IDENTIFICATION OF PROTEIN DOMAINS BY BIOINFORMATICS TOOLS IN LARGE SCALE

Nicolay, R.B.S.\textsuperscript{1,2}; da Silva M.L.\textsuperscript{1}

\textsuperscript{1}Grupo de Biologia Computacional, Diretoria de Metrologia Aplicada às Ciências da Vida - Dimav, Instituto Nacional de Metrologia, Qualidade e Tecnologia - Inmetro, Duque de Caxias/RJ, Brazil.

\textsuperscript{2}Universidade Católica de Petrópolis – UCP, Petrópolis/RJ, Brazil.

The development of new methodologies in proteomics field are essential for treating biological sequence to identify biological information from omics raw data. Phylogenetic analysis methods can aid the prediction of proteins structures and their function by searching for sequence patterns. Our goals consists in demonstrating a new strategy for identification of protein domains by the rearrangement of different methodologies. This strategy concatenates different algorithms for treating biological sequence and then perform phylogenetic analysis to extract relevant biological information from recognized patterns. First, we performed a search using local alignment techniques with BLAST and HMMER in order to identify a specific biological data: a protein sequence (GI|189438693) with two domains (Glycosyl Hydrolase family 10 and Carbohydrate Esterase family 1) in a metagenome sequenced raw data from the snail \textit{Achatina fulica} (available at the MG-Rast). Second, a new search using local and global alignment methods, performed with BLAST and COBALT, was applied to identify annotated data in order to refine the results using the NCBI and CAZy databanks and for further phylogenetic analysis. In phylogeny, we used distance and maximum likelihood methods to corroborate our results and infer the evolutionary distance between our local results and the annotated data and the probability of those results be represented in the nature, performed with MEGA and PhyML. Lastly, we performed a local search with BLAST using the annotated sequence against the \textit{Achatina fulica} metagenome sample to identify identity between both biological data and aiming the identification of the whole protein sequence inside the sample. The results demonstrated a potential sequence with a high rate of identity characterized as an endo-1,4-beta-xilanase from \textit{Bacteroides reticulotermitis} corresponding to ours expectations. The methodology demonstrated here was applied using other biological data and with other biological databases presenting potentially and precise results for further experiments and data analysis.

\textbf{Keywords:} Bioinformatics, phylogeny, metagenomics.

\textbf{Acknowledgements:} Capes-BioComputacional, CNPq.