BIOINFORMATICS

Vol. 19 no. 6 2003, pages 756–763
DOI: 10.1093/bioinformatics/btg070

Construction of reliable protein–protein interaction networks with a new interaction generality measure

Rintaro Saito[†], Harukazu Suzuki[*] and Yoshihide Hayashizaki

Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, and Genome Science Laboratory, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Received on September 9, 2002; revised on December 3, 2002; accepted on December 6, 2002

ABSTRACT

Motivation: Recent screening techniques have made large amounts of protein–protein interaction data available, from which biologically important information such as the function of uncharacterized proteins, the existence of novel protein complexes, and novel signal-transduction pathways can be discovered. However, experimental data on protein interactions contain many false positives, making these discoveries difficult. Therefore computational methods of assessing the reliability of each candidate protein–protein interaction are urgently needed.

Results: We developed a new ‘interaction generality’ measure (IG2) to assess the reliability of protein–protein interactions using only the topological properties of their interaction-network structure. Using yeast protein–protein interaction data, we showed that reliable protein–protein interactions had significantly lower IG2 values than less-reliable interactions, suggesting that IG2 values can be used to evaluate and filter interaction data to enable the construction of reliable protein–protein interaction networks.

Availability: The protein–protein interaction data used in this study along with the associated IG2 values are available at http://genome.gsc.riken.go.jp.

Contact: rgscerg@gsc.riken.go.jp

INTRODUCTION

As whole-genome and complete cDNA sequences became available for numerous organisms (Adams *et al.*, 2000; Goffeau *et al.*, 1996; Kawai *et al.*, 2001; Lander *et al.*, 2001; The *C. elegans* Sequencing Consortium, 1998; Venter *et al.*, 2001), the focus of many research efforts is shifting rapidly from genomics to proteomics. One of the most important approaches in proteomics is the large-scale analysis of protein–protein interactions, because most proteins function within complexes (Oliver, 2000; Pawson and Nash, 2000). High-throughput genome-wide screening for protein–protein interactions has been carried out in yeast, *Caenorhabditis elegans*, and higher organisms such as the mouse (Ito *et al.*, 2001; Suzuki *et al.*, 2001; Uetz *et al.*, 2000; Walhout *et al.*, 2000). Several successful computational analyses of interaction data have also been completed (Fellenberg *et al.*, 2000; Schwikowski *et al.*, 2000).

However, the publicly available protein–protein interaction data, especially those obtained from two-hybrid systems, include many false-positive interactions (Legrain *et al.*, 2001). Von Mering *et al.* (2002) estimate that approximately half the interactions obtained from high-throughput data may be false positives. These false positives may unnecessarily link unrelated proteins, resulting in huge apparent interaction clusters (Ito *et al.*, 2001), which complicate elucidation of the biological importance of these interactions. Therefore, a method to assess the reliability of each candidate protein–protein interaction is necessary. Earlier, we developed a simple computational method, which yielded an ‘interaction generality’ measure (IG1) that could be used to assess the reliability of experimentally identified interactions from just a list of interaction data (Saito *et al.*, 2002). The development of IG1 was based on the idea that interacting proteins that appear to have many other interacting partners which have no further interactions are likely to be false positives. However, our IG1 method was a simple method for evaluating the reliability of interactions that did not consider the topological properties of the protein interaction network beyond the target pair of proteins. Here we define a new interaction generality (IG2) measure that overcomes this problem and show that it can assess the reliability of putative protein–protein interactions with higher accuracy.

[*] To whom correspondence should be addressed.

[†] Present address: Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, Tsuruoka, Yamagata 997-0035, Japan.

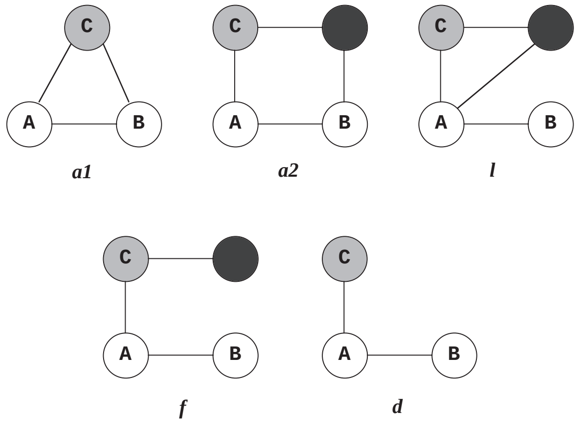


Fig. 1. Classification of a protein C that interacts with a target interacting protein pair A–B, according to the topological properties of its interaction network.

MATERIALS AND METHODS

Preparation of protein–protein interaction data

The publicly available protein interaction data of Ito *et al.* (2001); Uetz *et al.* (2000), and MIPS (Mewes *et al.*, 2000) were obtained from http://genome.c.kanazawa-u.ac.jp/Y2H/ (754 heterodimers), http://www.genome.ad.jp/brite/ (905 heterodimers), and http://www.mips.biochem.mpg.de/proj/yeast/ (2474 heterodimers), respectively. We assessed only heterodimers and considered the interactions of protein A (bait)–protein B (prey) and of protein B (bait)–protein A (prey) to represent a single interaction (these represent bidirectional interactions in a two-hybrid experiment). Combining these three data sets and removing redundancy from them yielded 3066 heterodimers, including 673 reproducible interactions. Interactions that were confirmed by coimmunoprecipitation assays and bidirectional interactions that were obtained from two-hybrid assays are considered to be reproducible.

The new interaction generality measure

The new interaction generality measure incorporates the topological properties of interactions around the target interacting pairs. The IG2 for the target interacting pair A–B is defined by the following procedure. Protein C, which interacts directly with the target interacting pair A–B, can be classified into one of five groups ($a1$, $a2$, l, f and d) according to the topological properties of its interaction network (Fig. 1). When C interacts with both A and B, it is classified as $a1$ (**a**lternative pathway from protein A to B through **1** protein). When C interacts with A but not B, and C also interacts with another protein that interacts with B, it is classified as $a2$ (**a**lternative pathway from protein A to B through **2** protein). When C is not classified as $a2$, interacts with A but not B, and interacts with at least one protein that interacts with A, it is classified as l (**l**ooping interaction). If C does not meet these three conditions and interacts with another protein, it is classified as f (**f**urther interaction). If C does not interact with any proteins except for either A or B, it is classified as d (**d**ead-end interaction).

Then, the numbers of proteins in the database that belong to each class are counted as $n = (Na1, Na2, Nl, Nf, Nd)$. We counted n's ($n_1, n_2, \ldots, n_p$, where p is the number of interactions in the given interaction network) for all p interactions. From the set of n's, we constructed this matrix:

$$N = \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_{p-1} \\ n_p \end{pmatrix} = \begin{pmatrix} Na1_1 & Na2_1 & Nl_1 & Nf_1 & Nd_1 \\ Na1_2 & Na2_2 & Nl_2 & Nf_2 & Nd_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Na1_{p-1} & Na2_{p-1} & Nl_{p-1} & Nf_{p-1} & Nd_{p-1} \\ Na1_p & Na2_p & Nl_p & Nf_p & Nd_p \end{pmatrix}$$

We defined the IG2 value for each interaction by applying principal-component analysis (Weller and Romney, 1990) to N in the following way. First, the averages of each column were subtracted from N, producing

$$N_c = N - \overline{N} = N - \begin{pmatrix} \overline{Na1} & \overline{Na2} & \overline{Nl} & \overline{Nf} & \overline{Nd} \\ \overline{Na1} & \overline{Na2} & \overline{Nl} & \overline{Nf} & \overline{Nd} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \overline{Na1} & \overline{Na2} & \overline{Nl} & \overline{Nf} & \overline{Nd} \end{pmatrix}$$

where $\overline{N}$ denotes the average of N and $N_c^{\mathrm{T}} N_c$ represents correlations between variables $(Na1, Na2, Nl, Nf, Nd)$. Note that $(1/p)N_c^T N_c$ represents the covariance matrix of these variables. Then we determined the matrix P that satisfied the following equations by singular-value decomposition (Weller and Romney, 1990):

$$N_c^{\mathrm{T}} N_c = PDP^{-1}, \quad P = (\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \boldsymbol{p}_4, \boldsymbol{p}_5),$$

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 \\ 0 & 0 & 0 & \lambda_4 & 0 \\ 0 & 0 & 0 & 0 & \lambda_5 \end{pmatrix}$$

where $\boldsymbol{p}_i$ and λ_i is each eigenvector and its corresponding eigenvalue of $N_c^{\mathrm{T}} N_c$ and they satisfy the following equations:

$$N_c^{\mathrm{T}} N_c \boldsymbol{p}_i = \lambda_i \boldsymbol{p}_i \ (i = 1, 2, 3, 4, 5)$$
$$\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5$$

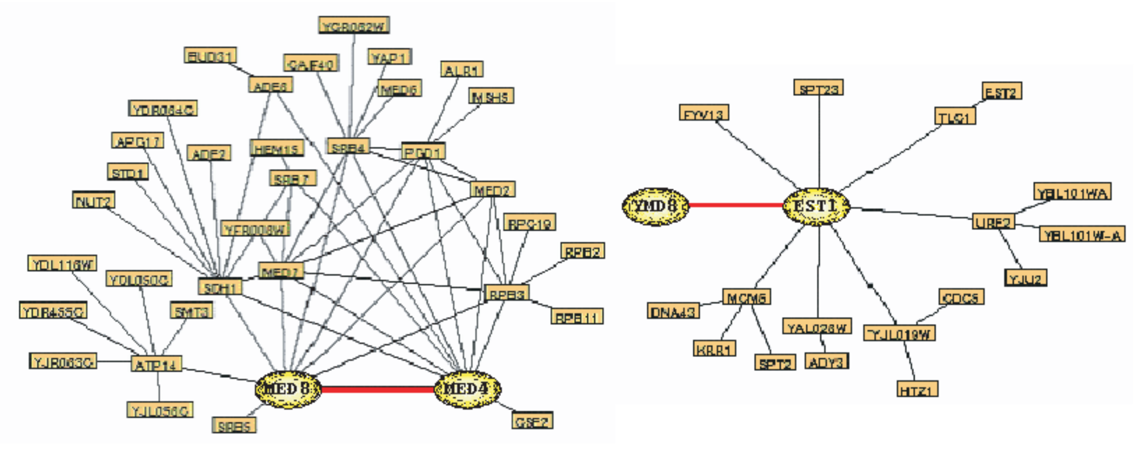


Fig. 2. Interaction networks involving the interacting pairs of yeast proteins MED8–MED4 and YMD8–EST1. Nodes and lines denote proteins and their interactions, respectively. Only proteins within two interaction steps of the target interacting pairs are shown. The MED8–MED4 interaction was experimentally verified by coimmunoprecipitation (Myers *et al.*, 1998), and the proteins were confirmed as components of a protein complex (Lorch *et al.*, 2000; Malik and Roeder, 2000), whereas the YMD8–EST1 interaction was indicated only in a high-throughput two-hybrid assay (Uetz *et al.*, 2000).

Singular-value decomposition considers correlations between the variables ($Na1$, $Na2$, Nl, Nf, Nd), and it summarizes the matrix N_c as vectors $\boldsymbol{p}_i$ ($i = 1, 2, 3, 4, 5$), which are orthogonal to each other. The $\lambda_i (i = 1, 2, 3, 4, 5)$ values indicate how well the matrix is summarized by each corresponding $\boldsymbol{p}_i$. As λ_1 has the greatest value among five λ's, $\boldsymbol{p}_1$ is the best vector that summarizes the matrix N_c. IG2 vectors for each interaction are defined as follows:

$$\text{IG2 vectors } = N_c P = (N - \overline{N})P = NP - K \quad (K = \overline{N}P)$$

IG2 values for each interaction are defined as

$$\begin{aligned} \text{IG2 values } &= N_c \boldsymbol{p}_1 = (N - \overline{N})\boldsymbol{p}_1 = N\boldsymbol{p}_1 - k_1 \\ &= Na1\boldsymbol{p}_{11} + Na2\boldsymbol{p}_{21} + Nl\boldsymbol{p}_{31} + Nf\boldsymbol{p}_{41} \\ &\quad + Nd\boldsymbol{p}_{51} - k_1 \\ &(k_1 = \overline{N}\boldsymbol{p}_1, \boldsymbol{p}_1 = (\boldsymbol{p}_{11}, \boldsymbol{p}_{21}, \boldsymbol{p}_{31}, \boldsymbol{p}_{41}, \boldsymbol{p}_{51})^{\mathrm{T}}) \end{aligned}$$

Thus, the topology of the protein–protein interaction structure is summarized by a single number IG2. In other words, characteristics of the topology are mapped to an IG2 value $\in R^1$.

The idea behind the mechanics of IG2 calculation is that interactions involving proteins that have many interacting partners are likely to be false positives, but highly interconnected sets of interactions or interactions forming a closed loop (such as the set of interaction pairs A–B, B–C, C–D, D–A) are likely to be true positives (Walhout *et al.*, 2000; Saito *et al.*, 2002). To distinguish the false and true interactions, proteins that interact with the target interaction pair were classified into five classes ($a1$, $a2$, l, f, and d) as mentioned above. Classes $a1$, $a2$, and l correspond to interactions forming a closed loop, and f and d do not. In particular, $a1$ is used for three proteins A, B, and C that interact with each other; if interactions having a large $a1$ value appear frequently in an interaction network, the proteins participating in the network are highly interconnected.

Implementation

Principal-component analysis was performed with R (a language and environment for statistical computing and graphics, http://www.r-project.org/). All other analyses were done by means of Perl scripts we developed.

RESULTS

Assessment of new interaction generality

The original IG1 measure was based on the idea that interactions observed in a complicated interaction network are likely to be true positives. The IG1 value was simply defined as the number of proteins that interact with only one of the target interacting pair (Saito *et al.*, 2002). Interactions with low IG1 values were more likely to be reproducible in independent assays. However, the topological properties of the protein interaction network beyond the target interacting pair were not considered in the IG1. For example, IG1 values for both MED8–MED4 and YMD8–EST1 interactions were three, even though MED8–MED4 seems to be involved in a more complicated interaction network and is experimentally more certain than the YMD8–EST1 interaction (Fig. 2). Actually, both the MED8 and MED4 proteins are known

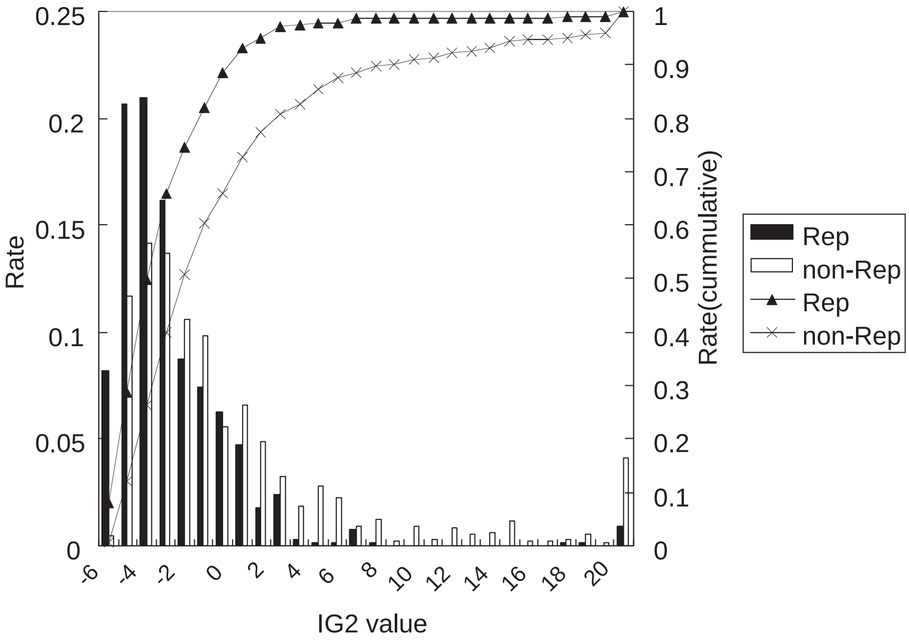


Fig. 3. Distributions of IG2 values for reproducible and non-reproducible interactions. The histogram shows the frequencies of interactions falling in the specified ranges of IG2 values. The lines show the cumulative proportion of interactions. Rep and non-Rep indicate reproducible and non-reproducible interactions, as defined in Materials and Methods.

to be transcription regulation mediators, whereas YMD8 and EST1 have different functions (chromatin chromosome structure/DNA synthesis and small molecule transport, respectively). To overcome this inaccuracy, we developed IG2, in which principal-component analysis of the topological properties is incorporated into the evaluation of the reliability of the interaction. First, we classified proteins that directly interact with the target interacting pair A–B into one of five groups ($a1$, $a2$, l, f and d) according to the topological properties of its interaction network (Fig. 1 and see Materials and Methods). Then, to convert the numbers of proteins that belong to each class ($Na1$, $Na2$, Nl, Nf, Nd) into a single IG2 value, each of the five numbers is weighted and summed. Principal-component analysis determines the weight for each number, based on the correlations between these numbers, so that a single summed value (IG2) can represent the five numbers ($Na1$, $Na2$, Nl, Nf, Nd). By applying the computational framework described in Materials and Methods to our interaction data set, $\boldsymbol{p}_1 = (-0.057, 0.0963, 0.179, 0.920, 0.331)^{\mathrm{T}}$ and $k_1 = 5.603$ were obtained. The IG2 value for the MED8–MED4 interaction (-4.17) is now very different from that for YMD8–EST1 (-0.34).

IG2 values ranging from 52.98 to –6.35 were obtained for all the interactions we collected. To investigate whether the IG2 value may be useful for assessing the validity of candidate protein–protein interaction pairs, distributions of IG2 values for reproducible and non-reproducible interactions were calculated. We expected that most reproducible interactions should be true positives, whereas the non-reproducible interactions should contain many false positives. As shown in Figure 3, the IG2 values for reproducible interactions are significantly lower than those of non-reproducible ones, suggesting that the IG2 value can be used to select reliable interactions (average IG2 values for reproducible and non-reproducible interactions are -2.904 and 0.817 respectively; $P < 1.37 \times 10^{-41}$).

Next, we investigated the mean IG2 value for the relatively reliable protein interaction data sets. Deane *et al.* (2002) found 3003 interactions that they considered reliable, by using information on gene expression and paralogous proteins. The average IG2 value for our interactions that are contained in Deane *et al.*'s data set is -1.07, and the average IG2 value for those that are not is $+0.80$ ($P < 1.1 \times 10^{-9}$), again showing that lower IG2 values tend to occur for true protein–protein interactions. In addition, Mering *et al.* showed that interactions confirmed by more than one method, such as by both two-hybrid and tagged proteins/mass spectrometry (protein complex data) methods, are reliable. The average IG2 value for protein–protein interactions that were verified by the protein-complexing experiments of Gavin *et al.* (2002) or Ho *et al.* (2002) is -2.48,

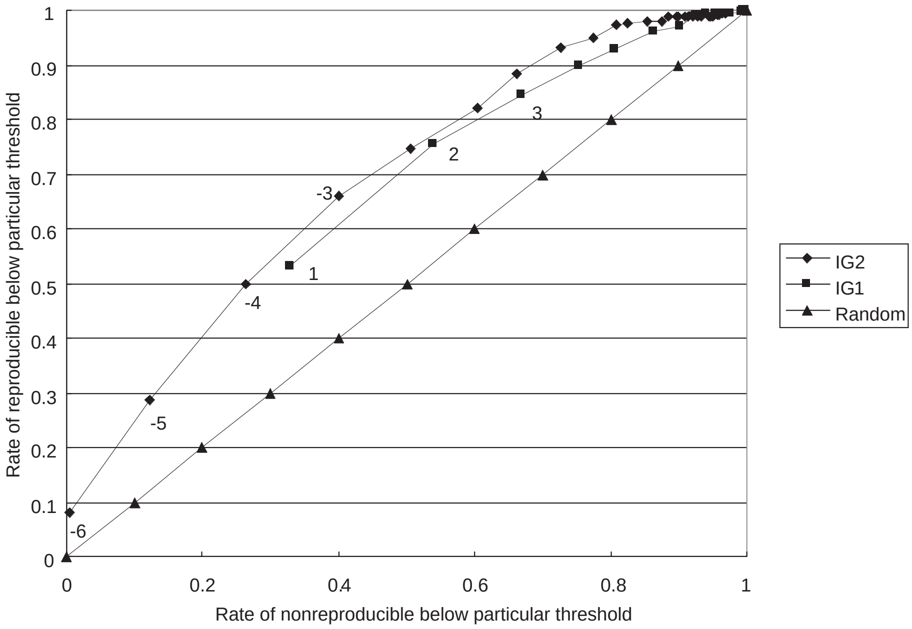


Fig. 4. Proportions of reproducible and non-reproducible interactions below various IG thresholds. Plots corresponding to IG thresholds from −6 to 20 (left to right) for IG2 (diamonds) and from 1 to 21 (left to right) for IG1 (squares) are shown. Some of the plots are labeled with corresponding IG threshold values. Theoretical proportions when the interactions are selected randomly regardless of their IG values are plotted in triangles.

whereas the average IG2 value for interactions not verified by their experiments is +0.39 ($P < 2.72 \times 10^{-16}$; complex data were converted to binary protein–protein interaction data by considering that the bait interacts directly with each protein in the complex; Bader and Hogue, 2002).

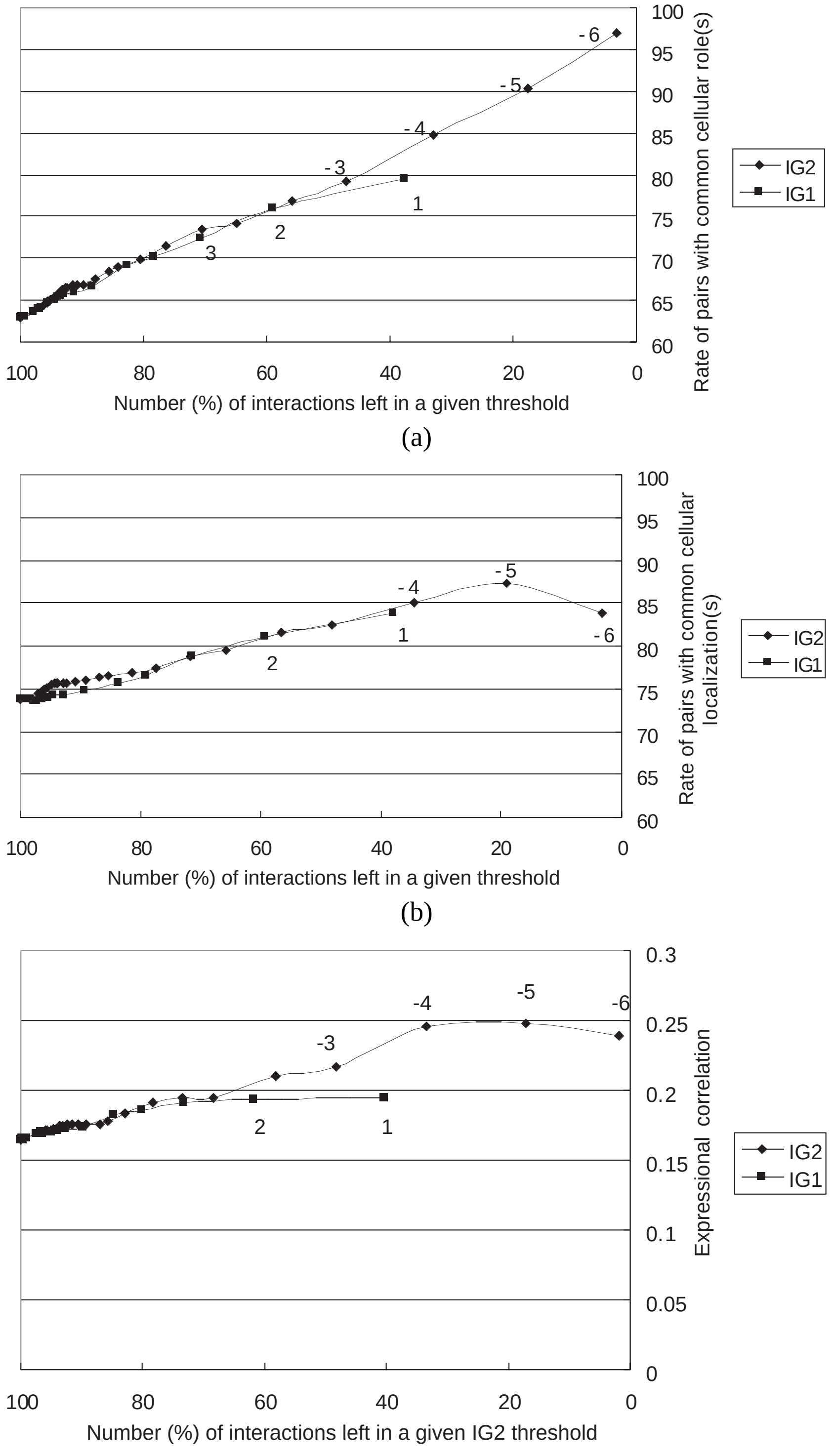
Another way to evaluate the utility of the IG2 measure is to calculate the proportions of reproducible and non-reproducible interactions falling below particular IG2 values. When we construct an interaction network by selecting interactions with IG2 values below a certain threshold, the IG2 measure is assumed to be useful if the set of interactions below that threshold contains many reproducible and few non-reproducible interactions. We used this approach to compare the IG1 and IG2 measures. Figure 4 shows the proportions of reproducible and non-reproducible interactions below various IG2 and IG1 thresholds, among all reproducible and non-reproducible interactions, respectively. IG2 (diamonds) performs better than IG1 (squares) for all thresholds, except where the proportion of reproducible interactions is close to 1. However, this region is not important in constructing a reliable interaction network, since the rate of non-reproducible interactions is also close to 1 in the region.

Most of the computational time needed to calculate IG2 is spent on calculating ($a1$, $a2$, l, f, and d) for each interaction. Theoretically, the computational time for calculating ($a1$, $a2$, l, f, and d) for each interaction scales as n^3, where n is the number of interacting partners for each protein, but the computational time for calculating IG1 is proportional to n. However in practice, proteins having many interaction partners are rather few, which reduces the calculation time. In fact, the real elapsed time to calculate IG2 for 3066 interactions was <1.5 h on a Pentium 4-based machine running Linux.

Functional associations and expressional correlations become clear in a reliable protein–protein interaction network

Interacting proteins generally share a common function and a common localization (‘guilt-by-association’ principle; Oliver, 2000). Approximately 63% of interacting proteins have at least one common cellular role (as defined in the Yeast Proteome Database; Costanzo *et al.*, 2001), and 73–76% of them have at least one common cellular localization (Hishigaki *et al.*, 2001; Schwikowski *et al.*, 2000). We investigated the accuracy of the IG2 measure by eliminating unreliable interactions and comparing its performance with that of IG1.

Figure 5a and b show the proportions of interacting protein pairs having common cellular roles and common localizations at various IG thresholds. As the IG2 threshold is decreased, the proportion of interacting pairs with common cellular roles and localizations increases,



(a)

(b)

(c)

Fig. 5. Evaluation of the reliability of a protein–protein interaction network constructed by eliminating protein–protein interactions having IG values below a given threshold. The network was evaluated by three measurements: (a) proportion of interacting proteins known to have a common cellular role; (b) proportion of interacting proteins known to have a common cellular localization; and (c) correlation of gene expression for the interacting proteins. Plots corresponding to IG thresholds of 20 to –6 (left to right) for IG2 (diamonds) and 21 to 1 (left to right) for IG1 (squares) are shown. Some of the plots are labeled with corresponding IG threshold values.

respectively, from 63% and 74% to >85%, indicating that interacting proteins are more likely to have common cellular roles and localizations. IG2 shows slightly better performance than IG1. However, unlike the IG1 value whose lower limit is 1, there is no lower limit for IG2. Therefore, reducing the IG2 threshold allows the proportion of interactions having a common cellular role to be increased to over 90% whereas the maximum proportion using the IG1 threshold is only 80%.

Some recent studies have shown significant correlation in the expression of genes that encode interacting proteins (Deane *et al.*, 2002; Ge *et al.*, 2001; Grigoriev, 2001; Jansen *et al.*, 2002; Kemmeren *et al.*, 2002; Mrowka *et al.*, 2001). Therefore, expressional correlation can also be used to assess the performance of IG1 and IG2. We calculated the average correlations of gene expression for protein partners whose interaction was considered to be reliable (i.e. below various IG thresholds), using expression data collected by Eisen *et al.* (1998). The average correlation of gene expression for the interactions significantly increased as the IG threshold become low, clearly demonstrating that the interaction network indeed becomes more reliable (Fig. 5c). IG2 is better than IG1 in selecting interacting pairs having a high degree of expressional correlation when the proportion of interactions left after eliminating unreliable interactions is below 60%.

DISCUSSION

We described a new interaction generality measure IG2, which is produced by a novel method for computationally assessing the reliability of candidate protein–protein interactions. The method includes principal-component analysis in evaluating the reliability of interactions, using five parameters ($\boldsymbol{p}_{11}, \boldsymbol{p}_{21}, \boldsymbol{p}_{31}, \boldsymbol{p}_{41}, \boldsymbol{p}_{51}$) for topological properties ($a1$, $a2$, l, f and d). In principle-component analysis, one can determine parameters and constants without needing to know whether each interaction is a true positive or a false positive. This is a great advantage because we do not know which of the non-reproducible interactions are indeed false positives.

The characteristics of IG1 (Saito *et al.*, 2002) are basically incorporated in IG2, because the parameters were effectively set to $\boldsymbol{p}_1 = (0, 0, 0, 0, 1)$ and $k_1 = -1$ in the calculating IG1 by ignoring the topological properties of the interaction network. We note that the parameter $\boldsymbol{p}_{11}$ for IG2 was determined to be a negative value in our analysis, which means that triangular interactions, shown as $a1$ in Figure 1, are likely to be valid. Actually, we found that triangular interactions do occur frequently in protein complexes (data not shown). In addition, $\boldsymbol{p}_{21}$ and $\boldsymbol{p}_{31}$ have relatively low values, which is consistent with the idea that interactions forming a closed loop might have an increased likelihood of being biologically relevant (Walhout *et al.*, 2000).

The usefulness of IG2 resides in that we can construct a protein network of desirable reliability in which new biological insights may wait to be uncovered. In addition, protein interactions deemed reliable by the IG2 measure may be useful for evaluating previously reported results obtained with unselected publicly available protein interactions, in which many false positives occur. Jeong *et al.* (2001) reported that lethal proteins (proteins whose deletion is lethal to the cell) are likely to have more interacting partners than non-lethal proteins do, suggesting that lethal proteins form relatively more extensive interaction complexes. We confirmed this tendency by using a reliable interaction data set (i.e. with an IG2 value < -1), in which we found that the mean number of interacting partners is 2.78 for lethal proteins and 1.94 for non-lethal ones. Recently, Maslov and Sneppen (2002) reported that proteins with many interacting partners are likely to interact with proteins with a few interacting partners. However, we could not confirm this tendency with the reliable interaction data set. This discrepancy may have occurred because Maslov and Sneppen used Ito's full data set for their analysis, which seems to include a relatively high proportion of false-positive interactions (Grigoriev, 2001; Ito *et al.*, 2001). Actually, all of the 116 interactions in the original protein interaction data set, which consist of protein pairs with many (>29) interacting partners and with few (<4) interacting partners, were removed from our reliable interaction data set (i.e. with an IG2 value < -1).

In proteomics studies it is definitely essential to construct reliable protein-interaction networks by integrating all available genome-wide interaction data sets. The IG2 measure may be useful for this purpose, at least for evaluating binary interaction data sets that have been obtained from biological experiments.

ACKNOWLEDGEMENTS

We thank Drs M. Harbers and S. Kondo for critically reading the manuscript. We also thank Drs D. Eisenberg and L. Salwinski for kindly sending us reliable protein–interaction data. The Yeast Proteome Database (YPD), constructed by Proteome Inc., was used in this study. This study was supported by a research grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y. H.

REFERENCES

Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A. Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Bader,G.D. and Hogue,C.W.V. (2002) Analyzing yeast protein–protein interaction data obtained from different resources. *Nat.*

Biotechnol., **20**, 991–997.

Costanzo,M.C., Crawford,M.E., Hirschman,J.E., Kranz,J.E., Olsen,P., Robertson,L.S., Skrzypek,M.S., Braun,B.R., Hopkins,K.L. Kondu,P. *et al.* (2001) YPD PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.*, **29**, 75–79.

Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, **1**, 349–356.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fellenberg,M., Albermann,K., Zollner,A., Mewes,H.W. and Hani,J. (2000) Integrative analysis of protein interaction data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 152–161.

Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M. Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Ge,H., Liu,Z., Church,G.M. and Vidal,M. (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.*, **29**, 482–486.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C. Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546–563–547.

Grigoriev,A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **29**, 3513–3519.

Hishigaki,H., Nakai,K., Ono,T., Tanigami,A. and Takagi,T. (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, **18**, 523–531.

Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K. Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jansen,R., Greenbaum,D. and Gerstein,M. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.*, **12**, 37–46.

Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y. Konno,H. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.

Kemmeren,P., van Berkum,N.L., Vilo,J., Bijma,T., Donders,R., Brazma,A. and Holstege,F.C. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell*, **9**, 1133–1143.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M. FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Legrain,P., Wojcik,J. and Gauthier,J.M. (2001) Protein–protein interaction maps: a lead towards cellular functions. *Trends Genet.*, **17**, 346–352.

Lorch,Y., Beve,J., Gustafsson,C.M., Myers,L.C. and Kornberg,R.D. (2000) Mediator–nucleosome interaction. *Mol. Cell*, **6**, 197–201.

Malik,S. and Roeder,R.G. (2000) Transcriptional regulation through mediator-like coactivators in yeast and metazoan cells. *Trends Biochem. Sci.*, **25**, 277–283.

Maslov,S. and Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.

Mewes,H.W., Frishman,D., Gruber,C., Geier,B., Haase,D., Kaps,A., Lemcke,K., Mannhaupt,G., Pfeiffer,F. Schuller,C. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **28**, 37–40.

Mrowka,R., Patzak,A. and Herzel,H. (2001) Is there a bias in proteome research? *Genome Res.*, **11**, 1971–1973.

Myers,L.C., Gustafsson,C.M., Bushnell,D.A., Lui,M., Erdjument-Bromage,H., Tempst,P. and Kornberg,R.D. (1998) The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain. *Genes Dev.*, **12**, 45–54.

Oliver,S. (2000) Guilt-by-association goes global. *Nature*, **403**, 601–603.

Pawson,T. and Nash,P. (2000) Protein–protein interactions define specificity in signal transduction. *Genes Dev.*, **14**, 1027–1047.

Saito,R., Suzuki,H. and Hayashizaki,Y. (2002) Interaction generality, a measurement to assess the reliability of a protein–protein interaction. *Nucleic Acids Res.*, **30**, 1163–1168.

Schwikowski,B., Uetz,P. and Fields,S. (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.

Suzuki,H., Fukunishi,Y., Kagawa,I., Saito,R., Oda,H., Endo,T., Kondo,S., Bono,H., Okazaki,Y. and Hayashizaki,Y. (2001) Protein–protein interaction panel using mouse full-length cDNAs. *Genome Res.*, **11**, 1758–1765.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode*C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M. Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.

Walhout,A.J., Sordella,R., Lu,X., Hartley,J.L., Temple,G.F., Brasch,M.A., Thierry-Mieg,N. and Vidal,M. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, **287**, 116–122.

Weller,S.C. and Romney,A.K. (1990) Principal component analysis. *Metric Scaling Correspondence Analysis*. SAGE, Newbury Park, CA, pp. 26–44.