Taxonomic classifier for β-glucosidase enzymes based on structural signatures

Letícia Xavier Silva Cantão, Luana Luiza Bastos, Leandro Liborio da Silva Matos, Marcos Augusto dos Santos, Raquel Melo Minardi

Universidade Federal de Minas Gerais - UFMG

Abstract

β-glucosidases are important enzymes for catalysis of hydrolysis of β-glucosidic bonds and can be applied in the field of biotechnology. They are involved in biofuel generation processes through the production of fermentable sugars. These enzymes can be found in many types of organisms such as bacteria, archaea, and eukaryotes. Structural signatures can be defined as patterns of distance between protein atoms. An example of a technique for calculating these distances is aCSM all software, where the calculation is performed for all atoms against all. The use of fresh new linear algebra techniques to construct logistic regression models estimates the probability associated with the occurrence of a given event in the face of a set of explanatory variables. The goal of this work is to construct a taxonomic classification model based on the structural signatures of β-glucosidases enzymes, considering the domain taxonomic level to which they belong, through the use of linear algebra and modified logistic regression. For the construction of the model, we used 162 PDB files of β-glucosidase enzymes from bacteria, archaea, and eukaryotes. From these PDBs files, we generated structural signatures using aCSM software, which resulted in a distance matrix with 7200 attributes and 162 entities. From this matrix, we build three datasets, one for each domain. To construct the classifier we divided each dataset into 80% for training and 20% for testing, applying linear algebra and modified logistic regression to calculate the probability of each entity belonging to its respective domain. We evaluated the constructed model through the (fa) - harmonic mean and the area under the ROC curve. The classifiers generated by the modified logistic regression models presented good results, being that the average harmonic average and the average area on the ROC curve for the archaea, bacteria, and eukaryotes enzyme classifiers were respectively: 0.88 / 0.89, 0.90 / 0.90, 0.89 / 0.88. The classifier proved to be efficient and promising for determining the taxonomic domain of β- glucosidase enzymes based on their structural signatures.

Funding: