Poster L-31

Large-scale analysis of the impact of alternative splicing on protein isoforms.

Authors:
Johannes Frey-Skött (*Stockholm Bioinformatics Center*)
Arne Elofsson (*Stockholm Bioinformatics Center*)

Short Abstract: Alternative splicing is the process where metazoan genomes expand their proteome. To study this phenomenon all genes in Ensembl that produce more than one transcript were extracted and the events occurring between the protein isoforms were identified. The events were studied for several features such as abundance and sequence identity.

Long Abstract:

In the post-genomic era, one surprising fact is that the size of the many metazoan genomes fell below of all expectations. The proteome size exceeded the number of protein coding genes by far and this difference is mainly due to the process of alternative splicing. During this process the exons may combine in alternative manners forming different transcripts thus creating the protein diversity observed in many higher species.

The phenomenon has been proposed to serve several functions. It has been observed that approximately one third of alternative transcripts creates candidates for nonsense-mediated decay (Lewis et al., 2003), thus regulation of splicing could determine which genes that are allowed to be translated into protein. Another suggestion is that the creation of a new isoform allow for evolution of new function without risking loss of the original and potentially crucial function. The process is then expected to reduce the loss of fitness while evolving new functionality (Modrek and Lee, 2002).

In general, alternative splicing may generate several different splice patterns, for instance exon skipping and mutually exclusive exons. However, all splicing patterns can be divided into two more simple forms; insertions/deletions (referred to as indels) and substitutions. Among indels a majority is thought to be deletions since an insertion of previously non-coding sequence may cause frame shifts and therefore give rise to inactive proteins. Still, indels are believed to generate negative regulators. In contrast, substitution events may produce two functional proteins more frequently, which could for instance be used to change the specificity of a substrate binding site in an enzyme (Kondrashov et al., 2002).

In order to study alternative splicing we strive to use a reliable data set and chose the Ensembl data to retrieve information on genes, their transcripts and included exons. We believe that their curation of data and strict inclusion criteria gives data that is highly reliable. From the peptide sequences we identify the nearest neighbor, which is defined by the pairwise alignment score from the program align in the Fasta package. By analyzing each isoform and its nearest neighbor, all events were extracted.

The features we have studied to determine the impact of alternative splicing on exon content,

fragment lengths a sequence similarity. Additionally, we have examined the frequency of indel and substitution events, their location in genes and their length distributions. Further, we show that most substituted exons show significant sequence similarity, indicating that the difference of the alternative functions is subtle. Although not all events we detect may be caused by alternative splicing but rather alternate promotors or alternate polyA sites, which we chose to not exclude in this study.

Our findings are that the events occurring in the internals of proteins are significantly shorter than those at the termini. Especially events occurring at the C-terminal are long probably because they experience lower selective pressure. At the N-terminal and in the internal rearrangements may cause larger disturbances, which lead to inactive or malfunctioning proteins.

We also find that substitutions are either short or have high sequence similarity. Among short substitutions (less than 50 amino acids) the sequence identity is fairly low while the identity increases with length. The length is not the only determining factor for how similar sequences are but also the location in the protein plays an important role. In the internal of the protein the sequence identity is higher and they are also kept short. This is probably due to the fact that they are involved in crucial interactions in the internal of the protein, thus creating stability. Altogether this indicates that substitutions are used to fine-tune function rather than creating new functionality.