Prediction of catalytic residues in proteins using machine-learning techniques

Natalia V. Petrova, Cathy H. Wu

One of the major goals of proteomics is to assign a function to every protein. The knowledge of the protein function is a key to determining the role it plays in the cell. The number of proteins, whose functions have been experimentally characterized, is growing linearly every year. Experimental data provide reliable (in most cases) information about protein functional residues as well as possible mechanism of protein function. Furthermore, analytical methods used for experimental characterization of protein function involve many man-hours. It is true that it can be reduced by either improving the existing or, perhaps, by the development of new methods in experimental biology. But, since the sizes of the protein sequence and protein structure databases are growing exponentially, the gap between experimentally characterized and uncharacterized proteins is also growing exponentially [*GenBank database statistics, http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html; PDB database statistics, http://www.rcsb.org/pdb/holdings.html*]. As a result, two major groups of computational methods are progressively developing: homology transfer of known experimental data (1) and prediction of protein function using various properties of proteins and amino acids (2). Prediction of the functional residues is a challenging and interesting task. The results of such prediction could be successfully used in many research areas such as drug design, experimental biology, and protein database annotations.

In this poster we present a novel method for the prediction of the catalytic residues in proteins using machine learning approach. Since for the complex dataset it is almost impossible to know *a priory* which classification algorithm is going to perform better, our first goal was to determine one of the best performing algorithms among machine learning techniques. Second, different authors seem to focus on different features of the protein in order to predict catalytic residues. Some of the authors use conservation of protein family/superfamily to define functional sites; others use additional information, such as solvent accessible surface of the residues, residue type, secondary structure, and so forth. Therefore, we found relevant features of the protein residues for the prediction of catalytic residues using our benchmarking dataset of enzymes with known catalytic sites and machine learning attribute selection algorithm.

The method can *predict* catalytic residues and 3D location of the active site for proteins with *unknown function*, provided that *structure of the protein is known*.