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

KiYoung Lee1,2,3,4, Han-Yu Chuang1,5, Andreas Beyer1,6, Min-Kyung Sung7, Won-Ki Huh7, Bonghee Lee2 and Trey Ideker1,5,*

1Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA, 2Center for Genomics and Proteomics, Lee Gil Ya Cancer and Diabetes Institute, Gachon University of Medicine and Science, Incheon 406-799, Republic of Korea, 3Structural Biology Laboratory, Salk Institute for Biology Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA, 4Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Republic of Korea, 5Bioinformatics Program, University of California San Diego, La Jolla, CA 92093, USA, 6Biotechnology Center, Technische Universität, 01062 Dresden, Germany and 7School 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...
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 pre-
diction algorithms in yeast cover as many as 22 distinct
cellular localizations (5,6). Not surprisingly, approaches
which limit their predictions to smaller numbers of loca-
izations 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 interac-
tion networks in yeast, fly, worm and human (25–34)
provides one means to at least partially address these chal-
enges. To interact physically, two proteins must localize
to the same or adjacent cellular compartments, suggesting
that interaction may serve as an indicator for co-locali-
zation. 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 rela-
tively few attempts to use interacting proteins in the pre-
diction 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-immunopre-
cipitation 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 data-
bases 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 infor-
mation 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 predic-
tions 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 pro-
tein’s primary sequence, depicting its amino acid composi-
tion and chemical properties. Occurrences of known
signaling motifs in the primary protein sequence, down-
loaded 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 interac-
tions (>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. endoplas-
mic reticulum) tend to interact with proteins in different
localizations (e.g. vacuole). To reflect the possibility
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 (endoplasmic 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 proteins based on their sequences.
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 ≤ k ≤ 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:

\[ \text{HPo}(k) = \frac{1}{n-k} \sum_{j=1}^{n-k} (\text{HPo}(R_j) - \text{HPo}(R_{j+k}))^2, \]

where \( \text{HPo}(R_j) \) 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:

\[ \text{PseuAA}(k) = \frac{1}{n-k} \sum_{j=1}^{n-k} \left[ \frac{1}{3} (U + V + W) \right], \]

where \( U = (\text{HPo}(R_j) - \text{HPo}(R_{j+k}))^2, \ V = (\text{HPil}(R_j) - \text{HPil}(R_{j+k}))^2 \) and \( W = (\text{SCM}(R_j) - \text{SCM}(R_{j+k}))^2 \). 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

<table>
<thead>
<tr>
<th>Species</th>
<th>Data set</th>
<th>Proteins</th>
<th>Localizations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Localization</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>Localization-known proteins</td>
<td>3914</td>
<td>5184</td>
</tr>
<tr>
<td></td>
<td>Localization-known and having interactions</td>
<td>3206</td>
<td>4284</td>
</tr>
<tr>
<td></td>
<td>Ambiguous</td>
<td>237</td>
<td>189335</td>
</tr>
<tr>
<td></td>
<td>Localization-unknown</td>
<td>1530</td>
<td>0</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>Localization-known</td>
<td>2187</td>
<td>2398</td>
</tr>
<tr>
<td></td>
<td>Localization-known and having interactions</td>
<td>1610</td>
<td>1778</td>
</tr>
<tr>
<td></td>
<td>Localization-unknown and having interactions</td>
<td>5656</td>
<td>0</td>
</tr>
<tr>
<td>Homo sapiens (13 localizations)</td>
<td>Localization-known</td>
<td>4570</td>
<td>5251</td>
</tr>
<tr>
<td></td>
<td>Localization-known and having interactions</td>
<td>2684</td>
<td>3093</td>
</tr>
<tr>
<td></td>
<td>Localization-unknown and having interactions</td>
<td>3767</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species</th>
<th>Data set</th>
<th>Proteins</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interaction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>BioGRID</td>
<td>5184</td>
<td>70700</td>
</tr>
<tr>
<td></td>
<td>DIP</td>
<td>4931</td>
<td>17471</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>5395</td>
<td>56035</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>BioGRID</td>
<td>7545</td>
<td>25463</td>
</tr>
<tr>
<td></td>
<td>DIP</td>
<td>7038</td>
<td>20719</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>BioGRID</td>
<td>7378</td>
<td>20968</td>
</tr>
</tbody>
</table>

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, mitochondrial-nuclear periphery, nucleolus, nucleolus, peroxisome, punctate composite, spindle pole, vacuolar membrane, vacuole.

\( ^b \) 12 DM localizations are actin, cell periphery, centrosome, cytosol, ER, golgi, lysosome, mitochondrial-nuclear peroxisome, vacuole.

\( ^c \) 13 HS localizations are actin, cell cortex, centrosome, cytosol, ER, golgi, lysosome, mitochondrial-nuclear peroxisome, plasma membrane, vacuole. Further details regarding localizations and interactions of SC, DM and HS are in Supplementary Figure S1 and Tables S1–S4.
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 \((\text{PLCP})\) matrix for each protein network \((\text{BiG} \ SC, \ DIP \ SC, \ SGD \ SC, \ \text{BiG} \ DM, \ DIP \ DM \text{ andBiG} \ HS)\) 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 \((l_i \text{ and } l_j)\). \(I_{ij}\) is defined as:

\[
I_{ij} = \frac{\sum_{a \in l_i, b \in l_j, a \neq b} g(a,b) \cdot N(a) \cdot N(b)}{N(l_i) \cdot N(l_j)}
\]

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: \(N^p\) and \(L^p\). \(N^p\) of protein \(P_i\) is defined as the weighted average of the \(S\) 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 \(C^p_i\)).

\[
H^p_i = \frac{1}{\sum_{k=1}^S w_{ki}} \sum_{k=1}^S w_{ki} S_k.
\]

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 \(\Psi_{ki}\) is the number of distinct localizations covered by proteins along a path from \(P_i\) to \(P_k\), and \(\rho\) 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.

\(L^p_i\) 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:

\[
L^p_i = [L^p_{i}(1), \ldots, L^p_{i}(y), \ldots, L^p_{i}(22)]
\]

\[
L^p_i(y) = \frac{1}{\sum_{k=1}^D w_{ki}} \sum_{k=1}^D \left[ w_{ki} \max_{l \in \Gamma_k} \max_{l \in \Gamma_k} \frac{p(l_x | l_{iy})}{w_{ii}} \right]
\]

where \(l_x\) is one element of the localization set \(\Gamma_k\) of \(P_k\), and \(p(l_x | l_{iy})\) is the conditional probability of label \(l_x\) given the label \(l_{iy}\) (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 \(9S\) vectors (AA, diAA, gapAA, three kinds of chemAA, psuAA, Motif, GO), the \(54 N\) vectors [\(= 9 S\) vectors \(\times 2\) (up to second neighborhood) \(\times 3\) (the number of network databases)], and the \(6 L\) vectors [\(= 2\) (up to second neighborhood) \(\times 3\) (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 \(\mu_l\) for label \(l\) is defined as:

\[
\mu_l = \left[ \frac{1}{k} \sum_{m=1}^k n_{ml} \times \phi_{ml} \right]^{\frac{1}{n_m}},
\]

where \(n_{ml}\) is the number of \(k\)-nearest neighbors that have label \(l\) according to sub-vector \(m\). \(\phi_{ml}(\sum \phi_{ml} = 1)\) 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 \(\phi_{ml}\), 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) pretreated with concanavalin A (Sigma, St. Louis, MO, USA) to ensure cell adhesion. Microscopy was performed on a Zeiss Axiocam MRm 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: \(S\) features only, \(N\) features only, \(L\) features only, all three features together (\(S + N + L\)) 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
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 $S$, $N$ and $L$ 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 $S$ 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 $S$ and network features $N$ and $L$ 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 Supplements.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 ($5 \times 10^{-4}$ false positive rate for Noc4 and $1 \times 10^{-3}$ 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) $S$ features only, (iii) $N^i$ only, (iv) $L^i$ only and (v) all three kinds of features. (b–e) The performance of the $N^0$ features for amino acid frequencies (b), chemical amino acid properties (c), and GO terms (d) as well as performance of the feature $L^0$ (e). Performance is based on the five interaction networks BiG, DIP, SGD, Combined, and Random (different color curves). The performance of other $N^0$ 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 $S$ 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.

### Extrapolation to higher eukaryotes

Obwohl die Genauigkeit des proteinbasierten DC-kNN-Verfahrens in höheren Eukaryoten höher ist, ist es auch möglich, dass die AUC-Performance mit dem Anwachsen der Komplexität der Proteininteraktionen sinkt. In einigen Fällen sind die AUC-Performance von DC-kNN-Verfahren mit weniger Proteininteraktionen höher als die ISort- oder PSLT2-basierte Methoden.

### Comparison with previous methods

In einem früheren Artikel (1) zeigten die Autoren, dass das DC-kNN-Verfahren signifikant überlegen ist im Vergleich zu ISort und PSLT2-basierten Methoden. Die AUC-Performanz von DC-kNN-Verfahren war in einigen Fällen höher als die ISort- oder PSLT2-basierte Methoden.

### Exemplary experiments


### Exemplary experiments

Um die Genauigkeit des proteinbasierten DC-kNN-Verfahrens in höheren Eukaryoten zu bestätigen, haben wir Beispiele von Proteininteraktionen für einheitliche Proteininteraktionen 27 Beispiele und von 24 zu 75 Beispiele und von 75 zu 275 Beispiele. Die AUC-Performanz wurde mit einer Laufzeit von 27 Minuten erreicht. Darüber hinaus zeigte das DC-kNN-Verfahren eine signifikant höhere AUC-Performanz als die ISort- oder PSLT2-basierte Methoden.

### Exemplary experiments

Um die Genauigkeit des proteinbasierten DC-kNN-Verfahrens in höheren Eukaryoten zu bestätigen, haben wir Beispiele von Proteininteraktionen für einheitliche Proteininteraktionen 27 Beispiele und von 24 zu 75 Beispiele und von 75 zu 275 Beispiele. Die AUC-Performanz wurde mit einer Laufzeit von 27 Minuten erreicht. Darüber hinaus zeigte das DC-kNN-Verfahren eine signifikant höhere AUC-Performanz als die ISort- oder PSLT2-basierte Methoden.
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. 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).

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).
both fly and human. The currently available fly and human networks, containing 25,463 and 20,968 interactions among 7,545 and 7,378 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, 1,709 fly and 2,684 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 7,058 (fly) and 4,366 (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.
marks denote the performance of the proposed method when applied to all nine single protein features without feature set selection. The red ‘X’ diamond represents the performance of a conventional approach using the fraction of known localizations for network proteins (b). The blue curves using the average degree of their available protein networks (a) Worm, fly, and human were mapped onto these yeast performance the leave-one-out approach. To avoid over fitting, the selected feature 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**


