Poster I-63

The Influence of Alternative Splice Sites in Protein Structure

Authors:
Elza H.A.B. Durham (*IME -USP/LICR*)
Richard Garrat (*IFSC-USP*)
Sandro J. de Souza (*LICR*)

Short Abstract: Alternative splicing of pre-mRNAs is one of the most important mechanisms that increase protein diversity in eukaryotes. Using information from expressed sequences and protein structures, we studied the structural impact of alternative splicing in proteins.

Long Abstract:

The fact that gene number is not significantly different between mammals and some invertebrates suggests that other mechanisms are being used to generate diversity, such as alternative splicing (AS) and post-translational modifications. Splicing is characterized by the removal of introns from the pre-mRNA. AS could be understood as a single gene originating different mRNA sequences which can occur by the use of alternative splice sites. The major types of AS are: intron retention (IR), alternative splice sites usage (AU), exon skipping (ES) and mutually exclusive exons.

It is know that some AS variants are tissue-specific and/or associated with several diseases in humans, as cancer. However, AS can create thousand of mRNA sequences and their functional viability has been questioned. Some studies indicate that variants with a frame shift and/or premature stop codons will be degraded. As some exceptions were found, the relationship between the amount of AS variants and their activity is still not clearly established. Some suggested that a high number of ESTs/mRNAs supporting a variant correlates with its functionality while others use the comparison between human and other organisms (mouse, rat) to exclude not functional sequences.

Many computational tools have been used to find and compare alternative splicing variants. Generally, cDNA, mRNA, ESTs and protein sequences that are public available are aligned against each other or against the genome to identify splicing isoforms. Most of this information is usually deposited in relational databases with open access. This can be used to join all sequence information related to variants as size, frame shift, insertions, deletions, repetitive elements and domains.

Some previous studies correlated the effect of alternative splicing in protein structures. Of them, some are focused on protein families while others do not cover all possible protein modifications caused by alternative splice sites. So, there still exists a lack of information about the protein structure modifications as a consequence of alternative splicing.

In this study we intend to identify and distinguish human protein structures modified by alternative splicing. In order to do it, mRNAs and EST sequences from UCSC were mapped to the human genome using BLAT and SIM4. All mapped sequences were deposited in a local database and the splicing boundaries from all sequences from a gene were compared to identify splicing variants. Those variants were assigned as IR, AU or ES events.

We constructed a pipeline where TBLASTN was performed between those variants (829.212 mRNA and EST sequences) and a set of 3.196 non-redundant PDB human sequences.

.Some BLAST parameters were carefully adjusted to allow gap opening and extension and identity was recalculated considering the gap size. Terminal regions without alignment were resubmitted to TBLASTN and the correct splice boundaries were assigned. Sequences with identity greater than 70% were included in our analysis, except for those containing stop codons.

Initially, the non-redundant PDB structures were related to 1.364 Unigene clusters allowing a directly association between the genes with alternative splicing sequences and their structural effects on proteins.

Events in proteins were separated in insertion and deletion, depending of the splicing sequence alignment. Proteins with deletions presented 7.427 donor and acceptor splice boundaries mapped into 1.662 structures (716 Unigene clusters) while insertion had 5.673 cases were related to 1.314 structures (585 Unigene clusters).

Other structural features were analyzed, as secondary structure (C, E and H) frequency in deletion and insertion boundaries and secondary structure patterns (combination of C, E and H) in deleted regions. Motility (measured through experimental B-factor values from PDB files) which is one feature expected to vary in determined regions of proteins was measured to deletion and insertion boundaries as to deleted regions. Spatial distance restraints between CA atoms which can be used by alternative splicing sequences to restraint the energy needed to fold a new protein, was measured in deleted regions and compare between the prototype and variant structures. Association between interaction regions (intra-protein and inter-protein) and diseases are also in course. We will discuss all of our results in the poster presentation.