Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*

Hui Ge[1], Zhihua Liu[2], George M. Church[3] & Marc Vidal[1]

Published online: 5 November 2001, DOI: 10.1038/ng776

© 2001 Nature Publishing Group http://genetics.nature.com

Genomic and proteomic approaches can provide hypotheses concerning function for the large number of genes predicted from genome sequences[1–5]. Because of the artificial nature of the assays, however, the information from these high-throughput approaches should be considered with caution. Although it is possible that more meaningful hypotheses could be formulated by integrating the data from various functional genomic and proteomic projects[6], it has yet to be seen to what extent the data can be correlated and how such integration can be achieved. We developed a 'transcriptome–interactome correlation mapping' strategy to compare the interactions between proteins encoded by genes that belong to common expression-profiling clusters with those between proteins encoded by genes that belong to different clusters. Using this strategy with currently available data sets for *Saccharomyces cerevisiae*, we provide the first global evidence that genes with similar expression profiles are more likely to encode interacting proteins. We show how this correlation between transcriptome and interactome data can be used to improve the quality of hypotheses based on the information from both approaches. The strategy described here may help to integrate other functional genomic and proteomic data, both in yeast and in higher organisms.

Expression profiling, protein–protein interaction mapping, protein-localization mapping and large-scale phenotypic analysis projects have been developed to various degrees for several model organisms[1–5]. We first focused our attempts on the integration of functional genomic approaches using expression profiles and protein-interaction maps for *S. cerevisiae*, mainly because these represent the two largest sets of available data. After collecting expression data, clustering analysis can be used to group genes according to the similarity of their expression across different experimental conditions and genetic backgrounds[7]. These expression clusters result in hypotheses of function based on the assumption that groups of genes that are co-expressed are likely to mediate related biological functions. Similarly, interaction clusters can be generated from protein–protein interaction mapping data and are indicative of the existence of protein complexes and/or signal transduction pathways[3].

To investigate a potential correlation between expression clusters and interaction clusters, we devised the transcriptome–interactome correlation mapping strategy (Fig. 1). We generated a two-dimensional (2-D) matrix by organizing the

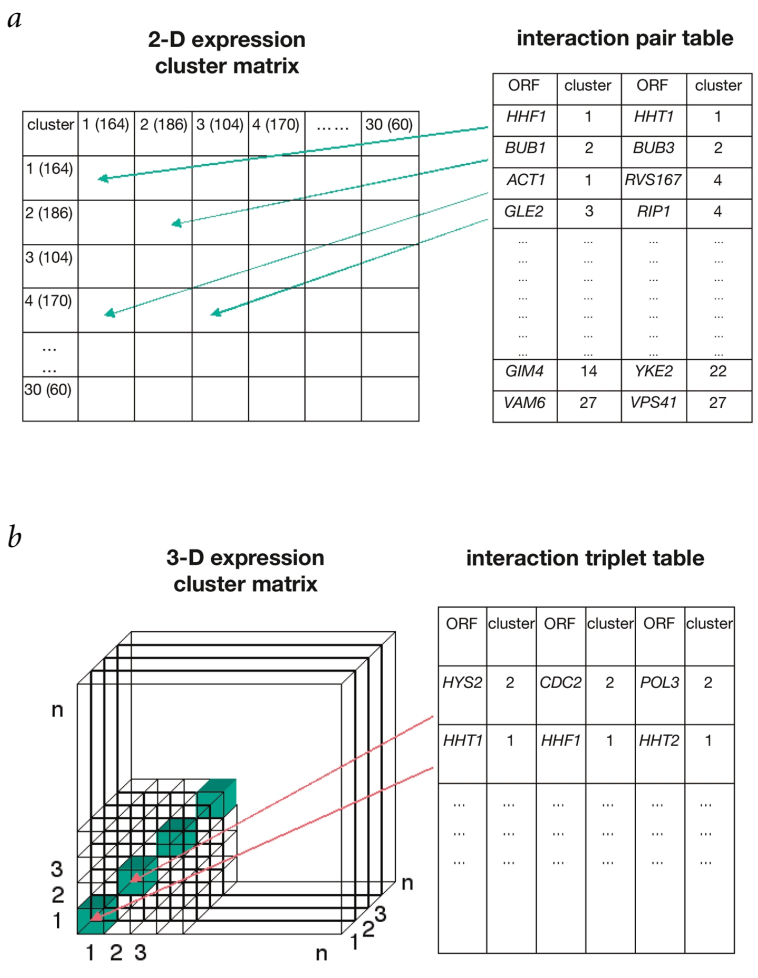

Fig. 1 A general strategy for transcriptome–interactome correlation mapping. ***a***, The 2-D matrix on the left shows all pair-wise combinations between the clusters of an expression profiling experiment. The numbers assigned to each cluster are indicated on the corresponding rows and columns of the matrix along with the number of genes each cluster contains (in parenthesis). The table on the right shows protein interaction pairs together with the expression cluster to which the corresponding genes belong. For each interaction pair, an arrow points to its corresponding square in the 2-D expression-cluster matrix. ***b***, As in *a*, a 3-D matrix can be generated to integrate triplets of interactions.

[1]*Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.* [2]*Department of Neurology, Brigham and Women's Hospital and Center for Neurologic Diseases, Harvard Medical School, Boston, Massachusetts, USA.* [3]*The Lipper Center for Computational Genetics and Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to M.V. (e-mail: marc_vidal@dfci.harvard.edu).*

© 2001 Nature Publishing Group http://genetics.nature.com

clusters derived from a set of related transcriptional profiling experiments into two identical axes. We then arranged in the matrix pairs of genes whose products can interact, according to the cluster(s) to which each gene belongs (Fig. 1*a*). For n clusters, the matrix arrangement results in n^2 squares, with each square representing all pair-wise combinations of genes either in a single cluster (diagonal or 'intracluster' squares) or between two different clusters (nondiagonal or 'intercluster' squares). Thus, for a matrix of n clusters, pairs of interactions can be assigned to their corresponding intracluster or intercluster squares. For each square, we calculated an index of protein interaction density (PID) as the ratio of the number of observed protein interaction pairs (IP) to the total number of possible pair-wise combinations of protein pairs (PP; PID=IP/PP). We reasoned that, for a given list of interactions and a set of expression-profiling conditions, significantly higher PIDs for intra- versus intercluster squares would reveal a correlation between transcription profiles and protein-interaction maps.

To generate a transcriptome–interactome correlation map, we first used the results of a clustering analysis[8] carried out by a *k*-means algorithm[9] with yeast cell-cycle expression data (Methods)[10]. For this analysis, we organized approximately 3,000 open reading frames (ORFs) showing significant transcriptional changes across two consecutive cell cycles into 30 clusters. The relatively high biological significance of these clusters is indicated by several facts[8]: (i) nearly half of the clusters show significant enrichment in genes that are known to mediate similar functions (ii) the promoters of many of the genes in these clusters contain related upstream regulatory binding sites and (iii) the 'tightness' of the clusters—that is, the average distance of all members of a cluster from the cluster mean—correlated with functional enrichment and the presence of potential binding sites (P value=0.02 and 0.006, respectively)[8]. For the protein–protein interaction data, we used 1,666 interaction pairs described in the literature and collected in YPD[11] and MIPS (the 'literature' data set)[12]. In this list, 335 protein pairs reside in the set of cell-cycle expression clusters and can be assigned to their corresponding intra- or intercluster squares. As a negative control, we randomized the 335 pairs of interactions and assigned the resulting gene pairs to their corresponding squares. For both sets, we calculated the PID of each square and represented it by a color gradient. Overall, the resulting transcriptome–interactome correlation map shows a high-density region along the diagonal, indicating that the combination of genes from the same clusters results in a higher PID (Fig. 2*a*). To confirm this, we calculated both the average PID of all intracluster squares and that of all intercluster squares. We found that the intracluster region has an average PID 5.8 times as high as the intercluster region (Fig. 2*b*). In contrast, the intra- and intercluster average PIDs are very similar for the randomized sample (Fig. 2*a*,*b*).

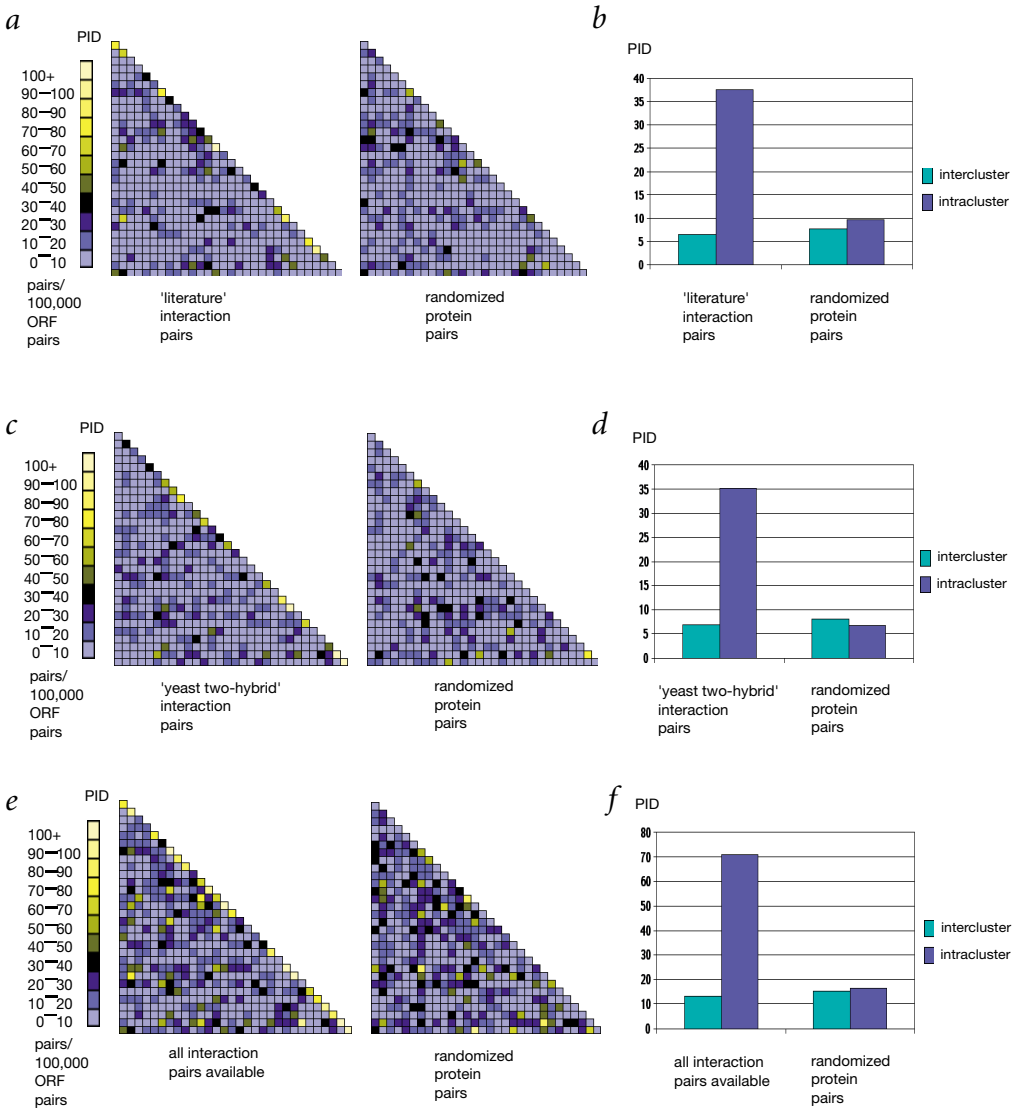

Fig. 2 Transcriptome–interactome correlation maps. ***a–f***, We calculated the protein interaction density (PID) for each square in the matrix as the ratio of interaction pairs assigned to the square (IP) over the total number of protein pairs possibly formed by combinations of the genes in the square (PP). PIDs are represented in the map by a color system, as indicated in the scale on the left side of *a*,*c*,*e*. Control maps can be generated by the same approach from randomized protein pairs (*a*,*c*,*e*, right side). The average PIDs from all intracluster squares (in the diagonal) and inter-cluster squares (outside the diagonal) can be calculated from the correlation maps (*b*,*d*,*f*). The unit of PID in each panel is 'interaction pairs/100,000 ORF pairs'. We constructed transcriptome–interactome correlation maps using cell-cycle expression–profiling clusters and protein interaction data from the literature (*a*,*b*), from genome-wide yeast two-hybrid screens (*c*,*d*) or from the combination of both (*e*,*f*).

© 2001 Nature Publishing Group http://genetics.nature.com

© 2001 Nature Publishing Group http://genetics.nature.com

We investigated the relevance of transcriptome–interactome correlation maps for interactions obtained from large-scale protein-interaction mapping projects. We combined the data from two maps[13–15]; in the resulting list of 1,709 potential protein–protein interactions, the genes for 347 interacting pairs could be assigned to one of the 30 clusters. The results are similar to those obtained for the 'literature' protein interaction data set (Fig. 2*c*). The average PID of the intracluster region is 5.1 times that of the intercluster region, whereas for the negative control the intra- and intercluster average PIDs are very similar (Fig. 2*d*). This correlation seems to apply to interactions identified by large-scale yeast two-hybrid mapping as well as those identified by conventional strategies. We combined the protein interaction data in a list of 3,222 pairs of yeast protein–protein interactions. The transcriptome–interactome correlation map generated from this combined protein interaction data set also shows higher PIDs for the intracluster squares (Fig. 2*e*,*f*). Notably, in addition to the intracluster squares, a few intercluster squares also seem to have high PIDs (Fig. 2*a*,*c*,*e*). These 'outliers' may indicate 'crosstalk' between different clusters, which may suggest crosstalk between different pathways; however, occasional high-PID squares also occur in the randomized controls (Fig. 2*b*,*d*,*f*).

We examined the statistical significance of the potential correlation between transcriptome and interactome data by comparing the total number of experimentally observed intracluster interactions to the number expected if one assumes a random distribution of interactions across the different clusters. In the combined interactome data set, 117 of a total of 670 interactions in the correlation map belonged to intracluster squares, whereas only 25 would be expected to do so in a random distribution (P=9.8×10^{-43}; Table 1). We conclude that a statistically significant correlation exists between expression clusters across the yeast cell cycle and large data sets of protein–protein interactions.

Table 1 • Statistical analysis

Group size	Data set	Total in map	Expected[a]	Observed[a]	P value[b]
pair	cell cycle	670	25	117	9.8×10^{-43}
	sporulation	309	46	115	1.1×10^{-21}
	stress response	731	44	165	2.0×10^{-49}
triplet	cell cycle	1,632	3	40	3.0×10^{-33}
	sporulation	495	12	57	2.1×10^{-22}
	stress response	2,716	11	67	2.2×10^{-31}

[a]The observed number of groups (k) whose members belong to the same cluster is compared with the expected number, assuming the interaction groups are randomly distributed. [b]The probability for obtaining at least k observed groups in the intracluster region by chance is calculated for each data set using cumulative binomial distribution.

This correlation could be specific to cell cycle–regulated genes. To evaluate this possibility, we used two other independent sets of expression profiling data generated from yeast cells undergoing meiosis[16] or subjected to various stresses[17]. We observed a similar statistically significant correlation in these two transcriptome–interactome correlation maps (Fig. 3 and Table 1). The correlation between transcriptome and interactome may thus apply to many different experimental situations.

Molecular complexes and signal-transduction pathways are often the result of several protein–protein interactions. To estimate the extent to which transcriptome–interactome correlation maps could be used for more complex sets of interactions, we determined the correlation between the expression clusters described above and triplets (series of two consecutive interactions such as A-B-C) of interacting proteins. We derived a list of interaction triplets from the list of 3,222 interaction pairs from the combined data and arranged them in 3-D matrices similar to the 2-D matrix described above (Fig. 1*b*). Of 1,632 triplets that can be analyzed from the cell-cycle expression–clustering data set, the observed number of triplets for which all three members belong to the same expression cluster (40) is significantly higher than that expected from a random distribution (3, P=3.0×10^{-33}; Table 1). We obtained similar results when interaction triplets were integrated with the meiotic and cell-stress transcriptome data (P=2.1×10^{-22} and 2.2×10^{-31}, respectively; Table 1). We therefore suggest that it should be possible to combine more complex sets of interactions with transcription-profiling clusters. Indeed, N-dimensional matrix settings could, in principle, be generated to overlap sets of N interactions (Methods).

The correlation described above suggested that interactome data could help to identify expression clusters with relatively greater biological relevance. To test this possibility, we calculated the P values

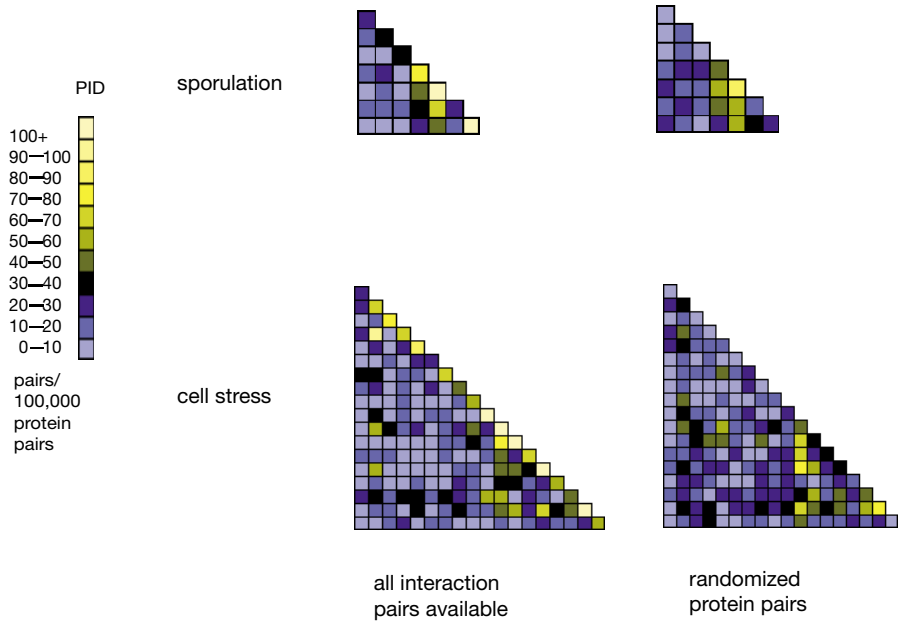

Fig. 3 Transcriptome–interactome correlation maps for sporulation and cell-stress data sets (see legend of Fig. 2). We constructed transcriptome–interactome correlation maps using sporulation (top) and cell-stress (bottom) profiling clusters and protein interaction data currently available.

© 2001 Nature Publishing Group http://genetics.nature.com

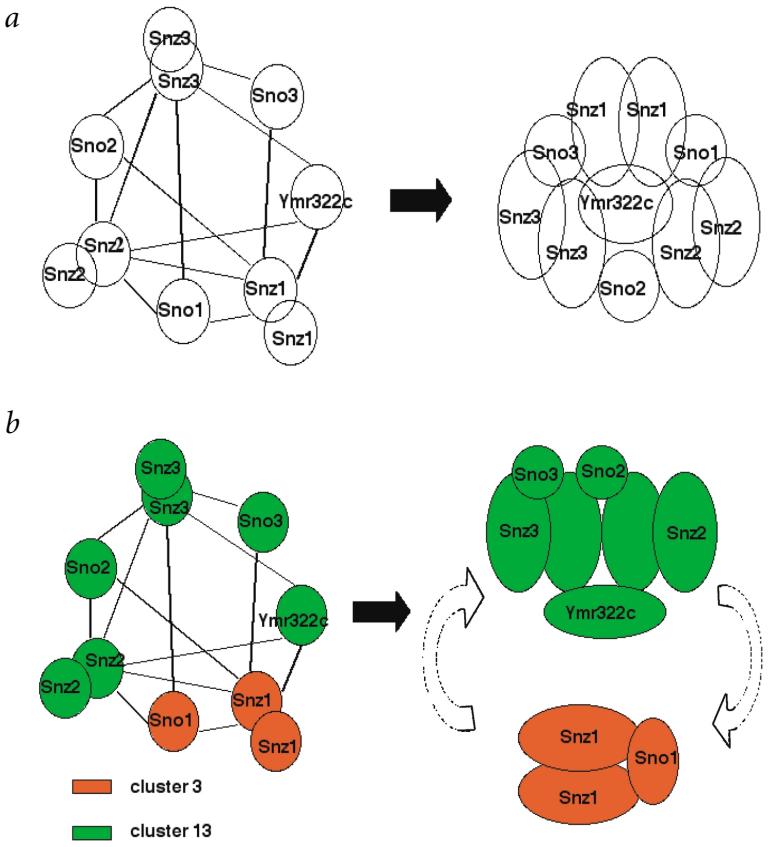

Fig. 4 Improved model from the integration of transcriptome and interactome data. ***a***, Interaction network obtained from the combined data set of protein–protein interactions. Circles represent proteins and lines represent two-hybrid interactions. ***b***, Improved model obtained by taking into account expression profiles.

for individual intracluster PIDs of the cell cycle and compared these values to different parameters used to characterize the functional clusters[8]. Clusters with a significantly higher PID tend to contain genes with potential DNA binding motifs and functional enrichment (Web Table A). In addition, clusters with a significantly higher PID tend be tighter, as measured by the average Euclidean distance (Web Table A). Thus, it is possible that significantly higher PIDs for individual squares in the transcriptome–interactome map may point to clusters containing genes that are more likely to be functionally linked.

The correlation between transcriptome and interactome data also suggests that their integration can be used to improve the hypotheses resulting from protein-interaction maps. For example, the crude protein-interaction map data seem to indicate that the Snz and Sno proteins (involved in the response to stress)[18] and the product of the uncharacterized *YMR322C*-predicted ORF form a complex (Fig. 4*a*). The correlation map reveals, however, that the corresponding seven genes belong to two different clusters in the stress experiments mentioned above[17]. Snz1 and Sno1 belong to expression cluster 13, whereas the other five genes are in expression cluster 3 (Fig. 4*b*). This suggests that two Snz/Sno subcomplexes may exist, one formed by Snz2, Sno2, Snz3, Sno3 and Ymr322C and the other by Snz1p and Sno1p. The intercluster interactions may thus represent false positives of the two-hybrid method and the corresponding proteins might not interact *in vivo*. On the other hand, the putative Snz/Sno complex may exist transiently, as the clustering analysis does not necessarily preclude partial overlaps of expression. In this case, it is possible that one subcomplex may regulate the other. The integration of data from expression profiling and yeast two-hybrid analysis suggests an improved model for these proteins (Fig. 4*b*) that is consistent with one previously reported[18]. In addition, the potential involvement of Ymr322C in the Snz2(3)/Sno2(3) complex is supported by their belonging to the same cluster (Fig. 4*b*). This example shows how predictions based on the combination of expression profile data and protein–protein interaction data can be more informative than those based on either approach alone.

The strategy described here reveals a global correlation between expression profiling and protein-interaction mapping data in yeast. Although one may expect that proteins that form complexes or interact in signal transduction pathways are encoded from co-expressed genes, there are well-described exceptions to this assumption. For example, across the cell cycle, the genes encoding cyclin-dependent kinases are uniformly transcribed, whereas those of their regulatory subunits, the cyclins, are tightly regulated. The correlation observed here thus shows that despite those exceptions, the integration of transcriptome and interactome data may help improve the hypotheses emerging from either approach individually. We suggest that similar transcriptome–interactome correlation maps could be used to unravel comparable correlations in other model organisms. Because data from other functional genomic approaches, such as protein-localization mapping and high-throughput loss-of-function analysis, can potentially be arranged in clusters[6], correlation maps might be adapted to the integration of data from these approaches as well. It is possible that such multidimensional integration will generate increasingly meaningful biological hypotheses.

Methods

Data source. We obtained protein–protein interaction pairs from different sources: http://www.mips.biochem.mpg.de/proj/yeast/tables/interaction/physical_interact.html, http://www.proteome.com and http://depts.washington.edu/yeastrc/th_11.htm (only the core data was used). The clustering analysis data sets of cell cycle–regulated genes, meiosis-regulated genes and cell stress–regulated genes are available at http://arep.med.harvard.edu/network_discovery, http://re-esposito.bsd.uchicago.edu and http://www.hsph.harvard.edu/geneexpression, respectively. We chose these three experiments because the cluster analysis contained relatively large numbers of genes, which enabled us to include large numbers of protein interaction pairs or triplets in our analysis.

Calculation of the protein interaction density. Each square of a transcriptome–interactome correlation matrix is defined by a ($k1$,$k2$) pair, where $k1$ and $k2$ refer to the numbers assigned to each of the expression profiling clusters organized in the matrix. For each square, we calculated the protein interaction density (PID) as IP/PP, where IP is the observed number of protein interaction pairs and PP is the number of all pairwise combinations of proteins. $PP(k1,k2)=n_{k1}n_{k2}$ (for intercluster squares, $k1<>k2$) or $PP(k1,k2)=n_{k1}(n_{k1}+1)/2$ (for intracluster squares, $k1=k2$), where n_{k1} and n_{k2} are the number of genes that belong to cluster $k1$ and cluster $k2$, respectively.

© 2001 Nature Publishing Group http://genetics.nature.com

© 2001 Nature Publishing Group http://genetics.nature.com

Statistical analysis. To determine whether the enrichment of protein interaction pairs in the intracluster region is statistically significant, we used the cumulative binomial distribution given by the formula:

$$P(i>i_0) = \sum_{i=i_0}^{I} p^i (1-p)^{I-i} \left[\frac{I!}{I!(I-i)!}\right]$$

where I is the total number of protein interaction pairs sampled, i_0 is the number of protein interaction pairs falling in the intracluster region and p is the probability of a protein interaction falling in the intracluster region, assuming the protein interaction pairs are randomly distributed. We calculated p by the following formula:

$$p = \frac{\sum_{k=1}^{K} [n_k(n_k+1)/2]}{T(T+1)/2}$$

where K is the total number of clusters, n_k is the number of genes in cluster k and T is the total number of genes in all clusters. For example, in the correlation map of cell-cycle expression profiling and 'literature' interaction data sets, K=30, T=2945, n_1=164, n_2=186, … n_{30}=60, I=335, i_0=62.

In principle, an N-dimensional setting ($N \geq 2$) could be generated to correlate the expression clusters with sets of N interactions. The same cumulative binomial distribution formula could be used to determine the statistical significance of co-expressed N interaction sets. The probability p of an N-interaction set having its members falling in the same expression cluster can be calculated as:

$$p = \frac{\sum_{k=1}^{K} \{\prod_{j=1}^{N} [(n_k+j-1)/j]\}}{\prod_{j=1}^{N} [(T+j-1)/j]}$$

where K is the total number of clusters, n_k is the number of genes in cluster k, and T is the total number of genes in all clusters. The expected number (E) of protein interaction pairs (or triplets) falling in intra-cluster squares (or cubes) is calculated as:

$$E = sp$$

where s is the total number of protein interaction pairs (or triplets) in the matrix and p is the probability of a protein interaction pair (or triplet) falling in the intracluster squares (or cubes), assuming the protein interaction pairs are randomly distributed.

Note: Supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Acknowledgments

We thank P. Lansburry for his support, Mr. and Ms. Fu for their financial help, R.E. Esposito for exchange of information, V. Rebel and members of the Vidal lab for insightful suggestions and M. Walhout, T. Brüls and N. Thierry-Mieg for reading the manuscript. This work was supported by grant 1 RO1 HG01715-01 from the National Human Genome Research Institute, awarded to M.V.

Received 9 July; accepted 10 October 2001.

1. Lockhart, D.J. & Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
2. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
3. Walhout, A.J.M. & Vidal, M. Protein interaction maps for model organisms. *Nature Rev. Mol. Cell. Biol.* **2**, 55–62 (2001).
4. Kumar, A. & Snyder, M. Emerging technologies in yeast genomics. *Nature Rev. Genet.* **2**, 302–312 (2001).
5. Sternberg, P.W. Working in the post-genomic *C. elegans* world. *Cell* **105**, 173-176 (2001).
6. Vidal, M. A biological atlas of functional maps. *Cell* **104**, 333–339 (2001).
7. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
8. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
9. Hartigan, J.A. *Clustering Algorithms* (Wiley, New York, 1975).
10. Cho, R.J. *et al*. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
11. Hodges, P.E., McKee, A.H., Davis, B.P., Payne, W.E. & Garrels, J.I. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.* **27**, 69–73 (1999).
12. Mewes, H.W. *et al*. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
13. Uetz, P. *et al*. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
14. Ito, T. *et al*. Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between yeast proteins. *Proc. Natl Acad. Sci. USA* **97**, 1143–1147 (2000).
15. Ito, T. *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574 (2001).
16. Primig, M. *et al*. The core meiotic transcriptome in budding yeasts. *Nature Genet.* **26**, 415–423 (2000).
17. Jelinsky, S.A., Estep, P., Church, G.M. & Samson, L.D. Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Mol. Cell. Biol.* **20**, 8157–8167 (2000).
18. Padilla, P.A., Fuge, E.K., Crawford, M.E., Errett, A. & Werner-Washburne, M. The highly conserved, coregulated SNO and SNZ gene families in *Saccharomyces cerevisiae* respond to nutrient limitation. *J. Bacteriol.* **180**, 5718–5726 (1998).