Pattarika Na Pikul 2011: Semi-Supervised Learning for Protein Function Classification. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Kitsana Waiyamai, Ph.D. 52 pages.

Protein function is one of active bioinformatics research topics. To predict protein function with high accuracy, traditional classification techniques require large amount of labeled data for training. Unfortunately, labeled data is very hard to obtain, while unlabeled data is abundant.

In this research, we develop a semi-supervised learning technique for protein function classification. Pairwise alignment and Jaccard coefficient are used for composing data selection criteria, while Square Correlation between objects in a cluster and selected data in each training round is considered as stopping criterion for a self-training algorithm. Experimental results using UniProtKB/Swiss-Prot which contains 17,407 labeled and 47,619 unlabeled protein sequences show that our technique yields good performance on both training and test sets. Our proposed method generate classifier with more accuracy, precision, recall and F-measure than which is using only Jaccard coefficient for data selection and minimum mean square error as stopping criterion 0.3%, 1.00%, 2.83%, 1.91% respectively. In addition, the overlapping between prediction results on unknown genes also demonstrates the effectiveness of the developed technique.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature