A Pattern Recognition Approach for Catalytic Site Residues Prediction using STING Structural Protein Descriptors

Salim, J.A.\textsuperscript{1,2}; de Moraes F. R.\textsuperscript{1}, Neshich, I.A.P.\textsuperscript{1}; Mazoni, I.\textsuperscript{1}; Jardine, J.G.\textsuperscript{1}; Von Zuben, F.\textsuperscript{2}; Neshich, G\textsuperscript{1}.

\textsuperscript{1}Grupo de Pesquisa de Biologia Computacional, Embrapa Informática Agropecuária, SP; \textsuperscript{2}Departamento de Engenharia de Computação e Automação Industrial, Unicamp, SP, Brazil

The enzymes perform their biological role through some specific amino acids, known as catalytic residues. Thus, the function and classification of a particular enzyme could be obtained by selective description of their catalytic site residues (CSR) separating them from the rest of the protein amino acids. Based on information from Catalytic Site Atlas (CSA) which has literature evidence on CSRs, excluding those annotated by sequence homology, protein structural descriptors of CSRs from STING database were extracted. Most if not all methods for the prediction of catalytic residues use primary sequence conservation. As the protein function is determined by its structure, this research project aims to identify which structural descriptors, available in STING database, are of utmost importance to discriminate catalytic residues. Java Protein Dossier allows for “intuitive” selection of descriptors and corresponding intervals which could retrieve only CSRs. However, this procedure is not always efficient and development of an automatic approach to classify CSRs is reported in this work. Two different techniques were employed; One using Decision Trees which produces a human readable rules that enable the interpretation of results in terms of detailed description of biological environment of enzymes CSRs based on selected parameters and another one using Artificial Neural Networks, which despite the fact that it does not reveal legible rules, is more capable of dealing with classification tasks in nonlinear scenarios. Our goal is to obtain a set of structural descriptors and their intervals of values which will enable us to build a “periodic table” of catalytic site residues.

Keywords: Catalytic Site Residues, Enzymes, Pattern Recognition, Machine Learning, STING Structural Protein Descriptors
Supported by: FAPESP and CAPES