Published online 4 October 2008

Nucleic Acids Research, 2008, Vol. 36, No. 20 ***e136***
doi:10.1093/nar/gkn619

Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species

KiYoung Lee[1,2,3,4], Han-Yu Chuang[1,5], Andreas Beyer[1,6], Min-Kyung Sung[7], Won-Ki Huh[7], Bonghee Lee[2] and Trey Ideker[1,5,*]

[1]Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA, [2]Center for Genomics and Proteomics, Lee Gil Ya Cancer and Diabetes Institute, Gachon University of Medicine and Science, Incheon 406-799, Republic of Korea, [3]Structural Biology Laboratory, Salk Institute for Biology Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA, [4]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Republic of Korea, [5]Bioinformatics Program, University of California San Diego, La Jolla, CA 92093, USA, [6]Biotechnology Center, Technische Universität, 01062 Dresden, Germany and [7]School of Biological Sciences, Research Center for Functional Cellulomics, Institute of Microbiology, Seoul National University, Seoul 151-747, Republic of Korea

Received April 18, 2008; Revised August 13, 2008; Accepted September 11, 2008

ABSTRACT

The function of a protein is intimately tied to its subcellular localization. Although localizations have been measured for many yeast proteins through systematic GFP fusions, similar studies in other branches of life are still forthcoming. In the interim, various machine-learning methods have been proposed to predict localization using physical characteristics of a protein, such as amino acid content, hydrophobicity, side-chain mass and domain composition. However, there has been comparatively little work on predicting localization using protein networks. Here, we predict protein localizations by integrating an extensive set of protein physical characteristics over a protein's extended protein–protein interaction neighborhood, using a classification framework called 'Divide and Conquer *k*-Nearest Neighbors' (DC-kNN). These predictions achieve significantly higher accuracy than two well-known methods for predicting protein localization in yeast. Using new GFP imaging experiments, we show that the network-based approach can extend and revise previous annotations made from high-throughput studies. Finally, we show that our approach remains highly predictive in higher eukaryotes such as fly and human, in which most localizations are unknown and the protein network coverage is less substantial.

INTRODUCTION

For a protein to operate properly, it must reside in the correct compartment of a cell. Knowing the subcellular localization of a protein, therefore, is an important step to understanding its function (1,2). In budding and fission yeast (1–4), systematic protein localization experiments have been carried out through GFP fusions to each open reading frame at the 3′- or 5′-end. Such studies have not yet been performed in higher eukaryotes such as *Caenorhabditis elegans*, *Drosophila melanogaster* or mammals, due to the larger proteome sizes and the technical difficulties associated with protein tagging in those species (5–7). In the interim, reliable and efficient computational methods are required to predict the subcellular localization of a newly identified protein.

A considerable number of classification methods have been developed for this purpose (5–24). Typically, these algorithms input a list of features with which to characterize a protein, such as its molecular weight, amino acid content, codon bias, hydrophobicity, side-chain mass and so on. During the training phase, they learn to recognize which features, or patterns of features, are best able to classify a set of gold-standard proteins whose localizations are well known. To date, amino acid content

*To whom correspondence should be addressed. Tel: +1 858 822 4665; Fax: +1 858 822 4246; Email: trey@bioeng.ucsd.edu.

has been a very successful and widely used feature (5,6,8,11–16). Other informative features have been protein sorting signal motifs near the N-terminus (18), as well as protein sequence motifs (7,9–12,16,24) and Gene Ontology terms (5). Classification of these features has relied on a variety of algorithms, including Least Distance Algorithms (20,21), an Artificial Neural Network (10), a Nearest Neighbor approach (5,14), a Markov Model (22), a Bayesian Network approach (9), Support Vector Machines (SVMs) (13,15,16) and Support Vector Data Description (SVDD) (6).

Early methods attempted to classify proteins into a small number of compartments, e.g. intracellular versus extracellular (19). More recently, many compartmental localizations have been defined, including not only membrane-enclosed organelles but also categories such as spindle pole or microtubule association. Current prediction algorithms in yeast cover as many as 22 distinct cellular localizations (5,6). Not surprisingly, approaches which limit their predictions to smaller numbers of localizations have performed better than approaches which attempt to predict many. Moreover, most of these studies have demonstrated their predictions assuming a single localization per protein within a single species such as yeast. Therefore, some open challenges for new methods development are to: (i) increase the classification accuracy when predicting across many cellular compartments; (ii) allow for multiple predictions per protein; and (iii) stabilize performance across many species, some of which may have far fewer data available for training and classification than does yeast.

The recent availability of large protein–protein interaction networks in yeast, fly, worm and human (25–34) provides one means to at least partially address these challenges. To interact physically, two proteins must localize to the same or adjacent cellular compartments, suggesting that interaction may serve as an indicator for co-localization. Integrated analysis of genome-wide protein localization and protein–protein interaction data in *Saccharomyces cerevisiae* (*SC*) supports this hypothesis, showing that interactions are strongly enriched between co-localized proteins (1). However, there have been relatively few attempts to use interacting proteins in the prediction of localization (7). Moreover, in recent years the numbers of protein interaction measurements have increased exponentially. This increase has been driven by various proteomics technologies, such as co-immunoprecipitation followed by tandem mass spectrometry, the yeast two-hybrid system and its variants, and large screens for genetic interactions (26,35,36). As a result, there were more than 170 000 protein interactions in the public databases as of this writing (http://www.ebi.ac.uk/intact/); prior to 2002 there were no more than several hundred. Given these developments, protein interactions have become a basic feature available for many proteins. It is therefore of significant interest to ask whether, and to what extent, protein interaction networks can impinge on the prediction of subcellular localization.

Here, we pursue a protein network-based approach for summarizing diverse sequence and functional information of interacting proteins into useful predictors of localization. A variant of the *k*-Nearest Neighbors classification algorithm (5,14) is developed to exploit the synergy between the physical characteristics of an individual protein and the properties of its interacting neighbors. After generating useful features based on single proteins and their neighbors, the method extracts the best combination of feature sets for each cellular localization. We apply this network-based prediction method to predict the localizations of 5681 *SC* proteins, in which a protein is not given a single annotation but is characterized by its predicted distribution across 22 subcellular compartments. Through further GFP imaging experiments, we show that the predictions can provide novel leads even when the localization of a protein has already been measured experimentally.

MATERIALS AND METHODS

Overview of protein network-based localization prediction

We integrated three major types of features to predict the localization of a protein, which we term ***S***, ***N*** and ***L*** (Figure 1). ***S*** (single protein) features, nine in total, were used to describe various characteristics of the protein. Seven of the nine ***S*** features were extracted from the protein's primary sequence, depicting its amino acid composition and chemical properties. Occurrences of known signaling motifs in the primary protein sequence, downloaded from cross-references in UniProt or FlyBase, was also used as one ***S*** feature. The final ***S*** feature encoded functional annotations of the protein downloaded from the Gene Ontology database. ***N*** and ***L*** are network-dependent: ***N*** summarizes the ***S*** features of the protein's extended network neighborhood, while ***L*** represents the distribution of known localizations in the neighborhood. Our modified *k*-Nearest Neighbor classifier, called DC-kNN, integrates the diverse information of all these features for each localization in a Divide-and-Conquer manner, in which a single kNN classifier is built using each type of feature and the predictions are made through majority voting of the kNN classifiers. A protein can be assigned to multiple localizations if the protein has an estimated probability over a meaningful threshold for each localization.

To generate the network features (***N*** and ***L***), we pooled protein–protein interactions for *SC* from the BioGRID (BiG) (37), the Database of Interacting Proteins (DIP) (38), and the Saccharomyces Genome Database (SGD) (39). Known localizations of 3914 proteins from Huh *et al*. (1) were used for ***L*** features (Table 1). Most interactions (>57%) in the protein networks connected known co-localized protein pairs, which implies a high degree of correlation between interaction and localization (Figure 2a and Supplementary Table S5; $P << 10^{-16}$ compared to 100 random networks of same topology). Among the three databases, BiG has the largest coverage and highest enrichment of co-localized proteins. We also found that proteins in some localizations (e.g. endoplasmic reticulum) tend to interact with proteins in different localizations (e.g. vacuole). To reflect the possibility

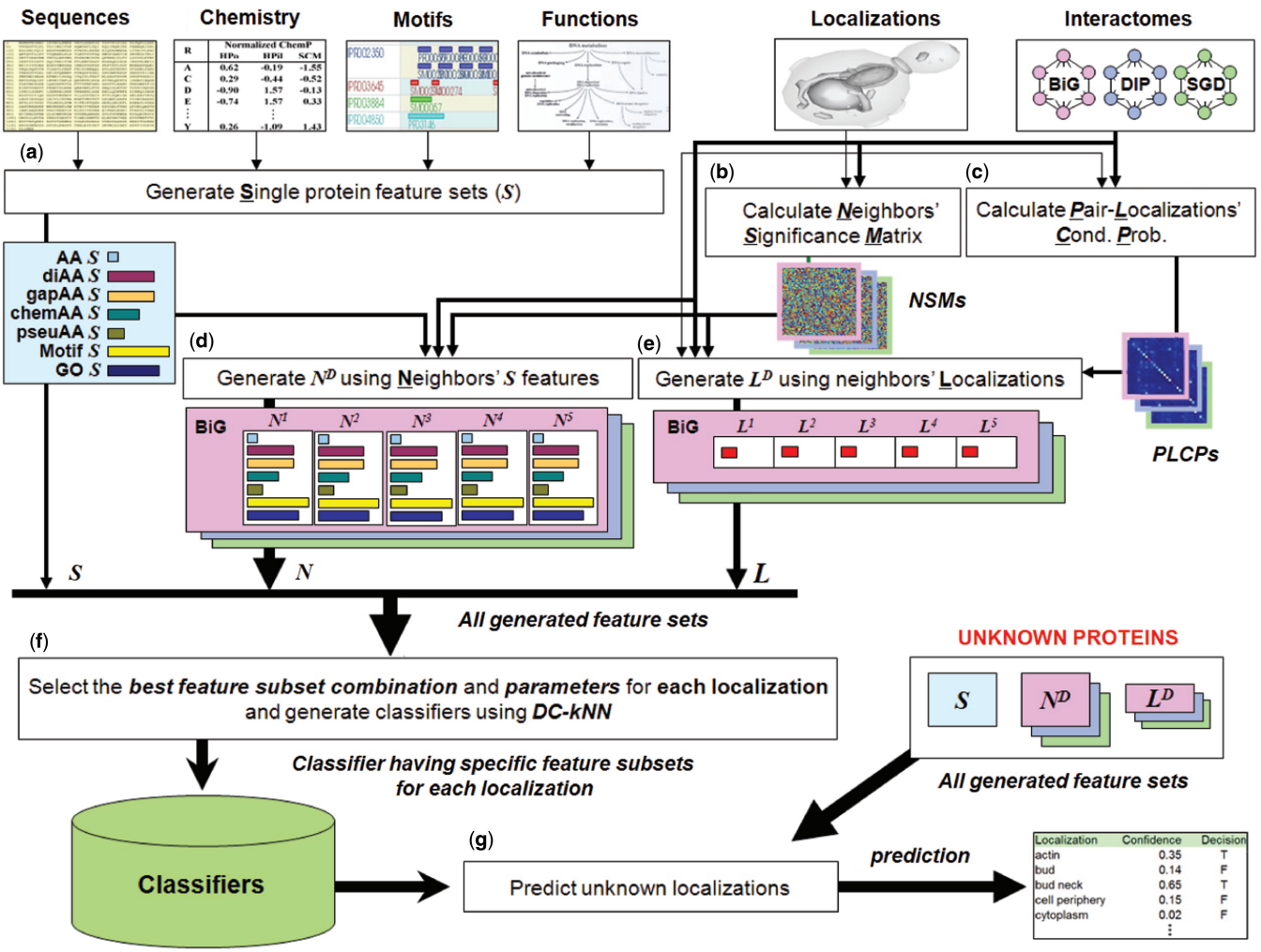

Figure 1. Schematic overview of the integrated network-based framework. (**a**) Generation of single-protein feature vectors (***S***s). Nine kinds of S_i (AA, diAA, gapAA, three kinds of chemAA, pseuAA, Motif and GO) were generated for each protein P_i based on its sequence, chemical properties, motifs and functions. (**b**) Calculation of Neighbors' Significance Matrixes (**NSM**s). These were calculated based on the number of distinct localizations covered by proteins falling along the path with the highest weight from a target protein to a neighbor protein (see Materials and methods section). (**c**) Calculation of **PLCP**s. They were calculated based on a weighted counting with normalization (see Materials and methods section). (**d**) Generation of network feature vector N_i^D. Each N_i^D was generated using up to D-th neighborhood's ***S***s with neighbors' significance degrees from **NSM**s. (**e**) Generation of network feature vector L_i^D. Each L_i^D was generated using P_i's network neighbors up to distance D, weighted by **NSM**s and **PLCP**s to reflect each neighbor's significance and the conditional probabilities of interactions between localization pairs, respectively. (**f**) Model selection for each localization. The best combination of feature sets was selected for each localization based on a forward approach with the DC-kNN classifier. (**g**) Prediction of unknown localizations. After generating all feature vectors using all known localization and network information, a confidence degree and a decision (on whether an unknown protein has a specific localization or not) were computed for each localization.

of interacting pairs being in different localizations, we incorporated such conditional probability into the ***L*** features.

Localization and network data

For *SC*, we downloaded the localization data of Huh *et al.* (1), who used GFP-tagging experiments to annotate 3914 proteins with up to 22 distinct localizations (Table 1). The 22 localizations are actin (actin cytoskeleton), bud, bud neck, cell periphery, cytoplasm, early Golgi (early Golgi/COPI), endosome, ER (endoplsmic reticulum), ER to Golgi (endoplasmic reticulum to Golgi), Golgi (Golgi apparatus), late Golgi (late Golgi/clathrin), lipid particle, microtubule, mitochondrion, nuclear periphery, nucleolus, nucleus, peroxisome, punctate composite, spindle pole, vacuolar membrane and vacuole (see Supplementary Table S1 for more information). The remaining 1530 *SC* proteins have no known localization at present and were designated 'localization-unknown'. For *DM* and *HS*, we first downloaded all proteins which had sequence information in FlyBase and UniProt, respectively. We assigned localization information to the 2187 *DM* and 4570 *HS*

Table 1. Data sources integrated to predict localization information

Species	Data set	Proteins	Localizations
Localization			
Saccharomyces cerevisiae (22 localizations[a])	Localization-known proteins	3914	5184
	Localization-known and having interactions	3206	4284
	Ambiguous	237	189 335
	Localization-unknown	1530	0
Drosophila melanogaster (12 localizations[b])	Localization-known	2187	2398
	Localization-known and having interactions	1610	1778
	Localization-unknown and having interactions	5656	0
Homo sapiens (13 localizations[c])	Localization-known	4570	5251
	Localization-known and having interactions	2684	3093
	Localization-unknown and having interactions	3767	0
Species	Data set	Proteins	Interactions
Interaction			
Saccharomyces cerevisiae	BioGRID	5184	70 700
	DIP	4931	17 471
	SGD	5395	56 035
Drosophila melanogaster	BioGRID	7545	25 463
	DIP	7038	20 719
Homo sapiens	BioGRID	7378	20 968
Feature	Description		
Protein feature			
Sequences	UniProt (for *SC* and *HS*) and FlyBase (for *DM*)		
Chemical property	Hydrophobicity, hydrophilicity and side-chain mass		
Motifs	InterPro		
Functions	InterPro and GO		

Here, we only considered the proteins with sequence information.
[a]22 *SC* localizations are actin, bud, bud neck, cell periphery, cytoplasm, early Golgi, endosome, ER, ER to Golgi, Golgi, late Golgi, lipid particle, microtubule, mitochondrion, nuclear periphery, nucleolus, nucleus, peroxisome, punctate composite, spindle pole, vacuolar membrane, vacuole.
[b]12 *DM* localizations are actin, cell periphery, centrosome, cytosol, ER, golgi, lysosome, mitochondrion, nucleolus, nucleus, peroxisome, vacuole.
[c]13 *HS* localizations are actin, cell cortex, centrosome, cytosol, ER, golgi, lysosome, mitochondrion, nucleolus, nucleus, peroxisome, plasma membrane, vacuole. Further details regarding localizations and interactions of *SC*, *DM* and *HS* are in Supplementary Figure S1 and Tables S1–S4.

proteins with GO cellular component annotations. To define the corresponding set of localization unknown proteins, we identified 5656 *DM* and 3767 *HS* proteins in the BiG protein network with sequences available but that did not have known localizations, i.e. missing GO annotations. For the interaction data, we downloaded the contents of BiG, DIP, and SGD for *SC*, of BiG and DIP for *DM* and of BiG for *HS*.

Generation of single protein feature vectors (*S*)

Using sequences from UniProt for *SC* and *HS* and FlyBase for *DM*, we generated three kinds of amino acid features for each protein: amino acid composition frequencies (AA), pair-coupled amino acid frequencies (diAA) and pair-coupled amino acid frequencies with a gap (length = 1) (gapAA). AA is a vector of length 20; the diAA and gapAA vectors contain 400 elements enumerating frequencies over all ordered amino acid pairs. For incorporating chemical properties, we generated three kinds of chemical amino acid compositions (chemAA) using normalized hydrophobicity (40) (HPo), hydrophilicity (41) (HPil) or side-chain mass (42) (SCM), respectively (see Supplementary Table S6 for the normalized values of each chemical property). The chemAA compositions were computed by scanning a window of length k along the amino acid sequence ($1 \le k \le 40$) and recording the mean squared difference in the chemical property value across all window positions. The k-th element of chemAA using hydrophobicity was defined as:

$$\mathrm{HPo}(k) = \frac{1}{n-k}\sum_{l=1}^{n-k}(\mathrm{HPo}(R_l) - \mathrm{HPo}(R_{l+k}))^2,$$

where $\mathrm{HPo}(R_l)$ is the normalized hydrophobicity value of the l-th residue, and n is the length of the protein sequence. The pseudo-amino acid composition (pseuAA) (43) was generated by combining the three chemical properties into one. Formally stated:

$$\mathrm{PseuAA}(k) = \frac{1}{n-k}\sum_{l=1}^{n-k}\frac{1}{3}[U+V+W],$$

where $U = (\mathrm{HPo}(R_l) - \mathrm{HPo}(R_{l+k}))^2$, $V = (\mathrm{HPil}(R_l) - \mathrm{HPil}(R_{l+k}))^2$ and $W = (\mathrm{SCM}(R_l) - \mathrm{SCM}(R_{l+k}))^2$. For the *Motif* and *GO* feature vectors, we downloaded InterPro Motifs and GO information from UniProt (*SC* and *HS* proteins) and FlyBase (*DM* proteins). After extracting the motif or GO set using all localization-known proteins for each species, we constructed a

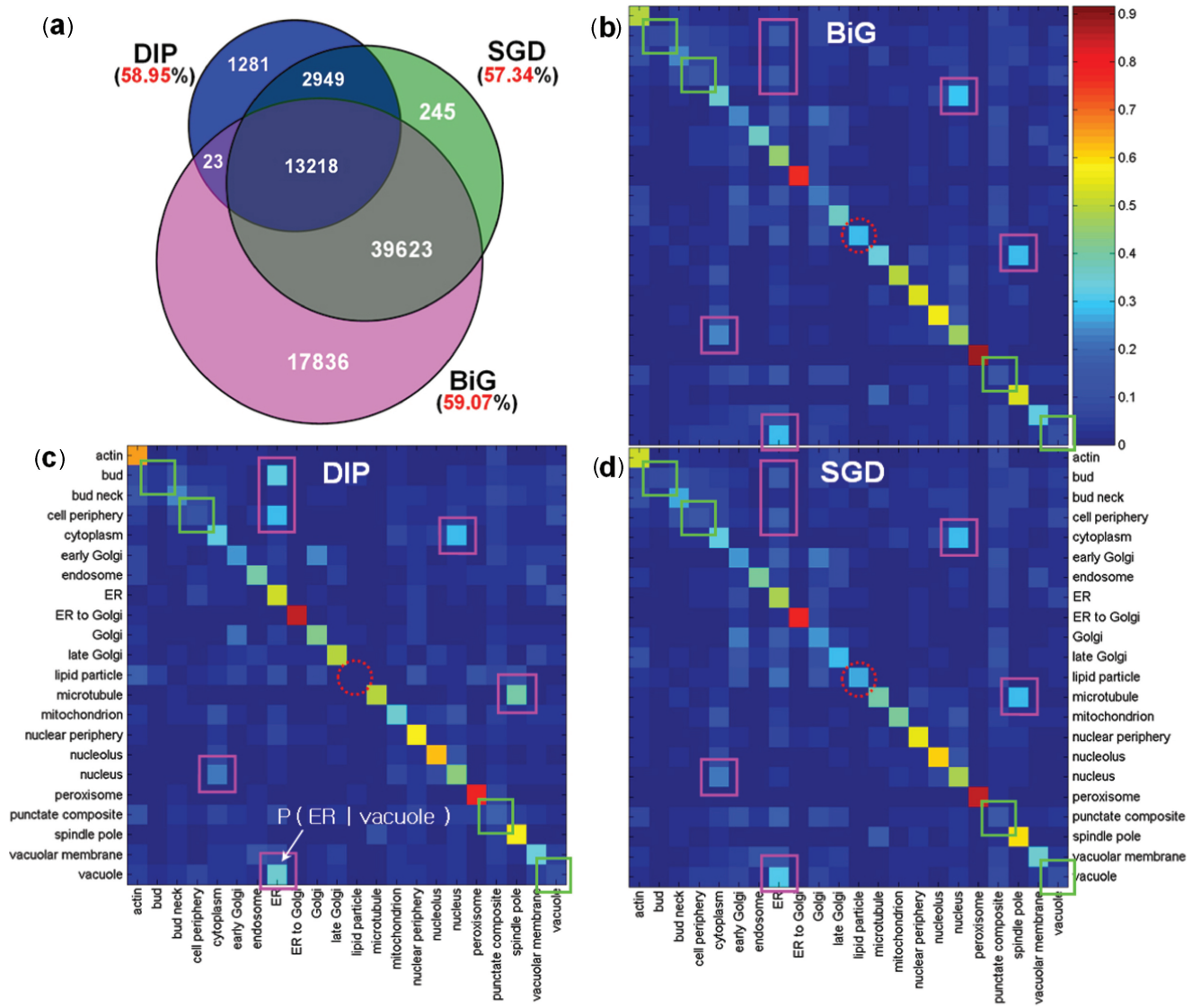

Figure 2. Correlation between known localizations and protein interactions of yeast proteins. (**a**) The number of interactions (inside the circles) and the fraction of interactions whose proteins share localization information (outside the circles) of three interaction databases: BiG, DIP and SGD. (**b–d**) The **PLCP**s of BiG, DIP and SGD, respectively. Given a protein at a particular localization (row), each cell corresponds to the conditional probability of the localization of its interacting partners (column). The squares on the diagonal (or off-diagonal) indicate the locations with relatively low (or high) degrees of location-sharing interactions within (or between) locations; the dotted circles on the diagonal indicate different patterns among three interaction databases for proteins in the lipid particle.

binary feature vector (5,6) in which each element was set to '1' if the protein had the corresponding motif (or GO) annotation, otherwise '0'. Note that GO terms also include cellular component annotations, which are also used as class labels especially for *DM* and *HS*. Thus, to reduce circularity we omitted these annotations while generating the *GO* feature vectors even though most previous studies used all three branches of GO terms (5,44).

Pair-localizations' conditional probability

We calculated a pair-localizations' conditional probability (**PLCP**) matrix for each protein network (BiG *SC*, DIP *SC*, SGD *SC*, BiG *DM*, DIP *DM* and BiG *HS*) to capture the probability of a protein being in localization l_j given that its interaction partner is in localization l_i:

$$P(l_j|l_i) = \frac{I_{ij}}{\sum_k I_{ik}},$$

I_{ij} is the normalized number of interactions between protein pairs spanning (l_i and l_j). I_{ij} is defined as:

$$I_{ij} = \frac{\sum_{a\in l_i, b\in l_j (a\neq b)} \frac{\phi(a,b)}{N(a)\times N(b)}}{N(l_i)+N(l_j)}$$

where $N(l_i)$ is the total number of proteins in localization l_i, $N(a)$ is the number of localizations in protein a, and

$\phi(a,b)$ is '1' if there is an interaction between proteins a and b; otherwise, zero.

Network-dependent interacting protein-group feature generation

In this study, we generated two kinds of network feature vector: $\boldsymbol{N^D}$ and $\boldsymbol{L^D}$. $\boldsymbol{N^D}$ of protein P_i is defined as the weighted average of the $\boldsymbol{S}$ feature vectors over proteins up to distance D from P_i in the network, including P_i itself (called the D-th neighborhood of P_i and represented by the variable C_i^D):

$$\mathbf{H}_i^D = \frac{1}{\sum w_{ki}} \sum_{P_k \in C_i^D} w_{ki} S_k.$$

The weightings w_{ki}, which make up the Neighbors' Significance Matrix (Figure 1), represent the significance of neighbor P_k, defined as:

$$w_{ki} = \frac{1}{\Psi_{ki} + \rho},$$

where Ψ_{ki} is the number of distinct localizations covered by proteins along a path from P_i to P_k, and ρ is a pseudo-counter for handling incompleteness of localization data (in this study $\rho = 1$ for *SC*, $\rho = 2$ for *HS* and $\rho = 3$ for *DM*; different values were used because the portions of known localizations for *DM* and *HS* are less than that of *SC*—see Figure 6b). Note that we assigned max w_{ki} among multiple paths from P_i to P_k and assigned less weight on a neighbor protein that interacts with other proteins having many distinct localizations.

$\boldsymbol{L}_i^D$ is a vector representing the probability that P_i has each of the 22 localizations, given the D-th neighborhood of protein P_i and considering the probabilities of interaction between proteins in distinct localization pairs:

$$\mathbf{L}_i^D = \left[L_i^D(1), \ldots\ L_i^D(y), \ldots\ L_i^D(22)\right]$$

$$L_i^D(y) = \frac{1}{\sum w_{ki}} \sum_{P_k \in C_i^D} \left[w_{ki} \max_{l_x \in \Gamma_k} p(l_y | l_x) \right]$$

where l_x is one element of the localization set $\boldsymbol{\Gamma}_k$ of P_k, and $p(l_y|l_x)$ is the conditional probability of label l_y given the label l_x (from the **PLCP** matrix). Note that we choose the maximum value among multiple choices for the conditional probability of each localization, owing to the multiple localization property. Moreover, to satisfy the symmetric property, we also include the single protein feature vector of input protein P_i when generating network feature vectors.

Divide-and-Conquer k-Nearest Neighbor Classifier

The DC-kNN has three main steps: dividing, choosing, and synthesizing. In the dividing step, the full feature vector is divided into m meaningful feature subvectors. In this study, each single protein feature set and each network-dependent protein group feature set were treated as meaningful sub-vectors, yielding $m = 69$ subvectors in total for yeast: the 9 $\boldsymbol{S}$ vectors (AA, diAA, gapAA, three kinds of chemAA, pseuAA, Motif, GO), the 54 $\boldsymbol{N}$ vectors [= 9 $\boldsymbol{S}$ vectors × 2 (up to second neighborhood) × 3 (the number of network databases)], and the 6 $\boldsymbol{L}$ vectors [= 2 (up to second neighborhood) × 3 (the number of network databases)]. In the choosing step, the k-nearest neighbors are chosen for each protein and subvector (in this study, $k = 5$). Finally, the synthesizing step averages the m sets of k neighbors with a weight on each set, and it generates a confidence for each label by means of a normalization process with m and k. Formally, the confidence μ_l for label l is defined as:

$$\mu_l = \left[\frac{1}{k} \sum_m n_{ml} \times \phi_m \right]^{\frac{1}{\sqrt{m}}},$$

where n_{ml} is the number of k-nearest neighbors that have label l according to sub-vector m. $\phi_m (\sum_m \phi_m = 1)$ is the weight of the m-th subvector. Instead of using all sub-vectors, DC-kNN finds the best combination of feature subvectors for each label, based on a forward approach. At each iteration, DC-kNN chooses the most predictive feature subvector among those remaining, i.e. the vector that shows the best AUC when added to the previously selected feature subvectors. In the first iteration, feature subvectors are used individually for finding the most predictive one. For the weights ϕ_m, DC-kNN uses the AUC obtained using each feature subvector alone. DC-kNN produces a confidence degree (0–1) and a decision on whether a protein has a specific localization or not, using a threshold based on a false positive rate (in this study, <0.01).

Microscopic localization analysis

Yeast cells grown to mid-logarithmic phase in SC medium were microscopically analyzed in 96-well glass bottom microplates (Whatman, Florham Park, NJ, USA) pre-treated with concanavalin A (Sigma, St. Louis, MO, USA) to ensure cell adhesion. Microscopy was performed on a Zeiss Axiovert 200M inverted microscope with a Plan-NeoFluar 100×/1.3 NA oil immersion objective. Images were recorded on a Zeiss Axiocam MRm with 2 × 2 binning. Fluorescence images for GFP were taken using a standard fluorescein isothiocyanate filter set (excitation band pass filter, 450–490 nm; beam splitter, 510 nm; emission band pass filter, 515–565 nm).

RESULTS AND DISCUSSION

Network information improves localization prediction in yeast

We compared the predictive performance of different features during prediction of localization: $\boldsymbol{S}$ features only, $\boldsymbol{N}$ features only, $\boldsymbol{L}$ features only, all three features together ($\boldsymbol{S} + \boldsymbol{N} + \boldsymbol{L}$) and random guesses. DC-kNN classification was used in all cases, and performance was evaluated using the technique of leave-one-out cross-validation (LOOCV). In every run of LOOCV, the known localization of one of the 3914 *SC* proteins in Huh *et al.* (1) was

designated as 'test' data and withheld during classifier training.

Three metrics, *Top-K*, *Total* and *Balanced*, were used to summarize the performance of the 3914 runs. The *Top-K* measure is the fraction of correctly predicted runs, in which the prediction is considered correct if at least one of the known localizations of the test protein is included in the top-K predicted localizations. We used $K = 3$ assuming most yeast proteins have less than or equal to three localizations (6). The *Total* measure is the fraction of correctly predicted localizations in the 3914 runs, counting all predictions for all proteins. The *Balanced* measure calculates the averaged fraction of correctly predicted localizations in distinct localizations (see Supplementary Figure S2 for the metrics used). The *Balanced* measure is used because predictions based on localization categories with few proteins are usually not as good as predictions based on localization categories with many proteins annotated. For the random guesses, we randomly permuted the assignment of localizations to proteins preserving both the number of localizations per protein and the number of proteins per localization; the measures (*Top-3*, *Total* and *Balanced*) were averaged over 30 runs.

Although all classifiers were clearly better than random (based on the background distribution of proteins in the 22 localizations; Figure 3a), the combination of all three features provided the highest predictive accuracy regardless of the measure. Moreover, according to the *Balanced* metric, either of the network features $\boldsymbol{N}$ or $\boldsymbol{L}$ achieved higher accuracy than $\boldsymbol{S}$ features. These results suggest that when the number of proteins was not sufficient to learn sequence-level rules for classifying smaller compartments like 'bud' or 'peroxisome', interaction networks provided one alternative to amplify the weak signals encoded in the individual protein sequences.

In all of the above cases, the network neighborhood was defined as a protein's immediate interactors ($\boldsymbol{N}^{\mathbf{1}}$ or $\boldsymbol{L}^{\mathbf{1}}$, designating network distance = 1). Next, we explored the impact of expanding a protein's network neighborhood to incorporate not only immediate neighbors, but all proteins at network distances up to and including distance D. As seen in Figure 3b–d, incorporating network information up to distance 2 generally improved the accuracy of the amino acid, chemical AA properties and GO features. However, network distances larger than 2 did not have a significant increase in performance, which is understandable given the diameter of the yeast network was six. Similar findings were observed for the $\boldsymbol{L}$ features (Figure 3e). The $\boldsymbol{L}$ features alone (*Total* accuracies range from 60% to 66% depending on the network used) outperformed any kind of $\boldsymbol{S}$ feature (42–55%), but their accuracies did not increase significantly when more than distance 2 neighbors were included.

Interestingly, a network pooled from all three interaction databases did not improve the performance over any single network alone (Figure 3b–e). It achieved equivalent performance as the SGD network and sometimes worse than the BiG network, indicating that the network quality played a bigger role than the coverage in generating useful $\boldsymbol{N}$ and $\boldsymbol{L}$ features. Overall, the BiG network had the best performance.

The best combination of single-protein features and network features for each localization

Using a subset of features may reduce the possibility of overfitting and therefore lead to a more robust classifier (45,46). To further optimize the predicted localizations, we applied a forward selection which combined feature sets of high predictive power from a pool of $\boldsymbol{S}$, $\boldsymbol{N}$ and $\boldsymbol{L}$ features from up to distance 2 network neighborhoods. During feature set selection, we used the common measure of Area Under receiver operator characteristic Curve (AUC) (47,48) to rank the predictive power of features and also to evaluate the performance of the resulting classifiers. To reduce overfitting further, we withheld two examples from each training round of cross-validation, and then used one for feature selection and one for performance reporting. Without feature selection, DC-kNN with all single-protein $\boldsymbol{S}$ features achieved 0.65 AUC averaged from the prediction of the 22 compartments. This accuracy increased to 0.79 if feature set selection for each localization was applied during classifier training using all single-protein features (Figure 4a).

Lastly, we explored the effect of selecting the best combination of single-protein features $\boldsymbol{S}$ and network features $\boldsymbol{N}$ and $\boldsymbol{L}$ for each localization separately. We found that selecting different features per localization using single and network features resulted in a dramatic increase in performance, with average AUC of 0.94 for the 22 localizations (see Supplementary Figures S4–S6 for the forward feature set selection, the ROC curves of each approach, and the selected feature sets for each compartment, respectively). This means that the combinatorial effect between single-protein features and network features is indispensable for capturing functional characteristics of proteins.

Another issue in the localization prediction of proteins might be the influence of homologous data in training data. To evaluate the influence of sequence similarity in the developed network-based approach, we checked the performance of DC-kNN with only nonhomologous yeast proteins (see Supplments.doc for more information). We observed similar performance (average AUC value of 0.94) with the previous result with all known yeast proteins. It implies that the network-based DC-kNN is insensitive to the presence of close sequence homologs in a training data set.

Novel localization predictions can revise previous high-throughput experiments

Based on its good performance, we applied this last method to comprehensively predict 5184 localizations for 3914 yeast proteins. Although these yeast predictions were in good agreement with the GFP localization experiments performed by Huh *et al*. (1) (as expected since the Huh data were used as features), to our surprise we found that for 61 proteins the predicted localizations were novel (Supplementary Tables S7 and S8). For example, Noc4/Ypr144c and Utp21/Ylr409c were localized to the nucleus by Huh *et al*. (1), whereas our predictions produced the highest signal (5×10^{-4} false positive rate for Noc4 and 1×10^{-3} for Utp21) at the nucleolus. To determine whether a nucleolar localization could be corroborated

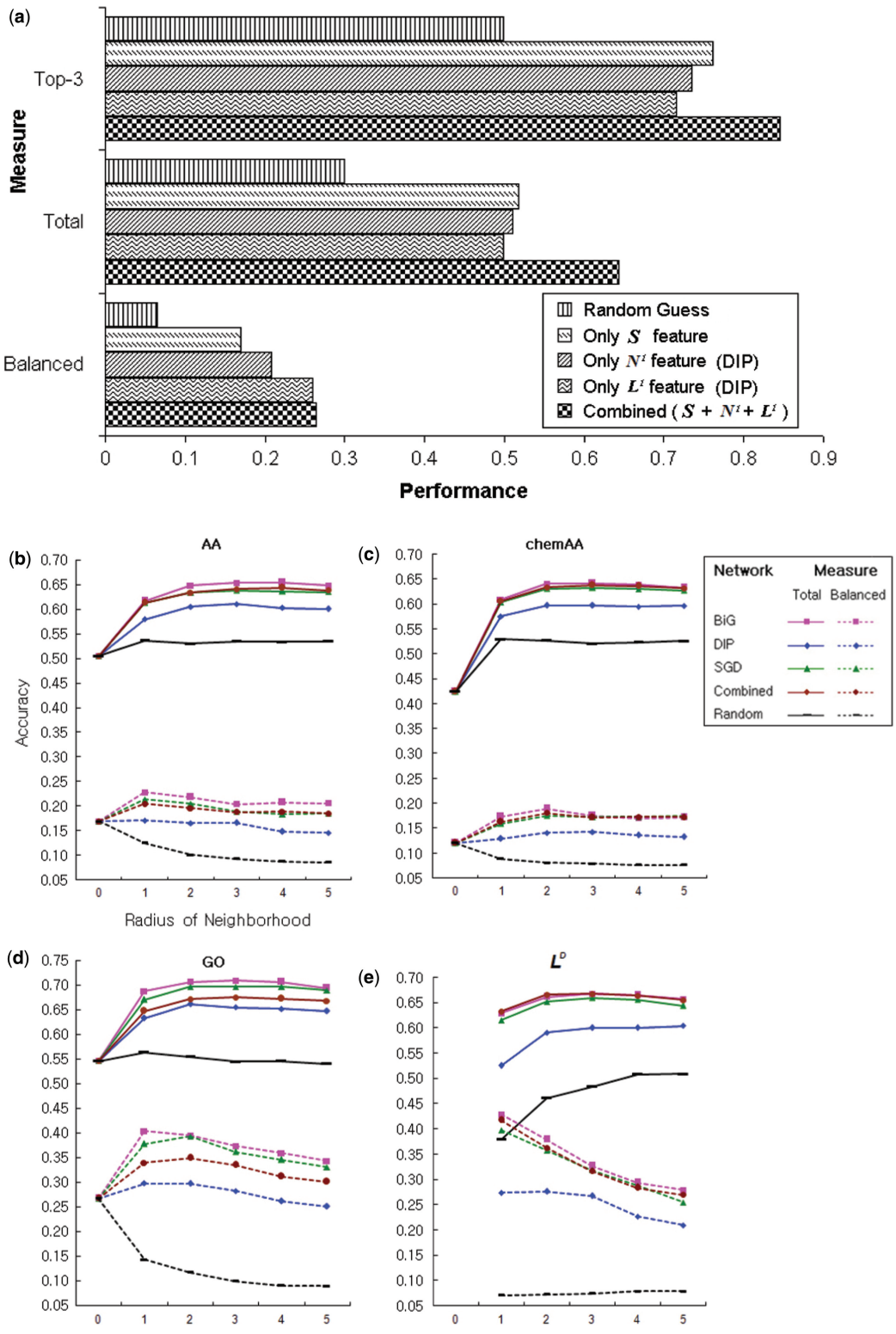

Figure 3. Usefulness of protein interaction networks. (**a**) The performance of five cases, including (i) random guess of localization, (ii) $\boldsymbol{S}$ features only, (iii) $\boldsymbol{N^I}$ only, (iv) $\boldsymbol{L^I}$ only and (v) all three kinds of features. (**b–e**) The performance of the $\boldsymbol{N^D}$ features for amino acid frequencies (**b**), chemical amino acid properties (**c**), and GO terms (**d**) as well as performance of the feature $\boldsymbol{L^D}$ (**e**). Performance is based on the five interaction networks BiG, DIP, SGD, Combined, and Random (different color curves). The performance of other $\boldsymbol{N^D}$ network features is shown in Supplementary Figure S3. The *x*-axis is the radius of neighborhood D; $D = 0$ means only the single protein feature vector $\boldsymbol{S}$ was used, which is a conventional approach. For Combined, the three interactome datasets BiG, DIP and SGD were pooled into a single network. For Random, localizations were randomly assigned on the BiG network. The solid lines and the dotted lines represent the *Total* and *Balanced* measures, respectively.

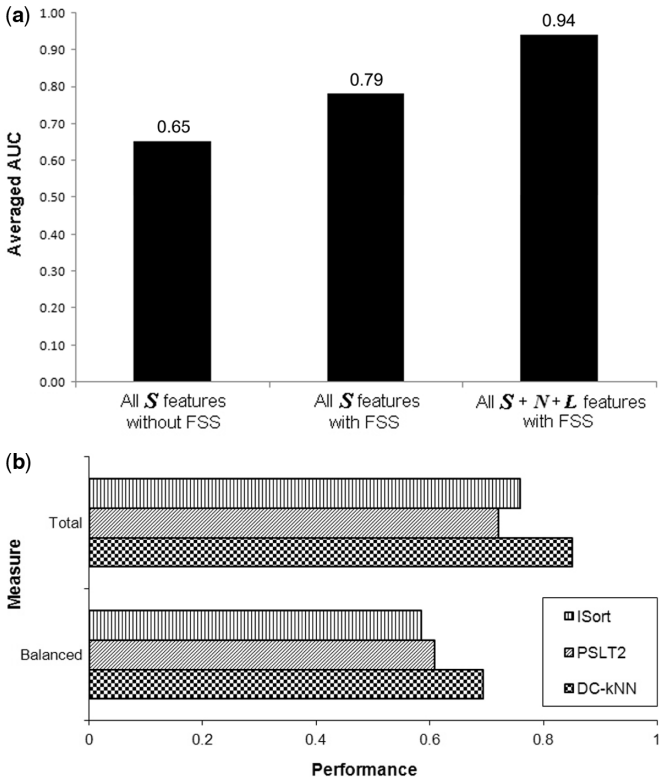

Figure 4. Performance of the network-based approach. (**a**) The averaged AUC values of three cases: (i) all ***S*** features without feature set selection (FSS), (ii) all ***S*** features with FSS and (iii) all ***S***, ***N*** and ***L*** features with FSS for each localization. (**b**) Performance comparison with two well-known methods. Performance is computing using the *Total* versus *Balanced* metrics (top three versus bottom three bars).

experimentally, we re-examined the strains containing GFP-tagged Noc4 and Utp21 using fluorescence microscopy (see Materials and methods section). The resulting images show that both proteins do indeed accumulate at the nucleolus with some spread to the nucleoplasm (Figure 5a and b). In some cases, therefore, it appears that network-based predictions can correct or complement the image readouts of high-throughput experiments. This power owes mainly to the fact that our framework synthesizes evidence from multiple interacting partners. For example, Noc4 interacts with many other proteins in the nucleolus, hence the prediction (Figure 5c).

In Huh *et al.* (1), 237 *SC* proteins had ambiguous image readouts for determining their localizations. Among these, 80 proteins were nonetheless annotated with 'low confidence' localizations and 157 were never annotated (1). Moreover, an additional 1530 yeast proteins could not be localized by the previous experiments owing to low GFP signals (1). We used the DC-kNN network-based classifier to predict the localization of all of these proteins (Supplementary Figure S7 and Tables S9–S10). For the 80 'low confidence' proteins in Huh *et al.* (1), our predicted localizations significantly overlapped with their assignments (Supplementary Figure S7c; $P < 2.0 \times 10^{-31}$ based on a hypergeometric distribution). We also found significant overlap between our predictions and the literature-curated annotations recorded in the cellular component branch of the GO database (see Supplementary Figure S7d and Table S11 for the overlap degree and the mapping relationship between 22 localizations and GO terms, respectively; $P < 2.6 \times 10^{-71}$).

Comparison with previous methods

We compared DC-kNN with two popular methods, ISort (5) and PSLT2 (7,17), for the prediction of yeast protein localization. ISort (5) is one of most comprehensive sequence-based methods and also the first of the few machine-learning-based methods to predict more than 15 compartments. PSLT2 (7) is a method that previously incorporated protein interaction networks into localization prediction. In the original PSLT2 paper (7), the authors demonstrated its accuracy in predicting *SC* proteins in nine general compartments. Therefore, we ran our method and ISort (5) for the same nine compartments with the same data used in the PSLT2 paper (7). Using both sequence and network features, DC-kNN significantly outperformed ISort and PSLT2 based on the *Total* and *Balanced* measures [*Top-K* and AUC measurements are not available in the PSLT2 paper (7)] (Figure 4b). Between ISort and PSLT2, ISort had higher *Total* accuracy but PSLT2 surpassed ISort in terms of the *Balanced* measure, which down-weights bigger compartments with more proteins (see Supplementary Table S12 for the performance of each compartment among three methods).

Extrapolation to higher eukaryotes

Given the power of protein network information to predict protein localization, an important question is whether a network-based approach can be extended to other eukaryotes with less network coverage than yeast. To address this question, we ran a series of simulations in which increasing numbers of interactions in the yeast network were successively removed. As expected, the performance of DC-kNN decreased as less network information was available (Figure 6a). However, the rate of decrease was gradual, such that when the average degree of the network was reduced by approximately half (27 versus 13), the associated decrease in AUC was 0.94–0.91. At an average degree of five, the AUC was still ~0.89. We note that the available protein networks for worm, fly and human are in this range (average degrees from 3 to 7; see Figure 6a). Thus, these results suggest that the protein network-based DC-kNN will achieve high accuracy in predicting protein localization in these species. At average degrees below three, the performance dropped more precipitously to approach 0.79, the AUC achievable without network information (***S*** features only).

Another potential problem is that in eukaryotes other than yeast, few known protein localizations are available for classifier training. Thus, our second simulation was to test the robustness of prediction as the number of proteins with known localization data was decreased. As expected, the AUC decreased when less localization data were available (Figure 6b), but with an even slower rate of degradation than that observed for loss of interaction

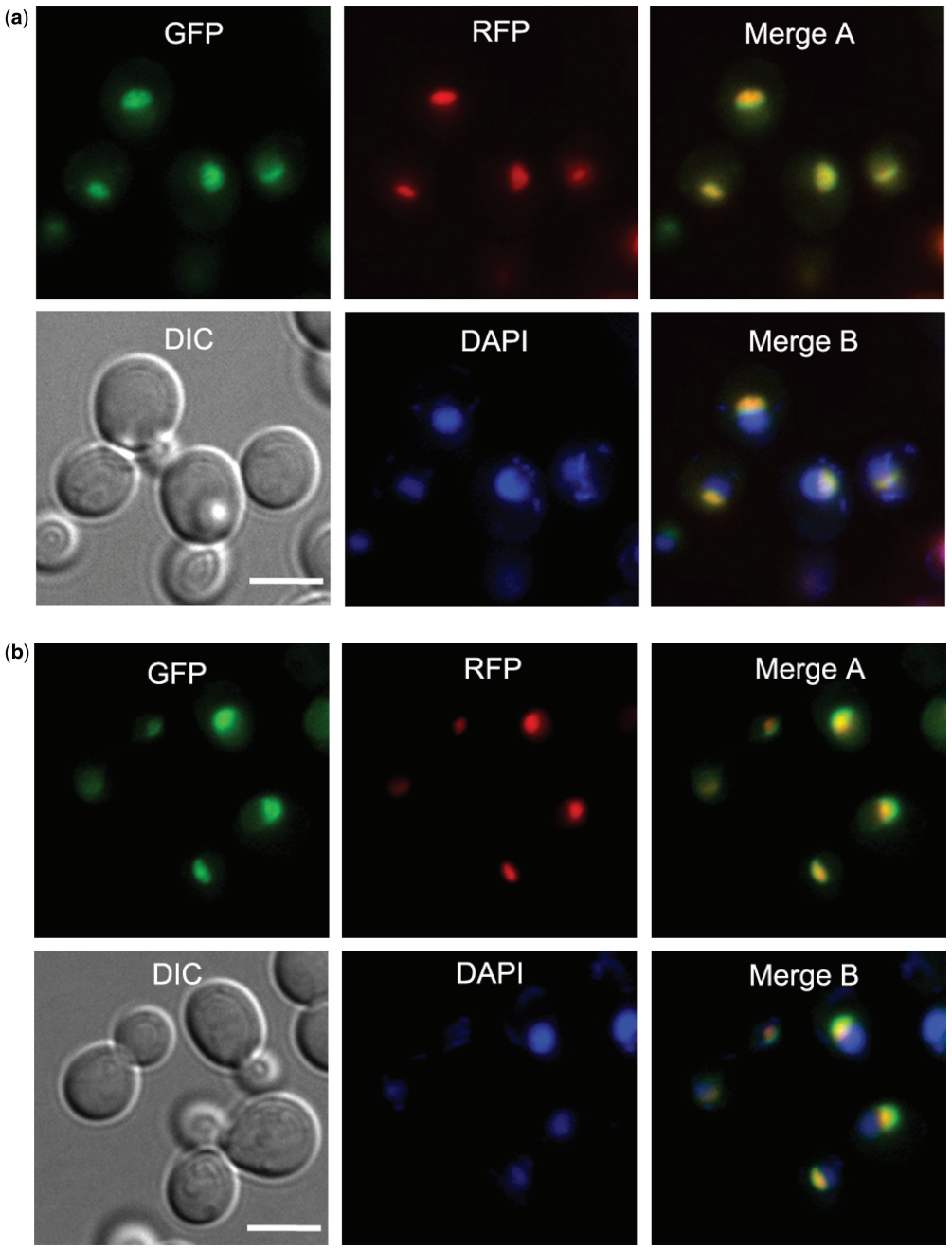

Figure 5. Validation of novel localizations for yeast proteins. New localization images for two yeast proteins, Noc4/Ypr144c (**a**) and Utp21/Ylr409c (**b**), for which the network-based prediction (nucleolus) was different than previously measured (nucleus) (1). The near-complete overlap area between the GFP and RFP images ('Merge A'), marking the protein and nucleolus, respectively, is consistent with a nucleolar localization (Sik1-RFP was used as a nucleolus marker). Here, DAPI is used for marking the nucleus, and 'Merge B' is the overlap among GFP, DAPI and RFP images. (**c**) Proteins that interact with Noc4/Ypr144c and their localizations. The values in the upper-left box represent the interacting protein pairs' localization purity (IPLP, or enrichment) among interacting protein pairs for distinct localizations (see the 'Supplements.doc' for more information). Panel (c) is drawn using Cytoscape (55).

data (Figure 6a). Dramatically, with only 1% of network proteins having known localizations, the network-based approach still achieved ~0.83 AUC, which is significantly higher than the ~0.65 AUC obtained from a conventional sequence-based approach. The improvement results from both the consideration of network features and the feature selection implemented in DC-kNN. These simulations suggest that the proposed network-based method can be applied to predict localization of proteins in higher eukaryotes where only little protein network information is available and only few proteins have previously determined localizations.

To cross-check these simulation results, we applied the proposed framework to predict protein localizations in

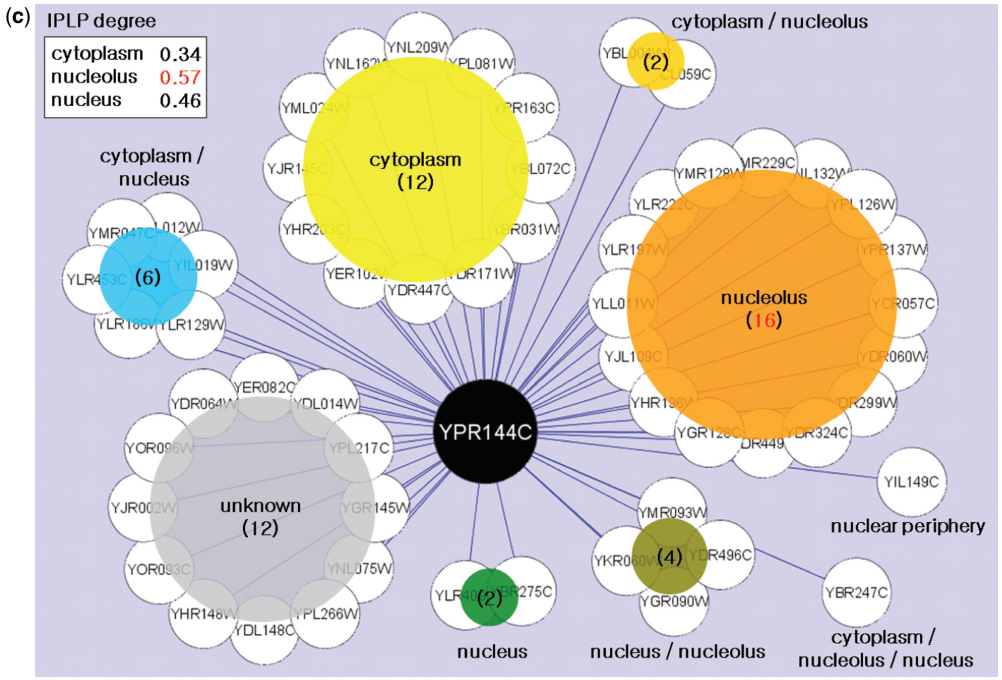

Figure 5. Continued.

both fly and human. The currently available fly and human networks, containing 25 463 and 20 968 interactions among 7545 and 7378 proteins, respectively, were downloaded from BiG. Because no high-throughput experimental studies have been conducted to measure the localizations of fly and human proteins, we trained the classifier using literature-curated protein localizations documented in the Cellular Component branch of the GO database. According to GO, 1709 fly and 2684 human proteins in the BiG network have known localizations covering 12 (fly) and 13 (human) cellular compartments in total. Approximately 77% (fly) or 64% (human) proteins had no known localizations, in contrast to only 33% of proteins in yeast (Supplementary Table S13). Nonetheless, consistent with our above simulation results, DC-kNN achieved ~0.88 (fly) or ~0.95 (human) AUC in cross validation (red 'X's in Figure 6b). In terms of network coverage, the performance in human was slightly higher than predicted in simulation (red 'X's in Figure 6a). (See Supplementary Figure S8 for forward feature set selection of fly and human and Supplementary Figures S9–S10 for selected feature sets for each compartment.) Overall, we predicted 7058 (fly) and 4366 (human) new localizations for proteins with no localizations previously known (see Supplementary Tables S14–S15 for all predicted results and Supplementary Figure S11 for distribution of the results).

In this work, we obtained an average AUC of 0.94 for yeast, 0.88 for fly and 0.95 for human (see the 'Supplements' for the discussion of the localization-specific predictions of yeast, fly and human proteins). The high performance of the proposed approach results from both the consideration of network features, in addition to single protein features, and the feature selection implemented in DC-kNN. The performance may be further improved by efforts to specify further details about the type of relationship each interaction represents. For instance, interactions fall into specific biological categories, including physical binding events, genetic interactions such as synthetic lethals or suppressor relationships, and functional associations. Each of these interaction types may have different capacity to predict specific protein localizations. Moreover, protein interactions are dynamic according to external stimuli or environmental conditions (49,50). Where condition-specific expression or interaction data are available, it would be of high interest to predict dynamic changes in protein localization. It is increasingly recognized that such changes are the cornerstone of many cellular regulatory events (51–54), such as the translocation of transcription factors to the nucleus or the trafficking of proteins to the vacuole or cellular membrane.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Dr Michael Hallett at McGill University and Dr Jenn-Kang Hwang at National Chiao Tung University for sharing their localization data sets, and Dr Kwang Hyung Lee and Dr Doheon Lee at KAIST for valuable discussion on this research. Our special thanks to Hyun-Min Kang at UCLA for advice on statistical analysis of results and Hye-Young Cho at KAIST for preparation of localization data sets.

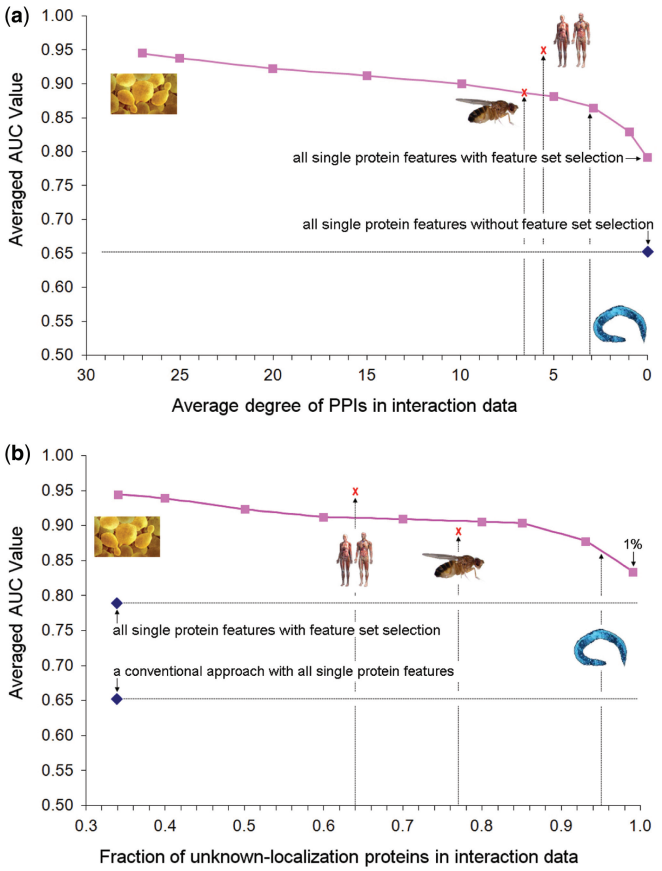

Figure 6. Performance of predicting yeast protein localization as the available interaction (**a**) or localization (**b**) data are eroded. In (a), interactions were randomly deleted to reduce the average degree of the yeast PPI network to that specified (*x*-axis). In (b), known yeast protein localizations were randomly deleted. In either case, AUC was estimated using the leave-one-out approach. To avoid over fitting, the selected feature sets were taken from Supplementary Figure S6 and not re-optimized. Worm, fly, and human were mapped onto these yeast performance curves using the average degree of their available protein networks (a) or the fraction of known localizations for network proteins (b). The blue diamond represents the performance of a conventional approach using all nine single protein features without feature set selection. The red 'X' marks denote the performance of the proposed method when applied to recover known protein localizations in fly and human, using LOOCV.

FUNDING

Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2006-352-D00171, partially); NIGMS (GM070743 to T.I.); Korea Science and Engineering Foundation (#2006-04090 to B.L.); 21C Frontier Functional Proteomics Project (FPR08A1-060) funded by the Ministry of Education, Science and Technology, Republic of Korea. Funding for open access charge: NIH/NIGMS (NIGMS is the National Institute of General Medical Sciences); grant no. 1 R01 GM070743.

Conflict of interest statement. None declared.

REFERENCES

1. Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
2. Matsuyama,A., Arai,R., Yashiroda,Y., Shirai,A., Kamata,A., Sekido,S., Kobayashi,Y., Hashimoto,A., Hamamoto,M., Hiraoka,Y. *et al.* (2006) ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. *Nat. Biotechnol.*, **24**, 841–847.
3. Ross-Macdonald,P., Coelho,P.S., Roemer,T., Agarwal,S., Kumar,A., Jansen,R., Cheung,K.H., Sheehan,A., Symoniatis,D., Umansky,L. *et al.* (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, **402**, 413–418.
4. Kumar,A., Agarwal,S., Heyman,J.A., Matson,S., Heidtman,M., Piccirillo,S., Umansky,L., Drawid,A., Jansen,R., Liu,Y. *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev.*, **16**, 707–719.
5. Chou,K.C. and Cai,Y.D. (2005) Predicting protein localization in budding yeast. *Bioinformatics*, **21**, 944–950.
6. Lee,K., Kim,D.W., Na,D., Lee,K.H. and Lee,D. (2006) PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic Acids Res.*, **34**, 4655–4666.
7. Scott,M.S., Calafell,S.J., Thomas,D.Y. and Hallett,M.T. (2005) Refining protein subcellular localization. *PLoS Comput. Biol.*, **1**, e66.
8. Bhasin,M. and Raghava,G.P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **32**, W414–W419.
9. Drawid,A. and 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**, 1059–1075.
10. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
11. Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., deFays,K., Lambert,C., Nakai,K. *et al.* (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.*, **31**, 3613–3617.
12. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
13. Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
14. Huang,Y. and Li,Y. (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics*, **20**, 21–28.
15. Park,K.J. and Kanehisa,M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656–1663.
16. Shatkay,H., Hoglund,A., Brady,S., Blum,T., Donnes,P. and Kohlbacher,O. (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics*, **23**, 1410–1417.
17. Scott,M.S., Thomas,D.Y. and Hallett,M.T. (2004) Predicting subcellular localization via protein motif co-occurrence. *Genome Res.*, **14**, 1957–1966.
18. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.
19. Nakashima,H. and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.*, **238**, 54–61.
20. Wang,M., Yang,J., Xu,Z.J. and Chou,K.C. (2005) SLLE for predicting membrane protein types. *J. Theor. Biol.*, **232**, 7–15.
21. Chou,K.C. and Elrod,D.W. (1999) Protein subcellular location prediction. *Protein Eng.*, **12**, 107–118.
22. Yuan,Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.*, **451**, 23–26.
23. Gardy,J.L. and Brinkman,F.S. (2006) Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.*, **4**, 741–751.
24. Mott,R., Schultz,J., Bork,P. and Ponting,C.P. (2002) Predicting protein cellular localization using a domain projection method. *Genome Res.*, **12**, 1168–1174.
25. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

26. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
27. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.
28. Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.
29. Krogan,N.J., Peng,W.T., Cagney,G., Robinson,M.D., Haw,R., Zhong,G., Guo,X., Zhang,X., Canadien,V., Richards,D.P. *et al.* (2004) High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell*, **13**, 225–239.
30. Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al.* (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.
31. Tong,A.H., Lesage,G., Bader,G.D., Ding,H., Xu,H., Xin,X., Young,J., Berriz,G.F., Brost,R.L., Chang,M. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
32. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
33. Mendelsohn,A.R. and Brent,R. (1999) Protein interaction methods—toward an endgame. *Science*, **284**, 1948–1950.
34. Formstecher,E., Aresta,S., Collura,V., Hamburger,A., Meil,A., Trehin,A., Reverdy,C., Betin,V., Maire,S., Brun,C. *et al.* (2005) Protein interaction mapping: a Drosophila case study. *Genome Res.*, **15**, 376–384.
35. Burckstummer,T., Bennett,K.L., Preradovic,A., Schutze,G., Hantschel,O., Superti-Furga,G. and Bauch,A. (2006) An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat. Methods*, **3**, 1013–1019.
36. Gavin,A.C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S., Dumpelfeld,B. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
37. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
38. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.
39. Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M. *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.*, **26**, 73–79.
40. Tanford,C. (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.*, **84**, 4240–4274.
41. Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.
42. Kawashima,S., Ogata,H. and Kanehisa,M. (1999) AAindex: amino acid index database. *Nucleic Acids Res.*, **27**, 368–369.
43. Chou,K.C. and Shen,H.B. (2007) Recent progress in protein subcellular location prediction. *Anal. Biochem.*, **370**, 1–16.
44. Chou,K.C. and Cai,Y.D. (2006) Predicting protein-protein interactions from sequences in a hybridization space. *J. Proteome Res.*, **5**, 316–322.
45. Lu,C., Devos,A., Suykens,J.A., Arus,C. and Van Huffel,S. (2007) Bagging linear sparse Bayesian learning models for variable selection in cancer diagnosis. *IEEE Trans. Inf. Technol. Biomed.*, **11**, 338–347.
46. Mao,K.Z. (2004) Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Trans. Syst. Man Cybern. B Cybern.*, **34**, 629–634.
47. Molodianovitch,K., Faraggi,D. and Reiser,B. (2006) Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. *Biom. J.*, **48**, 745–757.
48. Streiner,D.L. and Cairney,J. (2007) What's under the ROC? An introduction to receiver operating characteristics curves. *Can. J. Psychiatr.*, **52**, 121–128.
49. Gasch,A.P. and Werner-Washburne,M. (2002) The genomics of yeast responses to environmental stress and starvation. *Funct. Integr. Genomics*, **2**, 181–192.
50. Rinner,O., Mueller,L.N., Hubalek,M., Muller,M., Gstaiger,M. and Aebersold,R. (2007) An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat. Biotechnol.*, **25**, 345–352.
51. Liu,X.D. and Thiele,D.J. (1996) Oxidative stress induced heat shock factor phosphorylation and HSF-dependent activation of yeast metallothionein gene transcription. *Genes Dev.*, **10**, 592–603.
52. Mohri-Shiomi,A. and Garsin,D.A. (2008) Insulin signaling and the heat shock response modulate protein homeostasis in the Caenorhabditis elegans intestine during infection. *J. Biol. Chem.*, **283**, 194–201.
53. Conlin,L.K. and Nelson,H.C. (2007) The natural osmolyte trehalose is a positive regulator of the heat-induced activity of yeast heat shock transcription factor. *Mol. Cell Biol.*, **27**, 1505–1515.
54. Hahn,J.S., Hu,Z., Thiele,D.J. and Iyer,V.R. (2004) Genome-wide analysis of the biology of stress responses through heat shock transcription factor. *Mol. Cell Biol.*, **24**, 5249–5256.
55. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.