

Characterizing the Relationship between Protein-Fusion and Gene Co-Expression

Cary S. Gunther

csgunther@genomes.rockefeller.edu

Terry Gaasterland

gaasterl1@genomes.rockefeller.edu

Laboratory of Computational Genomics, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

Abstract

A pair of distinct proteins in one organism may most closely match different parts of the same protein in another organism. A comparison of all proteins from the genome of *Saccharomyces cerevisiae* and all proteins from 24 prokaryotic genomes yields 1010 pairs of yeast proteins whose homologs are parts of one protein from a prokaryotic genome. Marcotte *et al.* [12] showed that proteins related in this manner are more likely to interact than proteins chosen at random. In this paper, we investigated whether genes coding for such proteins are also likely to be concurrently transcribed. We identified 1010 fused pairs of proteins encoded in the yeast genome and analyzed expression of the corresponding genes at the transcriptional level. We found that the transcriptional profiles of fused gene pairs are significantly closer than those of randomly selected pairs. This finding is reproducible and established by multiple distance metrics. Moreover, such pairs frequently share additional biologically relevant properties. Thus, while protein fusion patterns are not predictive of co-expression, they are an important element in explaining co-expression. This justifies the use of curated protein fusion events to help characterize gene co-expression clusters.

Keywords: microarray, transcriptional profile, co-expression, protein interaction

1 Problem Definition and Background

The vastness of potential protein interaction space and heterogeneous quality of empirical data motivate the quest for methods to predict protein interactions, both to aid in the functional annotation of genomes and to help guide bench research. One such predictive method identifies “component” protein pairs in one organism that correspond to distinct parts of a single, “composite” protein in another organism. The composite protein is also referred to as a “Rosetta stone” sequence, to liken it to the famous multilingual artifact that was instrumental in the deciphering of Egyptian hieroglyphics. Finding a Rosetta stone relationship between a pair of proteins allows the inference that a meaningful, and possibly direct, interaction between the component proteins may exist [12]. Assessment of protein pairs identified by this approach suggests that they are in fact more likely to interact than randomly selected pairs [12, 7]; in addition, when Rosetta stone predictions agree with predictions based on correlated evolution or similar patterns of mRNA expression, the resulting prediction is as valid as experimental evidence [13, 15, 6].

In this study, we directly investigated the correlation between the Rosetta stone relationship and transcriptional regulation of the corresponding genes by using gene expression measurements performed on microarrays. Glass slide cDNA microarrays facilitate large scale investigations of patterns of mRNA abundance across entire genomes and broad classes of conditions (reviewed in [11]). A variety of techniques for evaluating shared behavior of genes and identifying genes that typify clusters have emerged in efforts to increase the relevance that can be assigned to array data ([5], reviewed in [11]).

Gasch *et al.* used microarrays to measure gene expression for most genes in the *Saccharomyces cerevisiae* genome under a wide array of conditions and to identify a set of genes commonly involved in the response to environmental stress [8]. We used the transcriptome-scale data generated by this study and the TANGO array database and analysis system [1, 26] to ask whether genes involved in a Rosetta, or fusion, relationship also tend towards co-expression at the mRNA level. Briefly, we found that three independent distance metrics demonstrate a statistically significant difference between gene expression correlation of fused gene pairs and that of randomly selected pairs. Moreover, fused gene pairs tend to share relationships represented by Gene Ontology terms [3] and KEGG pathway identifiers [10, 21].

2 Experimental Approach

To establish whether the genes encoding pairs of fused proteins are more likely to be co-expressed than random pairs of genes, we identified fused proteins in yeast based on 24 prokaryotic proteomes, computed gene expression profiles from 153 microarray hybridization datasets, and applied a probabilistic analysis to determine the correlation between fusion and co-expression.

Identification of fused proteins: Using PSIBLAST [2], we compared the yeast proteome to the proteomes of the 24 prokaryotes listed in Table I and compiled statistics on the startpoint, endpoint, and length of each HSP; percent identity over the stretch of each HSP; percent of the yeast protein included in each HSP; and percent of the prokaryotic protein included in each HSP. We then selected pairs of yeast proteins that generated HSPs with the same prokaryotic protein and for which these HSPs did not overlap by more than 20 residues of the prokaryotic sequence. We termed pairs for which at least 80% of each protein was included in the HSP “fused.” Pairs which did not meet this criterion we termed “mixed” to denote that at least one member of the pair consisted of multiple domains, not all of which were covered by the HSP. However, as initial studies did not discriminate between these two types of pairs, for the remainder of this study, the term “fused” will be used to refer to both categories.

Generation of transcriptional profiles: Gene expression data files generated with the Scanalyze software for 153 microarray 2-color hybridization experiments, along with jpeg files for spot visualization, were downloaded from the Stanford Microarray Database [25] and analyzed using our TANGO system [26], a relational database for microarray and DNA chip experiments accessed via web-form or command-line interface. TANGO performs tests of statistical significance on replicate experiments, supports several algorithms for clustering genes according to shared patterns of expression across experiments, and enables visualization of biological relationships among genes that co-cluster.

Probabilistic analysis of transcriptional profiles: We used a series of scripts available through TANGO to calculate the Euclidean and city-block distances, as well as the Pearson correlation coefficient, for each pair of ORFs on the Gasch arrays. The input for each distance measurement was a table of the log-transformed ratios of the expression levels of every ORF on the array under a test condition compared to baseline for each of the 153 experiments considered. We then used the R [23] programming environment to perform t-tests that compared the distance distributions for fused gene pairs to those for randomly selected pairs. Approximately 40,000 pairs were included in each randomly-selected sample.

3 Results

Previous work showed that protein pairs in one organism that correspond to single proteins in another organism are significantly more likely to interact than protein pairs picked at random [12]. We hypothesized that the genes encoding such fused protein pairs would tend to be transcriptionally co-regulated. To identify a set of pairs for use in testing this hypothesis, we used PSIBLAST [2] to compare the proteome of *Saccharomyces cerevisiae* with the proteomes of 24 prokaryotes (Table 1, see

Table 1: 24 prokaryotic targets used to identify fused yeast proteins.

Archaeoglobus fulgidus
Aquiflex aeolicus
Borrelia burgdorferi
Clostridium acetobutylicum
Chlamydomonas reinhardtii
Chlamydia trachomatis
Deinococcus radiodurans
Escherichia coli
Helicobacter pylori
Leishmania major
Methanococcus jannaschii
Mycobacterium tuberculosis
Neisseria meningitidis
Pseudomonas aeruginosa
Porphyromonas gingivalis
Pyrococcus horikoshii
Rickettsia prowazekii
Treponema pallidum
Ureaplasma urealyticum
Vibrio cholera
Mycoplasma genitalium
Mycoplasma pneumoniae
Methanococcus thermoautotrophicum
Synechocystis sp.

[27] for sequence sources). Limiting our results to at least 15% sequence identity, we identified 785 instances of pairs or triplets of yeast proteins that generated HSPs with the same prokaryotic protein and for which the corresponding segments of the prokaryotic proteins did not overlap by more than 20 residues. We imposed this constraint on maximal HSP overlap to ensure that pairs of yeast paralogs of the same prokaryotic protein were excluded from the list of fused pairs. Regarding each protein triplet as three protein pairs, this yielded a total of 1010 pairs of yeast proteins. To ensure that the protein pairs identified by our PSIBLAST output met the predictions of Marcotte *et al.*, we looked for occurrence of our fused pairs versus random pairs in two published comprehensive two-hybrid screens of the yeast proteome [19, 18, 9], as well as in the set of interactions examined by Schwikowski *et al.* [16] excluding those obtained from Proteome, Inc., and found that our fused pairs had indeed been experimentally found to interact more frequently than randomly chosen protein pairs (data not shown).

As a source of gene expression measurements, we chose the data set of Gasch *et al.* [8]. Briefly, these authors used cDNA arrays to assay transcriptional levels across roughly the entire yeast genome (~6200 genes) under a broad range of environmental stress conditions and in several mutant strains. Their data is publicly available via the Stanford Microarray Database [17, 25]. The variety of conditions employed in this study made it suitable for testing our hypothesis, as it reduced the likelihood of our finding chance associations between genes that are not generally co-expressed.

We uploaded data from 153 2-color glass-slide microarray hybridizations; i.e. cell source - treatment combinations from which mRNA had been isolated, reverse transcribed, labeled with two different dyes, and competitively hybridized to an array, into TANGO [25], our database and system for microarray

Table 2: Mean distances between pairs of proteins with associated p-values for seeing these sample distributions given a single dataset.

Distance Metric	Mean (random pairs)	Mean (fused)	Number of random pairs	p-value
Euclidean	14.25	13.78	38112	8.88×10^{-3}
City Block	131.6	120.3	37991	1.980×10^{-11}
Pearson	0.4806	0.4365	38267	$< 2.2 \times 10^{-16}$

and DNA chip analysis. Using the TANGO system, we generated profile clusters and K-means clusters of genes. Profile clustering assigns each arrayed gene to a “bin” that corresponds to its foldchange pattern across a sequence of hybridizations (supported in TANGO). K-means is a standard clustering technique in which an algorithm iteratively assigns genes to a user-defined number of clusters according to similarity of expression measurements [5]; K-means produces a set number of clusters of variable size. We then looked for co-clustering of members of fused gene pairs. Because of the large number of experiments analyzed, profile clustering generated predominantly singleton clusters (i.e., containing only one gene): the likelihood of exact co-expression of two genes in every instance of such a broad range of conditions is extremely low even for genes which are generally co-regulated. In addition, while the fused gene pairs did tend to co-cluster by the K-means approach, it was difficult to assign statistical significance to this result. This is because the K-means algorithm tended to generate a single, large cluster of mostly unregulated genes, and the size of this cluster meant that the expected frequency of co-clustering for randomly selected gene pairs was so high as to be near that of the fused gene pairs (data not shown).

To examine fused gene pair characteristics directly while avoiding the aforementioned difficulties, we generated distance matrices for each ORF represented on the arrays used by Gasch *et al.* against every other, using three distance metrics: Euclidian, city-block, and Pearson correlation coefficient. We then compared the distribution of distances between members of fused gene pairs to the distribution of approximately 40,000 randomly selected ORF pairs. By all three distance metrics, the average distance between the expression patterns of fused genes is significantly lower for fused pairs than for random pairs ($p=8.88e-3$ for Euclidean distances; $p=1.98e-11$ for city-block; $p<2.2e-16$ for Pearson; see Table 2 and Fig. 1). Although the differences in the means of the distributions are not large, the large sample size allows us to test whether the differences in the distributions overall are significant. To test the null hypothesis that the distributions are the same, we used the two-sided Kolmogorov-Smirnov test for goodness of fit. We applied the test to the fused pair and random pair distributions and rejected the null hypothesis at the 1% level for each distance metric. The test statistic $D=0.1507$ for city-block distance with $p<2.2e-16$; $D=0.1066$ for Euclidean distance with $p=3.727e-10$; for the Pearson metric, $D=0.186$ with $p<2.2e-16$.

As a control experiment, we also compared multiple samples of random pairs to each other and determined that the distance distributions for random samples of a given distance metric did not differ significantly (data not shown). To ensure that the randomly generated lists of ORF pairs were truly representative of the yeast genome as a whole, we also calculated the parametric mean of all pairs for the Euclidean and city-block distance metrics, and found essentially no difference between the parametric means and the means of our randomly selected samples (data not shown). Thus, fused gene pairs are significantly co-expressed in this gene expression data set.

We next investigated whether any of the statistics of the high scoring pairs (HSPs) of aligned sequences generated by PSIBLAST is predictive of the distance between the expression profiles of the fused genes. We examined percent-identity across the HSPs, percent coverage of the prokaryotic composite protein by the members of the yeast pair, and percent of the yeast proteins included in the

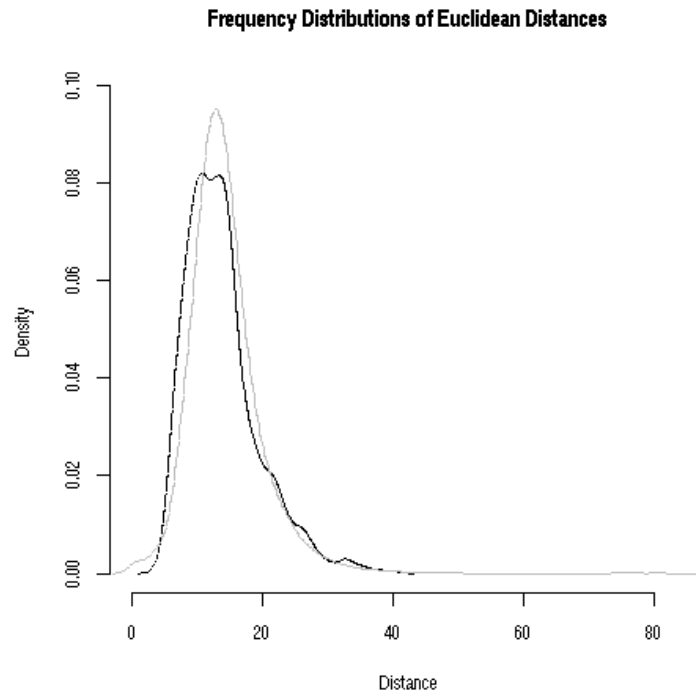


Figure 1: Frequency distributions of Euclidean distances for gene expression profiles of genes encoded fused protein pairs (black) and random pairs of genes (gray).

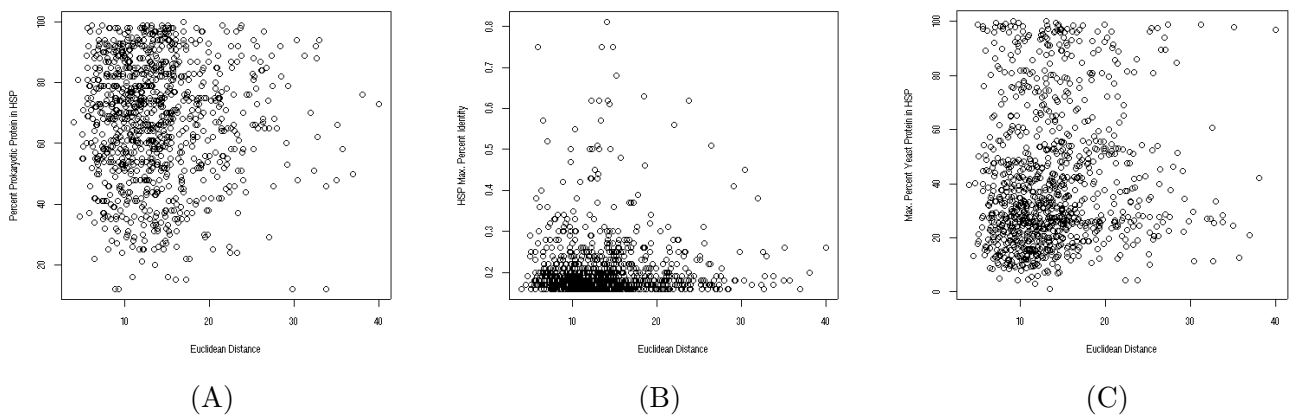


Figure 2: Distribution of Euclidean gene expression distance (x-axis) against psi-blast alignment properties (y-axis). (A) shows distance plotted against maximum percent identity of HSPs for fused proteins. (B) shows distance plotted against percent of the prokaryotic protein covered by the fused proteins. (C) shows distance plotted against maximum percent of fused protein covered by the prokaryotic protein. None of the distributions indicates a strong relationship between gene expression distance and alignment quality.

Table 3: Correspondence between GO terms of fused protein pairs and random protein pairs.

Shared GO term type	Biological process	Cellular component	Molecular function
Fused/mixed pairs	321/1010=31.8%	41/1010=4.1%	196/1010=19.4%
Random pairs	5319/37991=14.0%	115/37991=0.303%	6700/37991=17.6%
Odds ratio	2.27	13.4	1.10
Controlled for term attribution			
Controlled for term attribution	Biological process	Cellular component	Molecular function
Fused/mixed pairs	44.8%	21.1%	31.9%
Random pairs	33.8%	4.9%	41.0%
Ratio	1.33	4.36	0.779
“Top level” term sharing			
“Top level” term sharing	Biological process	Cellular component	Molecular function
Fused/mixed pairs	34/1010=3.4%	22/1010=2.2%	32/1010=3.2%
Random pairs	56/37991=0.15%	42/37991=0.11%	34/37991=0.089%
Ratio	~23	~20	~35

HSP (Fig. 2) but could not find an apparent relationship with any of the distance metrics. Indeed, calculation of the correlation of each of these three PSIBLAST statistics with the expression distances yielded near-zero values, indicating a lack of strong correlation in each case (data not shown). Using chromosome maps downloaded from the MIPS yeast genome database [14, 22], we determined that for 85 of the 1010, or 8.4%, of the fused pairs, both genes resided on the same chromosome, compared to approximately 5.5% for randomly selected pairs.

It seems reasonable that interacting proteins participate in one or more common tasks within the cell. Proteins sharing a Rosetta fusion relationship have previously been shown to interact with frequency above that of random proteins [12], and our examination of yeast microarray data suggests that the genes encoding such proteins tend to be co-expressed. Therefore, to further explore the functional relationships between members of fused pairs, we examined the sharing of Gene Ontology (GO) terms by members of these pairs.

GO terms derive from a controlled vocabulary constructed in an ongoing attempt to unify the way in which biological and biochemical events are described [3, 20]. Three categories of GO terms exist: cellular component, molecular function, and biological process. We compared the frequency with which members of fused pairs shared GO terms in each of these three categories with the frequency for randomly selected ORF pairs. For both the biological process and cellular component categories, the rate at which fused pairs shared GO terms is higher than that of randomly selected pairs, while for molecular function category, the rates are essentially indistinguishable (Table 3). However, association of ORFs with GO terms is an ongoing process. To eliminate any bias resulting from the possibility that more is known about genes in fused pairs, we accounted for the rates of attribution of GO terms to genes in fused pairs and randomly selected ORFs by selecting fused pairs for which both members had a term assigned in the given category. and generating a random set of pairs for which this was also true. This partially reduces the apparent contrast. To refine our examination of GO terms, we obtained the top level GO term, or that most commonly used to annotate an ORF, for each ORF from the *Saccharomyces* Genome Database [25], and excluded annotations of “process (,function, etc.) unknown.” The results of this second pass are also included in Table 3, and the results from all three term categories more firmly establish that functional properties are shared by members of fused gene pairs at a higher rate than across random pairs.

As an additional test of functional relatedness, we examined the participation of products of fused

Table 4: Numbers of shared KEGG pathways for fused protein pairs and random protein pairs.

	Fused pairs	Random pairs
Pairs in ≥ 1 shared KEGG pathway	36	78
Total # of shared pathways	66	100
Shared pathways/pair	$66/1010 = 6.5 \times 10^{-2}$	$100/37991 = 2.6 \times 10^{-3}$

gene pairs in pathways described by the KEGG database [21]. These pathways are an extension of the enzyme classification system and reflect both direct (i.e., physical) and indirect (e.g., epigenetic or substrate handling) interactions between proteins or gene products. For about 6.5% of fused pairs, both genes participate in one or more of the same KEGG pathways, compared to only 0.26% of random pairs (Table 4). This finding imputes additional biological relevance to the co-expression of genes in fused pairs and strengthens the assertion that the fusion, or Rosetta, relationship can in fact predict interaction between genes in a pair.

4 Discussion and Conclusions

We used the results of a PSIBLAST comparison of the *Saccharomyces cerevisiae* proteome against that of 24 fully-sequenced prokaryotes to identify pairs of genes that appeared to exist as “fused” genes in prokaryotes and then test for co-transcription of these genes. Our set of gene pairs is markedly smaller than that of Marcotte *et al.* [12], who initially identified the component/composite or Rosetta protein relationship; however, as we do not know the exact criteria by which the original gene pairs were identified, we cannot directly compare the test sets. Yeast is well-suited to this investigation in part because its genome is small by eukaryotic standards, and so each gene has a relatively small number of paralogs among which relationships are distributed. Moreover, as *S. cerevisiae* is unicellular, we do not have to account for tissue-specificity of gene expression profiles; rather, the combination of strain, set of conditions in which the cell lives, and history of the cell should be deterministic for its transcriptional profile.

The authors who originally identified the Rosetta relationship noted that protein pairs that share this relationship may sometimes not interact if the original fusion of the two domains in a single protein was for purposes of coordinate regulation rather than direct physical interaction [12]. Gene expression measurements should reflect this, and thus support the concept that the relationship is a biologically meaningful one. However, the inverse is also possible. Proteins for which the fusion served to effectively increase affinity and thus aid kinetics of interaction within the cell may appear to be unlinked if the gene which codes for one member of the pair does not need to be tightly regulated and is therefore constitutively expressed at a somewhat constant level, while the other member of the pair undergoes tighter fluctuations in transcriptional regulation. Moreover, small mutations that affect critical residues, e.g. in the active site of an enzyme, can markedly alter the function of a protein and necessitate a change in the regulation of its expression; however, such small sequence changes may not be sufficient to prevent such proteins from generating HSPs with the same prokaryotic protein. This effect will result in determination of some gene pairs as “fused” that do not really belong in this category, increasing the apparent mean distance between such gene pairs.

It is probable that we have underestimated the mean distance between randomly chosen ORF pairs, as the nomenclature of spots on the microarrays used by Gasch *et al.* [8] includes some synonyms for which our system did not account. This did not affect measurement of the distance distributions for fused gene pairs because these were a pre-defined set for which we had exact names. However, the existence of these synonyms means that some of the randomly selected pairs are essentially comparisons of an ORF with itself, which should result in a near-zero distance, the only difference resulting from

random biological variation or handling error. These near-zero values artificially lower the mean for randomly-selected pairs. Taking this effect into account would, if anything, increase the statistical significance of our comparison of distance metrics. As it stands, although the differences in means between fused pairs and randomly selected pairs are not large, our sample sizes are more than sufficient to firmly establish the statistical significance of the hypothesis that these two types of gene pairs differ in transcriptional profile distance.

The 1010 unique fusion pairs, or 785 unique instances of pairs and triplets, evaluated in this study are distributed among 676 unique yeast genes, indicating that some genes occur more than once. In addition, some gene pairs appear to fuse in more than one prokaryotic genome. As the number of fully sequenced prokaryotes increases, it will be interesting to determine whether pairs that fuse in multiple genomes share even closer functional relationships. In addition, as Eisenberg and colleagues have found that the corroboration of multiple predictive methods is a more certain indication of an interaction between members of a gene pair [13, 6], multiply-predicted interacting gene pairs may be even more likely to have closely related transcriptional profiles.

For assaying the functional relationships of fused pairs, it was necessary to set bounds on the levels of GO term and KEGG pathway matches. For the top level GO term analysis, only exact matches were considered; hence, biological process similarities such as “transcription from Pol II promoter” and “transcription from Pol III promoter” were considered to be non-matching, as were some high-level KEGG pathway matches. A more detailed comparison of word frequencies in description lines for these genes might illuminate such slightly more distant relationships. Additional study may also reveal that proteins from a particular class, e.g. enzymes, signaling proteins, or adaptor molecules, tend to fuse in prokaryotes more frequently than other proteins. Furthermore, it will be interesting to examine upstream regulatory regions of Rosetta-related genes for shared motifs and transcription factor binding sites.

In conclusion, we established that the mRNA expression patterns of members of fused gene pairs are significantly correlated. This correlation is significant, for a large number of degrees of freedom, whether the transcriptional patterns are compared using a Euclidean, city-block, or Pearson distance metric. Moreover, fused gene pairs share biological properties at a higher rate than randomly selected pairs, suggesting that the fusion relationship is functionally relevant. Our findings indicate that protein-fusion events help to explain, but not predict, co-expression patterns.

Acknowledgments

This work was supported by NCI grant #CA84699-01, the Burroughs-Wellcome Foundation, the Mathers Foundation, and NSF grant #DBI-9984882. In addition, C.S.G. was supported by MSTP grant #GM07739, NCI grant #T32 CA09763-26, and Rockefeller University institutional funds.

References

- [1] Altmann, C.R., Bell, E., Sczyrba, A., Pun, J., Bekiranov, S., Gaasterland, T., and Brivanlou, A.H., Microarray- based analysis of early development in *Xenopus laevis*, *Developmental Bio.*, 236:64–75, 2001.
- [2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25:3389–3402, 1997.
- [3] 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., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G.,

- Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.*, 25:25–29, 2000.
- [4] Ball, C.A., Dolinski, K., Dwight, S.S., Harris, M.A., Issel-Tarver, L., Kasarskis, A., Scafe, C.R., Sherlock, G., Binkley, G., Jin, H., Kaloper, M., Orr, S.D., Schroeder, M., Weng, S., Zhu, Y., Botstein, D., and Cherry, J.M., Integrating functional genomic information into the *Saccharomyces* Genome Database, *Nucleic Acids Res.*, 28:77–80, 2000.
 - [5] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–14868, 1998.
 - [6] Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O., Protein function in the post-genomic era, *Nature*, 405:823–826, 2000.
 - [7] Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzonis, C.A., Protein interaction maps for complete genomes based on gene fusion events, *Nature*, 402:86–90, 1999.
 - [8] Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O., Genomic expression programs in the response of yeast cells to environmental changes, *Mol. Biol. Cell.*, 11:4241–4257, 2000.
 - [9] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. U.S.A.*, 98:4569–4574, 2001.
 - [10] Kanehisa, M. and Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res.*, 28:27–30, 2000.
 - [11] Lockhart, D.J. and Winzler, E.A., Genomics, gene expression, and gene arrays, *Nature*, 405:827–836, 2000.
 - [12] Marcotte, E.M., Pellegrini, M., Ng, H., Rice, D.W., Yeates, T.O., and Eisenberg, D., Detecting protein function and protein-protein interactions from genome sequences, *Science*, 285:751–753, 1999.
 - [13] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D., A combined algorithm for genome-wide prediction of protein function, *Nature*, 402:83–86, 1999.
 - [14] Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaput, G., Pfeiffer, F., Schuller, C., Stocker, S., and Weil, B., MIPS: a database for genomes and protein sequences, *Nucleic Acids Res.*, 28:37–40, 2000.
 - [15] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O., Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proc. Natl. Acad. Sci. USA*, 96:4285–4288, 1999.
 - [16] Schwikowski, B., Uetz, P., and Fields, S., A network of protein-protein interactions in yeast, *Nat. Biotechnol.*, 18:1242–1243, 2000.
 - [17] Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A., Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., and Cherry, J.M., The Stanford Microarray Database, *Nucleic Acids Res.*, 29:152–155, 2001.

- [18] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochar, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J.M., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, 403:623–631, 2000.
- [19] Uetz, P. and Hughes, R.E., Systematic and large-scale two-hybrid screens. *Curr. Opin. Microbio.*, 3:303–308, 2000.
- [20] GO: <http://www.geneontology.org/>
- [21] KEGG: <http://www.genome.ad.jp/kegg/>
- [22] MIPS: <http://mips.gsf.de/proj/yeast/CYGD/db/>
- [23] R: <http://www.r-project.org>
- [24] *Saccharomyces* Genome Database: Cherry, J.M., Ball, C., Dolinski, K., Dwight, S., Harris, M., Matese, J.C., Sherlock, G., Binkley, G., Jin, H., Weng, S., and Botstein, D., <ftp://genome-ftp.stanford.edu/pub/yeast/SacchDB/> (August 2001)
- [25] Stanford Microarray Database: <http://genome-www4.stanford.edu/MicroArray/SMD/>
- [26] TANGO / Xenopus project: <http://arrays.rockefeller.edu/xenopus>
- [27] TIGR MICROBIAL GENOME LIST: <http://www.tigr.org>