Poster L-30

Features of sequence composition and population genetical measures of selection to analyse alternatively spliced exons and introns

Authors:
Ivana Vukusic (*University of Cologne, Institute for Genetics*)
Sushma-Nagaraja Grellscheid (*Institute of Human Genetics, International Centre for Life, University of Newcastle upon Tyne*)
Thomas Wiehe (*University of Cologne, Institute for Genetics*)

Short Abstract: We have developed a binary classifier based on Genetic Programming (GP) to predict whether a given gene sequence is spliced constitutively or alternatively. The prediction accuracies are greater than 85% on the dataset of retained introns. Furthermore we showed that skipped exons show traces of positive selection.

Long Abstract:

Features of sequence composition and population genetical measures of selection to analyse alternatively spliced exons and introns

Alternative pre-mRNA splicing is a major source of mammalian transcriptome and proteome diversity. Aberrant splicing is an important cause for genetic diseases and cancer. Until a few years ago it was believed that almost 95% of all genes undergo constitutive splicing, which always proceeds in a removal of introns which is followed by a merge of exons. It is now widely believed that alternative splicing is the rule rather than the exception and that up to 74 % of all human genes are alternatively spliced. Whether an exon or an intron will be included or excluded in the transcripts of a gene of a certain cell type is influenced by the information contained in the sequence of the exon and the flanking intronic region. It is commonly accepted that no single factor dictates whether or not an exon will be spliced into a transcript. Instead it is probably a combinatorial effect of various factors that include cis-acting sequences and trans-acting splicing factors.
To predict whether a given gene sequence is spliced constitutively or alternatively we used the technique of Genetic Programming (GP). GP is a sub-discipline of Machine Learning. Basic ideas of GP are inspired by the paradigm of Darwinian evolution. New programs are “bred” from a population of existing programs and subject to selection, mutation and recombination. We used the GP system “Discipulus”, a supervised learning system, which generates programs on data that describe a certain problem. The features provided to this system are in form of a “feature-matrix”, containing e.g. nucleotide composition, length, motifs etc.
After each GP run Discipulus collects the information, of how often each feature was used in the thirty best programs, in a so-called “input-impact”-table. This table can be used to reveal the “best features” for a certain classification problem.
The system has been tested on extended version of the AltSplice data base. Here, we concentrated on cassette exons (SCE) and retained introns (SIR) and analysed 27,519 constitutively spliced exons and 9641 cassette exons including their upstream and downstream introns; in addition we focused on the analysis of the difference of 33,316

constitutively spliced introns compared to 2712 retained introns. The classifier shows very high prediction accuracy on the SIR data: sensitivity is 91.4% and specificity is 81.9%. In contrast, on the SCE data the prediction accuracy is lower: sensitivity is 48.2% and specificity is 70.3%. This suggests that sequence properties, such as those collected in the GP feature matrix, are better suited to detect alternative splicing of introns than that of exons. A possible biological reason is that the constraints imposed by the genetic code affect (coding) exons but not introns.

During cross-validation we have collected and analysed the five input-impact-tables resulting from each GP run. A frequency value of 5 of a certain feature means that in all 5 GP runs the 30 best programs contained this feature. The most frequently used features of the SCE data are: Number of A residues (frequency value: 5), GGG sequences (frequency value: 3,6) and the number of C residues (frequency value: 1). Although every single run starts with a new population of randomly generated programs, a similar pattern occured in all other runs performed during cross-validation. The best feature on the SIR data set is the number of A residues (frequency value: 4,1), followed by GC divided by length (frequency value: 1.8) and the number of T residues (frequency value: 1.4) in accordance with the fact that exonic splicing enhancer tend to be purine rich sequences.

To see whether selection, positive or negative, acts differently in alternatively than constitutively spliced exons we extracted for our lists of exons all annotated sequence polymorphisms from the latest release of the HapMap database. A common measure to test for the presence of positive selection is Tajima’s D. We find that Tajima’s D is smaller in the European population compared to Africa. Also, Tajima's D is smaller in the skipped compared to the constitutive exon dataset in both populations, indicating an elevated level of positive directional selection in alternatively spliced genes. Linkage disequilibrium is higher in derived populations and in alternatively spliced genes in all populations. We also find a slightly elevated level of genetic diversity close to the splice boundaries in alternative exons. However, while these features indicate a general trend, the sequence polymorphism data are too sparse in order to be used as a predictor of alternative versus constitutively spliced exons in particular cases.