New annotation strategies: from sequence to 3D structure

Rafael Nicolay B. da Silva[1], Paulo José Miranda da Silva Iwakami Beltrão[1], Manuela Leal da Silva[1]

[1]*Instituto Nacional de Metrologia, Qualidade e Tecnologia - INMETRO*

The procedures applied to biological data annotation, nowadays, presents a computational network developed for analyze all kinds of data. Commonly, annotation algorithms perform comparisons between the raw data from a sample against renowned biological databanks. Annotation strategies are divided into sequential and structural, both strategies can be applied to increase the reliability for an analyzed dataset. This study will be presenting the identification of new molecular targets from sequential annotation of the feces from *Bradypus variegatus* metagenome sample. The sequential annotation strategy consists in the use of a known-function sequence, derived from renowned databanks, which is related to the function we want to identify within the sample. We performed local alignment techniques, in order to obtain contigs with high ratio of specificity. Further, we extracted the function-related sequences with short lengths to identify annotated sequences which presents a high percentage of identity, suggesting the presence of similar sequences inside the sample. At last, we performed a reverse search procedure using the annotated sequence, from previous step, against the biological sample. In this procedure, we extracted new fragments, besides the one employed to search for the annotated data, and performed a reference-based assembly and annotation of a new protein. As results, we identified a potential new enzyme characterized as a Glicosil Hydrolase family 8. The PSIPRED software was applied to predict secondary structures, the BLASTp suite was employed to perform local alignment techniques against the Protein Data Bank. The results revealed two potential enzymes, characterized as Cellulose synthase from *E. coli*, for templates with PDBIDs 3QXQ and 3QXF, both with 94% of coverage and 87%/85% of identity, respectively. Further, we applied MODELLER to perform the comparative modelling for generate 200 candidate models. The 3D structures generated from both templates were validated through different parameters. The best model presented values for Ramachandran's plot most favored and disallowed residue regions as 96.3% and 0%, 0.157Å for RMSD, -40830.668 for DOPEscore and 100% for GA341 score. The new strategy consists in the capability of automation for the whole annotation process, considering the reference-based assemble, the compilation of a new sequence and the creation of valid 3D models for further structural annotation, not described in the literature as an automated process yet. The next step consists in performing the structural annotation using ASAProt software (Automatized Structural Annotation of Proteins) and analyze the possibilities of experimental applications with the identified enzyme.

Financial support: CAPES and CNPq.