Detection and prediction of premature stop codon using mass spectrometry data at the protein level

Karla Cristina Tabosa Machado[1], Andre Fonseca[1], Sandro Jose de Souza[1], Gustavo Antônio de Souza[1]

1 UFRN

Abstract

The volume of public data regarding proteomics has increased significantly in the last few years, allowing to use its potential to improve the annotation of genomics data. The key stage in proteomics is to identify peptides and proteins, initially through a comparison of collected mass spectrometry (MS) data and theoretical sequences in a database. However, protein sequence databases report mostly a no-redundant number of known isoforms, while individual polymorphic variations are not represented. Nonsense mutations are characterized by the premature appearance of the stop codon in a gene, which could produce defective proteins. The objective of this paper was to define a computational approach which could allow the prediction, of such mutations in proteomic datasets, without previous knowledge of the genome of the sample and consequently, only using a reference protein sequence database for protein identification. It was proposed that when nonsense mutations are present, one could track the quantitative profile of each peptide of a given protein, in order to detect a drop of sequence coverage at the protein c-terminal, that could be explained by the presence of a premature stop codon. This method was developed with the use of Perl language programming scripts which divide the identified proteins with sequence coverage above 30% into bins, in which each bin is 5% of the size of the subject protein. The script verifies peptides structures to ensure that they are totally or partially included inside of each bin and then adds all spectral counts associated to each peptide. Public MS data from three publications investigating the proteomes of cell lines, ovary cancer and colon cancer were re-analyzed using the approach developed here. Among the findings, genes such as BUB3, CALR and PRMT5 appeared mutated in certain samples of patients with cancer. This data will be validated at a later date to confirm the presence of the stop codon in the gene.

Funding: UFRN