Scored Protein-Protein Interaction to Predict Subcellular Localizations for Yeast Using Diffusion Kernel

Ananda Mohan Mondal[1,*] and Jianjun Hu[2]

[1] Mathematics and Computer Science, Claflin University, Orangeburg, USA
amondal@claflin.edu
[2] Computer Science and Engineering, University of South Carolina, Columbia, USA
jianjunh@cec.sc.edu

Abstract. Network-based protein localization prediction is explored utilizing the protein-protein interaction score along with the network connectivity. Score-based diffusion kernel is introduced to solve the problem. Four different PPI networks, namely, co-expressed PPI, Genetic PPI, Physical PPI, and scored PPI are used for analysis. Our investigation shows that PPI score does have positive impact in predicting subcellular protein localization. At high average PPI score of 891, performance accuracy ranges from 0.78 for 'punctate composite' to 0.93 for 'nucleolus' and at low average PPI score of 169, performance accuracy ranges from 0.60 for 'cytoplasm' to 0.83 for 'mitochondrion'.

Keywords: Scored PPI, subcellular protein localization, protein localization, diffusion kernel, NetLoc.

1 Introduction

Precise targeting to designated subcellular localization is essential for proper protein function. Experimental determination of protein localization is costly [1, 2]. Computational algorithm can greatly help in predicting protein localizations, which in turn can infer protein functions. In the past decade, many computational algorithms have been developed for predicting subcellular localization of protein. These algorithms employ a variety of supervised machine learning techniques including support vector machines[3], neural networks [4], nearest neighbor classifier, Markov models, Bayesian networks [5, 6] etc. The existing prediction algorithms can be divided into four major categories [7] in terms of the evidences used: 1) algorithms based on targeting signals; 2) algorithms considering the preference or bias in terms of amino acid; 3) algorithms using localization information from other annotated proteins with indirect relationships such as functional annotation, phylogenetic profiling, homology and protein-protein interaction; and 4) algorithms that integrate multiple sources of information.

Recently, protein-protein correlation (PPC) networks such as PPI networks [8] and metabolic networks [9] have been used for localization prediction. One of the limitations of these methods that these are not capable of utilizing the inherent network

* Corresponding author.

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 647–655, 2013.

information that naturally appears among proteins [7, 10]. In their recent work, Mondal and Hu [7, 10-12] exploit the PPI network information in predicting subcellular protein localization using diffusion kernel. But these studies do not utilize the PPI score meaning they used score of unity for each PPI. PPI score, ranging between 150 and 999, in the STRING database reflects the confidence of functional association of two different proteins [13]. The higher is the score of a PPI, the higher is the probability of two proteins to be associated with the same function. If two proteins are associated with the same function, they are more likely to be localized at the same subcellular location. So, PPI scores in STRING database can better be utilized in predicting subcellular localization of proteins. It is thus interesting to explore score-based diffusion kernel algorithm in predicting protein subcellular localizations.

2 Diffusion Kernel-Based Logistic Regression using Scored PPI

2.1 Score-Based Diffusion Kernel

Score-based diffusion kernel is derived from weighted adjacency matrix of scored PPI network, where weight on an edge is the score of PPI connecting two proteins. Diffusion kernel *K,* to represent the scored interaction network, is defined using the following equation.

$$K = e^{\{W\}} \tag{1}$$

Where

$$W(i,j) = \begin{cases} w_{i,j} & if\ protein\ i\ interacts\ with\ protein\ j \\ -d_i & if\ prtoein\ i\ is\ the\ same\ as\ protein\ j \\ 0 & otherwise \end{cases}$$

Where d_i is the sum of weights of interactions with protein *i* and $e^{\{W\}}$ represents the matrix exponential of the matrix W. It is noticeable that $w_{i,j} = 1$ represents the diffusion kernel for non-score based PPI network or PPI with score of unity, which is used in [7, 10-12]. Kernel function $K(i, j)$ represents the similarity distance between protein *i* and protein *j* in the network.

2.2 Kernel-Based Logistic Regression (KLR) Model

For classification, we applied the diffusion kernel-based logistic regression (KLR) model [14] as used in [7, 10-12] to predict protein subcellular localization. Given a protein-protein interaction network with N proteins $X_1, \ldots, X_N$ with n of them $X_1, \ldots, X_n$ with unknown subcellular locations, the objective is to find subcellular locations of n unknown proteins using the locations of known proteins and protein-protein interaction network.

Let

$$X_{[-i]} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_N)$$

$$M_0(i) = \sum_{j \neq i, x_j \text{ known}} K(i,j) I\{x_j = 0\}$$

,

And $M_1(i) = \sum_{j \neq i, x_j \text{ known}} K(i,j) I\{x_j = 1\}$,

where $K(i,j)$ is the kernel function derived in section 2.2. Indicator $I(x_j = 0)$ represents that the interacting protein *j* does not have the location of interest and indicator $I(x_j = 1)$ represents that protein *j* does have the location of interest. Upon simplification, the KLR model is given by:

$$\log \frac{\Pr(X_i = 1 \mid X_{[-i]}, \theta)}{1 - \Pr(X_i = 1 \mid X_{[-i]}, \theta)} = \gamma + \delta M_0(i) + \eta M_1(i)$$

This means that the logit of $(X_i = 1 | X_{[-i]}, \theta)$, the probability of a protein targeting a location is linear based on the summed distances of proteins targeting to that location or other locations. This equation can be rewritten as

$$\Pr(X_i = 1 \mid X_{[-i]}, \theta) = \frac{1}{1 + e^{-(\gamma + \delta M_0(i) + \eta M_1(i))}}$$

Maximum likelihood estimation (MLE) method is used for estimating the parameters $\gamma, \delta,$ and η. It is noticeable that only the annotated proteins are used in the estimation procedure.

Fig. 1 presents the flow diagram of the scored PPI network-based framework for protein localization prediction employing the KLR model. First, a scored PPI network or weighted adjacency matrix is obtained from the list of scored PPIs. Then diffusion kernel is determined using eqn. (1).

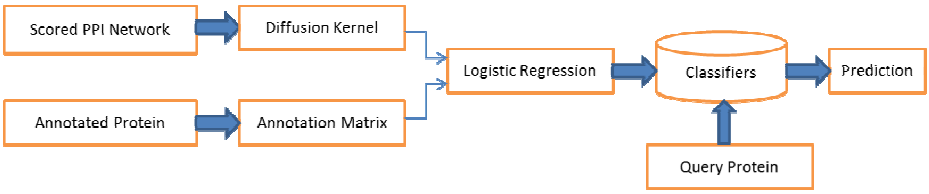

Fig. 1. Protein localization prediction using scored PPI network employing kernel-based logistic regression

Annotation matrix is developed from annotated proteins. This is an *m* by *n* matrix, consists of 1 (annotated) and 0 (not annotated), where *m* is the number of annotated proteins and *n* is the number of localizations. Finally, KLR model is developed from diffusion kernel and annotation matrix using logistic regression.

2.3 Performance Evaluation

The outputs of the KLR model are confidences for each protein to be localized at each of the locations. A threshold on confidence value is used to classify the proteins to be localized at a location or not. If the threshold is set to 0.7, then a protein with higher than 0.7 confidence will be labeled as positive prediction meaning that the protein belongs to this location, otherwise, negative prediction. The results of the KLR prediction algorithm can have varying true positive and true negative rate depending on the threshold value. This makes the comparison difficult. To avoid this difficulty, the AUC (Area Under the Curve) score was used to measure the prediction capability of the proposed KLR model. 5-fold cross-validation was used to calculate the AUC values.

3 Datasets

3.1 PPI Networks

Four different PPI networks for yeast are used in the present study: two networks, physical PPI and genetic PPI, are obtained from BioGRID database [15], one Scored PPI network from STRING database [13], and one co-expressed PPI network is derived from gene expression data of Stanford University [16]. Pearson correlation is used to derive co-expressed PPI from gene expression data. In this study, the networks are named as co-expressed PPI as COEXP, genetic PPI as GPPI, physical PPI as PPPI, and scored PPI as SPPI. Table-1 summarizes the topology of four network datasets. SPPI is the largest network (proteins: 6314; edges: 489934) and COEXP is the smallest (proteins: 2004; edges: 11954). SPPI is also the densest network (77.6 PPI/protein) followed by GPPI (19.73 PPI/protein), PPPI (9.31 PPI/protein), and COEXP (5.96 PPI/protein).

Table 1. Topology of protein-protein interaction networks

Property	**COEXP**	**GPPI**	**PPPI**	**SPPI**
No. of proteins	2004	5252	5477	6314
Edges	11954	103631	50997	489934
Average interactions per node	5.96	19.73	9.31	77.60

3.2 Annotated Proteins

The developed KLR model is applied to protein localization prediction of yeast proteins using the localization data of Huh et al. [1] as the basis for annotation. After removing ambiguous localization, we have 3919 proteins with 5191 localizations. Out of 22 locations, only 7 locations have more than 100 proteins with known subcellular

localization annotation. These locations are cell periphery, cytoplasm, ER (endoplasmic reticulum), mitochondrion, nucleolus, nucleus, and punctuate composite. We evaluated our network prediction model based on these 7 locations.

3.3 Networks in Terms of PPI Score

In COEXP, GPPI, and PPPI network all PPIs have the equal score of unity even though two proteins of different PPIs have different level of confidence of functional association. In order to investigate the relationship of these three types of PPIs with respect to functional association, we identified the PPIs for these three networks that are common with the STRING PPIs. Fig. 1 presents the cumulative distribution of PPI with PPI score for three different types of PPI. In case of GPPI, 50% PPIs have score larger than 375; in case of PPPI, 50% PPIs have score larger than 800; in case of COEXP 50 % PPIs have score larger than 850. So, two proteins in PPI of PPPI and COEXP network are more likely to be related to the same function. By definition, physical PPI means two proteins physically interact together to produce some products, and co-expressed PPI means two proteins have same level of expression in the same direction and as such they are more likely to be associated with the same function, which results in high PPI score for PPPI and COEXP.

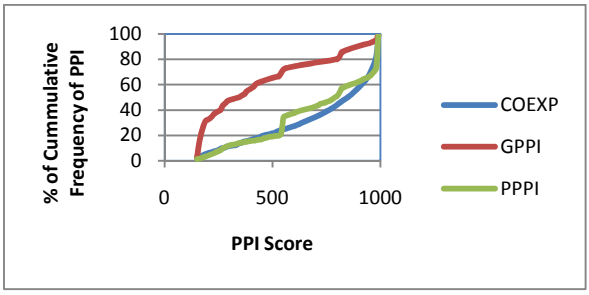

Fig. 2. Distribution of PPIs with PPI score for different types of PPI

4 Results and Discussion

In their previous work [7, 10-12] Mondal and Hu used NetLoc model to predict subcellular localization using PPI network without score. In the present work, we used NetLoc to explore it's capability of similar prediction but using scored PPI from STRING database [13].

In order to see the effect of PPI score in predicting subcellular localization, NetLoc model was applied to five equally divided scored PPI (SPPI) network. The PPIs in the whole SPPI network was sorted in descending order based on PPI score and then divided into five equal parts in terms of number of PPIs. Table 2 summarizes the topology of the divided networks along with the average PPI score and the number of annotated proteins corresponding to each of the divided network. It is noticeable that

topology of the divided networks are very similar in terms of every component of topology: number of PPI ranging from 97986 to 97990, number of protein ranging from 5394 to 5948, and average degree ranging from 33 to 36. The number of corresponding annotated protein is also similar ranging from 3728 to 3870. The major different in the five networks are average PPI score ranging from 169 to 891.

Table 2. Topology of 5 divided networks and average PPI score and number of annotated proteins for corresponding networks

Parts of Network	Network Topology			Avg Score	Ann Protein
	PPI	Protein	Avg Degree		
N1_5	97986	5394	36	891	3728
N2_5	97986	5948	33	575	3870
N3_5	97986	5860	33	359	3837
N4_5	97986	5944	33	230	3839
N5_5	97990	5809	34	169	3809

4.1 Model Performance in Predicting Individual Subcellular Localization

Fig. 3 presents the model performance in predicting 7 subcellular localizations for two of five equally divided networks, N1_5 with the largest average PPI score of 891 and N5_5 with the smallest average PPI score of 169. At high score, performance accuracy ranges from 0.78 for 'punctate composite' to 0.93 for 'nucleolus' and at low score, performance accuracy ranges from 0.60 for 'cytoplasm' to 0.83 for 'mitochondrion'. It is clear that for all locations, network N1_5 performs better than network N5_5. This demonstrates that PPI score does have influence in predicting subcellular localizations. It can be concluded that higher is the average PPI score better is the NetLoc performance.

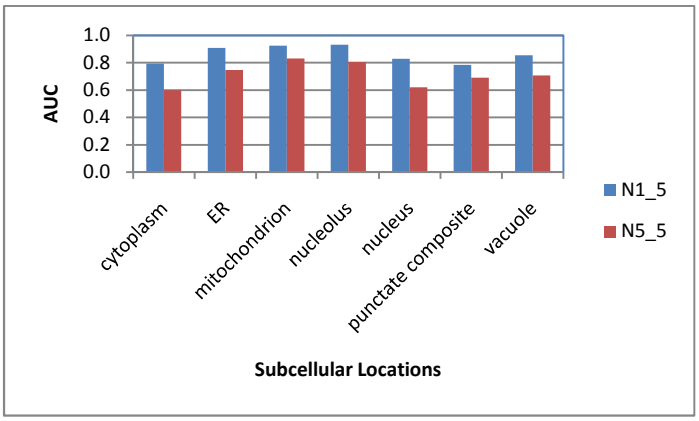

Fig. 3. NetLoc performance in predicting individual subcellular localization

4.2 Overall Performance with Scored PPI Networks

Fig. 4 shows the overall performance for five equally divided networks. It is clear that higher is the PPI score of the network better is the overall performance. This experiment definitely demonstrates that PPI scores have direct impact in predicting subcellular localization. PPI score in the STRING database [13] represents the level of confidence that two proteins can have the similar function. Higher is the score higher is the probability that two proteins are more likely to have the similar function. If two proteins have similar function they are more likely to be localized at the same subcellular compartment [7]. As a result network with high average PPI score produces better results in predicting subcellular localization.

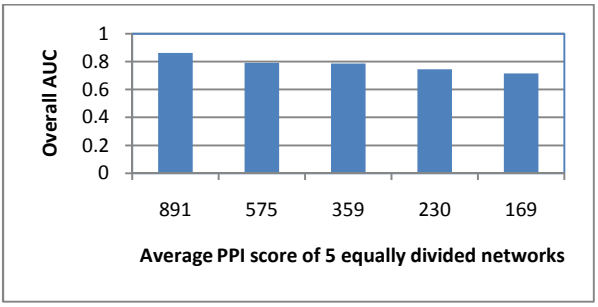

Fig. 3. NetLoc Performance for networks with different PPI scores. Overall AUC is evaluated based on 7 locations

4.3 Comparing Performance with Non-scored PPI Networks

In their previous work [7, 10-12], Mondal and Hu explored protein localization prediction using non-scored PPI networks, namely, COEXP, GPPI, and PPPI. In order to compare we need two networks composed of same number of proteins and same PPIs, one with PPI score and the other without PPI score. As mentioned in section 3.3, these networks can be obtained from the PPIs which are common with STRING PPIs. The derived COEXP contains 8017 PPIs, GPPI contains 85770 PPIs and PPPI contains 45739 PPIs. Fig. 5 shows the model performance for three different types of PPI with and without PPI score. It is evident that improvement upon using PPI score is very small ranging from 1.0% for PPPI to 2.6% for GPPI. This improvement is due to the score attached to each PPI. This proves that network connectivity (prediction without score) provides the most information in predicting protein subcellular localization and inclusion of score features on top of network connectivity have little influence.

4.4 Statistical Significance of Improvement in Prediction due to PPI Score

It is clear from Figs. 3 and 4 that PPI score does have positive impact in predicting protein localization. According to Fig. 5, the improvement in prediction due to score in PPI is very small (1% ~ 3%). Now the question is whether this improvement is statistically significant or it's happening by chance. In Fig. 5, improvement is happening to all three types of PPI. So, it's not happening by chance.

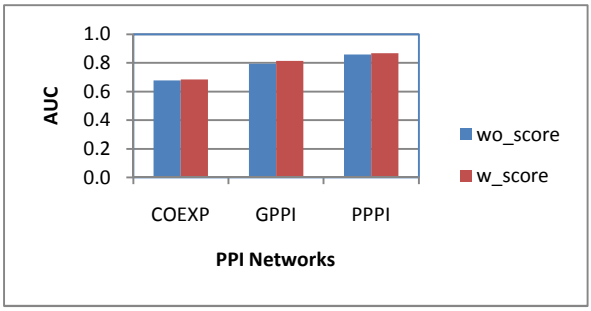

Fig. 4. Overall performance of different types of PPI with and without PPI score

5 Conclusion

A score-based diffusion kernel is introduced in predicting protein subcellular localization using scored protein-protein interaction network. Our investigation shows that PPI scores have direct impact in predicting subcellular localization: higher is the PPI score of the network better is the performance. Our results also show that network connectivity or network with non-scored PPI provides the most information in predicting protein localization. Inclusion of PPI score features on top of network connectivity has little influence.

Acknowledgment. This work was partially supported by NASA grant, Prime Award No: NNX12AI12A, Sub-award No: 520976-Claflin-Mondal and NSF Career Award DBI-0845381.

References

1. Huh, W.K., et al.: Global analysis of protein localization in budding yeast. Nature 425(6959), 686–691 (2003)
2. Agarwal, A.K., et al.: Genome-wide expression profiling of the response to polyene, pyrimidine, azole, and echinocandin antifungal agents in Saccharomyces cerevisiae. J. Biol. Chem. 278(37), 34998–35015 (2003)
3. Hua, S., Sun, Z.: Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17(8), 721–728 (2001)
4. Shen, H.B., Yang, J., Chou, K.C.: Methodology development for predicting subcellular localization and other attributes of proteins. Expert Rev. Proteomics 4(4), 453–463 (2007)
5. King, B.R., Guda, C.: ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes. Genome Biol. 8(5), R68 (2007)
6. Bulashevska, A., Eils, R.: Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinformatics 7, 298 (2006)
7. Mondal, A.M., Hu, J.: NetLoc: Network Based Protein Localization Prediction Using Protein-Protein Interaction and Co-expression Networks. In: IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2010), Hong Kong (2010)

8. Lee, K., et al.: Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. Nucleic Acids Res. 36(20), e136 (2008)
9. Mintz-Oron, S., et al.: Network-based prediction of metabolic enzymes' subcellular localization. Bioinformatics 25(12), i247–i252 (2009)
10. Mondal, A.M., Hu, J.: Network Based Prediction of Protein Localization Using Diffusion Kernel. International Journal of Data Mining and Bioinformatics (2011) (in press)
11. Mondal, A.M., Lin, J., Hu, J.: Network Based Subcellular Localization Prediction for Multi-Label Proteins. In: BIBM-International Workshop on Biomolecular Network Analysis (IWBNA) (2011)
12. Mondal, A.M., Hu, J.: Protein Localization by Integrating Multiple Protein Correlation Networks. In: The 2012 International Conference on Bioinformatics & Computational Biology (BIOCOMP 2012), Las Vegas, USA (2012)
13. von Mering, C., et al.: STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 33(database issue), D433–D437 (2005)
14. Lee, H., et al.: Diffusion kernel-based logistic regression models for protein function prediction. OMICS 10(1), 40–55 (2006)
15. Stark, C., et al.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34(database issue), D535–D539 (2006)
16. Spellman, P.T., et al.: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell. 9(12), 3273–3297 (1998)