Effective Protein Function Prediction with Meaningful Explanation Using Reactive Motifs

Thanapat Kangkachit\textsuperscript{1,2,3} and Kitsana Waiyamai\textsuperscript{1} and Philippe Lenca\textsuperscript{2,3}

\textsuperscript{1}DAKDL-Lab, Department of Computer Engineering, Faculty of Engineering Kasetsart University, Bangkok, Thailand
\textsuperscript{2}Telecom Bretagne, UMR 3192 Lab-STICC, Brest, France
\textsuperscript{3}Université européenne de Bretagne, France
\{g5185041, fengknw\}@ku.ac.th, philippe.lenca@telecom-bretagne.eu

Keywords: Protein function prediction, reactive motifs, fuzzy concept lattice, biochemical knowledge.

Abstract: Classification of protein functions is a major task in health science. Unfortunately, it is expensive and time-consuming. In addition, there are nowadays a huge number of protein sequences available in bioinformatics databases but very few with function annotations. As a consequence, important effort is done to develop computational techniques to classify protein functions. There are now efficient algorithms but they do not provide meaningful background which is very important for biologists. Our work addresses both issues: effective classification and biological meaning explanation. It is based on reactive motifs discovery which are directly extracted from binding and catalytic sites where the chemical reaction mechanism occurs. They thus provide better biological meaning than statistical-based motifs. Unfortunately, very few protein sequences contain sites information. We thus propose a fuzzy concept lattice-based (FCL) method for discovering more general reactive motifs. FCL is constructed to represent biochemical knowledge. The general reactive motifs are then used as input features for supervised algorithms to build a classification model taking into account the high-dimensional and class-imbalanced problems that characterize protein datasets. Experimental results show that this approach outperforms classical ones in term of coverage and F-measure with SVM algorithm. In addition, it provides more meaningful knowledge to biologists.

REFERENCES


