 1489–1498.
- Li, W., Lin, Y., and Liu, Y. 2007. The structure of weighted small-world network. *Phys. A Stat. Mech. Appl.* 376, 708–718.
- Li, W., Liu, C.-C., Zhang, T., et al. 2011. Integrative analysis of many weighted co-expression networks using tensor computation. *PLoS Comput. Biol.* 7, e1001106.
- Lin, S., Coutinho-Mansfield, G., Wang, D., et al. 2008. The splicing factor SC35 has an active role in transcriptional elongation. *Nat. Struct. Mol. Biol.* 15, 819–826.
- Lindstrom, D., Squazzo, S., Muster, N., et al. 2003. Dual roles for spt5 in pre-mRNA processing and transcription elongation revealed by identification of spt5-associated proteins. *Mol. Cell. Biol.* 23, 1368.
- Little, C., Nau, M., Carney, D., et al. 1983. Amplification and expression of the c-myc oncogene in human lung cancer cell lines. *Nature* 306, 194–196.
- Loflin, P., Chen, C., and Shyu, A. 1999. Unraveling a cytoplasmic role for hnRNP D in the *in vivo* mRNA destabilization directed by the AU-rich element. *Genes Dev.* 13, 1884.
- Mapendano, C.K., Lykke-Andersen, S., Kjems, J., et al. 2010. Crosstalk between mRNA 3′ end processing and transcription initiation. *Mol. Cell* 40, 410–422.
- Maslov, S., and Sneppen, K. 2002. Specificity and stability in topology of protein networks. *Science* 296, 910–913.
- Matlin, A., Clark, F., and Smith, C. 2005. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* 6, 386–398.
- Motwani, R., and Raghavan, P. 1995. *Randomized Algorithms*. Cambridge University Press, New York.
- Motzkin, T.S., and Straus, E.G. 1965. Maxima for graphs and a new proof of a theorem of Turán. *Can. J. Math.* 17, 533–540.
- Muñoz, M.J., Santangelo, M.S.P., Paronetto, M.P., et al. 2009. DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. *Cell* 137, 708–720.
- Nair, S., and Burley, S. 2003. X-ray structures of myc-max and mad-max recognizing DNA: molecular bases of regulation by proto-oncogenic transcription factors. *Cell* 112, 193–205.
- Nayler, O., Strätling, W., Bourquin, J., et al. 1998. SAF-B protein couples transcription and pre-mRNA splicing to SAR/MAR elements. *Nucleic Acids Res.* 26, 3542.
- Nevins, J. 2001. The Rb/E2F pathway and cancer. *Hum. Mol. Genet.* 10, 699.
- Ou, S., Wu, F., Harrich, D., et al. 1995. Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs. *J. Virol.* 69, 3584.

- Pandit, S., Wang, D., and Fu, X. 2008. Functional integration of transcriptional and RNA processing machineries. *Curr. Opin. Cell Biol.* 20, 260–265.
- Peng, R., Hawkins, I., Link, A., et al. 2006. The splicing factor PSF is part of a large complex that assembles in the absence of pre-mRNA and contains all five snRNPs. *RNA Biol.* 3, 69.
- Perales, R., and Bentley, D. 2009. Co-transcriptionality: the transcription elongation complex as a nexus for nuclear transactions. *Mol. Cell* 36, 178–191.
- Piva, F., Giulietti, M., Burini, A.B., et al. 2011. SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. *Hum. Mutat.* 33, 81–85.
- Rappsilber, J., Ryder, U., Lamond, A., et al. 2002. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* 12, 1231.
- Reed, R. 2003. Coupling transcription, splicing and mRNA export. *Curr. Opin. Cell Biol.* 15, 326–331.
- Rosonina, E., and Blencowe, B. 2002. Gene expression: the close coupling of transcription and splicing. *Curr. Biol.* 12, 319–321.
- Ruepp, A., Waegele, B., Lechner, M., et al. 2010. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* 38, D497–D501.
- Sandelin, A., Alkema, W., Engström, P., et al. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 32, D91–D94.
- Scherf, U., Ross, D.T., Waltham, M., et al. 2000. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24, 236–244.
- Shankavaram, U.T., Reinhold, W.C., Nishizuka, S., et al. 2007. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integrative microarray study. *Mol. Cancer Therap.* 6, 820–832.
- Stark, C., Breitkreutz, B., Reguly, T., et al. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535.
- Staunton, J.E., Slonim, D.K., Collier, H.A., et al. 2001. Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. USA* 98, 10787–10792.
- Suzuki, H., Igarashi, S., Nojima, M., et al. 2010. IGFBP7 is a p53-responsive gene specifically silenced in colorectal cancer with CpG island methylator phenotype. *Carcinogenesis* 31, 342.
- Tabach, Y., Brosh, R., Buganim, Y., et al. 2007. Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PLoS ONE* 2, e807.
- Thomas, D.J., Rosenbloom, K.R., Clawson, H., et al. 2007. The ENCODE project at UC Santa Cruz. *Nucleic Acids Res.* 35, D663–D667.
- Trapnell, C., Pachter, L., and Salzberg, S.L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111.
- Trapnell, C., Williams, B.A., Pertea, G., et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515.
- Tsay, A., Lovejoy, W., and Karger, D. 1999. Random sampling in cut, flow, and network design problems. *Math. Oper. Res.* 24, 383–413.
- Weinstein, J.N., Myers, T.G., O'Connor, P.M., et al. 1997. An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 343–349.
- Xu, M., Kao, M., Nunez-Iglesias, J., et al. 2008. An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. *BMC Genomics* 9, S12.
- Yan, X., Mehan, M.R., Huang, Y., et al. 2007. A graph-based approach to systematically reconstruct human transcriptional regulatory modules. *Bioinformatics* 23, i577–i586.
- Zhang, C., Dowd, D., Staal, A., et al. 2003. Nuclear coactivator-62 kDa/Ski-interacting protein is a nuclear matrix-associated coactivator that may couple vitamin D receptor-mediated transcription and RNA splicing. *J. Bio. Chem.* 278, 35325.
- Zhou, Q., and Sharp, P. 1996. Tat-SF1: cofactor for stimulation of transcriptional elongation by HIV-1 Tat. *Science* 274, 605.
- Zhou, X., Kao, M., Huang, H., et al. 2005. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol.* 23, 238–243.
- Zhou, Z., Licklider, L., Gygi, S., et al. 2002. Comprehensive proteomic analysis of the human spliceosome. *Nature* 419, 182–185.

Address correspondence to:
 Dr. Xianghong Jasmine Zhou
 Program in Computational Biology
 Department of Biological Sciences
 University of Southern California
 Los Angeles, CA 90089

E-mail: xjzhou@usc.edu