

# Chapter 5

## Computational Reconstruction of Protein–Protein Interaction Networks: Algorithms and Issues

Eric Franzosa, Bolan Linghu, and Yu Xia

### Abstract

Accurate mapping of protein–protein interaction networks in model organisms is a crucial first step toward subsequent quantitative study of the organization and evolution of biological systems. Data quality of experimental interactome maps can be assessed and improved by integrating multiple sources of evidence using machine learning methods. Here we describe the commonly used algorithms for predicting protein–protein interaction by genome data integration, and discuss several important yet often overlooked issues in computational reconstruction of protein–protein interaction networks.

**Key words:** Protein–protein interaction, machine learning, protein network, data integration, Naïve Bayes, logistic regression.

---

### 1. Introduction

In the past few years, significant progress has been made in genome-wide identification of protein–protein interactions, especially in model organisms such as *Saccharomyces cerevisiae* (1–6) and *Caenorhabditis elegans* (7), and also recently in human (8). With the availability of these experimental interactome maps, it is now possible for the first time to quantitatively study the organization and evolution of biological systems at the level of protein–protein interaction networks, and develop theoretical models that account for the observed statistical trends (9–11). This line of research depends crucially on the quality of the reconstructed protein–protein interaction networks, as measured by accuracy, completeness, and possible bias. The dependence of

49 derived organizational and evolutionary hypotheses on data  
50 quality is not always obvious; an excellent recent example is the  
51 observation that power-law topology of the interactome map  
52 depends on its completeness (12). Such studies underlie the  
53 importance of rigorous assessment and subsequent improvement  
54 of the quality of interactome maps by a combination of experi-  
55 mental and computational methods.

56 Here we focus on computational reconstruction of protein-  
57 protein interaction networks by integrating multiple sources of  
58 evidence (13–15). Such sources of evidence can be the interactome  
59 maps produced by different labs, other binary maps such as genetic  
60 interaction maps, or other genomic features suggestive of protein-  
61 protein interaction. The basic premise is simple: if multiple reliable  
62 sources of evidence all suggest that two proteins interact, then  
63 the probability that these two proteins interact is high. To make  
64 this intuition precise, we need to quantify the reliability of each  
65 source of evidence, taking into account data quality (as mentioned  
66 above), as well as redundancy and similarity among different  
67 sources of evidence. Machine learning methods provide a straight-  
68 forward solution to this issue. In machine learning, we specify the  
69 simplest possible model that, we believe, captures the dominant  
70 structure in the data. In our case, the model relates multiple sources  
71 of evidence to whether or not two proteins interact. We then fit the  
72 model to a training set (selected from a small gold-standard data  
73 set), adjusting the model parameters so as to maximize the agree-  
74 ment between the model and the data. The performance of the  
75 learned model on unseen data can be evaluated using a separate  
76 testing set, again selected from the gold-standard data set. Finally,  
77 we apply the model genome-wide to generate predictions. Here, the  
78 complexity of the data is captured by the choice of the model. Linear  
79 models and their variants have been widely used, because: (1) these  
80 models often capture the dominant structure in the data: noise,  
81 incompleteness, redundancy, and correlation; (2) many nonlinear  
82 structures in the data can become linear after appropriate data  
83 transformation; (3) these models are simple: efficient optimization  
84 methods exist to fit such models to the data, and over-fitting  
85 problem is usually minimal.

86 In **Section 2**, we describe the choice of gold-standard  
87 positive and negative interaction data sets, genomic features  
88 for predicting protein-protein interaction, machine learning  
89 methods for predicting protein-protein interaction, and ways  
90 to transform nonlinear structure in continuous and graph-  
91 based data into linear structure. In **Section 3**, we describe  
92 additional important issues in reconstructing the interactome:  
93 the choice of the size of positive and negative examples, dealing  
94 with features whose predictive power is difficult to quantify, and  
95 the effect of size and bias in the experimental interactome  
96 maps.

---

## 2. Methods

### 2.1. Gold-Standard Positive and Negative Interaction Data Sets

There are two different ways of defining protein–protein interactions. The first definition is more specific: two proteins interact when they share a physical binding interface. This is also called binary interaction, and can be detected with yeast two-hybrid experiments. The second definition is broader: two proteins interact when they are subunits of the same complex. This is also called co-complex memberships, and can be detected with pull-down experiments. Here we focus on the prediction of co-complex memberships in yeast, but the same framework also applies to the prediction of binary interactions.

The gold-standard positive data set, a set of protein pairs that are known to interact, is usually constructed from known protein complexes annotated in MIPS (14, 16). Gold-standard positive data sets constructed in this way, although highly useful, are not perfect: they are biased toward important, well-behaved proteins and protein complexes associated with pronounced phenotypes or diseases. Unless explicitly modeled, standard machine learning methods are not able to correct such biases.

The gold-standard negative data set, a set of protein pairs that are known not to interact, is much harder to construct (17). This is because negative results are typically neither published nor stored in any database. One way to solve this problem is to assume that proteins that localize in different cellular compartments do not interact (14). An alternative approach is to construct an approximate gold-standard negative data set as all protein pairs that do not belong to the gold-standard positive data set and to use co-localization information as one of the many features (18, 19). There are several advantages of this approach. First, co-localization information is treated in the same way as all other features. Second, a protein is estimated to interact on average with at most 10–20 proteins out of ~6,000 proteins in yeast. As a result, the vast majority (>99.5%) of the approximate gold-standard negative data sets are in fact true negatives. Third, gold-standard data sets do not need to be 100% accurate. A small amount of noise can be tolerated as long as the gold-standard data sets contain strong enough signals to guide the parameterization of the classifier.

### 2.2. Compiling a List of Genomic Features

Many protein pair features correlate with interaction. Such genomic features can be collected for each of the ~18 million yeast protein pairs. Here we list a representative subset of these features: (1) experimental physical and genetic interaction maps from different labs; (2) the mapping of interologs (20), i.e., conserved interactions between two proteins or domains, from another organism to yeast; (3) features based on comparative genomic evidence, such as

145 similarity of phylogenetic profiles (21) and gene neighborhood (22),  
 146 co-evolution (23), belonging to the same gene cluster (24), and the  
 147 existence of domain fusion events in another organism (25, 26);  
 148 (4) pair protein features that are derived from single protein features  
 149 such as function (14), localization (14), mRNA expression (27, 28),  
 150 abundance (15, 18), regulation (29), and phenotype (14, 15, 30);  
 151 (5) features based on 3D structural analysis, such as multimeric  
 152 threading (31).

153 Missing data is a serious problem and needs to be treated  
 154 differently depending on the missing data mechanism. However,  
 155 in many cases, if feature X contains missing data, simply creating a  
 156 new binary variable “X-is-missing” will work well in practice.  
 157

### 158 2.3. Naïve Bayes

159 Consider the following binary classification problem. Given a  
 160 training set of independently and identically distributed samples  
 161  $T = \{(x^{(i)}, y^{(i)}); i = 1, \dots, m\}$  of feature and binary class variables  
 162 from an unknown distribution  $D$ , estimate a classifier  $f(x)$  that  
 163 predicts the binary class variable  $y \in \{0, 1\}$  (whether or not the  
 164 protein pair interacts) from the features  $x$ . Without loss of general-  
 165 ity, suppose that we have two binary feature variables  $x = (x_1, x_2)$ ,  
 166 where  $x_1, x_2 \in \{0, 1\}$ . (We will discuss continuous feature vari-  
 167 ables later.) The goal here is to come up with a classifier  $f(x)$  that  
 168 minimizes the expected prediction error  $\mathbf{E}_{(x,y) \in D} \mathbf{1}\{y \neq f(x)\}$ ,  
 169 where  $\mathbf{1}\{X\}$  is equal to 1 when statement  $X$  is true, and 0  
 170 otherwise.

170 According to statistical decision theory, the optimal classifier  
 171  $f(x)$  as defined above can be written in the following way:  
 172

$$173 f(x) = \begin{cases} 1, & \text{when } \frac{p(y = 1|x)}{p(y = 0|x)} > 1 \\ 0, & \text{when } \frac{p(y = 1|x)}{p(y = 0|x)} < 1 \end{cases} \quad [1]$$

174 Now we make the Naïve Bayes assumption that features are  
 175 conditionally independent:  $p(x_1, x_2|y) = p(x_1|y)p(x_2|y)$ . Under  
 176 this assumption,  
 177

$$178 \frac{p(y = 1|x)}{p(y = 0|x)} = \frac{p(y = 1)p(x_1|y = 1)p(x_2|y = 1)}{p(y = 0)p(x_1|y = 0)p(x_2|y = 0)} \quad [2]$$

179 The five independent parameters in the above equation can be  
 180 easily estimated from the training set.  
 181

### 182 2.4. Logistic 183 Regression

184 Equation [2] is equivalent to the following equation:  
 185

$$186 \ln \frac{p(y = 1|x)}{p(y = 0|x)} = w_0 + w_1 x_1 + w_2 x_2 \quad [3]$$

This equation relates linearly the posterior log-odds of an interaction given the evidence with the presence or absence of each piece of evidence.

Naïve Bayes classifiers assume that features are conditionally independent. Such assumptions are often incorrect. In logistic regression, the linear model in Eq. [3] is fit to the data, without the extra assumption of conditional independence (32). The weights  $w_0$ ,  $w_1$ ,  $w_2$  are obtained by maximizing the following likelihood function:  $L_C(w_0, w_1, w_2) = \prod_{i=1}^m p(y^{(i)} | x^{(i)})$ .

## 2.5. SVM and Boosting

The above maximum likelihood (ML) estimate of the weights  $w_0$ ,  $w_1$ ,  $w_2$  is equivalent to minimizing the following function:  $\sum_{i=1}^m \phi_{LR}(\alpha^{(i)})$ , where  $\alpha = (2y - 1)(w_0 + w_1x_1 + w_2x_2)$  is called the margin, and the loss function  $\phi_{LR}(\alpha) = \ln(1 + e^{-\alpha})$  is a convex surrogate for 0–1 loss function  $\phi_{0-1}(\alpha) = I\{\alpha < 0\}$ . Let us now relax the requirement for ML estimation and consider other ways to estimate the weights. The different estimation methods generally aim at minimizing the empirical classification error  $\frac{1}{m} \sum_{i=1}^m \phi_{0-1}(\alpha^{(i)})$ , with 0–1 loss function surrogated by a convex loss function so as to make efficient global optimization possible. In the case of logistic regression, this convex surrogate loss function is  $\phi_{LR}(\alpha) = \ln(1 + e^{-\alpha})$ . But we are free to choose other appropriate convex surrogate loss functions; in particular, support vector machine (SVM) and AdaBoost use different loss functions (33):  $\phi_{SVM}(\alpha) = \max(1 - \alpha, 0)$ , and  $\phi_{AdaBoost}(\alpha) = e^{-\alpha}$ .

## 2.6. Regularization

In some cases even the linear model in Eq. [3] is too complex and causes over-fitting. For example, we usually have a small number of annotated protein–protein interactions, and a large number of genomic features most of which are irrelevant. In this case, we want to make the linear model even simpler by imposing additional constraints that only a small subset of all features has non-zero weights. Such regularization can be done in several different ways. For example, a feature selection step can be performed prior to the model-fitting step. Alternatively, a regularization term can be added to the model-fitting step to penalize complex models, as done in SVM. Finally, AdaBoost uses greedy optimization coupled with early stopping to control the complexity of the model.

## 2.7. Nonlinear Continuous and Graph-Based Features

We previously focused on binary features. A categorical feature with  $n$  categories can be easily decomposed into  $n$  binary features. What about continuous features, such as expression correlation? In general, the posterior log-odds of interaction may depend on these continuous features in a nonlinear way. However, we can convert a nonlinear continuous feature into several linear binary features by binning the data. For example, we can bin the expression correlation data into three binary features: expression-correlation-high, expression-correlation-medium, and

241 expression-correlation-low. We can then fit a linear model to the  
 242 transformed feature space, assigning three different weights to  
 243 protein pairs with high, medium, and low expression correlation.  
 244 Notice that even though the model is linear in the transformed  
 245 categorical feature space, it is actually nonlinear in the original  
 246 continuous feature space. This simple binning procedure allows  
 247 us to extend the linear model to many nonlinear cases. There are  
 248 also other more complex procedures, such as the kernel-based  
 249 methods (34).

250 Some genomic features are based on graphs such as interac-  
 251 tome maps and genetic interaction maps. Several different metrics  
 252 have been proposed to measure the distance between a pair of  
 253 proteins in these graphs, such as diffusion distance (35), linear  
 254 kernel (36, 37), and congruence score (38). These metrics can  
 255 then be combined with the rest of the genomic features to predict  
 256 protein–protein interaction.

## 258 **2.8. Decision Tree** 259 **and Random Forests**

260 Sometimes the dependence of protein–protein interaction on  
 261 genomic features is so complex that the linear relationship in  
 262 Eq. [3] is no longer valid. The most common machine learning  
 263 method that deals with such irreducible nonlinearity is decision  
 264 tree and its variants, such as random forests. These methods have  
 265 been successfully applied to the prediction of protein–protein  
 266 interaction (39).

---

## 268 **3. Notes**



- 270
- 271 1. *How large should the gold-standard negative set be?* We usually  
 272 fix the size of the gold-standard positive set to be a constant,  
 273 as determined by the MIPS complex catalog, but we are  
 274 free to vary the size of the gold-standard negative set. As the  
 275 gold-standard negative set gets bigger, the classifier applies a  
 276 stricter cut-off, and as a result predicts a smaller number of  
 277 positive interactions. For all classifiers except Naïve Bayes,  
 278 individual evidence weights will also change.

279 What, then, is the right choice for the negative example  
 280 size? Shall we pick the same number of negative examples  
 281 as positive examples? Or to the other extreme, shall we pick  
 282 a lot more negative examples than positive examples to  
 283 approximately preserve the ratio of positive to negative  
 284 interactions in the entire proteome? The right choice depends  
 285 on the prediction task at hand. If our task is not to correctly  
 286 predict *all* interactions but rather to come up with a list of  
 287 predicted interactions that are *accurate*, then none of the  
 288

## Predicting Protein–Protein Interactions

289 above two methods are appropriate. Rather, we should  
290 choose the appropriate negative example size so that there  
291 are roughly equal numbers of true and false positives in the  
292 predicted interactions (17).

- 293 2. *Features whose predictive powers are difficult to quantify.* It is  
294 sometimes difficult to assess in a quantitative way the predic-  
295 tive powers of certain features, such as functional similarity  
296 based on Gene Ontology annotations. Because a subset of the  
297 Gene Ontology annotations are themselves derived from  
298 interaction information, part of the observed correlation  
299 between functional similarity and interaction is spurious.  
300 One way to solve this problem is to exclude the subset of  
301 the Gene Ontology annotations that are derived from inter-  
302 action information (18).

303 Let us now consider a hypothetical situation where we  
304 do not know which subset of the Gene Ontology annota-  
305 tions are derived from interaction information. It then  
306 becomes impossible to quantify the predictive power of  
307 the functional similarity feature. However, this does not  
308 mean that this feature is not useful at all for making new  
309 predictions. Even without quantitative assessment, we can  
310 infer the usefulness of this feature based on biological  
311 common sense: interacting proteins should tend to share  
312 common biological functions. We argue that the best way  
313 to deal with this situation is to exclude the functional  
314 similarity feature from training-testing so as to obtain a  
315 conservative estimate of the prediction performance, but  
316 then to include the functional similarity feature in the  
317 integrated classifier for making final genome-wide  
318 predictions.

- 319 3. *Effect of size and bias in experimental interactome maps.* It  
320 is important to keep in mind that the statistical machine  
321 learning approach outlined here can only be applied  
322 straightforwardly to integrate large-scale, unbiased interac-  
323 tome mapping experiments, where the overlap with gold-  
324 standard data sets provides an accurate measurement of data  
325 quality. However, a significant fraction (34%) of the physical  
326 and genetic interactions contained in BioGRID (40) is from  
327 small-scale experiments, each mapping 100 or less interac-  
328 tions. It is difficult to assess individual small-scale data sets,  
329 but we can assess different methods by pooling together  
330 all data sets carried out using the same method. As shown  
331 in Table 5.1, the predictive power for co-complex member-  
332 ships decreases from affinity capture to two-hybrid to  
333 genetic interaction, as expected. At the same time, co-com-  
334 plexed proteins are significantly enriched for almost all  
335 methods.  
336

Franzosa, Linghu, and Xia

**Table 5.1**

**The most popular methods for mapping physical and genetic interactions, compiled from BioGRID (40). Methods are sorted by decreasing number of interactions deposited in BioGRID, and only methods with more than 1,500 interactions are shown. For each method, we compute the fold enrichment, i.e., the fraction of co-complexed protein pairs that are detected using this method, divided by the fraction of all protein pairs that are detected using this method. A fold enrichment larger than 1 indicates that the method is predictive for co-complex memberships**

Method	Number of interactions	Fold enrichment
Affinity capture – MS	18,747	166.7
Two-hybrid	9,642	75.4
Synthetic lethality	9,019	47.9
Synthetic growth defect	5,002	18.8
Affinity capture – western	3,523	354.8
Epistatic mini-array profile	3,416	7.5
Dosage rescue	2,442	138.4
Synthetic rescue	1,605	72.7
Phenotypic enhancement	1,425	80.8
Reconstituted complex	1,327	311.7

Many large-scale physical and genetic interaction mapping experiments are biased: these experiments are concerned with a specific subset of genes that share a common biological function or disease phenotype. Here, the use of a generic gold-standard positive data set is questionable, as it will tend to underestimate the data quality. For example, as shown in **Table 5.2**, there are three data sets with apparently unusually low prediction power for co-complex membership. However, close examination reveals that they are all biased maps that are concerned with a subset of the interactome. These sub-networks usually involve a specific function that is not previously well characterized and therefore underrepresented in the gold-standard positives, such as the proteins involved in DNA integrity and secretion, and membrane proteins. As a result of this bias, the quality of these data sets is significantly underestimated by standard machine learning methods. New methods are needed to accurately assess the quality of such biased interactome maps.

**Table 5.2**

**Large-scale physical and genetic interaction data sets, compiled from BioGRID. Data sets are sorted by decreasing number of interactions deposited in BioGRID, and only data sets with more than 1,000 interactions are shown. For each data set, we again compute the fold enrichment. A fold enrichment larger than 1 indicates that the data set is predictive for co-complex memberships, if we assume no biases in these data sets**

Data set	Method	Number of linteractions	Fold enrichment
Krogan et al., 2006 (6)	Affinity capture – MS	7,076	275.8
Gavin et al., 2006 (5)	Affinity capture – MS	6,531	287.2
Pan et al., 2006 (41)	Synthetic growth defect	4,533	0.3
Ito et al., 2001 (2)	Two-hybrid	3,959	43.6
Ho et al., 2002 (3)	Affinity capture – MS	3,596	82.2
Schuldiner et al., 2005 (42)	Epistatic mini-array profile	3,416	7.5
Tong et al., 2004 (43)	Synthetic lethality	3,411	10.7
Gavin et al., 2002 (4)	Affinity capture – MS	3,210	324.8
Miller et al., 2005 (44)	Two-hybrid	1,941	9.4

---

## Acknowledgments

Y.X. thanks Mark Gerstein for advice and support.

## References

1. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000, 403(6770):623–7.
2. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 2001, 98(8):4569–74.
3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figey D, Tyers M. Systematic identification

## Franzosa, Linghu, and Xia

- of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 2002, 415(6868):180–3.
4. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, 415(6868):141–7.
  5. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006, 440(7084):631–6.
  6. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandhi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O’Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006, 440(7084):637–43.
  7. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhautte J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M. A map of the interactome network of the metazoan *C. elegans*. *Science* 2004, 303(5657):540–3.
  8. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhautte J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005, 437(7062):1173–8.
  9. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature* 2000, 407(6804):651–4.
  10. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. *Science* 2002, 296(5568):750–2.
  11. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: Simple building blocks of complex networks. *Science* 2002, 298(5594):824–7.
  12. Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* 2005, 23(7):839–44.
  13. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. *Nature* 1999, 402(6757):83–6.
  14. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003, 302(5644):449–53.
  15. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res* 2005, 15(7):945–53.
  16. Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res* 1999, 27(1):44–8.
  17. Jansen R, Gerstein M. Analyzing protein function on a genomic scale: The importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol* 2004, 7(5):535–45.
  18. Xia Y, Lu LJ, Gerstein M. Integrated prediction of the helical membrane protein

## Predicting Protein-Protein Interactions

- interactome in yeast. *J Mol Biol* 2006, 357(1):339–49.
19. Ben-Hur A, Noble WS. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics* 2006, 7(Suppl 1):S2.
  20. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. *Genome Res* 2004, 14(6):1107–18.
  21. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999, 96(8):4285–8.
  22. Tamames J, Casari G, Ouzounis C, Valencia A. Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol* 1997, 44(1):66–73.
  23. Goh CS, Cohen FE. Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol* 2002, 324(1):177–92.
  24. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. Prolinks: A database of protein functional linkages derived from coevolution. *Genome Biol* 2004, 5(5):R35.
  25. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science* 1999, 285(5428):751–3.
  26. Enright AJ, Iliopoulos I, Kyripides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 1999, 402(6757):86–90.
  27. Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet* 2001, 29(4):482–6.
  28. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* 2002, 12(1):37–46.
  29. Yu H, Luscombe NM, Qian J, Gerstein M. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet* 2003, 19(8):422–7.
  30. Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M. Genomic analysis of essentiality within protein networks. *Trends Genet* 2004, 20(6):227–31.
  31. Lu L, Arakaki AK, Lu H, Skolnick J. Multimeric threading-based prediction of protein-protein interactions on a genomic scale: Application to the *Saccharomyces cerevisiae* proteome. *Genome Res* 2003, 13(6A):1146–54.
  32. Ng AY, Jordan MI. On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. *Adv Neural Inform Process Syst* 2002, 2(14):841–8.
  33. Zhang T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann Statist* 2004, 32(1):56–85.
  34. Ben-Hur A, Noble WS. Kernel methods for predicting protein-protein interactions. *Bioinformatics* 2005, 21(Suppl 1):i38–46.
  35. Kondor RI, Lafferty JD. Diffusion kernels on graphs and other discrete input spaces. In: *Proc 19th International Conf on Machine Learning*. Morgan Kaufmann Publishers Inc., 2002, pp. 315–22.
  36. Rives AW, Galitski T. Modular organization of cellular networks. *Proc Natl Acad Sci USA* 2003, 100(3):1128–33.
  37. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. *Bioinformatics* 2004, 20(16):2626–35.
  38. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS. Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol* 2005, 1:2005.0026.
  39. Lin N, Wu B, Jansen R, Gerstein M, Zhao H. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics* 2004, 5:154.
  40. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res* 2006, 34(Database issue):D535–9.
  41. Pan X, Ye P, Yuan DS, Wang X, Bader JS, Boeke JD. A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* 2006, 124(5):1069–81.
  42. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 2005, 123(3):507–19.
  43. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB,

Franzosa, Linghu, and Xia

529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576

Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast

genetic interaction network. *Science* 2004, 303(5659):808–13.

44. Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS, Fields S. Large-scale identification of yeast integral membrane protein interactions. *Proc Natl Acad Sci USA* 2005, 102(34):12123–8.

UNCORRECTED PROOF