doi:10.1016/j.jmb.2005.12.067 J. Mol. Biol. (2006) 357, 339–349

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Integrated Prediction of the Helical Membrane Protein Interactome in Yeast

Yu Xia[1], Long J. Lu[1] and Mark Gerstein[1,2]*

[1]*Department of Molecular Biophysics and Biochemistry Yale University New Haven CT 06520, USA*

[2]*Department of Computer Science, Yale University New Haven, CT 06520, USA*

**Corresponding author*

At least a quarter of all genes in most genomes contain putative transmembrane (TM) helices, and helical membrane protein interactions are a major component of the overall cellular interactome. However, current experimental techniques for large-scale detection of protein–protein interactions are biased against membrane proteins. Here, we define protein–protein interaction broadly as co-complexation, and develop a weighted-voting procedure to predict interactions among yeast helical membrane proteins by optimally combining evidence based on diverse genome-wide information such as sequence, function, localization, abundance, regulation, and phenotype. We use logistic regression to simultaneously optimize the weights of all evidence sources for best discrimination based on a set of known helical membrane protein interactions. The resulting integrated classifier not only significantly outperforms classifiers based on any single genomic feature, but also does better than a benchmark Naïve Bayes classifier (using a simplifying assumption of conditional independence among features). Finally, we apply the optimized classifier genome-wide, and construct a comprehensive map of predicted helical membrane protein interactome in yeast. This can serve as a guide for prioritizing further experimental validation efforts.

Keywords: helical membrane protein; protein–protein interaction; integrated prediction; Naïve Bayes; logistic regression

Introduction

Many fundamental cellular processes involve protein–protein interactions. The mapping of the interactome, the set of all interactions among proteins encoded in a genome, is not only an important step towards systematically defining protein function, but also a first step towards understanding the mechanisms of cell behavior.[1,2] In the past few years, significant progress has been made in genome-wide identification of protein–protein interactions, especially in the budding yeast *Saccharomyces cerevisiae*.[3–6]

In addition to the experimentally derived interaction network, it is also possible to predict protein interactions by using sequence, structural, and functional genomic information. For example, two proteins are more likely to interact if they share similar phylogenetic profiles,[7] are co-expressed,[8] have homologs in another organism that are known to interact,[9–12] or if a low energy 3D structural model for the complex can be built using multimeric threading.[13] Detailed reviews of these and many other individual methods for predicting protein interactions can be found elsewhere.[14,15] In addition, interaction prediction can be further improved by integrating different features.[16–20] Various machine learning methods have been applied, ranging from the simple Naïve Bayes[17] to the more sophisticated boosting[21] and decision tree-based methods.[22,23] In addition to protein interaction, predictions for other important protein and protein pair properties can also be improved by feature integration, such as subcellular localization,[24] protein function,[25,26] and genetic interaction.[27]

The membrane protein interactome, and helical membrane protein interactome in particular, is an important part of the overall interactome. Genomic studies suggest that membrane proteins make up ~25% to ~33% of the predicted proteins in an organism,[28–31] most of which are helical proteins. Unfortunately, mapping the membrane protein

Abbreviations used: TM, transmembrane; TF, transcription factor; ROC, receiver operating characteristic.

E-mail address of the corresponding author: mark.gerstein@yale.edu

interactome is difficult, as many experimental techniques for directly assaying protein–protein interactions genome-wide are thought to be biased against membrane proteins. For instance, the yeast two-hybrid system[32] is difficult to carry out for integral membrane proteins. In this assay, interacting proteins help reassemble a functional transcription factor, which becomes in a consecutive step bound to its target promoter for the activation of the corresponding reporter gene. Thus, interaction itself must take place in the cell nucleus. However, integral membrane proteins are anchored in the membrane and cannot be transported into the nucleus. Related considerations apply for other methodologies such as the protein chip[33] and large-scale pull-down experiments.[5,6] Several experimental techniques have been proposed to address this problem.[34] Furthermore, there exist new experimental techniques that detect interactions between individual transmembrane (TM)-helices.[35,36] Nevertheless, it has so far been difficult to experimentally construct a genome-wide map of membrane protein interactions and TM-helix interactions in yeast.

Membrane protein interactome is different from soluble protein interactome in several ways. First, the biophysical environment of membrane and soluble proteins are very different. A significant part of the interactions among integral membrane proteins occur within the lipid bilayer (Figure 1). Second, there are likely fewer potential interaction partners for membrane proteins than for their soluble counterparts. This is because soluble proteins are relatively free to move within the cell and, therefore, have the ability to interact with many proteins at different times and places. In contrast, a membrane protein's mobility is largely limited by the membrane. Third, membrane proteins interact with more restricted geometry than soluble ones, which can project a wide variety of different interfaces. For example, the overwhelming majority of helical membrane protein interactions are parallel and antiparallel helix-to-helix interactions (Figure 1). Thus, it may be easier to computationally model the structures and energetics of helical membrane protein interactions compared to soluble protein interactions in general. Furthermore, mapping helical membrane protein interactome provides a starting point for understanding TM-helix interactions in a genome-wide fashion.[37]

Here, we focus on integrating genomic features to predict yeast helical membrane protein interactions, which are the majority of yeast membrane protein interactions. We define protein–protein interaction broadly as co-complexation, a definition previously used in both experimental[5,6] and computational studies.[17,19] We make use of the rich and diverse genome-wide sequence, function, localization, abundance, regulation, and phenotype data that exist in yeast for interaction prediction. We first assemble a list of 14 features that are potential predictors for interaction, among which 11 show strong correlation with helical membrane protein interaction. We then combine these pieces of evidence together to form an integrated interaction prediction using a novel logistic regression classifier, which naturally deals with the redundancy and correlation among features. To calibrate the interaction classifier, we need two data sets: a list of helical membrane protein pairs that are known to interact with high confidence (gold-standard positive set), and a list of helical membrane protein pairs that are known not to interact with high confidence (gold-standard negative set).

Several challenges are posed for membrane proteins that are distinct from soluble proteins. First, membrane proteins often have characteristic functional genomic attributes. For example, membrane proteins are expressed at a lower level compared to soluble proteins.[38] As a result, the predictive power of each genomic feature needs to be re-assessed for membrane proteins. Second, it is difficult to define a gold-standard negative set for membrane protein interactions. In the case of soluble proteins, we can use the sub-cellular localization information and define two proteins to be non-interacting if they belong to different cellular compartments.[17] In the case of membrane proteins, however, membrane localization annotation is less accurate and complete: many membrane proteins are only annotated to localize in

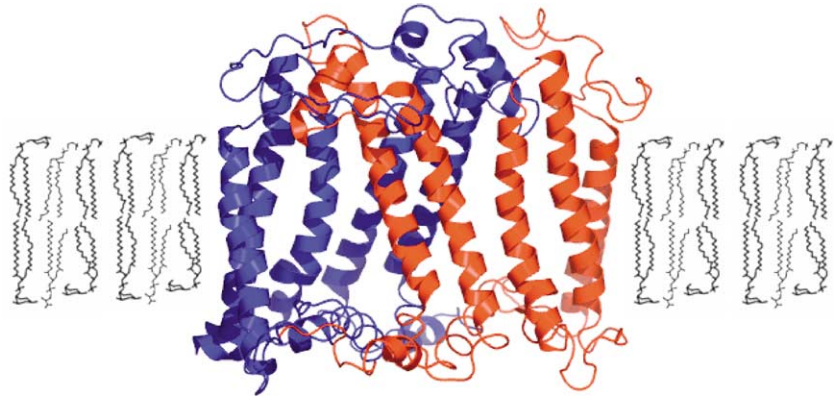

Figure 1. A 3D example of helical membrane protein–protein interaction.

"membrane". As a result, we chose to use an approximate gold-standard negative set for membrane protein interactions, and treat membrane co-localization as one of the many genomic features for integrated interaction prediction.

Results and Discussion

Identification of putative helical membrane proteins in yeast

We identified 1048 putative helical membrane proteins based on consensus predictions from two servers: TMHMM[30] and Phobius[39] (Supplemental Data, Table S2). TMHMM is widely used for predicting transmembrane topology from protein sequence information. It is particularly accurate for distinguishing membrane proteins from soluble proteins. The main problem with TMHMM is that it sometimes confuses signal peptides for TM-helices. A related prediction server, Phobius, addresses this problem by making joint predictions for TM-helices and signal peptides. To further reduce the number of false positives, we only consider the consensus predictions from the two servers. Our predictions are reasonably accurate: they contain roughly 72% of the proteins annotated with plasma membrane localization (the rest of the annotated plasma membrane proteins are presumably mostly beta-sheet proteins). As a comparison, less than 5% of the proteins annotated with cytoplasm localization show up in our predictions (Supplemental Data, Table S3).

Throughout this paper, we define protein–protein interaction as co-complexation, a definition also used elsewhere.[17,19,21] These include protein pairs with direct physical contact, as well as other protein pairs that belong to the same protein complex. Such definition for interaction has also been previously used in large-scale pull-down experiments.[5,6] Intuitively, we aim to carry out *in silico* pull-down experiments for yeast helical membrane proteins. The number of interactions in our gold-standard positive set is 304 (see Materials and Methods). The total number of possible pairwise interactions among the 1048 identified helical membrane proteins is 548,628. Estimating the actual size of the helical membrane protein interactome, however, is difficult. It has been estimated that each yeast protein physically interacts with an average of five partners,[40] and there are likely fewer potential interaction partners for membrane proteins than for their soluble counterparts due to restrictions imposed by membrane geometry. At the same time, the average number of co-complexed partners per protein should be considerably larger than that of physical interaction partners. Taken together, it is reasonable to assume that each helical membrane protein has on the order of ten co-complexed helical membrane protein partners, so the estimated number of actual helical membrane protein interactions is 5240. As a result, the percentage of all helical membrane protein pairs that interact is likely very small. This observation provides a justification for our construction of the approximate gold-standard negative set for interaction (see Materials and Methods).

Relating protein pair features to helical membrane protein interaction

We collected a list of 14 helical membrane protein pair features that potentially correlate with interaction based on diverse genome-wide information such as sequence, function, localization, abundance, regulation, and phenotype. Numerical features were first converted into categorical ones by binning the data into several distinct categories, such as from high to low. For each piece of evidence, i.e. a categorical feature taking on a particular value, we computed how frequently it occurs for helical membrane protein pairs known to interact (the gold-standard positive set), as well as how frequently it occurs for all helical membrane protein pairs. We then computed the fold enrichment, defined as the ratio of these two frequencies. A fold enrichment much larger than 1 indicates that the evidence is a good predictor for interaction. Likewise, a fold enrichment less than 1 indicates that the evidence is anti-predictive for interaction.

Function and localization

For each helical membrane protein pair, we calculated the similarity between the two functional categories as defined in the Gene Ontology (GO) database.[41] The procedure for calculating functional similarity is the same as described.[17] Protein pairs with high GO functional similarity are significantly enriched in the gold-standard positive set, and the enrichment gradually falls off with decreasing functional similarity (Figure 2(a)). Since interaction can be used to infer protein function,[2] in this calculation we only used the subset of functional assignments that is independent of interaction information to avoid circularity. A similar trend is observed when we calculate similarity between functional categories as defined in the MIPS database[42] (Figure 2(b)). Evidently, protein pairs with similar function are more likely to interact than expected by chance.[43]

Protein pairs that have been shown experimentally to localize to the same type of membrane (for example, plasma *versus* ER membrane) are five times more likely to interact than expected by chance (Figure 2(c)), based on the localization annotations from the *Saccharomyces* Genome Database (SGD).[44] This makes membrane co-localization a useful predictor for interaction. Still, only less than half of the protein pairs in the gold-standard positive set (126 out of 304) show membrane co-localization evidence. This is largely due to the fact that localization annotation for membrane proteins is more difficult and less complete compared to soluble proteins. In the case of soluble

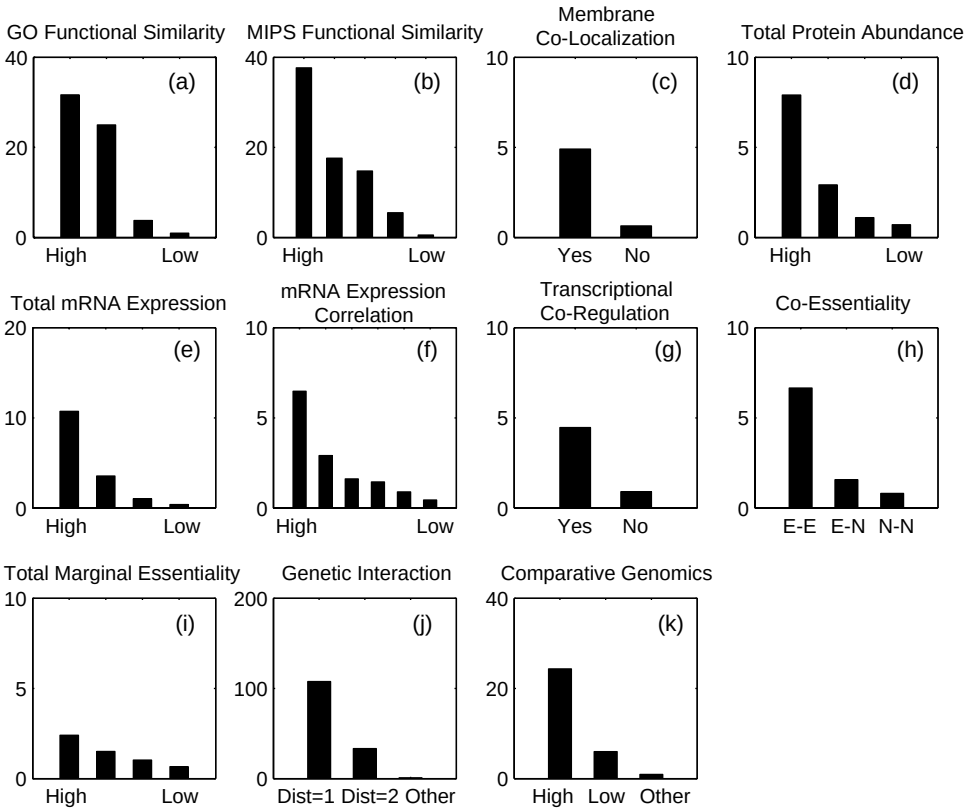

Figure 2. Fold enrichments of interacting proteins for features used in our integrated prediction. For each categorical value of a given feature, we compute the fold enrichment, i.e. the frequency it occurs for helical membrane protein pairs known to interact, divided by the frequency it occurs for all helical membrane protein pairs. The *X*-axis represents the possible categorical values associated with each feature, and *Y*-axis represents the fold enrichment associated with each categorical value.

proteins, sub-cellular localization information is reliable enough for construction of accurate gold-standard negative set.[17] In the case of membrane proteins, however, membrane co-localization can only be used as one of the many predictors.

Abundance

For each pair of helical membrane proteins, we calculated total protein abundance as the sum of the log abundance levels for the two proteins. Protein pairs with high total abundance level are more likely to interact than expected by chance (Figure 2(d)). This is because abundant membrane proteins tend to have more interaction partners, a trend also observed for yeast proteins in general.[45] A reasonable explanation for this is that abundant proteins are more likely to encounter other proteins by chance. A second factor could be that abundant proteins are more likely to be studied experimentally. The same trend is also observed for total mRNA expression (Figure 2(e)), which is not surprising since mRNA expression level correlates with protein abundance level.[46]

For each pair of helical membrane proteins, we also calculated relative protein abundance as the absolute difference between the log abundance levels for the two proteins. Since interacting proteins should be present in stoichiometrically equal amounts, we reasoned that protein pairs with similar abundance levels should be more likely to interact.[8] Surprisingly, we did not observe this trend (Figure 3(a)), possibly due to the large noise level associated with the protein abundance data. A similar analysis with mRNA expression levels, however, shows that protein pairs with similar mRNA expression levels are indeed more likely to interact (Figure 3(b)), but the trend is rather weak compared to other features. This can be explained by the fact that

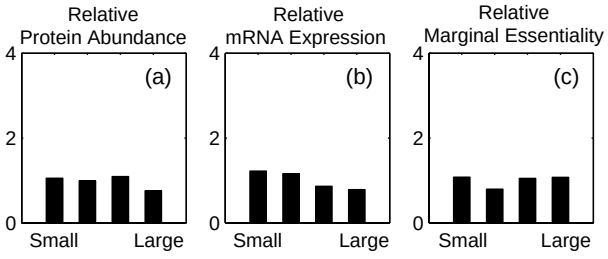

Figure 3. Fold enrichments of interacting proteins for features not used in our integrated prediction. These are computed using helical membrane protein pairs with complete feature information.

membrane proteins tend to express at a low level.[38] As a result, relative mRNA expression is very noisy and less useful as a predictor for protein interaction.

We further calculated the correlation of mRNA expression profiles over time-course experiments for helical membrane protein pairs. Interacting proteins tend to show correlated mRNA expression profiles (Figure 2(f)). However, mRNA expression correlation here is not as strong a predictor as in the case for soluble proteins,[17,47] again possibly due to the noise associated with low expression levels of membrane proteins.

Regulation

Much is known about gene regulation in yeast, especially at the level of transcriptional regulation, where transcription factors (TF) regulate the expression of target genes by binding DNA at specific promoter regions. A yeast transcriptional regulatory network has been constructed[48] by integrating known TF-target relationships[49] with results from large-scale experiments such as ChIP-chip.[50,51] From this we derived a list of protein pairs that are regulated by the same TF. Helical membrane protein pairs with transcriptional co-regulation evidence are nearly five times more likely to interact than expected by chance (Figure 2(g)). Thus, transcriptional co-regulation is a good predictor for interaction.

Phenotype

A yeast protein can be classified as essential or non-essential, based on the viability of the cell when the gene is knocked out. Based on essentiality annotations from the SGD database,[44] we observed that two helical membrane proteins are more likely to interact when they are both essential[17] (Figure 2(h)). This is because essential proteins tend to have more interactors.[52]

Marginal essentiality of a non-essential protein is a continuous measure for the degree of importance of this protein to the cell.[53] For each pair of helical membrane proteins, we calculated total marginal essentiality as the sum of the log marginal essentiality for the two proteins. Protein pairs with high total marginal essentiality are more likely to interact than expected by chance (Figure 2(i)). This is because membrane proteins with high marginal essentiality tend to have more interaction partners, a trend also observed for yeast proteins in general.[53] Furthermore, we calculated relative marginal essentiality as the absolute difference between the log marginal essentiality for the two proteins. We reasoned that interacting proteins should tend to have similar marginal essentiality, since deleting either protein would render the whole protein complex dysfunctional, thus reducing the fitness of the cell by a similar degree. Unfortunately, our calculations do not support this hypothesis (Figure 3(c)).

There exists an additional type of phenotype information, called genetic interaction, for pairs of genes. Two genes are said to interact genetically if a mutation in one gene either suppresses or enhances the phenotype of a mutation in the other gene. A prime example of genetic interaction is synthetic lethality associated with two non-essential genes, which individually are not essential, but when jointly knocked out are lethal.[54] We have constructed a yeast genetic interaction network by combining information from MIPS[42] and GRID databases.[55] We found that helical membrane protein pairs that genetically interact are much more likely to be members of the same complex than expected by chance (Figure 2(j)). The effect is transitive: two helical membrane proteins that each genetically interacts with a common third gene are themselves more likely to be members of the same complex than expected by chance as well (Figure 2(j)). This result is consistent with the observation that most ($\sim$80%) genetic interactions occur between genes from parallel pathways that are functionally redundant, rather than genes belonging to the same complex.[56] Even though proteins that genetically interact belong to the same complex only 20% of the time, this probability is still much larger than the probability that two proteins selected at random belong to the same complex (<0.01). Thus, genetic interaction is still a strong predictor for co-complexation.

Comparative genomics

There exist powerful methods to infer functional relatedness of genes based on different comparative genomic evidence. For example, functionally related proteins tend to co-occur in different genomes,[7] to be close together along the chromosome,[57] and to be fused together in another genome.[58,59] A database, Prolinks, has been constructed to collect and assess protein functional linkages inferred from these comparative genomic features.[60] In Prolinks, different comparative genomic features are combined statistically into one score. We then use this score as our single comparative genomic feature. The existence of such comparative genomic evidence is a strong predictor for yeast helical membrane protein interactions (Figure 2(k)).

Integrated prediction using logistic regression

Out of the 14 genomic features tested, 11 are good predictors for protein interaction (Figure 2): the fold enrichment significantly deviates from 1 for at least one categorical value of each feature (p-value <0.02, using χ^2 test). We further integrated these 11 features to predict helical membrane protein interaction using a weighted voting scheme. Each piece of evidence, i.e. a genomic feature taking on a particular categorical value, is assigned a different weight. For each helical membrane protein pair, a total weight is computed by summing up all

weights associated with different pieces of evidence. The protein pair is predicted to interact if the total weight exceeds a cut-off, and non-interact if otherwise. Each genomic feature can take two to six categorical values, and the total number of independent parameters for our classifier is 30. All parameters are simultaneously optimized using logistic regression for best prediction performance on the training set, and the resulting logistic regression classifier is evaluated using the test set.

Our gold-standard positive set contains 304 protein pairs, and our approximate gold-standard negative set contains 548,324 protein pairs. We vary the class size ratio by assigning unit weight to the gold-standard negative set, and a weight of k to the gold-standard positive set. In practice, we do this by duplicating the gold-standard positive set k times before combining it with the negative set. Intuitively, k is the cost of a false negative error relative to a false positive error. For each k, we estimate the true positive rate (fraction of gold-standard positives predicted as interaction) and the false positive rate (fraction of gold-standard negatives predicted as interaction) of the classifier using sevenfold cross-validation. By varying k and repeating the above cross-validation, we can plot true positive rate as a function of false positive rate, i.e. a receiver operating characteristic (ROC) curve for the classifier.

We plotted the ROC curve for logistic regression classifiers based on each of the features, as well as based on integration of all features (Figure 4). At the level of low false positive rates (<0.03), the integrated classifier significantly outperforms any of the individual classifiers. Among the individual classifiers, the ranks of prediction power are different at different levels of false positive rate. When the false positive rate is low (~ 0.002), the best individual classifiers are based on GO functional similarity, genetic interaction, MIPS functional similarity, and comparative genomic evidence. When the false positive rate is higher (~ 0.03), the best individual classifiers are based on MIPS functional similarity, GO functional similarity, genetic interaction, and membrane co-localization.

We selected $k=100$ for constructing our final integrated classifier. The corresponding cross-validated true positive rate is 0.385, and the false positive rate is 0.00734 (Figure 4). Assuming that the estimated number of helical membrane protein interactions is 5240, the number of true positives will be roughly 2017. The total number of positive predictions at $k=100$ is 4145, so the ratio of true positives to false positives for our predictions will be on the order of 1:1. This provides the rationale for our particular choice of k. The best individual classifier that can achieve similar true positive rate is MIPS functional similarity, but with a much higher false positive rate (~ 0.03, or more than four times higher than the integrated classifier) (Figure 4).

Our integrated logistic regression classifier slightly outperforms feature integration using Naïve Bayes. When we fix k to be 100, logistic regression reduces the total cost of misclassification by 20% (Table 1). When we fix the cross-validated true positive rate to be 0.38, the false positive rate for the logistic regression classifier is 28% smaller than that of the corresponding Naïve Bayes classifier (Table 1).

Our classifier is robust against perturbations on the gold standard positive dataset. Even when the positive dataset is reduced by as much as 30% upon deletion of the three largest complexes, the resulting optimized weights are still reasonably similar to the weights trained on the full positive dataset, with a correlation coefficient of 0.983. Genome-wide predictions are similar as well: 87% of the interactions predicted by the resulting classifier are also predicted by the original classifier trained on the full gold-standard dataset.

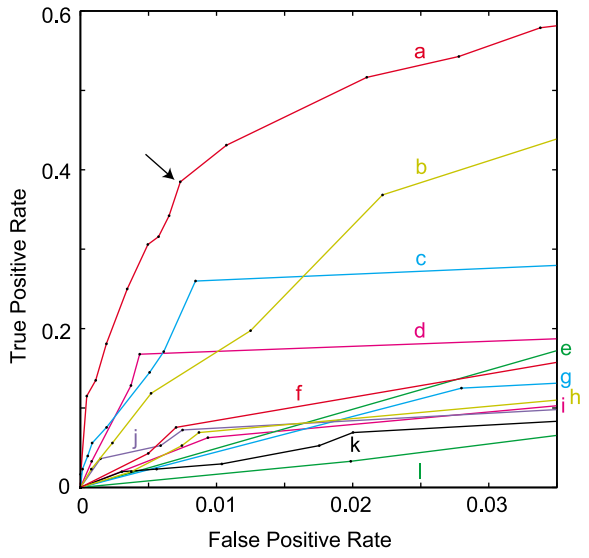

Figure 4. ROC curve for logistic regression classifiers based on all features combined, as well as each individual feature alone. (a) All features combined; (b) MIPS functional similarity; (c) GO functional similarity; (d) genetic interaction; (e) membrane co-localization; (f) total mRNA expression; (g) transcriptional co-regulation; (h) total protein abundance; (i) co-essentiality; (j) comparative genomic evidence; (k) mRNA expression correlation; (l) total marginal essentiality. The arrow represents a cost of $k=100$ used in our final integrated predictions.

A map of predicted helical membrane protein interactome in yeast

Our final integrated predictions consist of 4145 interactions among helical membrane proteins (Supplemental Data, Table S4), among which 120 are in the gold-standard positive set. The fraction of gold-standard positive set correctly predicted by our integrated classifier is 0.395. This is more than 50 times higher than that expected by chance, 0.00756 (Figure 5). This result is similar to the above cross-validated results, suggesting that overlearning is not an issue here.

Table 1. Comparison of prediction performances between logistic regression and Naïve Bayes, evaluated with sevenfold cross-validation

	Cost (k)[a]	TP	P	TPR	FP	N	FPR	Total cost of misclassification
Logistic regression	100	117	304	0.38	4022	548,324	0.0073	22,722
Naïve Bayes	100	152	304	0.50	13,520	548,324	0.025	28,720
Naïve Bayes	20	117	304	0.38	5630	548,324	0.0103	9370

True positive rate (TP) is defined as the number of true positives (P) divided by the total number of positives (TPR); false positive rate (FPR) is defined as the number of false positives (FP) divided by the total number of negatives (N).
[a] The cost of a false negative is k, and the cost of a false positive is 1.

In Figure 6 we show a high-confidence subset (based on logistic regression score, defined by the right-hand side of equation (1)) of the predicted as well as known helical membrane protein interactome. From the map it is apparent that our predictions can be grouped into four categories: (a) new interactions among known members of the known complexes; (b) new members for the known complexes; (c) new complexes; and (d) various other new interactions. The high-level organization of the map can be further visualized by labeling each of the large complexes with the consensus GO functional annotation of all members of the complex (Figure 6).

To further test the validity of our predictions, we compared our predictions against the core set of the Database of Interaction Proteins (DIP) database, a manually compiled set of protein interactions based on literature curation.[61] We found 52 new helical membrane protein interactions in the DIP-core set that are not present in our gold-standard positive set. Our list of 4145 predicted interactions correctly identifies 19 of these 52 new interactions, or a true positive rate of 36.5%. This is consistent with our cross-validated true positive rate estimate of 38.5%, further corroborating the quality of the predictions. These 19 new predictions that are further validated by literature curation are shown in Table 2.

Comparing predictions with large-scale experiments by Miller *et al.*

Recently, Miller *et al*. published a large-scale experimental screen to identify physical interactions between yeast integral membrane proteins by using a modified split-ubiquitin technique.[62] They identified 1949 putative non-self interactions among 705 proteins annotated as integral membrane; 556 of these 705 proteins were included in our list of putative helical membrane proteins, and the number of identified physical interactions among them is 1607. As a comparison, we predicted 3020 co-complexations among these proteins, and 79 of them agree with the experimental results. The overlap, although small, is significant (fold enrichment $=2.5$, p-value $<10^{-12}$), despite the fact that the goals of the two studies are different (physical interactions, *versus* co-complexations). This small overlap, similar to the small overlap found between independent genome-wide two-hybrid screens,[3,4] is likely due to a combination of false negative and false positive rates.[19,63] Furthermore, 155 interactions were identified to be of highest confidence by Miller *et al*., or used in their training set for such identification. Among these high confident interactions, we correctly predicted 35. The overlap here is even more significant (fold enrichment $=11.5$, p-value $<10^{-13}$). Note, however, that the latter comparison is compounded by the fact that some of the criteria Miller *et al*. used to ascribe confidence levels are similar to some of the genomic features we used for integrated prediction. Nevertheless, the significant overlap found between our predictions and their experimental results further demonstrates the validity of our integrated approach.

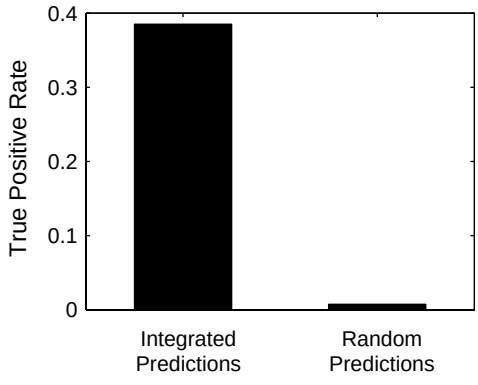

Figure 5. The true positive rate of our integrated predictions compared to random predictions.

Conclusion

By integrating a diverse set of genomic features based on function, localization, abundance, regulation, phenotype, and comparative genomics information, we constructed a high confidence map of predicted helical membrane protein interactome in yeast. Moreover, the presence or absence of correlation with interaction for each one of the 14 features tested, as well as the strength of the correlation, provides further biological insights. Logistic regression is used to optimally combine different pieces of evidence for interaction prediction. We predicted 4145 helical membrane protein interactions, or 0.756% of all possible interactions. As a comparison, our predictions cover 38.5% of the known interactions in the gold-standard positive set, using cross-validation. Moreover, our predic-

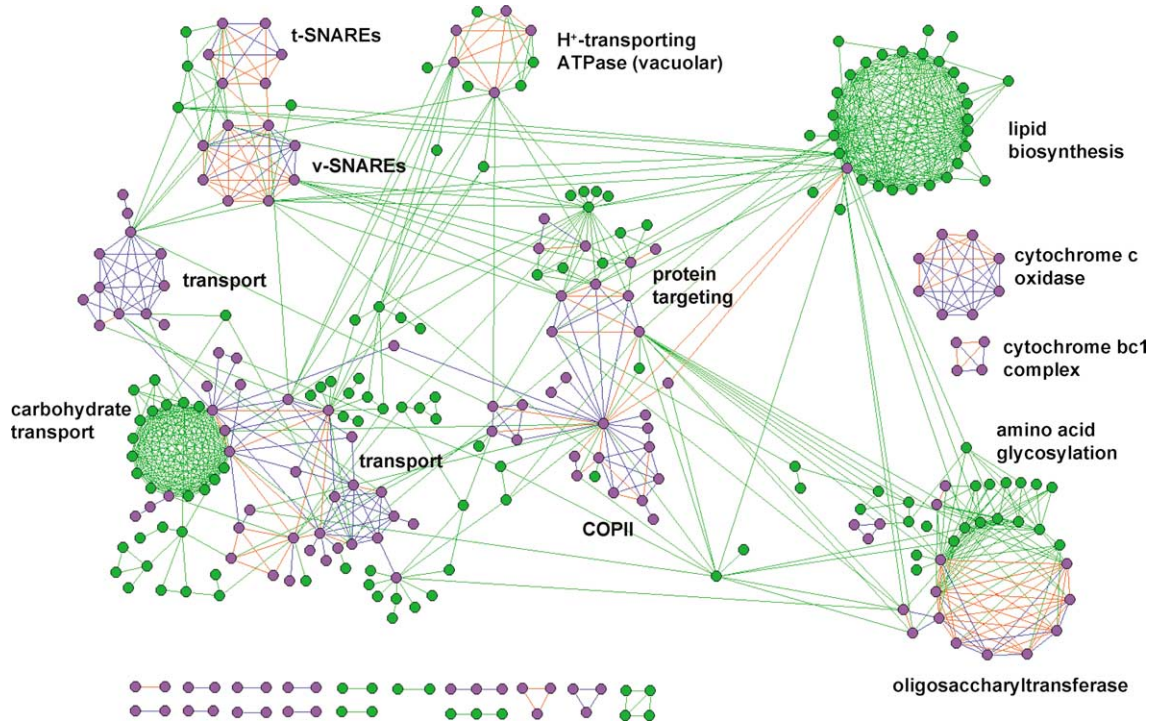

Figure 6. A map of known and a subset of predicted interactions among helical membrane proteins. Nodes represent helical membrane proteins, and edges represent interactions among them. Red edges represent known interactions that are also predicted to interact, blue edges represent other known interactions, and green edges represent ~700 top interaction predictions (ranked by descending logistic regression score) out of a total of 4145. Purple nodes represent helical membrane proteins that show up in the known interactions, and green nodes represent new helical membrane proteins.

tions correctly identify 36.5% of the 52 literature-curated new interactions not present in the gold-standard positive set, which is consistent with the cross-validated estimate, and further corroborate the quality of the predictions. This map can be used to prioritize further experimental validation efforts, and represent the first step towards understanding TM helix–helix interactions in a genome-wide fashion.

Table 2. A list of 19 new predicted interactions that are validated by literature

Protein 1	Protein 2	Logistic regression score
KAR2	SEC63	4.65
SEC28	SEC22	3.69
TLG2	SNC2	3.36
SED5	SEC22	3.07
SFT1	SED5	2.75
PEP1	VTI1	2.39
BET1	SED5	2.37
VPH2	VPH1	2.17
PEX10	PEX12	1.89
SEC28	BOS1	1.73
BET1	SEC28	1.61
MNN10	MNN11	1.25
ANP1	MNN9	1.21
SED5	VTI1	1.19
GOS1	SED5	0.82
AKR1	STE3	0.81
VTI1	PEP12	0.36
NYV1	VAM3	0.36
PEX22	PEX12	0.19

There are several important directions for future work. First, it is desirable to identify and remove non-specific interactions which are often biologically not interesting.[64] Second, it is useful to automate the identification of putative protein complexes based on predicted interactions. Third, it is important to study in detail the effects of sampling biases present in genomic datasets. For example, part of the observed correlation between total abundance and protein interaction is possibly due to experimental sampling bias. The magnitude of this sampling bias, and sampling biases in genomic datasets in general, warrants further investigation as more experimental data on membrane proteins become available.

Materials and Methods

Identifying putative helical membrane proteins in yeast

Yeast helical membrane proteins are identified based on consensus predictions from two servers, TMHMM[30] and Phobius,[39] with default parameters. We only consider those proteins with at least one predicted TM-helix based on both servers.

Genomic features for predicting helical membrane protein interactions

We collected a list of 14 protein pair properties that potentially correlates with helical membrane protein

interaction (Supplemental Data, Table S1). Missing values were replaced by its row average for continuous variables, or the most populated category for categorical variables (Supplemental Data, Table S1). We then convert all protein pair properties to categorical features by following two guiding rules: (1) there should be a large difference in fold enrichment between adjacent categories; (2) the number of points in each category should be reasonably large to ensure statistical significance. Finally, we apply various classification methods such as Naïve Bayes and logistic regression to integrate these categorical features for interaction prediction.

Known helical membrane protein interactions for training and testing

We collected a list of known yeast complexes and their protein components from the MIPS database.[42] We then compiled a gold-standard positive set for interaction by identifying all helical membrane protein pairs that belong to the same MIPS complex. In addition, we constructed an approximate gold-standard negative set for interaction by identifying all helical membrane protein pairs that do not belong to the gold-standard positive set. These gold-standard sets are used for training and testing interaction classifiers.

Naïve Bayes and logistic regression classifiers

For a helical membrane protein pair, we want to predict the class label y (1 if interacting, and 0 otherwise) by integrating features F. There are m categorical features: $F_1,\ldots,F_m$, where each feature F_j ($j=1,\ldots,m$) can take on r_j different values: $f_{j1},f_{j2},\ldots,f_{jr_j}$. The training set, $\{(F^{(i)}, y^{(i)}); i=1,\ldots,n\}$, contains n samples. Both Naïve Bayes and logistic regression classifiers can be expressed as the following weighted voting scheme:[65]

$$\log\frac{p(y=1|F)}{p(y=0|F)} = w_0 + \sum_{j=1}^{m}\sum_{k=1}^{r_j} w_{jk} I(F_j = f_{jk}) \qquad (1)$$

where I is the indicator function: $I(X)$ is equal to 1 when statement X is true, and 0 otherwise. $w_{11},\ldots$ are weights associated with each piece of evidence. $p(y=1|F)$ is the probability that the protein pair is interacting given the features. The protein pair is predicted to interact if and only if $p(y=1|F)$ is larger than 0.5.

In Naïve Bayes classifier, the weights are estimated from the training set in the following way:

$$w_0 = \log\frac{\sum_{i=1}^{n} I(y^{(i)}=1)}{\sum_{i=1}^{n} I(y^{(i)}=0)}$$

$$w_{jk} = \log\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r_j} I(y^{(i)}=1 \wedge F_j^{(i)}=f_{jk})\right) \Big/ \sum_{i=1}^{n} I(y^{(i)}=1)}{\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r_j} I(y^{(i)}=0 \wedge F_j^{(i)}=f_{jk})\right) \Big/ \sum_{i=1}^{n} I(y^{(i)}=0)} \qquad (2)$$

Here w_0 is the prior log-odds for interaction, and w_{jk} is the log likelihood ratio for feature F_j taking on the value f_{jk}. A crucial assumption made by the Naïve Bayes classifier is that features are conditionally independent given the class label y. This limitation can be overcome by using logistic regression. Here, all weights are chosen to optimize the following log-likelihood function for the training set, i.e. the log-probability of observing the data given the weights:

$$\log L(w_0, w_{11}, \ldots, w_{mr_m}) = \sum_{i=1}^{n}\left(I(y^{(i)}=1)\log p(y^{(i)}=1|F^{(i)}) + I(y^{(i)}=0)\log p(y^{(i)}=0|F^{(i)})\right) \qquad (3)$$

The right-hand side of the above equation measures the agreement between the actual class labels y and the predictions $p(y|F)$. Additional regularization terms can be added to penalize model complexity, thus reducing the problem of over-fitting. Here, we did not use any regularization terms, as it was shown in Results and Discussion that the over-fitting issue is negligible in this study.

Acknowledgements

We thank Matteo Pellegrini for providing the comparative genomic scores, and Donald Engelman, Haiyuan Yu, and Alberto Paccanaro for helpful discussions. Y.X. is a Fellow of the Jane Coffin Childs Memorial Fund for Medical Research. This work is supported by a grant from NIH/NIGMS (P50 GM62413-01).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.12.067

References

1. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, **402**, C47–C52.
2. Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, **405**, 823–826.
3. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R. *et al*. (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
4. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
5. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A. *et al*. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
6. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L. *et al*. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

7. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
8. Jansen, R., Greenbaum, D. & Gerstein, M. (2002). Relating whole-genome expression data with protein–protein interactions. *Genome Res.* **12**, 37–46.
9. Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A. *et al*. (2000). Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, **287**, 116–122.
10. Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J. *et al*. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or "interologs". *Genome Res.* **11**, 2120–2126.
11. Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J. D. *et al*. (2004). Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res.* **14**, 1107–1118.
12. McDermott, J., Bumgarner, R. & Samudrala, R. (2005). Functional annotation from predicted protein interaction networks. *Bioinformatics*, **21**, 3217–3226.
13. Lu, L., Arakaki, A. K., Lu, H. & Skolnick, J. (2003). Multimeric threading-based prediction of protein–protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res.* **13**, 1146–1154.
14. Valencia, A. & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* **12**, 368–373.
15. Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D. *et al*. (2004). Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem.* **73**, 1051–1087.
16. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
17. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S. *et al*. (2003). A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.
18. Ge, H., Walhout, A. J. & Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* **19**, 551–560.
19. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
20. Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I. & Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14**, 292–299.
21. Lu, L. J., Xia, Y., Paccanaro, A., Yu, H. & Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.* **15**, 945–953.
22. Lin, N., Wu, B., Jansen, R., Gerstein, M. & Zhao, H. (2004). Information assessment on predicting protein–protein interactions. *BMC Bioinformatics*, **5**, 154.
23. Zhang, L. V., Wong, S. L., King, O. D. & Roth, F. P. (2004). Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, **5**, 38.
24. Drawid, A. & Gerstein, M. (2000). A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.* **301**, 1059–1075.
25. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.
26. Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
27. Wong, S. L., Zhang, L. V., Tong, A. H., Li, Z., Goldberg, D. S., King, O. D. *et al*. (2004). Combining biological networks to predict genetic interactions. *Proc. Natl Acad. Sci. USA*, **101**, 15682–15687.
28. Jones, D. T. (1998). Do transmembrane protein superfolds exist? *FEBS Letters*, **423**, 281–2285.
29. Gerstein, M. (1998). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518–534.
30. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.
31. Wallin, E. & von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7**, 1029–1038.
32. Fields, S. & Song, O. (1989). A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.
33. Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P. *et al*. (2001). Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.
34. Stagljar, I. & Fields, S. (2002). Analysis of membrane protein interactions using yeast-based technologies. *Trends Biochem. Sci.* **27**, 559–563.
35. Russ, W. P. & Engelman, D. M. (1999). TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc. Natl Acad. Sci. USA*, **96**, 863–868.
36. Schneider, D. & Engelman, D. M. (2003). GALLEX, a measurement of heterologous association of transmembrane helices in a biological membrane. *J. Biol. Chem.* **278**, 3105–3111.
37. Lehnert, U., Xia, Y., Royce, T. E., Goh, C. S., Liu, Y., Senes, A. *et al*. (2004). Computational analysis of membrane proteins: genomic occurrence, structure prediction and helix interactions. *Quart. Rev. Biophys.* **37**, 121–146.
38. Drawid, A., Jansen, R. & Gerstein, M. (2000). Genome-wide analysis relating expression level with protein subcellular localization. *Trends Genet.* **16**, 426–430.
39. Kall, L., Krogh, A. & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036.
40. Kumar, A. & Snyder, M. (2002). Protein complexes take the bait. *Nature*, **415**, 123–124.
41. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M. *et al*. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*. **25**, 25–29.
42. Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G. *et al*. (2004). MIPS: analysis and annotation of proteins from whole genomes. *Nucl. Acids Res.* **32**, D41–D44.

43. Schwikowski, B., Uetz, P. & Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnol.* **18**, 1257–1261.
44. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T. *et al.* (1998). SGD: Saccharomyces Genome Database. *Nucl. Acids Res.* **26**, 73–79.
45. Yu, H., Zhu, X., Greenbaum, D., Karro, J. & Gerstein, M. (2004). TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucl. Acids Res.* **32**, 328–337.
46. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **4**, 117.
47. Ge, H., Liu, Z., Church, G. M. & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*.. *Nature Genet.* **29**, 482–486.
48. Yu, H., Luscombe, N. M., Qian, J. & Gerstein, M. (2003). Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.* **19**, 422–427.
49. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I. *et al.* (2001). The TRANSFAC system on gene expression regulation. *Nucl. Acids Res.* **29**, 281–283.
50. Horak, C. E. & Snyder, M. (2002). ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* **350**, 469–483.
51. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K. *et al.* (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
52. Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
53. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. & Gerstein, M. (2004). Genomic analysis of essentiality within protein networks. *Trends Genet.* **20**, 227–231.
54. Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H. *et al.* (2004). Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
55. Breitkreutz, B. J., Stark, C. & Tyers, M. (2003). The GRID: the General Repository for Interaction Datasets. *Genome Biol.* **4**, R23.
56. Kelley, R. & Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnol.* **23**, 561–566.
57. Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66–73.
58. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
59. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
60. Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O. & Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* **5**, R35.
61. Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M. & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* **30**, 303–305.
62. Miller, J. P., Lo, R. S., Ben-Hur, A., Desmarais, C., Stagljar, I., Noble, W. S. & Fields, S. (2005). Large-scale identification of yeast integral membrane protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 12123–12128.
63. Han, J. D., Dupuy, D., Bertin, N., Cusick, M. E. & Vidal, M. (2005). Effect of sampling on topology predictions of protein–protein interaction networks. *Nature Biotechnol.* **23**, 839–844.
64. Rives, A. W. & Galitski, T. (2003). Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA*, **100**, 1128–1133.
65. Ng, A. Y. & Jordan, M. I. (2002). On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. *Advan. Neural Informat. Process. Syst.* **14**, 605–610.

Edited by G. von Heijne

(Received 23 August 2005; received in revised form 19 December 2005; accepted 20 December 2005)
Available online 5 January 2006