International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 4, No. 3 (2015): 667-672
Research Article
Open Access

ISSN: 2320-0634

Protein Function Prediction Using Hypergeometric Distance with Background Frequency from Protein-Protein Interaction Network

Md. Khaled Ben Islam[1]*, Julia Rahman[2], Md. Al Mehedi Hasan[2] and Md. Abdur Rahim[1]

[1] *Department of Computer Science & Engineering, Pabna University of Science & Technology, Pabna, Bangladesh*
[2] *Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh*

*Corresponding author: Md. Khaled Ben Islam; email: mdkhaledben@gmail.com

Received: 13 April 2015 Accepted: 03 May 2015 Online: 09 May 2015

ABSTRACT

Due to the availability of large scale experimental data of protein-protein interaction, it remains a challenging task to predict the protein's function from those high-throughput datasets. Widespread presence of false positive interaction and annotation error in public protein databases are the primary barrier in this case. Considering this fact, we work on network-based statistical approach to predict protein function. Our work is based on the insight that if two proteins have significantly larger numbers of common interaction partners in the experimental network than what is expected from a random network, they have close functional associations and most frequently occurred function in the network has higher probability to be found in the neighbourhood of a protein. First we pre-process both BioGRID and DIP interaction datasets, and then employ the chi-square scoring method with and without considering hypergeometric distance of each interacting pairs. We performed leave-one-out validation to evaluate prediction performance in both cases and find that considering hypergeometric distance improves the performance significantly.

Keywords: Protein Interaction Network, Protein Function Prediction, Hypergeometric Distance, Chi Square.

1. INTRODUCTION

Protein is one of the influencer molecules of life, so it has become essential to know the functionality of it. It has great biomedical and pharmaceutical implications. Since the number of unannotated proteins of different organism is continuously increasing and there is still no sophisticated technique of functionality determination available at hand, accurate annotation of protein function becomes an open issue in the post genomic era. Experimental determination of protein functions i.e. wet lab techniques are relatively time consuming and costly. On the other hand, currently developed computational techniques are subject to various limitations. Though functionality can be predicted computationally in a certain extent, they are subject to manual verification in the wet lab experiment. Researchers are continuously trying to develop new computational approaches so that functionality of protein can be predicted more accurately and computational approach can be used to formulate biological hypotheses as well as to guide wet lab experiments through prioritization.

Predrag Radivojac *et al.* [1] reported on a community based initiative, Critical Assessment of Protein Function Annotation (CAFA), an experiment dedicated to evaluating computational tools for protein function prediction. Arvind Kumar Tiwari and Rajeev Srivastava [2] discussed on various computational techniques for function prediction. Different authors used different protein data source for function prediction. Some researchers used sequence data; some used structural data; some used protein interaction data; some used gene expression data or analysed pathway; even some used multiple data sources simultaneously. Our work is based on protein-protein interaction data. We choose

interaction data because, protein function is a context sensitive task and proteins interact with each other to perform their function. Hence the function of a protein may be inferred by looking at its interaction neighbourhood [3]. Moreover, different diseases can be easily analysed by analysing protein interaction data [4].

Several works have already been done to predict protein function using protein interaction network. Some researchers have tried to predict the functionality of a target protein based on its direct interacting neighbour proteins [5-6]. Similarly, Haretsugu Hishigaki *et al.* [7] used chi-square statistics by looking at all proteins within a specified radius in the network for functionality prediction and Tae-Ho Kang *et al.* [8] employed a modified chi-square measure. Many researchers also use graph based approaches with Gene Ontology based distance metrics for functionality prediction of a target protein [9-12]. Some researchers also first enriched the original protein interaction network and then employed a collective classification algorithm on the new network to predict protein function [13].

In this work, we have find out pairwise hypergeometric distance for all the interacting proteins and reconstruct the interaction network based on the distance for both BioGRID [14] and DIP [15] dataset and then employed chi square measure to include background frequency of annotations for protein function prediction..

2. MATERIALS AND METHODS

2.1. Interaction Datasets

We used two interaction datasets of budding yeast (*Saccharomyces cerevisiae)*. One set is taken from BioGrid database (release December 2014, version 3.2.119) and another set is taken from DIP (release January 2015). We pre-processed both datasets to fit our needs i.e. we have considered only physical interactions between proteins. We have also filtered out the interactions which are solely based on high throughput Yeast Two-hybrid assays from BioGRID dataset because, they are inherently error prone. Since we have to verify our prediction, we have kept only those interactions in which both interactors are already experimentally annotated.

2.2. Annotation Datasets

We used the annotation dataset of *Saccharomyces cerevisiae* from Gene Ontology Consortium [16] (release December 2014). As like [6] and [17], we have filtered out proteins that are annotated either by electronic means or have ambiguity in the evidence used to annotate the proteins. We have included only those proteins annotated with experimental evidence codes IDA, IEP, IGI, IMP, IPI, RGA and TAS. We have done this to avoid the uncertainty of misannotation in public protein databases. Alexandra M. Schnoes *et al.* [18] shown that there exists lot of misannotation in pubic protein databases when considering only computational annotation techniques.

2.3. Function Prediction Algorithm

Our approach is based on the two observations. Firstly, if two proteins have significantly larger numbers of common interaction partners in the experimental network than what is expected from a random network, they have close functional associations and most frequently occurred function in the network has higher probability to be found in the neighbourhood of a protein. This observation is found in the work presented in [3] and [19]. Secondly, when using neighbourhood information of a protein for function prediction, considering the background frequency of the target functions may improve the prediction accuracy.

To accomplish our task, we first calculated the hypergeometric distance for each interacting protein pairs like [3] and [19]. If $p_1$and p_2 are two interacting protein pair within an interaction network having total N proteins and m common interacting proteins, then the hypergeometric distance between them is-

$$P(N,p_1,p_2,m) = \frac{\binom{N}{m}\binom{N-m}{n_1-m}\binom{N-n_1}{n_2-m}}{\binom{N}{n_1}\binom{N}{n_2}}$$

$$= \frac{(N-n_1)!\,(N-n_2)!\,n_1!\,n_2!}{N!\,m!\,(n_1-m)!\,(n_2-m)!\,(N-n_1-n_2+m)!}$$

We then reconstruct the interaction network by keeping only those interacting pairs that satisfy a certain threshold level i.e. hypergeometric-distance <= Th_{cutoff} is considered significant. Cutoff threshold is calculated as-

$$\mathrm{Th_{cutoff}} = {}^{1}/_{N^2}$$

In the second phase, we employ chi square method [3] in both original datasets and reconstructed datasets for functionality prediction. For each target protein u, we use the following chi-square scoring function for scoring each possible functions and assign the top scored functions to the target protein. If we have to place k functions to wet lab experiment, then k functions will be assigned to an unannotated protein, with the k largest chi-square scores.

$$Chi\ Square\ score, S_x(u) = \frac{(f_x(u) - e_x(u))^2}{e_x(u)}$$

where,

$$e_x(u) = N(u) * \frac{Total\ Number\ of\ Protein\ with\ x\ function}{Total\ Number\ of\ Protein}$$

$N(u) = Total\ Neighbor\ of\ u\ protein$

$f_x(u) = Total\ Number\ of\ Neighbor\ of\ u\ protein\ with\ x\ functionality$

2.4. Assessment of Algorithm

To asses our approach of prediction, we have used the similar strategy that was used in the work [6] but, in a larger context.We have used the interaction datasets of budding yeast (*Saccharomyces cerevisiae*) collected from BioGRID and DIP. We have chosen yeast interaction dataset because it is one of the model organisms for evaluating computational techniques. We have analysed both dataset and found that many proteins are not well annotated; even many proteins are still un-annotated (Figure 1 and Figure 2). In this situation, we have discarded completely un-annotated proteins from our test dataset and use leave-one-out method to evaluate predictions performed by both simple Chi-Square method and Hypergeometric Distance with Background Frequency. In our case, a target protein is held out (i.e. its annotations are considered unknown) and a prediction is computed using the rest of the annotation information in the network as like [20].

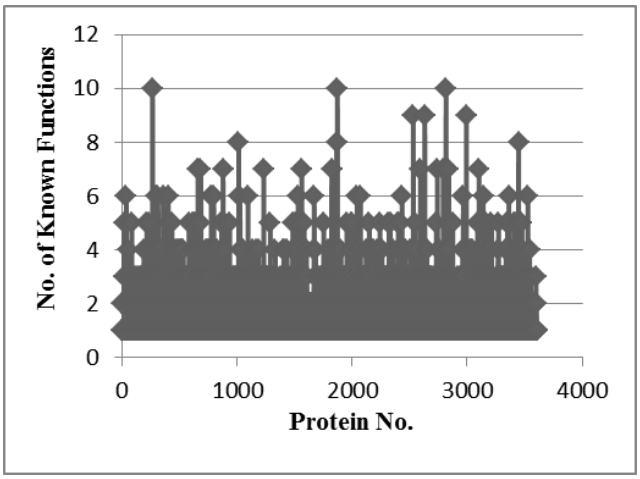

Figure 1: Distribution of known functions of interacting proteins in interaction network of Yeast collected from BioGRID Dataset v3.2.119

We have observed that a significant number of proteins of *Saccharomyces cerevisiae* perform several functions, and hence have multiple annotations. Hence, annotation prediction for a protein is a multi-label classification problem and prediction for a protein is a set of annotations. Therefore, the prediction can be fully correct, partially correct (with different levels of correctness) or fully incorrect. To facilitate all the cases, we have used Precision and Recall as performance measure using the following definitions presented in [21].

Precision (P): Precision is the proportion of predicted correct annotations to the total number of actual annotations,averaged over all target proteins.

$$Precision, P = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Z_i|}$$

Recall (R): Recall is the proportion of predicted correct annotations to the total number of predicted annotations, averaged over all target proteins.

$$Recall, R = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i|}$$

Where, Y is the set of predicted annotations for a target protein, Z is the set of actual annotations for a target protein, and n is the total number of target proteins.

Since computationally predicted functions are subject to experimental verification and most proteins of Yeast have exposed small numbers of functionality, so we have shown the prediction result for 1 to 10 functionality assignment.

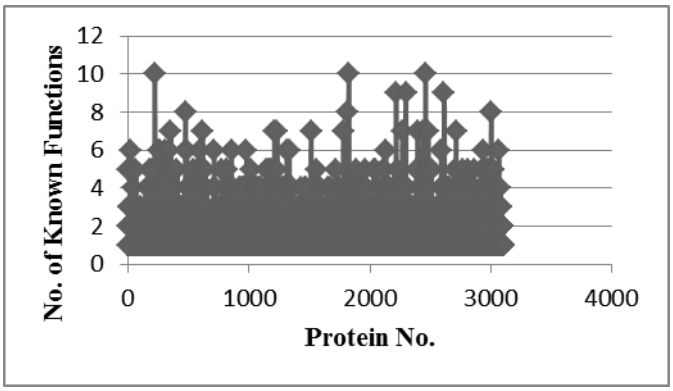

Figure 2: Distribution of known functions of interacting proteins in interaction network of Yeast collected from DIP Dataset (published on 20 December 2014)

3. RESULTS AND DISCUSSION

We have implemented both Hypergeometric Distance with Background Frequency and Chi-Square measure in Matlab and tested them on Yeast proteins interaction dataset of both BioGRID and DIP, considering the molecular functional aspects. In both cases, we have considered the exact match for gene ontology annotation as reported by [6] and [10].

The results show that Hypergeometric Distance with Background Frequency approach performed better than simple Chi-Square method in case of all performance metrics for both datasets (Figure 3 to Figure 6). It implies that, first considering the interactions with low hypergeometric distance i.e. interactions having biological significance with target function's background frequency in the network is positive for functionality prediction. From figure 3 and figure 4, we observe that our approach have predicted functions with higher precision in case of small number of functionality prediction target. On the other hand, higher recall rate have shown in case of larger number of function prediction target (Figure 5 and Figure 6). This is because most of the proteins have small number of experimentally known functions in our considered datasets. From the results shown in below figures, it is also clear that BioGRID dataset contains more false positive than DIP dataset, because both approaches have performed better in DIP than BioGRID dataset.

In addition, in computational approaches, we are more interested in predicting True Positive and False

Positive functionality than True Negative and False Negative functionality. Because of the large search space of possible functionality and high experimental cost to verify them, characterizing a protein using positive predictions is more feasible compared to using negative ones. This fact is also mentioned in paper [6]. Hence, in case of both BioGRID and DIP dataset, the trade-off between precision and recall shown by our approach is reasonable for wet lab experiment (Figure 7 and Figure 8).

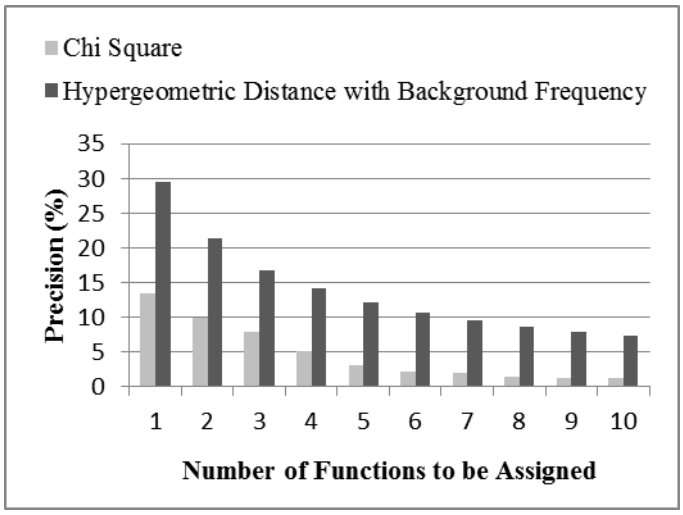

Figure 3: Precision in case of different top scored function assignment for Hypergeometric Distance with Background Frequency and Chi Square method using BioGRID dataset

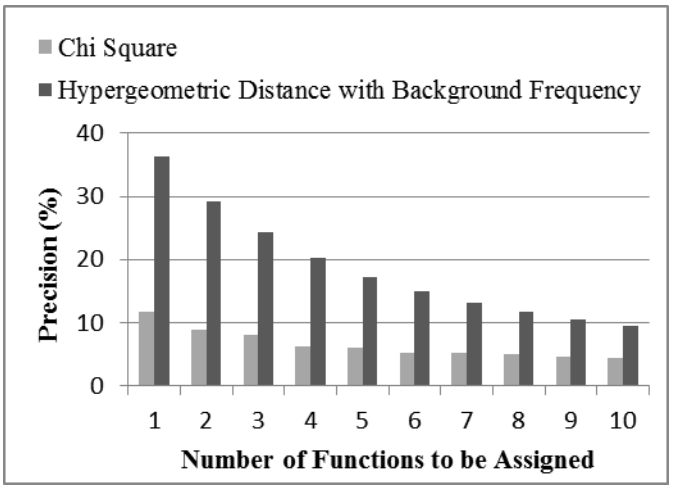

Figure 4: Precision in case of different top scored function assignment for Hypergeometric Distance with Background Frequency and Chi Square method using DIP dataset

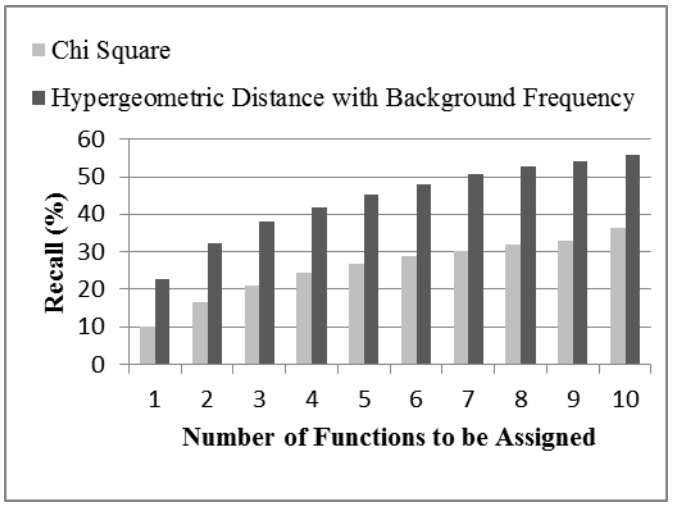

Figure 5: Recall in case of different top scored function assignment for Hypergeometric Distance with Background Frequency and Chi Square method using BioGRID dataset

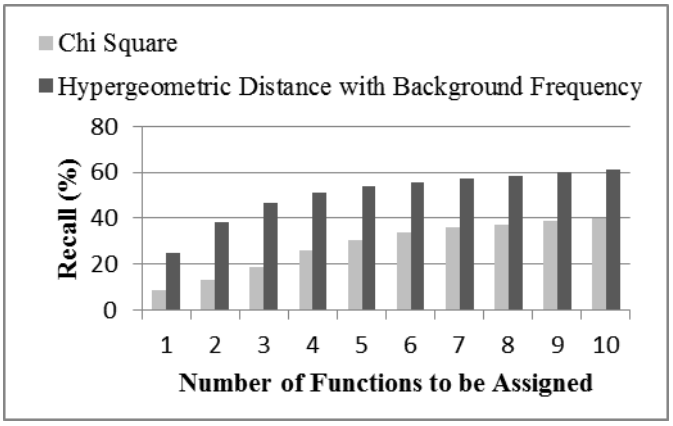

Figure 6: Recall in case of different top scored function assignment for Hypergeometric Distance with Background Frequency and Chi Square method using DIP dataset

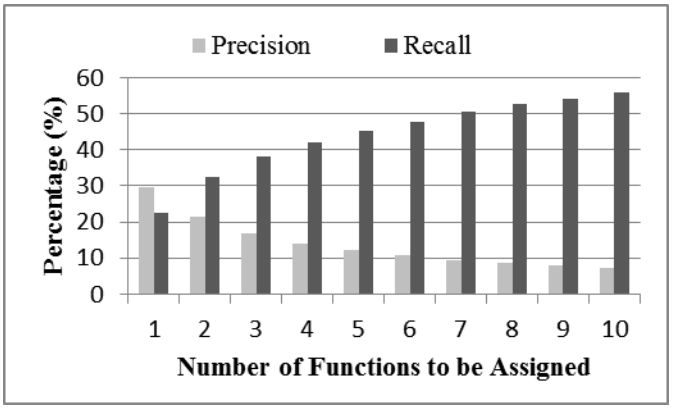

Figure 7: Precision and Recall level in case of different top scored function assignment for Hypergeometric Distance with Background Frequency method using BioGRID dataset

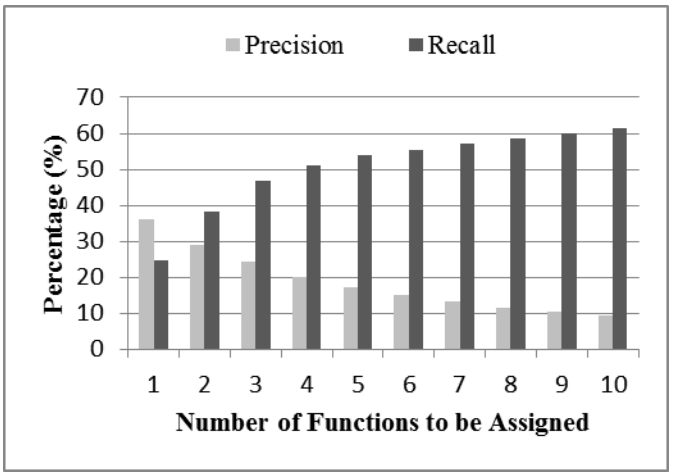

Figure 8: Precision and Recall level in case of different top scored function assignment for Hypergeometric Distance with Background Frequency method using DIP dataset

4. CONCLUSION

In this research work, we tried to predict protein functionality using two well-known protein-protein interaction datasets i.e. BioGRID and DIP. In this case we have used the facts that biologically significant interacting protein pairs have higher functional association and there is a close relationship between the background frequency of a target protein's functionality and neighbourhood of the protein. Results shows that if we consider both facts at the same time when predicting protein functionality, then performance improves significantly instead of considering only background frequency issue. We have experimented on Yeast interaction dataset and used leave-one-out cross validation method. We have used precision and recall as performance metric because our target was to check how precisely our approach can predict the target proteins actual functionality. Though the outcome was not good enough as expected but the exploitation of our considered fact will be very useful for future research and to predict the proteins functionality more accurately.

Acknowledgments

We thank Mohammed Nasser (Professor, Department of Statistics, University of Rajshahi, Bangladesh) for discussions during the development of our work.

REFERENCES

1. Predrag Radivojac, Wyatt T Clark, and Tal Ronnen Oron *et al.* (2013). A large-scale evaluation of computational protein function prediction. Nature Methods. 10: p. 221-228.
2. Arvind Kumar Tiwari and Rajeev Srivastava (2014). A Survey of Computational Intelligence Techniques in Protein Function Prediction. International Journal of Proteomics. 2014: (article id 845479), 22 pages.
3. Xiao-li Li, and See-Kiong Ng (2009). Biological Data Mining in Protein Interaction Network. IGI Global, USA, ISBN: 9781605663982.
4. Ke Jin, Gabriel Musso, and James Vlasblom *et al.* (2014). Yeast mitochondrial protein-protein interactions reveal diverse complexes and disease-relevant functional relationships. Journal of Proteome Research. 14: p. 1220–1237.
5. Benno Schwikowski, Peter Uetz, and Stanley Fields (2000). A network of protein-protein interactions in yeast. Nature Biotechnology. 18: p. 1257–1261.
6. Md. Khaled Ben Islam, Julia Rahman, and Md. Al Mehedi Hasan *et al.* (2015). Protein Function Prediction Using Neighbor Counting with Dynamic Threshold from Protein-Protein Interaction Network. Computational Biology and Bioinformatics. 3: p. 1-5.
7. Haretsugu Hishigaki, Kenta Nakai, and Toshihide Ono *et al.* (2001). Assessment of prediction accuracy of function from protein-protein interaction data. Yeast. 18: p. 523–531.
8. Tae-Ho Kang, Myung-Ho Yeo, and Jae-SooYoo (2009). A Novel Method for Functional Prediction of Proteinsfrom a Protein-Protein Interaction Network. Journal of the Korean Physical Society. 54: p. 1716-1720.
9. Behnam Neyshabur, Ahmadreza Khadem, and Somaye Hashemifar *et al.* (2013). NETAL: a new graph-based method for global alignment of protein-protein interaction networks. Bioinformatics. 29: p. 1654–1662.
10. Mengfei Cao, Hao Zhang, and Jisoo Park *et al.* (2013). Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks. PLoS One 8(10): e76339.
11. Kire Trivodaliev, Ilinka Ivanoska, and Slobodan Kalajdziski *et al.* (2015). Novel Gene Ontology Based Distance Metric for Function Prediction via Clustering in Protein Interaction Networks. ICT Innovations 2014. Advances in Intelligent Systems and Computing. 311: p. 167-176.
12. Darren Davis, Omer Nebil Yaveroğlu, Noël Malod-Dognin *et al.* (2015). Topology-function conservation in protein–protein interaction networks. Bioinformatics. 31: p. 1-8.
13. Wei Xiong, Hui Liu, and Jihong Guan *et al.*(2013). Protein function prediction by collectiveclassification with explicit and implicit edges inprotein-protein interaction networks. BMC Bioinformatics. 14(Suppl 12):S4.
14. Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly *et al.* (2006). Biogrid: A General Repository for Interaction Datasets. Nucleic Acids Research. 34: D535-9.
15. Xenarios I, Salwinski L, Duan XJ *et al.* (2002). DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. Nucleic Acids Research, Database Issue. 30: 303-5.
16. Michael Ashburner, Catherine A. Ball, and Judith A. Blake *et al.* (2000). Gene Ontology: Tool for the unification of biology. Nature Genetics. 25: p. 25 – 29.
17. Ömer Sinan Saraç, Volkan Atalay, and Rengul Cetin-Atalay (2010). GOPred: GO molecular function prediction by combined classifiers. PloS One. 5(8): e12382.
18. Schnoes AM, Brown SD, Dodevski I *et al.* (2009). Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. PLoSComputBiol. 5(12): e1000605.
19. Manoj Pratim Samanta and Shoudan Liang (2003). Predicting protein functions from redundancies inlarge-scale protein interaction networks. PNAS. 100: p. 12579-12583.
20. Petko Bogdanov, and Ambuj K. Singh (2010). Molecular Function Prediction using Neighborhood Features, IEEE/ACMTransactions on Computational Biology and Bioinformatics. 7: p. 208-217.
21. Mohammad S Sorower (2010). A Literature Survey on Algorithms for Multi-label Learning. Ph.D Qualifying Review Paper, Oregon State University.

*****