Computational Identification of Orthologous Proteoforms between Human and Murine

Letícia Graziela Costa Santos de Mattos, Fabio Passetti

Laboratory of Gene Expression Regulation, Carlos Chagas Institute, Fundação Oswaldo Cruz (Fiocruz), Curitiba, PR, Brazil

Abstract

The advent of next-generation sequencers in Transcriptomics and mass spectrometry in Proteomics has resulted in a large volume of available data. These datasets became integrated into several Bioinformatics studies in an area called Proteogenomics. Thus, the fraction of messenger RNAs (mRNA) that are effectively translated into proteins has been deeply studied. Alternative splicing (AS) is a molecular event that may occur during mRNA maturation in more than 95% of human genes. AS might produce several mRNA isoforms that can change amino acids sequence, and consequently different proteoforms. In this context, the main goal of this project is to incorporate algorithms to our methodologies that allow us to identify orthologous proteoforms between humans and murine. For this purpose, we used transcriptome data from the Ensembl project to create a sequence repository using the ternary matrices methodology, which was developed by our research group. This customized sequence repository, created for human and murine datasets, was used to identify AS isoforms. In the human datasets, we were able to identify 22, 242 splicing variants in 61, 122 genes. According to the Ensembl Transcript Support Level (TSL) parameter, 41, 080 were ranked as reliable and 181, 382 with lower reliability. The gene that presented more AS variants was MAPK10, with 192 transcripts. Other human genes with known splice variants such as BCL2L1, KLF6, and TMP2 were also detected in our database. In the mouse transcriptome datasets, we found 126, 679 splicing variants in 43, 976 genes. From these variants, 43, 985 were classified as reliable and 82, 694 with lower reliability, based on TSL. Henceforth, for our next steps, we aim to identify expressed orthologous proteoforms using RNA-Seq data and proteomic shotgun data from human and murine healthy tissues. Additionally, we intend to select a list of proteoforms for experimental validation.

Funding: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; Conselho Nacional de Desenvolvimento Científico e Tecnológico