Protein Localization by Integrating Multiple Protein Correlation Networks

Ananda Mohan Mondal
Department of Mathematics and Computer Science
Claflin University
Orangeburg, SC 29115, USA
amondal@claflin.edu

Jianjun Hu*
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29208, USA
jianjunh@cse.sc.edu

Abstract

We explored how integration of different protein-protein correlation (PPC) networks improves the performance of a network based classifier, NetLoc, in predicting protein subcellular localization. We investigated different integration approaches and evaluated their performance. Results showed that integration of different PPC networks improves NetLoc performance ranging from 3% to 49% depending on the base networks for integration and the integration approaches.

***Keywords*: protein localization; protein-protein correlation network; network integration; diffusion kernel;PPI network.**

Conference: BIOCOMP2012

1. INTRODUCTION

Literature shows that integrating multiple evidences can greatly improve the prediction accuracy [1-3] of classifiers for predicting protein localization. Wolf-PSort [1] achieved competitive results by combining features from PSORT, iPsort, amino acid content, and sequence length. Drawid and Gerstein [2] proposed a naïve Bayesian classifier to integrate features including motifs, sequence properties, and whole-genome gene expression features. Recently, Scott et al. [3] proposed a two-level Bayesian network approach to integrate information from InterPro motifs, targeting signals, and protein interacting partner relationships.

In our previous work [4], we proposed a network based approach for protein localization prediction. We showed that different protein-protein correlation networks such as physical protein-protein interaction (PPPI), genetic PPI (GPPI), mixed PPI (MPPI), and co-expressed PPI (COEXP) carry different levels of localization information and the performance of the proposed algorithm, NetLoc [4], depends on the topological characteristics such as connectivity and percentages of co-localized PPIs in the network [5]. Figure 1 presents the distribution of PPIs among 4 different networks: PPPI(P), GPPI(G), MPPI(M), and COEXP70(C). Most of the PPIs of each network are not shared by other networks. For example, in PPPI network 43363 out of 50997 PPIs are not shared by other three networks. Similarly, GPPI has 103631 PPIs and 95120 PPIs are not shared by other three networks. So, integration of different networks would change the topological characteristics of the resulting network and may improve the prediction performance.

In the present study, we developed a PPC network based integration framework for protein localization prediction. This method is inspired by the successful application of network integration methods in protein/gene function prediction [6]. Integration of different networks may or may not change the scope of the resulting network from the original networks depending on the integration approach. The scope of a network in the present context is concerned with either the number of proteins in the network or the number of annotated proteins in the network or the number of PPIs in the network. Our objective is to find a unified network, with maximum scope in terms of network proteins and annotated proteins for a species by combining all available networks, which could be used as the standard network for protein localization prediction for that species.

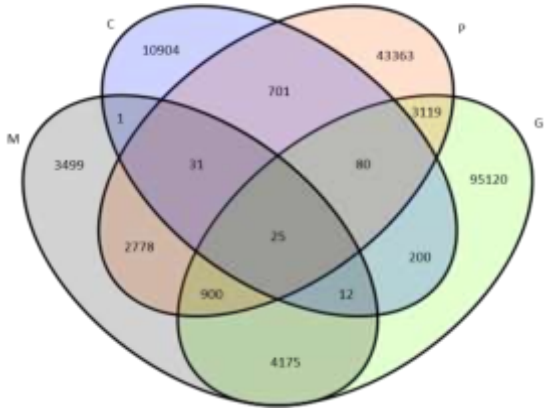


Figure 1. Distribution of PPIs in different networks

2. Unified Network for a Species

Different kinds of PPI networks exist for a species and they provide different level of information for protein localization as mentioned earlier. One reasonable question to ask is how to come up with a unified network for network based classifiers such as NetLoc, which can be then used as the standard network for predicting protein subcellular localization for that species. Before enumerating the properties of the unified network, we define the following terms:

Co-Localized PPIs (coPPIs): PPIs for which both proteins are localized at the same location.

Non-co-localized PPIs (ncPPIs): PPIs for which two proteins are localized at two different locations.

Signal to Noise Ratio (SNR): Ratio of coPPIs to ncPPIs.

Density of coPPI (DCOP): Number of coPPI per annotated protein.

Based on the results presented in [5, 7], criteria for a unified network for protein localization include: i) the network should have high values of SNR and DCOP [7]; ii) the network should have large connected components [5]; iii) the network should have maximum possible scope with respect to the number of proteins and the number of annotated proteins i.e., network with most of the proteins in a genome; and iv) the network should have more coPPIs. The more the coPPIs the better is the network [7].

Answers to the following questions would help in finding the unified network for a species. Question-1: which type of PPIs carries more information about protein localization? Question-2: does removing some PPIs from any network improve the performance? Question-3: how does the integration approach affect the performance? Question-4: which approach should we use to integrate different networks?

3. Data and Methods

3.1 Datasets

We conducted experiments on data sets for Saccharomyces cerevisiae used by Mondal and Hu [4, 5, 7]. Two networks, physical PPI (PPPI) network and genetic PPI (GPPI) network, are obtained from BioGRID [8], mixed PPI (MPPI) network is from MIPS [9] and the co-expression (COEXP) network is from gene expression data of Stanford University [10]. PPPI contains only physical interactions whereas MPPI contains both physical and genetic interactions. MPPI has much less interactions since it has not been updated since 2006.

The localization data of Huh et al. [11] was used as the basis for annotation. The experiment was carried out using high-resolution localization (22 locations) for networks COEXP70, GPPI, MPPI and PPPI. Table 1 shows the summary of the four network datasets used in this study. In terms of the number of interactions, GPPI is the largest network followed by PPPI, COEXP70 and MPPI. Considering the number of proteins, PPPI is the largest network followed by GPPI, MPPI and COEXP70. GPPI is the densest graph, meaning it has the highest values in terms of the average degree of nodes, followed by PPPI, COEXP70 and MPPI. The PPPI network has the largest number of proteins with annotated localization followed by GPPI, MPPI, and COEXP70.

TABLE 1. PPC Networks and Annotation

Property	COEXP70	GPPI	MPPI	PPPI
Number of PPIs	11954	103631	11421	50997
Number of Proteins	2004	5252	4319	5477
Average Degree of Nodes	11.92	39.46	5.28	18.62
Number of Annotated Proteins	1479	3732	3026	3803
Localization	1961	4947	4049	5039

3.2 Integration Approaches

3.2.1 Integration without changing the scope of the base network

In this approach, a network is firstly selected as the base network. Interactions from other networks that fit into the base network are imported to the base network. This integration does not change the scope of the base network in terms of network proteins and annotated proteins. The only changes are the number of PPIs or edges in the integrated network. This integration can be carried out in two different methods. In the first method, all types of PPIs from other networks that fit into the base network are imported and in the second method, only the coPPIs from other networks that fit into the base network are imported. In the second method we are avoiding importing noises or ncPPIs to maintain lower level of noise in the integrated network. For subsequent discussion, the scope of the integrated network in first method is called scope-1 and in second method it is called scope-2.

Table 2 summarizes the network structures before and after integration without changing the scope of the base network in terms of network proteins and annotated proteins. For example, for integration considering MPPI as the base network, the number of network proteins (4319) and annotated proteins (3026) in the resulting integrated network remains the same as the base network. Integration using scope-1 produces a network with 119965 PPIs and scope-2 produces a network with 49066 PPIs. It is clear that integrated network with scope-2 is more connected (more edges or PPIs) than the base network (49066 > 11421) and network with scope-1 is more connected than network with scope-2 (119965 > 49066) as expected.

Table 2 Networks upon integration without changing the scope of the base network

	Proteins		PPIs		
Networks	Network	Annotated	Base	Scope-1	Scope-2
COEXP70	2004	1479	11954	34688	20468
GPPI	5252	3732	103631	157423	124839
MPPI	4319	3026	11421	119965	49066
PPPI	5477	3803	50997	158983	83457

3.2.2 Integration with changing the scope of the base network

The resulting network upon union of two or more networks would have different scopes than the original networks in terms of network proteins and annotated proteins. In general, union of two or more networks would broaden the scope by increasing the numbers of both network proteins and annotated proteins. Two different methods are employed to integrate the networks in union approach. In the first method, the resulting network proteins are union of four original networks and the resulting annotated proteins are union of annotated proteins of four original networks. In the second method, a network is considered as the base and only the coPPIs from other networks are imported where coPPIs are determined based on the resulting annotated proteins found in the first method. In the second method we are avoiding importing noises or ncPPIs to maintain lower level of noise in the integrated network. For subsequent discussion, scope in the first union method is called scope-3 and that in the second is called scope-4.

Table 3 summarizes the network structures before and after integration with changing the scope of the base network in terms of network proteins and annotated proteins. By definition, integrated networks in scope-3 have only one value for each network for each of the network attributes such as network proteins (= 6079), annotated proteins (= 3899), and number of PPIs (=164908). For completeness, the same value is shown for each of the base networks. Integration using both scope-3 and scope-4 increases the scope in terms of network proteins and annotated proteins but the increase is less in scope-4. For example, for COEXP70, network proteins increase from 2004 to 4296 in scope-4 and 2004 to 6079 in scope-3. Similarly, annotated proteins increase from 1479 to 3771 in scope-4 and 1479 to 3899 in scope-3. Scope-4 produces integrated networks of different sizes ranging from 62203 PPIs for COEXP70 to 126807 PPIs for GPPI. In general, integrated networks are more connected (more PPIs or edges) than the base network.

Table 3 Networks upon integration with changing the scope of the base network.

	Network Proteins			Annotated Proteins			PPIs		
Networks	Base	Scope-3	Scope-4	Base	Scope-3	Scope-4	Base	Scope-3	Scope-4
COEXP70	2004	6079	4296	1479	3899	3771	11954	164908	62203
GPPI	5252	6079	5389	3732	3899	3869	103631	164908	126807
MPPI	4319	6079	5132	3026	3899	3839	11421	164908	62375
PPPI	5477	6079	5544	3803	3899	3870	50997	164908	84057

3.3 Classification Algorithm

We applied the diffusion kernel-based logistic regression (KLR) model [12] as used in [4, 5, 7] to predict protein subcellular localization. The KLR model based subcellular prediction problem can be formulated as in [12]. 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 task is to assign subcellular location labels to the n unknown proteins based on the location labels of known proteins and the protein-protein interaction network.

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

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

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 for calculating the similarity distances between two proteins in the network. $I(x_j = 0)$ is an indicator which indicates the interacting protein j does not have the location of interest and $I(x_j = 1)$ indicates that protein j does have the location of interest. Diffusion kernel K, to represent the interaction network, is defined using the following equation.

$$K = e^{\{\tau L\}}$$

Where

$$L(i,j) = \begin{cases} 1 & \textit{if protein i interacts with protein j} \\ -d_i & \textit{if prtoein i is the same as protein j} \\ 0 & \textit{otherwise} \end{cases}$$

Where d_i is the number of interaction partners of protein i, τ is the diffusion constant, and $e^{\{L\}}$ represents the matrix exponential of the Laplacian matrix L. Then the KLR model is given by:

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

which means that the logit of $Pr(X_i = 1 | X_{[-i]}, \theta)$, the probability of a protein targeting a location L is linear based on the summed distances of proteins targeting to L or other location. We then have:

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

The parameters $\gamma, \delta,$ and η can be estimated using the maximum likelihood estimation (MLE) method. Note that here only the annotated proteins are used in the estimation procedure.

Fig. 2 presents the schematic overview of the network-based framework for protein localization prediction using the KLR model by integrating different PPC networks. First, an integrated network is obtained by combining different PPC networks using one of the four scopes. Then diffusion kernel type feature, which is a square matrix consisting of 1 (interaction) and 0 (no interaction), is developed for the integrated network.

Figure 2. Protein localization prediction using the KLR model by integrating PPC networks.

Annotation matrix, which 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, is developed from annotated proteins. KLR model is developed using kernel type features and annotation matrix using logistic regression. The KLR model produces confidences for each protein for all locations. Then a threshold on confidences is used to classify the proteins to be localized at a location or not.

4. Results and Discussion

4.1 Quality of PPI

To answer question-1, we need to find the quality of each type of PPI network, which depends on how much information is carried out by that type in predicting protein localization. There are three fundamental types of PPIs used in the present study – physical PPI, genetic PPI, and co-expressed PPI. In order to determine the quality of different types of PPI, we need to fix the scope of networks with respect to i) number of network proteins (same number of same proteins), ii) number of annotated proteins (same number of same proteins) and iii) number of PPIs (same number of PPIs but different types). Table 4 shows the common proteins among three fundamental networks and the corresponding PPIs in different networks.

Table 4: Numbers of common proteins and corresponding PPIs

Item	COEXP70	GPPI	PPPI
Original PPIs	11954	103631	50997
Original Network Proteins	2004	5252	5477
Original Annotated Proteins	1479	3732	3803
Common Network Proteins	1710	1710	1710
Common Annotated Proteins	1390	1390	1390
PPIs wrt common proteins	9007	12369	10136

It is clear that there are 1710 network proteins and 1390 annotated proteins which are common among three fundamental networks but they have different number of PPIs (COEXP70:9007, GPPI:12369, PPPI:10136). Now, NetLoc performance is determined by selecting a fixed number of PPIs (6000, 7000, 8000, 9000) randomly for each network. For each selection 10 sets of PPIs are selected and then the mean and standard deviation of 10 performances are determined. Table 5 shows the statistics of performances for 10 experiments for each selection of edges. It is clear that for a specific number of edges, performance for 10 experiments are very close for each network since the standard deviations are very small compared to mean values. Figure 3 shows the trend of performance with different types of PPIs. It is clear that for a specific number of edges/PPIs, physical PPI produces the best performance, followed by Co-expressed PPI and then genetic PPI. For example, at edge equal to 7000, AUC values are 0.7696 for PPPI, 0.7295 for COEXP70, and 0.6852 for GPPI. This trend increases with the increase of number of edges in the network. It can be concluded from this experiment that physical PPI has the highest contribution to predicting protein localization followed by co-expressed PPI and then by genetic PPI. So, Physical PPI network could be used as the basis for unified network.

Table 5: Statistics of AUC values for 10 sets

	COEXP70		GPPI		PPPI	
#Edges	Mean	S.D.	Mean	S.D.	Mean	S.D.
6000	0.7211	0.0072	0.6757	0.0089	0.7612	0.0085
7000	0.7295	0.0042	0.6852	0.0084	0.7696	0.0062
8000	0.7374	0.0054	0.6883	0.0077	0.7730	0.0037
9000	0.7460	0.0001	0.6960	0.0066	0.7844	0.0040

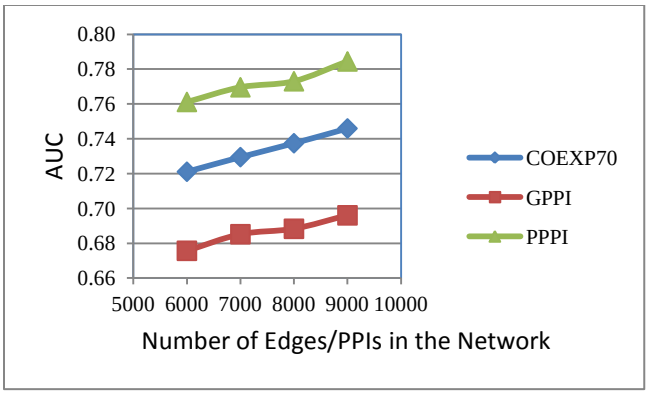
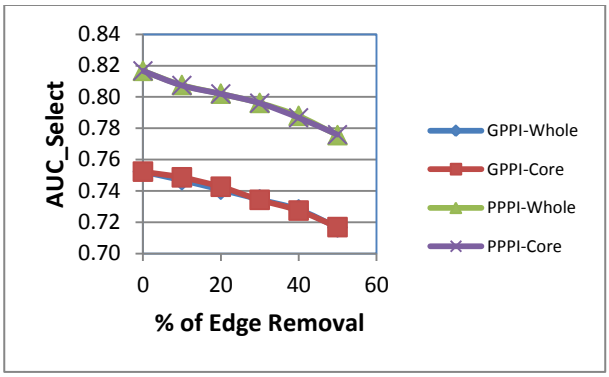


Figure 3. Contribution of PPI types in predicting protein localization.

4.2 Effect of Removing Some Interactions

Both GPPI and PPPI are composed of only one connected component, table 4 of [4]. Any of these two networks could be a good candidate as the basis of a unified network. GPPI (network proteins = 5252, annotated proteins = 3732, PPIs = 103631) is the densest network or it has too many PPIs. On the other hand PPPI (network proteins = 5477, annotated proteins = 3803, PPIs = 50997) has less PPIs (about 50% of GPPI) and lower annotation coverage (69.44% < 71.06%). But PPPI produces better results than GPPI (AUC: 0.82 > 0.75), figure 2 of [4]. This suggests that for a unified network, we may not need too many interactions. Then question arises, does removing some PPIs from GPPI network improve the performance (Question-2)? Removing edges from the whole network makes some of the proteins isolated from the network, specially, proteins with single-degree of interaction. These single-degree proteins are located at the edge of the network. In order to avoid producing isolated proteins, removal is also carried out from the core of the network. The core for the present study is composed of proteins with at least degree equal to 4.

Figure 4 shows the performance after removing edges from the whole network and from the core for both GPPI and PPPI networks. It is clear that removal of edges deteriorates the performance for both networks. But there is hardly any difference in performance in two different removal approaches. This suggests that for a unified network, we should not remove any edges or PPIs from any network.



Whole: *represents removal from the whole network*
Core: *represents removal from the core of the network*

Figure 4. Effect of edge removal.

4.3 Effect of Changing the Scope of the Base Network

Table 6 summarizes the performance of integrated networks without changing the scope of the base network considering all locations (22 locations). It is clear that NetLoc performance significantly improves upon network integration. Using scope-1, performance improvement ranges from 3% for PPPI to 28% for COEXP70 and using scope-2, it ranges from 10% for PPPI to 36% for COEXP70. Two main reasons for improvement are- (i) each network becomes more connected (more edges) upon integration (Table 2) and (ii) increase in values for either SNR or DCOP or both. In our earlier study, we showed that NetLoc performance improves with the increase of SNR and DCOP [7]. In integration using scope-1, values of SNR for some integrated networks are slightly decreased from the corresponding base network but values for DCOP are significantly increased for each of the integrated networks compare to base networks, which in turn improve the performance of integrated networks. For a specific base network, values of DCOP for integrated networks using both scope-1 and scope-2 remain the same (14.16 for PPPI) but value of SNR in scope-2 (3.869 for PPPI) is significantly higher than that in scope-1 (0.987 for PPPI). As a result, scope-2 produces better results than scope-1 in general.

Table 6 NetLoc performance upon integration without changing the scope of the base network

	SNR			DCOP			AUC_All			Improve	
Networks	Base	Scope-1	Scope-2	Base	Scope-1	Scope-2	Base	Scope-1	Scope-2	Scope-1	Scope-2
COEXP70	1.451	1.147	4.388	2.84	8.60	8.60	0.6407	0.8229	0.8728	28%	36%
GPPI	0.806	0.959	1.352	8.38	14.06	14.06	0.7851	0.8813	0.9086	12%	16%
MPPI	0.996	0.961	11.709	1.16	13.60	13.60	0.7132	0.8692	0.9496	22%	33%
PPPI	1.537	0.987	3.869	5.63	14.16	14.16	0.8525	0.8787	0.9401	3%	10%

4.4 Effect of Changing the Scope of the Base Network

Table 7 summarizes the performance of integrated networks with changing scope of the base network considering all locations (22 locations). It is clear that NetLoc performance also significantly improves upon network integration with changing scope of the base network. Using scope-3, performance improvement ranges from 3% for PPPI network to 37% for COEXP70 and using scope-4, it ranges from 10% for PPPI to 49% for COEXP70. As explained earlier, the improvement in the performance is due to increase either in SNR or DCOP or in both. For example, for base network COEXP70, integration using scope-3 decreases SNR from 1.451 to 0.973 but increases DCOP significantly from 2.84 to 13.97, which in turn improves the performance from 0.6407 to 0.8809. In integration using scope-4, a significant increase happened to both SNR (from 1.451 to 18.784) and DCOP (from 2.84 to 14.44), which results in huge improvement in performance from 0.6407 to 0.9562.

Table 7 NetLoc performance upon integration with changing the scope of the base network

	SNR			DCOP			AUC_All			Improve	
Networks	Base	Scope-3	Scope-4	Base	Scope-3	Scope-4	Base	Scope-3	Scope-4	Scope-3	Scope-4
COEXP70	1.451	0.973	18.784	2.84	13.97	14.44	0.6407	0.8809	0.9562	37%	49%
GPPI	0.806	0.973	1.402	8.38	13.97	14.07	0.7851	0.8809	0.9041	12%	15%
MPPI	0.996	0.973	15.497	1.16	13.97	14.18	0.7132	0.8809	0.9565	24%	34%
PPPI	1.537	0.973	3.913	5.63	13.97	14.07	0.8525	0.8809	0.9351	3%	10%

4.5 Identifying Unified Network

Figure 5 presents the performance of integrated networks using four different scopes compare to base network. It is clear that integration improves performance in all methods of integration. Now the question is which integrated network should we select as the unified network or which approach should we use for integration (Question-4).

Unified Network based on Performance

Integration using Scope-2 produces better performance than scope-1 for all networks since scope-2 comes with better signals (relatively more co-localized PPIs) than scope-1. Similarly, scope-4 produces better performance than scope-3 for all networks. Considering performance, integrated networks using scope-2 and scope-4 are possible candidates for unified network. Out of 8 integrated networks, integration using scope-4 with base network MPPI produces the best performance of AUC = 0.9565 (Figure 5). So, integrated network obtained from MPPI network using scope-4 can be considered as the unified network.

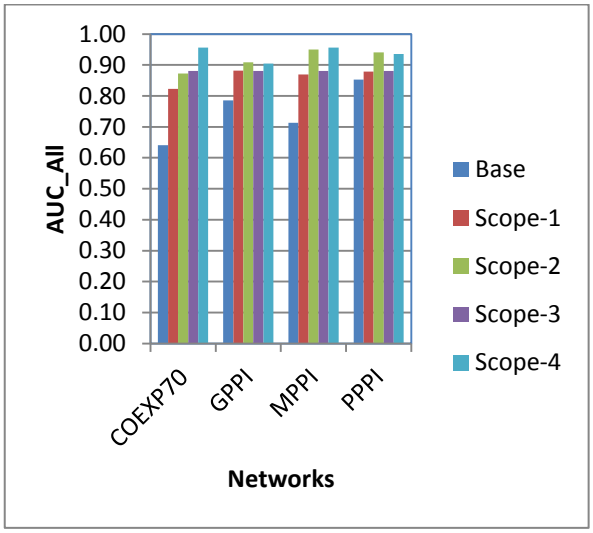


Figure 5. NetLoc performance upon integration with different scopes.

Unified Network based on Scope

Integration using scope-1 and scope-2 has the minimum scope, which is the same as base network, in terms of both network proteins (Figure 6) and annotated proteins (Figure 7) for each of the base networks. Integration using scope-3 has the maximum scope in terms of both network proteins (Figure 6) and annotated proteins (Figure 7), which are same for each of the base network. Integration using scope-4 has the intermediate scope in terms of both network proteins (Figure 6) and annotated proteins (Figure 7) for each of the base networks. So, considering scope, integrated network using scope-3 can be used as the unified network.

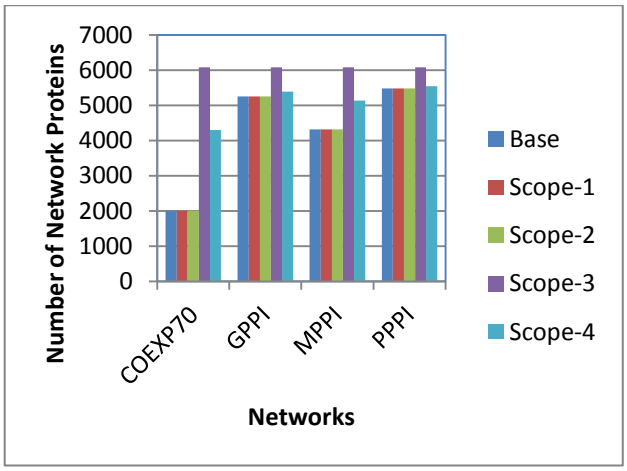


Figure 6 Network proteins upon integration with different scopes.

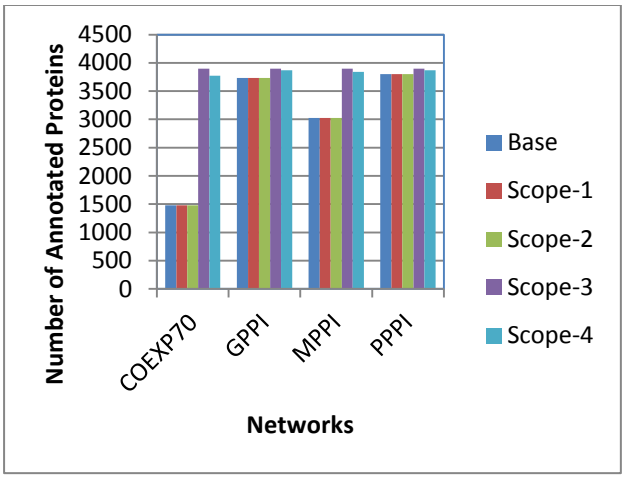


Figure 7 Annotated proteins upon integration with different scopes.

Balanced Unified Network

Unified networks based on performance and on scope represent networks based on two extremes. The first unified network produces a maximum performance of AUC = 0.9565 with a scope of 5132 network proteins and 3839 annotated proteins. The latter produces a performance of AUC = 0.8809 with the maximum scope of 6079 network proteins and 3899 annotated proteins. A balanced unified network is the one that provides a balance between performance and scope. Considering base PPPI, integration using scope-4 achieved a performance of AUC = 0.9351 with a scope of 5544 network proteins and 3870 annotated proteins. This network has both scope and performance in between the two unified networks based on two extremes. So, integrated network obtained from PPPI network using scope-4 can be considered as the balanced unified network for predicting protein localization. The overall performance (AUC = 0.9351) is improved by 10% over the individual best performance (AUC = 0.8525) with base network PPPI. This proves our hypothesis that the unified network should be based on high quality network which is physical PPI in the present study (Figure 3).

5. Conclusion

Different kinds of integration approaches are explored to observe the influence of integrating different PPC networks on the performance of a classifier, NetLoc, to predict protein localization. Our results showed that integration of different networks significantly improves NetLoc performance. This study also showed that physical PPI has the highest contribution to predicting protein localization followed by co-expressed PPI and by genetic PPI. Finally, we proposed a balanced unified network based on performance and scope of the integrated networks, and we found that the balanced unified network is based on a base network with the best quality, which is physical PPI.

Acknowledgment

This work is partially supported by NSF Career Award DBI-0845381, HBCU-UP grant HRD-0713853, and Center for Excellence in Teaching of Claflin University.

[1] P. Horton, et al., " WoLF PSORT: protein localization predictor," Nucleic Acids Research, vol. 35, pp. W585-W587, 2007.

[2] A. Drawid and M. Gerstein, "A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome," J Mol Biol, vol. 301, pp. 1059-75, Aug 25 2000.

[3] M. S. Scott, et al., "Refining protein subcellular localization," PLoS Comput Biol, vol. 1, p. e66, Nov 2005.

[4] A. M. Mondal and J. Hu, "NetLoc: Network Based Protein Localization Prediction Using Protein-Protein Interaction and Co-expression Networks," in IEEE International Conference on Bioinformatics & Biomedicine (BIBM2010), Hong Kong, 2010, pp. 142-148.

[5] A. M. Mondal and J. Hu, "Network Based Prediction of Protein Localization Using Diffusion Kernel," International Journal of Data Mining and Bioinformatics 2011.

[6] T. Hawkins, et al., " New paradigm in protein function prediction for large scale omics analysis " Mol. Biosyst., vol. 4, pp. 223-231, 2008.

[7] A. M. Mondal, et al., "Network Based Subcellular Localization Prediction for Multi-Label Proteins," in BIBM-International Workshop on Biomolecular Network Analysis (IWBNA), 2011.

[8] C. Stark, et al., "BioGRID: a general repository for interaction datasets," Nucleic Acids Res, vol. 34, pp. D535-9, Jan 1 2006.

[9] U. Guldener, et al., "MPact: the MIPS protein interaction resource on yeast," Nucleic Acids Res, vol. 34, pp. D436-41, Jan 1 2006.

[10] P. T. Spellman, et al., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," Mol Biol Cell, vol. 9, pp. 3273-97, Dec 1998.

[11] W. K. Huh, et al., "Global analysis of protein localization in budding yeast," Nature, vol. 425, pp. 686-91, Oct 16 2003.

[12] H. Lee, et al., "Diffusion kernel-based logistic regression models for protein function prediction," OMICS, vol. 10, pp. 40-55, Spring 2006.