Statistical analysis of splicing microarray data

Murlidharan T Nair
Departments of Biology and Computer Science/Informatics
Indiana University South Bend,
1700 Mishawaka Ave,
South Bend, IN 46634-7111
574-520-5068
mnair@iusb.edu

Proposal submitted for Indiana University's Faculty Research Grant

March 3 2006

Alternative splicing involves a combinatorial editing wherein two or more exons are joined together. Alternative splicing is a major source of proteome diversity. Querying the uniquely edited forms (mRNA isoforms) is reflective of events that occur during transcription as well as those that occur during the processing of the transcript. This processing of the primary transcript which is responsible for removing the introns is carried out by the splicing machinery. The resulting processed transcript is either translated into a protein or may serve as a functional RNA transcript. Splicing depends on the proper recognition of exons and needs to carried-out with extreme fidelity. Mutations annotated on the human genome reveals that about 10% affect canonical splice site sequence. Thus being able to detect the levels of each isoform in the cell significantly increases our understanding of the state of the cell.

Conventional microarrys that have been in use for sometime now are capable of detecting only the primary transcript and do not convey the complete picture of the state of the transcript that a cell harbors as a result of splicing. Since each transcript that result from splicing is capable for being translated into a protein product, the proteome picture that conventional arrays portray is not accurate.

In our recent study we constructed splicing arrays to examine ~1500 mRNA isoforms from a set of genes that had been implicated in prostate cancer. The goal of the work was to identify signature mRNA isoforms that are characteristic of prostate cancer. In the study we also used the DASL assay (cDNA-mediated annealing, selection, extension and ligation) to circumvent the problem when dealing with partially degraded biological samples. The data so obtained was reflective of the levels of isoforms that were targeted for each gene

The objective here is to determine a subset of isoforms using the statistical method of multiple comparisons that uniquely differentiate themselves from others. Multiple comparisons permit a unique way to compare data and selectively pick only those that actually show a significant difference. The comparisons will be done in an iterative manner, each time building and assessing a model to explain the data. Simplest model with the minimum number of parameters required to explain the data will be taken as optimal.