BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 5 2006, pages 606–613
doi:10.1093/bioinformatics/btk028

Data and text mining

Inferring global levels of alternative splicing isoforms using a generative model of microarray data

Ofer Shai[1,*], Quaid D. Morris[2], Benjamin J. Blencowe[2] and Brendan J. Frey[1]

[1]Department of Electrical and Computer Engineering, University of Toronto,Toronto, Canada, M5S 3G8 and [2]Banting and Best Department of Medical Research, University of Toronto,Toronto, Canada, M5G 1L6

Received on November 15, 2005; revised on December 21, 2005; accepted on December 23, 2005
Advance Access publication January 10, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Alternative splicing (AS) is a frequent step in metozoan gene expression whereby the exons of genes are spliced in different combinations to generate multiple isoforms of mature mRNA. AS functions to enrich an organism's proteomic complexity and regulates gene expression. Despite its importance, the mechanisms underlying AS and its regulation are not well understood, especially in the context of global gene expression patterns. We present here an algorithm referred to as the Generative model for the Alternative Splicing Array Platform (GenASAP) that can predict the levels of AS for thousands of exon skipping events using data generated from custom microarrays. GenASAP uses Bayesian learning in an unsupervised probability model to accurately predict AS levels from the microarray data. GenASAP is capable of learning the hybridization profiles of microarray data, while modeling noise processes and missing or aberrant data. GenASAP has been successfully applied to the global discovery and analysis of AS in mammalian cells and tissues.

Results: GenASAP was applied to data obtained from a custom microarray designed for the monitoring of 3126 AS events in mouse cells and tissues. The microarray design included probes specific for exon body and junction sequences formed by the splicing of exons. Our results show that GenASAP provides accurate predictions for over one-third of the total events, as verified by independent RT–PCR assays.

Contact: ofer@psi.toronto.edu

Supplementary information: http://www.psi.toronto.edu/GenASAP

1 INTRODUCTION

In higher eukaryotes, genes are often composed of multiple exons which must be spliced together to generate mature mRNA that can be translated into protein. Transcription of protein coding genes initially results in a primary transcript from which the intervening sequences or introns are removed and the exons are spliced together. The exons of some genes can be spliced in different combinations to result in structurally and functionally distinct mRNA isoforms. This process, known as alternative splicing (AS), is capable of resulting in orders of magnitude more mRNAs and corresponding proteins than there are genes (Black, 2000; Maniatis and Tasic, 2002; Matlin *et al*., 2005). It is currently estimated that ∼74% or more of human genes contain one or more alternative exons (Johnson *et al*., 2003). These findings indicate that AS could account for much of the increased complexity associated with higher eukaryotes, which cannot be accounted for by differences in gene counts (Black, 2000).

Other than contributing to the expansion of an organism's genetic repertoire, AS is known to play critical roles in the regulation of development, cellular differentiation, maintenance of the differentiated state and apoptosis. In addition, disruption of splicing is frequently associated with human diseases (Blencowe, 2000; Cartegni *et al*., 2002). The mechanisms underlying AS and its regulation are relatively poorly understood. In most previous studies, AS was studied on a case-by-case basis, whereas technology for analyzing AS on a global scale has only recently been introduced.

In recent years, numerous studies have investigated global properties of AS using databases of expressed sequence tags (ESTs) and complementary DNA (cDNA) sequence data (reviewed by Matlin *et al*., 2005). While these studies have been revealing, they are often constrained by the numbers of available EST/cDNA sequences, since these frequently do not represent the entire length of a gene or are only available from a limited number of cell or tissue sources. Additionally, many ESTs/cDNAs are derived from cell lines or tumor tissue, and therefore may not represent physiologically relevant splicing patterns (Wang *et al*., 2003b; Xu and Lee, 2003; Hui *et al*., 2004). More recently, microarrays have been implemented in large-scale studies of AS (Johnson *et al*., 2003; Pan *et al*., 2004; Blanchette *et al*., 2005; Relógio *et al*., 2005; Ule *et al*., 2005). The common approach is to design probes for the exon bodies and junctions that are involved in the AS events being studied. The array data are then analyzed and compared across samples and the differential expression is used to detect the existence of, or changes in, mRNA isoform levels.

We present here an algorithm that allows the quantitative estimation of relative mRNA isoform levels in mammalian cells and tissues. The algorithm was developed for the analysis of data from custom DNA microarrays specifically designed for this task (Pan *et al*., 2004). The algorithm, which we have named GenASAP (Generative model for the Alternative Splicing Array Platform), is based on a generative probabilistic model and uses machine learning to estimate relative AS levels in an unsupervised fashion. While it was designed for the prediction of single exon skipping AS events, GenASAP is an extendable model and is capable of handling any number of mRNA splice isoforms and probes. The results produced by GenASAP have already been used to reveal important global regulatory features of AS in adult mammalian tissues (Pan *et al*., 2004).

*To whom correspondence should be addressed.

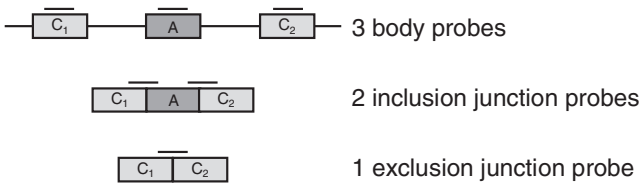

Fig. 1. Each AS event on the array is analyzed using six probes, as shown by the black lines. Three body probes monitor the three exons involved in the event, two junction probes monitor the inclusion isoform and one junction probe monitors the exclusion isoform.

2 ALTERNATIVE SPLICING MICROARRAY DATASET

We developed a custom DNA microarray for surveying AS levels on a large scale. This microarray contains probes for 3126 exon skipping events from 2647 unique mouse genes. The AS events were primarily selected on the basis of having strong EST/cDNA-based support (i.e. multiple independent ESTs/cDNAs sequences revealed skipping or inclusion of each alternative exon) (Pan *et al*., 2004, 2005).

The microarray contained multiple probes for each AS event, allowing for redundancy in the measurements and enabling quantitative analysis. Each event is analyzed using six probes, as shown in Figure 1. Three body probes are used to monitor each of the three exons involved in the AS event—the constitutive (always included) exons, C_1 and C_2, and the alternative exon, A. Two junction probes are used to monitor the two junctions formed by the inclusion of the alternative exon and one junction probe monitors the junction formed by the exclusion of the alternative exon and joining of the constitutive exons. In addition to the aforementioned six probes, a probe was included to monitor the intron located between the C_1 exon and the A exon. While it was anticipated that this probe would provide useful information on global levels of unspliced pre-mRNA, few genes appeared to accumulate significant levels of unspliced mRNA in our initial analyses of mouse tissue AS and we have since omitted intron probes from our more recent microarray designs. A full description of the dataset is available in Pan *et al*. (2004).

Using microarray data generated from the sets of six microarray probes specific for each AS event, our goal was to generate a robust algorithm for estimating the relative levels of tissue-specific AS for each alternative exon. The complexity of the noise processes, interactions between the unknown AS levels and the presence of other hidden variables such as probe sensitivity make the analysis of the microarray data set well suited for the application of a probability model. We describe such a model in this paper and demonstrate its reliability by comparison with measurements of AS levels from the same tissue samples using RT–PCR assays.

3 GENASAP—A GENERATIVE MODEL FOR THE ALTERNATIVE SPLICING ARRAY PLATFORM

We would like to infer the relative levels of the two isoforms contributing to the array measurements. To achieve this, we model the process that generated the array measurements, learn the maximum likelihood (ML) parameter settings and infer the conditional distribution of the isoform levels given the array data and the estimated parameters. While the model presented below can be used for evaluating any arbitrary number of isoforms and probes, we use the array described in Section 2 as an example to simplify the notation and discussion. Throughout the document, lowercase letters represent scalar variables, bold lowercase letters represent vectors and uppercase letters represent matrices.

We assume that there is a linear relationship between the intensity measured by the probe and the abundance of target mRNA containing the probe binding sequences. Therefore, we model the intensity measured at each probe as a linear combination of the abundance of the two isoforms, plus noise. This can be written as $x_i = \lambda_{i1}s_1 + \lambda_{i2}s_2 + \epsilon_i$, where x_i is one of the six real-value intensity measurements pertaining to a six probe AS event set from the microarray, s_1 and s_2 are the two unknown real-value abundances of the mRNA isoforms, λ_{i1} and λ_{i2} are the estimated affinity between the two mRNA isoforms and probe i, and ϵ_i is the additive noise component for probe i. Ideally, we would like to learn a hybridization profile (i.e. a set of probe affinities) for each six probe set targeting a single AS event across the samples. However, this is not possible for two reasons. First, in our particular dataset there are only 10 tissue samples, which is too few to confidently learn a profile for each AS event. Second, and more importantly, although the AS events represented on the array were mined from EST data found to represent alternative isoforms, it is likely that not all events exhibit variable splicing levels across the 10 tissues. Learning separate hybridization profiles for each AS event would therefore lead to overfitting, and we resign ourselves to learning global hybridization profiles, shared among all AS events.

To accurately infer the relative levels of the mRNA isoforms, it is crucial to have an appropriate noise model. Microarray noise has been previously shown to be scale dependent (Rocke and Durbin, 2001). Data preprocessing techniques, such as VSN (Huber *et al*., 2002; Durbin and Rocke, 2004), reduce this effect by transforming the intensity data to a log or $\sinh^{-1}$ domain.[1] However, for the model's linear isoform combination assumption to be valid, we must maintain the microarray measurements in the intensity domain. Additionally, we must also account for outlying measurements resulting from faulty probes, aberrations on the array surface, non-specific binding and other experimentally introduced errors using an outlier model. To account for the scale-dependent noise and outlying measurements, we rewrite the model as

$$x_i = \left(r\left(\sum_j \lambda_{ij} s_j + \varepsilon_i \right) \right)^{1-o_i} (\zeta_i)^{o_i}, \qquad (1)$$

where the subscript j indexes isoforms, the scale factor r is a real number accounting for noise levels at the measured intensity, ζ_i is a pure noise component representing an outlying measurement and, the binary indicator, $o_i \in \{0, 1\}$ identifies a probe measurement as being an outlier ($o_i = 1$) or valid ($o_i = 0$).

The model is shown graphically as a Bayesian network in Figure 2. A Bayesian network is a directed acyclic graph where each vertex represents a variable in the model and the directed edges represent conditional dependencies. (Pearl, 1988). The Bayesian network visually demonstrates that each array observation, x_i, is dependent on the isoform abundances, $\mathbf{s}$, the scale factor, r, and an outlier indicators, $\mathbf{o}$. The network's roots do not depend on other variables and are independent a priori.

[1] $\sin h^{-1}$ is a log-like domain that is defined for negative values and is approximately linear near the origin.

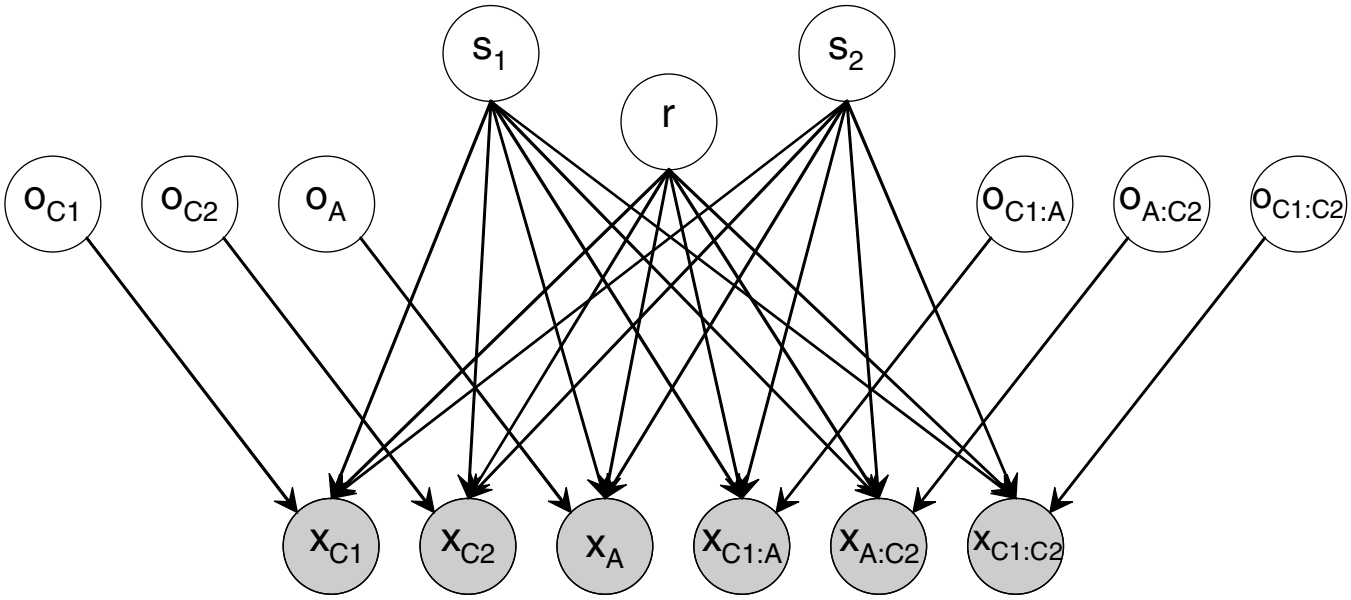

Fig. 2. A generative probability model for the inference of alternative splicing levels from the microarray data, shown as a Bayesian network. The six observed microarray intensities are modeled as a linear combination of the two isoform levels, s_1 and s_2, affected by scale-dependent noise, r. The model allows observations to be marked as outliers, as indicated by the binary indicator variables $\{o_{C_1}, o_{C_2}, o_A, o_{C_1:A}, o_{A:C_2}, o_{C_1:C_2}\}$, and be associated with the outlier process.

Under the assumption of independent, zero-mean, normally distributed noise, the conditional probability of the data given the isoform levels, scale and outlier indicators, can be written as

$$P(\mathbf{x} \mid \mathbf{s}, r, \mathbf{o}) = \prod_i \mathcal{N}\left(x_i; r \sum_j \lambda_{ij} s_j, r^2 \psi_i\right)^{1-o_i} \mathcal{N}(x_i; \mathcal{E}_i, \mathcal{V}_i)^{o_i}, \quad (2)$$

where $\mathcal{N}(x;\ \mu,\ \sigma^2)$ indicates the density of point x under normal distribution with mean μ and variance σ^2, the variance of the noise at probe i is given by ψ_i, and the mean and variance of the outlier model are given by $\mathcal{E}_i$ and $\mathcal{V}_i$, respectively.

Due to the biological interpretation of the variables and parameters in the model, there are certain positivity constraints that must be met. First, the isoform abundances, $\mathbf{s}$, may not take negative values. Also, the hybridization profiles, Λ (a matrix whose *i-th*, *j-th* element is λ_{ij}) may not assume negative values, since the presence of an isoform should not reduce the measured intensity. The constraint on the isoform abundances is enforced by setting its prior to a truncated Gaussian distribution, as given by

$$P(\mathbf{s}) = \mathcal{Z}^{-1} \mathcal{N}(\mathbf{s}, 0, I)[\mathbf{s} \geq 0], \quad (3)$$

where $[\cdot]$ is the indicator function such that $[\mathbf{s} \geq 0] = 1$ if $\forall j,\ s_j \geq 0$, and $[\mathbf{s} \geq 0] = 0$ otherwise, I is the identity matrix and $\mathcal{Z} = \int_{\mathbf{s}} \mathcal{N}(\mathbf{s}, 0, I)[\mathbf{s} \geq 0] d\mathbf{s}$ is the partition function. The truncated Gaussian distribution enables much of the analysis to be carried out analytically, while satisfying the constraints. Other non-negative prior distributions may be used, however, such as the gamma distribution.

To completely specify the joint probability, we require priors over the remaining noise processes, $\mathbf{o}$ and r. The prior for the indicator variable $\mathbf{o}$ is parameterized as $P(o_i = 1) = \gamma_i$, where γ_i is a learned parameter reflecting the probability of each type of probe to be an outlier a priori. For computational efficiency, we select r from a discrete set of possible values, $r \in \{\rho_1, \rho_2, \ldots, \rho_C\}$ and set the a priori probability to the uniform probability: $P(r = \rho_k) = 1/C$.

3.1 Inferring isoform levels

We next address the task of inferring the relative levels of the mRNA isoforms. The parameters of the model are shared among all AS events on the arrays and comprised the noise variances, Ψ, outlier probabilities, γ, the set of possible values for the scale factor, $\{\rho_1, \rho_2, \ldots, \rho_C\}$, the outlier model's mean and variance, $\mathcal{E}_i$ and $\mathcal{V}_i$, and the hybridization profiles, Λ. Additionally, the generative model contains observed and hidden (latent) variables that are unique for each AS event studied. The observed variables are the microarray measurements, $\mathbf{x}$, and the latent variables include the isoform levels, $\mathbf{s}$, outlier indicators, $\mathbf{o}$, and the scale factor, $\mathbf{r}$.

If the parameters of the model are specified and known, the posterior distribution of the latent variables can be obtained using Bayes' Rule:

$$P(\mathbf{s}, r, \mathbf{o} \mid \mathbf{x}) = \frac{P(\mathbf{s}, r, \mathbf{o}, \mathbf{x})}{P(\mathbf{x})} = \frac{P(\mathbf{s}, r, \mathbf{o}, \mathbf{x})}{\sum_r \sum_{\mathbf{o}} \int_{\mathbf{s}} P(\mathbf{s}, r, \mathbf{o}, \mathbf{x})}. \quad (4)$$

While the parameters could be set by hand, according to a predetermined strategy, we estimate the parameter values that maximize the likelihood of the model, $P(\mathbf{x})$. For example, while an intuitive form of the hybridization matrix, Λ, reflects the hybridization profiles (i.e. the exclusion isoform should bind strongly to the C_1, C_2 and C_1–C_2, probes, while hybridizing weakly or none at all to the other probes), we do not know the precise form and levels of cross hybridization between probes. As we show in Section 4.3, the ML estimate of the hybridization profiles agrees more closely with the independent validation.

Computing the likelihood of the model, $P(\mathbf{x})$, requires summing over the latent variables, as shown in Equation (4). In the GenASAP model, exact computation of $P(\mathbf{x})$ is infeasible. $P(\mathbf{x})$ is a mixture model, where each mixture component corresponds to a particular configuration of r and $\mathbf{o}$. Each mixture component is given by

$P(\mathbf{x} \mid r, \mathbf{o}) = \int_{\mathbf{s}} P(\mathbf{x} \mid \mathbf{s}, r, \mathbf{o}) P(\mathbf{s})$, which is a truncated Gaussian. Integrating over $\mathbf{s}$ thus requires evaluating the partition function of a multivariate truncated Gaussian. This integral is not analytically computable in the general case (Cozman and Krotkov, 1995). However, the partition function is computable under special circumstances, such as when the variables are independent.

When the exact posterior cannot be computed, it is common to employ sampling methods. We found these methods to be too computationally intensive due to the necessity to generate thousands of samples for each AS event during inference. The approach we take is to use variational methods. Variational methods use approximate inference to circumvent calculating the exact posterior. Variational learning proved to be significantly faster, as approximate inference can be carried out analytically.

3.1.1 Variational learning in GenASAP Variational methods are most easily explained in the context of the Expectation–Maximization (EM) algorithm (Dempster *et al*., 1977). The EM algorithm finds the ML estimate of the parameters by starting with an arbitrary parameter setting. EM iteratively computes the posterior distribution over hidden variables in the *E*-step and maximizes the log-likelihood of the model with respect to the parameters, while keeping the posterior fixed, in the *M*-step. When the posterior cannot be computed, variational EM replaces the exact posterior with a computationally tractable approximation, and maximizes a lower bound of the log-likelihood.

Under the mean field approximation, the approximate distribution is chosen such that all variables are independent (Neal and Hinton, 1998). This type of distribution is often easily modeled and computed but sacrifices knowledge of the structure inherent in the model. In GenASAP, we make a partial mean field approximation that retains much of the structure of the true posterior. The approximate posterior is given by

$$\begin{aligned} P(\mathbf{s}_t, r_t, \mathbf{o}_t \mid \mathbf{x}_t) &= P(r_t \mid \mathbf{x}_t) P(\mathbf{o}_t \mid \mathbf{x}_t, r_t) P(\mathbf{s}_t \mid \mathbf{x}_t, r_t, \mathbf{o}_t) \\ &\approx Q(r_t) Q(\mathbf{o}_t \mid r_t) \prod_i Q(\mathbf{s}_{i,t} \mid r_t, \mathbf{o}_t), \end{aligned} \tag{5}$$

where the subscript t indexes over the AS events on the array (each represented by six probes targeting two isoforms). Note that the Q distribution depends on the observed data, $\mathbf{x}$ indirectly through the variational parameters (see below). $Q(r_t)$ and $Q(\mathbf{o}_t \mid r_t)$ are discrete distributions, and together represent the responsibility of each of the mixture components. $Q(s_{jt} \mid r_t, \mathbf{o}_t)$ is parameterized as $Q(s_{jt} \mid r_t, \mathbf{o}_t) \propto \mathcal{N}(s_{jt}; \mu_{jt\mathbf{o}r}; \sigma_{jt\mathbf{o}r})[s_{jt} \geq 0]$. Thus, the a posteriori interdependence within $\mathbf{s}$ is disregarded, but the dependence of $\mathbf{s}$ on r and $\mathbf{o}$ is retained in the approximation. To avoid overfitting, we constrain $Q(\mathbf{o}_t \mid r_t)$ to those configurations where at most two of the probes are marked as outliers.

Since we cannot compute the likelihood or log-likelihood of the model, we can no longer maximize it. We can, however, make use of Jensen's inequality to maximize a lower bound of the log-likelihood (Neal and Hinton, 1998):

$$\begin{aligned} \log P(\mathbf{x}) &= \log \left(\int_{\mathbf{s}} \sum_{\mathbf{o}} \sum_{\mathbf{r}} Q(\mathbf{s}, \mathbf{o}, r) \frac{P(\mathbf{s}, \mathbf{o}, r, \mathbf{x})}{Q(\mathbf{s}, \mathbf{o}, r)} \right) \\ &\geq \int_{\mathbf{s}} \sum_{\mathbf{o}} \sum_{r} Q(\mathbf{s}, \mathbf{o}, r) \log \left(\frac{P(\mathbf{s}, \mathbf{o}, r, \mathbf{x})}{Q(\mathbf{s}, \mathbf{o}, r)} \right) \\ &= -\mathcal{F}(P, Q). \end{aligned} \tag{6}$$

The term $\mathcal{F}(P, Q)$ is the free energy of the model[2]. Minimizing the free energy is thus equivalent to maximizing a lower bound of the log-likelihood of the model. The variational EM learning in GenASAP proceeds as follows:

(1) Initialize model parameters.

(2) *E*: step: Minimize $\mathcal{F}(P, Q)$ with respect to the variational parameters of the Q distribution, $\{\mu_{jt\mathbf{o}r}\}$, $\{\sigma_{jt\mathbf{o}r}\}$, $\{Q(\mathbf{o}_t \mid r)\}$, and $\{Q(r_t)\}$, while keeping the model parameters fixed.

(3) *M*: step: Minimize $\mathcal{F}(P, Q)$ with respect to the model parameters, Λ, Ψ, and $\{\rho_1, \rho_2, \ldots, \rho_C\}$, while keeping the variational parameters fixed.

(4) Repeat steps 2 and 3 until convergence.

The minimizations in steps 2 and 3 can be carried out by setting partial derivatives of the free energy to zero, while enforcing the constraint that the Q distribution must be positive and integrate to 1. The variational updates for steps 2 and 3 are available in the Supplementary information. After convergence, we are left with an estimation of the optimal parameters for the model and an approximation of the posterior distribution. The setting of $\mathbf{s}_t$, r_t and $\mathbf{o}_t$ that maximizes $Q(\mathbf{s}_t, r_t, \mathbf{o}_t)$ approximates the maximum a posteriori (MAP) estimate of the outlier indicator, scaling factor and, most usefully, the isoform levels, which are used as estimates of the mRNA isoform abundances.

4 INTERPRETING THE MODEL'S POSTERIOR TO PREDICT ALTERNATIVE SPLICING

The results presented in this paper were obtained using two stages of learning. In the first stage, the hybridization profile, Λ, is learned on a subset of the data for which both the constitutive exon body probes, C_1 and C_2, measured higher expression than the 75th percentile of the intron probes. In the second step, Λ is kept fixed, and we introduce the additional constraint that the noise is isotropic ($\Psi = \psi I$) and learn on the entire dataset. The constraint on the noise is introduced to prevent the model from using only a subset of the six probes for making the final set of predictions. The analysis was repeated 20 times with different initial parameter settings to evaluate consistency, and confidence intervals on the results were computed. Each trial was allowed to run to convergence, which was determined by a proportional change of <0.0001% of the free energy between iterations.

The learned set of hybridization profiles, including confidence intervals, are shown in Figure 3a. The profiles fit well with our intuition of what would be expected based on detection of the two splice isoforms. The two sets of hybridization profiles clearly identify the two isoforms by accounting for hybridization to the exon body and junction probes. Moreover, the learned profiles account for unanticipated trends in the data, exhibiting increased cross hybridization in the junction probes, and no hybridization of the excluded isoform to the alternative exon. For instance, while the exclusion junction probe C_1–C_2 was designed to monitor the exclusion isoform, the entirety of its sequence is present in the inclusion

[2] $\mathcal{F}(P, Q)$ is equivalent to the free energy of the system from statistical physics, if we consider the negative log joint probability, $-\log P(\mathbf{s}, r, \mathbf{o}, \mathbf{x})$, as the energy function of the system and the approximation, $Q(\mathbf{s}, r, \mathbf{o})$, as a distribution over the states of the system. The free energy is also equivalent to the Kullback–Liebler (KL) divergence, or relative entropy, between the joint distribution and the approximate posterior.

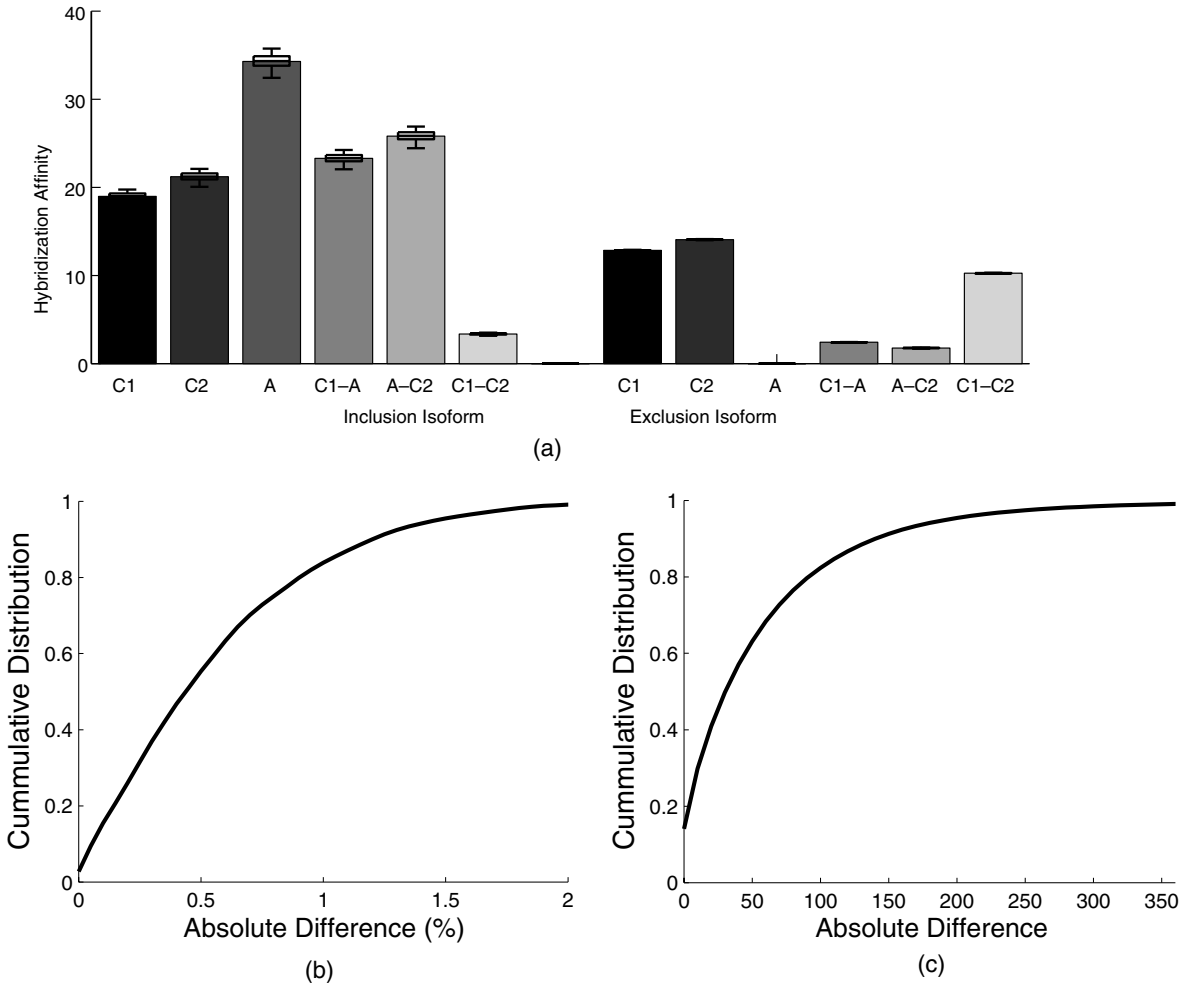

Fig. 3. (**a**) The profiles learned by GenASAP. Each bar represents the mean weight over 20 trials. Overlaying the bars are box-plots showing the consistency of the learned profiles. The center line of each box represents the median value, the edges of the boxes represent the 25th and 75th percentiles and the extended bars show the entire range of the parameters. (**b**) The percent exclusion predictions from each trial were compared with all other trials to evaluate consistency. The cumulative distribution of the absolute difference in percent exclusion across the trials is shown. (**c**) Similarly, the AS tissue ranks assigned to the predictions from each trial were compared with all other trials. The cumulative distribution of the absolute difference in ranks across the trials is shown.

isoform, albeit separated by the alternative exon. So, while we would not expect the inclusion isoform to bind to the exclusion junction well, it is interesting to note that some cross hybridization is detected by the model.

We do not expect all the inferred abundances to be equally accurate. Specifically, since most tissues only express a small subset of the genes, many of the events on the array would have little or no transcription. Additionally, many events would have poor agreement with the model due to the use of global hybridization profiles. A scoring criterion was used that ranks each AS event based on its transcription level and fit to the model. The prediction score, S, is given by

$$S_t = \frac{\sum_i \frac{|x_{it} - r_t \lambda_i \mathbf{s}_t|}{r_t \|\mathbf{s}\|}}{\frac{x_{it} + r_t \lambda_i \mathbf{s}_t}{2}}. \quad (7)$$

This criterion can be interpreted as the signal-to-noise ratio, scaled by the estimated transcript levels to further account for the expected low signal content of low abundance measurements. The scores of all the AS events across the 10 tissues are sorted and ranked, and each predicted AS level value receives an AS tissue rank—a number ranging from 1 to the total number of surveyed AS events (31 260).

Since it is the relative amount of the isoforms that is of most interest, we use the inferred distribution of the isoform abundances to obtain an estimate for the relative levels of AS isoforms. We refer to the isoforms that contain and skip the alternative exon as the inclusion and exclusion isoforms, respectively. Intuitively, the percent of the excluded alternatively spliced isoform is given by $S_{\text{ex}}/(S_{\text{ex}} + S_{\text{inc}})$, where s_{ex} and s_{inc} are the MAP estimation for the exclusion and inclusion isoforms, respectively.

4.1 Reproducibility of GenASAP predictions

Recall that each of the 20 trials was initialized randomly with a different set of parameters, potentially leading to different local minimum of the free energy. Yet, the hybridization profiles learned over the 20 trials exhibit remarkable reproducibility. As shown in Figure 3a, the included isoform displays a slight variation across the 20 trials, with a variance of $\sim$2% of the mean of the learned profile.

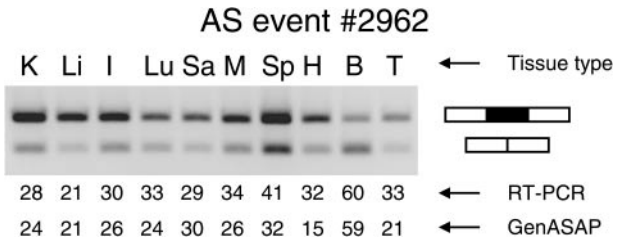

Fig. 4. Sample RT–PCR gels for an AS event across the 10 mouse tissues. Bands of sizes expected for the included and excluded mRNA isoforms are indicated. The labels above the gels indicate the tissues: Kidney, Liver, Intestine, Lung, Salivary gland, skeletal Muscle, Spleen, Heart, Brain and Testis. Below the gel, measurements for percent exclusion estimated from the RT–PCR and GenASAP are shown.

The learned profile for the excluded isoform is even more consistent, with variance $\ll 1\%$ of the mean. As we show next, this consistency led to nearly identical predictions for the relative isoform levels across the 20 trials.

Comparing the percent exclusion predictions across the 20 trials reveals a high level of consistency. The correlation of the predictions across the trials was found to be 0.9989 ± 0.0009. All the predictions from each trial were compared with the equivalent predictions from all other trials. This yielded 5 939 400 comparisons as follows: 3126 events across 10 tissues gives 31 260 predictions. A total of 20 trials yields 190 pair-wise comparisons for each prediction. Of the comparisons, 84% are ±1% exclusion, and >99% are ±2% exclusion, demonstrating insensitivity of GenASAP to the initial parameter settings (Fig. 3b).

The last evaluation for consistency is for the AS tissue ranks. As with the hybridization profiles and predictions, the assigned ranks are consistent across the random restarts. All the ranks from each trial were compared with the equivalent predictions from all other trials, again yielding 5 939 400 comparisons. Of the comparisons, 80% shift by <0.3% of the range of AS tissue ranks, and 99% of the comparisons shift by <1.2% (Fig. 3c).

4.2 RT–PCR validation

To evaluate the results obtained from GenASAP, we carried out 241 RT–PCR assays covering a wide range of percent exclusion values and AS tissue ranks. RT–PCR assays are often used as a semi-quantitative (i.e. noisy) method to validate microarray data. Figure 4 shows a sample RT–PCR assay carried out for an AS event across the 10 mouse tissues. Primer pairs were designed to have matching T_m (59°C) and were targeted to constant exon sequences flanking each alternative exon. Gel images were recorded using a Syngene gel documentation system and quantified with Gene Snap software. Each column produced two measurements corresponding to the two isoforms, and the RT–PCR-measured AS levels were calculated as $I_{ex}/(I_{ex} + I_{inc})$, where I_{ex} and I_{inc} are the measured intensity of the exclusion and inclusion isoforms, respectively.

While evaluating GenASAP's performance, it was evident that the unsupervised percent exclusion measurement presented in Section 4, while correctly predicting the trends in the data (Fig. 5b), produced biased predictions (Fig. 5c). This bias seems to have arisen from two sources. First, the model does not allow for a constant offset in the microarray measurement. Second, there is an inherent degeneracy in the model. One could, e.g. multiply all the inferred s_1 levels, while dividing the corresponding hybridization profile by the same constant to achieve the same reconstruction. We correct this bias by applying a simple transformation to the predictions. We estimate the transformation by minimizing the root mean squared error (RMSE) between the final predictions given by $a_1 S_{ex}/(S_{ex} + a_2 S_{inc}) + a_3$ and the RT–PCR measured percent exclusion based on the top 50 RT–PCR measurements as determined by the AS tissue ranks assigned by GenASAP. The two affine transform parameters, a_1 and a_3 correct for a constant offset and multiplier across all predictions, while the parameter a_2 corrects for the degeneracy in the model. We refer to the percent exclusion predictions obtained from the above transformation as the semi-supervised predictions.

We used three criteria to evaluate the performance of GenASAP. Each criterion is designed to assess the level of information available for different types of data analysis. The first criterion is correlation between the RT–PCR measurements and GenASAP's predictions, as measured using the Pearson correlation coefficient. This criterion assesses GenASAP's ability to correctly predict trends and patterns in the AS data. Figure 5b shows that the correlation between the RT–PCR measured percent exclusion and the semi-supervised predictions is >0.85 for the top 5000 predictions and >0.75 for the top 10 000 predictions as determined by AS tissue ranks (solid gray line). The correlation coefficient measure is only slightly lower for the unsupervised predictions, as it is invariant to the a_1 and a_3 transformation parameters. The discrepancy between the unsupervised and semi-supervised predictions arise solely from the a_2 parameter which corrects for a constant and arbitrary scaling of the isoform levels. In many applications of GenASAP, where a few samples are compared for changes in AS levels, the unsupervised predictions would suffice to provide accurate predictions of trends.

The second criterion used is the RMSE between the RT–PCR measured percent exclusion and GenASAP's predictions. The RMSE plot in Figure 5c shows that there is a high level of agreement between GenASAP's semi-supervised predictions and the RT–PCR measurements. We were able to establish an RMSE of <14% for the top 5000 predictions and <17% for the top 10 000. The RMSE measure is, of course, highly dependent on the transformation applied to the predictions, and the unsupervised predictions do not preform as well as the semi-supervised predictions. It is therefore advisable to carry out RT–PCR validation for new experiments and adjust for bias in the predictions if the accurate values of percent exclusions are required.

The third and final criterion used is GenASAP's ability to correctly predict high, medium and low percent exclusion. The domain of percent exclusion was divided to three overlapping regions of low exclusion (% <35%), medium exclusion (25% $\leq$% $\leq$75%) and high exclusion (% >65%). The overlap is necessary so as to avoid situations where a prediction may be close to a boundary, yet on the wrong side and therefore marked as erroneous. These situations would arise with any arbitrarily specified thresholds, but are avoided using overlapping regions. Figure 5d shows that 91% of the predictions fall into the correct category for the top 5000 predictions, and 86% of the events are correctly categorized for the top 10 000 predictions. Interestingly, when considering the entire set of 31 260 predictions, >80% of the predictions are correctly categorized. The task of categorizing predictions is inherently simpler than predicting

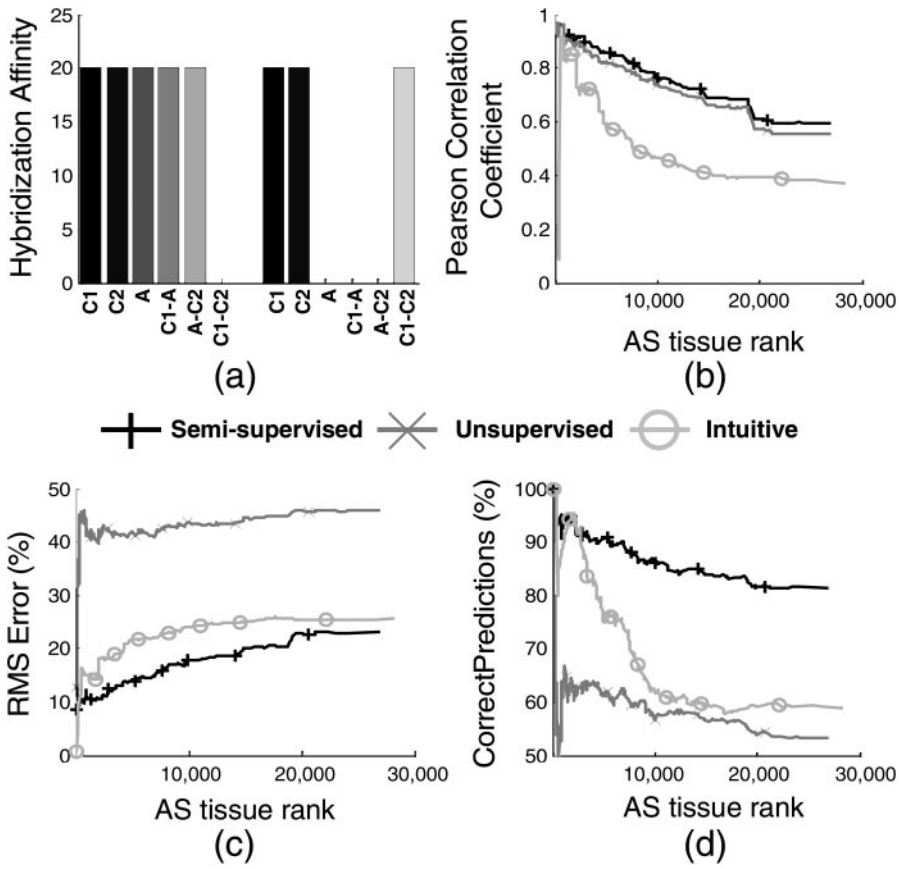

Fig. 5. (**a**) An intuitive hybridization profile, where both isoforms bind evenly and exclusively to the probes designed for them. (**b**) Cumulative correlation betweeen GenASAP predictions and the RT–PCR measurements. The correlation is evaluated at numerous AS tissue rank cutoffs and plotted as a function of rank. (**c**) Similarly, the RMSE between the RT–PCR measurements and GenASAP predictions is plotted as a function of rank cutoff. (**d**) The domain of percent exclusion was divided to three overlapping regions of low exclusion (% <35%), medium exclusion (25% $\leq$ % $\leqslant$ 75%) and high exclusion (% >65%). The portion of correct predictions as a function of rank is shown, where correct predictions are defined as those predictions that fall into the same region as the RT–PCR measurements.

the actual level of AS, and GenASAP also performs well in this respect.

4.3 The value of learning

We mentioned in Section 3.1 that learning the hybridization profile improves the performance of GenASAP. A set of predictions obtained using the intuitive hybridization profile is shown in Figure 5a. This hybridization profile is representative of how an ideal binding profile may appear. In this profile, the isoforms bind evenly and exclusively to the corresponding body and junction probes with no cross hybridization. The rest of the parameters in the GenASAP model were allowed to assume their ML values. The algorithm was executed and the predictions ranked. As in the case of the prediction obtained using the learned profile (learned predictions), the predictions obtained using the intuitive profile (intuitive predictions) were fitted to the RT–PCR measurements based on the top 50 ranked predictions, as explained in Section 4.

The predictions based on the intuitively selected hybridization profiles were evaluated using the same three criteria discussed in Section 4.2, and results are shown in Figure 5b, c and d. A comparison of the learned predictions (black line) greatly outperform the intuitive prediction (light gray line) for every criteria and in every rank cutoff. The intuitive predictions are similar to what one would obtain by simply comparing the exclusion junction probe, C_1–C_2, with the three inclusion isoforms, C_1–A, A and A–C_2, while allowing for outliers. Even when testing the abilities to categorize the events based on high, medium and low exclusion, the intuitive predictions do not perform satisfactorily, with only the top 5000 predictions scoring above the 80% accuracy mark, compared with the learned predictions, which are all above the 80% accuracy mark.

5 DISCUSSION

We have demonstrated the use of a generative model (GenASAP) for modeling the processes involved in generating microarray data for the analysis of alternative splicing levels in mammalian mRNA samples. GenASAP is able to extract thousands of accurate predictions of AS levels using a microarray designed for the study of AS. Although our array was designed to target two splice isoforms using six probes, GenASAP is an extendable model, capable of handling additional numbers and types of isoforms and probe sets, and including new noise models. Using unsupervised variational learning, GenASAP estimates the ML setting of the parameters that influence the detection of relative splice isoform levels

in microarray data. The resulting predictions for the relative splice isoform levels obtained from GenASAP correlate well with RT–PCR assays. The agreement between RT–PCR and the GenASAP predictions was found to be accurate for about one-third of the total number of surveyed events. This number most likely represents the expected number of events that one would expect to express at levels for which confident predictions are obtainable.

One of the key features that makes GenASAP capable of generating accurate predictions is the outlier model. The outlier model detects probes that exhibit aberrant behavior and disregards these probes when inferring splice isoform levels. In our experiments we found that ~5% of all probes were marked as outliers. In identifying outliers, the exon body probes proved to be critical. The redundant C_1 and C_2 probes allow the model to estimate the overall transcript levels contributed by the two mRNA isoforms monitored using the microarray and thereby detect outliers as discrepancies between these levels. In selected cases we found events for which the exclusion junction probe was marked as providing an outlying measurement, yet the overall percent exclusion was predicted accurately, by relying on the constant body probes. This effectively demonstrates the importance of the C_1 and C_2 for generating predictions in the presence of noise and outliers.

Another important feature of GenASAP is that it provide estimates of the quality of its predictions in the form of the tissue-specific ranks. The ranks provided by GenASAP allowed us to focus on the more accurate portion of the data, a crucial feature for further data analysis, such as inferring important global quantitative features of AS (Pan *et al*., 2004, 2005). Indeed, GenASAP outperforms common supervised methods, such as nearest neighbor, logistic regression and support vector machines, which were trained using the available RT–PCR measurements (see Supplementary information).

GenASAP represents the first and currently the only analysis tool for inferring quantitative information from AS microarrays. Previous microarray studies have attempted either qualitative detection of novel AS events (Clark *et al*., 2002; Johnson *et al*., 2003; Wang *et al*., 2003a; Cline *et al*., 2005; Le *et al*., 2004) or identification of increased or decreased AS levels in particular samples, as compared with a control sample (Blanchette *et al*., 2005; Ule *et al*., 2005). Fehlbaum and colleagues at ExonHit Therapeutics have succeeded in extracting quantitative estimates of AS levels, but their system requires the use of extensive control experiments employing titration of *in vitro* synthesized transcripts for their analysis (Fehlbaum *et al*., 2005). GenASAP, on the other hand, requires no information other than the microarray data generated using the AS microarray platform. GenASAP represents a powerful tool for the investigation of the global regulatory properties of AS in diverse biological contexts.

Further information on GenASAP and its implementation is available at http://www.psi.toronto.edu/GenASAP.

ACKNOWLEDGEMENTS

We would like to thank Qun Pan and Naveed Mohammad for designing the microarrays and Christine Misquitta and Arneet Saltzman for providing RT–PCR data.

This research was supported by an NSERC Discovery grant and CIHR Net grant to B.J.F, and a CIHR Operating grant and a Premier Research Excellence Award to B.J.B. Q.D.M. was supported by NSERC post-doctoral fellowship, and O.S. was supported by an NSERC graduate fellowship.

Conflict of Interest: none declared.

REFERENCES

Black,D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.

Blanchette,M. *et al*. (2005) Global analysis of positive and negative pre-mRNA splicing regulators in *Drosophila*. *Genes Dev*., **19**, 1306–1314.

Blencowe,B.J. (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases [Erratum (2000) Trends Biochem. Sci., **25**, 228.]. *Trends Biochem. Sci*., **25**, 106–110.

Cartegni,L. *et al*. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Gen*., **3**, 285–298.

Clark,T.A. *et al*. (2002) Genomewide analysis of mRNA processing in yeast using splicing specific microarrays. *Science*, **296**, 907–910.

Cline,A.S. *et al*. (2005) ANOSVA: a statistical method for detecting splice variation from expression data. *Bioinformatics*, **21** (Suppl. 1), i107–i115.

Cozman,F. and Krotkov,E. (1995) Truncated Gaussians as tolerance sets. In *Proceedings of the 5th Workshop on Artifcial Intelligence and Statistics*. Springer Verlag, NY, pp. 161–167.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc*., **B39**, 1–38.

Durbin,B.P. and Rocke,D.M. (2004) Variance-stabilizing transformations for two-color microarrays. *Bioinformatics*, **20**, 660–667.

Fehlbaum,P. *et al*. (2005) A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Res*, **33**, e4.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1), S96–S104.

Hui,L. *et al*. (2004) Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogne*, **23**, 3013–3023.

Johnson,J.M. *et al*. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.

Le,K. *et al*. (2004) Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res*., **32**, e180.

Matlin,A.J. *et al*. (2005) Unterstanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell. Biol*., **6**, 386–398.

Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.

Neal,R.M. and Hinton,G.E. (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan,M. (ed.), *Learning in Graphical Models*, MIT Press, Cambridge, pp. 355–368.

Pan,Q. *et al*. (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell*, **16**, 929–941.

Pan,Q. *et al*. (2005) Alternative splicing of conserved exons is frequently species specific in human and mouse. *Trends Genet*, **21**, 73–77.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems*. 2nd edn. Kaufmann, San Francisco, CA.

Relógio,A. *et al*. (2005) Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *J. Biol. Chem*., **280**, 4779–4784.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Compu. Biol*., **8**, 557–569.

Ule,J. *et al*. (2005) Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet*., **37**, 844–852.

Wang,H. *et al*. (2003a) Gene structure-based splice variant deconvolution using a microarray platform. *Bioinfomatics*, **19** (suppl. one), i315–i322.

Wang,Z. (2003b) Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res*., **63**, 655–657.

Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res*., **31**, 5635–5643.