Network Based Prediction of Protein Localization Using Diffusion Kernel

Abstract: We present NetLoc, a novel diffusion kernel-based logistic regression algorithm for predicting protein subcellular localization using four types of protein networks including physical PPI networks, genetic PPI networks, mixed PPI networks, and co-expression networks. NetLoc is applied to yeast protein localization prediction. The results showed that protein networks can provide rich information for protein localization prediction, achieving AUC score of 0.93. We also showed that networks with high connectivity and high percentage of co-localized PPI lead to better prediction performance. Investigation showed that NetLoc is a very robust approach which can produce good performance (AUC = 0.75) only using 30% of original interactions and capable of producing overall accuracy greater than 0.5 only with 20% annotation coverage. Compared to the previous network feature based prediction algorithm which achieved AUC scores of 0.49 and 0.52 on the yeast PPI network, NetLoc achieved significantly better overall performance with the AUC of 0.74.

Keywords: NetLoc, network based protein localization; protein localization prediction; protein localization; protein subcellular localization; protein-protein interaction; PPI; PPI networks, genetic networks, co-expression networks; kernel-based logistic regression; KLR ; diffusion kernel; data mining; bioinformatics.

1. Introduction

Proper protein functions are closely influenced by its precise targeting to designated subcellular localization. Computational prediction of protein localizations can greatly help to infer protein functions. However, experimental determination of protein localization is costly (Huh et al., 2003, Kumar et al., 2002) and has been conducted for a few model organisms such as human, mouse, and yeast. In the past decade, many algorithms have been developed for computational prediction of protein subcellular locations (Casadio et al., 2008, Emanuelsson et al., 2007, Gardy and Brinkman, 2006, Lee et al., 2006b). These algorithms employ a variety of supervised machine learning techniques including neural networks (Shen et al., 2007, Emanuelsson et al., 2000), nearest neighbor classifier, Markov models, Bayesian networks (King and Guda, 2007, Bulashevska and Eils, 2006), expert rules, meta-classifiers (Jin et al., 2008, Liu et al., 2007), and the support vector machines (Lorena and de Carvalho, 2007, Hua and Sun, 2001). While algorithm variation can tune up the prediction performance, the most critical factor for accurate prediction is to integrate different sources of data (information) to infer the subcellular location of a protein. Current prediction algorithms can be classified into four categories in terms of the evidences used: 1) algorithms based on targeting signals such as PSORT (Nakai and Horton, 1999) and TargetP (Emanuelsson et al., 2000). However, due to limited experimental targeting signal data and the low coverage of targeting signal prediction algorithms, the performances of these approaches are not satisfactory; 2) algorithms considering the preference or bias in terms of amino acid composition (Nanni and Lumini, 2008, Yu et al., 2004) or protein domains (Chou and Cai, 2004, Shi et al., 2007, Mott et al., 2002) of the proteins in specific subcellular compartments. Using composition information has the disadvantage of losing sequence order information and is not specific enough for precise prediction; 3) algorithms using localization information from other annotated proteins with indirect relationships such as functional annotation (Szafron et al., 2004), phylogenetic profiling (Marcotte et al., 2000), homology (Yu et al., 2006), and protein-protein interaction (Zhang et al., 2008); 4) algorithms that integrate multiple sources of information. Drawid and Gerstein's (2000) naïve Bayesian predictor uses signal motifs, gene expression patterns, and overall-sequence properties. Scott et al.'s (2005) Bayesian network predictor incorporates protein motifs, targeting signals, and protein-protein interaction data.

Recently, protein-protein correlation (PPC) networks have been used for localization prediction. Lee et al. (2008) 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. (2009) used metabolic networks for localization prediction using constraint-based models. However, it is difficult to incorporate other

information into the prediction model. In addition, genetic interaction networks and co-expression networks also carry information for localization prediction but remain unexplored. It is also not clear what topological characteristics of networks affect their potential for localization prediction.

Here we introduced a network (Srinivasan et al., 2007) based protein localization prediction algorithm NetLoc by combining diffusion kernel with logistic regression to build a prediction model. It can be applied to a variety of protein-protein correlation networks such as physical or genetic PPI network, and co-expression network. For all these networks, connected protein pairs tend to be localized in the same subcellular compartments. We applied NetLoc to genome wide yeast protein localization using PPI, and COEXP networks. In a cross-validation test of predicting known subcellular localization of 3804 proteins of Yeast, NetLoc is shown to achieve high accuracy with AUC values ranging from 0.71 to 0.93 for cytoplasm, ER, mitochondrion, nucleolus, nucleus, punctuate composite, and vacuole using only physical PPI network. We also found that the number of connected components and the co-localization degree of protein-pairs strongly affect the prediction performance using the proposed network prediction models.

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, which are derived either from protein sequences, or from external functional information such as gene ontology or physichemical properties. However, one apparent limitation of these methods is that it is not easy to exploit rich network information that naturally appears among proteins. For example, two proteins that interact physically will very likely be located within the same organelle. Thus protein-protein interaction networks are very informative for protein localization prediction. Another example is the gene co-expression network which describes whether two genes/proteins show similar gene expression behaviors indicating that they are regulated by the same set of transcription factors. So if two proteins are controlled by the same transcription factor, they are most likely to be involved in the same biological pathway and then likely to be located within the same compartment. It is thus interesting to explore non-feature based prediction algorithms for protein localization prediction.

Another issue of current protein localization prediction 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. An ideal prediction algorithm should be able to output probabilistic scores for all locations for each protein so that multi-location proteins can also be predicted with different confidence.

The basic idea of our approach is to utilize the information of protein-protein correlation network structure in 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. For example, a protein-protein interaction (PPI) network provides a neighborhood structure among the proteins. If two proteins interact, they are neighbors of each other. The localizations of its neighbors 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 localized to the same location. A confidence or probability about the fact that the protein is localized at a certain location will 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 (Lee et al., 2006a).

2.2. KLR logistic regression model

We applied the diffusion kernel-based logistic regression (KLR) model (Lee et al., 2006a) to predict protein subcellular localization based on the locations of all other proteins within function linkage

networks. This method has the unique advantage of considering the subcellular location labels of all the related proteins. It is desirable because signaling peptides that direct proteins to different locations usually share some similarity, e.g. the signal peptides targeting outer membrane and plasma membrane share the N-terminal secretary signals.

The KLR model based subcellular prediction problem can be formulated as follows (Lee et al., 2006a). 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 $X_{[-i]} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_N)$, $M_0(i) = \sum_{j \neq i, x_j\ known} K(i,j) I\{x_j = 0\}$, and

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

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:

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

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:

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

The parameters γ, δ, η can be estimated using the maximum likelihood estimation (MLE) method. Note that here only the annotated proteins are used in the estimation procedure.

The KLR model has been successfully applied to protein function prediction. However, comparing with that application, KLR is especially suitable for protein localization prediction due to the following factors: 1) there are much fewer locations than protein function categories and the correlation among the subcellular locations are much stronger than protein functions; 2) the location is a much broader classification than the protein function, which means that the network neighborhood topology may provide sufficient evidence for its inference.

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 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 confidence for each protein for a particular localization. Predictions are made for un-annotated proteins based on some threshold on confidence value.

3. Experimental results

3.1. Dataset preparation

Four protein networks for Saccharomyces cerevisiae are used in the present study: two networks, physical PPI network and genetic PPI network, are obtained from BioGRID (Stark et al., 2006), another PPI network is from MIPS (Guldener et al., 2006) and one co-expression network is from gene expression data of Stanford University (Spellman et al., 1998). In this study, the networks are named as physical PPI (PPPI), genetic PPI (GPPI), mixed PPI (MPPI) and COEXP respectively. PPPI contains only physical interactions whereas MPPI contains both physical and genetic interactions. MPPI has much less interactions due to its latest update is in 2006.

NetLoc is applied to protein localization prediction of Saccharomyces cerevisiae proteins using the localization data of (Huh et al., 2003) as the basis for annotation. They annotated 4160 proteins with 22 distinct localizations. Out of these localizations, only 7 of them have more than 100 proteins with known subcellular localization annotation. These localizations are cytoplasm, ER (endoplasmic reticulum), mitochondrion, nucleolus, nucleus, punctuate composite and vacuole. We evaluated our network prediction

model based on these 7 localizations. The original dataset has 4160 unique proteins annotated with 5380 localizations. Some proteins are annotated with multiple locations. We removed those proteins with ambiguous localization and 3923 proteins are left with 5191 localization annotations.

Table 1 Datasets of protein correlation networks

Property	PPPI	MPPI	GPPI	COEXP
No. of proteins	5477	4319	5252	2004
Edges	50997	11421	103631	11954
Average interactions per node	9.31	2.64	19.73	5.96

Table 1 shows the summary of four network datasets used for this study. In terms of the number of interactions, GPPI is the largest network followed by PPPI, COEXP70 and MPPI. On the other hand, in terms of proteins, PPPI is the largest network followed by GPPI, MPPI and COEXP70. GPPI is the densest graph followed by PPPI, COEXP70 and MPPI.

3.2. Performance evaluation

In the KLR logistic regression model, for each subcellular localization, all proteins are predicted with a confidence level which indicates how likely a protein belongs to this location. If the threshold is set to 0.5, then a protein with higher than 0.5 confidence will be labeled as positive prediction –belonging to this location, otherwise, negative. Based on this cutoff value, the resulting prediction algorithm can have varying true positive and true negative rate, which makes the comparison difficult. For the present analysis, the AUC (Area Under the Curve) score was used to measure the prediction capability of the proposed KLR model using network information. 5-fold cross-validation was used to calculate the AUC values for the classifiers.

3.3. Localization prediction using co-expression network

Co-expression network is prepared based on the gene expression patterns of Yeast. We first calculate the correlation coefficients of gene pairs in terms of their gene expression levels across several conditions. Then we derive a co-expression network given a threshold coefficient value. The motivation to use COEXP for localization prediction is that co-expressed proteins are expected to occur within the same subcellular compartment.

Table 2 shows the properties of the co-expression networks derived with different cutoff coefficient. For each of the network, we ran our prediction algorithm and evaluated their performance in terms of the AUC scores using 5-fold cross-validation. It can be observed that with larger cutoff threshold, less proteins and interactions remain in the network. The best prediction performance is achieved when the correlation coefficient threshold is set to 0.7 with considerable coverage of proteins.

Table 2 Co-expression networks and classification accuracy on 7 localizations

Item	COEXP60	COEXP65	COEXP70	COEXP75	COEXP80
Interactions	58988	26120	11954	4792	1528
Proteins	4434	3180	2004	1122	567
Average interactions per protein	13.30	8.21	5.96	4.27	2.69
AUC	0.6928	0.7273	0.7489	0.7391	0.7444

3.4. Localization prediction using PPPI, GPPI and MPPI networks compared to COEXP networks

The prediction performance of NetLoc using individual networks for the selected 7 localizations is shown in Figure 2 and Table 3. For PPPI network, AUC varies between 0.71 and 0.93 among which 4 classes have AUC > 0.80 and 1 class (nucleolus) has AUC > 0.90. For GPPI network, AUC varies between 0.63 and 0.89 with 3 classes having AUC > 0.80 and none having AUC > 0.90. For MPPI network, AUC varies between 0.61 and 0.81 with 1 class (nucleolus) having AUC > 0.80 and none having AUC > 0.90. For COEXP70 network, AUC varies between 0.66 and 0.90 with 2 classes having AUC > 0.80 and 1 class (nucleolus) having AUC > 0.90. Overall AUC values for PPPI, GPPI, MPPI, and COEXP70 are 0.82, 0.75, 0.75, and 0.69 respectively. The prediction performance shows that the PPPI network gives the best result for localization prediction.

The prediction performance of NetLoc is competitive compared to other localization prediction algorithms that only use single-protein features. For example, It was reported (Lee et al., 2008) that the single protein feature based methods achieved prediction performance of about 0.65 and 0.79 (AUC score) without or with feature selection on the same yeast dataset as used here. NetLoc achieved AUC score of 0.82 for the 7 selected locations and AUC score of 0.85 for all 22 locations. Compared to Lee *et al.*'s (2008) network feature based method which achieved AUC score of 0.49 and 0.52 using two types of PPI network features (L and N features) from DIP dataset (Xenarios et al., 2000), NetLoc achieved AUC score of 0.74 on the same dataset.

Table 3 Summary of performances with different PPC networks for selected 7 localizations

Network	Classes/Localizations			
	AUC > 0.60	AUC > 0.70	AUC > 0.80	AUC > 0.90
PPPI	7	7	4	1
GPPI	7	4	3	0
MPPI	7	3	1	0
COEXP70	7	5	2	1

3.5. Network topology versus localization prediction

The performance of NetLoc depends on a variety of topological properties of the network such as graph connectivity, density of edges, and the co-localization ratio of protein pairs. Table 4 summarizes the topological properties of four PPC networks along with their prediction performance.

Table 4 Summary of graphical structure for different protein networks

Item	**PPPI**	**GPPI**	**MPPI**	**COEXP70**
Nodes (Proteins)	5477	5252	4319	2004
Edges (PPIs)	50997	103631	11421	11954
Node Pairs	15m	13.7m	9m	2m
Connected Component	1	1	75	136
Nodes in Largest Comp	5477	5252	4158	1612
% Nodes in Largest Comp	100%	100%	96.%	80.44%
Performance	0.8525	0.7851	0.7132	0.6407

PPPI and GPPI networks have one connected component. COEXP70 has 136 connected components and MPPI has 75 connected components. In COEXP70, the largest component is composed of 80% of total nodes and in MPPI, the largest component is composed of 96% of total nodes. The performance on these four networks suggests that the number of connected component has direct impact on performance. In general, a network with only one connected component performs better than a network with more connected components. Another factor that also affects prediction performance is the percentage of PPIs going to the same location. While GPPI and MPPI networks have about same percent (30%) of PPIs going to the same location (Table 6), but GPPI produces better performance (0.7851) than MPPI (0.7132) because GPPI is composed of only one connected component and MPPI is composed of 75 connected components.

3.6. Effect of network connectivity on NetLoc performance

In order to check the effect of connectivity on NetLoc performance, we removed 5%, 10%, 20%, 50%, and 70% edges from the original network and evaluated the resulting performance. For each removal, 10 random sets of edges (PPIs) are removed, performance is evaluated with the remaining network after each set of removal and then the average of 10 performances is taken. Table 5 summarizes the network characteristics for PPPI network and the performance with different level of connectivity. It is found that the NetLoc performance for selected locations decreases from 0.81 to 0.75 when the percentage of edges decreased from 100% to 30%. This proves that networks with more connections/interactions in general produces better results in predicting protein localization. This also proves the hypothesis made earlier in section 3.5 that network with more connected components deteriorates NetLoc's performance.

Table 5 Network characteristics and NetLoc performance with different percent of edges (PPIs) in PPPI network.

Edges in the Network	100%	95%	90%	80%	50%	30%
# of Component	1	40	59	126	512	1027
# of Nodes in Largest Component	5477	5438	5419	5351	4965	4435
% Nodes in Largest Component	100.0	99.3	98.9	97.7	90.7	81.0
Lowest Degree	1	0	0	0	0	0
Highest Degree	2546	2422	2304	2032	1278	762
AUC, Selected Locations	0.8116	0.8094	0.8065	0.8026	0.7768	0.7488

3.7. Annotation coverage on NetLoc performance

A robust model for predicting protein localization based on PPI network should produce better performance if new annotations are added to the network. In the following experiments, we tested 1) the effect on prediction performance by adding additional annotations; 2) how the annotation coverage affects the prediction performance. The experiment is carried out with both low-resolution localization (5 locations) and high-resolution localization (22 locations). Five locations in low resolution localization 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).

Adding additional annotations improves NetLoc prediction performance

The annotated proteins are divided into 5 mutually exclusive equal-sized groups (pseudo randomly) i.e., 20% annotated proteins in each group. Then we left each annotated group out and for the remaining 4 groups of annotated proteins, we compare their 10-fold cross validation prediction performance with that of the network with the left out 20% annotations. For example, one test set is composed of 3042 (80%) annotated proteins and 761 (20%) leave-out annotated proteins. We run 10-fold cross-validation on the 3042 annotated proteins using the PPI network. The number of test proteins for one fold is 304 and the number of corresponding annotated proteins is 2738. First, localization predictions for 304 proteins are obtained using PPPI network and 2738 annotated proteins. Then, predictions for the same 304 proteins are obtained using PPPI network and 3499 (= 2738 + 761) annotated proteins. This procedure is repeated for each of the folds. Now we have two sets of prediction for 3042 proteins without and with 20% additional annotation. MCC (Matthews Correlation Coefficient) values are evaluated for these two sets of prediction by comparing with the actual experimental annotation. The whole procedure is repeated for 5 set of pairs and average of 5 sets are taken to eliminate the biasness of any set.

Figure 3 presents the prediction performance results (MCC) for high-resolution localizations before and after adding new annotations. It is clear that adding new annotations does improve the performance for all locations. Similar results were observed with low-resolution localizations. These results showed the effectiveness and robustness of the network approach for protein localization prediction.

Higher annotation coverage gives higher prediction performance

Here we tested the influence of the annotation coverage on the prediction performance. Only low-resolution localizations were included since many of the high-resolution locations have too few annotated proteins. Five sets of annotated proteins are created with varying degrees of annotation coverage 100%, 80%, 60%, 40%, and 20%. Other than 100% coverage, annotated proteins are randomly selected for the required coverage for five times, which gives 5 different sets of annotated proteins of same size. 10-fold cross validation is carried out for each of the 5 sets of annotations. The average of the 5 sets is calculated to avoid sampling bias. The whole procedure is repeated for each annotation coverage level.

Figure 4 and 5 show MCC values for the 5 low-resolution locations and the overall accuracy for different annotation coverage. It is clear that both MCC and overall accuracy are increased with the increase of annotation coverage as expected. It is also noticeable that PPI network is capable of producing overall accuracy greater than 0.5 (non-random) with only 20% annotation coverage. This shows the effectiveness of network approach in predicting protein localization.

4. Discussion

This paper investigates the performance of the proposed diffusion kernel based logistic regression model for predicting protein localizations using only protein-protein correlation network information. We have shown that the proposed NetLoc approach is robust, can achieve high prediction accuracy, and showed that network topological characteristics such as connectivity may affect the prediction performance.

Another important factor that may affect the prediction performance is the correlation of interactions as regard to co-localization. Table 6 shows the percentages of protein pairs of which both proteins go to the same location along with the prediction performance (AUC score) using the networks. PPPI has the highest percentage of co-localized protein pairs: 41.95% of protein pairs co-localize. Together with the high connectivity, NetLoc has the best performance on the PPPI network (AUC = 0.8525). GPPI network also has only one connected component, but its co-localized proteins only cover 30.18% of all protein pairs. So its performance (AUC = 0.7851) is lower than using PPPI network. Compared with GPPI network, both MPPI and COEXP70 networks have similar percentages of co-localized protein pairs, but they are distributed in much more disconnected patches with 75 connected components for MPPI and 136 connected components for COEXP70. The prediction performances are thus inferior to that of PPPI network. In general, the more protein pairs go to the same location, the better the prediction performance given equal number of connected components.

Table 6 Protein pairs targeting the same location and prediction performance

Network	Total PPI	Connected Component	PPI at Same Loc	%PPI at Same Loc	AUC
PPPI	50997	1	21395	41.95	0.8525
GPPI	103631	1	31279	30.18	0.7851
MPPI	11421	75	3501	30.65	0.7132
COEXP70	11954	136	4206	35.18	0.6407

Comparing the influence of network connectivity and co-localization percentage, the former seems to have a large effect. For example, the percentage of PPIs going to the same localization in COEXP70 is 35.18%, which is greater than that of MPPI (30.65%), However, it has much more connected components (136) compared to MPPI (75). As a result, COEXP70 produces poor performance.

We found that NetLoc is a highly robust approach in predicting protein localization, which can produce good performance (AUC = 0.75) with only 30% of original interactions and is capable of producing overall accuracy greater than 0.5 only with 20% annotation coverage.

Our experiments showed that diffusion kernel based network prediction model in NetLoc achieved better prediction performance than the method using network based features as used in previous work (Lee et al., 2008). N features of Lee et al. (2008) using weighted average of single-protein features was shown to be worse than the L features using weighted voting of neighbors within a certain distance. However, the weights are calculated from conditional probabilities. NetLoc used weighted voting of all proteins in the network in which the weights are optimized using logistic regression, which makes it exploits better the network information for localization prediction.

The cross-validation results showed comparable performance of popular amino acid composition based features. However, a main advantage of the network method is that it has the capability of integrating multiple networks to make prediction. Our preliminary experiments showed that by combining two networks, PPPI and GPPI, we can further improve the prediction performance. Moreover, the diffusion kernel based prediction model can be used to determine the contribution of each of the protein-protein networks in protein localization. Another ongoing work is to integrate NetLoc with other feature based methods to build an ensemble prediction algorithm. Since, in feature-based methods, it is very difficult to differentiate cytoplasmic proteins from nucleus proteins, our protein correlation network approach could be very helpful.

5. Conclusion

A diffusion kernel based logistic regression (KLR) model for protein subcellular localization prediction using protein-protein correlation networks has been proposed. Four types of networks including physical interaction, genetic interaction, mixed interaction, and co-expression network have been used for protein localization prediction of yeast. Results indicated that all these four networks carry protein co-localization information with their interactions (edges) and can thus be used for localization prediction. Experiments showed that the physical interaction network has the highest connectivity and highest percentage of co-localized protein pairs, which leads to best prediction performance. Genetic interaction network has the second best localization prediction performance. Co-expression network has the least information for localization prediction due to its lower connectivity with many isolated patches. It was found that network topology strongly affects the NetLoc prediction performance. In particular, the number of connected components, the average degree of nodes, and the percentage of co-localized protein-pairs all play important role for the prediction performance. Our experiments showed that the proposed network approach is highly robust in predicting protein localization as regard to the network connectivity and annotation protein coverage.

References

BLUM, T., BRIESEMEISTER, S. & KOHLBACHER, O. 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, 10**,** 274.

BULASHEVSKA, A. & EILS, R. 2006. Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinformatics,* 7**,** 298.
CASADIO, R., MARTELLI, P. L. & PIERLEONI, A. 2008. The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief Funct Genomic Proteomic,* 7**,** 63-73.
CHOU, K. C. & CAI, Y. D. 2004. Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. *J Cell Biochem,* 91**,** 1197-203.
DRAWID, A. & GERSTEIN, M. 2000. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol,* 301**,** 1059-75.
EMANUELSSON, O., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc,* 2**,** 953-71.
EMANUELSSON, O., NIELSEN, H., BRUNAK, S. & VON HEIJNE, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol,* 300**,** 1005-16.
GARDY, J. L. & BRINKMAN, F. S. 2006. Methods for predicting bacterial protein subcellular localization. *Nat Rev Microbiol,* 4**,** 741-51.
GULDENER, U., MUNSTERKOTTER, M., OESTERHELD, M., PAGEL, P., RUEPP, A., MEWES, H. W. & STUMPFLEN, V. 2006. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res,* 34**,** D436-41.
HUA, S. & SUN, Z. 2001. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics,* 17**,** 721-8.
HUH, W. K., FALVO, J. V., GERKE, L. C., CARROLL, A. S., HOWSON, R. W., WEISSMAN, J. S. & O'SHEA, E. K. 2003. Global analysis of protein localization in budding yeast. *Nature,* 425**,** 686-91.
JIN, Y. H., NIU, B., FENG, K. Y., LU, W. C., CAI, Y. D. & LI, G. Z. 2008. Predicting subcellular localization with AdaBoost Learner. *Protein Pept Lett,* 15**,** 286-9.
KING, B. R. & GUDA, C. 2007. ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes. *Genome Biol,* 8**,** R68.
KUMAR, A., AGARWAL, S., HEYMAN, J. A., MATSON, S., HEIDTMAN, M., PICCIRILLO, S., UMANSKY, L., DRAWID, A., JANSEN, R., LIU, Y., CHEUNG, K. H., MILLER, P., GERSTEIN, M., ROEDER, G. S. & SNYDER, M. 2002. Subcellular localization of the yeast proteome. *Genes Dev,* 16**,** 707-19.
LEE, H., TU, Z., DENG, M., SUN, F. & CHEN, T. 2006a. Diffusion kernel-based logistic regression models for protein function prediction. *OMICS,* 10**,** 40-55.
LEE, K., CHUANG, H. Y., BEYER, A., SUNG, M. K., HUH, W. K., LEE, B. & IDEKER, T. 2008. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res,* 36**,** e136.
LEE, K., KIM, D. W., NA, D., LEE, K. H. & LEE, D. 2006b. PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic Acids Res,* 34**,** 4655-66.
LIU, J., KANG, S., TANG, C., ELLIS, L. B. & LI, T. 2007. Meta-prediction of protein subcellular localization with reduced voting. *Nucleic Acids Res,* 35**,** e96.
LODISH, H., BERK, A. & ZIPURSKY, S. L. 2000. *Molecular Cell Biology,* New York.
LORENA, A. C. & DE CARVALHO, A. C. 2007. Protein cellular localization prediction with Support Vector Machines and Decision Trees. *Comput Biol Med,* 37**,** 115-25.
MARCOTTE, E. M., XENARIOS, I., VAN DER BLIEK, A. M. & EISENBERG, D. 2000. Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U S A,* 97**,** 12115-20.
MINTZ-ORON, S., AHARONI, A., RUPPIN, E. & SHLOMI, T. 2009. Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics,* 25**,** i247-52.
MOTT, R., SCHULTZ, J., BORK, P. & PONTING, C. P. 2002. Predicting protein cellular localization using a domain projection method. *Genome Res,* 12**,** 1168-74.
NAKAI, K. & HORTON, P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci,* 24**,** 34-6.
NANNI, L. & LUMINI, A. 2008. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids,* 34**,** 653-60.
SCOTT, M. S., CALAFELL, S. J., THOMAS, D. Y. & HALLETT, M. T. 2005. Refining protein subcellular localization. *PLoS Comput Biol,* 1**,** e66.
SHEN, H. B., YANG, J. & CHOU, K. C. 2007. Methodology development for predicting subcellular localization and other attributes of proteins. *Expert Rev Proteomics,* 4**,** 453-63.
SHI, J. Y., ZHANG, S. W., PAN, Q., CHENG, Y. M. & XIE, J. 2007. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids,* 33**,** 69-74.
SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTEIN, D. & FUTCHER, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell,* 9**,** 3273-97.
SRINIVASAN, B. S., SHAH, N. H., FLANNICK, J. A., ABELIUK, E., NOVAK, A. F. & BATZOGLOU, S. 2007. Current progress in network research: toward reference networks for key model organisms. *Brief Bioinform,* 8**,** 318-32.

STARK, C., BREITKREUTZ, B. J., REGULY, T., BOUCHER, L., BREITKREUTZ, A. & TYERS, M. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res,* 34**,** D535-9.

SZAFRON, D., LU, P., GREINER, R., WISHART, D. S., POULIN, B., EISNER, R., LU, Z., ANVIK, J., MACDONELL, C., FYSHE, A. & MEEUWIS, D. 2004. Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. *Nucleic Acids Res,* 32**,** W365-71.

XENARIOS, I., RICE, D. W., SALWINSKI, L., BARON, M. K., MARCOTTE, E. M. & EISENBERG, D. 2000. DIP: the database of interacting proteins. *Nucleic Acids Res,* 28**,** 289-91.

YU, C. S., CHEN, Y. C., LU, C. H. & HWANG, J. K. 2006. Prediction of protein subcellular localization. *Proteins,* 64**,** 643-51.

YU, C. S., LIN, C. J. & HWANG, J. K. 2004. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci,* 13**,** 1402-6.

ZHANG, S., XIA, X. F., SHEN, J. C. & SUN, Z. R. 2008. Eukaryotic protein subcellular localization prediction based on sequence conservation and protein-protein interaction. *Progress in Biochemistry and Biophysics,* 35**,** 531-535.

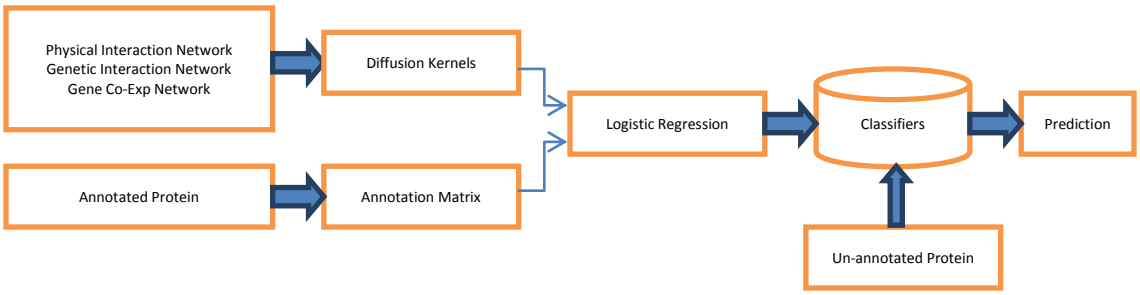

Figure 1 Protein localization prediction using the KLR model and protein networks.

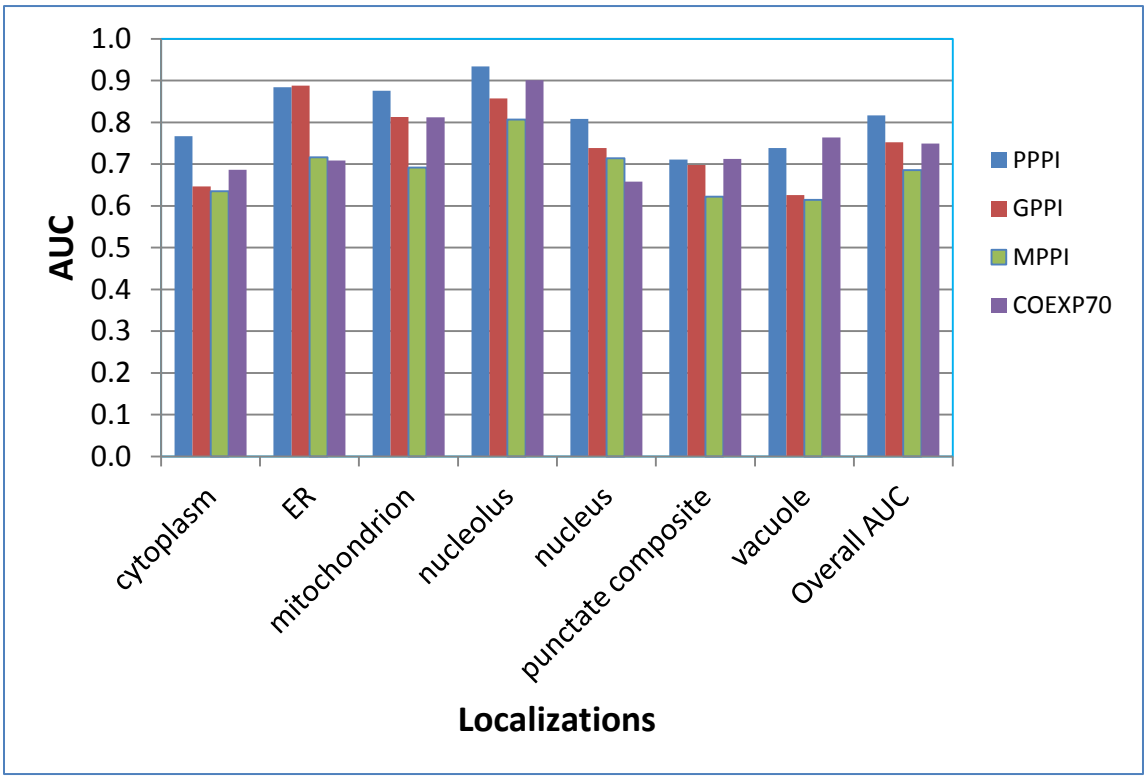

Figure 2 Performances of individual networks for selected 7 localizations with more than 100 proteins.

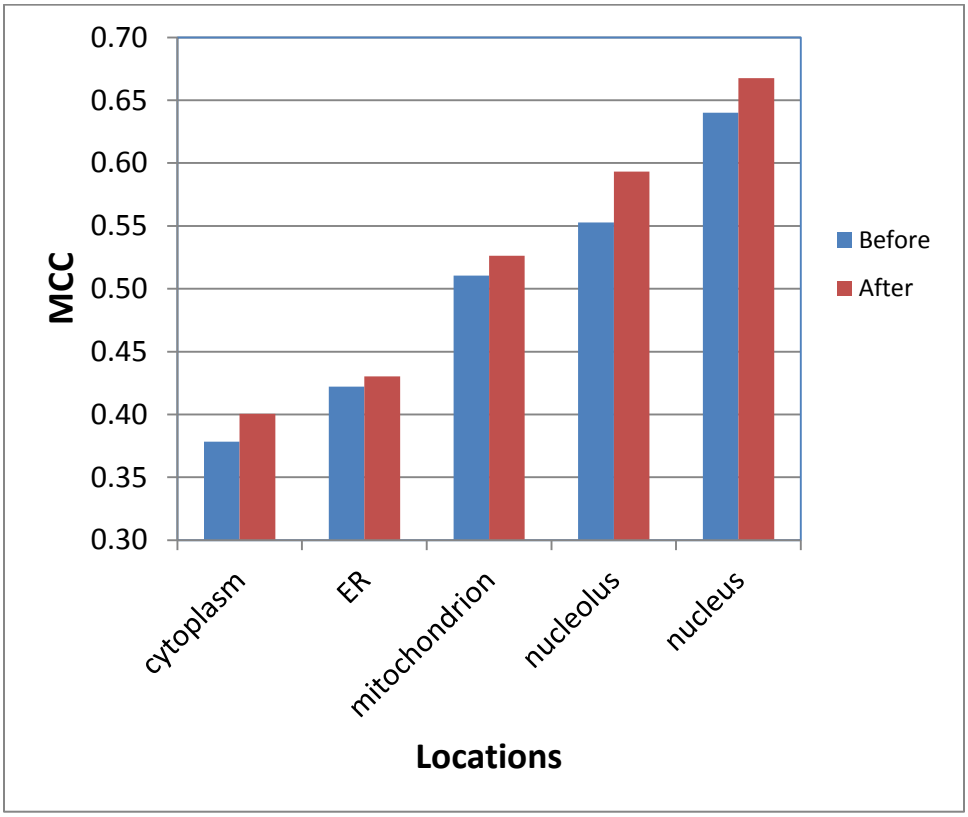

Figure 3 MCC values for 5 different locations for high-resolution localization before and after adding 20% new annotations. Locations with less than 100 proteins are not included in the test set due to their low coverage.

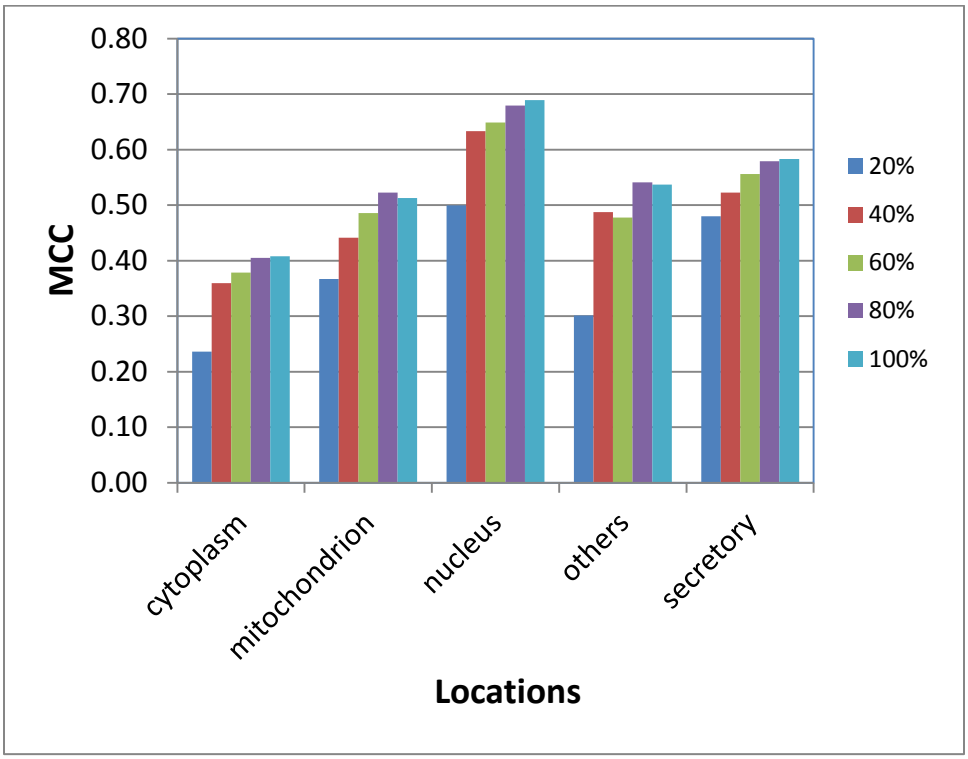

Figure 4 MCC scores for low-resolution locations with different annotation coverage.

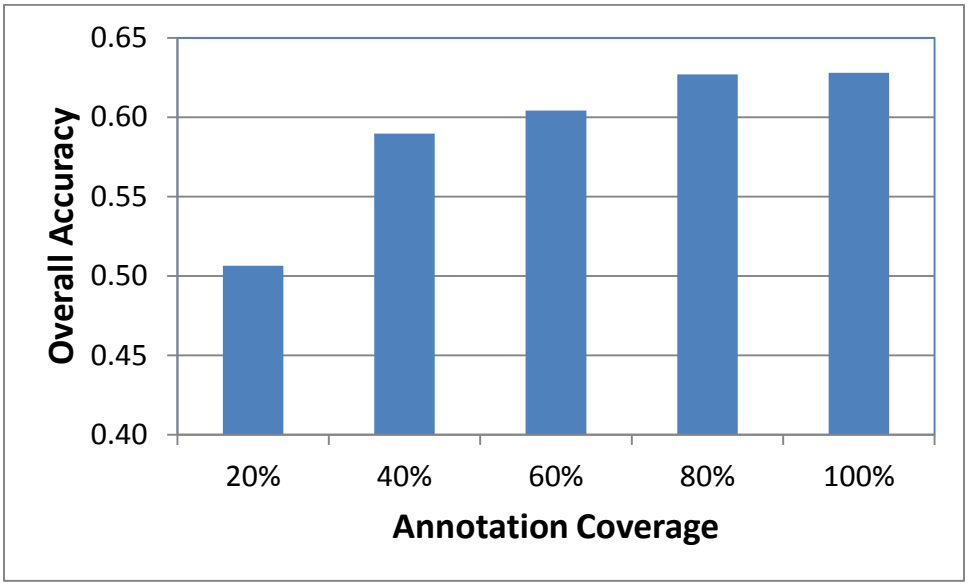

Figure 5 Overall accuracy for different annotation coverage for PPPI network.