Does mapping reveal correlation between gene expression and protein–protein interaction?

Genome-wide approaches make systematic inferences about function, regulation and interaction of genes and their corresponding protein products. The challenge is to integrate different sources of information[1–3], such as mRNA abundance[4] and protein–protein interaction data[5], to derive new, biologically relevant and testable hypotheses.

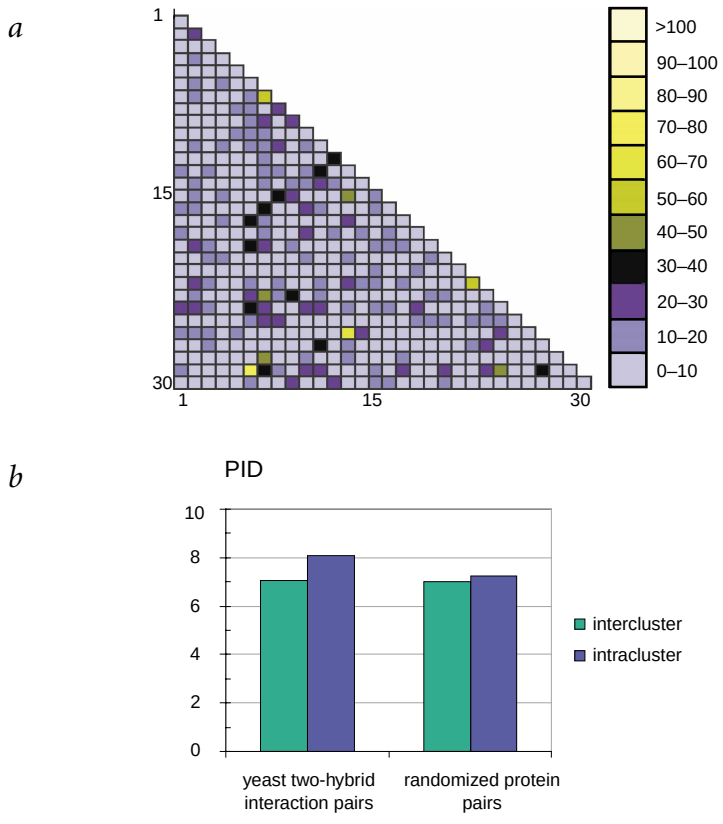
Ge *et al.*[6] carried out a large-scale mapping analysis of gene expression and protein–protein interaction data in the yeast *Saccharomyces cerevisiae*. The authors contrasted patterns of pair-wise combinations of genes within the same expression cluster (intracluster) and between different expression clusters (intercluster), and focused on an important biological problem: the relationship between coordinately expressed genes and interaction of their protein products.

Ge *et al.*[6] reported a significantly higher fraction of protein interaction densities (PIDs), that is, the number of observed protein interaction pairs over the total number of possible pair-wise combinations, in intracluster protein pairs as compared with intercluster pairs. They interpreted their findings as evidence that genes with similar expression profiles are more likely to encode interacting proteins. Analyzing two different protein–protein interaction databases, one derived from literature searches[7] (Munich Information Center for Protein Sequences database) and the other from genome-wide yeast two-hybrid (Y2H) experiments[8,9], the authors found that both data collections gave similar results. This contrasts with other observations of substantial differences between the literature survey data and Y2H assays[10,11]. Furthermore, the extent of the correlation between the transcription and protein interaction reported in Ge *et al.*[6] is markedly higher than that in a similar, previously reported analysis[11].

Here, we wish to point out that these discrepancies can be resolved. Though it does not concern the potential usefulness of the algorithm applied by Ge *et al.*[6], we find that their analysis favors an alternative explanation.

Ge *et al.*[6] attributed their results of generally higher PID values in intracluster pairs versus intercluster pairs to the global pattern of correlation between expression-profiling and protein-interaction data in yeast. Using the protein–protein interaction data from Y2H assays[8,9] and mapping the data corresponding to the clusters introduced in Tavazoie *et al.*[12], we were able to reproduce the findings. But we found that approximately 67% of the intracluster pairs constituted protein self-interactions. Although self-interacting proteins are valid in principle, they should have been excluded from the study under discussion[6] because self-interacting pairs have identical expression patterns by definition. As the authors did not exclude protein self-interactions, we studied the extent to which self-interactions might explain the unusually high intracluster PID values. We assessed the change in global patterns of correlation by computing R, the ratio of average intracluster PIDs over average intercluster PIDs (see Figure). When self-interactions were excluded, the number of intracluster protein pairs did not differ significantly from the random expectation ($P = 0.093$ at 5% significance level, binomial distribution), and $R \approx 1.1$ was close to $R \approx 1$ expected for random pairs. It is, therefore, implausible that interactions between distinct proteins would give rise to $R > 5$ as observed by Ge *et al.*[6]

We finally wish to point out that the relationship between coordinately expressed yeast genes and Y2H protein interactions can be identified in an alternative analysis. A histogram of correlation coefficients (r) between mRNA abundance levels for protein pairs can be used to test for positively or negatively regulated pairs compared with random controls. For instance, using gene-expression



Transcriptome–interactome correlation map. ***a***, Protein interaction density (PID) matrix for gene-expression and yeast two-hybrid protein–protein interactions. The rows and columns correspond to clusters of genes with similar mRNA abundance levels during the cell cycle, and the color of each matrix element encodes the PID value (scaled by a factor of 10^5). Following Ge *et al.*[6], PIDs were computed as the scaled ratio between the observed number of interaction pairs and the number of all pair-wise combinations of proteins, but with the exclusion of protein self-interactions. ***b***, PID values for intercluster and intracluster protein pairs (scaled by a factor of 10^5). Protein self-interactions were excluded, PIDs for intra- and intercluster elements of the PID matrix (*a*) were averaged separately, and the arithmetic mean is shown. The average PID inter- to intracluster ratio of about 1.1 was close to the value expected for randomly interacting pairs, and the number of intracluster pairs did not differ significantly from the random expectation ($P = 0.093$, binomial distribution). *a*,*b* correspond to Fig. 2c,*d* in Ge *et al.*[6].

data of the yeast's cell cycle[13] and Y2H[8,9] data, we found a significant shift toward positive r values for interacting non-self protein pairs ($P < 10^{-7}$, Kolmogorov–Smirnov test) when compared with random controls.

In conclusion, we found that the mapping approach may fail to identify a significant correlation between coordinated gene expression and protein interaction for non-self interactions, whereas a correlation effect was observed using alternative methods[11]. The high proportion of self-interactions may be of biological interest in its own right, for example, in the formation of regulatory homodimers[14].

Ralf Mrowka[1], Wolfram Liebermeister[2] & Dirk Holste[3]

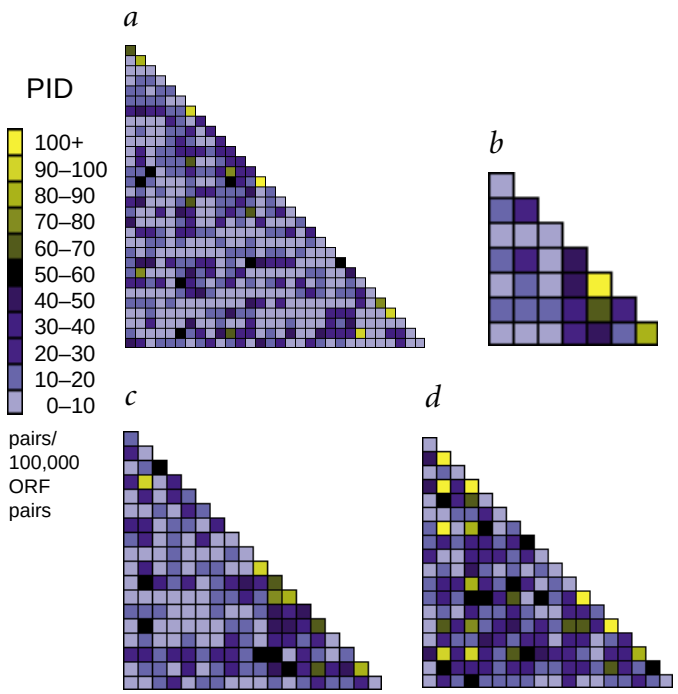
[1]Johannes Müller Institute for Physiology, Humboldt University, Tucholskystr. 2, D-10117 Berlin, Germany. [2]Max Planck Institute for Molecular Genetics, Berlin, Germany. [3]Department of Biology, Massachusetts Institute for Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to R.M. (e-mail: ralf.mrowka@charite.de).

1. Vidal, M. *Cell* **104**, 333–339 (2001).
2. Marcotte, E.M. *et al. Nature* **402**, 83–86 (1999).
3. Pilpel, Y. *et al. Nature Genet.* **29**, 153–159 (2001).
4. Lockhart, D.J. & Winzeler, E.A. *Nature* **405**, 827–836 (2000).
5. Legrain, P. *et al. Trends Genet.* **17**, 346–352 (2001).
6. Ge, H. *et al. Nat. Genet.* **29**, 482–486 (2001).
7. Mewes, H.W. *et al. Nucleic Acids Res.* **28**, 37–40 (2000).
8. Uetz, P. *et al. Nature* **403**, 623–627 (2000).
9. Ito, T. *et al. Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
10. von Mering, C. *et al. Nature* **417**, 399–403 (2002).
11. Mrowka, R. *et al. Genome Res.* **11**, 1971–1973 (2001).
12. Tavazoie, S. *et al. Nat. Genet.* **22**, 281–285 (1999).
13. Cho, R.J. *et al. Mol. Cell* **2**, 65–73 (1998).
14. Wolberger, C. *Curr. Opin. Struct. Biol.* **6**, 62–68 (1996).

—In Reply

In their correspondence entitled 'Does mapping reveal correlation between gene expression and protein-protein expression', Mrowka *et al.* state that homodimers should be removed from protein–protein interaction data sets in comparative analyses between transcriptome and interactome. We agree with this comment, and we greatly appreciate their input. Mrowka *et al.* also propose that the inclusion of homodimers in the interactome data sets analyzed in Ge *et al.*[1] might have led to the observation that large-scale two-hybrid (Y2H) data are similar to protein–protein interaction data previously published and archived in Yeast Proteome Database (YPD) and Munich Information Center for Protein Sequences (MIPS) in terms of their correlation with clusters obtained from expression profiling experiments. Furthermore, Mrowka *et al.* question the validity of the mapping approach described in Ge *et al.*[1] as a general method to identify correlation between coordinated gene expression and heterodimeric protein interactions.

To address those concerns, we carried out statistical analyses for each combination of transcriptome and interactome (pairs of interactors) data sets that were analyzed in Ge *et al.*[1], excluding the homodimers (see Table). The resulting P values showed that, overall, the correlation between transcriptome and interactome data is statistically significant ($P \leq 0.05$), with the exception of one combination—cell-cycle expression profiling data[2] versus Y2H data obtained by Uetz *et al.*[3] and Ito *et al.*[4] (we included only the 'core' data from Ito *et al.*[4], that is, interactions that were found at least three times). Thus, we conclude that, as in Ge *et al.*[1] but also in Grigoriev[5], Mrowka *et al.*[6], Jansen *et al.*[7] and Kemmeren *et al.*[8], pairs of genes that encode protein–protein interaction partners tend to be co-expressed. We also agree with Mrowka *et al.* that significant differences exist between the YPD/MIPS and the Uetz *et al.*[3]/Ito *et al.*[4] core data sets. But we point out that the statistical analysis presented in Ge *et al.*[1] pertained only to the combined YPD/MIPS and the Uetz *et al.*[3]/Ito *et al.*[4] core data sets (see Table and compare to Table 1 in Ge *et al.*[1]). Finally, as in Ge *et al.*[1], we carried out statistical analyses on triplets of interactors for each combination of transcriptome–interactome data sets, excluding the homodimers (see Table). Again, the resulting P values showed that, overall, the correlation between transcriptome and interactome data was significant.



Transcriptome–interactome correlation maps. The protein interaction density (PID) for each square in the matrix was calculated as the ratio of interaction pairs assigned to the square over the total number of protein pairs possibly formed by combinations of the gene products in the square. PIDs are represented in the map by a color system as indicated in the scale. The unit of PID in each panel is interaction pairs per 100,000 ORF pairs. Transcriptome–interactome correlation maps were constructed using different combinations of expression-profiling clusters and protein–protein interaction data sets. ***a***, Cell-cycle expression-profiling clusters and combined protein interaction data (from YPD/MIPS and from genome-wide yeast two-hybrid screens); ***b***, sporulation expression-profiling clusters and combined protein interaction data; ***c***, cell-stress expression-profiling clusters and combined protein interaction data; ***d***, cell-stress expression-profiling clusters and protein interaction data from a large-scale pull-down experiment by Ho *et al.*[9].