Analysis of alternative splicing with microarrays

Jingyi Hui[1], Shivendra Kishore[1], Amit Khanna[2] and Stefan Stamm[1,2]

[1]University of Erlangen, Fahrstrasse 17, 91052 Erlangen and [2]Department of Molecular and Cellular Biochemistry, B283 Biomedical Biological Sciences Research Building, 741 South Limestone, University of Kentucky, College of Medicine, Lexington, KY 40536-0509
e-mail: Stefan@stamms-lab.net

Summary/Abstract

Alternative splicing is one of the most important post-transcriptional processing steps that enhances genomic information by generating multiple RNA isoforms from a single gene. Recently, microarrays have been developed that can detect changes in splice site selection. Currently, the biggest challenge for analysis of alternative splicing with microarrays is the bioinformatics analysis of array data and their low reproducibility by RT-PCR. Despite these problems, microarrays revealed an unexpected number of expressed RNAs, showed changes of alternative splicing in diseases and indicated that a splicing factor regulates a biological meaningful set of genes.

Key words: alternative splicing, microarrays, RNA isoforms, experimental validation, exon, intron

1. Alternative splicing and gene expression

Almost all protein-coding genes in higher eukaryotes have introns, which are removed during pre-mRNA processing by RNA splicing. It is estimated that in more than 88% of human genes, splicing can be alternative, i.e. the cell decides whether to remove part of the pre-mRNA as an intron or include this part in the mature mRNA as an alternative exon (1). Since most alternative exons encode protein parts, the alternative splicing mechanism allows the creation of multiple proteins from a single gene, which increases the coding potential of the genome. Alternative splicing generates protein isoforms with different biological properties, such as altered protein:protein interaction, subcellular localisation, or catalytic ability (2). Bioinformatic analyses estimated that a quarter of alternative exons introduce premature stop codons that either lead to generation of truncated isoforms or to the degradation of the mRNA in nonsense-mediated decay (3). However, recent array analyses show that these variants are generated at a low level (4).

The importance of alternative splicing is evident from the large number of diseases that are associated with or caused by the wrong selection of alternative exons (5) (6). It has been estimated that up to 50% of disease mutations localized in exons can change splicing patterns (7). Naturally occurring single nucleotide polymorphisms (SNP) result in different alternative splicing patterns within a population, which has consequences for human health. This is illustrated by CYP2D6, a member of the cytochrome P450 family of proteins. Due to an intronic SNP that changes splicing patterns, up to 10% of caucasians express non-functional CYP2D6 and cannot metabolize certain prescription drugs (8, 9). In addition, drugs like acetaminophen act on a specific isoform generated by alternative splicing (10), demonstrating the role of alternative splicing in drug action.

The functional differences between isoforms created by alternative splicing are often subtle (2, 7). The first reports of large-scale analyses in tumor tissues indicate that the combination of numerous subtle changes in isoform ratio contribute to the disease phenotpye (11) and not surprisingly, changes in alternative splicing are frequently found in cancer (12). Finally, low abundant splicing variants are specific for individual species, where they most likely contribute to a pool of isoforms that evolution can act on (13). This implicates that not all human-specific isoforms have been identified, an assumption that is supported by the identification of new mRNA isoforms in global array analyses (14).

The importance of alternative splicing for human gene regulation makes it necessary to create global high-throughput analysis tools. Isoform-sensitive arrays have now been developed and applied to analyze alternative splicing. We review the different systems and studies that have been performed, which reveals the current experimental limitations.

2. Detection of splicing variants with microarrays

2.1 Structure of splice variants

Alternative splicing pattern can be classified in five basic categories: cassette exons, alternative 5’ and 3’ splice sites, retained introns and mutually exclusive exons (Figure 1A). Combination of these events and more complicated splicing patters are possible (15). Cassette exons are the most frequent form of alternative splicing (7). In addition to alternative splicing, alternative mRNA isoforms can be generated by alternative promoter usage, alternative polyadenylation and RNA editing (16). Typically a gene generates several different mRNA isoforms with often severe differences in expression levels.

2.2 Systems for exon detection

cDNA arrays cannot discriminate between splicing variants and detect the mixture of isoforms. Therefore, specific array formats have been developed that discriminate splicing events. These high-throughput analyses of splice variants was performed in two basic systems: bead-based fiber-optic arrays and oligonucleotide arrays in slide format.

2.2.1 Fiber-optic arrays

Bead-based fiber-optic arrays are sold by Illumina. In this technique, arrays of beads are randomly assembled onto patterned optical fiber bundles. Each bead contains an oligonucleotide probe, which can detect a complementary probe. Only specified splicing variants can be detected using this system, which requires the ligation of oligonucleotides prior to detection. The system is named RASL for RNA-mediated annealing, selection and ligation. Typically, oligonucleotides will be designed that ligate across exon-boundaries to detect differences in isoforms. One of these oligonucleotides contains an index-sequence, which allows identification on the array. The oligonucleotides are annealed on the mRNA, which is captured on a solid phase using biotinylated oligo dT.

In a next step, these oligonucleotides are ligated and amplified by PCR. The PCR products are then hybridized to the array. In a subsequent step, the array is decoded, i.e. the beads binding to the index-sequence are identified by hybridization to colored beads carrying the known index-sequence. The major advantage of the system is its high sensitivity and reproducibility by RT-PCR, the drawback are the limited number of events that can be studied (11, 17). Since the detection relies on the amplification of short RNA parts, fractionated RNA resulting for example from storage in paraffin embedded sections can be used (11).

2.2.2 Glass arrays

The majority of currently used arrays to study alternative splicing use oligonucleotides that are attached to glass slides. These slides can be produced by ink-jet printing (Agilent, Exonhit) or by photolithography (Affymetrix). The major advantage of ink-jet printing is that it can be easily customized, since it does not require the generation of photolithograpical masks. The drawback is the smaller number of spots per array (currently around 150,000). The major advantage of arrays generated by photolithography is their high number of spots (currently around 5,500,000). A compromise between the two systems is maskless photolithography offered by Nimblegen, which creates custom arrays with about 300,000 spots.

2.2.3 Probe designs

Array designs differ in the nature of probesets. Probes can be arranged at even spacings in tiling arrays; they can target only exon bodies or exon junctions (Figure 1B). The use of exon junction probes (Figure 1B, dotted lines) allows the direct probing of exon:exon junctions, whereas tiling arrays indicate the relative change of exon expression. Custom-made designs often combine different types of probes. The scale of the probes is another feature that differentiate various designs. Probsets can be genome-wide, addressing either all exon-exon junctions (14) or all currently known exons (18). More recently, highly focused designs were used to study a smaller number of better-characterized genes (11). A large portion of currently known exons is derived from EST predictions which is reflected in a bias towards the 5’ and 3’ ends of genes in array designs based on these databases. This problem can be overcome by genome-wide tiling arrays. Their usage identified a large number of new transcripts that are not annotated in the current databases, which indicates that current database-generated genome-wide arrays will be incomplete (19). In all slide-based systems, the mRNA is transcribed into cDNA. During this reversed transcription step, fluorescent dyes are incorporated into the cDNA that is then hybridized to the array. The details of the labeling and detection procedure have been recently reviewed (20).

2.2.4 Array Designs using exon junction oligos

The detection of a splicing event with a combination of exon-junction and exon-body probes is illustrated in Figure 1B. Typically, several probes are made against exons (exon body probe) and a junction probe hybridizes half to the end of one exon and half to the beginning of the next exon. An increase of exon usage is indicated by a simultaneous increase of signal for its junction and body probes, whereas there is no change for the signal from the body probes from adjacent exons. In an ideal case, the signal from the junction oligonucleotide detecting the joining of both constitutive exons would decrease proportionally. However in our laboratory we have never observed such a perfect case. The combined signal from the body probes detecting the constitutive exons indicates the general transcipt level, which allows discrimination between alternative splicing and a simple change in transcript abundance. There is flexibility in the design of exon body probes that can be optimized for a similar hybridization temperature (21). However, there is almost no flexibility in the design of exon junction probes. One solution for this problem is the usage of several exon junction probes that are offset by 1-2 nucleotides, which allows their combined analysis (14). It is also difficult to detect small exons with oligonucleotides as the exons are too short to allow the design of body probes. Such short exons are frequently found as 3-nt long variations of alternative 3’ splice sites (22).

Commercially available Affymetrix designs do not contain junction probes. Here, a change in exon usage is detected by a change in splicing index, which is the logarithm of the ratio of the exon signal to the total signal from the gene ($\log_2$(exon/total)).

2.2.5 Array designs using tiling probes

Tiling arrays can cover the complete genome and contain 25-mer oligonucleotides that are currently spaced 35 bp apart, which defines the resolution of the arrays. Due to technical progress, this resolution will increase. For each oligonucleotide, a nucleotide with a mismatch serves as a control. One major advantage of tiling arrays is that they are unbiased, i.e. they do not rely on previous experiments collected in databases. This advantage became apparent when the use of tiling arrays spaced 5nt apart showed that more than half of human gene expression is not yet annotated (23).

3 Analysis tools

The most difficult part of an array experiment is the data analysis and verification. Currently, several algorithms are used without a single program emerging as a standard application, which highlights the difficulty in the analysis. Arrays detecting alternative splicing are more complicated than cDNA arrays, since they detect multiple products from one gene and have to discriminate between changes in splicing and changes in overall gene expression.

3.1 PLIER

The PLIER (Probe Logarithmic Error Intensity Estimate) method produces an improved signal by accounting for experimentally observed patterns in probe behavior and handling errors at low and high signal values. The PLIER algorithm was developed and released by Affymetrix in 2004. Many commercially available software packages that analyze microarray data are using PLIER (e.g Avandis (Strand Genomic) and ArrayAssist (Stratagene)). The PLIER algorithm produces an improved gene expression value that is a summary value for a probe set, which is done by incorporating experimental observations of feature behavior. PLIER uses a probe affinity parameter, which represents the strength of a signal produced at a specific concentration for a given probe. Calculation based on the data across the arrays defines the probe affinities and the error model employed by PLIER assumes the error is proportional to observed intensity, rather than to background-subtracted intensity. However, the derivation of the method also assumes that the error of the mismatch probe is the reciprocal of the error of the perfect match probe.

3.2 MIDAS

TIGR's Midas (Microarray Data Analysis System) is a Java based application that offers an interface to design microarray data analysis protocols combining one or more normalization and filtering steps. This assumes that the data from individual hybridizations is treated in a uniform and reproducible manner. MIDAS harbors the normalization modules that includes locally weighted linear regression (loess; (24) (25) and total intensity normalization. These can be linked with filters, including low-intensity cutoff, intensity-dependent Z-score cutoffs, and replicate consistency trimming, creating a highly customizable method for preparing expression data for subsequent comparison and analysis. Data analysis methods are constructed using a graphical scripting language and can be saved for application to other datasets. Scatterplots generated by the program illustrate the effects of each algorithm on the data.

3.3 ASPIRE

ASPIRE (Analysis of Splicing by Isoform Reciprocity) was designed to identify reciprocal splicing changes between two samples i.e. it is normalized to steady-state levels. This approach allows to identify changes in alternative splicing with high sensitivity and to discriminate them from changes in RNA stability. Data quantification is based on the change in the fraction of exon inclusion (26).

3.4 GEnASAP

GenASAP (Alternative Splicing Array Platform) predicts the level of alternate splicing for exon skipping events detectable on custom microarray chips. It uses Bayesian learning in an unsupervised probability model to accurately predict alternate splicing levels from the microarray data. It reads the hybridization profiles of microarray data, while modeling noise processes and missing or aberrant data. It has been applied to the global discovery and analysis of AS in mammalian cells and tissues (27).

Different models have been recently reviewed (28).

3.5 Commercial software

The algorithms have been implemented in several commercially available programs that are summarized below:

XRAY (Biotiquesystems)

The program can run in EXCEL spreadsheets and analyses gene expression and alternative splicing events. It uses a mixed model ANOVA algorithm to discriminate between changes in alternative splicing and gene expression.

Genomatix's ChipInspector (Genomatix)
carries out significance analysis on the single exon probe level. Exon probes with significant expression ratios are annotated based on the Genomatix proprietary genome annotation and analysis system ElDorado.

Partek Genomics Suite (Partek)
performs statistical analysis and allows visualization of the result. The program annotates all results and provides hyperlinks to internet databases of splicing.

Arrays from the Agilent platforms are distributed by Exonhit, which offers SpliceArray Analysis Tool (SAT). The SpliceArray Analysis Tool is an Excel application, that allows identification of changes in splicing by indicating expression values, fold changes, pValues and Pearson correlation.

4 Experiments with splice-site sensitive microarrays

Published experiments using microarrays that were designed to identify splicing variants are summarized in Table 1. In almost all array experiments performed, changes in alternative splicing were validated by RT-PCR. The validation rates range from 35% (18) to 100% (26). In the validation, only false positive events were detected. It is likely that the false negative detection rates are in the same range. Slightly higher validation rates were observed when real time-RT-PCR was used (29), which probably reflects the ability of real time PCR to detect changes over a larger range of RNA concentrations. The published data show that there is no significant difference between the available analysis programs in predicting splicing events that can be validated by RT-PCR. The highest validation rates are achieved when the result of the array experiment can be combined with other experimental data, such as binding signatures of regulatory factors (30).
The major outcomes of the current array analyses are the identification of functionally related targets of the splicing factor NOVA-1 (30), the finding that tissue-specific exons are flanked by highly conserved intronic parts (31) and the description of widespread changes of alternative splicing in human tumors (11, 32).