Marcel Jentsch 02.11.2010

Alternative Splicing Detection Algorithms for Affymetrix Exon Array Data

1. Background

Generating a variety of transcripts from a single gene is common in many organisms. This process, called alternative splicing, can be found in all classes of eukaryotes. In lower eukaryotes only a small fraction of genes undergoes alternative splicing, while in higher eukaryotes nearly all genes are alternatively spliced[1]. The function of this process is to increase the protein diversity [2]. Alternative splicing coordinates physiologically meaningful changes in protein isoform expression and is a key mechanism to generate the complex proteome of multicellular organisms [3]. Protein isoforms can play different roles in the cell, because they differ in binding properties, stability or localization [3]. Different isoforms can be found in different tissues [4,5] or several types of cancer [6,7].

The microarray platform, GeneChip® Exon 1.0 ST, developed by Affymetrix, allows measuring the expression of each exon of a transcript [8]. These microarrays also allow to detect changes in the gene expression caused by the presence of different isoforms of the genes [20] and the measurement of genes that lack a ploy-A tail (e.g. some Histones [21]). This is achieved by more than 5.5 million oligonucleodide probes. This probes are distributed across the human exome, so that on average each exon is covered by four probes. The random primer technique used leads to evenly distributed probes and avoids the bias towards 3' ends, which was present in 3' arrays. Thus, exon arrays are an enhancement of the standard microarray platforms.

Computational analysis of exon microarray data includes data normalization, quality control measures and methods for alternative splicing detection. Many methods for standard 3' microarrays can be adapted to exon arrays. But this is not true for the detection of alternative splicing. The most outstanding difference between gene level and exon level analysis emerges in the detection of differential expression [9] on exon level. Only if the expression of an exon is set in relation with the expression of the corresponding gene, in two different experimental conditions, differential exon expression can be detected.

Many alternative splicing detection algorithms have been proposed [10, 11, 12, 13, 14, 15, 16, 17, 18]. Their aim is to detect different exon expression patterns of a gene within different conditions. Most of these methods are based on geometric measures, such as linear correlation, or analysis of variance.

1.1 Focus of this work

This work will concentrate on alternative splicing detection algorithms. We will perform a detailed comprehensive evaluation of alternative splicing prediction methods for Affymetrix exon arrays. We will do this to detect unknown and to analyze known problems, like inherent dependencies of the prediction on the variability of exon expression or on the number of exons per gene. We want to characterize which method is in which situation the appropriate one. Beside the characterization of the methods another goal of this work is the development of new, and hopefully better, alternative splicing detection algorithms.

The third aim of this work will be the question how all different results of these different methods can be aggregated into one meta-result. We will use an ensemble of the score statistics derived from the previously mentioned methods and non-parametric statistical methods for aggregating independent measurements, like Rank Product, to detect alternative splicing. We will also propose a workflow, which we will derive from the characterization of the different methods. The workflow will chose the appropriate method for each gene to analyze experimental data. We will do this to improve sensitivity and specificity. We will test this approach both on artificial and real data.

We will characterize and compare these methods by studying their behavior with respect to a number of simulated data sets exhibiting certain properties. The use of artificial data has the advantage that we can change many parameters (exons per gene, amount of replicates, ...) and directly observe the behavior of the different methods, the disadvantage is that artificial data only can mimic the real world. Thus, we will also test the methods on tissue data with literature confirmed events [16].

2. Methods for detection of alternative splicing

The advantage of Affymetrix exons arrays is that all exons of a gene are covered by probes. One main task in the analysis of exon array data is the detection of alternative splicing events. We will shortly introduce some methods and give a foresight on our ideas.

2.1 Proposed Methods

A variety of splicing prediction methods has been proposed for Affymetrix exon arrays. We selected ten approaches (*SI [13], SPLICE [14], PAC [10], MADS [10], MIDAS [10], FIRMA [11], Rank Product [15], ARH [16], ANOSVA [17]* and *Correlation [18])* for a detailed analysis. These 10 methods can be grouped due to the character of the prediction: scores (Splicing Index, SPLICE, PAC, FIRMA and Correlation), tests (ANOSVA, MiDAS and MADS), non-parametric statistics (Rank Product) and information theoretic concept (ARH). Some methods make prediction on the exon-level (SI, SPLICE, PAC, MADS, MiDAS, FIRMA and Rank Product) while the other methods make gene-level predictions (ARH and ANOSVA).

2.1 New approaches

Apart from those published methods, we also plan to explore three new ideas for solving the problem.

These are:

(1) The first approach is adapting clustering algorithms, like k-means, to exon array data. This is a geometric approach. We interpret the gene's exon expression values as a vector in a multidimensional space. We partition all replicated measurements of exons of a gene into two clusters. The better these partition fits the original partition into control and treatment, the more probable is it that this partitioning is due to an alternative splicing event.

(2) The second approach is to use the Kullback-Leibler-Divergence. Like ARH we will use the information theoretic concept to detect alternative splicing. The Kullback-Leibler-Divergence is defined to be $D_{KL}(P||Q) = sum(P(i)log(P(i)/Q(i)))$. We will compute the exon splicing probabilities P and Q with respect to treatment and control. The Kullback-Leibler-Divergence is a non-symmetric difference between two probability distributions. The larger $D_{KL}(P||Q)$ the more probable is a alternative splicing event. While we work on this approach we will also have a look at the potentials of cross entropie and perplexity.

(3) Third, the idea is to adapt the model of Purdom et al. [11], originally developed to generate artificial data. We will also use this model to generate artificial data. While we use the same model for generating artificial data and developing a new method, we have to find a alternative way to test this approach.

3. Identifying and using the potentials of the different methods

The task of the presented methods is identifying genes with different splicing in two experimental conditions, such as treatment and control. How can we use this methods to get the best solution?

3.1 Characterization of the different methods by using artificial data and tissue data with literature confirmed events

We will compare all the previously mentioned alternative splicing detection algorithms. Approaches for the comparison of the methods can be found in Purdom et al. [11], Beffa et al. [19] and Rasche et al. [16].

We will mainly concentrate on how the algorithms behave on artificial data. The most methods, for example, perform better if a gene contains only few exons.

Due to the fact that we will work mainly on artificial data, we can alternate many parameters. What influence has changing a parameter on the score statistics? How does the score statistic depend on the number of exons per gene, the number of replicates, the intensity of up-and down- splicing, position of the spliced exon within the gene, or the number of alternatively spliced exons per gene. This is only a short list of questions we can try to answer. In the end we will have a list of advantages and disadvantage of the different methods. This knowledge can be used to choose the appropriate method in the right situation.

In the last step we will compare the alternative splicing detection methods on real data, tissue data with literature confirmed events[16]. We will use this data as a gold standard.

3.2 Meta-Scoring and an alternative splicing analysis workflow

To improve the prediction, the idea is to use a set of different score statistics, from different methods, to build an ensemble. On this ensemble we will use non-parametric statistics, like Rank Product [15], to identify the genes that are most likely alternatively spliced. Our hypothesis is that such an ensemble will show improved sensitivity and specificity. Not all methods will be, by all means, part of the ensemble. We will maybe skip some alternatively splicing detection methods based on the results of the previously performed comparison with respect to previously done analysis of performance.

Alternatively, with the knowledge we will gather about the different alternative splicing detection methods, we like to build an intelligent workflow for alternative splicing detection. The idea is that for each gene in a experiment the appropriate methods are chosen. The choice can for example

depend on the number of exons of a gene. The aim is again to improve specificity and sensitivity.
Also a combination of both approaches would be a promising opportunity.
We will test this approach on artificial and real data.

References

[1] Lewin, B.(2002) Molekularbiologie der Gene. Spektrum, 711-740

[2] Maniatis, T. and Tasic, B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoas. Nature, 418, 236-243

[3] Stamm, S., Ben-Ari, S., Rafalska, I.,Tang, Z., Zhang, Z., Toiber, D., Thanaraj, T. A., Soreq, H.et al. (2005) Function of alternative splicing. Gene, 344, 1-20

[4] Pan, Q. et al. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-thoughput sequencing. Nature Genetics, 40, 1413-1415

[5] Wanger, E. T. et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature, 456, 470-476

[6] Brinkman, B. (2004) Splice variants as cancer biomarkers. Clin. Biochem., 37, 584-594

[7] Venables, J. (2004) Aberrant and alternative splicing in cancer. Cancer Res., 64, 7647-7654

[8] Gardina, P. J. et al. (2006) Alternative splicing and differential gene expression in colon cancer detected by a whole genom exon array. BMC Genomics, 7, 325-???

[9] Zimmermann, K. et al. (2010) Analysis of Affymetrix exon arrays. Technical Report.

[10] Exon Array Whitepaper Collection (2005) Alternative transcript analysis methods for exon arrays. Technical Report 1.1, Affymetrix, Inc.

[11] Purdom, E. et al. (2008) FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics, 24, 1707-1714

[12] Xing, Y. et al. (2006) Probe selection and expression index computation of Affymetrix Exon Arrrays. PloS ONE, 1, e88

[13] Srinivasan, K. et al. (2005) Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods, 37, 345-359

[14] Hu, G. K., et al. (2001) Predicting splice variant from DNA chip expression data. Genome

Res., 9, 1093-1105

[15] Breitling, R. et al. (2004) Rank Products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett., 573 (1-3), 83-92

[16] Rasche, A. et al. (2009) ARH: predicting splice variants from genome-wide data with modified entropy. Bioinformatics, 26, 84-90

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

[18] Shah, S. H. and Pallas, J. A. (2009) Identifying differential exon splicing using linear models and correlation coefficients. BMC Bioinformatics, 10, 26

[19] Beffa, C. D. et al. (2008) Dissecting an alternative splicing analysis workflow for for genechip exon 1.0 st Affymetrix arrays. BMC Genomics, 9, 571

[20] Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D., Jensen, R., Majewski, J. (2008) Gene expression and isoform variation analysis using affymetrix exon arrays. BMC Genomics, 9, 529

[21] Adesnik, M., Salditt, M., Thomas, W., Darnell, J.E. (1972) Evidence that all messenger RNA molecules (except histone messenger RNA) contain Poly (A) sequences and that the Poly (A) has a nuclear function. Journal of molecular biology, 71, 21

[22] Kullback, S., Leibler, R.A. (1951) On information and sufficiency. Annals of Mathematical Statistics, 22(1), 79-86