Computational Biology and Bioinformatics
2015; 3(1): 1-5
Published online January 31, 2015 (http://www.sciencepublishinggroup.com/j/cbb)
doi: 10.11648/j.cbb.20150301.11
ISSN: 2330-8265 (Print); ISSN: 2330-8281 (Online)

Protein Function Prediction Using Neighbor Counting with Dynamic Threshold from Protein-Protein Interaction Network

Md. Khaled Ben Islam[1, *], Julia Rahman[2], Md. Al Mehedi Hasan[2], Mohammed Nasser[3]

[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
[3]Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh

Email address:
mdkhaledben@gmail.com (M. K. B. Islam), juliacse06@gmail.com (J. Rahman), mehedi_ru@yahoo.com (M. A. M. Hasan), mnasser.ru@gmail.com (M. Nasser)

To cite this article:
Md. Khaled Ben Islam, Julia Rahman, Md. Al Mehedi Hasan, Mohammed Nasser. Protein Function Prediction Using Neighbor Counting with Dynamic Threshold from Protein-Protein Interaction Network. *Computational Biology and Bioinformatics.* Vol. 3, No. 1, 2015, pp. 1-5. doi: 10.11648/j.cbb.20150301.11

Abstract: In recent years, a large number of proteins of different organisms have been discovered but due to high experimental cost and uncertain time boundary, yet it is not possible to find out all of the functionalities of those proteins. With the recent advent of huge protein-protein interactions, it becomes an opportunity to computationally predict a protein's functionality based on its interacting partners. In this work, we mainly try to find out a way by which we can predict functionality of a target protein with low computational complexity. We propose a simple approach for protein function prediction based on Classical Neighbor Counting method. We also investigate the functional dependency of a protein to its direct neighbors in the interaction network. We find that when majority of its interacting partners have more experimentally known annotation, then more accurately we can predict a protein's functionality using Neighbor Counting technique.

Keywords: Protein Function Prediction, Neighbor Counting with Dynamic Threshold, Protein-Protein Interaction Network

1. Introduction

Discovering the functions of proteins in living Organism is essential to understand our life at a molecular level. Though Protein sequences have been published at a dramatic rate but a large fraction of newly discovered proteins have no experimentally validated functional annotation.

Experimentally determination of protein function is expensive, time consuming and some experiments cannot be performed in some organisms for a variety of biological or ethical reasons. So, successful computational predictive methods have an important role to play in this regard. Besides computational function prediction can also be used to formulate biological hypotheses and guide wet lab experiments through prioritization.

Several Computational techniques have been developed that can be used to predict protein function including analyzing gene expression pattern [1-2], phylogenetic profiles [3-4], protein sequences [5-6], and protein structures [7-8]. In case of protein function prediction by analyzing gene expression pattern, it is considered that co-expressed proteins may have related functions. In case of analyzing phylogenetic profile, evolutionary history of proteins is used for inferring functionality to unannotated protein. In case of protein function prediction using protein sequence, sequence similarity measures, homologies are primarily used. In case of function prediction using protein structure, structural alignment may be used; even they differ in their sequence data. In [9-10], authors also discuss about other methods of computational function prediction. Each of these approaches has achieved some success in some particular case, but in general, yet they cannot be reliably used to predict proteins functionality.

Since availability of protein-protein interaction data is increasing and proteins interact with each other for a common purpose, so a protein's functionality may be predicted based on the functionality of its interacting neighbors.

Several attempts have already been taken to predict the protein's functionality using its interaction network. In [11], authors tried to predict the functionality of an un-annotated protein directly from its neighbors. In [12], authors use χ^2 statistics by looking at all proteins within a specified radius, though it doesn't take into account the underlying topology of the PPI network. In [13], authors introduce a new metric based on a graph diffusion property to transfer function to unannotated protein from already annotated interacting partners.

In this work, we propose an approach based on neighbor's functional frequency count with dynamic threshold to annotate a target protein and investigate the influence of direct neighbors for assigning function to an un-annotated protein.

2. Materials and Methods

2.1. Datasets

We have used the molecular interactions of *Saccharomyces cerevisiae* from the BioGrid database [14] (release November 2014, version 3.2.118). We have collected the annotation dataset for *Saccharomyces cerevisiae* from Gene Ontology Consortium [15] (release November 2014). We have considered only physical interactions and have discarded those physical interactions which are solely based on high-throughput Yeast Two-hybrid and Protein-RNA assays. We have filtered out the Y2H high throughput assay because they are inherently error prone. We have filtered out Protein-RNA interactions because our main focus was only on the protein's functionality prediction. Initially there were 139693 raw physical interactions. Since we have to verify our prediction, we have kept only those interactions in which interactors are already annotated.

As like [16], we have also 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 database. In [17], authors show that there exits lots of misannotation in pubic protein databases when considering only computational annotation techniques. In our prepared datasets, there were 48835 Non-Redundant Interactions between 3594 unique annotated proteins.

2.2. Proposed Algorithm

We propose an approach for assigning function to an un-annotated protein based on its neighbor's functional annotation's frequency. Actually it's a modification of Classical Neighbor Counting method [11, 18]. The main concept of our approach is based on the fact that the functionalities shown by more neighbors have likely to be shown also by the target proteins and this likeliness varies with the target protein and its respective annotated neighbors.

For each target protein p in the interaction network, each function f_i F is given a score based on the frequency of its occurrence in the direct neighbors of p, where F is the set of functionalities shown by all neighbors of p. The functions having a minimum threshold frequency are assigned to the target protein. This minimum threshold frequency is adjustable for each target protein.

Scoring Function,

$$S_f(p) = \sum_{n \epsilon N_p} \delta(n, f)$$

Where,
(n, f) = 1 if neighbor n has function f, 0 otherwise;
N_p refers to the set of direct interacting neighbors of protein p.

Threshold,

$$Th = \begin{cases} i_{th}\ ; if \max\left(S_f(p)\right) \geq\ i_{th} \\ \max\left(S_f(p)\right)\ ; otherwise \end{cases}$$

Where,
i_{th} denotes the initial threshold;
Th denotes the threshold, adjusted after score calculation.

2.3. Assessment of Algorithms

In our used Interaction dataset of *Saccharomyces cerevisiae,* many annotations are of low frequency, many proteins are not well annotated; even many proteins are still un-annotated. From Fig. 1, we see that most of the proteins have only 2 or 3 known functions i.e. GO annotation id. Other proteins have small numbers of known functions. In this situation, we discard completely un-annotated proteins from our test dataset and use *leave-one-out* method to evaluate predictions performed by both Classical Neighbor Counting (CNC) and Neighbor Counting with Dynamic Threshold method (NCDT). 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 [19].

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 use Precision and Recall as performance measure besides Accuracy calculation using the following definitions presented in [20].

Accuracy (A): Accuracy for each target protein is defined as the proportion of the predicted correct annotations to the total number (predicted and actual) of annotations for that protein. Overall accuracy is the average across all target proteins.

$$Accuracy, A = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}$$

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.

In addition, majority of the proteins interact with around 1 to 30 proteins directly in our filtered interaction network. Some of the proteins directly interact with large number of other proteins, but this type of proteins is low in our considered protein-protein interaction network. So we show the predicted result for 1 to 15 threshold level. Our approach will be able to predict the functionality of those proteins but we have to initiate the initial threshold level accordingly for better prediction performance.

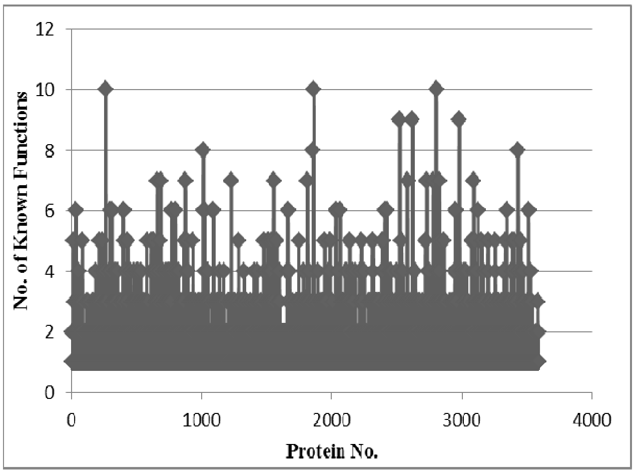

Fig. 1 *Distribution of known functions of interacting proteins in our filtered interaction network of BioGRID Dataset v3.2.118*

3. Results and Discussion

We have implemented both Classical Neighbor Counting (CNC) [11] and our proposed Neighbor Counting with Dynamic Threshold (NCDT) methods in MatLab and have tested them on *Saccharomyces cerevisiae*'s filtered interaction dataset for molecular functional aspects collected from BioGRID database. In all cases, we have considered the exact match for gene ontology annotation like [13].

The results show that adding Dynamic Threshold option improves the performance of Classic Neighbor Counting method for protein function prediction (Fig. 2, Fig. 3 and Fig. 4). In our prepared dataset of *Saccharomyces cerevisiae*, Classical Neighbor Counting considers most of the annotations of neighboring proteins as negative for lots of test proteins in higher degree of majority count (Fig. 5). On the other hand, from Fig. 6, we observe that NCDT balance the ratio of the percentage of correct positive predictions and the percentage of positive annotation retrieved among the actual functionalities of the test proteins.

In addition, we are more interested in predicting True Positive and False Positive annotations than True Negative and False Negative annotations. Because of the large 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.

From Fig. 6, we see that our approach can be easily used to predict the possible functionalities of a target protein, those functionalities may be verified in future by wet-lab experiment. We also see that when majority of a test protein's interacting partners have more experimentally known annotation, then more accurately we can predict a protein's functionality using neighbor counting technique.

Besides, Functional annotations for the proteins of most organisms including *Saccharomyces cerevisiae* are incomplete at present. Not only that, interaction network of different organisms becomes more visible continuously by wet lab experiment. Therefore, annotations that are currently identified as false positive (predicted as a function of the target protein but that is actually not part of the target protein's annotation), may have a chance that will be experimentally verified in future.

To predict the functionality of an unannotated protein using Neighbor Counting with Dynamic Threshold, we have to perform around N * G_N operation for scoring G_N unique GO annotation ID of N interacting proteins and additional one constant operation for threshold level adjustment. This is very low computational cost for predicting functionality of a protein by considering the interacting proteins because at least we have to evaluate the functionalities of all the neighbor proteins. Though our approach takes a little more computation time but it is ignorable because of its higher prediction performance comparing Classical Neighbor Counting.

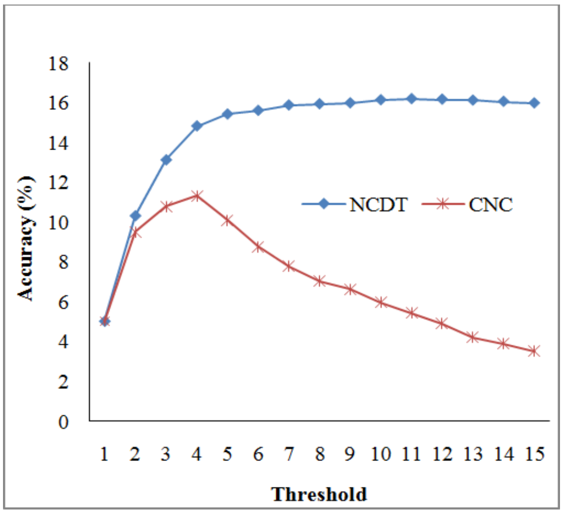

Fig. 2 *Status of Accuracy in Neighbor Counting with Dynamic Threshold (NCDT) and Classical Neighbor Counting (CNC) with respect to different threshold level (Initial Threshold for NCDT)*

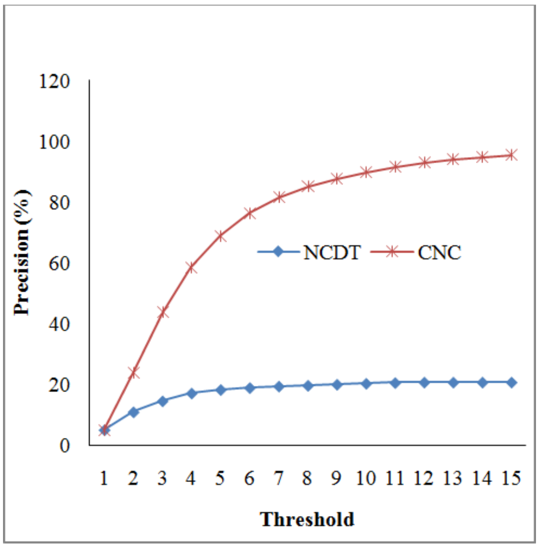

Fig. 3 *Status of Precision in Neighbor Counting with Dynamic Threshold (NCDT) and Classical Neighbor Counting (CNC) with respect to different threshold level (Initial Threshold for NCDT)*

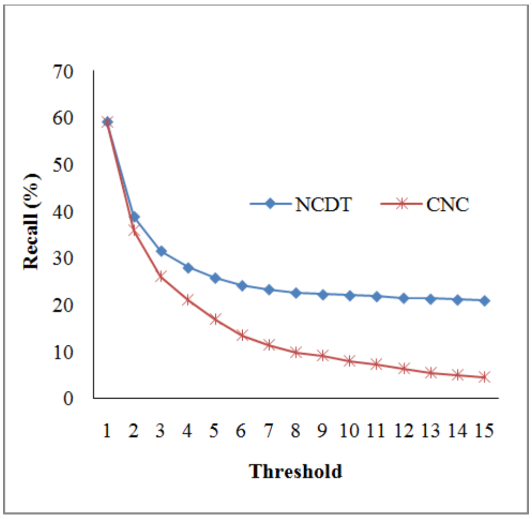

Fig. 4 *Status of Recall in Neighbor Counting with Dynamic Threshold (NCDT) and Classical Neighbor Counting (CNC) with respect to different threshold level (Initial Threshold for NCDT)*

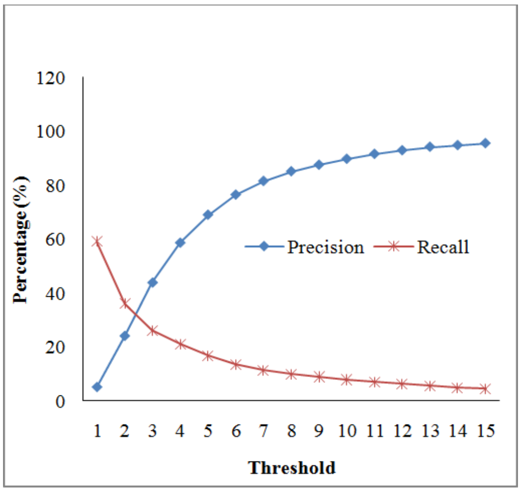

Fig. 5 *Precision and Recall results for the Classical Neighbor Counting (CNC) on filtered Saccharomyces cerevisiae dataset*

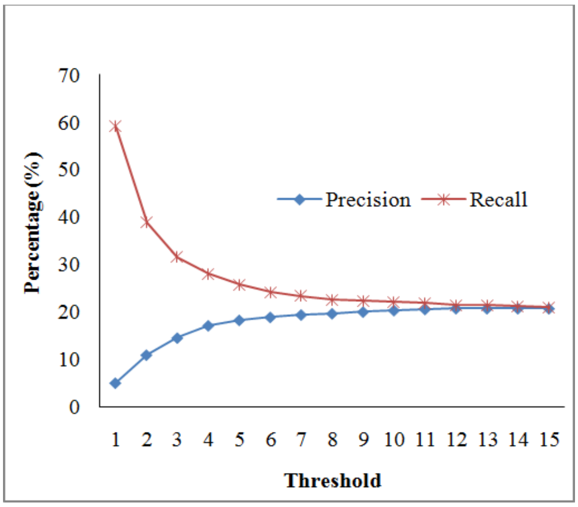

Fig. 6 *Precision and Recall results for the Neighbor Counting with Dynamic Threshold (NCDT) on filtered Saccharomyces cerevisiae dataset*

4. Conclusion and Future Work

In this work, we propose a computationally simple approach for protein function prediction by exploiting the fact that, proteins interacts with each other for a common purpose and therefore functionality of any interacting protein can be predicted from its interacting partners. Basically it's an enhancement to the Classical Neighbor Counting method. Our approach takes into account the possibility that if any functionality appears in a certain number of neighbors or more, then the probability of showing that functionality by the target protein is higher and this number changes with the number of interacting proteins and their verified annotations.

We have tested our proposed approach and Classical Neighbor Counting Method on protein-protein interaction dataset of *Saccharomyces cerevisiae* organism i.e. baker's yeast using the leave-one-out cross-validation. Results show that our approach outperforms the *Classical Neighbor Counting* methods for prediction of protein functions. In addition, we have observed that, to predict possible functionality of a target protein based only on the direct neighbors, more interacting partners with their more experimentally known functionality is beneficial.

Finally it is noted that to improve the reliability of the function prediction, we must consider the annotation states of a function in the whole interaction network and indirect interaction of a protein with others, not only direct neighbors in addition to adjustable threshold level of function assignment.

References

[1] Xing-Ming Zhao, Yong Wang, Luonan Chen, Kazuyuki Aihara, "Gene function prediction using labeled and unlabeled data", BMC Bioinformatics 2008, 9:57.

[2] Loc Tran, "Hypergraph and protein function prediction with gene expression data", Computing Research Repository, Cornell University, abs/1212.0388, 2012.

[3] Appala Raju Kotaru, Ramesh C. Joshi, “Classification of Phylogenetic Profiles for Protein Function Prediction: An SVM Approach”, Contemporary Computing, ISBN: 978-3-642-03547-0, pp 510-520, 2009.

[4] Monique Marlene Morin, “Phylogenetic Networks Simulation, Characterization, and Reconstruction”, PhD Dissertation, The University of New Mexico, December 2007.

[5] Lee Sael, Meghana Chitale, Daisuke Kihara, “Structure- and sequence-based function prediction for non-homologous proteins”, Journal of Structural and Functional Genomics, Volume 13, Issue 2, pp. 111-123, June 2012.

[6] Zheng Wang, Renzhi Cao, Jianlin Cheng, “Three-Level Prediction of Protein Function by Combining Profile-Sequence Search, Profile-Profile Search, and Domain Co-Occurrence Networks”, BMC Bioinformatics 2013, 14 (Suppl 3):S3.

[7] Laskowski RA, Watson JD, Thornton JM, “ProFunc: a server for predicting protein function from 3D structure”, Nucleic Acids Research, Vol. 33:W89-W93, 2005.

[8] Dariya S. Glazer, Randall J. Radmer, Russ B. Altman, “Improving structure-based function prediction using molecular dynamics”, Structure, vol. 17, no. 7, pp. 919–929, 2009.

[9] Arvind Kumar Tiwari, Rajeev Srivastava, “A Survey of Computational Intelligence Techniques in Protein Function Prediction”, International Journal of Proteomics, 2014:845479, 2014.

[10] Predrag Radivojac et al., “A large-scale evaluation of computational protein function prediction”, Nature Methods 10, 221–227, 2013.

[11] Benno Schwikowski, Peter Uetz, Stanley Fields, “A network of protein-protein interactions in yeast”, Nature Biotechnology 18, 1257–1261, 2000.

[12] Haretsugu Hishigaki, Kenta Nakai, Toshihide Ono, Akira Tanigami, Toshihisa Takagi, “Assessment of prediction accuracy of protein function from protein-protein interaction data.”, Yeast, Vol. 18, Issue 6, pp. 523–531, April 2001.

[13] Cao M, Zhang H, Park J, Daniels NM, Crovella ME, et al. “Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks”, PLoS One 8(10): e76339. doi:10.1371/journal.pone.0076339, 2013.

[14] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Mike Tyers, “Biogrid: A General Repository for Interaction Datasets”, Nucleic Acids Research, Vol. 34:D535-9, 2006.

[15] Michael Ashburner, et al., “Gene Ontology: Tool for the unification of biology”, Nature Genetics 25, 25 – 29, 2000.

[16] Ömer Sinan Saraç, Volkan Atalay, Rengul Cetin-Atalay, “GOPred: GO molecular function prediction by combined classifiers”, PloS One, 5(8): e12382, doi:10.1371/journal.pone.0012382, 2010.

[17] Schnoes AM, Brown SD, Dodevski I, Babbitt PC, Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. PLoS Comput Biol 5(12): e1000605. doi:10.1371/journal.pcbi.1000605, 2009.

[18] Hon Nian Chua, Limsoon Wong, “Predicting Protein Functions from Protein Interaction Networks”, International Journal of Knowledge Discovery in Bioinformatics, Vol. 3, Issue 4, pp. 50-70, 2012.

[19] Petko Bogdanov, Ambuj K. Singh, “Molecular Function Prediction using Neighborhood Features”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 7, Issue 2, pp. 208-217, April 2010.

[20] Mohammad S Sorower, “A Literature Survey on Algorithms for Multi-label Learning”, Ph.D Qualifying Review Paper, Oregon State University, 2010.