Protein Function Prediction by Integrating Different Data Sources

Liang Lan, Nemanja Djuric, Yuhong Guo, Slobodan Vucetic*
Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA
*To whom correspondence should be addressed: vucetic@ist.temple.edu

1. INTRODUCTION

Protein function annotation is a key challenge in the post-genomic era. Experimental determination of protein functions is accurate, but time-consuming and resource-intensive. With the advent of high-throughput technologies, a variety of large-scale datasets are becoming available that can be very useful for protein function prediction. As participants in the 2011 Critical Assessment of Function Annotation (CAFA) challenge, we used data mining techniques to predict Gene Ontology (GO) functions of human proteins by integrating three different biological data sources containing information about protein sequence, gene expression, and protein-protein interactions.

Sequence similarity has been considered as the most useful metric for structural or functional annotation, following the widely accepted hypothesis that proteins with similar sequences have similar structure and function. In this work we explored to what extent the similarity in gene expression and protein-protein interactions also implies functional similarity. Furthermore, we explored whether combining the similarity metrics from multiple sources can increase the accuracy of functional annotation.

2. METHODOLOGY

To calculate the likelihood that protein p has function f, we used a weighted variant of k-nearest neighbor (k-NN) algorithm, as used previously [1]. The prediction score of a function f for protein p is calculated as

$$score(p, f) = \sum_{p' \in N_k(p)} sim(p, p') \cdot I(f \in functions(p')), \quad (1)$$

where, $sim(p,p')$ denotes the similarity score between proteins p and p', I is an indicator function that returns 1 if p' is experimentally annotated with f and 0 otherwise, and $N_k(p)$ are the k nearest neighbors of p according to metric sim. We used this scoring algorithm in CAFA challenge due to its simplicity of implementation on multiple data sources, straightforward integration of multiple scores, and its competitive accuracy with more complex algorithms such as Support Vector Machines.

Similarity scores for three different data sources were calculated in the following way. For protein sequence data source, the similarity score was calculated as percent identity divided by 100. For microarray data source, we used the Pearson correlation between the normalized gene expressions to measure the similarity score between two proteins. In protein-protein interaction (PPI) data source, the similarity score was set to 1 if the two proteins interacted and 0 otherwise.

Using equation (1) we obtained several scores for each pair (p,f). In particular, one score was obtained using sequence similarity, $score^{SEQ}(p,f)$, and one using PPI, $score^{PPI}(p,f)$. We used J microarray data sets, thus we obtained J gene expression scores, $score_j^{EXP}(p,f)$, $j = 1 \ldots J$. Given the J+2 scores for a pair (p,f), an open question is what is the best way to integrate them into a single score. We considered several approaches that calculate final score as a weighted average of individual scores,

$$score(p, f) = w^{SEQ} \cdot score^{SEQ}(p, f) + w^{PPI} \cdot score^{PPI}(p, f) + \sum_{j=1}^{J} w_j \cdot score_j^{EXP}(p, f), \quad (2)$$

where w^{SEQ}, w^{PPI}, and w_j are the corresponding weights. We studied several schemes including assigning different weights to different functions f and assigning the same weights to all functions, weight optimization by likelihood maximization and weight optimization by large margin approaches. We also considered enhancing (1) with the functional similarity scheme proposed in [1]. Interestingly, in our experiments, none of these approaches worked consistently and significantly better than the simple averaging. As a result, for CAFA challenge we decided to give the same weight to all 3 data sources and used w^{SEQ}= 1/3, w^{PPI}=1/3, and w_j=1/(3J).

3. RESULTS

We focused on integration of multiple data sources for function prediction of human proteins. There were 8714 annotated human proteins in the CAFA training set. We obtained sequence identity scores for all pairs of CAFA proteins. For gene expression data, we downloaded 392 Affymetrix GPL96 Platform microarray datasets from GEO. For PPI we used physical interactions between human proteins listed in OPHID database. In total, 2869 of the annotated CAFA human proteins were covered by all 3 data sources. For evaluation, we used only GO functions annotated by more than 10 out of the 2869 proteins. This resulted in 240 Molecular Function (MF) and 1123 Biological Process (BP) GO terms.

In Table 1 we show the average AUC for MF and BP predictions. We used leave-one-protein-out cross-validation to evaluate the performance. We used k=20 as neighborhood in all experiments. We show three versions of AUC for protein sequence data, depending on how many sequences were considered to find the k nearest neighbors: (ver.1) only 2,869 overlapping human proteins, (ver.2) all 8,714 human proteins, (ver.3) all 36924 training CAFA proteins. This allowed us to test how useful it is to transfer functions from paralogous (versions 1 and 2) and both paralogous and orthologous (version 3) proteins.

The results show that sequence similarity is consistently superior to gene expression and PPI data and that it is beneficial to transfer functions to human proteins from their orthologues. Gene expression is more useful for MF prediction, while PPI is more useful for BP prediction. Integration of all 3 data sources improved AUC significantly on both MF and BP terms.

Table 1. Average AUC for 240 MF and 1123 BP terms

Data Source	MF terms	BP terms
Microarray data	0.6442	0.6279
PPI data	0.6283	0.6671
Protein Sequence data, ver.1	0.7636	0.6642
Protein Sequence data, ver.2	0.7896	0.6921
Protein Sequence data, ver.3	0.8396	0.7537
Integrating 3 data sources, ver.1	**0.8134**	**0.7468**
Integrating 3 data sources, ver.2	**0.8494**	**0.7939**
Integrating 3 data sources, ver.3	**0.8788**	**0.8165**

To get further insights, in Figure 1 we compare AUCs of sequence similarity scores (ver.3) and integrated scores (ver.3) for each MF (left panel) and BP (right panel) function.

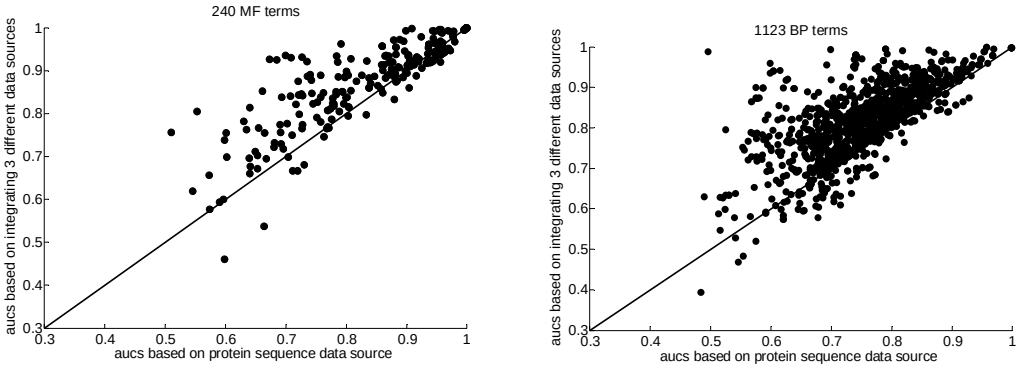

Figure 1. Accuracy comparison on 240 MF and 1123 BP functions

4. ACKNOWLEDGEMENTS

We thank Dr. Predrag Radivojac from Indiana University for providing us with the sequence similarity data for CAFA proteins and human protein-protein interaction data.

5. REFERENCES

[1] Pandey G, Myers CL, Kumar V (2009) Incorporating functional inter-relationships into protein function prediction algorithms. BMC Bioinformatics 10: 142