NetLoc: Network Based Protein Localization Prediction Using Protein-Protein Interaction, Genetic Interaction, and Co-expression Networks

Ananda M. Mondal[1,2], Jianjun Hu[1]

[1]*Department of Computer Science and Engineering, University of South Carolina, 301 Main St, Columbia, SC, 29036*

[2]*Department of Mathematics and Computer Science, Claflin University, 400 Magnolia St, Orangeburg, SC 29115*

E-mail: *ammondal@cec.sc.edu, jianjunh@cse.sc.edu*

Abstract

Recent study shows that protein-protein interaction network based features can significantly improve the prediction of protein subcellular localization. However, it is unclear whether network prediction models or other types of protein-protein correlation networks would also improve localization prediction. We present NetLoc, a novel network based algorithm for predicting protein subcellular localization using four types of protein networks including physical protein-protein interaction (PPPI) network, genetic interaction network (GPPI), and co-expression network (COEXP). Diffusion kernel-based logistic regression (KLR) is used to develop the prediction model. We applied NetLoc to yeast protein localization prediction. The results showed that protein networks can provide rich information for protein localization prediction, achieving prediction performance up to AUC score of 0.93. We also showed that networks with high connectivity and high percentage of interacting protein pairs targeting the same location lead to better prediction performance. In terms of localization prediction performance, PPPI is better than GPPI which is better than COEXP. The classification performance (AUC) with PPPI network ranges between 0.71 and 0.93 for 7 locations. The overall balanced performance is 0.82 which is significantly better than the performance (0.49 and 0.57) of the previous network feature based classification algorithm evaluated on the same yeast dataset using leave-one-out cross-validation.

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 [1;2] 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 [3-7]. These algorithms employ a variety of supervised machine learning techniques including neural networks [8-10], nearest neighbor classifier, Markov models, Bayesian networks [11;12], expert rules, meta-classifiers [13;14], and the support vector machines [15-17]. 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 three categories in terms of the evidences used: 1) algorithms based on targeting signals such as pSORT [18] and TargetP [10]. 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 [19;20] or protein domains [21-23] 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 [24], phylogenetic profiling [25], homology [26], and protein-protein interaction [27]; 4) algorithms that integrate multiple sources of information. Drawid et al's naïve Bayesian predictor [28] uses signal motifs, gene expression patterns, and overall-sequence properties. Scott et. al.'s Bayesian network predictor [29] 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. [30] 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. [31] 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 [32;33] 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 networks, and co-expression networks. 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 3807 proteins of Yeast, NetLoc is shown to achieve high accuracy with AUC values ranging from 0.77 to 0.93 for cytoplasm, ER, mitochondrion, nucleolus, and nucleus 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, and 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 others. 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 [34].

2.2. KLR logistic regression model

We applied the diffusion kernel-based logistic regression (KLR) model [34] to predicting 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 [34]. 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 $X_{[-i]} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_N)$,

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

And $M_1(i) = \sum_{j \neq i, x_j \text{ known}} K(i, j) I\{x_j = 1\}$, where $K(i, j)$ is the kernel function 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:

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

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.

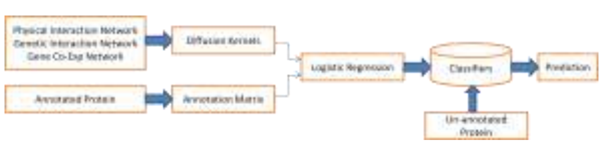

Figure 1. Protein localization prediction using the KLR model and protein networks

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 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 [35], another PPI network is from MIPS [36] and one co-expression network is from gene expression data of Stanford University [37]. 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. [1] 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 cell periphery, cytoplasm, ER (endoplasmic reticulum), mitochondrion, nucleolus, nucleus, and punctate composite. 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 multi-locations). We removed those proteins with ambiguous localization and 3923 proteins are left with 5191 localization annotations.

Table1 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.

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

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 the KLR model using network information. 5-fold cross-validation was used to calculate the AUC value 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 calculated the correlation coefficients of gene pairs in terms of their gene expression levels across several conditions. Then we can 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 as 0.7 with considerable coverage of proteins.

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

Item	COEXP 60	COEXP 65	COEXP 70	COEXP 75	COEXP 80
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 [30] 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 [30] network feature based method which achieved AUC score of 0.49 and 0.52 using two sets of PPI network (MPPI) features (L and N features) from DIP dataset [38] , NetLoc achieved AUC score of 0.7 on the same dataset.

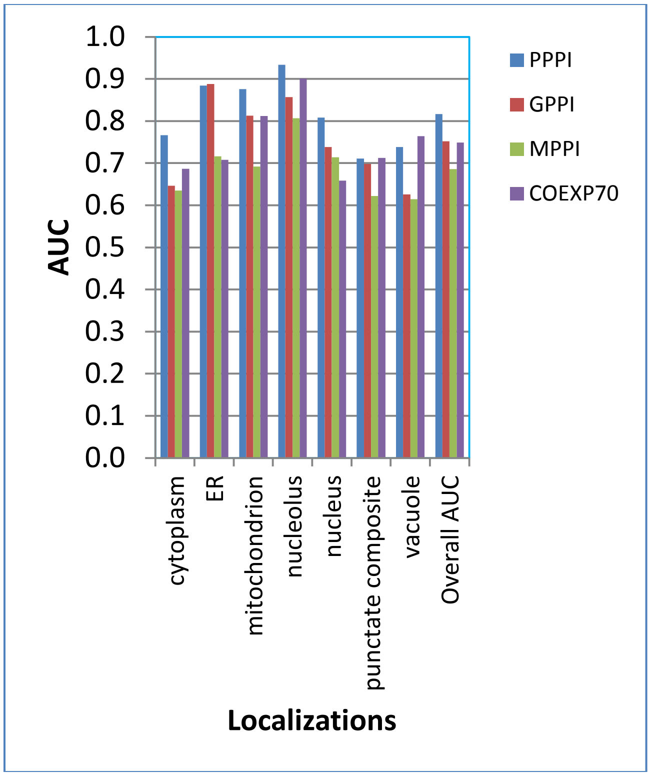

Figure 2. Performances of individual networks for selected 7 localizations with more than 100 proteins.

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 the graph connectivity, density of edges, as well as the co-localization ratio of protein pairs. Table-4 summarizes the topological properties of the four PPC networks along with their prediction performance. 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. A network with only one connected component performs better than the network with more than one connected component. For the same percentage of PPIs going to the same location, the network with only one connected component gives better results than the network with more than one connected components. For example, GPPI and MPPI have about same percent (30%) of PPIs going to the same location, 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.

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

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 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 5 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 5. 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 percentages, 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.

Our experiments showed that diffusion kernel based network predication model in NetLoc achieved better prediction performance than the method using network based features as used in previous work [30]. N features of Lee *et al.* [30] 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 to better exploit 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 our 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, discussed here, 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 network in protein localization. Another ongoing work is to integrate NetLoc with other feature based methods to build an ensemble prediction algorithm. Since feature based method 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, and co-expression network have been used for 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 the best prediction performance. Genetic interaction network has the second best localization prediction. Co-expression network has the least information for localization prediction due to its lower connectivity with many isolated patches. The 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.

Acknowledgement

This work is supported by NSF Career Award DBI-0845381.

References

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] R. Casadio, P. L. Martelli, and A. Pierleoni, "The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation," ***Brief Funct. Genomic. Proteomic.*****, vol. 7, no. 1, pp. 63-73, Jan.2008.**

[4] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, "Locating proteins in the cell using TargetP, SignalP and related tools," ***Nature Protocols*****, vol. 2, no. 4, pp. 953-971, 2007.**

[5] J. L. Gardy and F. S. L. Brinkman, "Methods for predicting bacterial protein subcellular

localization," *Nature Reviews Microbiology*, vol. 4, no. 10, pp. 741-751, Oct.2006.

[6] J. Sprenger, J. L. Fink, and R. D. Teasdale, "Evaluation and comparison of mammalian subcellular localization prediction methods," *Bmc Bioinformatics*, vol. 7 2006.

[7] 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.

[8] H. B. Shen, J. Yang, and K. C. Chou, "Methodology development for predicting subcellular localization and other attributes of proteins," *Expert Review of Proteomics*, vol. 4, no. 4, pp. 453-463, Aug.2007.

[9] T. Thireou and M. Reczko, "Bidirectional Long Short-Term Memory Networks for predicting the subcellular localization of eukaryotic proteins," *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 441-446, July2007.

[10] 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.

[11] 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.

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

[13] Y. H. Jin, B. Niu, K. Y. Feng, W. C. Lu, Y. D. Cai, and G. Z. Li, "Predicting subcellular localization with AdaBoost Learner," *Protein and Peptide Letters*, vol. 15, no. 3, pp. 286-289, Mar.2008.

[14] J. Liu, S. L. Kang, C. N. Tang, L. B. M. Ellis, and T. B. Li, "Meta-prediction of protein subcellular localization with reduced voting," *Nucleic Acids Research*, vol. 35, no. 15 Aug.2007.

[15] 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.

[16] D. Sarda, G. H. Chua, K. B. Li, and A. Krishnan, "pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties," *Bmc Bioinformatics*, vol. 6 June2005.

[17] 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.

[18] K. Nakai and P. Horton, "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization," *Trends in Biochemical Sciences*, vol. 24, no. 1, pp. 34-35, Jan.1999.

[19] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653-660, May2008.

[20] C. S. Yu, C. J. Lin, and J. K. Hwang, "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions," *Protein Science*, vol. 13, no. 5, pp. 1402-1406, May2004.

[21] K. C. Chou and Y. D. Cai, "Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition," *Journal of Cellular Biochemistry*, vol. 91, no. 6, pp. 1197-1203, Apr.2004.

[22] J. Y. Shi, S. W. Zhang, Q. Pan, Y. M. Cheng, and J. Xie, "Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition," *Amino Acids*, vol. 33, no. 1, pp. 69-74, July2007.

[23] R. Mott, J. Schultz, P. Bork, and C. P. Ponting, "Predicting protein cellular localization using a domain projection method," *Genome Research*, vol. 12, no. 8, pp. 1168-1174, Aug.2002.

[24] D. Szafron, P. Lu, R. Greiner, D. S. Wishart, B. Poulin, R. Eisner, Z. Lu, J. Anvik, C. Macdonell, A. Fyshe, and D. Meeuwis, "Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations," *Nucleic Acids Research*, vol. 32, p. W365-W371, July2004.

[25] E. M. Marcotte, I. Xenarios, A. M. van der Bliek, and D. Eisenberg, "Localizing proteins in the cell from their phylogenetic profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 22, pp. 12115-12120, Oct.2000.

[26] C. S. Yu, Y. C. Chen, C. H. Lu, and J. K. Hwang, "Prediction of protein subcellular localization," *Proteins-Structure Function and Bioinformatics*, vol. 64, no. 3, pp. 643-651, Aug.2006.

[27] S. Zhang, X. F. Xia, J. C. Shen, and Z. R. Sun, "Eukaryotic protein subcellular localization prediction based on sequence conservation and protein-protein interaction," *Progress in*

***Biochemistry and Biophysics*, vol. 35, no. 5, pp. 531-535, May2008.**

[28] A. Drawid and M. Gerstein, "A Bayesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome," *Journal of Molecular Biology*, vol. 301, no. 4, pp. 1059-1075, Aug.2000.

[29] M. S. Scott, S. J. Calafell, D. Y. Thomas, and M. T. Hallett, "Refining protein subcellular localization," *PLoS Comput Biol*, vol. 1, no. 6, pp. 518-528, Nov.2005.

[30] 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.

[31] 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.

[32] G. Y. Cui, Y. Chen, D. S. Huang, and K. Han, "An algorithm for finding functional modules and protein complexes in protein-protein interaction networks," *Journal of Biomedicine and Biotechnology*, 2008.

[33] B. S. Srinivasan, N. H. Shah, J. A. Flannick, E. Abeliuk, A. F. Novak, and S. Batzoglou, "Current progress in network research: toward reference networks for key model organisms," *Briefings in Bioinformatics*, vol. 8, no. 5, pp. 318-332, Sept.2007.

[34] 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.

[35] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. Database issue, p. D535-D539, Jan.2006.

[36] U. Guldener, M. Munsterkotter, M. Oesterheld, P. Pagel, A. Ruepp, H. W. Mewes, and V. Stumpflen, "MPact: the MIPS protein interaction resource on yeast," *Nucleic Acids Res.*, vol. 34, no. Database issue, p. D436-D441, Jan.2006.

[37] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol. Biol. Cell*, vol. 9, no. 12, pp. 3273-3297, Dec.1998.

[38] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the database of interacting proteins," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 289-291, Jan.2000.