Analysis of the Yeast Transcriptome with Structural and Functional Categories: Characterizing Highly Expressed Proteins

Ronald Jansen

&

Mark Gerstein

Department of Molecular Biophysics & Biochemistry
266 Whitney Avenue, Yale University
PO Box 208114, New Haven, CT 06520
(203) 432-6105, FAX (360) 838-7861
Mark.Gerstein@yale.edu

(Revised Version)

ABSTRACT

We analyzed ten genome expression data sets by large-scale cross-referencing against broad structural and functional categories. The data sets, generated by different techniques (for instance, SAGE and gene chips), provide various representations of the yeast transcriptome (the set of all yeast genes, weighted by transcript abundance). Our analysis enabled us to determine features more prevalent in the transcriptome than the genome -- i.e., those that are common to highly expressed proteins. Starting with simplest categories, we find that, relative to the genome, the transcriptome is enriched in Ala and Gly and depleted in Asn and very long proteins. We find, furthermore, that protein length and maximum expression level have a roughly inverse relationship. To relate expression level and protein structure, we assigned transmembrane helices and known folds (using PSI-blast) to each protein in the genome; this allowed us to determine that the transcriptome is enriched in mixed alpha-beta structures and depleted in membrane proteins relative to the genome. In particular, some enzymatic folds, such as the TIM barrel and the G3P dehydrogenase fold, are much more prevalent in the transcriptome than the genome, whereas others, such as the protein-kinase and leucine-zipper folds, are depleted. The TIM-barrel, in fact, is overwhelmingly the "top fold" in the transcriptome, while it only ranks fifth in the genome. The most highly enriched functional categories in the transcriptome (based on the MIPS system) are energy production and protein synthesis, while categories such as transcription, transport, and signaling are depleted. Furthermore, for a given functional category, transcriptome enrichment varies quite substantially between the different expression data sets, with a variation an order of magnitude larger than for the other categories cross-referenced (e.g., amino acids). One can readily see how the enrichment and depletion of the various functional categories relates directly to that of particular folds. Further information can be found at bioinfo.mbb.yale.edu/genome/expression.

INTRODUCTION

Whole-genome expression experiments have become important tools in functional genomics. The result of these experiments, the expression levels of all the genes in the genome, has been dubbed the transcriptome (1). Many of the initial expression experiments have focused on the eukaryote yeast for technical reasons as well as the fact that it is a widely studied model organism with a known genome sequence (2). Quantitative profiles of the yeast transcriptome have been determined for a variety of conditions using serial analysis of gene expression (SAGE) (1) as well as gene chip technology (8, 9, 10, 11). Brown and colleagues have developed cDNA microarrays to conduct time-course experiments measuring the expression changes of yeast genes in response to a variety of conditions (3, 4, 5; 6, 7). Researchers have also started to investigate quantitative *protein* abundance profiles for yeast, using such approaches as fusion proteins (12) and 2D-gels (13).

Various approaches have been proposed to interpret the wealth of data generated by these experiments. Algorithms to cluster genes into functionally related groups have been proposed (14, 15, 16, 17). Roth et al. (10), van Helden et al. (18), and Brazma et al. (19) have introduced new ways to identify regulatory regions located upstream of genes. Gerstein proposed an initial ranking of protein folds in terms of their expression levels (20). A number of proposals have been made for the archiving and management of expression data (21).

Here we present another way to interpret gene expression data. We perform large-scale "cross-referencing" of expression data against a number of structural and functional categories. These categories include (i) simple characteristics shared by all proteins, their amino-acid composition and length; (ii) aspects of protein structure, fold family and number of transmembrane helices; and (iii) broad functional classes. The correlation of expression level with these categories gives us insight into the characteristics of highly expressed proteins and also suggests some interesting conclusions about the overall biochemistry of the yeast cell. More specifically, we compare the composition of all our categories in the transcriptome with that in the genome. We find that the transcriptome is notably enriched with certain types of proteins -- e.g., those rich in Ala and Gly, those with a mixed alpha-beta structure, and those associated with energy production and protein synthesis -- and depleted in others -- e.g., Asn-rich proteins, membrane proteins, very long proteins, and transcription factors and transport proteins.

Expression Data

We based our analysis of the yeast transcriptome on a diverse set of publicly available expression experiments, which are summarized in Table 1 a. Including data sets derived from different experimental techniques potentially reduces the bias introduced by focusing on one particular experiment. We focused more on data from DNA chips and SAGE technology rather than cDNA microarray experiments, since DNA chips and SAGE allow a better measurement of the absolute number of transcript copies for an open reading frame (ORF), facilitating direct comparisons between ORFs. In contrast, cDNA microarrays mainly measure expression level *changes* of a given ORF as ratios to

a reference point, and ORF-to-ORF comparisons at a given time point are more problematic.

In presenting our data, we decided, for convenience, to use the data set generated by Holstege et al. (9) as the main reference. This data set represents the average of two transcript abundance level measurements for most yeast genes. Furthermore, the authors report that 99% of these transcripts exhibited a less than 2-fold change in the two measurements. We also extensively used the SAGE data sets (1), which give expression profiles of a large but less complete subset of the yeast genome in different conditions; the gene chip data generated by Roth et al. (10), which represent profiles of the yeast transcriptome for different conditions; and the gene chip data by Jelinsky et al. (11), who investigated expression profiles before and after the yeast cell is subjected to an alkylating agent (we only used the first, more typical, profile).

General Approach

Most of our analyses have the same basic structure, which is schematized in Table 1 b. First, we compute the genome composition of a specific category, then we compute its composition in the transcriptome, and finally we determine its enrichment in the transcriptome, that is, the relative difference between transcriptome and genome composition. For computing transcriptome compositions we weight each gene with its respective expression level. With the term "genome" we refer, strictly speaking, only to the set of open reading frames which are covered by each particular expression experiment. In this sense, the "genomes" covered by two different expression

experiments might include different yeast ORFs, and therefore their composition (of a particular amino acid, for instance) might be different -- though in practice these differences are generally very small.

Our complete results and additional information (such as genome and transcriptome compositions, and number of proteins per category) are available on the internet at http://bioinfo.mbb.yale.edu/genome/expression.

RESULTS

Transcriptome Composition of Amino Acids

One of the simplest attributes associated with a protein is its amino acid composition. The amino acid compositions of the genome and the transcriptome differ significantly for some amino acids. This is shown in Figure 1 a. The amino acids are ordered along the x-axis in the order of increasing transcriptome enrichment for the reference data set by Holstege et al. (9). Although the results vary between the different expression data sets, they all follow a general trend. Most notably, the composition of Ala increases by about 30 to 40 % whereas the composition of Asn decreases by about 20 %. The transcriptome is also significantly enriched in Gly and Val and the positively charged amino acids, Arg and Lys.

As mentioned before, the data from cDNA microarrays, as given by ratios of red and green fluorescence intensities, is primarily used for the measurement of expression level

changes. These data are less suitable for absolute expression level measurements. For purely illustrative purposes, we analyzed amino-acid enrichment in the transcriptome using the red fluorescence intensity less the background intensity of the cDNA microarray data set as a crude approximation of the absolute expression level (Figure 1 b). Although the results for the enrichment of amino acid composition have a trend similar to that in figure 1 a, the magnitudes are much smaller (as expected). It can also be observed that there appears to be little difference in the amino acid composition of the transcriptome for different time points measured during the diauxic shift experiment, suggesting that even though the precise proteins that make up the transcriptome change in different conditions, the overall amino acid composition remains very similar. This is also suggested by the fact that there is little variance in transcriptome amino acid composition between DNA chip experiments in different conditions -- i.e., between the different data sets of Roth et al. (10).

Relationship between Gene Length and Expression Level

Figure 2 shows the relationship between protein length (measured by the number of residues in the sequence) and expression level for the reference data set. It is obvious that there is no direct relationship between these two quantities. However, it seems that protein length is in some way an upper limit for the expression level of the corresponding gene. The straight line in figure 2 represents the fit of a hyperbolic function through the maximum protein length at a given expression level. If the maximum protein length for a given expression level were inversely proportional to the expression level, the slope of this line would be equal to about -1. We find the slope to be about -0.7 for the data set of

Holstege et al. (9). We find similar relationships for the other data sets (details in caption to figure 2). The expression level of a short gene is dependent on the rate of transcription of RNA polymerase in relation to the rate of mRNA degradation. However, for a long gene, the overall rate of transcription might also be affected by the processivity of RNA polymerase -- i.e. by the chance that the polymerase falls off.

Transcriptome Composition of Membrane Proteins

Another aspect of protein structure we analyzed was the occurrence of membrane proteins in the transcriptome. Membrane proteins are often classified in terms of the number of hydrophobic transmembrane (TM) helices they contain. We identified yeast ORFs coding for membrane proteins using a standard hydropathy scale and a sliding window, as described previously (20) (further details in the caption to figure 3 a). Based on their most hydrophobic segment, we divided the predicted membrane proteins into "sure" and "marginal" candidates (using the MaxH approach also described in the caption) and then classified them further based on the number of TM-helices they contain. Figure 3 a shows how the composition of ORFs with "sure" transmembrane regions changes from genome to transcriptome. For comparison we also show the relative enrichment of soluble proteins (for which no transmembrane region is predicted). The results show that, in general, helical membrane proteins are underrepresented in the transcriptome relative to the genome, whereas soluble proteins are enriched by about 22%. Furthermore, some classes of membrane proteins are more highly enriched than others, for instance, those with four TM-helices are more enriched than those with one or

two TM-helices. However, for many of the membrane structure categories there is considerable variation between the different experiments.

Transcriptome Composition of Fold Classes

In the previous section we compared the transcriptome enrichment of membrane and soluble proteins. Here we subdivide soluble proteins further according to their folds. To do this, we matched the PDB structure database (22) against the yeast genome using an iterative database search program (PSI-blast) (23) (see caption to figure 3 b for more details on our fold assignment methods). Overall we found a total of 2305 domain level matches in 1710 distinct ORFs (about 27% of the genome). We classified these structure matches into one of 344 folds using the Structural Classification of Proteins (SCOP) (24). In addition, each fold is further grouped into one of six soluble protein classes – for instance, all-alpha, all-beta, alpha/beta, etc. (25).

For each domain match we looked at the expression level of the corresponding ORF. Figure 3 b shows the relative differences of the composition of protein fold classes between genome and transcriptome. The fold classes are sorted along the x-axis in the order of increasing transcriptome enrichment for the reference data set. We observe an increase in the fraction of mixed alpha and beta folds (alpha+beta and alpha/beta) while the fraction of the other fold classes decreases. It is also interesting to note that while the all-alpha class is depleted in the transcriptome, the most helix-favoring amino acid, Ala (26), is greatly enriched (see figure 1 a).

For fold class composition, the results for the SAGE experiments and the gene chip experiments differ quite a bit from each other. This may be attributed to the substantially smaller number of ORFs covered by the SAGE experiments, which sample the structure matches in a somewhat biased fashion. Furthermore, the much greater enrichment of mixed helix and sheet structures in the SAGE experiments may, to some degree, result from the fact that these proteins tend to be longer (27) and the SAGE experiment is somewhat weighted towards longer proteins.

Top Folds in the Transcriptome

Figure 4 shows the top-ten most highly expressed protein folds in yeast. Their exact fractions in the transcriptome are listed for the reference data set of Holstege et al. (9) and schematized with rankings for the other sets. The ranking of the most common folds in the transcriptome and the genome are very different. The most common transcriptome fold, by a large margin, is the TIM-barrel (8% vs. 5% for the second ranked fold), which is by contrast only ranked fifth in the genome. Many of the other common folds in the transcriptome also have a mixed alpha/beta structure and are associated with enzymatic functions, for instance, the P-loop NTP hydrolase, ferrodoxin, Rossmann, thioredoxin, and G3P dehydrogenase folds. In particular, the G3P dehydrogenase fold is greatly enriched in the transcriptome relative to the genome, increasing from 0.2% to 2.7%. Common folds in the genome that are depleted in the transcriptome include the protein kinase (catalytic core), long helix oligomers (the Leu-zipper fold), and the Zn2-C6 DNA binding domain. This makes sense since these folds act as "switches" in signaling and transcription-factor functionality and thus do not need to be present in large quantities. In

contrast, cytosolic enzymes are needed in bulk to ensure high throughput in synthetic and energy-producing pathways (see figure 5).

The top-folds analysis is a relatively "fine-grained" measurement, dividing the transcriptome into many categories, thus making the differences between the various experiments more apparent. Some of these differences may be explained by the different conditions probed by each experiment; others may reflect the natural variability of the experiments. However, in all cases the most common transcriptome fold always remains the TIM-barrel. These fine-grained differences are also evident in the analysis of cDNA microarray data for the diauxic shift in yeast (5), which shows that the fold class composition does not change much over the time course of the experiment, but the ranking of the most common folds by expression level does. (Data not shown; related data in ref. 19; absolute expression levels are approximated as explained in the caption to figure 1 b.)

It is well known that protein abundance can vary quite significantly for a given mRNA transcript abundance level. Recent large-scale studies suggest that there is only a weak linear relationship between mRNA and protein abundance for many genes, especially for weakly expressed genes (13). On the other hand, mRNA abundance is certainly still a better measure of protein abundance than genome content. From our results it seems clear that the distribution of folds in the cell's proteins is very different from that in the genome complement.

Transcriptome Composition of Functions

To analyze the transcriptome in terms of broad functional categories, we used the functional categorization of the Munich Information Center for Protein Sequences (MIPS) (28, 29, 30), which divides proteins amongst a hierarchy of functional categories (for instance, “synthesis”, “metabolism” etc. on the top level of the hierarchy).

Figure 5 shows the transcriptome enrichment of the various functional categories at the top level of the MIPS system. The functional categories are sorted along the x-axis in the order of increasing transcriptome enrichment for the reference data set. We observe an increase in the number of the proteins in the category "protein synthesis" of about 200 - 500 % depending on the data set. This is considerably larger than the change for the structural categories or simpler categories such as amino acid composition (5-fold vs. 40%). The transcriptome is also notably enriched in proteins associated with energy production, cell structure, and protein synthesis (most often ribosomal proteins). None of the other broad categories are as greatly depleted as these are enriched. However, it is worth noting that the depleted categories include transcription factors and signaling and transport proteins. Furthermore, the fraction of unclassified proteins in the transcriptome is lower than in the genome, perhaps because the more highly expressed genes are easier to study experimentally. There is also great variability between the different experiments; depending on the experiment, the most highly enriched MIPS category is different. (For instance, the most highly enriched category is “protein synthesis” for the reference data set by Holstege et al. (9), but “energy” for some of the SAGE data sets.)

DISCUSSION AND CONCLUSION

It is clear from our results that the structural and functional categories we investigated are differently distributed in the transcriptome and the genome. That is, the proteins of highly transcribed genes have on average different characteristics than the unweighted protein complement in the genome. There are variations between the different expression experiments, but we can observe some general trends in how structural and functional features occur in the transcriptome. In particular, we find that the transcriptome is enriched in Ala, Gly and, to a lesser extent, positively charged residues, soluble folds with combinations of helices and sheets, and proteins involved in protein synthesis (in particular ribosomal proteins), cell structure, and energy production. Likewise, it is depleted in membrane proteins, transport, transcription, and signaling proteins, very long proteins, and those rich in Asn. Common sense, as well as a number of previous surveys, suggests that many of these structural and functional categories are interrelated (31, 32). Thus, for instance, proteins involved with protein synthesis or energy production are often enzymes, which tend to be associated with alpha/beta architectures. Likewise, membrane proteins tend to have less charged residues than soluble ones and also tend to have transport or signaling functions.

Looking at the variability of the transcriptome enrichment between experiments, it is particularly interesting to note that the greatest variability can be observed for the MIPS functional categories while the variability of amino acid composition is an order of magnitude lower. It seems that the usage of amino acids is very similar even when differential gene expression occurs to accommodate different functional tasks in the cell.

This indicates that the cell might have to meet general requirements in its amino acid usage.

One requirement might be energy expenditure. In the metabolism of the yeast cell, Ala, which is the most enriched amino acid in the transcriptome, is synthesized directly in one step from pyruvate, a precursor of the TCA cycle. In contrast, Asn, the most depleted amino acid in the transcriptome, follows a more involved route. It is synthesized in two steps from oxaloacetate, the last component in the TCA cycle; in addition, the conversion of aspartate (Asp) to asparagine (Asn) involves conversion of ATP to AMP. This is the only step in amino acid biosynthesis in which two pyrophosphates are consumed at the same time. Thus, by strongly favoring Ala over Asn in highly expressed proteins, it seems that the cell has adapted to these energetic realities in the course of evolution. Further research could elaborate on this anecdotal evidence by looking comprehensively at the metabolic network in the cell.

In the context of Asn, it is also interesting to note that in some organisms (notably some archeons) Asn-tRNA is produced by an alternative pathway (transamidation) from Asp-tRNA (45). In the mitochondria of yeast, Gln-tRNA is synthesized by transamidation from Glu-tRNA; this might be related to the depletion of Gln in the yeast transcriptome.

It is worth emphasizing that this study uses mRNA abundance rather than protein abundance in the cell. It is to be hoped that techniques for large-scale protein abundance

measurement will be developed that will provide us with better view of the cellular machinery.

ACKNOWLEDGEMENTS

We would like to thank Hedi Hegyi for help with matching PDB structures to the yeast genome. MG would like to thank the Keck foundation and the NIH (grant 2P01GM54160-04) for support.

REFERENCES

1. 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) *Cell*, **88**(2)**,** 243-51.

2. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996) *Science*, **274**(5287)**,** 546, 563-7.

3. Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995) *Science*, **270**(5235)**,** 467-70.

4. Shalon, D., Smith, S. J. and Brown, P. O. (1996) *Genome Res*, **6**(7)**,** 639-45.

5. DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997) *Science*, **278**(5338)**,** 680-6.

6. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. and Herskowitz, I. (1998) *Science*, **282**(5389)**,** 699-705.

7. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) *Mol Biol Cell*, **9**(12)**,** 3273-97.

8. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E. L. (1996) *Nat Biotechnol*, **14**(13)**,** 1675-80.

9. 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) *Cell*, **95**(5)**,** 717-28.

10. Roth, F. P., Hughes, J. D., Estep, P. W. and Church, G. M. (1998) *Nat Biotechnol*, **16**(10)**,** 939-45.

11. Jelinsky, S. A. and Samson, L. D. (1999) *Proc Natl Acad Sci U S A*, **96**(4)**,** 1486-91.

12. Ross-Macdonald, P., Sheehan, A., Roeder, G. S. and Snyder, M. (1997) *Proc Natl Acad Sci U S A*, **94**(1)**,** 190-5.

13. Gygi, S. P., Rochon, Y., Franza, B. R. and Aebersold, R. (1999) *Mol Cell Biol*, **19**(3)**,** 1720-30.

14. Michaels, G. S., Carr, D. B., Askenazi, M., Fuhrman, S., Wen, X. and Somogyi, R. (1998) *Pac Symp Biocomput***,** 42-53.

15. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) *Proc Natl Acad Sci U S A*, **95**(25)**,** 14863-8.

16. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999) *Proc Natl Acad Sci U S A*, **96**(6)**,** 2907-12.

17. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M. (1999) *Nat Genet*, **22**(3)**,** 281-5.

18 van Helden, J., Andre, B., and Collado-Vides, J. (1998) *J Mol Biol,* **281**(5):827-42.

19. Brazma, A., Jonassen, I., Vilo, J. and Ukkonen, E. (1998) *Genome Res*, **8**(11)**,** 1202-15.

20. Gerstein, M. (1998) *Proteins*, **33**(4)**,** 518-34.

21. Ermolaeva, O., Rastogi, M., Pruitt, K. D., Schuler, G. D., Bittner, M. L., Chen, Y., Simon, R., Meltzer, P., Trent, J. M. and Boguski, M. S. (1998) *Nat Genet*, **20**(1)**,** 19-23.

22. Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. and Abola, E. E. (1998) *Acta Crystallogr D Biol Crystallogr*, **54**(1 (Pt 6))**,** 1078-84.

23. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) *Nucleic Acids Res*, **25**(17)**,** 3389-402.

24. Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) *J Mol Biol*, **247**(4)**,** 536-40.

25. Levitt, M. and Chothia, C. (1976) *Nature*, **261**(5561)**,** 552-8.

26. Chakrabartty, A., Kortemme, T. and Baldwin, R. L. (1994) *Protein Science*, **3**(5)**,** 843-52.

27. Gerstein, M. (1998) *Fold Des*, **3**(6)**,** 497-512.

28. Mewes, H. W., Hani, J., Pfeiffer, F. and Frishman, D. (1998) *Nucleic Acids Res*, **26**(1)**,** 33-7.

29. Frishman, D., Heumann, K., Lesk, A. and Mewes, H. W. (1998) *Bioinformatics*, **14**(7)**,** 551-61.

30. Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. and Frishman, D. (1999) *Nucleic Acids Res*, **27**(1)**,** 44-8.

31. Hegyi, H. and Gerstein, M. (1999) *J Mol Biol*, **288**(1)**,** 147-64.

32. Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. and Thornton, J. M. (1998) *Structure*, **6**(7)**,** 875-84.

33. Engelman, D. M., Steitz, T. A. and Goldman, A. (1986) *Annu Rev Biophys Biophys Chem*, **15,** 321-53.

34. Arkin, I., Brunger, A. and Engelman, D. (1997) *Proteins*, **28,** 465-466.

35. Wallin, E. and von Heijne, G. (1998) *Protein Sci*, **7**(4)**,** 1029-38.

36. Tomb, J.-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karpk, P. D., Smith, H. O., Fraser, C. M. and Venter, J. C. (1997) *Nature*, **388,** 539-547.

37. Boyd, D., Schierle, C. and Beckwith, J. (1998) *Protein Sci*, **7**(1)**,** 201-5.

38. Klein, P., Kanehisa, M. and DeLisi, C. (1985) *Biochim Biophys Acta*, **815**(3)**,** 468-76.

39. Lipman, D. J. and Pearson, W. R. (1985) *Science*, **227,** 1435-1441.

40. Altschul, S., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.*, **215,** 403-410.

41. Pearson, W. R. (1997) *Comput Appl Biosci*, **13**(4)**,** 325-32.

42. Pearson, W. R. (1998) *J Mol Biol*, **276**(1)**,** 71-84.

43. Gerstein, M. G. (1997) *J Mol Biol*, **274**, 562-576.

44. Teichmann, S., Park, J. and Chothia, C. (1998) *Proc. Natl. Acad. Sci.*, **95,** 14658-63.

45. Tumbula, D., Vothknecht, U. C., Kim, H. S., Ibba, M., Min, B., Li, T., Pelaschier, J., Stathopoulos, C., Becker, H. and Soll, D. (1999) *Genetics*, **152**(4)**,** 1269-76.

FIGURES AND TABLES

Figure 1, Transcriptome Enrichment of Amino Acids

In **Part A**, the amino acids are ordered along the x-axis according to the transcriptome enrichment found for the reference data set of Holstege *et al*. (9). Although the results vary between the different expression data sets, they all follow a general trend. Most notably, the composition of Ala increases by about 30 to 40 % whereas the composition of Asn decreases by ~ 20 %. The transcriptome is also significantly enriched in Gly and the positively charged amino acids, Arg and Lys.

Part B shows the transcriptome enrichment calculated for the cDNA microarray expression data of the diauxic shift in yeast (5). The data from this experiment is primarily used for the measurement of expression level *changes* and we show the transcriptome enrichment only for purely illustrative purposes. Here we use the red fluorescence intensity minus the background intensity as measured by DeRisi et al. (5) as a crude approximation of the absolute expression level of a given ORF. We look at both time point 1 (fermentation) and time point 7 (respiration) of the experiment.

Figure 2, Dependence of Expression Level on Gene Length.

We plotted protein length versus expression level for the reference data set of Holstege et al. (9). (For the other data sets, see http://bioinfo.mbb.yale.edu/genome/expression).

Each point on the graph represents one ORF and the axes of the graph are on a logarithmic scale. It is obvious that there is no strong positive or negative correlation between protein length and expression level (correlation coefficient is -0.16). However, it seems that protein length is related to the upper limit of the expression level possible for a given group of ORFs. A rough way to characterize this upper limit is to fit the hypberbolic function $L = (K/E)^A$ through the maximum protein lengths L (in units of amino acid residues) at given expression levels E (in units of transcripts per cell); K and A are constants. For the reference set of Holstege et al., parameter A was determined to be about 0.7 and K to be about $4.7 \cdot 10^4$. The table below lists the values for parameters A and K for all data sets.

Data Set	Holstege et al. (9)	Jelinsky et al. (11)	Roth et al., mat. type a (10)	Roth et al., mat. type alpha (10)	Roth et al., galactose (10)	Roth et al., heat shock (10)	SAGE, G2/M phase (1)	SAGE, log phase (1)	SAGE, S phase (1)
A	0.72	0.59	0.61	0.63	0.65	0.68	0.51	0.55	0.52
K	$4.7 \cdot 10^4$	$2.8 \cdot 10^5$	$7.4 \cdot 10^4$	$5.0 \cdot 10^4$	$3.4 \cdot 10^4$	$2.8 \cdot 10^4$	$2.6 \cdot 10^6$	$1.2 \cdot 10^4$	$1.7 \cdot 10^6$

As can be seen in figure 2 (especially on the left-hand side), the expression data is discrete, which makes the functional fit possible; this is due to the resolution limit of the experimental data (0.1 copies per cell for the data set of Holstege et al. (9)). Different data discretizations affect the slope of the straight line somewhat (that is, parameter A), but the general trend -- protein length is related to maximum expression level -- can always be observed.

Figure 3, Transcriptome Enrichment of Structural Classes

Part A shows the transcriptome enrichment of membrane proteins compared with soluble ones. We identified yeast ORFs coding for membrane proteins using the GES hydrophobicity scale (33). The values from this scale in a window of size 20 (the typical size of a transmembrane helix) were averaged and then compared against a cutoff of -1 kcal/mole. A value under this cutoff was taken to indicate the existence of a transmembrane helix. Initial hydrophobic stretches corresponding to signal sequences for membrane insertion were excluded. (These have the pattern of a charged residue within the first seven, followed by a stretch of 14 with an average hydrophobicity under the cutoff.) These parameters have been used, tested, and refined in surveys of membrane proteins in genomes (34, 35, 36, 20). "Sure" membrane proteins had at least one TM-segment with an average hydrophobicity less than -2 kcal/mole. "Marginal" membrane proteins had GES-identified TM-helices but did not fulfill this "MinH" criteria. This approach is similar to Boyd & Beckwith's MaxH criteria (37) and to the approach of Klein et al. (38).

Part B shows the transcriptome enrichment of soluble fold classes. The fold classes are sorted along the x-axis in the order of increasing transcriptome enrichment for the reference data set. To assign folds to the yeast genome, we followed a protocol similar to the one described previously, matching the PDB structure database against the yeast genome using both PSI-blast and FASTA (31, 39, 40, 23, 41, 42, 43). We used the following parameters in our PSI-blast searches: an inclusion threshold (h) of 10^{-5}, the

maximum number of iterations (j) of 10, and a final e-value cutoff of 10^{-4}. These parameters are somewhat stricter than those used in previous PSI-blast analyses -- e.g., our inclusion parameter is about 1/20 of that in Teichmann et al. (1998) (44) (who used $h = 5 \cdot 10^{-4}$ and $j = 20$); the inclusion parameter determines to which degree further homologs of a sequence are included at the next PSI-blast iteration. (A higher value leads to the inclusion of more sequences and greater coverage. However, an inclusion too high can lead to a corrupted profile and spurious matches.) We monitored our parameter settings by looking at how many domains were assigned to two different protein folds (obviously an erroneous assignment) and made sure this number was virtually nil. For the FASTA searches we used the usual e-value cutoff of 10^{-2} used in previous analyses (43).

Figure 4, The top-ten most highly expressed protein folds in yeast

The folds are listed from top to bottom in the order of decreasing transcriptome composition for the reference data set of Holstege et al. (9). In the left half of the table we first list the protein fold, then its fold class and the identifier for a representative structure in the Protein Data Bank (PDB) (22). In the columns "genome", "transcriptome" and "transcriptome enrichment", we list the genome and transcriptome compositions and the transcriptome enrichment of each fold. The right half of the table shows the rankings of each fold based on its transcriptome composition in the different expression data sets. For comparison we also show the ranking in the genome -- i.e. based purely on the level of duplication within the genome. The genome compositions are calculated with respect to the ORFs for which expression levels in the reference data set exist. Their exact

fractions in the transcriptome are listed for the reference data set and are schematized with rankings for the other sets. The ranking of the most common folds in the transcriptome and the genome are different. For instance, the most common transcriptome fold by a large margin (8% vs. 5% for the 2nd ranked fold) is the TIM-barrel, which is only ranked fifth in the genome.

—

* The second domain of this two-domain protein represents a G3P dehydrogenase-like fold.

Figure 5, Transcriptome Enrichment of MIPS categories

To analyze the transcriptome in terms of broad functional categories, we categorized the yeast ORFs using the functional categorization provided by MIPS (28, 29, 30). The functional categories are sorted along the x-axis in the order of increasing transcriptome enrichment for the reference data set.

Table 1, Overview of Data and Methods

Part A, Overview of the expression data sets used in our analysis.

The columns "reference" and "URL" provide the literature reference and the internet address of the data sets. Column "# ORFs covered" shows for how many different yeast ORFs expression levels were measured in the respective experiment. The column labeled "technology" shows the technology with which the data sets were obtained. All the data from the expression experiments as well as the soluble and membrane fold assignments were homogenized and relationalized and stored in a simple database.

We focused more on data from DNA chips (9, 10, 11) and the SAGE technology (1) than that from cDNA microarray experiments (5) since the former techniques allow a better measurement of the absolute number of transcript copies for a gene. In presenting our data, we decided, for convenience, to use the data set generated by Holstege et al. (9) as the main reference. For the SAGE data set we only considered SAGE tags that occur at most once per genome and fall into an ORF (rather than upstream regions) (1).

Part B, the general approach in our calculations

First, we calculate the genome composition of a specific feature *F*, *G(F)*. Then, we compute the composition of feature F in the transcriptome, *T(F)*; this is achieved by weighting the count of feature *F* with the expression level e_i of the corresponding ORF *i*. Finally, *D(F)* yields the transcriptome enrichment of feature *F*, the relative difference between its transcriptome and genome compositions. The table shows the calculation of the transcriptome enrichment *D(F)* for the amino acid Ala and the TIM-barrel fold as examples based on the data set by Holstege et al (9). To be consistent, we include only those ORFs in our calculations (of both the transcriptome and the genome composition) for which an expression level e_i exists. Because the set of ORFs for which expression levels were measured vary between the different experiments (see part A), different genome compositions are obtained for each experiment. However, these differences are generally very small and do not influence the results significantly.

Figure 1 a

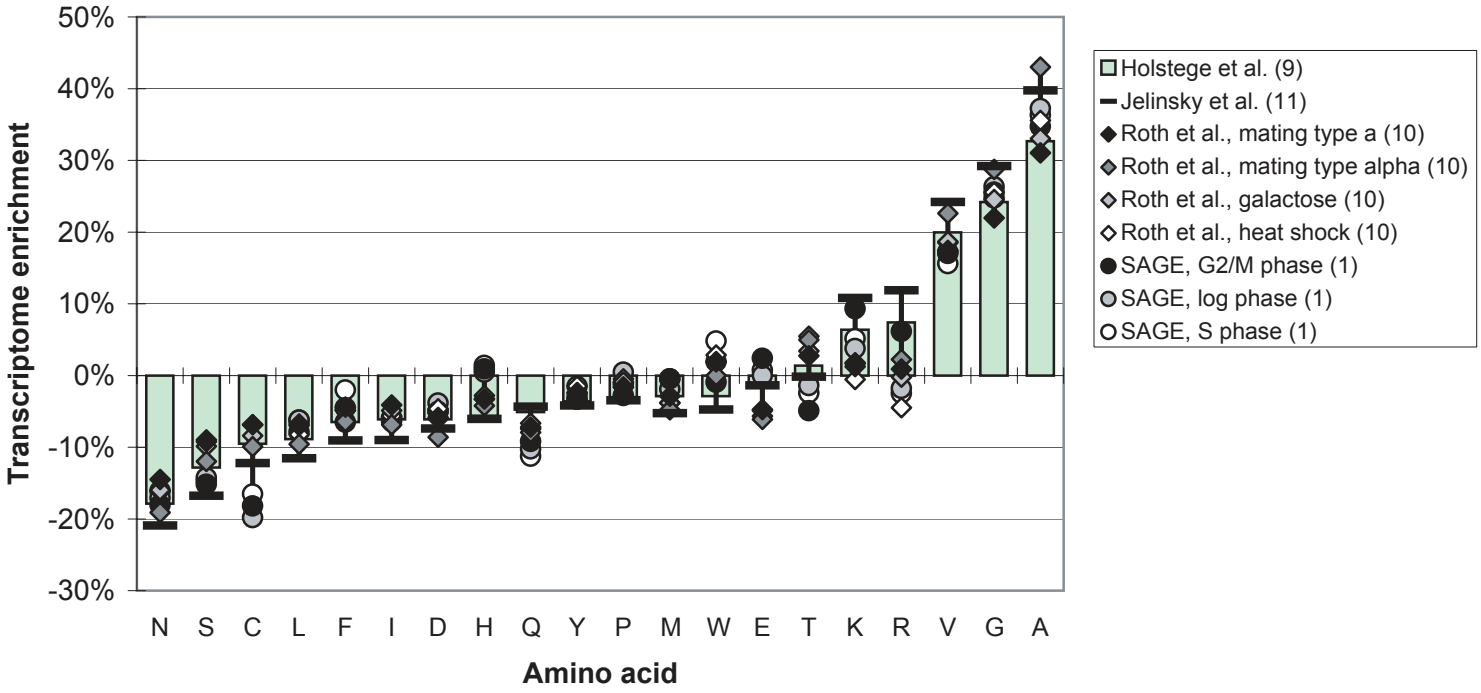


Figure 1 b

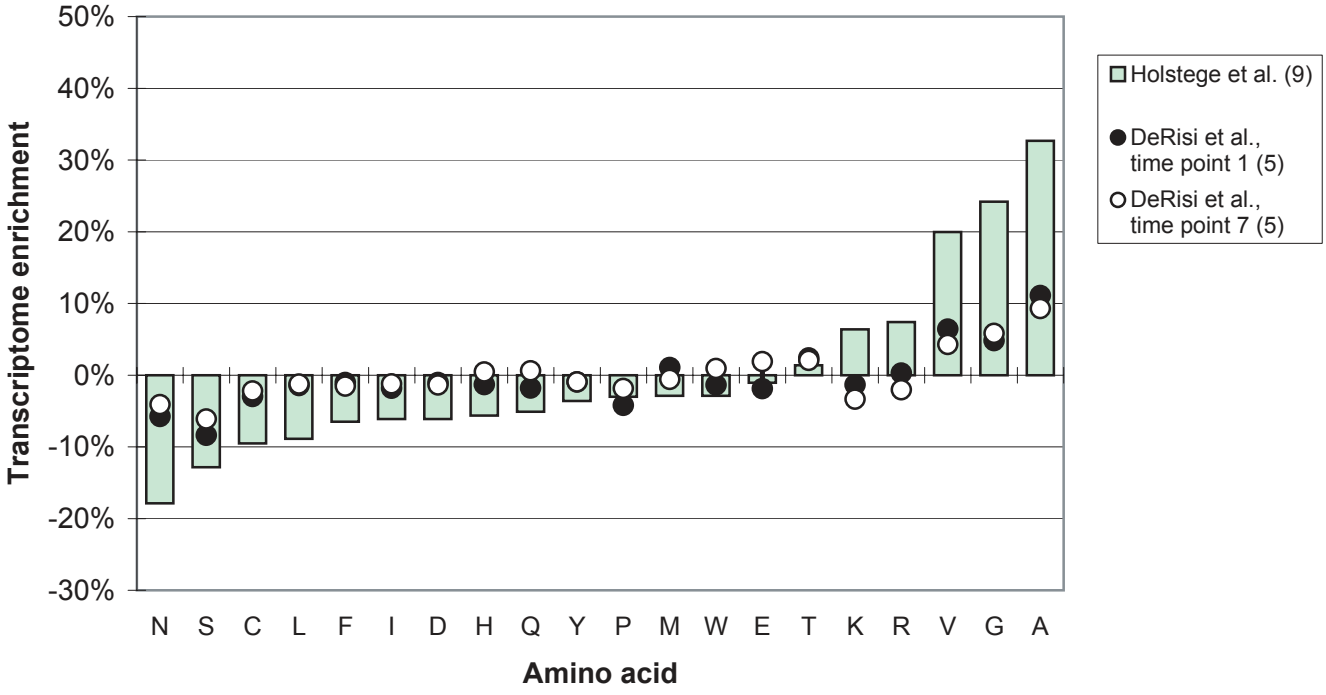


Figure 2

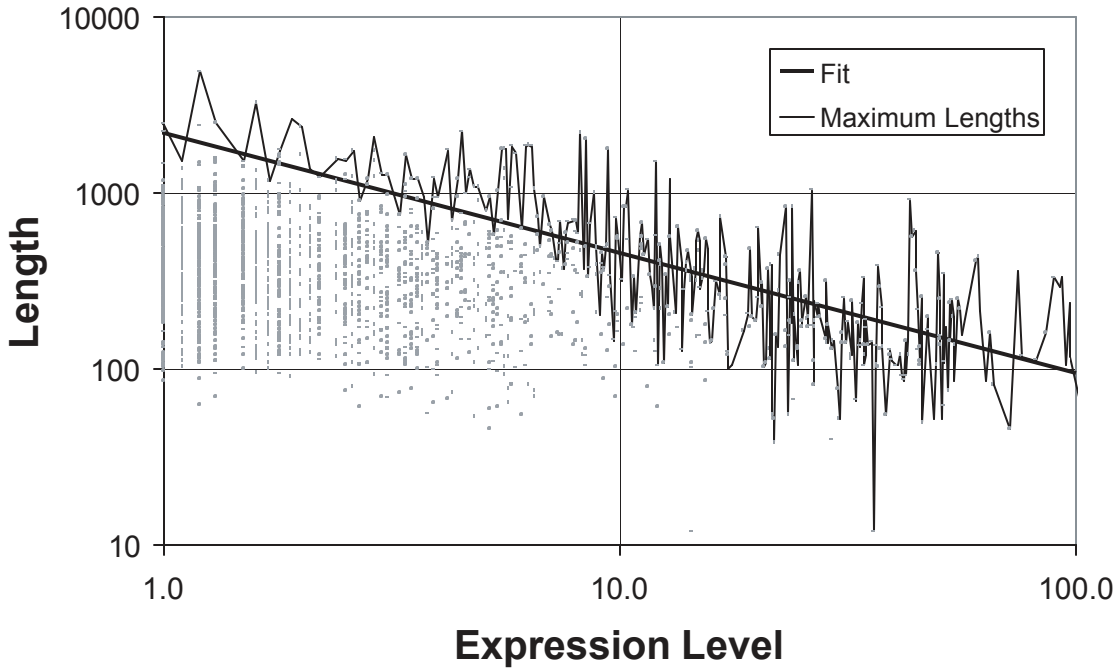

10000
1000
100
10
Length
1.0
10.0
100.0
Expression Level
Fit
Maximum Lengths

Figure 3 a

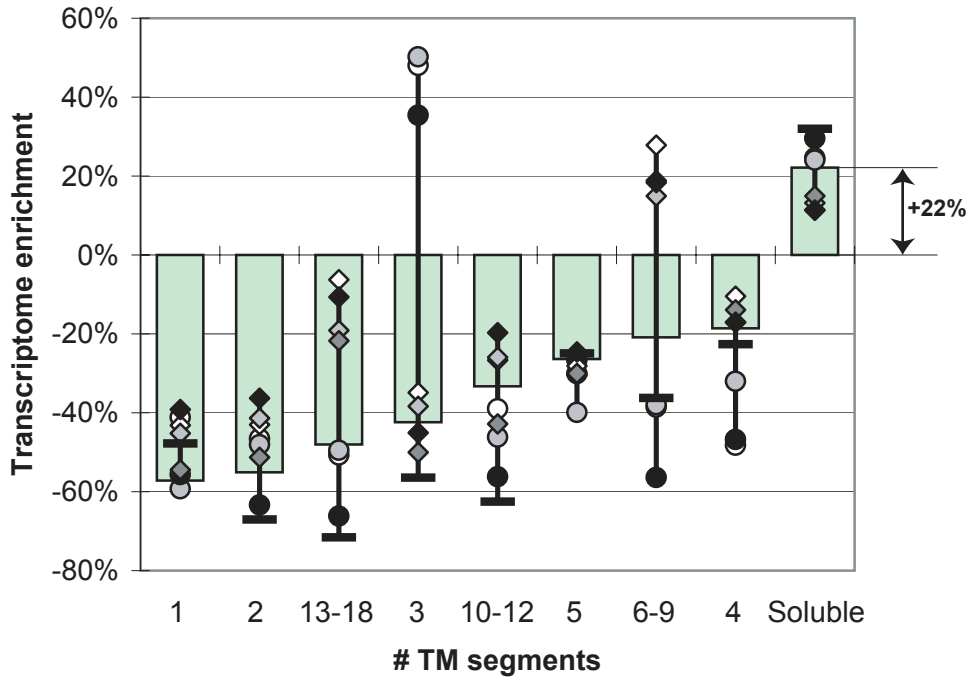


Figure 3 b

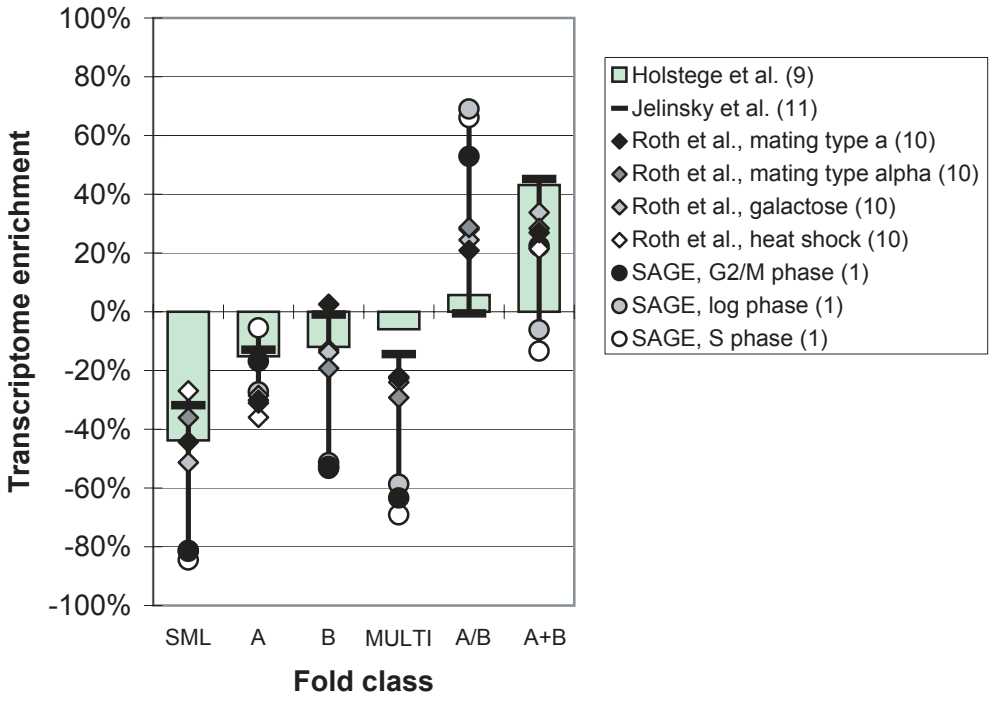
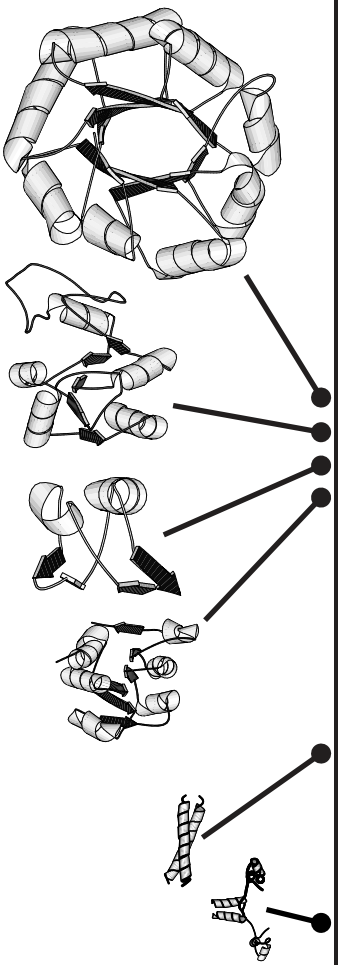


Figure 4



Fold	Fold Class	Rep. PDB	Genome composition [%]	Transcriptome composition [%]	Transcriptome Enrichment [%]	Rank: Genome	Holstege et al. (9)	Jelinsky et al. (11)	Roth et al., mating type a (10)	Roth et al., mating type alpha (10)	Roth et al., galactose (10)	Roth et al., heat shock (10)	SAGE, G2/M phase (1)	SAGE, log phase (1)	SAGE, S phase (1)
TIM barrel	/	1byb	4.2	8.3	+98	5	1	1	1	1	1	1	1	1	1
P-loop NTP hydrolases	/	1gky	5.8	5.2	-11	3	2	2	4	4	4	5	5	6	7
Ferredoxin like	+	1fxd	3.9	3.4	-14	6	3	7	11	9	8	10	4	10	11
Rossmann fold	/	1xel	3.3	3.3	0	8	4	3	3	3	2	2	19	15	9
7-bladed -propeller		1mda	6.4	2.9	-55	2	5	4	5	6	6	7	9	9	16
superhelix		2bct	4.4	2.7	-37	4	6	11	15	16	12	12	8	5	8
Thioredoxin fold	/	2trx	1.7	2.7	+63	14	7	6	8	2	5	4	11	11	6
G3P dehydrogenase-like	+	1drw*	0.2	2.7	+1316	81	8	12	2	5	3	3	35	19	31
grasp	+	1igd	0.6	2.6	+348	36	9	10	21	9	18	21	83	123	121
HSP70 C-term. fragment	multi	1dky	0.8	2.6	+231	31	10	16	17	11	16	12	48	25	56
Leu-zipper fold		1zta	3.8	2.1	-46	7	15	8	14	21	15	19	21	20	33
Protein kinases (cat. core)	multi	1hcl	6.8	1.6	-77	1	18	19	9	16	11	15	13	15	17
/ hydrolases	/	2ace	2.2	0.9	-62	10	32	31	25	26	21	23	26	26	27
Zn2/C6 DNA-bind. dom.	sml	1aw6	2.6	0.3	-89	9	75	94	27	50	32	40	48	39	50

Figure 5

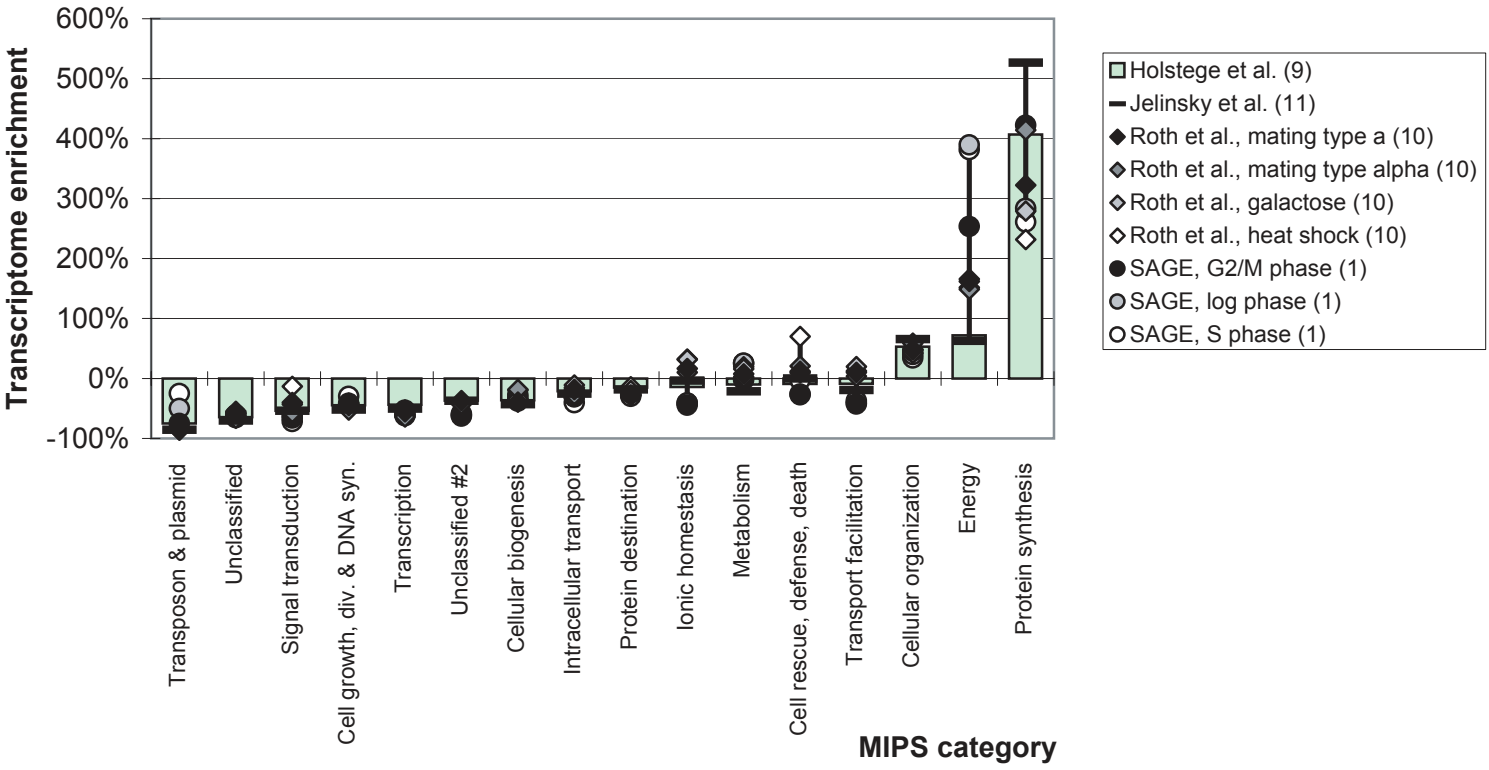

600%
500%
400%
300%
200%
100%
0%
-100%
Transcriptome enrichment
Transposon & plasmid
Unclassified
Signal transduction
Cell growth, div. & DNA syn.
Transcription
Unclassified #2
Cellular biogenesis
Intracellular transport
Protein destination
Ionic homestasis
Metabolism
Cell rescue, defense, death
Transport facilitation
Cellular organization
Energy
Protein synthesis
MIPS category
Holstege et al. (9)
Jelinsky et al. (11)
Roth et al., mating type a (10)
Roth et al., mating type alpha (10)
Roth et al., galactose (10)
Roth et al., heat shock (10)
SAGE, G2/M phase (1)
SAGE, log phase (1)
SAGE, S phase (1)

Table 1 a

Technology	*Reference*	*URL (http://…)*	*# ORFs covered*	*Description*
Gene chip	Holstege, Jennings et al. 1998 (9)	web.wi.mit.edu/young/ expression	5455	Vegetative growth
	Jelinsky and Samson 1999 (11)	www.hsph.harvard.edu/ geneexpression	6281	Response to alkylating agent
	Roth, Hughes et al. 1998 (10)	arep.med.harvard.edu/ mrnadata/expression.html	6214	Mat. type a, glucose (30 deg. C)
				Mat. type α, glucose (30 deg. C)
				Mat. type a, galactose (30 deg. C)
				Mat. type a, glucose (39 deg. C)
Serial Analysis of Gene Expression (SAGE)	Velculescu, Zhang et al. 1997 (1)	www.sagenet.org/ yeast/yeastintro.htm	3005	Yeast transcriptome – G2/M phase
				Log phase
				S phase
cDNA microarray	DeRisi, Iyer et al. 1997 (5)	cmgm.stanford.edu/ pbrown/explore	6153	Time course of diauxic shift

Table 1 b

$$G(F) = \frac{\sum_{\text{orf } i} n_i(F)}{\sum_F \sum_{\text{orf } i} n_i(F)} \qquad T(F) = \frac{\sum_{\text{orf } i} e_i n_i(F)}{\sum_F \sum_{\text{orf } i} e_i n_i(F)} \qquad D(F) = \frac{T(F) - G(F)}{G(F)}$$

Feature	$\sum_{\text{orf } i} n_i(F)$	$\sum_F \sum_{\text{orf } i} n_i(F)$	*G(F)*	$\sum_{\text{orf } i} e_i n_i(F)$	$\sum_F \sum_{\text{orf } i} e_i n_i(F)$	*T(F)*	*D(F)*
Amino acids, in particular Ala	Number of Ala in yeast genome	Number of amino acids in yeast	Genome composition of Ala in yeast	Number of Ala weighted by expression	Number of amino acids weighted by expression	Transcriptome composition of Ala in yeast	Relative enrichment of Ala in transcriptome
Numbers	141890	2574876	5.5%	347801	4758441	7.3%	32.7%
Folds, in particular the TIM-barrel (3.1)	Number of TIM-barrel fold matches in yeast genome	Number of matches with all folds in yeast genome	Genome composition of TIM-barrel fold matches	Number of TIM-barrel fold matches weighted by expression	Number of matches with all folds weighted by expression	Transcriptome composition of TIM-barrel fold matches	Enrichment of TIM-barrel fold matches in transcriptome
Numbers	65	1560	4.2%	389	4709	8.3%	97.8%