High-throughput Protein Functional Prediction by Data Science Approach

Yi-Wei Liu, Wen-Hung Liao and Jia-Ming Chang
Department of Computer Science
National Chengchi University, Taiwan
{104753013, whliao, jmchang}@cs.nccu.edu.tw

Abstract

Biological information has grown explosively with the accomplishment of Human Genome Project and Next-generation sequencing. Annotating protein function with wet lab experiment is time-consuming, so many proteins' functions are still unknown. Fortunately, computational function prediction can help wet lab formulate biological hypotheses and prioritize experiments. Gene Ontology (GO) is the framework for unifying the representation of gene function and classifying these functions into three domains namely, Biological Process Ontology, Cellular Component Ontology, and Molecular Function Ontology. Each domain is a hierarchical tree composed of labels known as GO terms. Protein function prediction can be considered as a multiple label classification problem, i.e., given a protein sequence, predict its GO terms. We proposed a protein function prediction framework based on its homology sequence structure, which is believed to contain protein family information and designed various voting mechanisms to resolve the multiple label prediction problem.

1 Introduction

1.1 Background

Proteins are biological macromolecules which are composed of one or more long chains of amino acids. They play a central role in our cell, tissues, organs, and bodies as they perform vital functions within organisms and take up more than 50% of the dry weight of cells. Examples of important functions of protein include catalytic activity, regulation of metabolism, muscle contraction, structural support, antibacterial and antiviral defense, transporting molecules, and storage. If we know what function a protein carried, we can understand life at the molecular level as well as the molecular mechanisms of disease. But the speed of annotating protein function from wet lab experiments is too slow compared with the growth of protein sequence data as illustrated in Fig. 1. Fortunately, computational function prediction can help wet lab formulate biological hypotheses and prioritize experiments.

Gene Ontology (GO) is the main framework for unifying the representation of gene function. It is a large project initiated by Gene Ontology Consortium in 1998. The Gene Ontology project is composed of the annotations, which represent the function terms, and Gene Ontology, which models biological aspects in a structured way.

Gene Ontology classify functions into three different domains, namely, Biological Process Ontology (BPO), Cellular Component Ontology (CCO), and Molecular Function Ontology (MFO). BPO describes the biological process where gene product participates in, for example, transporting oxygen. CCO describes the location of the gene product, for example, inner membrane, periplasmic, or extracellular space. MFO defines what the gene product can do or its ability. Terms are linked to each other with a hierarchical structure. The relationships are graphed as directed edges and the terms are graphed as nodes. Fig. 2 depicts an example of BPO.

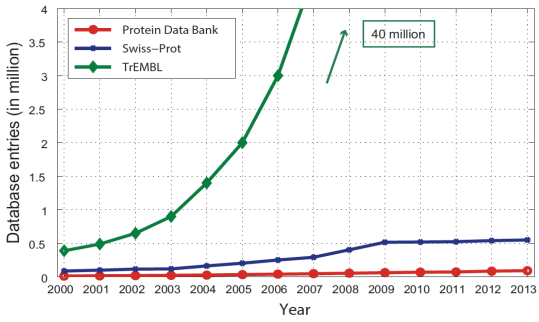

Figure 1: The growth of protein database, adopted from [2][p. 11]

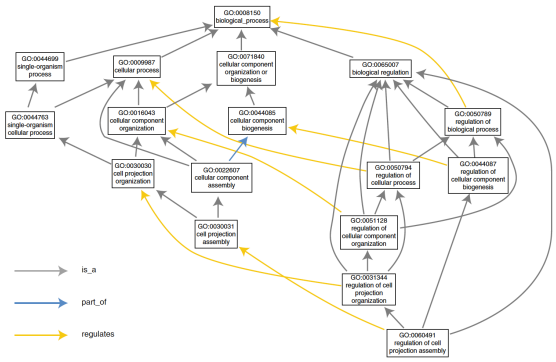

Figure 2: Example of Gene Ontology, adopted from [1] [p. 26]

1.2 Challenges

In this research, we focus on protein function prediction problem. We want to utilize information from protein sequence to predict its functions. This problem can be considered as a multiple label classification problem, because we can treat these gene ontology terms as labels and information from protein as input data. However, it differs from traditional multiple label classification problem in that these labels are hierarchical. Fig. 3 gives an illustration of the protein function prediction problem.

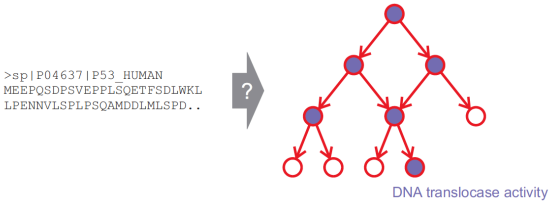

Figure 3: Protein Function Prediction Problem, adopted from [2][p. 8]

1.3 Out contributions

We proposed a new protein function prediction framework extended from PSLDoc [3] and PSLDoc2 [4], which have been employed to predict protein subcellular localization with superior performance. We designed three voting mechanisms to resolve the multiple label predict problem. One of them incorporates global features from homologous protein and local features from protein family. The performance of these voting strategies has been evaluated using benchmark database. All methods demonstrated better results than the baseline models.

2 Materials and Methods

We propose a framework composed of three stages to predict Gene Ontology with protein sequence information. Firstly, we extract gapped-dipeptides from PSIBLAST position-specific scoring matrix result. Secondly, Principal Component Analysis (PCA) is employed to reduce gapped-dipeptides features to lower dimension. Finally, we use different weighted voting strategies and other pertinent information to predict Gene Ontology annotations. The details of feature representation, feature reduction, CATH information, Gene Ontology prediction, data sets, evaluation measures, and baseline models are described in the following sections.

2.1 Feature representation by gapped-dipeptides

We apply the feature representation scheme from PSLDoc and use the same efficient homology extension approach adopted by PSLDoc2. Each protein is represented by a feature vector based on gapped-dipeptides and position-specific scoring matrix TFPSSM weighting scheme, called TFPSSM vector. TFPSSM vector is generated by the evolutionary information from PSSM, which is obtained from PSI-BLAST. Fig. 4 shows an example of TFPSSM vector. The default parameter settings with PSI-BLAST on PSLDoc will take a considerable amount of time. Instead, we use the fast parameter settings with PSIBLAST in this research.

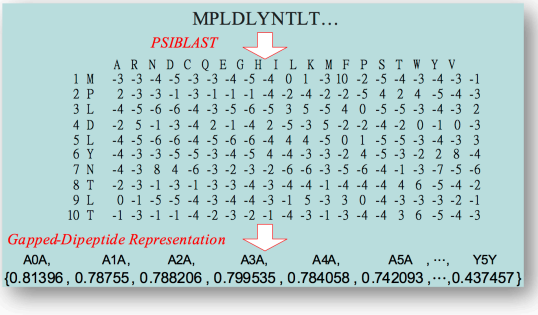

Figure 4: TFPSSM, adopted from [3][p. 6]

2.2 Feature reduction by Principal Component Analysis

Both PSLDoc and PSLDoc2 perform feature reduction on TFPSSM. PSLDoc utilized probabilistic latent semantic analysis (PLSA) [5] and PSLDoc2 employed correspondence analysis (CA), which is a generalized principal component analysis. In this study, we use PCA to reduce TFPSSM's feature dimensions, which explain 95% of the variance.

2.3 CATH information

CATH-Gene3D [6] is a database that stores protein information of known proteins sequences based on CATH [7] protein structure classification schemes. CATH-Gene3D data are clustered into functional families, and those proteins having the same functional family (FunFam) will possess similar sequences, structures, and functions.

We adopt the HMMer model of FunFam released on CATH Gene3D web server to predict the FunFam of query protein and protein in training data. In our experiment, we set best domain E -value threshold as 10^{-5}. That is, when HMM scan reports a FunFam best domain E-value smaller than 10^{-5}, we will consider this protein as one of this FunFam.

2.4 Gene Ontology prediction by weighted vote and nearest neighbor algorithm

We designed three approaches to utilize TFPSSM vector and nearest-neighbor algorithm to predict protein function.

2.4.1 TFPSSM-1NN

The first method is TFPSSM with one nearest neighbor. We will find query protein's nearest neighbor in the training data. Distance metric is the Euclidean distance. Then, the query protein will be predicted as the same GO terms of its nearest neighbor.

Because we want all proteins to be predicted and simplify the method, the confidence score of prediction is simply set to one.

2.4.2 TFPSSM-25%

The second method is TFPSSM with K nearest neighbors and weighted voting based on Euclidean distance. The number K is dynamically selected based on the third quartile of 1-nearest neighbor distance in the training data. After selecting K nearest neighbors, we use the inverse of distance as the weight to vote GO terms. In the end, these voting results will be normalized to the range 0 to 1.

2.4.3 TFPSSM-CATH

The third method is TFPSSM with K nearest neighbors and weighted voting based on CATH FunFams intersection amount. K is chosen in the same way described in TFPSSM-25%, but a different voting system is employed. With CATH FunFam HMMer model, we can obtain information regarding each proteins' FunFams, so we use the intersection amount of query protein and K nearest neighbor proteins as voting weight. Similarly, these voting results will be normalized to 0 to 1.

2.5 System architecture

In this study, we propose a protein function prediction framework based on homology extension and protein family. Fig. 5 illustrates our current system architecture.

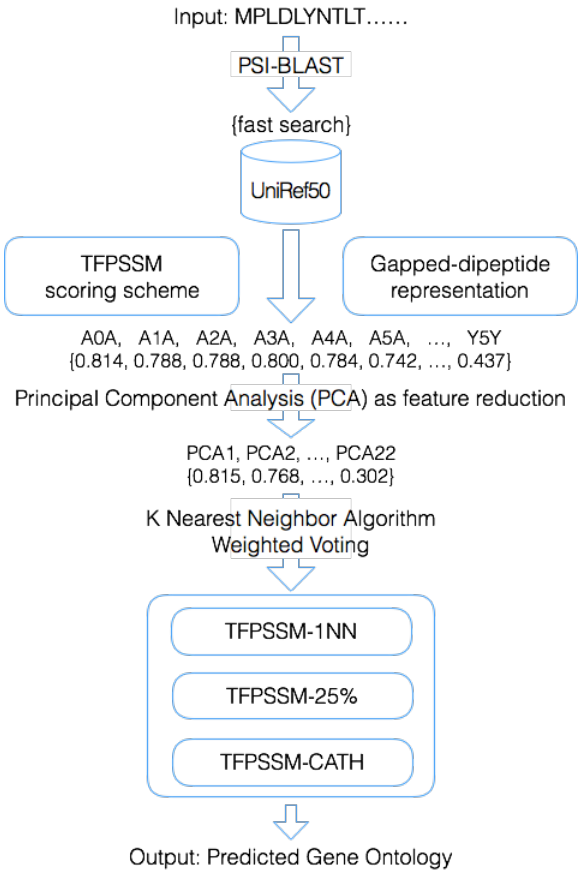

Figure 5: System architecture of proposed method for prediction of protein functions.

2.6 Data sets

We use data from the second and the third Critical Assessment of Functional Annotation (CAFA) in our experiments. CAFA is established by Function Special Interest Group (Function-SIG), whose aim is to advance the protein function prediction research by comparing all published computational methods in an unbiased evaluation. At the end of CAFA2, FunctionSIG released the training data, the testing data, and the evaluation metrics of each method that have been evaluated. We use the CAFA2 training data to train our model and predict the CAFA2 benchmark dataset, in order to compare our framework with other methods. Additionally, we include the training data from CAFA3. Because the CAFA3 is still in the evaluation phase, there are no ground truth labels of the testing data. Still, we can use CAFA3 training data to verify the robustness of our proposed methods using cross-validation. Table 1 gives a short summary of each dataset, including the number of protein

Table 1: Statistics of the Datasets. The number of protein sequence, number of GO term, and median annotation number of each protein in BPO, CCO, and MFO of each dataset.

Dataset	Type	# of seq.	# of GO term	# of median annotated GO terms
CAFA2 Training Dataset	BPO	40,728	15,838	25
	CCO	40,571	1,892	9
	MFO	26,056	5,480	8
CAFA3 Training Dataset	BPO	50,813	19,682	29
	CCO	49,328	2,426	10
	MFO	35,086	6,366	8
CAFA2 Testing Dataset	BPO	860	6,540	29
	CCO	1,259	833	11
	MFO	421	1,501	8

sequence, the number of GO term, and the median annotation number of each protein in BPO, CCO, and MFO.

2.7 Evaluation measures

There are two major evaluation types of protein function prediction, namely, protein-centric, and term-centric [9]. In this research, we focus on protein-centric evaluation. The prediction result for each term will have a predicted score between 0 and 1, which is considered as a confidence score. Precision (pr), recall (rc), and the resulting F_{max} are defined as follows:

$$\mathrm{pr}(\tau) = \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \frac{\sum_f \mathbb{1}\left(f \in P_i(\tau) \wedge f \in T_i\right)}{\sum_f \mathbb{1}\left(f \in P_i(\tau)\right)},$$

$$\mathrm{rc}(\tau) = \frac{1}{n_e} \sum_{i=1}^{n_e} \frac{\sum_f \mathbb{1}\left(f \in P_i(\tau) \wedge f \in T_i\right)}{\sum_f \mathbb{1}\left(f \in T_i\right)},$$

$$F_{\max} = \max_{\tau} \left\{ \frac{2 \cdot \mathrm{pr}(\tau) \cdot \mathrm{rc}(\tau)}{\mathrm{pr}(\tau) + \mathrm{rc}(\tau)} \right\},$$

where $P_i(\tau)$ is the set of terms that have predicted scores greater than or equal to τ for a protein i, T_i is the experimentally determined set of terms for that protein, $\mathbb{1}(\cdot)$ is the indicator function, $m(\tau)$ is the number of protein with at least one predicted score greater than or equal to τ, and N is the protein number of the entire testing set.

2.8 Baseline models

Two baseline models, namely, Naïve and BLAST are selected for comparative performance evaluation in our experiments. Sources for these two baseline models are adopted from the Matlab evaluation codes for the second CAFA experiment [8].

The Naïve method predicts terms based on the frequency in the training data, and the normalized frequency will be the score of the predicted term. As a result, in Naïve method, each query protein will be predicted to the same result. The BLAST method predicts terms based on Basic Local Alignment Search Tool (BLAST) searching result against the training data. BLAST will first return the high local alignment identity proteins of the query protein, it will then predict term based on these hits proteins, and convert e value to score.

3 Results and discussion

Table 2 summarizes our performance on training dataset (five-fold validation) and CAFA2 testing dataset using the three approaches we discussed earlier, respectively. The threshold of TFPSSM-25% and TFPSSM-CATH to dynamically selecting K nearest neighbors is the third quartile of 1-nearest neighbor distance from non-redundant training data protein.

3.1 CAFA2 and CAFA3 training dataset five-fold validation

TFPSSM-1NN demonstrated superior performance than the two baseline models on both CAFA2 and CAFA3 datasets, a clear indication that TFPSSM vector representation is effective in addressing protein function prediction problem. Both TFPSSM-25% and TFPSSM-CATH (derived from TFPSSM-1NN) achieve better performance than TFPSSM-1NN.

3.2 CAFA2 testing dataset

The performance on CAFA2 testing dataset is worse than those of CAFA2 and CAFA3 training dataset in five-fold validation. In addition, our performance is worse than that of the Naïve method in CCO category. The paper detailing CAFA2 results [9] provides some explanations for this observation. One reason is that the GO terms annotated on CAFA2 testing protein are more general than terms annotated on training data protein. The phenomenon is more prominent the CCO domain, thereby giving Naïve method more competitive advantages.

Table 2: The performance of CAFA3 training, CAFA2 training and CAFA2 testing.

Type	Method	Fmax (%)	Precision (%)	Recall (%)	Tau
BPO	BLAST	35.8 / 34.9 / 25.2	32.2 / 30.6 / 21.0	40.5 / 40.9 / 31.4	0.52 / 0.51 / 0.49
	Naïve	29.4 / 29.8 / 28.5	28.9 / 28.2 / 27.7	29.9 / 31.7 / 29.3	0.16 / 0.16 / 0.19
	TFPSSM-1NN	39.6 / 38.3 / 27.6	39.3 / 38.0 / 26.0	40.0 / 38.6 / 29.5	1 / 1 / 1
	TFPSSM-25%	41.8 / 40.5 / 30.4	42.1 / 41.0 / 31.1	41.6 / 40.0 / 29.7	0.31 / 0.31 / 0.3
	TFPSSM-CATH	**42.5** / **41.1** / **30.7**	43.6 / 41.9 / 28.2	41.4 / 40.3 / 33.6	0.38 / 0.36 / 0.29
CCO	BLAST	51.1 / 53.2 / 34.7	46.1 / 47.9 / 28.5	57.3 / 59.8 / 44.3	0.47 / 0.46 / 0.44
	Naïve	57.8 / 60.0 / **45.0**	61.8 / 67.6 / 39.1	54.3 / 54.1 / 53.2	0.33 / 0.45 / 0.31
	TFPSSM-1NN	62.8 / 65.4 / 41.4	62.6 / 65.4 / 38.8	63.0 / 65.5 / 44.3	1 / 1 / 1
	TFPSSM-25%	64.9 / 67.0 / 42.7	65.6 / 67.3 / 39.5	64.2 / 66.6 / 46.6	0.39 / 0.36 / 0.33
	TFPSSM-CATH	**65.0** / **67.3** / 42.7	65.0 / 66.8 / 39.5	65.0 / 67.8 / 46.4	0.41 / 0.37 / 0.34
MFO	BLAST	49.3 / 50.8 / 45.2	46.6 / 50.0 / 48.2	52.3 / 51.8 / 42.5	0.49 / 0.49 / 0.47
	Naïve	27.7 / 25.0 / 32.3	27.8 / 24.5 / 36.3	27.7 / 25.6 / 29.0	0.15 / 0.14 / 0.17
	TFPSSM-1NN	53.5 / 54.6 / 39.8	53.2 / 54.6 / 40.3	53.7 / 54.7 / 39.2	1 / 1 / 1
	TFPSSM-25%	53.5 / 54.2 / 43.3	55.3 / 55.8 / 48.6	51.9 / 52.7 / 39.0	0.36 / 0.35 / 0.34
	TFPSSM-CATH	**56.7** / **57.9** / **47.1**	56.4 / 58.1 / 47.4	56.9 / 57.7 / 46.8	0.4 / 0.43 / 0.39

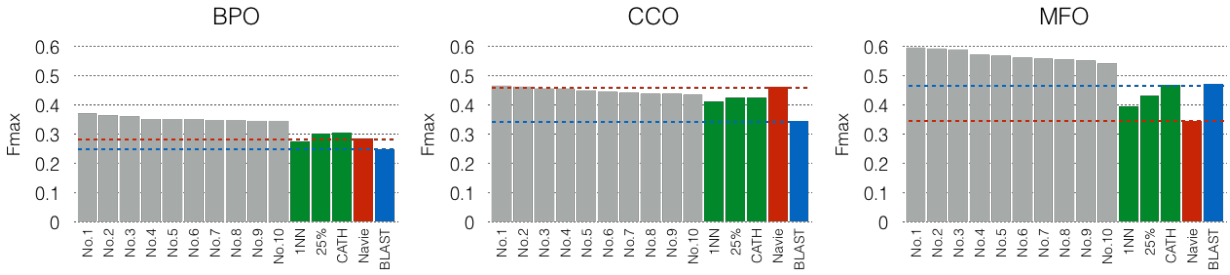

Figure 6: Performance compared with top-ten and baseline models on CAFA2 testing dataset

4 Conclusions

The framework we proposed exhibited superior performance compared with baseline models: BLAST and Naïve, in CAFA2 and CAFA3 training data with fivefold validation. However, there is still a lot of room for improvement if we compared our results with leading methodologies. Fig. 6 contrasts the performance of top-10 entries, baseline models, and our method. Even so, our framework still is competitive in the CCO category.

The third method (TFPSSM-CATH) combined protein family information and sequence structure, so it should demonstrate better performance than others. On average, the performance of TFPSSM-CATH is better than the other two voting schemes, but the difference is marginal.

In the future, we will design better voting strategies for TFPSSM-CATH and incorporate other machine learning techniques to achieve better performance.

Acknowledgments

This research was supported by the Ministry of Science and Technology (MOST 105-2218-E-004-003). We would also like to thank three anonymous reviewers their comments.

References

[1] Christophe Dessimoz and Nives Škunca. *The Gene Ontology Handbook*. Springer, 2016.

[2] Predrag Radivojac. A (not so) quick introduction to protein function prediction. 2013.

[3] Jia-Ming Chang, Emily Chia-Yu Su, Allan Lo, Hua-Sheng Chiu, Ting-Yi Sung, and Wen-Lian Hsu. Psldoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. *Proteins: Structure, Function, and Bioinformatics*, 72(2):693–710, 2008.

[4] Jia-Ming Chang, Jean-Francois Taly, Ionas Erb, Ting-Yi Sung, Wen-Lian Hsu, Chuan Yi Tang, Cedric Notredame, and Emily Chia-Yu Su. Efficient and interpretable prediction of protein functional classes by correspondence analysis and compact set relations. *PloS one*, 8(10):e75542, 2013.

[5] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.

[6] Ian Sillitoe, Tony E Lewis, Alison Cuff, Sayoni Das, Paul Ashford, Natalie L Dawson, Nicholas Furnham, Roman A Laskowski, David Lee, Jonathan G Lees, et al. Cath: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research*, 43(D1):D376–D381, 2015.

[7] Christine A Orengo, AD Michie, S Jones, David T Jones, MB Swindells, and Janet M Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8): 1093–1109, 1997.

[8] Yuxiang Jiang. CAFA2: Matlab Evaluation codes for the 2nd CAFA experiment, GitHub repository, https://github.com/yuxjiang/CAFA2, 2016.

[9] Yuxiang Jiang, Tal Ronnen Oron, Wyatt T Clark, Asma R Bankapur, Daniel D'Andrea, Rosalba Lepore, Christopher S Funk, Indika Kahanda, Karin M Verspoor, Asa Ben-Hur, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1):184, 2016.