Journal of Theoretical Biology 263 (2010) 471–480

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Alternative splicing regulatory network reconstruction from exon array data

Kun Qu, Anastasia M. Yesnik, Peter J. Ortoleva *

Center for Cell and Virus Theory, Department of Chemistry, Indiana University, Bloomington, IN 47405, USA

ARTICLE INFO

Article history:
Received 2 June 2009
Received in revised form
14 November 2009
Accepted 22 December 2009
Available online 5 January 2010

Keywords:
Regulatory network
Alternative splicing
Microarray
Exon expression

ABSTRACT

Pre-mRNA alternative splicing (AS) allows individual genes to produce multiple types of mRNA and associated protein isoforms. While AS regulation enables the production of the hundreds of thousands of types of proteins needed for the normal functioning of the human cell, it also presents many opportunities for the onset of cancer and other diseases. The AS process is known to be regulated by a group of serine/arginine rich (SR) proteins, heterogeneous nuclear ribonucleoproteins (hnRNPs), and small nuclear ribonucleoprotein (snRNP) particles through a complex assembly. Each gene-exon is regulated by one or multiple splicing regulators, from which one may hypothesize the existence of an alternative splicing regulatory network (SRN). The SRN contains a list of gene-exons, for each of which the factors that up/down regulate them are provided. Since defects in the SRN play key roles in human disease, a reconstruction of human SRN could be used to facilitate the design of diagnostic and therapeutic strategies. In this paper, we present a methodology to automate genome-wide SRN reconstruction. We construct SRN based on an extensive correlation analysis of human exon expression microarray data, conventional gene expression microarray profiles, and an experimentally verified AS and transcriptional regulatory interaction training set. This SRN reconstruction methodology is demonstrated and software (AutoNet) that automates the reconstruction of SRN is developed. A genome-wide SRN was constructed for normal human cells and an assessment of the reliability of each predicted interaction is provided. Human SRN we constructed are free available from our web portal: https://ruby.chem.indiana.edu/~scorenfl/srn_results/lookup0.php

1. Introduction

The expression of genetic information of mammalian cells consists of three steps: (1) transcription of genetic codes from DNA to pre-mRNA; (2) pre-mRNA splicing to create one or more alternative mRNAs; and (3) mRNA translation. Pre-mRNA transcription is regulated by a special group of proteins known as transcription factors (TFs). TFs bind to a gene's promoter region to enhance/repress its expression. If the binding of a TF facilitates the binding of RNA polymerases and therefore enhances pre-mRNA expression, the TF is considered to up-regulate the transcription of this gene, and conversely for down-regulation. Since every gene is regulated by one or more TFs and every TF is encoded from one or more genes, there exists a transcriptional regulatory network (TRN) which contains (1) a list of genes for each of which the TFs up/down regulate it; and (2) a list of TFs for each of which genes that encode it or its component. The discovery of TRNs greatly advances our understanding of mechanisms of cellular processes and responses, and is of great importance in biotechnical applications, particularly in delineating regulatory abnormalities in cancers and other diseases from a genome-wide perspective. Recently, numerous TRN reconstruction approaches have been presented via inference from experimentally verified training sets (Cartegni and Krainer, 2002; Qu et al., 2007; Sayyed-Ahmad et al., 2007), gene expression microarray analysis (Chang et al., 2008; Chen et al., 2006; Gardner et al., 2003; Gutierrez-Rios et al., 2003; Huang et al., 2007; Li et al., 2004; Sano et al., 2006; Sayyed-Ahmad et al., 2007; Zhou et al., 2005; Zou and Conzen, 2005), gene ontology (Tuncay et al., 2006), and phylogenetic similarity analysis (Pazos and Valencia, 2001). We have developed a TRN discovery system that integrates all the above methods with a Bayesian integration (Tuncay et al., 2006; Ortoleva, 2007; Qu and Ortoleva, 2008; Qu et al., 2007; Sayyed-Ahmad et al., 2007; Sun et al., 2007).

The human cell has about 25,000 genes, yet it creates hundreds of thousands of distinct proteins. This implies that there are a similar number of distinct mRNAs. The gap in these numbers is made up by splicing, i.e. transcription of a given gene creates one pre-mRNA; then splicing cleaves this pre-mRNA and rejoins segments to create multiple distinct mRNAs per gene. This feature, known as alternative splicing (AS), allows the generation of a large number of mRNA and protein isoforms from many fewer genes. There is evidence that AS occurs in over 80% of human genes, and at least 15%, and perhaps as many as 50%, of human genetic diseases arise from AS abnormality (Matlin et al., 2005).

* Corresponding author. Tel.: +1 812 855 2717; fax: +1 812 855 8300.
E-mail addresses: kqu@indiana.edu (K. Qu), ayesnik@indiana.edu (A.M. Yesnik), ortoleva@indiana.edu (P.J. Ortoleva).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.12.025

Therefore, understanding the mechanisms AS regulation is of great importance.

Studies on AS include identifying the proteins and other molecules involved in splicing (Graveley, 2000; Matlin et al., 2005; Stamm et al., 2005); locating exon and intron positions and splicing regulator binding sites (Das et al., 2007; Graveley, 2000; Matlin et al., 2005) (which are known as exon splicing enhancers (ESEs) or silencers (ESSs) and intron splicing enhancers (ISEs) or silencer (ISSs)); and delineating the mechanisms of how serine/arginine rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs) bind to ESEs/ESSs/ISEs/ISSs and facilitate recruiting spliceosomes, how spliceosomes are assembled, and how AS eventually takes place (Graveley, 2000; Matlin et al., 2005; Stamm et al., 2005).

Presently, pre-mRNA AS is considered to be regulated by a group of splicing regulatory factors (SRFs), including SR proteins and hnRNPs, and facilitated by small nuclear ribonucleoprotein (snRNP) particles (splicesome). SRFs bind to regulatory sites on pre-mRNA (ESE, ESS, ISE, or ISS) and help in recruiting spliceosome. If the splicesome's recruitment enhances the exon inclusion in a given type of mRNA, the SRFs are up-regulating the exon AS, conversely if they inhibit the exon inclusion, the SRFs are down-regulating. As in TRNs, it is reasonable to hypothesize the existence of an alternative splicing regulatory network (SRN), which contains a list of exons for each of which the SRFs up/down regulating is provided. The SRN controls the way that pre-mRNA is cleaved and rejoined to create the mRNAs. These proteins are then processed into the enzymes and regulatory molecules that underlie most cellular behavior. Therefore, the construction of an SRN is critical in identifying abnormalities underlying the onset and progression of cancer from a genome-wide perspective.

There has been a major investment in genome-wide exon junction microarray profiling technologies (Castle et al., 2003; Cuperlovic-Culf et al., 2006; Johnson et al., 2003; Xing et al., 2006), and hundreds of thousands of exon AS are identified. Given that exon microarrays monitor several hundred thousand mRNA types simultaneously, we believed such data could reveal many differences between normal and diseased cells; Due to the vastness of the network and the lack of a reliable training set of SRF/exon interactions, no software for reconstructing the network of AS regulatory processes from exon array and other data have been available. Here we present a SRN reconstruction algorithm and an accompanying training set that we believe to be a first step in automated SRN reconstruction. To appreciate the complexity of the cellular control network and identify regulatory abnormalities in AS, we developed a methodology to reconstruct networks of SRF/exon regulatory interactions based an extensive analysis of human exon junction microarray expression and conventional gene expression profiles. In this paper, we construct an *a priori* SRN using correlations between SRF/exon interactions and a training set of experimentally verified data based on a hypothesis that exons are regulated by same set of SRFs if their expression profiles are highly correlated. Based on this *a priori* network and exon junction and conventional gene microarray data, we create the "AutoNet" analyzer to construct a final prediction of SRN with minimum microarray inconsistency by generalizing our earlier TRND system (Tuncay et al., 2006; Ortoleva, 2007; Qu et al., 2007; Sayyed-Ahmad et al., 2007; Sun et al., 2007). Predicted AS regulatory interactions are evaluated using a standard statistical criterion. High quality predictions are archived in our alternative splicing regulatory database available from our website (https://ruby.chem.indiana.edu/~scorenfl/srn_results/lookup0.php).

Reconstructed SRNs, with hundreds of thousands of regulatory interactions discovered, will provide insight into the role of AS in carcinogenesis. As abnormality in the AS is one of the main reasons for carcinogenesis initiation and onset/propagation of other genetic diseases, regulatory network abnormalities of exons and their specific regulatory SRFs are high-value therapeutic targets in our search for the origins of cancer. In this way, we believe that the proposed automated genome-wide SRN software will enable a paradigm shift in our ability to identify targets that minimize uncertainties due to indirect causes otherwise undetected. Reconstructed SRNs, with hundreds of thousands of regulatory interactions discovered, will provide specific insight into the role of alternative splicing in carcinogenesis. The networks created could be used to generate a refined diagnosis and treatment regime, the latter involving multiple genes, drugs, nutritional and other factors in an optimized balance. We believe the accuracy and genome-wide character of our SRN-based treatment design strategies will facilitate cancer research and clinical practice. Potential strengths of this approach are that SRN treatment discovery is genome-wide in scope, based on multiple types of regulatory processes, fully automated, and integrates multiple large datasets (e.g. exon array profiles, proteomics, and clinical data). We believe the SRN-based treatment discovery could avoid side-effects and resistance to treatment strategies originally arrived at from an understanding of only one or a few genes and other factors, while the complexity and scope of their coupling to other genes across the wider cell regulatory network was ignored.

2. Methods

2.1. Overview of SRN reconstruction algorithm

The extensive complexity of a human regulatory network and the limited amount of validated AS and gene expression regulatory information requires an extensive set of expression data and multiple methodologies to arrive at a reconstructed genome-wide SRN. Our methodologies are integrated by the workflow of Fig. 1. The training set is assembled from experimentally validated exon regulatory information (Table 1) to provide minimal, but reliable, information on the structure of the network. We use the correlation method, the training set, and an extensive set of exon expression data to construct an *a priori* network for normal human cells. There is presently limited direct experimentally verified SRF/exon interaction information. Our network reconstruction software AutoNet (Section 4) is used to predict regulator (e.g. transcription factors or SRFs)/respond element (e.g. genes or exons) interactions, based on an analysis of respond element expression profiles (e.g. conventional gene or exon expression microarray datasets). AutoNet contains a package to minimize error due to the presence of noise in respond element expression profiles. Details on our SRN reconstruction algorithm and individual modules in Fig. 1 are as follows.

2.2. Exon data and experimentally verified AS regulatory interactions

Exon–exon junction microarray data profile (GSE740) used was obtained from NIH Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). This dataset includes 50–54 sample types across five chip patterns in five platforms (notably GPL543, GPL544, GPL545, GPL546, and GPL547, see also Table 2). Some sample types (e.g. brain, liver, and prostate) were conducted in all the five chips, and some not (e.g. melanoma, lung carcinoma and melanoma). To arrive at a united profile of exon expression levels under different samples, we selected only those samples types that were tested in all the five chips, and then integrated all the exons into a single profile (Table 1). As summarized in Table 1, our

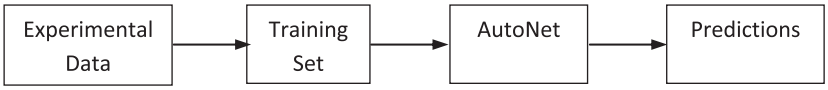

Fig. 1. Schematic workflow for SRN reconstruction. Experimental data used includes gene and exon expression microarray data and experimentally verified SRF/exon interactions. A training set for AutoNet is constructed based on experimental data. AutoNet predicts splicing regulator factor activities, and a correlation method integrated in AutoNet is used to predict regulator/respond element interactions.

Table 1
Summary of experimentally verified SRF/exon interactions.

Gene	MAPT	MATP	MAPT	SMN1	CASP9	EWSR1	BCL6	BRCA1	BRCA1	FGFR1
Exon	2	3	10	7	3, 4, 5, 6	2, 4, 5	1	5	18	1
SRp38	0	0	0	0	0	−1	0	0	0	0
ASF/SF2	0	−1	(1)	1	−1	0	0	0	1	0
SC35	−1	−1	−1	1	0	0	0	0	0	0
SRp20	0	−1	−1	1	0	0	1	0	0	0
SRp75	0	1	−1	1	0	0	0	0	0	0
SRp40	−1	0	−1	1	0	0	0	0	0	0
SRp55	−1	1	−1	1	0	0	0	−1	0	−1
9G8	0	0	−1	1	0	0	0	0	0	0
SRp30c	−1	−1	(−1)	1	0	0	0	0	0	0
SRp54	0	0	−1	0	0	0	0	0	0	0
Tra2β	−1	−1	1	1	0	0	0	0	0	0
U2AF	0	−1	−1	0	0	0	0	0	0	0
PTB	−1	−1	−1	0	0	0	0	0	0	0
hnPNPG	0	0	−1	0	0	0	0	0	0	0
CELF3	−1	0	1	0	0	0	0	0	0	0
CELF4	1	0	1	0	0	0	0	0	0	0
SWAP	0	−1	−1	0	0	0	0	0	0	0
Nova1	1	−1	−1	0	0	0	0	0	0	0
hnRNPA1	0	−1	0	0	0	0	0	0	0	0
SLM1	1	1	−1	0	0	0	0	0	0	0
SLM2	1	1	−1	0	0	0	0	0	0	0
nPTB	1	0	0	0	0	0	0	0	0	0
KSRP	−1	−1	−1	0	0	0	0	0	0	0

Up-regulation of the exon inclusion is indicated by +1 and down-regulation −1. 0 represents unknown regulation. Interaction in ( ) implies low confidence. Information was gathered from the literature (Arikan et al., 2002; Kondo et al., 2004; Li et al., 2003; Mabon and Misteli, 2005; Wang et al., 2004; Wang et al., 2005; Wu et al., 2006).

Table 2
Summary of information content in exon data profiles used.

Exon expression profile	Number of genes	Number of exons	Number of sample types
GPL543	1230	13413	54
GPL544	2506	23107	50
GPL545	2648	2307	54
GPL546	2574	23107	52
GPL547	1319	12961	54
Integrated exon profile	10273	95695	47

All data were taken from NIH GEO GSE740.

united exon expression profile contains 95,695 exons on 10,273 genes, all of which have expression levels in the 47 samples. Experimentally verified AS regulations were gathered from the literature and summarized in Table 1. While some details are different between different human tissues or cell lines, we made the hypothesis that many human tissues and cell lines share large regions of a common TRN and SRN. Detailed data and sample information are summarized in *Appendix A*.

2.3. Data pre-processing and a priori network reconstruction via AutoNet

The first steps of our SRN reconstruction algorithm is to normalize the exon expression profiles using conventional gene microarray expression data and then construct a preliminary SRN as a training set for final network reconstruction and quality assessment. We focused on the SRFs as they were generally believed to be key factors regulating splicing. We implemented an algorithm to discover each exon's regulation by a specific SRF protein(s) and the sense (up/down) of each regulation. Our algorithm constructs the quantity b_{gxf} which, for simplicity here, is 0 or ± 1 if exon x of gene g is non-activated or activated/repressed by factor f. Let Ψ_{gxs} be the measured exon expression level of gene g, exon x in sample s. We defined g to be the reference gene, i.e. E_{gs} is the gene expression level for gene g in sample s. With this, the "normalized expression level" ψ_{gxs} is defined via

$$\psi_{gxs} = \Psi_{gxs}/E_{gs} \tag{1}$$

Using these normalized exon expression profiles, a training set is constructed as follows. The algorithm is based on a hypothesis that exons with highly correlated expression profile are regulated by the same set of SRFs. Here we define correlation of two exons (gene i exon j and gene k exon l) $C_{ij,kl}$ via

$$C_{ij,kl} = \frac{\langle (\psi_{g_i x_j s} - \overline{\psi}_{g_i x_j})(\psi_{g_k x_l s} - \overline{\psi}_{g_k x_l}) \rangle}{\theta_{g_i x_j} \theta_{g_k x_l}} \tag{2}$$

where $\overline{\psi}_{g_i x_j}$ is the average expression of gene i, exon j over all the samples, and $\langle \cdots \rangle$ indicates an average over all samples and

$$\theta^2_{g_i x_j} = \langle (\psi_{g_i x_j s} - \overline{\psi}_{g_i x_j})^2 \rangle \tag{3}$$

The training set is constructed by calculating correlations of the expression of exons in Table 1 with those of all other exons. If an exon's profile is strongly correlated with one of those in Table 1, it is assumed to be regulated similarly to that exon.

Table 3
Gene expression datasets and normalized exon expression profiles.

Gene expression profile name	Number of genes	Number of exons	Number of sample types
GDS1096	13341	NA	36
Normalized exon expression I	5655	57946	26
GDS596	13341	NA	79
Normalized exon expression II	5655	57946	31
GDS422	9143	NA	12
Normalized exon expression III	4668	49447	10

GDS1096, GDS596 and GDS422 are three gene expression datasets obtained from NIH GEO. These datasets contain 36, 79 and 12 cell samples, respectively. For example, GDS1096 contains 36 cell samples, only 26 of which are also found in our original exon expression profile. We compare the exon expression and their correspondence gene expression via Eq. (1) over these 26 common cell samples and generate a normalized exon expression I profile. Normalized exon expression II and III profiles are generated in the same way.

To carry out this procedure for a representative set of human cell types, we use conventional gene expression profiles from NIH GEO (Table 3). Similarly with expression profiles in Table 1, some cell samples are conducted in both the conventional gene expression and exon expression profiles, some are not. In order to normalize the exon expression based on particular gene expression profiles (here GDS1096, GDS596 and GDS422, Table 3), only those cell types that are probed in both profiles are selected. Therefore, based on the above three gene expression profiles, we can generate three normalized exon expression profiles, each corresponding to its gene expression dataset. We use these three datasets to check the consistency of our predictions.

2.4. Automated microarray-based network reconstruction (AutoNet)

AutoNet is our statistical regulatory-based module for automatically reconstructing TRNs, SRNs or other networks, consisting of a set of respond elements (e.g. expressing genes) and the factor (e.g. transcription factors) that regulate them. AutoNet is a generalization of our earlier software FTF (Tuncay et al., 2006; Qu et al., 2007; Sayyed-Ahmad et al., 2007; Sun et al., 2007) designed to reconstruct the TRN from a training set of promoter site transcription factor regulatory interactions and profiles of gene expression. Similarly, AutoNet uses a training set of regulatory factor/respond element interactions and related microarray profiles to generate the network of regulator/responder interactions. The input array data are a list of respond elements for each of which the respond expression profile over the various conditions or sample types are provided. The output regulatory network is a list of respond elements, for each of which the regulators and the sense (up/down) of the regulation are specified. Thus, AutoNet with a preliminary SRN training set and exon expression microarray profiles can be used to reconstruct a richer SRN that incorporates the possibility of multiple splicing regulator interactions, each of a large set of regulating given respond elements.

The input to AutoNet is an *a priori* network and respond elements' expression profiles. The output is suggestions for improving the network, and predicted profiles for regulator activities across the set of conditions/samples. AutoNet is based on the following notions:

- microarray expression as well as exon expression profile data are usually error-prone and only semi-quantitative; thus a consensus method is needed whereby results from a variety of genes are synthesized to derive information on a given gene;
- a method based on regulators has the advantage that microarray noise and errors in given training set can be overcome by statistics—i.e. the regulation of many respond elements through a given regulator, or a small subset thereof;
- due to data uncertainty, it is usually not likely (except in rare cases where hundreds of microarrays are available) that there is enough information to obtain both regulatory network structure and the associated transcription and RNA degradation rate coefficients simultaneously; and
- network discovery requires many automated trials of possible networks so the algorithm must be extremely efficient.

Thus the objective of AutoNet is to discover regulatory network structure by taking advantage of the statistical robustness allowed by a regulator-based analysis.

The essential AutoNet equations are as follows. Taking construction of a TRN for an example, consider a system with N_g genes. Then,

$$T_n^r - T_n^s = \sum_{i=1}^{N_g} H(m_i^r - m_i^s) b_{in} \Psi_{in} \tag{4}$$

T_n^r the activity of TF n at condition or time r, m_i^r the cDNA microarray response for gene i at condition r, b_{in} the regulatory network matrix ($b_{in} >$ or < 0 for gene i up/down regulated by TF n, $b_{in}=0$ for no regulation)

$H(x) = \pm 1$ for $x >$ or < 0, $=0$ for $x=0$

$$\Psi_{in} = \text{normalized weight—e.g. } \Psi_{in} = \frac{2^{L_i}}{2^{L_i}-1} \cdot \frac{1}{M_n} \tag{5}$$

for L_i=number of TFs regulating gene i, and M_n is a normalized factor that is the number of genes TF n regulates.

The advantage of this analysis is that the T_n^r are obtained directly, i.e. no differential equations for them must be solved. The statistical weight Ψ_{in} accounts for the likelihood that a gene controlled by many TFs will not reflect the activity of any one of them. Finally, $\sum_{i=1}^{N_i} \Psi_{in} = 1$, i.e. Ψ_{in} has the character of a normalized probability.

One can compute T_n^r for TF n from one of the genes it regulates by keeping the $T_{n'}^r (n' \neq n)$ as obtained above. The correlation of this T_n^r and the one constructed as above gives a measure of how well the regulation of gene i by TF n is characterized by b_{in}. Such considerations are the basis of the gene ranking in AutoNet. As AutoNet is fast, many alternatives b_{in} can be tested and improvements on the *a priori* regulatory network are suggested.

AutoNet had been successfully applied to reconstruct bacterial and mammalian TRNs (Tuncay et al., 2006; Qu et al., 2007; Sayyed-Ahmad et al., 2007; Sun et al., 2007), and consists one of the key approaches we used in our TRN discovery system (https://systemsbiology.indiana.edu/trnd/pages/menuMain.php). In the present study, we expend AutoNet and applied it to SRN reconstruction using exon array data normalized as described above. As demonstrated in our earlier studies (Tuncay et al., 2006;

Qu et al., 2007; Sayyed-Ahmad et al., 2007; Sun et al., 2007), this can yield a genome-wide regulatory network. However, the earlier version of AutoNet was designed for networks involving a few thousand of respond elements. In contrast for alternative splicing regulation there are effectively hundreds of thousands of respond elements. Therefore, we developed a data/array healing framework optimized for extremely large networks, training sets, and array datasets. We integrated a correlation method for generating large set of SRN predictions with the previous version of AutoNet.

Correlation of predicted regulator activity with exon expression over all the samples is calculated. Here we define correlation of splicing regulator n activity T_n^s and exons (gene i exon j) expression $\psi_{g_i x_j s}$ over all the sample s as $C_{n,ij}$

$$C_{n,ij} = \frac{\langle (T_n^s - \overline{T}_n^s)(\psi_{g_i x_j s} - \overline{\psi}_{g_i x_j}) \rangle}{\theta_n \theta_{g_i x_j}} \tag{6}$$

where $\overline{T}_n^s$ is the average activity of regulator n over all the samples and θ_n is defined as

$$\theta_n^2 = \langle (T_n^s - \overline{T}_n^s)^2 \rangle \tag{7}$$

and average is taking over all the samples.

2.5. SRN prediction

For each normalized exon expression profiles (I, II and III in Table 3), we can predict SFR/exon interactions by comparing their correlation $C_{n,ij}$ with a threshold correlation C_{th}.

If $C_{n,ij} \geq C_{th}$ then we say SRF n up-regulate gene i exon j, and $b_{n,ij} = 1$

If $C_{n,ij} \leq -C_{th}$ then we say SRF n down-regulate gene i exon j, and $b_{n,ij} = -1$

Otherwise we say there is no apparent regulation.

2.6. Confidence measure

We hypothesize that a viable measure of confidence is the absolute value of the sum of the correlation from all SRNs, which in our case 3. Suppose we have M source data sets and generate M SRNs, and because of the nonuniformity of the array data used to generate the SRNs, for a given SRF/exon pair there maybe $0-M$ predictions. Suppose there are $N_{total,n,ij}$ SRNs predict an interaction of a specific SRF n/gene i exon j (either + or −). We define a confidence of a prediction $Q_{n,ij}$ as

$$Q_{n,ij} = \left| \sum_{m=1}^{N_{total,n,ij}} C_{n,ij,m} \right| \tag{8}$$

where $C_{n,ij,m}$ is the predicted correlation of SFR n and gene i exon j expression in SRN m. In our case, for any prediction, $N_{total,n,ij}$ is larger or equals to 1 and smaller or equals to M, which is 3. For each prediction, there is a confidence to evaluate its reliability. We set a confidence threshold value Q_{th}. If $Q_{n,ij} \geq Q_{th}$ then we say the regulatory interaction prediction is a high quality prediction.

3. Results

3.1. High quality splicing regulatory predictions

A predicted SRN is a list of exons for each of which a list of splicing regulator factors' (SRFs') up/down regulation is provided. A high quality regulatory prediction is a prediction with high confidence (defined in METHODS *VI*). Those predictions are archived in a database available at our website: https://ruby.chem.indiana.edu/~scorenfl/srn_results/lookup0.php. High quality predictions of splicing regulatory interactions and statistics of the predicted SRNs are provided in Table 4.

To assess the confidence threshold value Q_{th} that we consider a prediction to be a high quality prediction, we display a probability density distribution of confidence for all predictions (Fig. 2(a)). We set the correlation threshold $C_{th} = 0$, therefore for any SRF/exon, there is a prediction. From Fig. 2(a), it is clear that when confidence is low, from 0 to 0.8, the density is high and does not change much, which means at low confidence, interactions are almost equally distributed. A significant density drop happens when confidence increases from 0.8 to 1.75, and the change is very steep; when a confidence is above 1.75, the density only changes slightly. A lower confidence percentage function $A(Q_{n,ij})$ is defined such that

$A(Q_{n,ij})$ = percentage of interactions whose confidence is $\leq Q_{n,ij}$ (Fig. 2(b))

We chose our confidence threshold $Q_{th} = 1.24$, the point at which the probability of confidence changes most rapidly (Fig. 2(c)). From Fig. 2(b), it is seen that our high quality predictions, whose confidence $Q_{n,ij} \geq Q_{th}$, is the best 12% over all the predictions.

Predictions of SRF/exon interactions in SRN I, II and III are not always the same. If a prediction $b_{n,ij}$ (SRF n, gene i exon j) in SRN I equals that in SRN II, we say that predictions of SRF n/gene i exon j interaction in SRN I and SRN II are consistent, otherwise it is considered as inconsistent. A concept of consistency includes two aspects: (1) predictions of SRF n/gene i exon j interaction are made in both or all SRNs; and (2) their predictions agree. A prediction that is consistent in two SRNs is more reliable than that which is not consistency in any pair SRNs, and predictions that are consistent in all three SRNs have the highest reliability.

In our case, for any prediction in SRNs, it could be +1, −1 or 0, therefore there are 26 possibilities (exclude of 0, 0, 0). With all the possibilities, there are five categories, which are (I) the interaction is only predicted in one SRN; (II) the interaction is predicted in two SRNs and their predictions agree; (III) the interaction is predicted in two SRNs and their predictions disagree; (IV) the interaction is predicted in three SRNs and their predictions agree (most reliable); (V) the interaction is predicted in three SRNs and predictions in two SRNs agree, but disagree with that in the third SRN. Relative amounts of interactions in each category are shown in the pie chart below. Fig. 3 contains two pie charts, which show the percentages of interactions in each category in all the predictions with no confidence restriction, and in high quality predictions.

Table 4
SRNs I, II, and III are generated based on an analysis of normalized exon expression profiles I, II and III, respectively (Table 3).

Predicted SRN	Number of genes	Number of exons	Total number of predictions	Number of high quality predictions
SRN I	3074	13911	1069 854	169737
SRN II	4465	29101	1078 808	169974
SRN III	4186	28265	1064 234	157436

Our methodology predicts 169 737, 169 974, and 157 436 high quality regulatory interactions for SRNs I, II, and III. We archived our predictions on our web portal: https://ruby.chem.indiana.edu/~scorenfl/srn_results/lookup0.php.

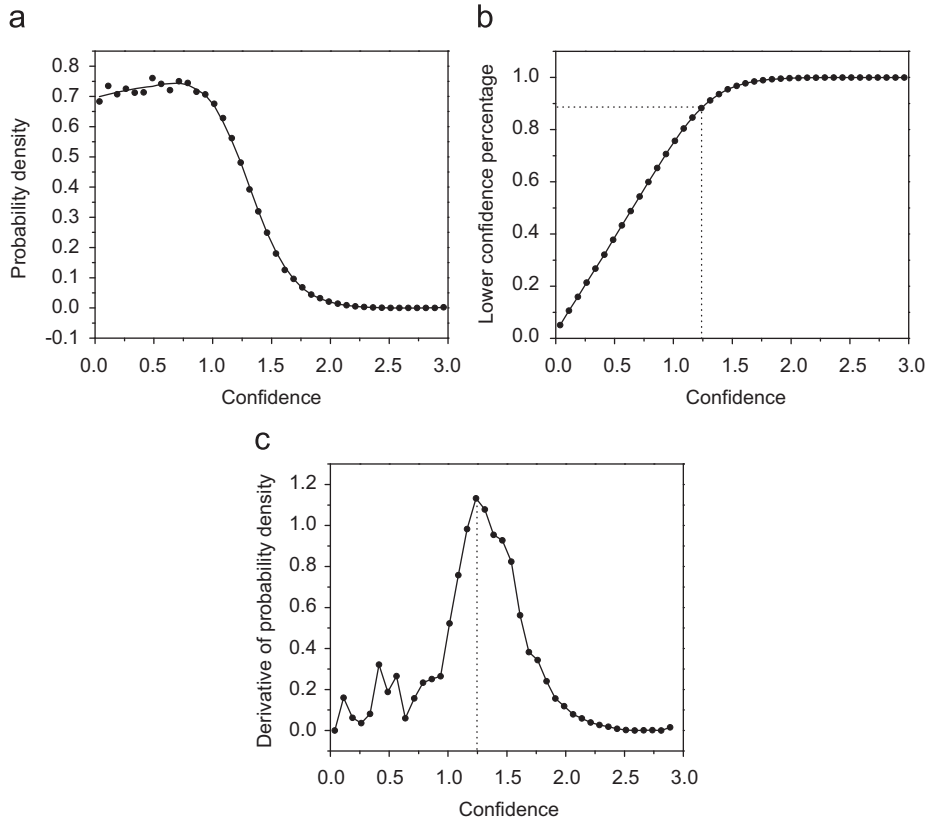

Fig. 2. (a) Probability density distribution of confidence for all the predictions. The *x*-axis is the confidence value, and the *y*-axis is the probability density; (b) the *x*-axis is confidence value, and the *y*-axis is the lower confidence percentage $A(Q_{n,ij})$, which indicates the percentage of interactions whose confidence are equal or smaller than $Q_{n,ij}$. A quality cut-off is set in dotted line where $Q_{th}=1.24$ and (c) the derivative of the probability density versus confidence. The high quality prediction confidence threshold is chosen to be the point at which the derivative is at its maximum.

From Fig. 3, we see, (1) within all the predictions, only 32.78% of all the interactions are the most reliable (category IV), while in high quality interactions, this percentage increases to 87.52%; (2) inconsistent prediction (categories III and V) consists 3.38%+ 36.49%=39.87% in all the predictions, but only 4.96% in high quality predictions. A higher consistent rate as well as a low inconsistent rate in high quality predictions showing a threshold of prediction confidence control does improve the quality of predictions.

3.2. *Predictions with higher correlation are more likely to be reliable*

Correlation methods have been widely used in statistics to indicate specific relationships between variables, notably in TF/gene regulation (Tuncay et al., 2006; Sayyed-Ahmad et al., 2007; Sun et al., 2007). We use a correlation approach to predict SRF/exon interactions. However, whether a higher correlation will result in a higher reliability, higher confidence and therefore a better prediction still is not demonstrated. Before we answer this question, we first define an average confidence

$$\overline{Q}=\left|\sum_{q=1}^{q_{total}^{x,x+dx}} Q\right| \tag{9}$$

For every interaction, we will have M correlation values $C_{n,ij,m}$ and a confidence value $Q_{n,ij}$ assigned to evaluate its reliability, exclusive the interactions in the training set. Suppose there are $dq_{total}^{x,x+dx}$ interactions whose absolute value of the correlations lies between x and $x\pm dx$, what is the average confidence of these $dq_{total}^{x,x+dx}$ interactions? And does this average confidence increase as absolute correlation increases?

From Fig. 4, it is seen that the higher the absolute correlation is, the higher the confidence, and therefore the better the prediction. Thus a prediction with higher correlation is more likely to be reliable. We set a quality control thresh value $Q_{th}=1.24$ to distinguish between a high and low quality prediction. From Fig. 4, we see when absolute correlation range is 0.70–0.75, the absolute correlation of a prediction lies between 0.70 and 0.75, on average, and its confidence is at $\sim$1.24. Thus, it is reasonable for us to set a threshold on correlation $C_{th}=0.75$. When a prediction has an absolute correlation larger than 0.75, it is very likely that its confidence is larger than 1.24, and therefore very likely to be a high quality prediction. The determination of a cutoff correlation indicating a high quality prediction is critical for applying our methodology and assessing the reliability of predictions made by new datasets. From Fig. 4, we see 0.75 is a reasonable cutoff.

We can only make a statement that "a prediction with higher correlation is more likely to be a better prediction", and not assert that "a prediction with higher correlation is a better prediction". It is possible that exceptional cases exist such as two interactions $b_{n1,ij}$ and $b_{n2,ij}$, the absolute correlation of $b_{n1,ij}$ is larger than that of $b_{n2,ij}$, but the confidence of predicted interaction $b_{n1,ij}$ is less than that of $b_{n2,ij}$. Two likely origins of exceptions could be errors in the input array data, and violations of our regulatory model. However, if our method is reliable, these exceptional cases should be statistically insignificant. In this study, we predicted 169,975 SRF/exon high quality interactions. Within all the pairs of predictions, 96.4% pairs of interactions are normal, which means the absolute correlation of one interaction is greater/less than the other and so is the confidence. Only 3.6% pairs are exceptional.

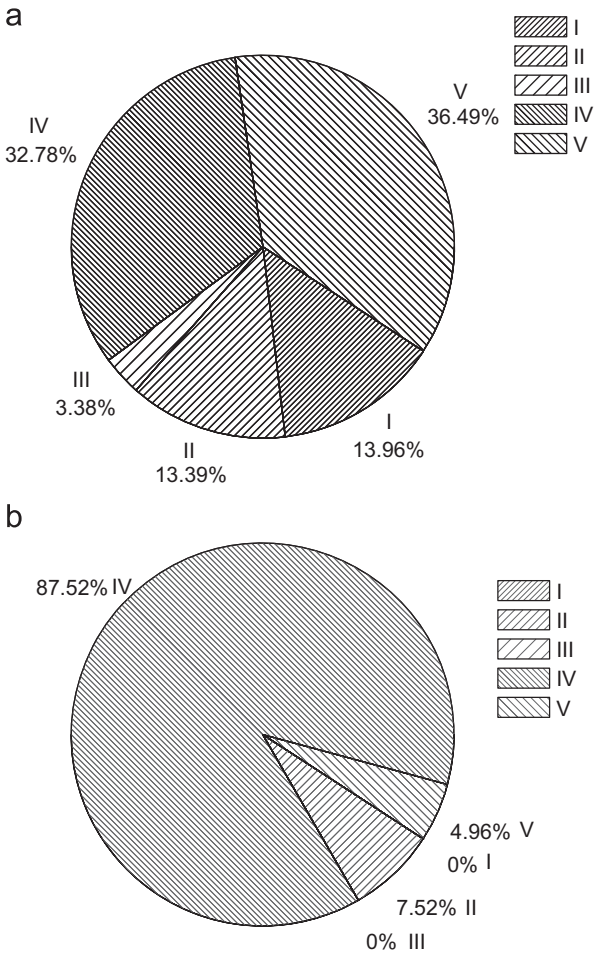

Fig. 3. (a) Pie chart of percentage of interactions in each category. Interactions are predicted with no confidence restriction and (b) pie chart of percentage of high quality interactions in each category. Interactions are predicted with a confidence threshold of 1.24.

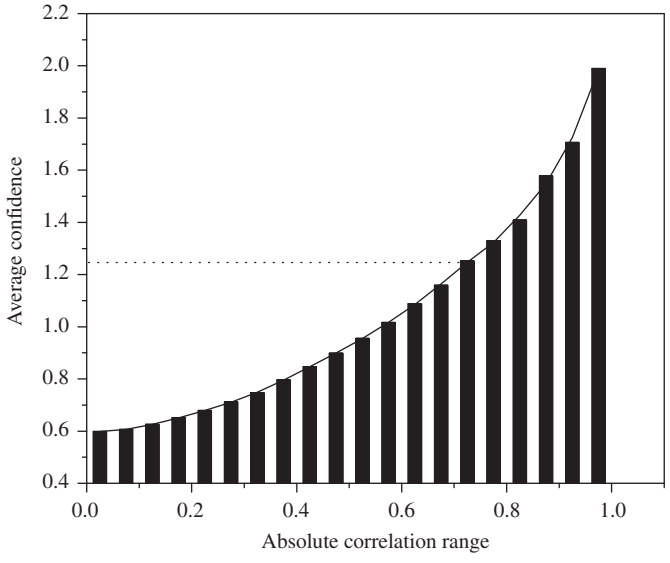

Fig. 4. Histogram of average confidence range as a function of absolute correlation. Absolute correlation of 20 ranges is calculated, and those ranges are defined as $0.025*(2N+1)\pm 0.025$, where N is an integer from 0 to 19.

3.3. Higher threshold results in better predictions

As discussed in Methods, we use a threshold C_{th} to determine whether a given interaction prediction is reliable, and we see from Fig. 4 that 0.75 is a reasonable choice. However, whether or not an increase of threshold C_{th} results in better predictions still remain to be demonstrated. Suppose at a fixed threshold C_{th}, we found I_{total} interactions are predicted in all three SRNs, and we compute an average confidence $\overline{Q}(C_{th})$ via the following equation

$$\overline{Q}(C_{th}) = \frac{1}{I_{total}} \sum_{i=1}^{I_{total}} Q_i \tag{10}$$

From Fig. 5, we see the larger the correlation threshold, the higher the average confidence, therefore the more reliable the predictions. A high quality prediction requires a confidence larger than 1.24 (as defined above), and from Fig. 5, we see, average confidence 1.24 lies between correlation threshold of 0.70–0.75, similar to Fig. 4, we conclude that a cutoff on correlation at 0.75 is reasonable, so that a prediction whose absolute correlation is larger than 0.75 is very likely to be a high quality prediction.

Probability distribution for different thresholds is shown in Fig. 6. It is seen that an increase of correlation threshold results in better predictions by increasing the probability density of higher confidence interactions, and decreasing that of the lower confidence ones.

3.4. Increase confidence yields agreement between SRNs reconstructed from data of different platforms

To assess the relationship between SRNs constructed from one dataset versus another, we constructed correlation diagrams of SRN I and SRN II, and SRN II and SRN III. Fig. 7 shows the correlation of non-quality screened predictions (frames a and c) and high quality predictions (frames b and d) in SRN I, II and III. SRN I and II are generated via normalized exon expressions I and II. The normalization of exon expressions I and II uses conventional gene expressions GDS1096 and GDS596, and both were conducted on a same platform GPL96. Fig. 7(a) shows the similarity between SRN I and II as assessed by a correlation diagram of their non-quality screened predictions in two aspects: (1) most predictions exist in the first and third quadrants (consistent predictions), and there are much fewer in the second and forth quadrants (inconsistent predictions); (2) the

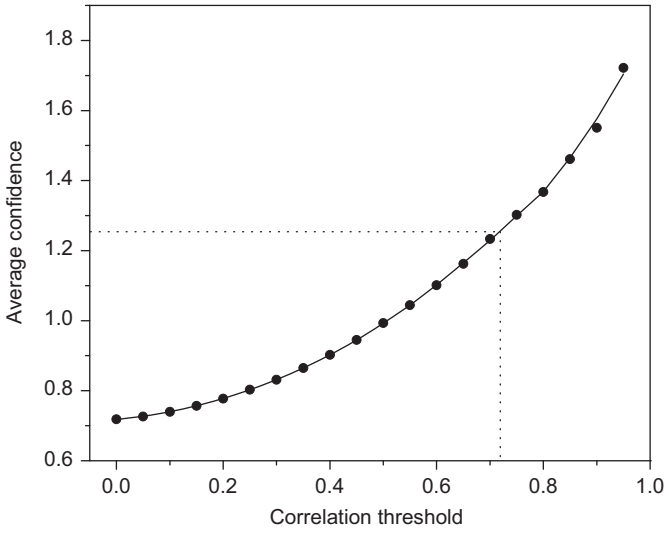

Fig. 5. Average confidence as a function of correlation threshold value.

higher the absolute value of the correlations, the more likely that the predictions are consistent (i.e. lie on a 45-degree line), conversely, the smaller the correlation, the greater the displacement from the 45-degree line. After we set a threshold value on the confidence, high quality predictions are selected and points cluster close to the 45-degree line, shown in Fig. 7(b). We see most of the high quality predictions cluster in the first and third quadrants, and therefore low quality predictions are effectively screened out by assign this threshold value on confidence. However, since SRNs I and II were constructed using a similar conventional microarray platform, this improvement is not surprising. The question still remains regarding correlation of SRNs constructed using data from different platforms, and the ability of our methodology to screen out low quality predictions to arrive at similar networks using data from different platforms still needs to be demonstrated. SRN II and SRN III are constructed from different platforms. The conventional gene microarray used in normalizing the exon expression that generates SRN III is GDS422, which was conducted on platform GPL91, different from that of GDS1096 (SRN I) and GDS596 (SRN II). As seen from Fig. 7(c), SRN II and SRN III are not similar: low quality prediction resided in all four quadrants, and there are many in the second quadrant, which represents a negative regulation predicted in SRN II, but a positive regulation predicted in SRN III. High quality predictions are more consistent as is shown in Fig. 7(d). The majority of high quality predictions are located in the first and third quadrants. This shows our methodology arrived at essentially the same SRN even when different array platforms were used (i.e. the technologies and our approach capture the underling biology).

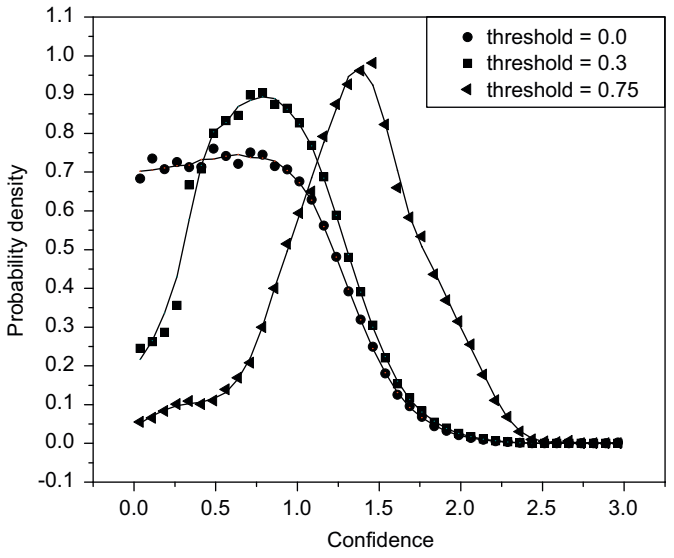

Fig. 6. Probability density distribution of confidence under correlation threshold 0 (circle), 0.3 (square), and 0.75 (triangle). The *x*-axis is confidence, and the *y*-axis is probability density.

3.5. *Assessing the reliability of applying the method to a new dataset*

Our methodology can be applied to reconstructed SRNs via new datasets is demonstrated in Section 4, we now explore the use of our findings to address the reliability of predictions on a new system and associated sets of array data. In particular, can the reliability cut off determined in the above exploratory study be transferred to be the cut off on the single dataset prediction? To establishing the cut off is a complex issue, however if the new

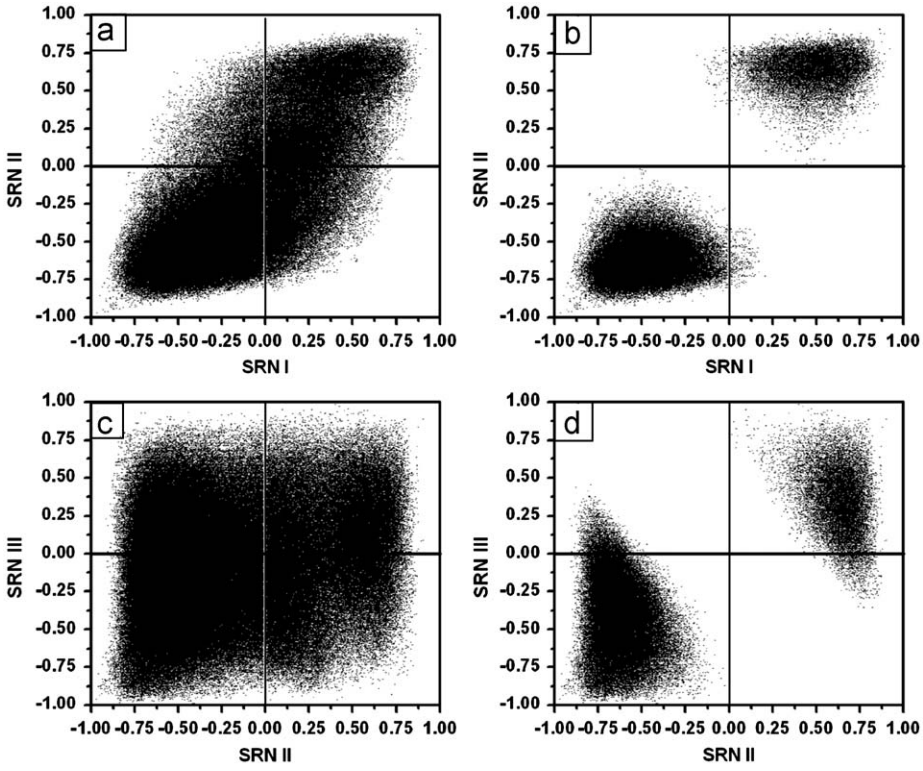

Fig. 7. (a) Correlation of non-quality screened predictions between SRN I and II; (b) correlation of high quality prediction between SRN I and II. Showing that as confidence increases, the correlation between predictions from different datasets improves; (c) correlation of non-quality screened predictions between SRN II and III and (d) correlation of high quality predictions between SRN II and III. Showing that even when SRNs constructed using different platforms (GPL96 versus GPL91), they still agree when confidence is high. The *x*-axes are the correlation predicted in one SRN, and the *y*-axes are those of the same interaction predicted in the other SRN.

dataset is comparable in accuracy, richness, and number of arrays to one of those in the above exploratory study, then we expect a similar confidence density distribution as presented in Section 1. Therefore, as we discussed in Sections 2 and Section 3, 0.75 is a reasonable absolute correlation cut off, in order to keep an average confidence of high quality predictions above a critical value of 1.24 (corresponding to the top 12%). Another cut off could be obtained by retaining only the top 12% highest correlation predictions.

4. Discussion

4.1. Innovation of our array-based reconstruction methodology

Given that genes encode all the information needed to run a cell, there has been a tremendous investment in understanding the biological function of each of the roughly 25,000 human genes and their encoding proteins based on total RNA and exon microarray technologies. Early attempts to interpret this data were based on gene–gene network and other highly simplified models (Chang et al., 2008; Chen et al., 2006; Gardner et al., 2003; Gutierrez-Rios et al., 2003; Huang et al., 2007; Li et al., 2004; Sano et al., 2006; Zhou et al., 2005; Zou and Conzen, 2005). While they are clear key parts of a cancer research strategy, we suggest that it is equally important to understand the genome-wide system of molecular signals that leads to normal versus abnormal patterns of gene and exon expressions. In our earlier work (Tuncay et al., 2006; Qu and Ortoleva, 2008; Qu et al., 2007; Sayyed-Ahmad et al., 2007; Sun et al., 2007) we deliver computational system biology technologies for reconstructing the network of processes regulating gene expression, and for the identification of gene regulatory subnetworks underlying the onset and progression of cancer. In this effort, we make a major advance in network discovery technology via constructing SRN based on gene and exon microarray analysis. Our unique methodology takes gene and exon microarray profiles as input and automatically generates alternative splicing regulatory processes as output. In comparing to other methodologies in reconstructing SRN via a sequence analysis (Modrek and Lee, 2002; Stamm et al., 2006; Takeda et al., 2007; Thanaraj et al., 2004), our methodology is unique in three aspects: (1) our methodology meets the great challenge in analyzing massive available microarray profiles; (2) our predictions are on a genome-wide perspective and therefore not limited to a few genes, exons or SRFs; and (3) our results will enable the computer-aided diagnosis and treatment of cancer by comparing hundreds of thousands of regulatory interactions in normal versus abnormal cells.

4.2. Assessment of the methodology

The methodology developed in this paper is based on several basic hypotheses: (1) SRN can be reconstructed based on available experimentally verified SRF/exon interactions and gene and exon microarray expression profiles; (2) higher confidence results in better predictions, and a threshold on confidence screens out low quality predictions; (3) a threshold value on confidence could be transferred to a threshold value on single correlation, and a higher correlation results in better prediction. These hypotheses are the rationale for our methodology and they were all validated in this paper and our earlier studies (Tuncay et al., 2006; Qu and Ortoleva, 2008; Qu et al., 2007; Sayyed-Ahmad et al., 2007; Sun et al., 2007). Our methodology yields genome-wide SRF/exon regulatory networks. Using our novel SRN reconstruction methodology and widely available gene microarray profiles with the limited experimentally verified training set of SRF/exon interactions, we predict genome-wide SRF/exon regulatory interactions. The SRNs we created also provide reliability scores (confidence) for all possible SRF/exon interactions. The higher the confidence scores the more reliable the prediction.

4.3. Genome-wide SRN reconstruction and its potential application

The development of strategies for preventing, diagnosing, and treating cancer and other diseases would be greatly facilitated by the availability of technologies for reconstructing the network regulating normal and abnormal cellular processes. Major controls on cellular activity are TF/gene interactions and SRF/exon interactions. The former controls the rate at which genes are transcribed into RNAs, while the latter controls the way that RNAs are cleaved and rejoined to create the mRNAs that are translated into proteins. The proteins are processed into the enzymes and regulatory molecules that underlie much of cellular behavior. Earlier we have shown that a key aspect of cancer is the existence of subnetworks of genes, the cross-talk among which can lead to runaway feedback we believe underlies the onset and progression of cellular abnormalities (Qu et al., 2007). To appreciate the complexity of the cellular control network, and the challenge that we face in identifying abnormalities in it that underlie cancer onset and progression, we developed methodologies that reconstructs genome-wide TRNs (Tuncay et al., 2006; Sayyed-Ahmad et al., 2007; Sun et al., 2007), and in this paper SRNs. Reconstructing networks of TF/gene and SRF/exon regulatory interactions are the first step before one can reliably identify the patient-specific cause of cancer onset and progression, and minimize uncertainty due to indirect causes otherwise missed in the millions of processes regulating the genome-wide network. As cancer involves an abnormality in the cellular regulatory network, genes/exons and the TFs/SRFs proteins that interact with them are high-value targets for therapeutic intervention and our search for the origins of cancer. In this way, we believe that our methodology will be a paradigm shift in identifying targets via this automated, genome-wide approach. Furthermore we believe the treatment that could be discovered will avoid adverse side-effects which would have been arrived at from an understanding of only one or a few genes and other factors, when the complexity and scope of their coupling to other genes across the wider cell regulatory network is ignored.

5. Conclusion

In this study, a unique methodology that constructs genome-wide alternative splicing regulatory networks (SRN) based on an analysis of exon and gene microarray expression profiles was presented. The rationale for the methodology is tested by statistical analyses of the predictions. Each prediction is assigned a confidence, which indicates its reliability. All predictions together with their confidence values are archived on our website: https://ruby.chem.indiana.edu/~scorenfl/srn_results/lookup0.php. Contrasting SRNs reconstructed using array data from normal versus diseased tissues, this methodology could identify splicing regulatory abnormalities underlying cancer and other diseases.

Acknowlegments

We thank S. Corenflos for assistance in archiving our predictions on our webportal. This project is supported in part by the United States Department of Energy (Genomics: Genome to Life Program), Oak Ridge Institute for Science and Education (Student

Research Participation Program), and the College of Arts and Sciences (Indiana University) as general support of the Center for Cell and Virus Theory.

References

Arikan, M.C., Memmott, J., Broderick, J.A., Lafyatis, R., Screaton, G., Stamm, S., Andreadis, A., 2002. Modulation of the membrane-binding projection domain of tau protein: splicing regulation of exon 3. Molecular Brain Research 101, 109–121.

Cartegni, L., Krainer, A., 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in *SMN*2 causes spinal muscular atrophy in the absence of *SMN*1. Nature 30, 377–384.

Castle, J., Garrett-Engele, P., Armour, C., Duenwald, S., Loerch, P., Meyer, M., Schadt, E., Stoughton, R., Parrish, M., Shoemaker, D., Johnson, J., 2003. Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. Biology 4 (R66), 1–13.

Chang, C.Q., Ding, Z., Hung, Y.S., Fung, P.C.W., 2008. Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. Bioinformatics 24, 1349–1358.

Chen, X.H., Chen, M., Ning, K.D., 2006. BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. Bioinformatics 22, 2952–2954.

Cuperlovic-Culf, M., Belacel, N., Culf, A., Ouellette, R., 2006. Microarray analysis of alternative splicing. Journal of Integrative Biology 10, 344–357.

Das, D., Clark, T., Schweitzer, A., Yamamoto, M., Marr, H., Arribere, J., Minovitsky, S., Poliakov, A., Dubchak, I., Blume, J., Conboy, J., 2007. A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. Nucleic Acids Research 35, 4845–4857.

Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J., 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301, 102–105.

Graveley, B.R., 2000. Sorting out the complexity of SR protein functions. RNA 6, 1197–1211.

Gutierrez-Rios, R.M., Rosenblueth, D.A., Loza, J.A., Huerta, A.M., Glasner, J.D., Blattner, F.R., Collado-Vides, J., 2003. Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. Genome Research 13, 2435–2443.

Huang, Z., Li, J., Su, H., Watts, G.S., Chen, H., 2007. Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining. Decision Support Systems 43, 1207–1225.

Johnson, J., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P., Armour, C., Santos, R., Schadt, E., Stoughton, R., Shoemaker, D., 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon-junction microarrrays. Science 302, 2141–2144.

Kondo, S., Yamamoto, N., Murakami, T., Okumura, M., Mayeda, A., Imaizumi, K., 2004. Tra2 beta, SF2/ASF and SRp30c modulate the function of an exonic splicing enhancer in exon 10 of tau pre-mRNA. Genes to Cells 9, 121–130.

Li, H.Q., Zhou, M., Cui, Y., 2004. Ranking gene regulatory network models with microarray data and Bayesian network. Data Mining and Knowledge Management 3327, 109–118.

Li, K., Arikan, M.C., Andreadis, A., 2003. Modulation of the membrane-binding domain of tau protein: splicing regulation of exon 2. Molecular Brain Research 116, 94–105.

Mabon, S.A., Misteli, T., 2005. Differential recruitment of pre-mRNA splicing factors to alternatively spliced transcripts in vivo. Plos Biology 3, 1893–1901.

Matlin, A., Clark, F., Smith, C., 2005. Understanding alternative splicing: towards a cellular code. Nature 6, 386–398.

Modrek, B., Lee, C., 2002. A genomic view of alternative splicing. Nature Genetics 30, 13–19.

Ortoleva, P., Automated transcriptional regulatory network discovery applied to diabetes, vol. 346,628.00. Government, Indiana University Bloomington CCVT 2007, pp. 1–73.

Pazos, F., Valencia, A., 2001. Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Engineering 14, 609–614.

Qu, K., Ortoleva, P., 2008. Understanding stem cell differentiation through self-organization theory. Journal of Theoretical Biology 250, 606–620.

Qu, K., Abi Haidar, A., Fan, J., Ensman, L., Tuncay, K., Jolly, M., Ortoleva, P., 2007. Cancer onset and progression: a genome-wide, nonlinear dynamical systems perspective on onconetworks. Journal of Theoretical Biology 246, 234–244.

Sano, R., Ogata, Y., Sakurai, N., Aoki, K., Suzuki, H., Saito, K., Shibata, D., 2006. Elucidation of regulatory network for isoprenoid biosynthesis using correlation coefficients from microarray expression data. Plant and Cell Physiology 47 S51–S51.

Sayyed-Ahmad, A., Tuncay, K., Ortoleva, P.J., 2007. Transcriptional regulatory network refinement and quantification through kinetic modeling, gene expression microarray data and information theory. BMC Bioinformatics 8:20.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T., Soreq, H., 2005. Function of alternative splicing. Gene 344, 1–20.

Stamm, S., Riethoven, J.J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N., Thanaraj, T., 2006. ASD: a bioinformatics resource on alternative splicing. Nucleic Acids Research 34, D46–D55.

Sun, J., Tuncay, K., Haidar, A.A., Stanley, F., Trelinski, M., Ortoleva, P., 2007. Transcriptional regulatory network discovery via multiple method integration: application to E. coli K12. Algorithms in Molecular Biology 2:2.

Takeda, J., Suzuki, Y., Nakao, M., Kuroda, T., Sugano, S., Gojobori, T., Imanishi, T., 2007. H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-invitational. Nucleic Acids Research 35, D104–D109.

Thanaraj, T.A., Stamm, S., Clark, F., Riethoven, J.J., Le Texier, V., Muilu, J., 2004. ASD: the alternative splicing database. Nucleic Acids Research 32, D64–D69.

Tuncay, K., Ensmann, L., Sun, J., Haidarn, A.A., Stanley, F., Trelinskin, M., Ortoleva, P., 2006. Transcriptional regulatory networks via gene ontology and expression data. In silico Biology 7.

Wang, J.N., Gao, Q.S., Wang, Y.Z., Lafyatis, R., Stamm, S., Andreadis, A., 2004. Tau exon 10, whose missplicing causes frontotemporal dementia, is regulated by an intricate interplay of cis elements and trans factors. Journal of Neurochemistry 88, 1078–1090.

Wang, Y.Z., Wang, J.N., Gao, L., Lafyatis, R., Stamm, S., Andreadis, A., 2005. Tau exons 2 and 10, which are misregulated in neurodegenerative diseases, are partly regulated by silencers which bind a SRp30c center dot SRp55 complex that either recruits or antagonizes htra2 beta 1. Journal of Biological Chemistry 280, 14230–14239.

Wu, J.Y., Kar, A., Kuo, D., Yu, B., Havlioglu, N., 2006. SRp54 (SFRS11), a regulator for tau exon 10 alternative splicing identified by an expression cloning strategy. Molecular and Cellular Biology 26, 6739–6747.

Xing, Y., Kapur, K., Wong, W., 2006. Probe selection and expression index computation of affymetrix exon arrays. PLoS 1 (e88), 1–9.

Zhou, X.H.J., Kao, M.C.J., Huang, H.Y., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O.M., Finch, C.E., Morgan, T.E., Wong, W.H., 2005. Functional annotation and network reconstruction through cross-platform integration of microarray data. Nature Biotechnology 23, 238–243.

Zou, M., Conzen, S.D., 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21, 71–79.