Metabolic pathway prediction of enzymes: a machine learning approach

Rodrigo de Oliveira Almeida[1], Guilherme Targino Valente[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas – Campus Muriaé – MG, [2]Universidade Estadual Paulista "Júlio de Mesquita Filho" - Faculdade de Ciências Agronômicas – Botucatu - SP

For long time, enzymes properties have been applied on several areas, since pharmacy to food industry. Nowadays, to establish an appropriate enzyme function is hard due to lots of biochemical procedures to be required. Bioinformatics is also focus to define protein functions by homologies or structural analysis; however those strategies are not applicable for most of proteins. Since the advance of sequencers along the last years, biological data generated are increasing fast and nowadays it is necessary more efficient tools to analyze this high amount of data. Thereby, machine learn is an interesting tool to help analyze those big data. The present study aims to construct models able to predict the metabolic pathway of enzymes based only on amino acid sequence properties. Protein sequences from four fungi (*Agaricus bisporus*, *Aureobasidium subglaciale*, *Saccharomyces cerevisiae* and *Talaromyces stipitatus*) were downloaded from Uniprot (http://www.uniprot.org/) and high similarity sequences (99%) were removed using the software CD-Hit. Data mining from protein annotations were performed to split in enzymes or non-enzyme proteins, which are the input dataset. For each metabolism (aminoacids, co-factors and vitamins, drug response, glycan, lipid and nucleotides) it was constructed a positive and a negative dataset. Around 1,200 protein attributes were generated using the R packages "Peptides" and "protr". Relevant attributes were selected using Weka software tools. After this process, all datasets were normalized, the positive dataset was undersampled and balanced datasets were done; after that each dataset was submitted to supervised training using Weka software to generate prediction models. It was used 6 classifier algorithms (J48, Random Forest, RepTree, Sequential Minimal Optimization, Voted Perceptron and Multilayer Perceptron) to generate models for each metabolism and final models were generated using the MetaVote or MetaVoteBagging. All scripts were written in R language and ran in parallel using a Shell script to improve the time of performance. The averages of correctly classified instances for training were 87.07, 90.63, 97.14, 94.71, 95.02 and 86.73% (metabolism of aminoacids, co-factors and vitamins, drug response, glycan, lipid and nucleotides, respectively). The final models were applied on 2,607 enzymes sequences with unknown metabolic pathway (from the same organisms) to classify them. Those models were able to assign metabolic pathways for most of unlabeled enzymes, which the results of prediction ≥0.7 of probability show a mean of 6.82, 9.90, 1.66, 18.15, 17.36 and 2.01% of enzymes classified in metabolism of aminoacids, co-factors and vitamins, drug response, glycan, lipid and nucleotides, respectively.

Funding Support: BIOEN FAPESP