EFFICIENT AND FAST PROTEIN FUNCTION PREDICTION

N. Pavlovic*[*], I. Ivanoska, S. Kalajdziski
Ss. Cyril and Methodious University, Faculty of Electrical Engineering and Information Technologies, Karpos 2 bb, 1000 Skopje, Macedonia
e-mail: natasa.pavlovik@feit.ukim.edu.mk
[*]Corresponding author

Keywords: *Gene Ontology, CLUS, Ray protein descriptor, Voxel protein descriptor, Predicting protein function*.

Motivation and Aim: Proteins are the most important parts of all living organism. They can be found in all living systems, starting from bacteria and viruses, to the human. They are responsible for the biological processes that happen in the cell. Therefore, the knowledge on protein function is of great significance. Mainly protein function is determined in laboratory, it is expensive and time consuming task. Our motivation was to apply data mining methods to speed up the protein function determination for newly discovered proteins. Once the protein structure is known, the potential function of each protein can be predicted by comparing their structures. Two similar structures are very likely to have similar functions. The computer can not recognize the diversity of different protein structures; therefore we will use protein descriptors for unifying the representation of the structure.

Methods and Algorithms: Hierarchical multi-label classification (HMC) is a classification where instances may belong to multiple, hierarchically organized classes at the same time. It often happens that one protein has more than one parent. Therefore, as a system for structural and hierarchical representation of proteins and gene products which supports the DAG hierarchy, Gene Ontology (GO) is used. Here, we use two protein descriptors: Voxel and Ray-based descriptor. Voxel protein descriptor transforms the protein's tertiary structure into N–dimensional feature vector, and additionally gives some other protein structural features. Ray-based protein descriptor transforms the protein backbone into M–dimensional feature vector. CLUS is a system for predicting protein function that created the model (decision tree) trained by using the protein descriptor datasets. As an algorithm for hierarchical multi-label classification, the CLUS-HMC algorithm is used.

Results: The datasets for both the Voxel and Ray-based protein descriptors were consisted of 24.502 proteins. 10 fold cross validation was used for training and then testing the model. The CLUS system calculates the error measures which appear in the hierarchy. A score often used for comparison is the area between PR curve and the recall axis (AUPRC). The close the AUPRC is to 1.0, the better the model is. When Voxel protein descriptor is used, average AUPRC (weighted)=0.90, while when Ray-based descriptor is used, average AUPRC (weighted)=0.83. By comparison of the scores gained for Voxel and Ray-based protein descriptors, a new direction for further predicting of protein function is given.

Conclusion: By comparison of these values (0.90 > 0.83), can be concluded that the PR curve gained from the Voxel protein descriptor dataset covers bigger area with the PR axis. That means that by using the Voxel descriptor in generating a dataset used for protein's function prediction, better and more accurate results are gained than using the Ray-based descriptor.

Availability: The software and data are available on request from the authors.