Protein Function Prediction Using Motifs, Sequence Features, Alignment Scores and Classifier Combination

Zehra Cataltepe[1], Ugur Ayan[2], Eser Aygun[3]

Keywords:function prediction, motifs, Prints, Prosite, PROFEAT, alignment scores

1 Introduction.

Protein function prediction [1] is one of the most important problems in bioinformatics. Advances in solutions of this problem could lead to better understanding of how some diseases occur and how to prevent or treat them on a person to person basis.

Gene Ontology (GO) [2] is one of the common ways of defining protein function. There are three different main GO classes: Molecular Function, Biological Process and Cellular Component. Since Molecular Function is one of the most used classes in the literature, in this study, we concentrate on it.

Protein function determination is usually done based on the sequences in each class. Machine Learning algorithms need features extracted out of each sequence in order to train a classifier. Feature extraction can be done based on the a) physiochemical properties [3], b) sequence alignment scores between a sequence and the training data or c) the fingerprints/motifs (for example as in PRINTS [4] and PROSITE [5] databases) existing in each sequence.

In this study, our aim is to compare success of function prediction when sequence physiochemical properties, alignment scores or motifs are used.

2 Data

Sequences: We used Gene Ontology [2] first level Molecular Function classes of sequences of E. Coli and H. Pylori organisms. We only concentrated on the Catalytic Activity (GO:0003824) and Binding (GO:0005488) classes. For E. Coli there were 298 sequences in Binding and 827 sequences in Catalytic Activity. For H. Pylori, there were 195 Binding and 280 Catalytic Activity sequences.

Features: We computed three different sets of features for each organism. a) The PROFEAT extracted features. b) ClustalW alignment scores of each sequence with all the remaining sequences for the organism c) The PRINTS and PROSITE motifs that exist in each of the sequences. We represented the PRINTS and PROSITE motif information in vector space representation as in document categorization. Each sequence is treated like a document and each motif is treated like a word in a document.

The dimensionality of each input vector for PROFEAT, PRINTS and PROSITE features were 1447, 1901 and 2023 respectively. The dimensionality of input vector when ClustalW aligments were used was equal to the number of data points for the training set.

[1]Istanbul Technical University, Computer Engineering Department, Istanbul, Turkey. E-mail: cataltepe@itu.edu.tr

[2]Kultur University and Istanbul Technical University, Computer Engineering Departments, E-mail: ugur.ayan@gmail.com

[3]Istanbul Technical University, Computer Engineering Department, Istanbul, Turkey. E-mail: eser.aygun@gmail.com

Classification Algorithms: We experimented with a number of different algorithms, Naive Bayes, Support Vector Classifier and kNN. Since kNN gave the best results it is included in the results. We used kNN with k=3, 5 and adaptive and show them as 3NN, 5NN and Adptv NN respectively.

3 Results

The table below shows the 10-fold cross validation accuracies. ClustalW, PROFEAT, PROSITE and PRINTS results are shown under ClsW, PROF, PROS and PRI columns respectively. The error bars on each accuracy reported is about 1 %. Results are not very promising. A classifier that assigns every sequence to the most occurring class would get accuracies of 74 % for E. Coli sequences and 59 % for H. Pylori sequences respectively. Only for H. Pylori, using ClustalW alignment performed better than the classifier that would assign every sequence to the majority class.

We also experimented with different dimensionality reduction techniques, tfidf-like representation as in [6], and also classifier combination using majority voting, but were not able to get better results than shown below.

	E. Coli				H. Pylori			
	ClsW	PROF	PROS	PRI	ClsW	PROF	PROS	PRI
3NN	69	69	67	73	62	60	58	58
5NN	70	69	68	73	62	60	60	58
Adptv NN	74	74	73	69	63	55	57	59

Acknowledgements All authors were supported by The Scientific and Technological Research Council of Turkey (TUBITAK) EEEAG research grant no 105E164.

References

[1] Whisstock, J.C. and Lesk, A.M. 2003, Prediction of protein function from protein sequence and structure, *Q Rev Biophys*, **36**, 307-340.

[2] Ashburner, M., Ball, C.A., Blake, J.A., et. al. 2000, Gene ontology: tool for the unification of biology, *Nat Genet.*, **25:1**, 25-29.

[3] Li, Z. R., Lin, H. H., Han, L. Y. *et al.* 2006, PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, *Nucleic Acids Research*, **34**, W32-W37.

[4] Attwood, T.K., Mitchell, A., Gaulton, A., et. al. 2006, The PRINTS protein fingerprint database: functional and evolutionary applications, *Encyclopaedia of Genetics, Genomics, Proteomics and Bioinformatics*, John Wiley & Sons.

[5] Hulo N., Bairoch A., Bulliard V., et. al. 2006, The PROSITE database, *Nucleic Acids Res.* 34:D227-D230.

[6] Riad, M.R. 2000, *Representative based protein sequence clustering*, M.S. Thesis, Simon Fraser University, Canada.