Data analysis of alternative splicing microarrays

Miroslava Cuperlovic-Culf[1,3], Nabil Belacel[2], Adrian S. Culf[1,3] and Rodney J. Ouellette[1,3]

[1] Atlantic Cancer Research Institute, 35 Providence Street, Moncton, NB E1C 8X3, Canada
[2] National Research Council Canada, Institute for Information Technology-e-Health, 55 Crowley Farm Road, Suite 1100-C, Scientific Park Moncton, NB E1A 7R1, Canada
[3] Atlantic Microarray Facility, 35 Providence Street, Moncton, NB E1C 8X3, Canada

The importance of alternative splicing in drug and biomarker discovery is best understood through several example genes. For most genes, the identification, detection and particularly quantification of isoforms in different tissues and conditions remain to be carried out. As a result, the focus in drug and biomarker development is increasingly on high-throughput studies of alternative splicing. Initial strategies for the parallel analysis of alternative splicing by microarrays have been recently published. The design specificities and goals of alternative splicing microarrays, in terms of identification and quantification of multiple mRNAs from one gene, are promoting the development of novel methods of analysis.

Splicing of precursor mRNA (pre-mRNA) is a step in mRNA transcription involving removal of the introns and ligation of the flanking exons. Alternative pre-mRNA splicing (AS) leads to vast complexity in mRNA isoforms by generating several mRNAs from a single gene [1,2]. Thus, AS is widely assumed to be a key step in the creation of proteomic diversity in complex organisms [2,3]. There are examples of both conserved and species-specific splice forms [4,5], and over 60% of all human genes are alternatively spliced (see Refs [6,7] and references therein). Specific AS modalities are shown in Figure 1.

Variant transcripts generated through AS, similar to those initiated from distinct promoters, are often tissue- and/or development-specific, resulting in effects only in some cell types or during particular developmental stages. Whereas changes in promoter activity predominantly alter the expression levels of mRNA, however, AS can affect the sequence and the structure of the gene product by inserting or deleting sections of the pre-mRNA (Figure 1), possibly changing the reading frame of the mRNA.

AS-induced transcript changes fall into three categories: introduction of stop codons in the mRNA, changes in the resulting protein sequence and structure, and alterations in the 5′- and 3′-untranslated regions of the mRNA. The effects of these changes range from a complete loss of function of the protein to subtle effects, such as altered transcript localization, stability and translation [8–12]. Large-scale genomics studies suggest that AS has a key role in the production of functional complexity in the human genome and that at least 10–30% of alternatively spliced genes are tissue-specific [1,13,14]. Thus, it has been proposed that, apart from diversifying protein function, AS provides an additional layer of control in the development and function of healthy tissue [3]. This review deals with microarray methods developed for the high-throughput analysis of alternative splicing.

Significance of AS in drug and biomarker discovery

Alterations in AS have been linked to changes in cell physiology, developmental regulation and cancer [15]. The control of AS can be deregulated in human disease as a consequence of alterations in signaling cascades, in splicing regulators or in the spliced genes [15,16]. For example, at least 15% of the single base-pair mutations that cause human genetic disease arise from RNA splicing defects [17–19].

Understanding the diversity created by AS is essential for the discovery of both new biomarkers of disease and future therapeutic strategies.

Corresponding author: Cuperlovic-Culf, M. (miroslavac@health.nb.ca)

1359-6446/06/$ - see front matter. Crown Copyright © 2006 Published by Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2006.09.011

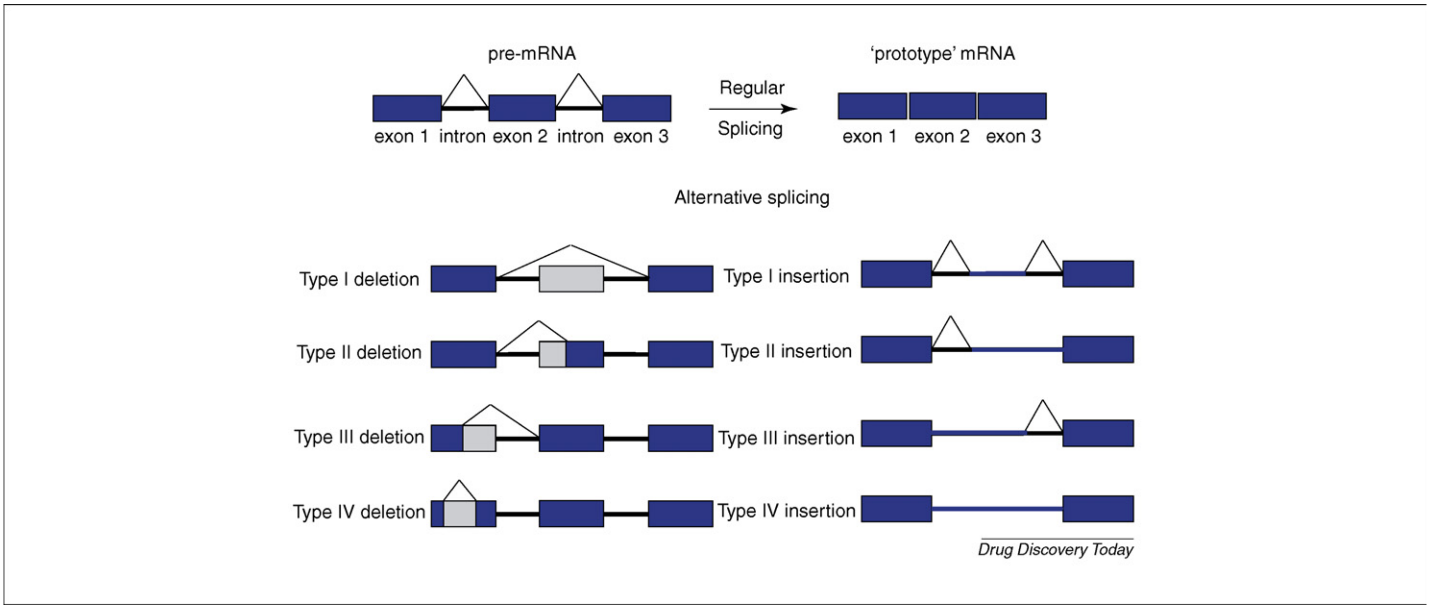


FIGURE 1

Alternative splicing modalities. The definition of alternative splicing modalities is based on the assumption that there is a 'prototype' mRNA that defines exons and introns in a gene. Introns can then be defined as segments of a gene that are removed before nuclear export of the prototype mRNA and thus do not contribute to protein synthesis. Exons are, in this context, defined as DNA sequences that contain information that is exported from the nucleus. Sections that are retained in mRNA are shown in blue. Deletions represent AS events where part of an exon (type II–IV) or complete exons (type I or cassette exon) are removed. Insertions represent AS modalities in which part of an intron (type I–III) or complete introns (type IV) are retained in the final mRNA product.

Examples of AS biomarkers

One of the first well-established examples of the effects of AS on gene function was provided by the BCL family of proteins, which are involved in the regulation of apoptosis. Two forms of Bcl-x are produced by means of an alternative donor splice site: the short form, Bcl-x_S, is pro-apoptotic, whereas the long form, Bcl-x_L, is anti-apoptotic. The Bcl-x_S:Bcl-x_L ratio determines cell viability, and subtle changes to this ratio alter cell fate [20,21].

Another example of phenotype-specific AS and possibly different function of mRNA isoforms is the gene *Pax5*, which is the principal regulator of B-cell development. Our group [22] has observed that individuals with lymphoma express only one splice form of the gene *Pax5*, whereas healthy donors express various isoforms. Although the relevance of this observation is not fully understood, it indicates that AS is altered during tumorigenesis of B cells.

The potential of mRNA isoforms in diagnosis has been demonstrated for Alzheimer's disease, where the ratio of acetylcholine esterase mRNA isoforms is used to predict the treatment outcome in affected individuals [23]. Many other isolated examples of disease-related AS are known [24–27]. Similar to the type of biomarker discovery that has been facilitated by high-throughput gene expression, however, the determination of optimal biomarkers based on AS will be possible only when high-throughput analysis of all AS forms across different tissues and disease types is available [28].

AS-aided drug discovery

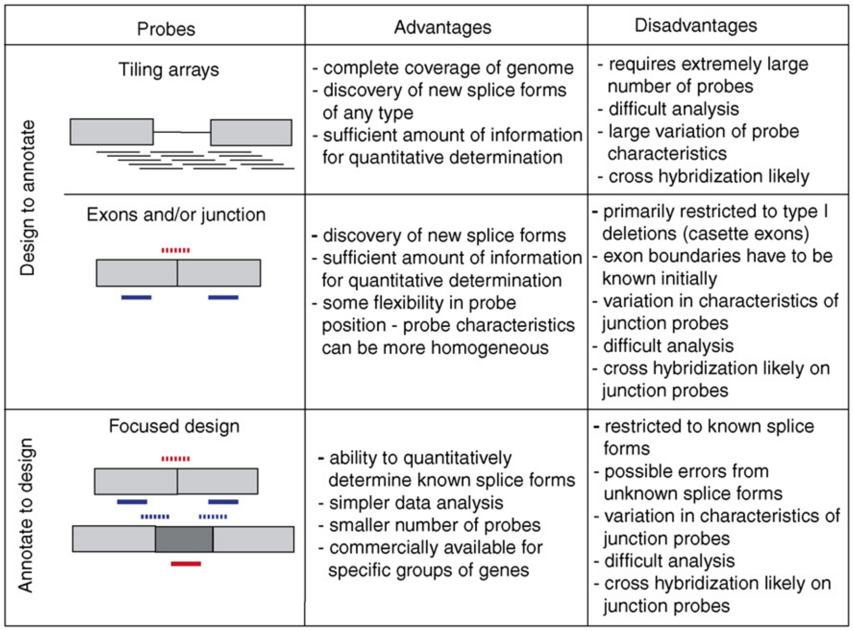
Understanding the diversity in mRNAs and proteins created by AS is also essential for future drug discovery efforts. Pharmacogenomics, defined as the determination and analysis of the genome (DNA) and its products (RNA and proteins) in relation to drug response, searches for more-focused drug targets while minimizing side-effects. AS provides another layer of opportunity for selecting drug targets. If AS is overlooked, drug discovery processes are exposed to only a fraction of the actual proteome and therefore miss potential protein or mRNA targets. In addition, AS can lead to altered drug metabolism because drugs can be bonded or digested by different isoforms of the targeted protein or mRNA [16].

Furthermore, there are many examples of protein isoforms with different and even opposite functions. Targeting the wrong isoform of the protein can have detrimental effects. Examples include the apoptosis genes *BCL2* [24] and *TP53* [25], where treatment of the wrong isoform can lead to increased proliferation of cancer cells rather than cell death. Another well-known example is the *VEGF* gene involved in angiogenesis [26]. Only the larger molecular mass isoforms, out of at least five mRNA isoforms of *VEGF*, are responsible for tumor angiogenesis and are thus an appropriate drug target [27]. Once more, high-throughput analysis of AS forms will probably result in numerous new drug targets and will give much more detailed information about the beneficial effects and side-effects of existing drugs.

Microarray-based AS methods and probe design

The highly parallel and sensitive nature of microarrays makes them ideal for monitoring gene expression on a tissue-specific, genome-wide level and provides the possibility of identification and quantification of isoforms. Several groups have shown that microarrays can be used to analyze pre-mRNA splicing [4,6,9, 10,29].

There are two main groups of microarray methods, resulting in different designs of probes that can be used to determine and/or quantify different AS forms: the first group includes 'annotate-to-design' methods [30–32], whereas the second group includes



	Probes	Advantages	Disadvantages
Design to annotate	Tiling arrays	- complete coverage of genome - discovery of new splice forms of any type - sufficient amount of information for quantitative determination	- requires extremely large number of probes - difficult analysis - large variation of probe characteristics - cross hybridization likely
	Exons and/or junction	- discovery of new splice forms - sufficient amount of information for quantitative determination - some flexibility in probe position - probe characteristics can be more homogeneous	- primarily restricted to type I deletions (casette exons) - exon boundaries have to be known initially - variation in characteristics of junction probes - difficult analysis - cross hybridization likely on junction probes
Annotate to design	Focused design	- ability to quantitatively determine known splice forms - simpler data analysis - smaller number of probes - commercially available for specific groups of genes	- restricted to known splice forms - possible errors from unknown splice forms - variation in characteristics of junction probes - difficult analysis - cross hybridization likely on junction probes

Drug Discovery Today

FIGURE 2

Methods used to design AS microarray probes. In the 'annotate-to-design' approach [30], splicing is exhaustively annotated and an array is designed to provide quantitative information on the expression levels of known RNA isoforms of interest [31,32]. In the 'design-to-annotate' approach, the goal is to discover new mRNA isoforms and new examples of alternative exons, in other words, to annotate gene structure and function with splicing-level resolution [33,6]. Blue lines indicate exon probes; red lines indicate junction probes. Tiling probes (black) extend across the whole genome and do not make any distinction between exon and intron probes.

'design-to-annotate' approaches [33,6]. In addition to these two types of design, there are three different microarray platforms: tiling, exon or junction, and focused (Figure 2). In principle, all three platforms facilitate the detection and quantification of mRNA splice forms.

The identification of new splice forms is possible with tiling and exon or junction arrays, but tiling arrays are the only method that can determine type IV exon deletions and intron insertions (Figure 1). In addition, the tiling method is independent of the pre-determination of gene exons. Thus, tiling arrays provide a very promising tool for the detection of splicing variants. The probe design for tiling arrays is the same for AS applications and for other applications such as ChIP–Chip analysis [34]. Tiling arrays have been used successfully to analyze the prevalence of exon skipping in chromosomes 21 and 22 (see Ref. [34] and references therein). Other than this set of experiments, however, the tiling method has not been widely used in AS owing to technical and analytical problems resulting from the extremely large number of probes.

Unlike tiling arrays, exon or junction arrays depend on knowledge of either isoforms (for quantitative analysis by focused design) or at least exon boundaries in the gene. Thus, the essential first step in the determination – namely the design of AS probes for exon and junction arrays – requires the characterization of gene exons (Figure 1). Several software tools are available to align an mRNA sequence with the genome sequence in order to determine exon boundaries (e.g. Sim4 [35]). It should be noted that these tools can determine only exons of known mRNA sequences; thus, if the true full-length mRNA sequence (Figure 1, prototype mRNA) has not been determined, exons will be missed.

The optimal probe length and position in terms of specificity and sensitivity present a difficult problem, particularly for the design of junction probes. Optimization experiments for junction and exon probes have been carried out by several groups, leading to general agreement about exon probes but differences over the optimal junction probe design [29,30,36]. The optimal probe length depends on the specific sequence of the gene; thus, it is unlikely that a particular length will be ideal for all genes and all junctions. Furthermore, the optimal probe length depends on the array platform and the RNA amplification method used. For this reason (i.e. variations in probe-binding energies), corrections for differences in probe specificity and sensitivity have to be included in the qualitative and/or quantitative analyses. Focused arrays require that gene isoforms are known prior to probe design, making these arrays the most accurate for quantitative analysis but unable to perform isoform discovery. The design of probes for focused arrays face the same problems as the design of junction probes and, thus, similar correction factors have to be included in their analysis.

Data analysis methods for AS microarrays

Perhaps the biggest challenge of AS array experiments is data analysis and biological interpretation. Experiments based on AS impose a complex 'one gene, multiple products' scheme [37], bringing challenges to the data analysis step with several issues that need to be resolved for qualitative and quantitative analysis. First, the

TABLE 1

Comparison of basic properties and features of AS microarray methods[a]

Method	Quantitative analysis possible	New splice form discovery possible	Optimal array design
Probe correction method	Yes	Yes	Any
ASAP	No	Yes	Any
Splicing index	No	Yes	Any
SPLICE	somewhat	Yes	Any
ANOSVA	Not accurately	No	Any
Sequence-based splice variant deconvolution	Yes (for two splice form system)	No	Any (but splice forms have to be known)
GenASAP	Yes	No	Focused

[a] The prime interest is on the possibility of using particular methods to identify new isoforms and to quantify isoforms.

analysis must be able to distinguish between changes in splicing and changes in overall gene expression. Deconvolution of differences in gene expression, AS and probe sensitivity has been a difficult problem in this respect. Second, the analysis must distinguish among different splice forms and ideally enable expression information (quantification) to be generated for different isoforms.

The methods outlined below present some of the tools that have been used on complete genomes. In this fast developing area and in the limited space available, it is not possible to give an exhaustive review of all of the tools, and some very interesting methods, such as those developed specifically for tiling arrays, are regrettably not included. The methods that we focus on deal specifically with the identification and/or quantification of mRNA (Table 1).

Standard gene analysis tools

All data analysis tools developed for standard gene microarray experiments can be used, in principle, in the analyses of AS microarrays. Both clustering and supervised analysis are possible for exon and junction data. But because each probe can represent a sum of intensities from several isoforms and because different probes represent expression of the same gene, there is the possibility of erroneous results. Furthermore, the information that can be extracted from clusters of exon and junction probes is hard to interpret: it suggests which gene probes are co-expressed but it does not provide any information about mRNA isoform sequences, quantities or coexpression.

The design of exon and particularly junction probes does not allow complete uniformity of probe properties across genes, leading to possible differences in probe sensitivities and errors in analysis. Nevertheless, AS is known to be significant in sample classifications, and the clustering analysis of cancer samples using limited-size AS microarrays shows excellent results and has been even proposed as a diagnostic tool [38–42].

Probe correction method

Johnson *et al.* [6] have used 36-mer DNA exon junction probes designed for all multi-exon human RefSeq mRNA sequences (~10 000 genes) to measure expression across 52 diverse tissues (available at GEO database, GSE740: http://www.ncbi.nlm.nih.gov/geo/). They introduced a correction factor that is dependent on the probe characteristics and independent of the hybridized tissue and thus can be determined from the median of probe intensities across all tissues.

Differences between modeled probe intensities (Box 1a) – determined from gene abundance (the average intensity of all probes for one gene) and probe binding affinity – and the actual intensities were used to generate prediction scores from 0 (no difference in probe expression relative to gene abundance) to 3 (highest difference in probe expression), which effectively rank the AS events measured by each probe. These scores were used in clustering of tissues by agglomerative hierarchical clustering with a cosine similarity measure (which measures pattern similarity rather than distance). Similar tissues were found to have similar patterns of mRNA isoform expression, and the resulting clusters were similar to clusters of samples obtained from gene expression levels. Prediction of differences in the expression of splice forms was confirmed in 45% of genes analyzed by PCR with reverse transcription.

This method, based on junction probes, has several shortcomings, however, including difficulties in distinguishing between two splicing events in the same or different transcripts, an inability to detect two equally expressed transcripts, and problems with cross-hybridization [6]. Several specific methods have been subsequently proposed, but some of the problems described by Johnson *et al.* [6] remain.

AS annotation project

The ‘alternative splicing annotation project’ (ASAP) method developed by Le *et al.* [43] is aimed at distinguishing between changes in AS and changes in gene expression and has been applied to obtain statistically significant evidence of tissue-specific AS from microarray data. Differences in the expression of mRNA isoforms in two samples is recognized from anticorrelation in the expression of two different probe sets relative to the sample pool. Using a pool of two samples as a reference for analysis of expression for each gene probe provides an intrinsic error correction for probe sensitivity and allows the straightforward observation of differences in AS of a gene as a negative correlation between the relative probe intensities in two samples.

Le *et al.* [43] also used the correlation coefficient, *r*, for hierarchical clustering of tissues. The clusters showed a clear division of AS samples in the five tissues analyzed. Clustering of gene probes using *r* yielded groups distinct from those produced by gene expression clustering of the same microarray data set. In addition, some genes that showed no tissue specificity by gene expression showed strongly tissue-specific AS. Similar results have been observed in other studies using different analysis methods.

BOX 1

Mathematical definitions of described methods for the analysis of alternative splicing microarray data.

(a) ASAP

$$P_{tj} = \theta_t \sum_f \omega_{tf} \varphi_{fj} + \nu_j + \varepsilon$$

where P_{tj} is the probe response for a specific probe j to a specific tissue sample t, θ_t is the gene expression level in tissue sample t, ω_{tf} is the fraction of the total gene transcript for the splice form f in tissue t, φ_{fj} is the sensitivity of probe j, and ν_j and ε are baseline and error terms.

(b) Splicing index, S_i (Figure I)

$$\text{single-label system}: \; S_i = \log_2 \frac{E}{G}$$

$$\text{dual label system}: \; S_i = \log_2 \frac{SE}{RE} - \log_2 \frac{SG}{RG}$$

where E is the individual exon signal, G is the gene signal, SE is the sample individual exon signal, SG is the sample gene signal, RE is the reference exon signal and RG is the reference gene signal.

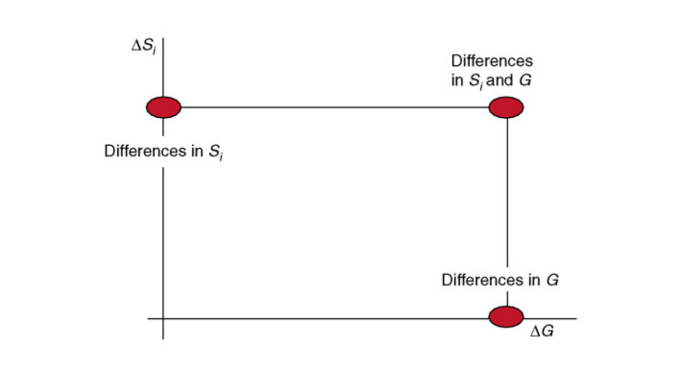


FIGURE I

Shows main application for the splicing index. Groups of genes schematically presented (as red ovals) show genes with changes in splicing (Si) and/or gene expression (Gi).

(c) SPLICE

$$RSS_{i,x} = D_{i,x} / \text{Avg}D_{i,x}$$

Where $\text{RSS}_{i,x}$ is the relative signal strength value of probe pair (PM, MM) i within probe set i in tissue x, $D_{i,x}$ is the PM-MM difference value of probe pair i in tissue x, PM is the GeneChip perfect match probe, MM is the GeneChip mismatch probe, and $\text{Avg}D_{i,x}$ is the trimmed mean PM-MM difference value of all probe pairs of probe set i in tissue x.

$$FR_{i,x} = \ln\left(\frac{RSS_{i,x}}{AvgRSS_{i,(n-x)}}\right)$$

where $\text{FR}_{i,x}$ is the final log ratio of probe i in tissue x and $\text{AvgRSS}_{i,(n-x)}$ is the average relative signal strength for probe i in all tissues other than x.

(d) ANOSVA

$$y_{ijkl} = +\alpha_i + \beta_j + \gamma_{ij} + \text{error}$$

where y_{ijkl} is the observed *log* intensity of probe k of probe set i, measured in experiment j of experiment set i, μ is the baseline intensity level for all probes in all experiments; α_i is the average probe affinity of probe i, β_j is the average target concentration for each experiment set, and γij is the interaction effect for each combination of the probe and concentration factors.

(e) GenASAP

$$x_{jtk} = r_{jt}\{s_{jt1}\lambda_{ik} + s_{jt2}\lambda_{2k} + [(1 - o_{jtk})\psi_k + o_{jtk}\phi]e_{jtk}\}$$

where x_{jtk} is the preprocessed expression level for probe k in gene j and tissue t, r_{jt} is the scaling factor, λ_{1k} and λ_{2k} are the probe predetermined profiles for isoforms 1 and 2, s_{jt1}, s_{jt2} is the isoform concentration, o_{jtk} indicates whether probe k is an outlier, and e_{jtk}, ψ_k and ϕ are error factors.

$$\%ASex = \frac{100 \cdot E[s_{jt1}]}{E[s_{jt1}] + E[s_{jt2}]}$$

where %ASex is the percentage of alternatively spliced exon exclusion, and $E[s_{jt1}]$ and $E[s_{jt2}]$ are the expected values of the isoforms levels.

Although the method of Le *et al.* [43] provides a very simple tool for investigating AS arrays, it has several drawbacks. It is only a discovery method for detecting possible AS and the results require validation by other methods. Cross-hybridization, among other effects, might cause changes in probe intensities, resulting in false-positive results. Thus, this method is considered to be a qualitative analysis tool.

Splicing index

When the difference in isoform expression between two samples is the only information required, the most straightforward approach is to use a splicing index [30,44]. The splicing index is calculated as the expression of an individual exon relative to the expression of the gene (Box 1b). The splicing index is most appropriate when analyzing quantitative changes for a specific known isoform. The same method can be used to determine and to quantify novel isoforms, although the presence of multiple isoforms with some overlapping regions will result in erroneous results.

The splicing index is usually presented as a graph of changes in the splicing index in two samples versus gene signal changes in the same samples (Box 1b). Thus, groups of genes that show changes in mRNA isoforms, changes in gene expression, or both, can be used for the initial estimate of changes in splicing events across different samples (Box 1b, Figure I). This method is, however, qualitative: it does not provide information about possible multiple mRNA splice forms. It cannot distinguish among multiple probe differences in one isoform and multiple isoforms. In addition, the splicing index method does not provide any sequence information about the different splice forms and thus cannot be used for a detailed description of novel splice forms.

SPLICE and NEIGHBORHOOD algorithms

One of the first efforts at using microarrays for AS analysis explored the possibility of using the standard multiprobe set up of Affymetrix GeneChip arrays (http://www.affymetrix.com) rather than special AS microarrays [45]. Each gene on a standard GeneChip is monitored by 20 pairs of 25-base oligonucleotides. The hypothesis of Hu *et al.* [45] was that large changes in gene expression detected in the same tissue with some probes of a gene relative to other probes of that gene indicate AS in the region of the gene where the altered probe expression is observed.

The SPLICE approach used by Hu *et al.* [45] is to identify groups of probes that cluster spatially in the genome with expression levels similar to each other but differing from the average expression level of the gene. The signal for a probe is calculated relative to the average probe signal for that gene (Box 1c, $RSS_{i,x}$). Final analysis is done with the probe signal for a given tissue relative to the average value of probe signals for all tissues (Box 1c, $FR_{i,x}$). The probes with an absolute FR value greater than a defined value are selected as candidates for AS location.

The accuracy of this analysis is improved with the NEIGHBORHOOD algorithm, which takes into consideration the expression levels of adjacent probes. If several adjacent probes survive selection by the algorithm, they potentially represent an extended region of AS. Although this work clearly shows that, in principle, AS can be detected by a standard Affymetrix microarray approach, many AS events have been missed primarily because the multiple probes were designed only for accurate detection of a gene and not for detection of mRNA isoforms. Additional errors in this study come from technical limitations such as 3′-labeling bias.

Analysis of splice variation

Analysis of splice variation (ANOSVA) uses a statistical testing principle to separate potential splice variation from gene expression data [46]. The ANOSVA model is based on two-way analysis of variance (ANOVA) [47], where the observed data are fitted to a linear model of two input quantities – probe set and experiment set factors – and can be used to determine the importance of each factor to the model (Box 1d). The likelihood of AS is determined from the interaction between the probe effect and the experiment effect. If the interaction term is significantly different from zero, then AS on that probe is statistically possible.

The advantage of ANOSVA is that it does not require transcript information and thus can be used when the level of annotation is poor. This method should be used only as a predictive tool, however, because factors other than splice variation, such as cross-hybridization, can result in significant interaction effects that are key to the method. In addition, there is a limit to the sensitivity of ANOSVA: failure to reject the hypothesis that there is AS means only that there is not enough evidence to prove that AS has occurred. Preliminary evaluation of the method has not yielded good performance for exon array data [46]; thus, the method should be used with caution.

Gene-sequence-based splice variant deconvolution

The idea underlying gene-sequence-based splice variant deconvolution is that the intensity of each AS microarray probe reflects the total concentration of all mRNAs containing sequence complementary to that probe [48]. Each probe is characterized by an affinity term. Assuming that gene sequence information for the splice form of the gene is available, feature concentrations and probe affinities can then be deconvoluted from the intensity measurements of multiple gene probes. The maximum likelihood estimation framework is finally used for the iterative determination of optimal values for probe affinities and feature concentrations [48]. Although this method is very accurate for the one-gene/two-isoform titration curve, it is less than perfect for the more complex three-isoform system. In addition, its prerequisite is to have complete information about the number and structure of all possible splice forms in the system, which, in most cases, is impossible to attain. Thus, the method is good for gene analysis in two-isoform systems, but should not be used for multi-isoform systems and is not even intended for the discovery of new splice forms.

GenASAP

The generative model for alternative splicing array platform (GenASAP) [3,49] (http://www.psi.utoronto.ca/~ofer/AS-supp.pdf) is a probabilistic model for the inference of AS levels from microarray data. It was developed to generate automatically the percentage contribution of each isoform to overall gene expression in addition to the confidence level of the value determined from microarray data. The intensity measured is modeled as a weighted sum of the overall abundance of the two isoforms. The calculations use machine learning and bayesian network approaches. Several features of the model make it very attractive. Bayesian estimation of parameters can account for the problem of a large number of unknowns in relation to the number of data points (samples). The model includes a possibility for expression-dependent noise. Finally, it accounts for aberrant observations with an outlier model. Although the model is usually presented for cassette exon AS events, it can be used for any type of AS and possibly multiple types with different probe selection as long as the focused probe design is used.

Pan *et al.* [3,9] used focused probe design with six probes for each excluded exon (Figures 1 and 2, three junctions, three exons) designed to distinguish among 3126 sequence-verified cassette-type AS events in ten different mouse tissues. The GenASAP algorithm [3] was used to determine tissue-specific rank for each of the two cassette exon isoforms. In this application, the preprocessed expression level for a gene probe for a given tissue is given in terms of an isoform profile, the number of isoforms, and several correction and scaling factors (Box 1e). Variables that are shared across all genes and tissues are inferred from the training set by using bayesian analysis. The variation in the expectation maximization algorithm is then used to determine gene- and tissue-specific values. After convergence, the distributions over the variables are used to compute the expected values of the isoform levels, which are then combined to estimate the percentage contribution of each isoform (Box 1f, %ASex). GenASAP performs better than various supervised methods for AS detection and is currently one of the most advanced methods for the analysis of focused AS arrays.

Conclusions

Even with our current limited knowledge of AS events, it is clear that diagnoses of cellular states and focused drug discovery require a study of AS. The gaps in our knowledge of the expression of mRNA isoforms are likely to be filled by high-throughput analysis of AS with AS microarrays. This application of DNA microarrays is still in development and there remain some unresolved issues:

- the large amount of data presents analytical and technical problems;
- splice junction arrays cannot determine whether two splicing events observed in one sample occur in the same or in distinct transcripts if the two isoforms are expressed at similar levels

(e.g. two-exon deletions could happen together in one splice form or individually in two splice forms);

- although novel isoforms can be determined, complete sequences of novel isoforms remain impossible to determine accurately, particularly when multiple mRNA isoforms of a gene are present;
- AS microarrays require an even more robust detection system than do regular arrays: problems with degradation and fluorescence quenching present a more serious issue in AS microarrays.

Despite these problems, the early experiments are promising and it is increasingly clear that AS microarrays provide a good approach to high-throughput analysis of AS. Future development will need to focus more on the development of AS microarray methods and data analysis tools that will be accessible to researchers interested in both cellular profiling using AS and analysis of novel AS forms. The current analytical methods, developed over the past few years, provide an excellent base for the future development of more quantitative and widely available tools.

Acknowledgements

MCC, ASC and RJO would like to acknowledge the generous support of the Atlantic Innovation Fund (ACOA) and the Dr. Georges L. Dumont Hospital Foundation. NB would like to acknowledge the support from National Research Council of Canada, e-health, Saint John.

References

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

2 Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* 72, 291–336

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

4 Pan, Q. *et al.* (2005) Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.* 277, 73–77

5 Modrek, B. and Lee, C.J. (2003) Alternative splicing in the human mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* 34, 177–180

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

7 Stamm, S. *et al.* (2005) Function of alternative splicing. *Gene* 344, 1–20

8 Lewis, B.P. *et al.* (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U. S. A.* 100, 189–192

9 Pan, Q. *et al.* (2006) Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.* 20, 153–158

10 Modrek, B. and Lee, C.J. (2002) A genomic view of alternative splicing. *Nat. Genet.* 30, 13–19

11 Ladd, A.N. and Cooper, T.A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.* 3 reviews0008

12 Davis, M.J. *et al.* (2006) Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. *Plos Genet.* 2, 0554–0563

13 Xu, Q. *et al.* (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.* 30, 3754–3766

14 Yeo, G. *et al.* (2004) Variation in alternative splicing across human tissues. *Genome Biol.* 5, R74

15 Faustino, N.A. and Cooper, T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev.* 17, 419–437

16 Bracco, L. and Kearsey, J. (2003) The relevance of alternative RNA splicing to pharmacogenomics. *Trends Biotechnol.* 21, 346–353

17 Krawczak, M. *et al.* (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* 90, 121–122

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

19 Levanon, E.Y. and Sorek, R. (2003) The importance of alternative splicing in the drug discovery process. *Drug Discov. Targets* 2, 109–114

20 Rohrbach, S. *et al.* (2005) Apoptosis-modulating interaction of the neuregulin/erbB pathway with antracyclines in regulating Bcl-xS and Bcl-xL in cardiomyocytes. *J. Mol. Cell. Cardiol.* 38, 485–493

21 Taylor, J.K. *et al.* (1999) Induction of endogenous Bcl-xS through the control of Bcl-x pre-mRNA splicing by antisense oligonucleotides. *Nat Biotechol* 17, 1097–1100

22 Robichaud, G.A. *et al.* (2004) Human Pax-5 C-terminal isoforms possess distinct transactivation properties and are differentially modulated in normal and malignant B cells. *J. Biol. Chem.* 279, 49956–49963

23 Darreh-Shori, T. *et al.* (2004) Long-lasting acetylcholinesterase splice variations in anticholinesterase-treated Alzheimer's disease patients. *J. Neurochem.* 88, 1102–1113

24 Akgul, C. *et al.* (2004) Alternative splicing of Bcl-2-related genes: functional consequences and potential therapeutic applications. *Cell. Mol. Life Sci.* 61, 89–99

25 Mills, A.A. (2005) p53: link to the past, bridge to the future. *Genes Dev.* 19, 2091–2099

26 Frelin, C. *et al.* (2000) Vascular endothelial growth factors and angiogenesis. *Ann. Endocrinol.* 61, 70–74

27 Zhang, L. *et al.* (2002) Vector-based RNAi, a novel tool for isoform-specific knock-down of VEGF and anti-angiogenesis gene therapy of cancer. *Biochem. Biophys. Res. Commun.* 303, 1169–1178

28 Brinkman, B.M.N. (2004) Splice variants as cancer biomarkers. *Clin. Biochem.* 37, 584–594

29 Castle, J. *et al.* (2003) Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol.* 4, R66

30 Srinivasan, K. *et al.* (2005) Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods* 37, 345–359

31 Davis, C.A. *et al.* (2000) Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res.* 28, 1700–1706

32 Spingola, M. *et al.* (1999) Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* 5, 221–234

33 Shoemaker, D.D. *et al.* (2001) Experimental annotation of the human genome using microarray technology. *Nature* 409, 922–927

34 Mockler, T.C. *et al.* (2005) Application of DNA tiling arrays for whole-genome analysis. *Genomics* 85, 1–15

35 Florea, L. *et al.* (1998) A computer program for aligning a cDNA sequence with the genomic DNA sequence. *Genome Res.* 8, 967–974

36 Fehlbaum, P. *et al.* (2005) A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Res.* 33, e47

37 Lee, C. and Roy, M. (2004) Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol.* 5, 231

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

39 Kan, Z. *et al.* (2005) Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles. *Nucleic Acids Res.* 33, 5659–5666

40 Okumura, M. *et al.* (2005) Candidates for tumor-specific alternative splicing. *Biochem. Biophys. Res. Commun.* 334, 23–29

41 Kirschbaum-Slager, N. *et al.* (2005) Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. *Physiol. Genomics* 21, 423–432

42 Kirschbaum-Slager, N. *et al.* (2004) Splicing factors are differentially expressed in tumors. *Genet. Mol. Res.* 3, 512–520

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

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

45 Hu, G.K. *et al.* (2001) Predicting splice variants from DNA chip expression data. *Genome Res.* 11, 1237–1245

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

47 Cuperlovic-Culf, M. *et al.* (2005) Determination of tumour marker genes from gene expression data. *Drug Discov. Today* 10, 429

48 Wang, H. *et al.* (2003) Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics* 19 (Suppl. 1), i315–i322

49 Shai, O. *et al.* (2006) Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics* 22, 606–613