Network Based Subcellular Localization Prediction for Multi-Label Proteins

Ananda Mohan Mondal[1,2], Jhih-rong Lin[1], and Jianjun Hu[1,*]
[1]*Machine Learning and Evolution Laboratory*
Department of Computer Science and Engineering, University of South Carolina, SC 29208, USA.
[2]*Department of Mathematics and Computer Science, Claflin University, SC 29115, USA*
ammondal@cec.sc.edu: jianjunh@cse.sc.edu

Abstract

Many proteins are sorted to multiple subcellular localizations within the cell. However, computational prediction of multi-location proteins remains a challenging task. Here we applied a logistic regression and diffusion kernel based algorithm NetLoc for predicting multiplex proteins and explored its capability and limitations. Experiment shows that the overall and true success rates for physical protein-protein interaction network are 65% and 41% respectively, and for mixed PPI network these values are 88% and 75% respectively. Our study also showed that the performance of NetLoc in predicting protein localization is limited by the network characteristics such as ratio of the number of co-localized protein-protein interactions (coPPI) to the number of non-co-localized PPI (ncPPI) and the density of annotated coPPI in the network. For a given network with a specific number of proteins, NetLoc performance increases with higher coPPI/ncPPI ratio and higher density of annotated coPPI.

1. Introduction

With increasing number of genes are sequenced, computational predictions of protein localizations can greatly help to infer their functions. However, experimental determination of protein localization is costly [1;2]. In the past decade, many algorithms have been developed for computational prediction of protein subcellular locations [3] using a variety of supervised machine learning techniques including neural networks [4], nearest neighbor classifier, Markov models, Bayesian networks [5;6], expert rules, meta-classifiers, and the support vector machines [7;8].

Most of these current protein localization prediction algorithms focus on single-location proteins and relatively much less effort has been made to address proteins which are localized to multiple subcellular locations. Chou and Cai [9] first attempted to classify multiplex proteins in yeast with a hybridized model including gene ontology, functional domain and pseudo-amino acid composition. In their method, they did not use any threshold for the predictor for multiplex protein localization. They considered three rankings in evaluating overall success rate: i) rank I – considers the prediction with the highest score, ii) rank II – considers the predictions with two highest scores, and iii) rank III – considers the predictions with three highest scores. Yang and Lu [10] developed a multi-label classifier using SVM to predict multiplex proteins using amino acid compositions alone. Chou and Shen developed Euk-mPLoc [11] a fusion classifier for Eukaryotic protein localization for multiplex protein established by hybridizing the gene ontology approach and pseudo amino acid composition approach. They developed an improved version, Euk-mPLoc 2.0 [12], by incorporating functional domain information with the previous model, Euk-mPLoc. Finally, Chou et al. [13] developed multi-label K-nearest neighbor classifier called iLoc-Euk for predicting subcellular protein localization for Eukaryotic proteins.

Recently, protein-protein correlation (PPC) networks have been used for localization prediction. Lee et al. [14] used PPI networks for localization prediction by deriving some network-specific features combined with other traditional features such as amino acid composition. This method however only used limited information (neighbor proteins) of the network. Mintz-Oron et al. [15] used metabolic networks for localization prediction using constraint-based models. We [16] applied protein-protein interaction (PPI) network for single-location protein localization prediction using diffusion kernel based NetLoc algorithm.

In this study, first we explored the effect of coPPI (co-localized PPI) and ncPPI (non-co-localized PPI) on the prediction performance of NetLoc and identified two major factors, namely, SNR (signal to noise ratio i.e. coPPI to ncPPI) and DCOP (density of annotated coPPI). Actually, the values of these two factors determine the quality of a network in predicting protein localization. Second, we extended NetLoc for predicting multiplex protein localization by introducing a method to determine the number of predicted locations. Finally, we showed how to improve the NetLoc performance in predicting multiplex protein localization by increasing the factors SNR and DCOP. We applied NetLoc to predict localization of genome wide yeast proteins using the PPI and COEXP

networks. In a leave-one-out cross-validation test of predicting known subcellular localization of 3803 proteins of Yeast both mono-locational and multi-locational, NetLoc is shown to achieve high overall success rate of 88%.

2. Diffusion kernel-based logistic regression for protein localization prediction

2.1. Motivation

Most of current protein subcellular localization prediction algorithms are developed using feature based methods using sequence information, gene ontology or physichemical properties. However, one limitation of these methods is that it is not easy to exploit rich network information that naturally appears among proteins such as protein-protein interaction networks and gene co-expression network. Another issue of current protein localization algorithms is the lack of capability to predict multi-location proteins. Most researchers explicitly remove these proteins in their data preprocessing steps before training their prediction algorithms.

The main idea of our approach is to utilize the information of protein-protein correlation network structure for predicting the localization of un-annotated proteins. This network can be based on protein-protein interaction, PFAM domain interaction, co-expressed gene interaction, genetic interaction, etc. The reason is interacting proteins tend to be localized to the same subcellular locations. Thus, the localizations of neighboring proteins in the PPI network carry some information about the localization of the un-annotated proteins. For example, if most of the neighbors of a protein have the same localization; it is more likely that the protein is also localized to the same location. A confidence or probability about the fact that the protein is localized at a certain location can be determined. Finally, the localization labels will be assigned to un-annotated proteins based on some threshold on confidence value.

The confidence of a protein to be localized at a specific location can be determined using two different approaches: a) considering only the localization information of the direct neighbors and b) considering the localization information of all the proteins in the network. First approach uses Markov Random Field (MRF) model to solve the problem. To solve the problem in second approach, diffusion kernel-based logistic regression (KLR) model is suitable. Literature shows that the KLR model performs better than MRF model [17].

2.2. KLR logistic regression model

We applied the diffusion kernel-based logistic regression (KLR) model [17] to predicting protein subcellular localization based on the locations of all other proteins within protein-protein correlation networks. This method has the unique advantage of considering the subcellular location labels of all the related proteins.

The KLR model based subcellular prediction problem can be formulated as follows [17]. 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 distances between two proteins in the network that have the same localization. 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.

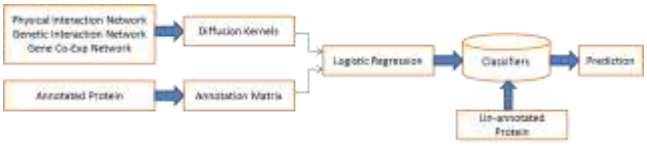

Figure 1. Protein localization prediction using the KLR model and protein networks

Figure 1 presents the schematic overview of the network-based framework for protein localization prediction using the KLR model and protein networks. Diffusion kernel type feature, which is a square matrix consists of 1 (interaction) and 0 (no interaction), is developed for each of the 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.

2.3. Determination of the number of predicted locations

In the localization prediction of multi-location proteins, usually a threshold is needed to determine the number of predicted subcellular locations. In NetLoc, given a query protein, a probability value is calculated for each location indicating the confidence that the protein is localized to that location. Thus 22 probability scores will be calculated when 22 different locations are considered. A cutoff probability score is needed to determine the number of predicted locations. Since the probabilities for the top K locations of different proteins are different, it is not suitable to use an absolute probability value as the threshold. In this study, we first normalized the 22 probabilities for a protein by dividing them by the largest value of the 22 probabilities. Then, a value between 0 and 1 is selected as the cutoff threshold for determining the number of predicted locations for a given protein based on the overall prediction performance for a given network. If the normalized probability for a predicted location is greater than the threshold, it is reported as a valid predicted location. The number of valid predicted locations is thus determined.

3. Results and discussion

3.1. Datasets

We conducted experiments on data sets for Saccharomyces cerevisiae used by Mondal and Hu [16]. Two networks, physical PPI (PPPI) network and genetic PPI (GPPI) network, are obtained from BioGRID (Stark et al., 2006), mixed PPI (MPPI) network is from MIPS (Guldener et al., 2006) and the co-expression (COEXP) network is from gene expression data of Stanford University (Spellman et al., 1998). 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. [1] 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. Only the PPPI network was used for low-resolution localizations (5 locations) and was denoted as PPPI5. The five locations in low-resolution are: i) cytoplasm, ii) mitochondrion, iii) nucleus (consists of 3 locations: nucleus, nucleolus, and nuclear periphery), iv) secretory (consists of 9 locations: cell periphery, early Golgi, endosome, ER, ER to Golgi, Golgi, late Golgi, vacuolar membrane, and vacuole), and v) others (consists of 8 locations: actin, bud, bud neck, lipid particle, microtubule, peroxisome, punctate composition, and spindle pole) (Blum et al., 2009, Lodish et al., 2000).

Table 1. PPI networks and annotation

Property	COEXP70	GPPI	MPPI	PPPI	PPPI5
Number of PPIs	11954	103631	11421	50997	50997
Number of Proteins	2004	5252	4319	5477	5477
Degree of Nodes	11.92	39.46	5.28	18.62	18.62
Number of Annotated Proteins	1479	3732	3026	3803	3803
Localization	1961	4947	4049	5039	4854

Table 1 shows the summary of the five network datasets used in this study. In terms of the number of interactions, GPPI is the largest network followed by PPPI, COEXP70 and MPPI. When considering the number of proteins, PPPI is the largest network followed by GPPI, MPPI and COEXP70. GPPI is the densest graph followed by PPPI, COEXP70 and MPPI. PPPI network has the largest number of proteins with annotated localization followed by GPPI, MPPI, and COEXP70. The only difference between PPPI and PPPI5 is that the later has less number of localizations.

Table 2 shows the distribution of multi-localized proteins. For example, in PPPI network, out of 3803 annotated proteins, 2647 target 1 location, 1085 target 2 locations, 63 target 3 locations, 7 target 4 locations, and 1 targets 5 locations.

Table 2. Distribution of multi-localized proteins

Number of Locations	COEXP70	GPPI	MPPI	PPPI	PPPI5
1	1029	2598	2072	2647	2787
2	421	1062	892	1085	982
3	27	64	56	63	33
4	1	7	5	7	1
5	1	1	1	1	-
Total Proteins	***1479***	***3732***	***3026***	***3803***	***3803***

3.2. Effect of coPPI and ncPPI on NetLoc performance

In our previous study (Mondal and Hu 2010), we showed that NetLoc performance depends on network topology features such as network connectivity and degree of interactions, and the percentage of co-localized PPI (coPPI) (Table 5 of Mondal and Hu, 2010). Here we explore the behavior of the model with respect to not only co-localized PPI but also non-co-localized PPI (ncPPI) in a

network. Studies have shown that the probability that a pair of interacting proteins have the same function is higher than the probability that they have different functions (Schwikowski *et al.* 2000). The same analogy is applicable to protein localization, which implies that more coPPIs may increase the performance on network based localization prediction.

In order to check the effect of coPPI and ncPPI on the NetLoc performance, we evaluated their performance by removing all the coPPI or ncPPI from the network respectively and compared them with the performance of the original network. Performance was evaluated for selected locations as used in (Mondal and Hu, 2010) using 5-fold cross-validation. Figure 2 summarizes the results without coPPI, without ncPPI and with all PPIs for different networks.

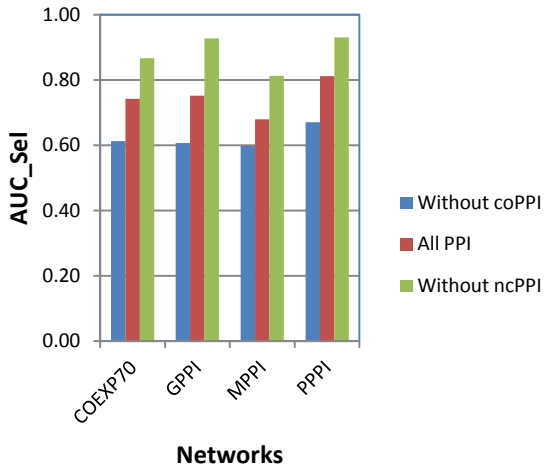

Figure 2. NetLoc performance without coPPI and without ncPPI compare to original network

Without coPPI, NetLoc performance deteriorates 12% for MPPI and 19% for GPPI compared to that of the original network respectively. When ncPPIs are removed, the performance improves from original network by 15% for PPPI and 23% for GPPI. Thus, NetLoc performance for a network depends on the coPPI/ncPPI ratio, a signal to noise ratio (SNR). Usually for a given network, the higher the value of SNR, the better the performance. Another factor that may affect the prediction performance is the density of coPPI (DCOP) measured by the number of coPPI per annotated protein. For a network, the higher, the value of DCOP, the better the performance. Table 3 shows the NetLoc performance considering all PPIs for different networks along with the corresponding SNR and DCOP values.

A network with high values of both SNR and DCOP would perform better, for example PPPI has high value of SNR (= 1.537) and DCOP (= 5.63) and the NetLoc performance is also high (AUC = 0.8167). When SNR values are similar, higher DCOP will ensure better performance. For example, the value of SNR for GPPI (SNR = 0.806) is less than that of MPPI (SNR = 0.996) but GPPI performs better (AUC = 0.7523) than MPPI (AUC = 0.6858) due to its very high value of DCOP (8.38) compared to MPPI (DCOP = 1.16).

Table 3. NetLoc performance and corresponding SNR and DCOP values

Network	DCOP	SNR	AUC_Sel
COEXP70	2.84	1.451	0.7489
GPPI	8.38	0.806	0.7523
MPPI	1.16	0.996	0.6858
PPPI	5.63	1.537	0.8167

3.3. NetLoc as multiplex protein localization predictor

Performance Evaluation

Leave-one-out cross-validation was used to evaluate the performance of NetLoc for predicting both monoplex and multiplex protein localization. Two measurements are defined to evaluate the performance: Overall success rate and true success rate.

Overall success rate vs. true success rate: In evaluating overall success rate, a 2-location protein is considered as 2 instances of a single-location protein, a 3-location protein as 3 single-location proteins and so on. For example, a protein has three locations such as cytoplasm, nucleus, and ER: if the predicted locations are cytoplasm, ER, and vacuole then overall success rate is 2/3 whereas true success rate is zero, which is defined as the number of proteins of which the predicted locations matches exactly with the true locations. True success rate is much stricter than the overall success rate. If a prediction is under-prediction or over-prediction, true success rate is zero in either case (If the number of predicted locations is more than the number of true locations, it is called over-prediction. Conversely, it is called under-prediction).

Since the prediction performance of NetLoc on multiplex proteins depends on the selection of threshold probability for a given network, we evaluate the overall success rate and true success rate based on two types of prediction: a) top-k prediction and b) optimal prediction.

Success rates based on top-k prediction

In top-k prediction, i) for 1-locational protein top-1 prediction is considered, ii) for 2-locational protein top-2 predictions are considered and so on. This ideal situation is based on the assumption that the number of true locations is already known. In order to see how SNR and DCOP play important role, the success rates for selected locations, i.e. locations with more than 100 proteins are shown in Table

4. Cytoplasm (SNR = 0.50, DCOP = 5.73) and nucleus (SNR = 0.50, DCOP = 6.75) have higher values for both SNR and DCOP compared to other locations, thus producing higher success rates of 94.86% and 86.39% respectively. Punctate composite (SNR = 0.05, DCOP = 0.96) and vacuole (SNR = 0.04, DCOP = 0.35) have very low values for both SNR and DCOP, thus producing very low success rates of 8.09% and 1.96% respectively. Success rate for nucleolus (53.05%) with SNR equal to 0.30 is higher than mitochondrion (36.63%) with SNR equal to 0.38 because the nucleolus has much higher DCOP (7.82) than mitochondrion (1.89).

Table 4. Success rates for individual locations and corresponding SNR and DCOP values

Locations	Protein	coPPI	ncPPI	DCOP	SNR	Success
cell periphery	128	244	1619	1.91	0.15	24.22%
cytoplasm	1731	9917	20022	5.73	0.50	94.86%
ER	283	886	2164	3.13	0.41	31.80%
mitochondrion	486	920	2415	1.89	0.38	36.63%
nucleolus	164	1283	4259	7.82	0.30	53.05%
nucleus	1411	9521	19228	6.75	0.50	86.39%
punctate composite	136	131	2386	0.96	0.05	8.09%
vacuole	153	53	1273	0.35	0.04	1.96%

It is evident from above discussion that the NetLoc performance is mostly influenced by two factors: SNR and DCOP. For a network, if one of these values becomes higher it can perform better, for example PPPI and PPPI5 have the same number of PPI and annotated proteins but PPPI considers high-resolution 22 locations and PPPI5 considers low-resolution 5 locations. In the latter case, low-resolution location 'nucleus' consolidates 3 high-resolution locations; 'secretory' consolidates 9 high-resolution locations, and 'others' consolidates 8 high-resolution locations. Because of this consolidation the number of coPPI increases which in turn increases the value of SNR from 1.537 to 2.093 and DCOP from 5.63 to 6.28 (Table 5). This improves the overall success rate from 68.96% to 73.90%. Among four networks with 22 localizations, COEXP70 performs the worst and PPPI performs the best. It is clear that PPPI has high values for both SNR and DCOP, and as a result, it performs better. It is noticeable that COEXP70 has higher values for SNR and DCOP than those of MPPI, but COEXP70 performs worse than MPPI. This is because MPPI is more connected (75 connected components) than COEXP70 (136 connected components) as mentioned in table 4 of Mondal and Hu (2010). GPPI has lower SNR than MPPI (0.806 < 0.996), but performs better than MPPI. This is because GPPI has much higher value of DCOP than MPPI (8.38 >> 1.16). The values of SNR and DCOP play the similar role for true success rate also. The value of SNR can be increased by removing some ncPPI from the network which in turn would improve the success rate, which is explored later.

Table 5. Success rates for different networks and corresponding SNR and DCOP values

			Success Rate	
Network	**DCOP**	**SNR**	**Overall**	**True**
COEXP70	2.84	1.451	57.78%	48.48%
GPPI	8.38	0.806	60.12%	51.02%
MPPI	1.16	0.996	58.98%	49.41%
PPPI	5.63	1.537	68.96%	61.82%
PPPI5	6.28	2.093	73.90%	67.21%

Overall success rate based on top-k prediction for the best network i.e., PPPI (success rate = 68.96%), is higher than two existing predictors i) Euk-mPLoc (success rate = 39.26%) and ii) Euk-mPLoc 2.0 (success rate = 64.17%). From the top-k success rate it is clear that NetLoc is a good candidate for multi-label protein localization prediction. Now the question is what threshold on normalized prediction probability should be used for NetLoc to be a predictor for multi-label protein localization. Following section discusses how a threshold can be selected at the optimum prediction quality for a network.

3.4. Thresholds for optimum prediction quality

In practice, the number of true locations of a query protein is not known and a threshold value is needed to determine the number of predicted locations for each query protein. An optimal threshold is one that can achieve the best overall prediction performance.

Let $L_p = \{l_1, l_2, \cdots l_m\}$ represents the set of m different subcellular predicted locations for a protein by NetLoc using threshold ω and $L_t = \{l_1, l_2, \cdots l_n\}$ forms the set of n true subcellular locations for the same protein. Now define a quality control function as used in (Chou and Shen, 2007) for the protein P as:

$Q_P(\omega) = H_s - H_m^o$

Where H_s represents the number of successful hits and H_m^o represents the number of miss-hits and over-hits using NetLoc in predicting the localization for the query protein, and these can be formulated as:

$$H_s = \|L_p \cap L_t\|$$

$$\begin{aligned} H_m^o &= \|L_p \cup L_t\| - \|L_p \cap L_t\| \\ &= \|L_p\| + \|L_t\| - 2\|L_p \cap L_t\| \\ &= m + n - 2\|L_p \cap L_t\| \end{aligned}$$

Now, $\|L_p \cap L_t\|$ can be evaluated as:

$$\|L_p \cap L_t\| = \sum_{i}^{m} \delta_i(L_p^i, L_t)$$

Where L_p^i represents the i^{th} component of L_p and the delta function is given by

$$\delta_i(L_p^i, L_t) = \begin{cases} 1, & \text{if } L_p^i \in L_t \\ 0, & \text{if } L_p^i \notin L_t \end{cases}$$

Finally, the quality control function can be expressed as

$$Q_P(\omega) = 3\sum_i^m \delta_i(L_p^i, L_t) - (m+n)$$

The overall quality control function for the NetLoc is given by

$$\mathbb{Q}(\omega) = \sum_{P \text{ known}} Q_P(\omega)$$

And the optimal value for ω is given by

$$\omega^* = \arg\max_\omega \{\mathbb{Q}(\omega)\}$$

Figure 3 shows the overall prediction quality with thresholds for different networks. Other than COEXP70, each network has one maximum (optimum) quality value. COEXP70 has two optima at threshold values 0.83 and 0.45. It is noticeable that each network maintains the same level of true success rate for the thresholds from 1 down to the optimum threshold for the network (Figure 4). For COEXP70, this trend is maintained from 1 down to the threshold at the 1st optimum, 0.83. So, the optimum threshold equal to 0.83 for COEXP70 is used for next section.

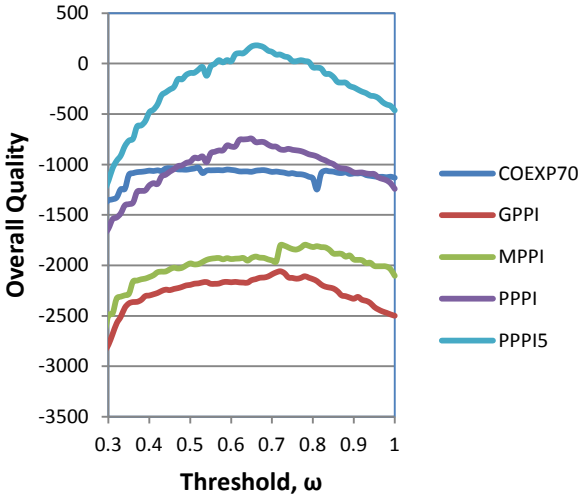

Figure 3. Overall prediction quality as a function of threshold for original networks

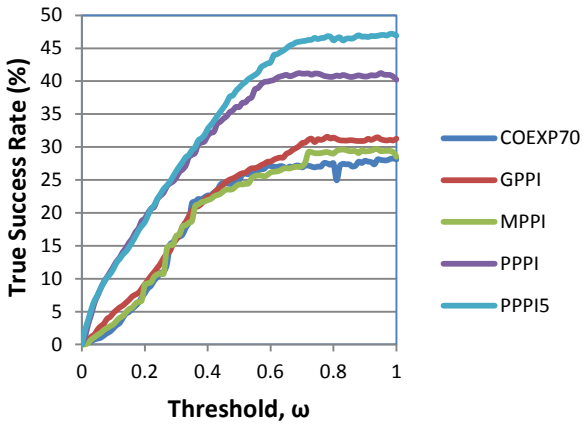

Figure 4. True success rate as a function of threshold for original networks

Table 6 summarizes the results based on optimum quality. Since the four networks with 22 localizations have different sizes in terms of the number of proteins and the number of PPIs, it is not possible to compare them in terms of optimum quality. PPPI and PPPI5 have the same size and PPPI5 has higher values for controlling factors than PPPI (SNR: 2.093 > 1.537; DCOP: 6.28 > 5.63). As a result PPPI5 has a higher prediction quality than PPPI (181 > -744) and higher success rate (73.09% > 65.09%). The suggested values for threshold for different networks based on optimum quality are: COEXP70→ 0.83, GPPI→ 0.72, MPPI→ 0.72, PPPI→ 0.65, and PPPI5→ 0.66.

Table 6. Thresholds and success rates based on optimum quality

Network	COEXP70	GPPI	MPPI	PPPI	PPPI5
Optimum quality	-1063	-2058	-1803	-744	181
Optimum threshold	0.83	0.72	0.72	0.65	0.66
True success (%)	27.45	31.08	29.21	40.94	45.41
Overall success (%)	45.89	58.60	55.64	65.09	73.09

3.5. Improving success rates by combining different networks

It is clear from the earlier discussion that PPPI is the best network in terms of quality of predicting multiplex proteins considering 22 locations (Table 6) and NetLoc performance improves with the increase of SNR and DCOP. There are two different ways that can be employed to increase the value of SNR and DCOP for PPPI network: a) by importing coPPIs from other three networks (COEXP70, GPPI, MPPI) into PPPI network, and b) by removing ncPPI from the resulting PPPI network in (a).

The resulting augmented networks are named as PPPI3CO and PPPI3CONOR respectively.

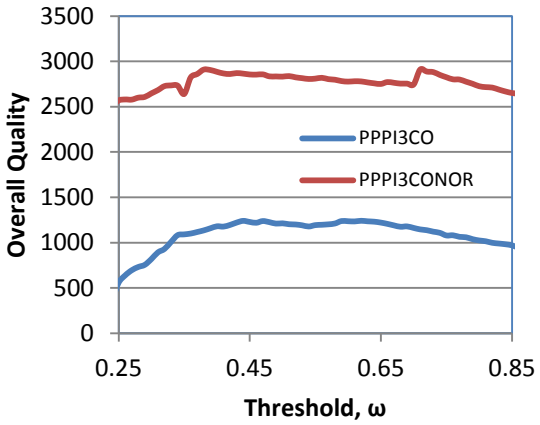

Figure 5. Overall prediction quality for augmented PPPI networks

Figure 5 shows the overall prediction quality for the two augmented PPPI networks. There are two optimal thresholds, namely 0.44 and 0.62 for PPPI3CO and 0.38 and 0.72 for PPPI3CONOR (Table 7). Variations in true success rates between two optimum thresholds are 2% for PPPI3CO and 0.50% for PPPI3CONOR which are insignificant (Figure 6). So, the lower value of two optimum thresholds can be used as the working thresholds.

Table 7. Thresholds and success rates for PPPI and augmented PPPI networks

Network	PPPI	PPPI3CO	PPPI3CONOR
Optimum Quality	-744	(1240, 1241)	(2909, 2887)
Optimum Threshold	0.65	(0.44, 0.62)	(0.38, 0.72)
True Success (%)	40.94	(56.96, 59.08)	(74.65, 75.10)
Overall Success (%)	65.09	(81.37, 75.47)	(88.43, 83.09)
DCOP	5.63	10.82	10.82
SNR	1.537	2.957	∞

Table 7 and figure 6 compare the results for augmented PPPI networks with original PPPI network. It is clear that addition of coPPI from other three networks to PPPI network (resulting network is PPPI3CO) increases the value of DCOP from 5.63 to 10.86 and value of SNR from 1.537 to 2.957. As a result prediction quality increases from -744 to 1240, true success rate increases from 41% to 57% and overall success rate increases from 65% to 81%. Further removing ncPPI (resulting network is PPPI3CONOR), the value of DCOP remains the same at 10.82 but there is an increase in SNR from 2.957 to ∞. As a result, we see further increase in quality from 1240 to 2909, true success rate from 57% to 75%, and overall success rate from 81% to 88%.

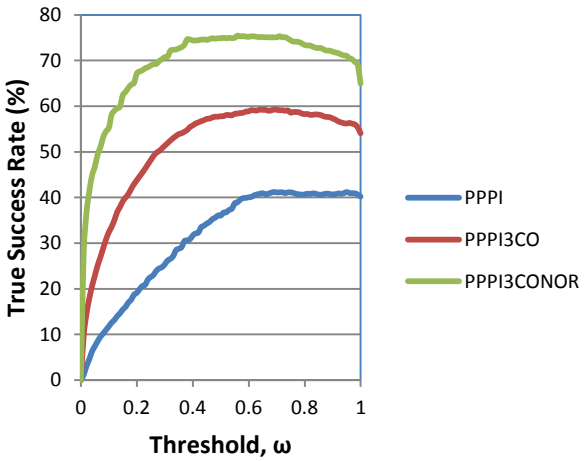

Figure 6. True success rates of original PPPI and augmented PPPI networks

3.6. Comparison with other classifiers

In the present study, we used the same set of experimental annotation of yeast as used by Chou and Cai (2005) for predicting multiplex protein localization. They reported their results for 3875 different proteins with 5132 localizations. Our analysis is based on 3803 different proteins with 5039 localizations. One of the limitations of network-based prediction is that if a protein is not in the network, then the predictor cannot predict for that protein. As a result, the number of proteins in the present study is lightly less than that used by Chou and Cai (2005).

Table 8. Overall success rate of leave-one-out cross-validation

Scope*	Success Rate	
	Chou and Cai, 2005	Our Approach
Ranking I	3596/5132 = 70.07%	3555/5039 = 70.55%
Ranking I + II	4328/5132 = 84.33%	4564/5039 = 90.57%
Ranking I + II + III	4627/5132 = 90.16%	4698/5039 = 93.23%

*Definition of scope. Ranking I considers top-1 prediction; Ranking II considers top-2 prediction; Ranking III considers top-3 prediction.

It is clear from table 8 that our approach, NetLoc, produces better results in all of the three ranking scheme. This shows that PPC network alone provides rich information about protein localization.

4. Conclusion

We have applied a diffusion kernel based logistic regression classifier for predicting subcellular locations of

proteins that target multiple locations based on the protein-protein interaction network. Experimental results showed that this network based method can achieve high accuracy with overall and true success rate of 88.43% and 74.65% respectively when multiple networks are used. Two factors are identified as important network features that determine the prediction performance, including the ratio of the number of co-localized protein-protein interactions (coPPI) to the number of non-co-localized PPI (ncPPI) and the density of annotated coPPI in the network. Networks with larger values for either or both features tend to allow the Netloc algorithm to achieve higher prediction accuracy.

Acknowledgement

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.

Bibliography

[1] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea, "Global analysis of protein localization in budding yeast," *Nature*, vol. 425, no. 6959, pp. 686-691, Oct.2003.

[2] A. Kumar, S. Agarwal, J. A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, K. H. Cheung, P. Miller, M. Gerstein, G. S. Roeder, and M. Snyder, "Subcellular localization of the yeast proteome," *Genes Dev.*, vol. 16, no. 6, pp. 707-719, Mar.2002.

[3] K. Lee, D. W. Kim, D. Na, K. H. Lee, and D. Lee, "PLPD: reliable protein localization prediction from imbalanced and overlapped datasets," *Nucleic Acids Research*, vol. 34, no. 17, pp. 4655-4666, Oct.2006.

[4] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence," *Journal of Molecular Biology*, vol. 300, no. 4, pp. 1005-1016, July2000.

[5] B. R. King and C. Guda, "ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes," *Genome Biology*, vol. 8, no. 5 2007.

[6] A. Bulashevska and R. Eils, "Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains," *Bmc Bioinformatics*, vol. 7 June2006.

[7] A. C. Lorena and A. C. P. L. de Carvalho, "Protein cellular localization prediction with support vector machines and decision trees," *Computers in Biology and Medicine*, vol. 37, no. 2, pp. 115-125, Feb.2007.

[8] S. J. Hua and Z. R. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, no. 8, pp. 721-728, Aug.2001.

[9] K. C. Chou and Y. D. Cai, "Predicting protein localization in budding yeast," *Bioinformatics*, vol. 21, no. 7, pp. 944-950, Apr.2005.

[10] Yang Yang and Bao-Liang Lu, "Prediction of Protein Subcellular Multi-locations with a Min-Max Modular Support Vector Machine," *Lecture Notes in Computer Science*, vol. 3973/2006 2006.

[11] K. C. Chou and H. B. Shen, "Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites," *J. Proteome. Res.*, vol. 6, no. 5, pp. 1728-1734, May2007.

[12] K. C. Chou and H. B. Shen, "A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0," *PLoS One*, vol. 5, no. 4, p. e9931, 2010.

[13] K. C. Chou, Z. C. Wu, and X. Xiao, "iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS One*, vol. 6, no. 3, p. e18258, 2011.

[14] K. Lee, H. Y. Chuang, A. Beyer, M. K. Sung, W. K. Huh, B. Lee, and T. Ideker, "Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species," *Nucleic Acids Res.*, vol. 36, no. 20, p. e136, Nov.2008.

[15] S. Mintz-Oron, A. Aharoni, E. Ruppin, and T. Shlomi, "Network-based prediction of metabolic enzymes' subcellular localization," *Bioinformatics.*, vol. 25, no. 12, p. i247-i252, June2009.

[16] Ananda Mondal and Jianjun Hu, "NetLoc: Network Based Protein Localization Prediction Using Protein-Protein Interaction and Co-expression Networks," IEEE International Conference on Bioinformatics & Biomedicine (BIBM2010), 2010.

[17] H. Lee, Z. D. Tu, M. H. Deng, F. Z. Sun, and T. Chen, "Diffusion kernel-based logistic regression models for protein function prediction," *Omics-A Journal of Integrative Biology*, vol. 10, no. 1, pp. 40-55, Mar.2006.