# Relating Whole-Genome Expression Data with Protein-Protein Interactions

Ronald Jansen,[1,4] Dov Greenbaum,[2,4] and Mark Gerstein[1,3,5]

[1]Departments of Molecular Biophysics and Biochemistry, [2]Genetics, and [3]Computer Science, Yale University, New Haven, Connecticut 06520, USA

We investigate the relationship of protein-protein interactions with mRNA expression levels, by integrating a variety of data sources for yeast. We focus on known protein complexes that have clearly defined interactions between their subunits. We find that subunits of the same protein complex show significant coexpression, both in terms of similarities of absolute mRNA levels and expression profiles, e.g., we can often see subunits of a complex having correlated patterns of expression over a time course. We classify the yeast protein complexes as either permanent or transient, with permanent ones being maintained through most cellular conditions. We find that, generally, permanent complexes, such as the ribosome and proteasome, have a particularly strong relationship with expression, while transient ones do not. However, we note that several transient complexes, such as the RNA polymerase II holoenzyme and the replication complex, can be subdivided into smaller permanent ones, which do have a strong relationship to gene expression. We also investigated the interactions in aggregated, genome-wide data sets, such as the comprehensive yeast two-hybrid experiments, and found them to have only a weak relationship with gene expression, similar to that of transient complexes. (Further details on genecensus.org/expression/interactions and bioinfo.mbb.yale.edu/expression/interactions.)

Analysis of gene expression data is currently one of the most exciting areas in genomics. Computationally, it involves clustering and grouping individual expression measurements and interrelating them to other sources of information, such as phenotypes, functional classifications, or cellular responses (Golub et al. 1999; Brown et al. 2000; Califano et al. 2000; Gaasterland and Bekiranov 2000; Raychaudhuri et al. 2001; Subrahmanyam et al. 2001). In particular, functional assignment of uncharacterized genes can take place through transferring the annotation from a characterized gene (gathered from databases such as the catalog of the Munich Information Center for Protein Sequences (MIPS) or Gene Ontology (Ashburner et al. 2000; Mewes et al. 2000)) to an uncharacterized gene when their expression profiles are strongly related by a similarity criterion (such as the correlation coefficient). While this procedure is usually not sufficient to unambiguously determine the function of an uncharacterized gene, it can be the starting point (e.g., in target selection) for further genetic experiments, functional characterization, or high-throughput proteomic analysis (Luscombe et al. 1998; Westhead et al. 1999; Christendat et al. 2000a,b; Eisenberg et al. 2000; Emili and Cagney 2000; Gerstein and Jansen 2000).

An important component of functional annotation is characterizing protein interactions, as these often circumscribe (or effectively define) protein function. Moreover, protein interactions can be described more precisely than protein functions. Thus, rather than directly dealing with the general relationship between protein function and expression, we look here at a subproblem: the relationship between mRNA expression and protein-protein interactions, especially those in protein complexes. A priori it seems reasonable that there should be a well-defined relationship between the expression levels of the subunits in a complex: because the functionality of many complexes hinges on the presence of all the subunits, a haphazard and independent expression of any one subunit would be energetically costly. For instance, the components of the ribosome are regulated in a complex way but there is usually agreement that they should be present in equimolar amounts, although this has not yet been measured directly (Woolford et al. 1991; Planta et al. 1997; Li et al. 1999; Nomura 1999).

We investigate this relationship for many of the known protein complexes in a comprehensive, global fashion by interrelating many of the yeast data sets for protein interactions and expression. The diversity and number of yeast experiments provide high-quality data under varied conditions. Additionally, we investigate the relationship between other types of protein-protein interactions (e.g., aggregated physical and genetic interactions) and mRNA expression. Our work follows up on many recent analyses of protein-protein interactions (Fellenberg et al. 2000; Hishigaki et al. 2001; Teichmann et al. 2001; Walhout and Vidal 2001).

In general, our goal was to integrate and cross-correlate already existing data from different sources and find general trends in it. This is an exploratory study prior to any type of prediction. In a sense, this study can be understood as an exploration of the knowledge already implicit in the current data but not yet obvious because, previously, it has not yet been integrated and put together in this way.

## RESULTS

In our survey of existing data, we used two different approaches to analyze the two different types of expression data available: the computation of normalized differences for *absolute* expression levels and a more standard analysis of the correlation of profiles of *relative* expression levels (expression

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-MAIL Mark.Gerstein@Yale.edu ; FAX (360) 838-7861.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.205602.

ratios). We explain these two approaches in more detail in the following two sections.

## Calculation of Normalized Differences between Absolute Expression Levels

To compare absolute mRNA expression levels between subunits of a protein complex, we define the normalized difference $D_{ij}$ as follows:

$$D_{ij} = \frac{|E_i - E_j|}{E_i + E_j}$$

where $E_i$ and $E_j$ are the mRNA expression levels of subunits $i$ and $j$. This quantity defines the difference as a fraction of the sum of the expression levels, thus allowing for a comparison of gene pairs of both high and low expression. Values for the normalized difference range from 0 to 1.

For a group of $N$ proteins in a complex, we generally compute the normalized difference not only for the pairs that are in direct physical contact, but for all $(N^2 - N)/2$ theoretically possible pairs, thus arriving at a distribution of normalized differences of these pairs for each complex. We can then investigate this distribution of normalized differences and compare it with those among randomly chosen proteins. In our following discussion, we often refer to the median of the $(N^2 - N)/2$ protein pairs as a key summarizing statistic.

In general, we assume stoichiometric ratios of 1:1 between subunits, although the above equation could be adjusted to account for other ratios. But even then, as shown in the Methods section below, we would not expect this quantity to always be close to zero because of the relationship between mRNA and protein abundance and also the noise in the expression data.

It should also be noted that there are obviously many limitations in treating GeneChip and SAGE data, our input (see Methods), as absolute measurements of mRNA expression (Schadt et al. 2000).

To judge the statistical significance of normalized differences for particular groups of proteins, we compare them to the control distribution of randomly chosen protein pairs (Fig. 1). An interesting theoretical aspect in this context is that if $E_i$ and $E_j$ are random variables with an exponential distribution (which is a close approximation to the actual distribution of expression of levels in the reference expression set), then $D_{ij}$ is distributed uniformly between 0 and 1 (Pitman 1993). This explains why we can observe a nearly uniform distribution of normalized differences for randomly selected pairs of proteins (Fig. 1).

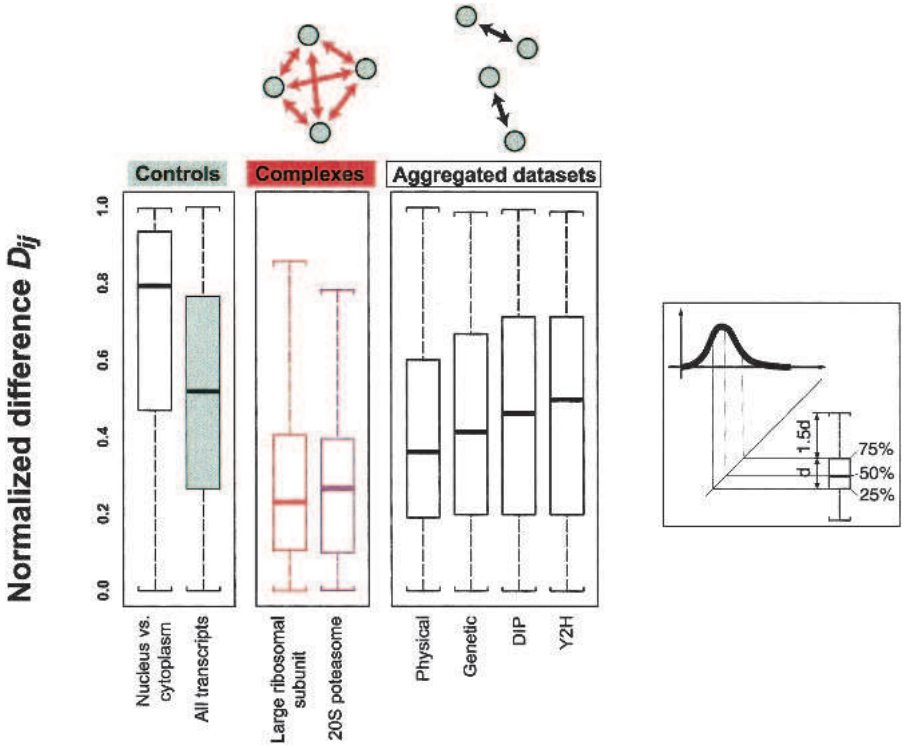


**Figure 1** Distributions of normalized differences for various groups of proteins in boxplot representation. The normalized difference $D_{ij}$ is a measure of the relative similarity of two absolute gene expression levels $E_i$ and $E_j$. The *middle* panel shows the distribution for two protein complexes (the large ribosomal subunit and the 20S proteasome). Note that we considered all theoretically possible protein pairs within the protein complex (as indicated in the schematic drawing above the panel). The *right* panel shows the distribution for the aggregated data sets of protein-protein interactions (Y2H is yeast two-hybrid) (Bader and Hogue 2000; Cagney et al. 2000; Fellenberg et al. 2000; Ito et al. 2000; Schwikowski et al. 2000; Uetz et al. 2000; Uetz and Hughes 2000; Xenarios 2000; Ito et al. 2001). Unlike in the complexes, where we consider interactions among a whole group of proteins, the interactions in the aggregated data sets are specific to individual protein pairs (see schematic drawing). The *left* panel shows two control distributions of the normalized difference, on the left for pairs of nuclear and cytoplasmic proteins, which presumably, because of spatial separation, do not interact, and on the right for any random protein pair ("all transcripts") in yeast. The distribution of nuclear versus cytoplasmic proteins is strongly skewed toward one (the maximum value of the normalized difference), which is partially explained by the fact that cytoplasmic proteins tend to have higher expression levels than cytoplasmic ones (Drawid 2000; Drawid and Gerstein 2000). The distribution of all transcripts is nearly uniform (with a median of 0.5) (see Methods). The complexes distributions are clearly skewed toward zero with medians between 0.2 and 0.3. The medians of the distributions of the aggregated data sets are still somewhat smaller than the control median, most notably for the physical interactions data set; on the other hand, there is virtually no difference between the control and the distribution of the yeast two-hybrid data set.

The aggregated data, obviously, includes some interactions implied by the complexes, with the degree of intersection ranging from 35% for the physical interactions to ~6% for Y2H.

## Correlation of Expression Profiles for Relative Expression Levels

Analysis of expression profiles may be more useful than that of absolute levels for characterizing interacting proteins that exist in unequal but stoichiometrically related amounts (e.g., 3:1) as it refers to the relative shape of expression profiles. It can be carried out on data from cDNA microarrays (such as the Rosetta data) because only relative rather than absolute expression levels are necessary. Specifically, we look at the distribution of Pearson correlation coefficients for pairs of

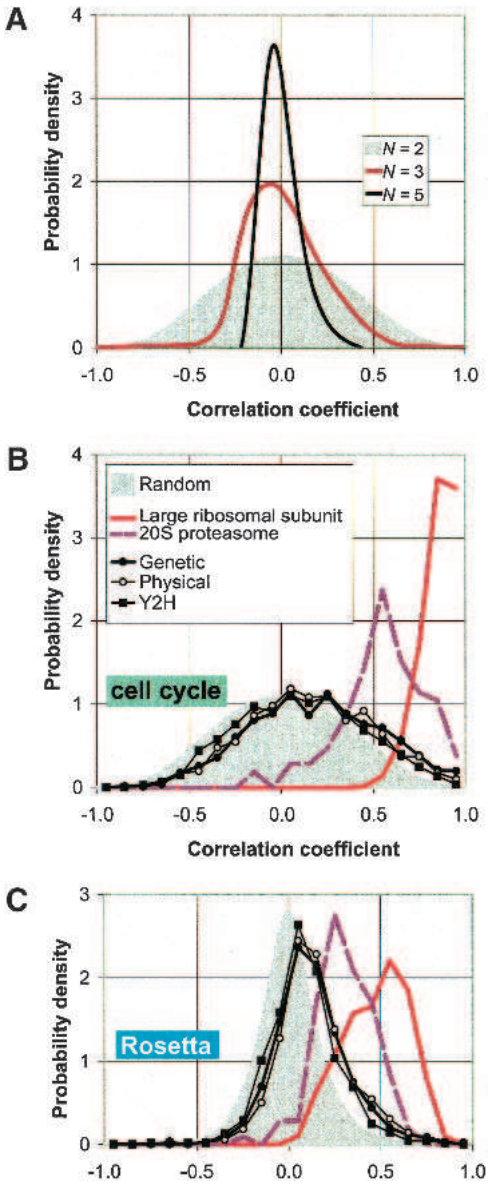


**Figure 2** Distributions of correlation coefficients between expression profiles. In *A*, we show distributions of the average correlation $\bar{\rho}_N$ of $N$ genes for the cell cycle experiments. The gray curve in the background represents the case $N = 2$ (i.e., simply the distribution of pair-wise correlations). In the case of $N > 2$, $\bar{\rho}_N$ is defined as the average of all possible $(N^2 - N)/2$ pairwise correlations among the $N$ genes. We show here, as examples, the distributions for $N = 3$ and $N = 5$. The distributions obviously become narrower, reflecting the fact that it becomes more unlikely to find large groups of strongly correlated genes at random as $N$ increases.

These distributions provide a suitable control for the observed correlations between pairs of genes ($N = 2$) or for the average correlations among the subunits of a complex ($N > 2$).

We have developed a method to efficiently sample the distribution curves $f(\rho_N)$ (see Methods). Based on the distribution function of $f(\rho_N)$ we can calculate a one-sided *P*-value:

$$P(\bar{\rho}_N) = \int_{\bar{\rho}_N}^{1} f(\rho_N) d\rho_N$$

This *P*-value then represents the chance that a group of $N$ randomly selected genes could exhibit an average correlation greater than or equal to that of a complex with $N$ proteins (see Fig. 3). (*B* and *C*) The distribution of pair-wise correlations for both the cell cycle (Cho et al. 1998) and the Rosetta experiments (Hughes et al. 2000) in two protein complexes (the ribosome and the proteasome) as well as for the aggregated data sets (genetic, physical and yeast two-hybrid). The gray curves in the background are the control distributions for $N = 2$ as explained above. The distributions for the ribosome and the proteasome are strongly shifted to the right of the control; this effect is much weaker for the data sets of aggregated interactions.

genes as the measure of similarity. Other measures of similarity are possible as well (D'haeseleer 1997; Wen et al. 1998; Heyer et al. 1999; Qian et al. 2001).

As the input for our procedure, we use the expression vectors or profiles of all the subunits of a complex and then compute their pair-wise correlations. Like for the normalized difference, we compute the correlation coefficients for all protein pairs in a complex, thus gaining a distribution of correlation coefficients. If the complex consists of $N$ subunits, this yields $(N^2 - N)/2$ different combinations of protein pairs and thus correlation coefficients. To summarize these distributions, we calculate the "average correlation" (meaning the average of all pair-wise correlations within a complex). As a suitable control to assess statistical significance, we use the distributions of correlation coefficients for random groups of proteins and their averages (see Methods). We would expect correlations of close to one for subunits in a tight complex. However, as we show in the Methods section, this will not be exactly the case because of the relationship between mRNA and protein abundances.

## Specific Complexes

We first outline some results obtained for specific protein complexes, then we proceed to a more general overview of complexes.

### *Ribosome*

It has long been known that the mRNA expression levels of the ribosomal proteins are strongly correlated with one another (Johannes et al. 1999). Figure 1 shows the observed distribution of normalized differences for protein pairs in the large subunit of the cytoplasmic ribosome. The median of this distribution is 0.23, much lower than the median of 0.5 for randomly selected protein pairs. While there is a wide range of normalized differences (which may partially result from the fact that many proteins in the ribosome are known not to be expressed in a 1:1 ratio [Kruiswijk et al. 1978]), the ribosomal distribution is clearly skewed toward zero. Considering pairs of duplicated ORFs in the ribosome as one subunit by summing their expression levels leads to an even lower median normalized difference of 0.20 (data not shown). Distributions of the correlation coefficients for protein pairs within the large ribosomal subunit are shown in Figure 2. For both the cell cycle and the Rosetta data, the correlations tend to be much higher than the random control.

Similar observations can be made for the proteins in the small cytoplasmic ribosome. Key statistics are summarized in Figure 3 in comparison to those for other protein complexes. Furthermore, the two separate ribosome particles are strongly coregulated. In fact, the large and the small ribosomal particles cannot be differentiated by our measures of expression similarity.

### *Proteasome*

A second example of a complex whose individual subunits are strongly coregulated is the proteasome, which is involved in protein degradation and responsible for the rapid breakdown of ubiquitinated proteins. Like the ribosome, the 26S proteasome can be divided into two subparticles: the 20S and the 19S (or 19S/22S regulatory particle). The 20S particle is present as a dimer in the center of the complex structure and contains the catalytic core, whereas two 19S particles are attached to both ends of the 20S particle dimer (Coux et al. 1996; Wilkinson et al. 1999).

The distribution of the normalized differences for all possible protein pairs in the 20S proteasome is shown in Figure 1. Like the ribosome, it is clearly skewed toward zero, compared to the control, with a median of 0.29. Figure 2 shows the distribution of correlation coefficients, which is strongly shifted to the right of the control, though to a lesser extent than that for the ribosome. An investigation of the crystal structure of 20S particle (Whitby et al. 2000) did not reveal any relationship with the gene expression differences (e.g., proteins with slightly more random correlations tending to be more on the surface of the particle).

Similar results can be observed for the 19S particle of the proteasome (Fig. 3A). Also, in terms of both measures of coexpression (normalized differences and correlation of expression profiles), the 19S and the 20S particles of the proteasome form a single unit that is difficult to separate. Part of the reason for this may be that the common classification into 19S and 20S particles is based on the purification procedure for the proteasome (M. Hochstrasser, pers. comm.) and thus does not necessarily reflect functional or biochemical properties in a direct way.

One subunit, Doa4p, exhibits a very low average correlation (−0.02). Biochemical studies have previously shown that not all proteasomes have Doa4p bound and that the Doa4p-proteasome interaction is more likely to be transitory (Papa and Hochstrasser 1993; Papa et al. 1999).

### *RNA Polymerase II Holoenzyme*

We have seen above that the ribosome and proteasome can be regarded as strongly associated and coregulated multiparticle complexes. However, in some cases a complex contains more loosely associated components. An example is the RNA polymerase II holoenzyme, which contains the core RNA polymerase II together with the more loosely associated SRB complex (Kornberg's mediator) and other smaller components (such as the SWIF/SNF complex and the TAFIIs).

It is known that, unlike the RNA polymerase II core enzyme, the SRB complex and the other holoenzyme components are only needed for the transcription of a fraction of genes (Holstege et al. 1998). In other words, the holoenzyme is an example of a complex of transitory nature with a permanent core. This permanent-and-transitory structure is clearly evident in the gene expression analysis. For the core enzyme, the average correlation in both the cell cycle and Rosetta data sets are significantly higher than for the random control (Fig. 3). However, for the SRB complex and a variety of other, smaller components (e.g., the TAFIIs) the average correlations are virtually indistinguishable from the random control.

### *Replication Complex*

Another example of a transient complex is the replication complex, which binds to DNA and is needed for the initiation of replication. The replication complex can be subdivided into a number of subcomponents: the MCM proteins, the

A

| Class | Complex | Ave. mRNA exp. [copies/cell] | Median norm. diff. [Dij] | ρ (cell cycle) | ρ (Rosetta) | -log(P) (cell cycle) | -log(P) (Rosetta) | Complex size [# ORFs] |
|---|---|---|---|---|---|---|---|---|
| Permanent | *Cytoplasmic ribosome* | 38.6 | 0.27 | 0.80 | 0.41 | | | 139 |
| | Cytoplasmic ribosomal small subunit | 38.8 | 0.27 | 0.80 | 0.41 | | | 57 |
| | Cytoplasmic ribosomal large subunit | 38.3 | 0.22 | 0.81 | 0.43 | | | 81 |
| | *Mitochondrial ribosome* | 1.3 | 0.25 | 0.49 | 0.35 | | | 49 |
| | Mitochondrial ribosomal small subunit | 1.3 | 0.26 | 0.42 | 0.28 | | | 14 |
| | Mitochondrial ribosomal large subunit | 1.3 | 0.25 | 0.51 | 0.39 | | | 32 |
| | *26S proteasome* | 4.4 | 0.29 | 0.43 | 0.22 | | | 36 |
| | 19S proteasome | 4.1 | 0.23 | 0.51 | 0.29 | | | 18 |
| | 20S proteasome | 5.5 | 0.27 | 0.50 | 0.25 | | | 15 |
| | Cytochrome bc1 complex | 3.6 | 0.35 | 0.54 | 0.58 | 3.7 | | 10 |
| | F0/F1 ATP synthase | 3.9 | 0.21 | 0.64 | 0.48 | | | 18 |
| | Succinate dehydrogenase complex | 1.2 | 0.28 | 0.68 | 0.42 | 2.9 | 3.7 | 4 |
| | Cytochrome c oxidase | 4.4 | 0.14 | 0.76 | 0.38 | | | 11 |
| | RNA polymerase I | 5.0 | 0.27 | 0.39 | 0.30 | | | 14 |
| | RNA polymerase III | 4.5 | 0.40 | 0.27 | 0.32 | 2.9 | | 13 |
| | *RNA polymerase II holoenzyme* | 4.4 | 0.29 | 0.43 | 0.22 | 1.0 | | 36 |
| | RNA polymerase II | 4.2 | 0.35 | 0.26 | 0.18 | 2.7 | | 13 |
| Transient | Kornberg's mediator (SRB) complex | 0.9 | 0.37 | -0.01 | 0.04 | 0.1 | 1.3 | 21 |
| | SAGA complex | 1.0 | 0.29 | 0.12 | 0.05 | 1.2 | 1.0 | 13 |
| | CCR4 complex | 1.3 | 0.38 | 0.10 | 0.00 | 0.9 | 0.2 | 13 |
| | SWI/SNF transcription activator complex | 0.8 | 0.33 | 0.29 | 0.07 | 2.2 | 1.3 | 11 |
| | TAFIIs | 1.2 | 0.29 | 0.07 | 0.05 | 0.6 | 1.1 | 12 |
| | RSC complex | 0.8 | 0.27 | 0.28 | 0.16 | 2.1 | 3.0 | 10 |
| | Pre-replication complex | 0.6 | 0.26 | 0.16 | 0.14 | 1.7 | | 14 |
| | Replication complex | 0.6 | 0.23 | 0.05 | 0.11 | 0.7 | | 19 |
| | Anaphase promoting complex (APC) | 0.6 | 0.37 | 0.02 | 0.05 | 0.2 | 0.9 | 12 |
| Other | H+-transporting ATPase, vacuolar | 9.2 | 0.36 | 0.63 | 0.12 | | 3.0 | 15 |
| | TRAPP (Transport Protein Particle) complex | 1.2 | 0.35 | 0.07 | 0.11 | 0.5 | 1.9 | 10 |
| | Nuclear pore complex (NPC) | 1.3 | 0.17 | 0.14 | 0.09 | 3.0 | | 29 |

**Figure 3** (See facing page for legend.)

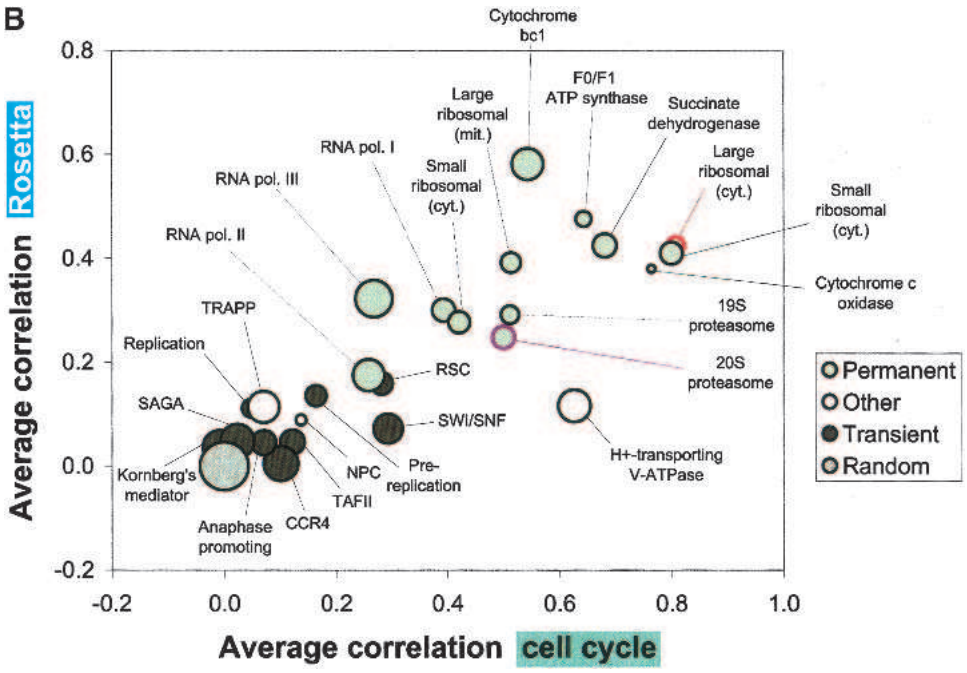


**Figure 3** (*A*) Consolidates various key statistics shown in Figures 1 and 2 for the ribosome and proteasome as well as for a large number of protein complexes. We list all protein complexes from the MIPS catalog having at least 10 open reading frames (ORFs). The complexes are divided into three classes: permanent, transient, or other (see below). Some complexes can be divided into smaller subcomplexes (e.g., the ribosomes) as indicated. The table lists (from left to right) the average expression level of the complex, the median normalized difference (see Fig. 1A), the average correlation for the cell cycle and Rosetta experiments (see Fig. 2), the negative logarithm of the *P*-value of the average correlations in both experiments (see Fig. 2), and the size of the complex in terms of the number of ORFs.

In general, the *P*-values for the average correlations are very low for most of the permanent protein complexes [accordingly, $-\log_{10}(P)$ is very high], indicating that these averages are significantly greater than for random groups of proteins of the same size. The same cannot be observed for the transient protein complexes, for which the correlation averages are usually much smaller.

The section "other" at the bottom of *A* contains complexes that are either difficult to classify as permanent/transient or for which, as a result of very small turnover rates, down-regulations of mRNA levels take a very long time to affect protein abundance. The H+-transporting ATPase can be thought of as containing a mixture of permanent and transient components at the same time (P. Kane, pers. comm.). The nuclear pore complex (NPC) and the TRAPP complex are known to have low turnover rates (Bucci and Wente 1997; Winey et al. 1997; Sacher et al. 1998; Barrowman et al. 2000). The NPC has relatively small average correlations, but this still yields *P*-values of $10^{-3}$ (cell cycle) and $<10^{-4}$ (Rosetta) because the nuclear pore complex is a relatively large aggregation of proteins, and even these weak average correlations are very unlikely to occur for random groups of proteins of this size. The TRAPP protein complex, while existing throughout the cell cycle, has a low turnover rate and as such its mRNA expression data would not be sufficient for our analysis.

The RNA polymerase holoenzyme is composed of both permanent and transient components. Note that the MIPS complexes catalog does not include the SWI/SNF chromatin-remodeling complex and a subset of basal transcription factors (Wilson et al. 1996) as part of the holoenzyme, thus we list them separately here.

The list does not include those categories from the MIPS complexes catalog that do not really represent protein complexes per se, but rather aggregations of disparate proteins that are involved in similar types of complex interactions, such as the "actin-associated" and "tubulin-associated" protein groups.

(*B*) Shows a graphical representation of part of the protein complex statistics from *A*. The abscissa and ordinate represent the average correlations in the cell cycle and the Rosetta data, while the bubble sizes are a function of the normalized differences (larger bubbles represent larger normalized differences). In general, the permanent complexes tend to be located in the upper right region of the plot, whereas transient complexes are closer to the random control in the lower left.

origin recognition complex, and the DNA polymerases δ and ε (Aparicio et al. 1997).

As a whole, the replication complex exhibits a low average correlation not significantly different from that of the random control (Fig. 3, 4). However, Figure 4 shows how the entire complex breaks into subcomponents in terms of correlations in the cell-cycle experiment. The individual correlations for each of the subcomponents are much higher than that of the complex as a whole. This indicates that the replication complex is composed of independent units in terms of expression regulation. Using the permanent-transient terminology, each subcomponent behaves similarly to an independent permanent complex, whereas the replication complex as a whole can be characterized as transient. The permanent subcomponents can be seen to come together to form a transient functional entity. (Note, this effect is more evident in the cell cycle experiment than the Rosetta data, as it should only be observable in a synchronized population of cells, not those averaged across the cell cycle.)

## Complexes in General: Permanent versus Transient

In discussing the specific examples above, we have found the permanent or transient nature of the association to be an important feature. This distinction is, in fact, valuable in a more general context. As shown in Figure 3, we have a priori formalized a division between "permanent" complexes, which are maintained throughout the cell cycle and most

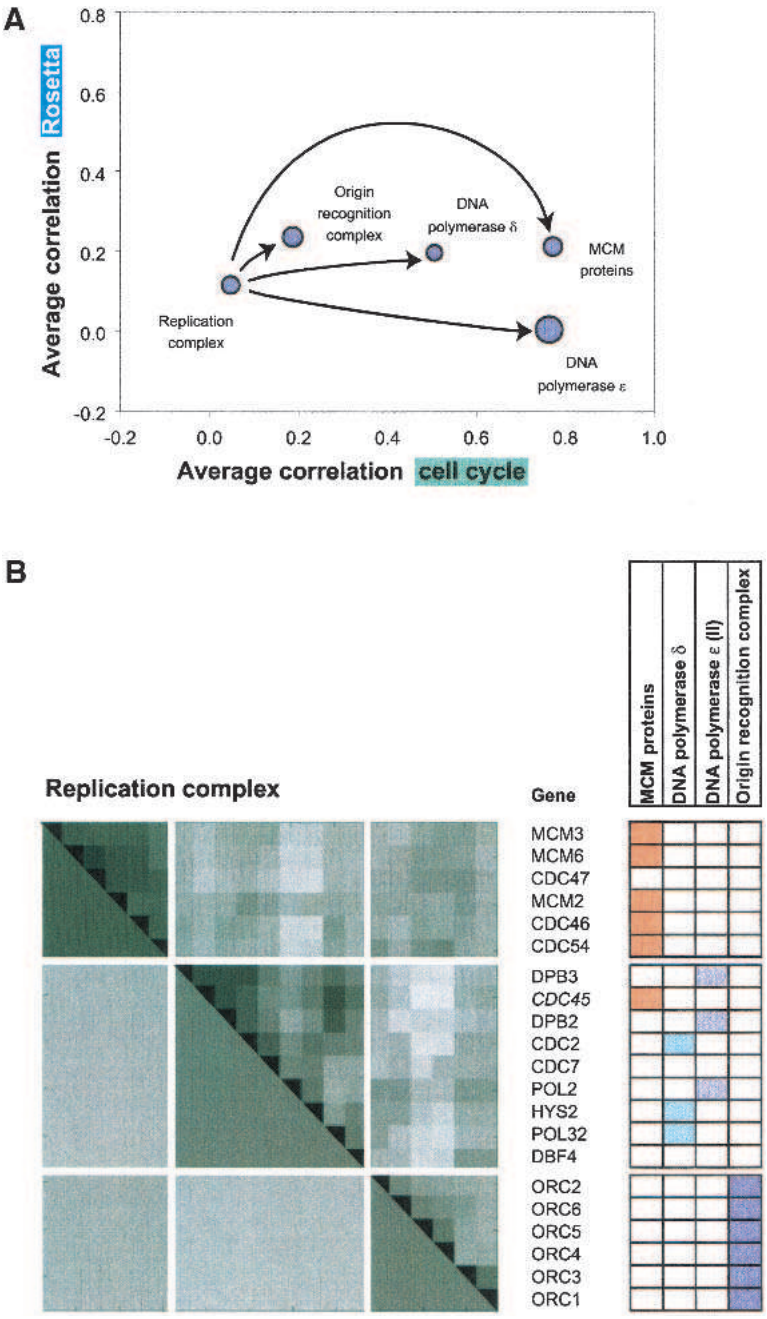


**Figure 4** (*A*) A representation of the replication complex and its components on the same coordinates as the protein complexes in Figure 3B. The transient replication complex can be decomposed into smaller complexes: the origin recognition complex, the MCM proteins, and the DNA polymerases δ and ε. Whereas the whole replication complex exhibits an average correlation close to zero (in both the cell cycle and the Rosetta data), the four smaller complexes show greater correlations in the cell cycle experiment. The four subcomplexes behave more like permanent complexes than the replication complex as a whole.

(*B*) The correlation coefficient matrix for the subunits of the replication complex derived from the cell cycle data. The upper triangle of the correlation matrix shows the individual correlation coefficients for particular gene pairs (with darker colors indicating higher correlations). The lower triangle shows the average correlations for subgroups of proteins (representing the MCM proteins, the two DNA polymerases, and the origin of the replication complex) within the complex as a whole. The table on the right side shows which genes belong to which subgroups in different colors. The genes were ordered with unsupervised clustering (average linkage) without regard to their classification according to the three subgroups. It can be seen that this order reflects the separation according to the subgroups very well (only the proteins in the two DNA polymerase cannot be separated into two groups). An exception is the CDC45 protein that belongs to the MCM proteins but tends to cluster with the DNA polymerases.

cellular conditions, and "transient" ones, which we define here as a group of proteins that do not consistently maintain their interactions. That is, the existence of a transient complex is temporal and specific to a part of the cell cycle or a subset of cellular states. We are aware that the division into the two absolute categories "permanent" and "transient" is perhaps somewhat oversimplifying as there can be varying degrees and combinations of these attributes (see Discussion).

In Figure 3, we show a general classification of the large MIPS complexes into permanent and transient classes, together with key statistics (details of the classification method are given in the legend). We list all complexes with more than 10 subunits (which together account for ~80% of all the theoretically possible protein-protein interactions in the MIPS complexes), with smaller complexes listed on our website. Figure 3B shows a graphical representation of the complex list, synthesizing the correlations for both the Rosetta and cell-cycle experiments with the normalized differences. It clearly shows that there is a greater tendency for permanent complexes to have higher average correlations than for transient ones.

Comparing the average correlations in Figure 3A against random controls allows us to derive *P*-values for the statistical significance of the correlation. As shown in the figure, these are $<10^{-4}$ for most of the permanent complexes. On the other hand, they are considerably higher, and thus less significant,

for transient complexes. The separation between permanent and transient complexes is also evident in terms of the normalized difference statistics, although not as strongly.

### Aggregated Protein-Protein Interaction Sets

From our analysis above, it seems reasonable to conclude that there is indeed a strong relationship between mRNA expression and the protein-protein interactions in "permanent" complexes. This raises the question whether or not similar observations can be made for other types of protein-protein interactions. We briefly summarize here the degree to which the interactions in the aggregated interaction data sets, such as the yeast two-hybrid data, are related to expression.

Figure 1 shows the distribution of normalized differences, and Figure 2 shows the distributions of correlation coefficients between interacting proteins in the aggregated data sets. The distributions of normalized differences are relatively similar to those of the transient protein complexes. The physical interactions show the smallest median normalized difference while the yeast two-hybrid interactions have a median normalized difference closest to the random control (~0.5). Figure 2 shows that the correlation distributions for the aggregated data sets are fairly similar among themselves and only slightly shifted toward the right of the distribution curve for random protein pairs. This, again, is very similar to the behavior of transient protein complexes.

Thus, overall, it seems fair to conclude that the aggregated protein-protein interactions are related to mRNA expression in a similar fashion as the transient protein complexes.

## DISCUSSION AND CONCLUSION

We have investigated the relationship of protein-protein interactions and mRNA expression levels, integrating and surveying a variety of data sources for yeast. We have focused our investigation on the protein interactions within specific complexes. While we have demonstrated a strong relationship between expression data and most permanent protein complexes, this relationship is much weaker for transient protein complexes as well as for the aggregated sets of protein-protein interactions (i.e., physical, genetic, and yeast two-hybrid interactions).

### Issues with Permanent-Transient Classification

Our complex classification scheme—separating most complexes into either permanent or transient—while useful, cannot account for all complexes in the MIPS database. Some complexes may not clearly fit into the permanent-transient classification. We list a few of these as "other" in Figure 3. Moreover, the complexes list is a compilation of current biochemical knowledge and therefore reflects its inherent limitations (sometimes not all subunits are known or some proteins are mistakenly assigned to a complex).

Of course, even for the complexes that we do classify, the terms "transient" and "permanent" are somewhat of an oversimplification. In particular, our detailed discussions of the RNA polymerase II holoenzyme and the replication complex above are precisely two examples where our simplified terminology fails to completely explain the situation since these complexes are somewhere between fully "transient" and "permanent".

One can think about the distinction between permanent and transient in terms of the mathematical model introduced in the Methods section. Whenever a complex is formed, its subunits tend to be expressed at equimolar protein concentrations: $P_i \approx P_j$ and $dP_i/dt \approx dP_j/dt$ (where $P_i$ and $P_j$ are the protein concentrations of two subunits $i$ and $j$). If the complex is "permanent", then these conditions should be approximately or vaguely met. If the complex is "transient", then these conditions can be relaxed in those situations where the complex is not formed. There are some complexes that are always formed ("permanent") whereas the "transient" complexes are only formed under particular conditions. There can be different degrees of being transient: for instance, complexes that are formed under 80% of conditions or those that are formed under 20% of conditions. The transient complex formed under 80% of conditions behaves almost like "permanent" (i.e., 100% of conditions), whereas the transient complex formed only 20% of the time would be expected to show less significant normalized differences and correlations.

If one goes as far as to accept the premise that the subunits in a complex should be present at equimolar amounts, then it is perhaps circular reasoning to say that they should also be coexpressed.

### Complexes versus the Aggregated Interactions: The Need for Structures

We found it difficult to discern expression-based relationships in the aggregated data sets. This may be because of the generalized and heterogeneous nature of the aggregated data sets, (e.g., inconsistent physiological conditions, false positives, and false negatives). Moreover, both the aggregated sets and the transient complexes suffer partially from the limited amount of mRNA expression data as their interactions may occur under particular physiological conditions that may not be sampled by mRNA expression data. Our results, thus, illustrate the difficulty in drawing general conclusions for the pair-wise interaction sets and highlight the important role clearly resolved crystal structures of complexes, detailing protein interactions between subunits, have in studying protein-protein interactions.

### Noise in the Expression and Interaction Data

In general, the interactions in the aggregated data sets exhibited surprisingly little deviation from randomness in terms of the coexpression of interaction pairs. This was most strongly observed for the yeast two-hybrid data. It is true that, overall, this deviation from randomness is statistically significant. All the same, the gene expression data and the aggregated protein interaction data do not reinforce each other strongly and it seems that the prediction of these type of interactions from expression data would be of little benefit.

Perhaps the most optimistic view of this situation is that the strong degree of independence of the two types of data makes both of them suitable for use in machine-learning approaches to characterize genes of unknown function: if they were strongly correlated, then one type of data could perhaps well replace the other since it represents very similar information. A negative view would be that the reason for the surprisingly weak relationship between the aggregated interactions and mRNA expression are to be found in the problems with the either the expression or the interaction data.

We feel confident that our results are robust to the noise in the expression data for the following reasons. With respect to the correlation analysis of expression profiles, roughly the same results (in terms of statistical significances) can be ob-

tained for two independent data sets (the cell-cycle time course and the Rosetta knockout series). The normalized difference analysis is perhaps more sensitive to problems with the data, in particular, considering that the measurement of absolute expression levels with gene chips is problematic to start with. However, we have looked at an integrated data set from various chip experiments and the SAGE data, thus averaging out errors to some degree (see Methods). In addition, for both the correlation and the normalized difference analysis, we have concentrated on the statistical significance of distributions rather than relying on the error-prone data for individual protein pairs, thus observing more robust, aggregate trends for whole complexes and groups of proteins.

Part of the aggregated data, in particular the yeast two-hybrid data, represent a relatively new approach to studying protein-protein interactions and it is interesting to note that it, obviously, includes some interactions implied by the complexes. However, the degree of intersection with possible complexes interactions ranges from 35% for the physical interactions to only ~6% for the yeast two-hybrid data (as a fraction of the number of interactions in the aggregated datasets). This is surprisingly low, given that the yeast two-hybrid data is from experiments that covered the complete genome (Uetz et al. 2000; Ito et al. 2001). Independently, Ito et al. (2001) reported that only a small fraction of the previous yeast two-hybrid data (Uetz et al. 2000) overlapped with their own yeast two-hybrid results. (Although Ito and colleagues assumed that their core data was similar in quality as the Uetz data, the fraction of interactions present in both data sets was only 16.8% for the Ito core and 20.4% for the Uetz data).

### mRNA versus Protein Expression

The coregulation of subunits in a protein complex should be primarily observable in terms of protein abundance and only indirectly in terms of mRNA expression. Several recent studies have attempted to investigate the relationship between mRNA and protein expression levels in yeast cells and found them to be correlated to various degrees (Anderson and Seilhamer 1997; Futcher et al. 1999; Gygi et al. 1999; Greenbaum et al. 2001; Lian et al. 2001). Generally, post-transcriptional regulation is more difficult to investigate given the sparse data resources currently available for protein abundance levels. It is possible that in some situations coregulation occurs mostly on the protein level, almost independent of cellular mRNA levels. Particularly, those permanent complexes that do not have high levels of correlation in our analysis may be indicative of translational or post-translational control and could be a starting point for further experimental investigation. See Methods section for further discussion.

## METHODS

### Interactions Data Sources

The primary focus of this paper is the interactions occurring within specific complexes. These were obtained from the MIPS complexes catalog (Fellenberg et al. 2000), which represents a carefully annotated, comprehensive data set of protein complexes culled from the scientific literature. In addition, we looked at other types of protein-protein interactions from large "aggregated" data sets collecting many heterogeneous pair-wise interactions. We collected these from the MIPS catalogs of physical and genetic interactions (Fellenberg et al. 2000), databases of interacting proteins (DIP and BIND) (Bader and Hogue 2000; Xenarios 2000), and a comprehensive collection of yeast two-hybrid experiments (Cagney et al. 2000; Ito et al. 2000; Schwikowski et al. 2000; Uetz et al. 2000; Uetz and Hughes 2000; Ito et al. 2001). These interactions are subdivided into groups based on their method of discovery. They include physical interactions (e.g., collected through co-immunoprecipitation and copurification), genetic interactions (e.g., determined through genetic means such as synthetic lethality or suppression experiments), and yeast two-hybrid pairs.

### Expression Data Sources

We included two different types of expression measurements in our analysis: absolute expression levels in vegetative yeast cells as determined by SAGE or gene chip experiments, and profiles of ratio-type expression data from microarray experiments. For the first type, we use a comprehensive reference set, which we merged and scaled together from a variety of Affymetrix GeneChip and SAGE data sets (Velculescu 1997; Holstege 1998; Roth et al. 1998; Jelinsky and Samson 1999) into a single representative data source (scaling details on our website; Greenbaum et al. 2001). For the expression profiles, we focused on two different data sets: a cell cycle experiment (Cho et al. 1998) and the Rosetta yeast compendium (Hughes et al. 2000). The two data sets provide a fairly good sampling of the possible cellular states of yeast and represent different experimental methodologies. The cell-cycle data contains expression profiles obtained from synchronized cells over the course of two cell cycles, whereas the Rosetta data contains genome-wide expression ratios for 300 stationary cell states, which are derived from 280 gene deletions and the 20 drug interaction experiments.

### Efficient Calculation of the Average Correlations

For two expression ratio profiles $\mathbf{X_i}$ and $\mathbf{X_j}$ (transformed to average 0 and standard deviation 1), the Pearson correlation coefficient $\rho_{ij}$ is given by the dot product:

$$\rho_{ij} = \frac{1}{M-1}\mathbf{X_i} \cdot \mathbf{X_j},$$

where $M$ is the number of elements in the profiles $\mathbf{X_i}$ and $\mathbf{X_j}$. The profile $\mathbf{X}$ can be computed as a 'Z-score' from the measured expression ratio profile $\mathbf{x}$, through the relation $X_k = (x_k - \bar{x})/\sigma_x$, where $\bar{x}$ denotes the average and $\sigma_x$ the standard deviation of values in $\mathbf{x}$, and $X_k$ and $x_k$ are the kth components of their respective profiles.

Given a group of $N$ genes, we can compute the correlation coefficient matrix $\mathbf{R}$, where each element $\rho_{ij}$ of the matrix denotes the Pearson correlation coefficient between genes $i$ and $j$. We can then compute the average correlation coefficient $\bar{\rho}$ by averaging the matrix elements (excluding the main diagonal). This statistic gives an idea of the overall similarity of the expression profiles in a group of genes. Although there are $O(N^2)$ elements in $\mathbf{R}$, the computation time for $\bar{\rho}$ can be kept proportional to $O(N)$ by using the linearity of the correlation to calculate $\bar{\rho}$ as follows:

$$\bar{\rho} = \frac{1}{N^2 - N}\left(\sum_{i,j}^{N} R_{ij} - N\right) = \frac{1}{N^2 - N}\left(\frac{1}{M-1}\mathbf{X_T} \cdot \mathbf{X_T} - N\right),$$

where

$$\mathbf{X_T} = \sum_{n=1}^{N} \mathbf{X}_n$$

is the sum of all expression profiles in the group of $N$ genes.

### Kinetic Model of the Relationship between Protein and mRNA Concentration

For a protein complex that is perfectly coregulated we can

assume that its components are present at equimolar amounts and change similarly over time. So for the protein concentrations $P_i$ and $P_j$ of two different subunits $i$ and $j$ we would get: $P_i \approx P_j$ and $dP_i/dt \approx dP_j/dt$. Using a simple model for the relationship between mRNA and protein concentrations, we can see how even under these ideal conditions similarity measures based on the mRNA concentrations would deviate from perfect results. For instance, a linear kinetic model for the protein concentration $P_i$ and the mRNA concentration $R_i$ of a subunit $i$ in a complex is given by:

$$\frac{dP_i}{dt} = k_{Ri}R_i - k_{Pi}P_i$$

where $k_{Ri}$ is an mRNA translation rate constant and $k_{Pi}$ is a protein degradation constant.

## Why Expression Profile Correlations Have to Be Less Than One

For two subunits in a complex with $P_i = P_j \equiv P$ and $dP_i/dt = dP_j/dt$, we can deduce:

$$k_{Ri}R_i(t) = k_{Rj}R_j(t) + (k_{Pi} - k_{Pj})P(t)$$

It is clear that only under the strong assumption that the two protein degradation constants are equal ($k_{Pi} = k_{Pj}$)

$$\frac{R_i(t)}{R_j(t)} = \frac{k_{Rj}}{k_{Ri}} = const$$

from which would follow $corr(R_i, R_j) = 1$. Otherwise, $corr(R_i,R_j) < 1$.

## Why Normalized Differences Are Greater Than Zero

Furthermore, assuming steady-state (that is, $dP_i/dt = dP_j/dt = 0$), we can deduce the following relationship for the relationship between the mRNA levels of two complex subunits:

$$R_i = \frac{k_{Rj}}{k_{Pj}}\frac{k_{Pi}}{k_{Ri}}R_j$$

Thus, the two mRNA expression levels are only expected to be equal if the ratios of the rate constants for translation and degradation are the same for both proteins. This is not necessarily the case for the subunits of a complex and therefore normalized differences should not be expected to be zero.

It is clear that the arguments above are based on a variety of simplifying assumptions. In reality, there are additional factors (such as the noise in the expression data, the stochastic nature of gene expression) that add even more difficulty to the analysis of mRNA levels.

## Supplementary Information

Additional information can be found at genecensus.org/expression/interactions and bioinfo.mbb.yale.edu/expression/interactions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

# ACKNOWLEDGMENTS

MG acknowledges support by the Keck Foundation. RJ is supported by an IBM PhD Fellowship. The authors wish to thank Mark Hochstrasser and Jiang Qian for stimulating discussions.

# REFERENCES

Anderson, L. and Seilhamer, J. 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18:** 533–537.

Aparicio, O.M., Weinstein, D.M., and Bell, S.P. 1997. Components and dynamics of DNA replication complexes in *S. cerevisiae*: Redistribution of MCM proteins and Cdc45p during S phase. *Cell* **91:** 59–69.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Bader, G.D. and Hogue, C.W. 2000. BIND—A data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16:** 465–477.

Barrowman, J., Sacher, M., and Ferro-Novick, S. 2000. TRAPP stably associates with the Golgi and is required for vesicle docking. *EMBO J.* **19:** 862–869.

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Jr., and Haussler, D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.* **97:** 262–267.

Bucci, M. and Wente, S.R. 1997. In vivo dynamics of nuclear pore complexes in yeast. *J. Cell. Biol.* **136:** 1185–1199.

Cagney, G., Uetz, P., and Fields, S. 2000. High-throughput screening for protein-protein interactions using two-hybrid assay. *Methods Enzymol.* **328:** 3–14.

Califano, A., Stolovitzky, G., and Tu, Y. 2000. Analysis of gene expression microarrays for phenotype classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8:** 75–85.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* **2:** 65–73.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Gerstein, M., Arrowsmith, C.H., and Edwards, A.M. 2000a. Structural proteomics: Prospects for high throughput sample preparation. *Prog. Biophys. Mol. Biol.* **73:** 339–345.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., et al. 2000b. Structural proteomics of an archaeon. *Nat. Struct. Biol.* **7:** 903–909.

Coux, O., Tanaka, K., and Goldberg, A.L. 1996. Structure and functions of the 20S and 26S proteasomes. *Annu. Rev. Biochem.* **65:** 801–847.

D'haeseleer, P., Wen,X., Fuhrman, S., and Somogyi, R. 1997. Mining the gene expression matrix: Inferring gene relationships from large scale gene expression data. In *Information processing in cells and tissues*. (ed. P.M. Holcombe), pp. 203–212. Plenum, New York.

Drawid, A., Jansen, R., and Gerstein, M. 2000. Genome-wide analysis relating expression level with protein subcellular localization. *Trends Genet.* **16:** 426–430.

Drawid, A. and Gerstein, M. 2000. A Bayesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome. *J. Mol. Biol.* **301:** 1059–1075.

Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. 2000. Protein function in the post-genomic era. *Nature* **405:** 823–826.

Emili, A.Q. and Cagney, G. 2000. Large-scale functional analysis using peptide or protein arrays. *Nat. Biotechnol.* **18:** 393–397.

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.

Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S., and Garrels, J.I. 1999. A sampling of the yeast proteome. *Mol. Cell. Biol.* **19:** 7357–7368.

Gaasterland, T. and Bekiranov, S. 2000. Making the most of microarray data. *Nat. Genet.* **24:** 204–206.

Gerstein, M. and Jansen, R. 2000. The current excitement in bioinformatics-analysis of whole-genome expression data: How does it relate to protein structure and function? *Curr. Opin. Struct. Biol.* **10:** 574–584.

Greenbaum, D., Jansen, R., and Gerstein, M. 2002. Analysis of mRNA expression and protein abundance data: An approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics*. (in press).

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286:** 531–537.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. 1999. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19:** 1720–1730.
Heyer, L.J., Kruglyak, S., and Yooseph, S. 1999. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Res.* **9:** 1106–1115.
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.
Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95:** 717–728.
Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102:** 109–126.
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.* **98:** 4569–4574.
Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. 2000. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci.* **97:** 1143–1147.
Jansen, R. and Gerstein, M. 2000. Analysis of the yeast transcriptome with structural and functional categories: Characterizing highly expressed proteins. *Nucleic Acids Res.* **28:** 1481–1488.
Jelinsky, S.A. and Samson, L.D. 1999. Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl. Acad. Sci.* **96:** 1486–1491.
Johannes, G., Carter, M.S., Eisen, M.B., Brown, P.O, and Sarnow, P. 1999. Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proc. Natl. Acad. Sci.* **96:** 13118–13123.
Kruiswijk, T., Planta, R.J., and Mager, W.H. 1978. Quantitative analysis of the protein composition of yeast ribosomes. *Eur. J. Biochem.* **83:** 245–252.
Li, B., Nierras, C.R., and Warner, J.R. 1999. *Mol. Cell. Biol.* **19:** 5393–5404.
Lian, Z., Wang, L., Yamaga, S., Bonds, W., Beazer-Barclay, Y., Kluger, Y., Gerstein, M., Newburger, P.E., Berliner, N., Weissman, S.M. 2001. Genomic and proteomic analysis of the myeloid differentiation program. *Blood* **98:** 513–524.
Luscombe, N.M., Laskowski, R.A., Westhead, D.R., Milburn, D., Jones, S., Karmirantzou, M., and Thornton, J.M. 1998. New tools and resources for analysing protein structures and their interactions. *Acta. Crystallogr. D. Biol. Crystallogr.* **54:** 1132–1138.
Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S., and Weil, B. 2000. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **28:** 37–40.
Nomura, M. 1999. Regulation of ribosome biosynthesis *in Escherichia coli* and *Saccharomyces cerevisiae*: Diversity and common principles. *J. Bacteriol.* **181:** 6857–6864.
Papa, F.R., Amerik, A.Y., and Hochstrasser, M. 1999. Interaction of the Doa4 deubiquitinating enzyme with the yeast 26S proteasome. *Mol. Biol. Cell* **10:** 741–756.
Papa, F.R. and Hochstrasser, M. 1993. The yeast DOA4 gene encodes a deubiquitinating enzyme related to a product of the human tre-2 oncogene. *Nature* **366:** 313–319.
Pitman, J. 1993. *Probability*. Springer-Verlag, New York.
Planta, R.J. 1997. Regulation of ribosome synthesis in yeast. 1997. *Yeast* **13:** 1505–1518.
Qian, J., Dolled-Filhart, M., Lin, J., and Gerstein, M. 2002. Beyond synexpression relationships: Clustering of time shifted and inverted gene expression profiles identifies new biologically relevant interactions. *J. Mol. Biol.* (in press).
Raychaudhuri, S., Sutphin, P.D., Chang, J.T., and Altman, R.B. 2001. Basic microarray analysis: Grouping and feature reduction. *Trends Biotechnol.* **19:** 189–193.
Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16:** 939–945.
Sacher, M., Jiang, Y., Barrowman, J., Scarpa, A., Burston, J., Zhang, L., Schieltz, D., Yates 3rd, J.R., Abeliovich, H., and Ferro-Novick, S. 1998. TRAPP, a highly conserved novel complex on the cis-Golgi that mediates vesicle docking and fusion. *EMBO J.* **17:** 2494–2503.
Schadt, E.E., Li, C., Su, C., and Wong, W.H. 2000. Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* **80:** 192–202.
Schwikowski, B., Uetz, P., and Fields, S. 2000. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18:** 1257–1261.
Subrahmanyam, Y.V., Yamaga, S., Prashar, Y., Lee, H.H., Hoe, N.P., Kluger, Y., Gerstein, M., Goguen, J.D., Newburger, P.E., and Weissman, S.M. 2001. RNA expression patterns change dramatically in human neutrophils exposed to bacteria. *Blood* **97:** 2457–2468.
Teichmann, S.A., Murzin, A.G., and Chothia, C. 2001. Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.* **11:** 354–363.
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.
Uetz, P. and Hughes, R.E. 2000. Systematic and large-scale two-hybrid screens. *Curr. Opin. Microbiol.* **3:** 303–308.
Velculescu V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E. Jr, Hieter, P., Vogelstein, B., and Kinzler, K.W. 1997. Characterization of the yeast transcriptome. *Cell* **88:** 243–251.
Walhout, A.J. and Vidal, M. 2001. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* **24:** 297–306.
Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., and Somogyi, R. 1998. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci.* **95:** 334–339.
Westhead, D.R., Slidel, T.W., Flores, T.P., and Thornton, J.M. 1999. Protein structural topology: Automated analysis and diagrammatic representation. *Protein Sci.* **8:** 897–904.
Whitby, F.G., Masters, E.I., Kramer, L., Knowlton, J.R., Yao, Y., Wang, C.C., and Hill, C.P. 2000. Structural basis for the activation of 20S proteasomes by 11S regulators. *Nature* **408:** 115–120.
Wilkinson, C.R., Penney, M., McGurk, G., Wallace, M., and Gordon, C. 1999. The 26S proteasome of the fission yeast *Schizosaccharomyces pombe*. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **354:** 1523–1532.
Wilson, C.J., Chao, D.M., Imbalzano, A.N., Schnitzler, G.R., Kingston, R.E., and Young, R.A. 1996. RNA polymerase II holoenzyme contains SWI/SNF regulators involved in chromatin remodeling. *Cell* **84:** 235–244.
Winey, M., Yarar, D., Giddings, T.H., Jr., and Mastronarde, D.N. 1997. Nuclear pore complex number and distribution throughout the *Saccharomyces cerevisiae* cell cycle by three-dimensional reconstruction from electron micrographs of nuclear envelopes. *Mol. Biol. Cell* **8:** 2119–2132.
Woolford, J.L., and Warner, J.R. 1991. In *The molecular and cellular biology of the yeast Saccharomyces*: Genome dynamics, protein synthesis, and energetics (eds., J.R. Broach, J.R. Pringle, and E.W. Jones), pp. 587–626, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., and Eisenberg, D. 2000. DIP: The database of interacting proteins. *Nucleic Acids Res.* **28:** 289–291.

*Received July 18, 2001; accepted in revised form November 2, 2001.*