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

Dov Greenbaum[3, *], Ronald Jansen[1, *], & Mark Gerstein[1, 2, †]

Department of Molecular Biophysics & Biochemistry[1] Computer Science[2] and Genetics[3]
266 Whitney Avenue, Yale University

PO Box 208114, New Haven, CT 06520
(203) 432-6105, FAX (360) 838-7861

Dov.Greenbaum@yale.edu
Ronald.Jansen@yale.edu
Mark.Gerstein@yale.edu

* These authors contributed equally to this work.
† To whom correspondence should be addressed

ABSTRACT

Motivation

Protein abundance is related to mRNA expression through many different cellular processes. Up to now, there have been conflicting results on how correlated the levels of these two quantities are. Given that expression and abundance data are significantly more complex and noisy than the underlying genomic sequence information, it is reasonable to simplify and average them in terms of broad proteomic categories and features (e.g. functions or secondary structures), for understanding their relationship. Furthermore, it will be essential to integrate, within a common framework, the results of many varied experiments by different investigators. This will allow one to survey the characteristics of highly expressed genes and proteins.

Results To this end, we outline a formalism for merging and scaling many different gene expression and protein abundance data sets into a comprehensive reference set, and we develop an approach for analyzing this in terms of broad categories, such as composition, function, structure and localization. As the various experiments are not always done using the same set of genes, sampling bias becomes a central issue, and our formalism is designed to explicitly show this and correct for it. We apply our formalism to the currently available gene expression and protein abundance data for yeast. Overall, we found substantial agreement between gene expression and protein abundance, in terms of the enrichment of structural and functional categories. This agreement, which was considerably greater than the simple correlation between these quantities for individual genes, reflects the way broad categories collect many individual measurements into simple, robust averages. In particular, we found that in comparison to the population of genes in the yeast genome, the cellular populations of transcripts and proteins (weighted by their respective abundances) were both enriched in: (i) the small amino acids Val, Gly, and Ala; (ii) low molecular weight proteins; (iii) helices and sheets relative to coils; (iv) cytoplasmic proteins relative to nuclear ones; and (v) proteins involved in "protein synthesis," "cell structure," and "energy production".

Supplementary Information http://genecensus.org/expression/translatome

Contact mark.gerstein@yale.edu

INTRODUCTION

With the recent popularity of high-throughput experimentation, biologists have begun to create a large inventory of scientific data (Claverie 1999; Einarson & Golemis 2000; Epstein & Butow 2000; Shapiro & Harris 2000). Much of this has come from expression experiments, partially fueled by the advent and continuous evolution of the microarray and Gene Chip systems. These experiments allow for large scale, comprehensive scans of gene expression within the cell (Schena *et al.* 1995; Eisen & Brown 1999; Ferea & Brown 1999; Lipshutz 1999). Expression data sets are currently the single richest source of information in genomics, and for yeast, expression information now dwarfs that in the sequence alone. However, "theory" has not kept up with experimentation in this area, and how to best interpret the vast amount of data generated by these experiments is still a very open question (Bassett *et al.* 1996; Wittes & Friedman 1999; Zhang 1999; Gerstein & Jansen 2000; Searls 2000; Sherlock 2000).

Genome-wide experimentation has also been used to directly measure the cellular population of proteins (protein abundance). (Anderson & Seilhamer 1997; Futcher *et al.* 1999; Gygi *et al.* 1999; Ross-Macdonald *et al.* 1999) Understanding how protein abundance is related to mRNA transcript levels is essential for interpreting gene expression and also, more generally, for understanding the interactions, structures and functions in a cellular system (Hatzimanikatis *et al.* 1999). Moreover, as protein concentration, rather than transcript population, is the more relevant variable with respect to enzyme activity, it is this quantity that connects genomics to the physical chemistry and dynamics of the cell (Kidd *et* al 2001). Finally, protein abundance levels may become invaluable for diagnostic methods as well as for determining new drug targets (Corthals 2000). High-throughput two-dimensional gel electrophoresis (2-DE), in conjunction with mass spectrometry, has been used to identify proteins that can then be quantified to determine protein abundance (Futcher *et al.* 1999; Gygi *et al.* 1999; Harry *et al.* 2000). Other technologies include using random integration of reporter transposons in yeast (Ross-Macdonald *et al.* 1999), and modifying the microarray concept for use with proteins (Lopez 2000; MacBeath & Schreiber 2000; Nelson *et al.* 2000; Zhu *et al.* 2000).

Gene expression is indirectly related to cellular protein abundance through the process of translation. The cell connects mRNA expression and protein abundance through translational control, which is primarily regulated at the initiation of translation (Lindahl & Hinnebusch 1992; Jackson & Wickens 1997; Day & Tuite 1998; McCarthy 1998). Much of this control is the result of multiple cis-acting elements in the mRNA (Jacobs Anderson & Parker 2000). There are large non-coding regions in each mRNA species devoted to regulation of that mRNA as well as its stability and degradation properties, including 5` and 3` UTRs, uORFs and uAUGs (Vilela *et al.* 1998; Vilela *et al.* 1999; Morris & Geballe 2000).

Previously, we surveyed the population of protein features -- such as folds, amino acid composition, and functions -- in yeast, and a number of the other recently sequenced genomes (Gerstein 1997; Gerstein 1998; Gerstein 1998; Gerstein & Hegyi 1998; Hegyi & Gerstein 1999; Das & Gerstein 2000; Lin & Gerstein 2000). Others have also done related work (Frishman & Mewes 1997; Tatusov *et al.* 1997; Jones 1998; Wallin & von Heijne 1998; Frishman & Mewes 1999; Wolf *et al.* 1999). Recently, we extended this concept to compare the population of features

in the yeast transcriptome to that in the genome (Drawid, et al. 2000, Jansen and Gerstein 2000). Here, we present a new methodology to compare the features of the mRNA expression population with the protein abundance population.

Precise terminology is essential for this comparison to be readily understandable. Unfortunately, one of the terms that immediately come to mind in relation to protein populations, "proteome", has in the past been used inconsistently. In particular, the term proteome can logically be used to describe all the distinctly different proteins in the genome (Qi *et al.* 1996; Cavalcoli *et al.* 1997; Fey *et al.* 1997; Garrels *et al.* 1997; Gaasterland 1999; Jones 1999; Sali 1999; Tekaia *et al.* 1999; Bairoch 2000; Cambillau & Claverie 2000; Doolittle 2000; Pandey & Mann 2000; Rubin *et al.* 2000) and, in this context, it is equivalent to what others may refer to as the coding part of the genome. However, in papers on 2D electrophoresis, it is often used to describe the sum total of proteins in a cell, taking into account the different levels of protein abundance for different proteins (Shevchenko *et al.* 1996; Gygi *et al.* 2000; Lopez 2000; Washburn & Yates 2000). In an effort to be clear, we propose the term "translatome" for this second usage of proteome.

With this definition, we are able to refer compactly to three different cellular populations. These are illustrated in figure 1.

(i) We use the term *genome* when we refer to the population of open reading frames, where each ORF counts once.

(ii) We use the term *transcriptome* when we refer to the population of mRNA transcripts. This term was originally coined by Velculescu et al. (1997). Note that each ORF may give rise to different numbers of transcripts. Consequently, the transcriptome is essentially the same as the genome but with each ORF weighted by its expression level.

(iii) The next level is the cellular population of proteins. As each protein represents a translated transcript, we make an analogy with the term transcriptome and use the term *translatome* as described above to describe this third population. Thus, the translatome is a subset of the genome where each ORF is weighted by its associated level of protein abundance.

Note that one could also less compactly call the translatome a "weighted proteome". However, doing so assumes one of the two aforementioned definitions of proteome. To avoid ambiguity, we studiously avoid the use the proteome altogether in the paper.

Differences between the translatome and the transcriptome exist given that transcripts from different genes can give rise to different numbers of proteins, due to different rates of translation and protein degradation. Post-transcriptional modifications further affect the translatome.

Although there are gene expression and protein abundance data sets for multiple organisms, we have chosen to work specifically on yeast. Besides having its whole genome sequenced (Goffeau 1996), yeast is also a powerful tool in genetics (Carlson 2000) due to, among other things, the two hybrid system, a robust and versatile technique used in discerning protein-protein interactions (Luban & Goff 1995; Young 1998; Ito *et al.* 2000).

In our analysis of the transcriptome and translatome, we focus on global protein features rather than

the comparison of individual genes. Previous analyses have shown that differences between mRNA expression and protein abundance level can be quite dramatic for individual genes. This may either be due to the noise in the data or to fundamental biological processes. However, our analyses show that the variation between transcriptome and translatome is much smaller for global properties that are computed by averaging over the properties of many individual genes.

METHODS

Data Sources Used

For our analysis we culled many divergent data sets, representing protein abundance and mRNA expression experiments and also other sources of genome annotation. These are all summarized in Table 1. Briefly, they included two protein abundance sets, measured via 2-dimensional gel electrophoresis and mass spectrometry. We termed these 2-DE #1 (Gygi *et al.* 1999) and 2-DE #2 (Futcher *et al.* 1999). These sets, while admittedly small in comparison to the size of expression data sets, represent the largest amount of information on protein abundance publicly available at the present. We also apply our methodology, with limited success, to the semi-quantitative Transposon insertion data set that measures the LacZ expression of fusion proteins (Ross-Macdonald *et al.* 1999). Although this set contains many more genes than either of the gel electrophoresis sets, and thus is an appealing source of protein abundance information, the more qualitative nature of the data makes comparisons with other data sets difficult.

Our mRNA expression data came from multiple laboratories that used either Gene Chip or SAGE technology. The Gene Chip sets included the Young Expression Set (Holstege *et al.* 1998), the Church Expression Set (Roth *et al.* 1998) and the Samson Expression Set (Jelinsky & Samson 1999). We used data representing the vegetative state of yeast from all of the above experiments. We also compiled two reference sets to be used in our comparisons, one for protein abundance and another for mRNA expression (summarized below). Finally, we used many different types of genome annotation in our analysis, which are summarized in Table1. In particular, the Munich Information Center for Protein Sequences (MIPS), a site containing a large number of databases (Mewes *et al.* 2000), proved to be an invaluable source of data specifically in regard to functional categories.

Biases in the Data

There is a caveat to the usage of data from high-throughput experimentation (i.e. microarrays and two-dimensional gel electrophoresis). With all high throughput expression studies there always exists the difficulty of maintaining consistent biological and processing conditions across the assay. Moreover, the databases that annotate the specific genes may not always be accurate (Ishii *et al* 2000). Gene chip experiments suffer with regard to cross hybridization and the saturation of probes for the highly expressed genes. SAGE data is not always reliable for assessing ORFs with low expression levels. With regard to 2D gels, although the technology has undergone many improvements since its introduction over a quarter century ago (Klose 1975; O'Farrell 1975), there remain many aspects of the procedure that introduce biases into the data. These include the inability to resolve membrane proteins (approximately 30% of the genome) and basic proteins (Gerstein 1998; Krogh *et al* 2001). Moreover, there exist some biases in the data that, as in any compilation, reflect the tendencies of the investigator. These include the lack of low abundance

proteins (Fey & Larsen 2001; Gygi *et al* 2000; Harry *et al* (2000)) and the differences between labs in sample preparation. In addition, the procedures for identification (i.e. MALDI-TOF) and quantification (i.e. ICAT) (Gygi *et al.* 1999) of the protein spots are much more recent and themselves subject to problems and uncertainties (Haynes & Yates 2000).

We are trying to correct for these biases in our analysis in two ways. First, we create reference mRNA expression and protein abundance datasets as a starting point for our analysis. We achieve this by scaling and averaging different mRNA and protein datasets into a combined reference, in an attempt to obtain a better estimate of the normal expression state of a yeast cell (we explain this procedure in more detail in the following section). This results in a correction of the biases that might be found in individual datasets. Second, in analyzing the reference datasets, we use a formalism and a graphical representation that shows the dependency of the results on the subset of genes for which experimental data is available, thus making sampling or selection biases explicit.

Data Set Scaling

A Reference set for mRNA Expression

With many different mRNA expression data sets available, it is worthwhile to integrate them into a single unified reference set, with the intention of reducing the noise and errors contained in the individual data sets and to obtain a unified estimate of the normal expression state in a cell.

We adopt an iterative scaling and merging formalism, which we summarize below. We present a more detailed review of the methods at the following web site: genecensus.org/expression/translatome.

We start with the values of one Gene Chip data set U_i where i is used throughout as a subscript to denote gene number. We then transform the values of the next Gene Chip data set X_i to Y_i with the following non-linear regression:

$$\min \sum_i (Y_i - U_i)^2 \qquad \text{with } Y_i = AX_i^B$$

where A and B are the parameters of the regression. Note that two Gene Chip sets may not be defined for the same set of genes, so we have to perform the fit only over the genes common to both sets. The motivation for scaling is that the dynamic range of observed expression levels varies somewhat between different data sets, although cell types and growth conditions are very similar. Reasons for disparity may include different calibration procedures for relating fluorescence intensity to a cellular concentration (measured in copies of transcripts per cell) or different protocols for harvesting and reverse-transcribing the cellular mRNA.

We then merge and average the data to create a new reference set V as follows:

$$\text{If} \qquad U_i \text{ and } Y_i \text{ are both defined for gene } i \text{ and } \frac{|Y_i - U_i|}{Y_i + U_i} < \alpha$$

Then $V_i = \frac{1}{2}(Y_i + U_i)$

Else if only Y_i exists, $V_i = Y_i$

Else $V_i = U_i$

As presented above, where only one data set has a value for the corresponding ORF, we incorporated that value and did not exclude it. When both data sets have values for an ORF, we averaged the values if they were within 15% of each other; otherwise, we just stayed with the original chip data set U_i. We used α = 15% in order to prevent outliers from skewing the result. This 15% value is a reasonable threshold for excluding outliers though other values (e.g. 10% or 20%) would give similar results (data not shown). Other data sets are subsequently included in the same procedure, continuing the iteration from the new expression values V_i. The initial iteration starts with the Young Expression Set as U_i since we have the highest confidence in its accuracy.

The SAGE data was not included in the above procedure since it is of a fundamentally different nature. An advantage of the SAGE technology over Gene Chips is that there is no possible signal saturation for high expression levels, as is possible for chips (Futcher *et al.* 1999). Conversely, SAGE values are less reliable for lowly expressed genes since there is a chance that one might not sequence a SAGE tag corresponding to such a gene altogether. Therefore, if after the last iteration, the average Gene Chip expression level V_i was both above a certain threshold β and below the SAGE expression level S_i for the same gene, it was replaced with the SAGE value; otherwise the average Gene Chip value was kept. This gave us our final expression set $\mathbf{w_{mRNA}}$. Our treatment of the SAGE data is modeled after that in Futcher et al. (1999), and like them, we used $\beta = 16$. This incorporation of the SAGE data into the reference data set ensures that the highly expressed outliers are as accurate as possible.

Rather than plain arithmetic averaging, this overall scaling procedure with the α cutoff avoids "artificial averages" that combine very different values for a particular gene. Some expression values might be statistical outliers. In addition, it may be possible that the expression levels of a variety of genes can only be within mutually exclusive ranges or modes, such as when two alternative pathways are switched on or off. Simply averaging these would give values that are less representative of the particular mode values. This situation is analogous to that in averaging together an ensemble of protein structures, say from an NMR structure determination. Each structure in the ensemble could be stereochemically correct, with all side-chain atoms in predefined rotamer configurations. However, an average of all structures in the ensemble could yield one that is stereochemically incorrect if this involved averaging over particular side-chains in different rotameric states.

With regard to our regression analysis, we have investigated both non-linear and linear fits but found a non-linear procedure to be more advantageous. The non-linear relationship between different expression datasets perhaps reflects saturation in one or more of the gene chips -- not an uncommon phenomenon. This non-linearity is immediately evident on scatter plots of two datasets against one another (see website). Accordingly, the non-linear fit produces a smaller residual than the linear fit: 98297 (non-linear) versus 122182 (linear) for the scaling of the Church dataset and 59828 (non-linear) versus 67462 (linear) for the Samson dataset.

A Reference Set for Protein abundance

We followed a similar procedure to calculate a reference protein abundance set from the two gel electrophoresis data sets. We first scaled the two data sets against the mRNA expression reference data set, getting regression parameters C_j and D_j:

$$\min \sum_i \left(P_{i,j} - C_j w_{mRNA,i}^{D_j} \right)^2$$

where the subscript j indicates the data set 2-DE #1 or 2-DE #2 respectively; $P_{i,j}$ is the protein abundance value in data set j, and $w_{mRNA,i}$ the corresponding reference expression value, and C_j and D_j are the parameters of the non-linear regression.

Using these parameters, we transformed the values of set 2-DE #2 onto 2-DE #1. Then we combined both sets into the reference protein set $\mathbf{w_{Prot}}$ by averaging them, if both values existed. Otherwise, by using the existing value, viz:

$$Q_{i,2} = C_1 \left(\frac{P_{i,2}}{C_2} \right)^{D_1/D_2}$$

$w_{Prot,i} = (P_{i,1} + Q_{i,2})/2$ if both $P_{i,1}$ and $Q_{i,2}$ exist.
Else if only $P_{i,1}$ exists, $w_{Prot,i} = P_{i,1}$
Else if $Q_{i,2}$ exists, $w_{Prot,i} = Q_{i,2}$.

Comparison of mRNA expression and protein abundance

Figure 2 shows a comparison of our two reference data sets for transcripts and proteins on a log-log graph. The correlation coefficient is 0.67. A previous study (Futcher *et al.* 1999), in which the data set 2-DE #2 was investigated, reported a higher correlation coefficient of 0.76. The disparity may be due to the fact that we are looking at a larger number of points. Inspection of Figure 2 also shows that the correlation for the data values, which were derived from averaging values from both 2-DE sets, is larger. It should be emphasized that there are many limitations in this analysis as both 2-DE sets represent relatively homogenous sets of proteins and there are only a small number of proteins in each set.

Figure 2b shows the outliers from Figure 2a from both above and below the dashed line. These outliers are representative of those genes for which their mRNA expression differs significantly from their protein abundance (i.e. either there is little mRNA expression yet significant protein abundance or significant mRNA expression yet minimal protein abundance). For each, we present a description of its function. With one exception all outliers are associated with the MIPS category: cellular organization (MIPS category 30).

Enrichment of Features

Formalism

Figure 2 focuses on individual proteins. In the next part of our analysis, we want to group a number of proteins together into various categories based on common features and characterize those features that are enriched in one population relative to another, i.e. the translatome population of proteins as measured by 2D gels relative to the transcriptome population of transcripts or the genome population of genes. To this end, we set up a formalism that could be applied universally to all the attributes that we were interested in. Due to the limitations of the experiments, the translatome, transcriptome, and genome populations are defined on different sets of genes, and sometimes we want to remove this "selection bias" by forcing them to be compared on exactly the same set of genes. This is a key aspect of our formalism as presented in figure 1.

We call an entity like [**w**, *G*] a "population", where *G* is a set describing a particular selection of genes from the genome and **w** is vector of weights associated with each element of this population. In particular, we focus on three main populations here:

(i) [**1**,G_{Gen}] is the population of genes in the genome, all 6280 genes weighted once (**w** = **1**).

(ii) [$\mathbf{w_{mRNA}}$, G_{mRNA}] is the observed population of the transcripts in the transcriptome, i.e. the 6249 genes in the reference expression set weighted by their reference expression value.

(iii)[$\mathbf{w_{Prot}}$, G_{Prot}] is the observed cellular population of the proteins in the translatome, i.e. the 181 genes in the reference abundance set weighted by their reference abundance value.

(The set of genes in the genome G_{Gen} is approximately equal to the genes in set G_{mRNA}, such that we can use both symbols interchangeably.) We can also use this notation to describe specific experiments -- e.g. [$\mathbf{w_{lacZ}}$, G_{lacZ}] describes the gene set and weights relating to the Transposon Abundance set.

Furthermore, we define F_j as the value of a feature *F* in ORF *j*. For example, *F* could be the composition of leucine (a real number) or a binary value (0 or 1) indicating whether an ORF contains a trans-membrane segment. Given these definitions, the weighted average of feature *F* in population [**w**, *G*] is:

$$\mu(F,[\mathbf{w},G]) \equiv \frac{\sum_{j\in G} w_j F_j}{\sum_{j\in G} w_j}$$

The weighted averages of two populations [**w**, *G*] and [**v**, *S*] can be compared by simply looking at their relative difference Δ:

$$\Delta(F,[\mathbf{v},S],[\mathbf{w},G]) = \frac{\mu(F,[\mathbf{v},S]) - \mu(F,[\mathbf{w},G])}{\mu(F,[\mathbf{w},G])}$$

where **v** and **w** are weights for the sets of ORFs *S* and *G* respectively. We call Δ the "enrichment" of feature *F* because it indicates whether *F* is enriched (if Δ is positive) or depleted (if Δ is

negative) in population [**v**, *S*] relative to [**w**, *G*].

Usually, the gene set *G* is defined by the particular experiment, for which the weight **w** was measured. However, it is also possible to combine the gene set associated with one experiment with expression levels from another set. One may want to do this to compute the enrichment only on the genes common to both populations, for which there are defined values for both **w** and **v**, viz: , Δ(F, [v, S ∩ G],[w, S ∩ G]). In practice, this is most relevant for comparing G_{Prot} and G_{mRNA}. Since G_{Prot} is completely a subset of G_{mRNA}, we need not explicitly deal with intersections if we calculate all statistics directly over G_{Prot}.

One can adjust the weight vectors to take into account different types of averaging. For instance, when computing the amino acid composition ($F = aa$) from the amino acid compositions of individual ORFs $F_j = aa_j \ (\forall j \in G)$, we weight by ORF length. In the case of expression weights, we have:

$$w_j = N_j \, w_{mRNA,j} \qquad \forall j \in G$$

where N_j is a measure of the length of ORF *j* (such as the number of amino acids.)

On the other hand, when computing the average molecular weight per amino acid, we need to normalize by the number of amino acids per ORF, which is equivalent to choosing the following weights:

$$w_j = \frac{w_{mRNA,j}}{N_j} \qquad \forall j \in G$$

Application of Methodology to Quantitative Abundance Sets

Having defined our formalism, we applied it to a diverse set of protein features in yeast.

Amino Acid Enrichment

As shown in Figure 3a, we used our methodology to measure the enrichment of individual amino acids in both the translatome and the transcriptome relative to the genome. The horizontal axis lists the amino acids while the vertical axis shows their percent enrichment. We list enrichments for both the reference protein abundance and mRNA expression sets in relation to the genome population. We found that three amino acids -- Valine, Glycine and Alanine -- were consistently enriched in both transcriptome and translatome populations.

In Figure 3a we compare different gene sets. In Figure 3b we focus mainly on the variation in enrichments when all the comparisons are restricted to the set of 181 genes ($G_{Prot} \cap G_{mRNA} = G_{Prot}$) common to all data sets. Thus, the differences between the populations now only reflect the effects of differential transcription of certain genes and differential translation of certain transcripts. We find here an enrichment specifically of Cysteine in the translatome in relation to the transcriptome. This enrichment may be the result of the stability associated with sulfur bridges.

To measure the statistical significance of the results on amino acid enrichment, we have performed a control analysis on a randomized dataset (Figure 3D). We randomly permutated the expression values of the ORFs 1000 times and then recomputed the enrichments. This allowed us to compute distributions for the amino acid enrichments and, from integrating these, one-sided p-values indicating the significance of the observed enrichments.

Biomass Enrichment

A corollary to amino acid enrichments is the determination of the average biomass of the transcriptome and translatome populations. We show this in Figure 3C. We found that the average molecular weight of a protein in both populations was, on average, lower than in the genome population. These preliminary observations suggest a cell preference to use less energetically expensive proteins for those that are highly transcribed or translated. However, we also found that the average molecular weight *per amino acid* differed much less between the transcriptome and the translatome on the one hand, and the genome on the other hand (though it was still slightly less). This finding indicates that lower molecular weights in the translatome and transcriptome populations relative to the genome population are predominantly due to greater expression of shorter proteins rather than the incorporation of smaller amino acids.

Secondary Structure Composition

We also used our methodology to study the enrichment of secondary-structural features. Secondary structural annotation was derived from structure prediction applied uniformly to all the ORFs in the yeast genome as described in Table 1. As shown in Figure 4A, all three populations – genome, transcriptome, and translatome – had a fairly similar composition of secondary structures -- sheets, helices, and coils. The differences between populations were marginal and based only on the small subset of genes. They do, though, point to a possible trend of depletion of random coils relative to alpha helices and beta sheets in the transcriptome and translatome.

We also found that transmembrane proteins were significantly depleted in the transcriptome (see website). To identify transmembrane (TM) proteins, we used the GES hydrophobicity scale as described previously (see caption to Table 1 (Gerstein 1998)). These results are consistent with our previous analyses (Jansen & Gerstein 2000). This analysis could not be extended to the translatome because the 181 genes in the protein abundance data set (G_{Prot}) do not contain any membrane proteins, which are difficult to detect in gel electrophoresis (Molloy 2000).

Subcellular Localization

A generalization of the transmembrane protein analysis is subcellular localization. We looked into the enrichment of proteins associated with the various subcellular compartments. This is shown in Figure 4C. For clarity, we divided the cell into five distinct subcellular compartments, as described in Table 1. We found that, in comparison to the genome, both the transcriptome and translatome are enriched in cytoplasmic proteins. This is true whether we make our comparisons in relation to the relatively large reference mRNA expression set or the smaller reference protein abundance set. As figure 4C shows, the 2D gel experiments are clearly biased towards proteins from the cytoplasm. However, in the biased subset G_{prot} transcription and translation lead to an even higher fraction of cytoplasmic proteins in the translatome.

Functional Categories

Finally, we compared the enrichment of various functional categories in both the translatome and the transcriptome (see Figure 4B). This gives us a broad yet informative view of the cell as a whole. As described in Table 1, we used the top-level of the MIPS scheme for the functional category definitions (Mewes *et al.* 2000). We found broad differences between the various populations, with some of the functional categories showing strikingly high enrichments. In particular, we found enrichments of the "cellular organization," "protein synthesis," and "energy production" categories.

Application to Semi-quantitative Protein Abundance Data Sets

We also tried to extend our methodology to cope with the semi-quantitative Transposon set. The qualitative nature of the set makes it impossible to compute statistical relationships between mRNA and protein populations as we did for both the 2D gel sets. We briefly summarize our approach.

Many ORFs in the Transposon dataset had multiple, sometimes inconsistent, measurements ranging from one (background) to four (strong) for various different transposon insertions. We took only those 450 ORFs that consistently yielded either background or strong. We then used this set in a binary fashion, interpreting an ORF as either on or off. We show the enrichments of amino acids computed from this filtered Transposon Abundance Set in Figure 3A. Overall, the enrichments from this set seemed to be attenuated in comparison to either the mRNA expression or protein abundance data.

Discussion and Conclusion

We developed a methodology for integrating many different types of gene expression and protein abundance into a common framework and applied this to a preliminary analysis of yeast. In particular, we developed a procedure for scaling and merging different mRNA and protein sets together and then computing the enrichment of various proteomic features in the population of transcripts and proteins implied by these scaled sets. We showed that by analyzing broad categories instead of individual noisy data points, we could find logical trends in the underlying data.

The comparison of the translatome with the transcriptome and the genome helps to better understand cellular processes. For this purpose, we compiled two reference sets, the mRNA reference expression set integrated from various Gene Chip and SAGE experiments, and the protein reference abundance set, collected from published 2D gel electrophoresis experiments. Our reference sets proved useful for our analysis of the composition and enrichments of protein features in the various stages of gene expression. We found many similar trends for general protein categories between these two sets.

To compare the translatome and the transcriptome, we devised a formalism to measure enrichments of data sets. With this formalism we measured the enrichments of amino acids, protein function

and secondary structures in the vegetative yeast cell. Other comparisons included looking at average biomasses, looking into subcellular localizations and a direct comparison of mRNA expression vs. protein abundance.

Overall Transcriptome and Translatome Similarity: Outliers Against Trend

The overall similarity we find between transcriptome and translatome contrasts somewhat with the weak correlation between mRNA expression and gene abundance as shown in figure 2 and reported previously (Futcher *et al.* 1999; Gygi *et al.* 1999). This reflects the way our system of overall categories collects many proteins into robust averages. It shows that variation between proteins is not systematic with respect to the categories. For example, individual transcription factors might have higher or lower protein abundance than one expects from their mRNA expression, but the category "transcription factors" as a whole has a similar representation in the transcriptome and translatome.

We used the reference data sets to compare mRNA expression and protein abundance for the 181 genes shared between the two sets -- the largest such comparison. While we found an overall correlation between the two data sets, indicating that mRNA expression may be closely related to protein abundance, we found some genes that bucked the trend. Possible explanations for the aberrant behavior of some of these outliers are presented. Those outliers that have higher levels of protein abundance than expected from their mRNA expression are dominated by alcohol dehydrogenases and Glyceraldehyde-3-phosphate (G3P) dehydrogenases. It is known that G3P dehyderogenase forms a bienzyme complex with alcohol dehydrogenase, thus, the similar abundance pattern of these two enzymes can be rationalized (Batke *et al.* 1992). Alcohol dehydrogenase is also a stress induced protein in many organisms (Matton *et al.* 1990; An *et al.* 1991; Millar *et al.* 1994), induced into action when the cell undergoes trauma, thus perhaps translated to a higher degree prophylactically (although the expression pattern of another stress-induced protein (HSP70) shows that this is not always the case). Translation-related proteins are more prominent in the outliers, with lower protein abundance than expected from mRNA expression.

While it is known that multiple features of an individual mRNA influence its expression and regulation, it is presently not clearly understood how. There are many non-coding regions in each mRNA species that are responsible for this regulation. These include upstream AUG codons (uAUGs), both 3' and 5' untranslated regions, upstream open reading frames (uORFs) and the overall secondary structure of mRNA. Presently it is unclear how these act to exert their control (Morris & Geballe 2000).

One might conceive of using "outliers" with significantly different transcriptional and translational behavior to find consensus regulatory sequences. One possible method would involve using predicted mRNA structures (Jaeger *et al.* 1990; Zuker 2000) to find consensus structural elements in these outliers. In particular, it might be worthwhile to investigate the secondary mRNA structure, to which the yeast translational machinery is known to be sensitive (McCarthy 1998).

The regulation of mRNA stability is certainly an additional factor causing strong disparities between gene expression and protein abundance. Presently, there are many structures within

mRNA that are thought to influence stability including, among others, stem loops, UTRs premature stops and uORFS (Klaff *et al.* 1996).

Overall Transcriptome and Translatome Similarity: Consistent Enrichments

We found the enrichments relative to the genome to be consistent between the translatome and the transcriptome. In particular we found that the amino acids Valine, Glycine and Alanine -- all relatively small amino acids -- are significantly enriched in both populations in comparison to the genome population. These results coincide with the previous conclusion that those amino acids are also the most highly abundant amino acids in soluble proteins (Nauchitel & Somorjai 1994). Conversely we found that Cysteine, Serine, Asparagine and Arginine were markedly depleted. Our transcriptome enrichments using the reference set were similar to results attained previously using individual mRNA expression data sets (Jansen & Gerstein 2000). In addition, we found that the translatome and the transcriptome both have lower molecular weight proteins in relation to the genome.

Furthermore, we found, in comparison to the genome population, that the translatome and transcriptome had a depletion of random coils, a relatively less structurally complex and, as such, less stable protein structure, to alpha helices and beta sheets. These results are from a small and potentially biased subset of proteins and so, in of themselves, may not be informative. Yet, it is possible that they point to a logical trend that may result from the cellular preference for stability and structural rigidity through more regular secondary structures (helices and sheets).

In relation to functional categories, we found three trends that were particularly notable: (i) The "cellular organization," "protein synthesis," and "energy production" categories were increasingly enriched as we moved from genome to transcriptome to translatome. This finding was true for either of the gene sets and reflects the great abundance of structural proteins, such as actin, and, in the case of the transcriptome, ribosomal proteins. (In the protein abundance set G_{Prot} ribosomal proteins are rather underrepresented.) (ii) Proteins with "unclassified function" are significantly depleted in the transcriptome and the translatome in relation to the genome, perhaps reflecting a bias against studying them. (iii) Proteins in the "transcription" and "cell growth, cell division, and DNA synthesis" categories were consistently depleted in the transcriptome and translatome population relative to the genome. This perhaps reflects the fact that many of these proteins, such as transcription factors, act as "switches." While many copies are needed in the genome to give different specificities, only small quantities of the protein are necessary to activate or deactivate a process. These results concur with previous calculations (Jansen & Gerstein 2000) wherein we found the transcriptome is enriched specifically with proteins involved in protein synthesis and energy.

As opposed to the genome population, where there is a wide distribution of products in all cellular compartments, mainly cytoplasmic proteins dominate the translatome and transcriptome. For instance, while the genome data set has the largest allocation of genes going to the nucleus, the bulk of the translatome and transcriptome populations are localized to the cytoplasm. Part of this effect may also be due to the gel-electrophoresis experimental process that favors the higher expressing cytoplasmic proteins, although a similar effect can clearly be observed in the

transcriptome data set, which does not have this experimental bias. This may be related to the enrichment of functional categories that are connected to cytoplasmic proteins, such as "protein synthesis."

Limitations Given the Small Size of the Protein Abundance Data

Even with the extended coverage made possible by merging many datasets together into our two reference sets, we still found that the largest complication in our analysis was the limited amount of data. This was, obviously, most applicable to the protein abundance measurements. In addition to giving us fewer data points for our statistics, the small number of protein abundance measurements potentially biased our statistical results towards certain protein families. The 181 proteins in G_{prot} are certainly not a random selection from the possible 6280 in yeast. They are, rather, skewed towards well-studied proteins that are highly expressed. Our methodology attempts to control for this gene-selection bias through our enrichment formalism, which allows one to rather precisely gauge various aspects of the bias.

Our results will certainly be more complete and definitive when larger proteomics datasets become available, which we anticipate to become available soon (Smith 2000). However, we believe that the essential formalism and approach that we develop will remain quite relevant for all future datasets.

Although the translatome data we used in our study is small in comparison to the information on the genome and transcriptome, many protein features in both the translatome and the transcriptome are dominated by the very highly expressed proteins (to which the 2-DE experiments are biased). Under this circumstance, it is often sufficient to look at this smaller number of dominating proteins to approximately characterize the whole population. This is similar in spirit to the development of the Codon Adaptation Index for Yeast (Sharp & Li 1987). While based on only 24 highly expressed proteins, it has proven to be robust in predicting expression levels for the entire genome. In contrast, the experimental bias in the selection of proteins with particular biophysical properties should be of more concern.

Future Directions

Besides the recapitulation of our computations with the release of new data, we also hope to expand this analysis to other organisms. While presently we have limited our study to yeast gene expression, there are other potential model organisms for which there are expression experiments. Moreover, we have also limited ourselves to Gene Chip experiments, but it may be worthwhile to analyze cDNA microarray data sets (DeRisi *et al.* 1997; Cho *et al.* 1998; Winzeler *et al.* 1999). We can use these sizeable microarray data sets to study changes in protein features over time.

Supplementary Material

Supplementary material is available at http://genecensus.org/expression/translatome

Acknowledgements

MG thanks the Keck foundation for support.

Figure and Table Captions

Table 1, Data Sets

This table provides an overview of the data sets used in our analysis. The table is divided into three sections. The first section at the top lists different mRNA expression sets. The second section in the middle shows the protein abundance data sets used. The third section at the bottom contains different annotations of protein features. The column "Data set" lists a shorthand reference to each data set used throughout this paper. The next columns contain a brief description of the data sets, the number of ORFs contained in each of them, the literature reference and the URL. In contrast to the other data we investigated, the reference expression and abundance data sets have been calculated for the purpose of our analysis (see text).

Some further information on the genome annotations:

Localization: Protein localization information from YPD, MIPS and SwissProt were merged, filtered and standardized (Bairoch & Apweiler 2000; Costanzo *et al.* 2000; Mewes *et al.* 2000) into five simplified compartments -- cytoplasm, nucleus, membrane, extracellular (including proteins in ER and golgi), and mitochondrial -- according to the protocol in Drawid et al. (2000). This yielded a standardized annotation of protein subcellular localization for 2133 out of 6280 ORFs.

Transmembrane segments: In 2710 out of 6280 yeast ORFs transmembrane segments are predicted to occur, ranging from low to high confidence (732 ORFs). The transmembrane prediction was performed as follows: The values from the scale for amino acids 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 on surveys of membrane protein in genomes. "Sure" membrane proteins had at least two TM-segments with an average hydrophobicity less than –2 kcal/mole. (Rost *et al.* 1995; Gerstein *et al.* 2000; Santoni *et al.* 2000; Senes *et al.* 2000).

Functions. MIPS functional categories have been assigned to 3519 out of 6194 ORFs. (The remainder are assigned to category '98' or '99', which corresponds to unclassified function.)

Figure 1, Schematic overview of the analysis

On the left side we outline the terms we use to describe the process of gene expression. The coding section of the genome is transcribed into a population of mRNA transcripts called the "transcriptome". The transcripts in turn are translated to a population of proteins; we use the term "translatome" for this protein population rather than the alternative "proteome" because the latter term may be confounded with the protein complement of the genome (which is not necessarily associated with a quantitative abundance level).

The matrix in the middle schematically shows an analysis of the three stages of expression. In general, we define a protein "population" as a set of genes associated with a corresponding number of expression or abundance levels ("weights"). In the matrix each row represents a weight and each column a gene set. In particular, we differentiate between the mRNA reference expression set ($G_{mRNA} = G_{Gen}$), which essentially covers the complete genome, and the reference protein abundance set (G_{Prot}) which contains the proteins in data sets 2-DE #1 and 2-DE #2 (see table 1) because the protein abundance set is a significantly smaller subset of the genome. By definition, this subset contains only proteins that can be identified by 2-D gel electrophoresis and is therefore biased in this sense. The enrichment figures throughout this paper, through a comparison of the right and left sides of this figure, show the results of the experimental biases of 2D gels on the data set.

Each pie chart represents a composition of a particular protein feature F (for instance, an amino acid composition) in a population (represented by the symbol μ). We can further look at the "enrichment" of this feature in one population relative to another (represented by the symbol Δ, see section "Methods" for an explanation of the formalism).

For simplification, we neglect the effects of post-transcriptional and post-translational modifications that might alter the features of proteins (they affect the expression levels but this is largely accounted for by the measurements). In this study we analyze protein features as they are represented in the genome.

Figure 2, mRNA expression levels vs. protein abundance levels

Part A of this figure shows the reference protein abundance levels plotted against the mRNA reference expression levels on a log-log scale; this plot is similar to the one reported by Futcher et al. (1999) earlier. The trend line is described by the equation $y = 5.20x^{0.61}$ where y represents the protein abundance level (in units of 10^3 copies/cell) and x the mRNA expression level (in units of copies/cell). The dashed lines indicate a distance of 1.85 standard deviations (in the log scale) from the trend line. The outliers beyond the dashed lines are listed in **Part B**. For each of these outlier ORFs we show a description of their function and their respective MIPS categories (the numbers are defined in Figure 4C). With one exception, all outliers are associated with cellular organization (MIPS category 30). Those outliers that have a high level of protein abundance relative to the expected amount of mRNA expression are dominated by the alcohol and G3P dehydrogenases. Translation-related proteins are prominent in the group of those proteins with low protein

abundance in relation to mRNA expression.

Figure 3, Amino Acid and Biomass Enrichment

Part A shows the amino acid enrichments between different populations as indicated by the legend to the right of the plot (the legend is ordered in the same way as the schematic illustration in Figure 1). The bars indicate the enrichment of the transcriptome relative to the genome, whereas the circles indicate the enrichment of the translatome relative to the genome. In addition, we also show the enrichment for protein abundance from the Transposon Abundance Set, represented by the circles with the line through them. It can be seen that the enrichments for the transcriptome and the translatome follow a similar trend despite their differences. In general, the amino acid enrichments seem to be more strongly emphasized in the translatome. In contrast, the enrichments for the Transposon Abundance Set seem to be very small. This may be due to the fact that the ORFs fused with *lacZ* produce different gene products than the original genes. In both the translatome and the transcriptome the amino acids Valine, Glycine and Alanine are strongly enriched. On the other end, the amino acids Asparagine, Cysteine and Serine are strongly depleted.

Part B shows a different view of amino acid enrichment from that contained in part A, now focusing on changes, and thus restricting the comparison to the genes common to all the datasets. The graph is ordered according to the enrichment from transcriptome to translatome (black squares). We focus here only on the changes for the abundance gene set (G_{Prot}) to exclude the effects that arise from looking at different subsets. In this view the enrichments from genome to transcriptome (white squares) and from genome to translatome (white diamonds) look more similar than do the analogous sets in Part A. To make comparison with Part A easier we again show the enrichment from genome to the transcriptome for the complete gene set (G_{Gen}, shown in bars).

Part C shows biomass enrichment. The left panel depicts the average molecular weight per ORF (in units of kDa) and the right panel, the average molecular weight per amino acid (in units of Daltons) in each of the three stages of gene expression. The numbers inside the circles indicate the average molecular weights. The values next to the arrows indicate the enrichments in biomass between different populations. Both the circle diameters and the arrow widths are functions of the corresponding values (the hollow arrow indicates a positive value). It is very clear that the average molecular weight per ORF is much lower in the translatome (by 20% or 15%) and transcriptome (by 29%) than in the genome. This relative depletion of biomass mainly takes place as a result of transcription; the effect of translation is less clear, depending on the populations compared. On the other hand, the depletion in the average molecular weight per amino acid (-3.3 % from genome to translatome) is an order of magnitude smaller than in the average weight per ORF. This shows that the yeast cell favors the expression of shorter ORFs over longer ones, and agrees with our earlier observation that there is a negative correlation between maximum ORF length and mRNA expression (Jansen & Gerstein 2000); it seems that this effect mainly takes place during transcription rather than translation.

Part D This plot shows that the amino acid enrichments are statistically significant. We have assessed significance by randomly permuting the expression levels among the genes and then recomputing the amino acid enrichments. This procedure can be repeated and used to generate

distributions of random enrichments that can then be compared against the observed enrichments. In the plot the gray bars represent the observed enrichments already shown in figure 3a. On top of the gray bars we show standard boxplots of enrichment distributions based on 1000 random permutations. (The middle line represents the distribution median. The upper and lower sides of the box coincide with the upper and lower quartiles. Outliers are shown as dots and defined as data points that are outside the range of the whiskers, the length of which is 1.5 the interquartile distance.) Based on the random distributions, we can compute one-sided P-values for the observed enrichments. Amino acids for which the P-values are less than 10^{-3} are shown in bold font.

Figure 4, Breakdown of the Transcriptome and Translatome in terms of Broad Categories relating to Structure, Localization, and Function

All of the subfigures are analogous to the schematic illustration in figure 1.

Part A represents the composition of secondary structure in the different populations. In general, the secondary structure compositions appear to be relatively stable across the different populations. The most notable change from genome to translatome is perhaps the depletion of coils -- that is, relatively unordered structures compared to the more structured helices and sheets -- by about 4%.

Part B represents the distribution of subcellular localizations associated with proteins in the various populations. We used standardized localizations developed earlier (Drawid & Gerstein 2000), which, in turn, were derived from the MIPS, YPD, and Swiss-Prot databases (Bairoch & Apweiler 2000; Costanzo *et al.* 2000; Mewes *et al.* 2000). The subcellular localization has been experimentally determined for less than half of the yeast proteins, so our analysis applies only to this subset. The most notable difference between genome, transcriptome and translatome is the strong enrichment of cytoplasmic proteins. This is in agreement with our previous observations (Drawid *et al.* 2000). This also explains to some degree the observations for the functional classes in part C. For example, the functional group "energy" is mostly dominated by the highly expressed glycolytic proteins found in the cytoplasm. The depletion of the functional group "transcription" makes sense in the light of the strong depletion for nuclear proteins. We have argued before (Drawid *et al.* 2000) that the number of proteins in a particular subcellular compartment may be roughly related to the size of the compartment. For instance, membrane proteins occupy the relatively small "two-dimensional" space in lipid bi-layers. We also performed a separate, independent calculation for a more comprehensive list of transmembrane segments, which were predicted computationally (see caption of Table 1). This largely confirms the result. (Data not shown.)

Part C shows the division of ORFs into different functional categories (according to the MIPS classification) in the various populations. Only the largest functional categories of the top level of the MIPS classification are shown. The group "Other" contains the smaller top-level categories lumped together. This “Other” group is different from the group "Unclassified," which contains genes without any functional description. One complication is that many genes have multiple functional classifications such that they may be counted in more than one category (this explains why the group "Unclassified" has only a size of 28% for the genome population although the

number of unclassified genes in the yeast genome is much larger).

Comparing the genome with the transcriptome and translatome compositions in general, it can be observed that if a functional class is *enriched* in the transcriptome relative to the genome, it is also enriched in the translatome. Specifically, the functional classes "metabolism", "energy", "protein synthesis" and "cellular organization" are enriched in transcriptome and translatome. On the other hand, the classes "cell growth, cell division and DNA synthesis" and "transcription" are depleted; in particular, this is the case for the "unclassified" group, indicating that a lot of the current biochemical knowledge is clearly skewed towards more highly expressed genes. Some of the differences between the complete gene set (G_{Gen}) and the protein abundance set (G_{Prot}) are obviously a result of the bias of electrophoresis experiments. In addition, the ribosomal proteins that make up an important highly expressed part of the class “protein synthesis” are underrepresented in the protein abundance set (G_{Prot}).

REFERENCES

An, H., R. K. Scopes, et al. (1991). Gel electrophoretic analysis of Zymomonas mobilis glycolytic and fermentative enzymes: identification of alcohol dehydrogenase II as a stress protein. *J Bacteriol* **173**(19): 5975-82.

Anderson, L. and J. Seilhamer (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18**(3-4): 533-7.

Bairoch, A. (2000). Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics* **16**(1): 48-64.

Bairoch, A. and R. Apweiler (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**(1): 45-8.

Bassett, D. E., Jr., M. A. Basrai, et al. (1996). Exploiting the complete yeast genome sequence. *Curr Opin Genet Dev* **6**(6): 763-6.

Batke, J., V. A. Benito, et al. (1992). A possible in vivo mechanism of intermediate transfer by glycolytic enzyme complexes: steady state fluorescence anisotropy analysis of an enzyme complex formation. *Arch Biochem Biophys* **296**(2): 654-9.

Cambillau, C. and J. M. Claverie (2000). Structural and Genomic Correlates of Hyperthermostability. *J Biol Chem* **275**(42): 32383-32386.

Carlson, M. (2000). The awesome power of yeast biochemical genomics. *Trends in Genetics* **16**(2): 49-51.

Cavalcoli, J. D., R. A. VanBogelen, et al. (1997). Unique identification of proteins from small genome organisms: theoretical feasibility of high throughput proteome analysis. *Electrophoresis* **18**(15): 2703-8.

Cho, R. J., M. J. Campbell, et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**(1): 65-73.

Claverie, J. M. (1999). Computational methods for the identification of differential and coordinated gene expression [In Process Citation]. *Hum Mol Genet* **8**(10): 1821-32.

Corthals, G., Wasinger VC, Hochstrasser DF, Sanchez JC (2000). The dynamic range of protein expression: a challenge for proteomic research. *Electrophoreisis* **21**(6): 1104-1115.

Costanzo, M. C., J. D. Hogan, et al. (2000). The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res* **28**(1): 73-6.

Das, R. and M. Gerstein (2000). The Stability of Thermophilic Proteins: A Study Based on Comprehensive Genome Comparison. *Functional & Integrative Genomics* **1**: 33-45.

Day, D. A. and M. F. Tuite (1998). Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *J Endocrinol* **157**(3): 361-71.

DeRisi, J. L., V. R. Iyer, et al. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338): 680-6.

Doolittle, W. F. (2000). The nature of the universal ancestor and the evolution of the proteome. *Curr Opin Struct Biol* **10**(3): 355-8.

Drawid, A. and M. Gerstein (2000). A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol* **301**(4): 1059-75.

Drawid, A., R. Jansen, et al. (2000). Gene Expression Levels are Correlated with Protein

Subcellular Localization (in Press). *Trends in Genetics*.
Einarson, M. and E. Golemis (2000). Encroaching genomics: adapting large-scale science to small academic laboratories. *Physiological Genomics* **2**(3): 85-92.
Eisen, M. B. and P. O. Brown (1999). DNA arrays for analysis of gene expression. *Methods Enzymol* **303**: 179-205.
Epstein, C. and R. Butow (2000). Microarray technology - enhanced versatility, persistent challenge. *Current Opinions Biotechnology* **11**(1): 36-41.
Ferea, T. and P. Brown (1999). Observing the living genome. *Current Opinions Genetic and Development* **9**(6): 715-722.
Fey, S. J. and P. M. Larsen (2001). 2D or not 2D. Two-dimensional gel electrophoresis. *Curr Opin Chem Biol* **5**(1): 26-33.
Fey, S. J., A. Nawrocki, et al. (1997). Proteome analysis of Saccharomyces cerevisiae: a methodological outline. *Electrophoresis* **18**(8): 1361-72.
Frishman, D. and H. W. Mewes (1997). Protein structural classes in five complete genomes [letter]. *Nat Struct Biol* **4**(8): 626-8.
Frishman, D. and H. W. Mewes (1999). Genome-based structural biology. *Prog Biophys Mol Biol* **72**(1): 1-17.
Futcher, B., G. Latter, et al. (1999). A sampling of the yeast proteome. *Mol Cell Biol 1999* **19**(11): 7357-68.
Gaasterland, T. (1999). Archaeal genomics. *Curr Opin Microbiol* **2**(5): 542-7.
Garrels, J. I., C. S. McLaughlin, et al. (1997). Proteome studies of Saccharomyces cerevisiae: identification and characterization of abundant proteins. *Electrophoresis* **18**(8): 1347-60.
Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol* **274**(4): 562-76.
Gerstein, M. (1998). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des* **3**(6): 497-512.
Gerstein, M. (1998). Patterns of Protein-Fold Usage in Eight Microbial Genomes: A Comprehensive Structural Census. *Proteins* **33**: 518-534.
Gerstein, M. (1998). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* **33**(4): 518-34.
Gerstein, M. and H. Hegyi (1998). Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol Rev* **22**(4): 277-304.
Gerstein, M. and R. Jansen (2000). The current excitement in bioinformatics, analysis of whole-genome expression data: How does it relate to protein structure and function (In press). *Current Opinions in Structural Biology*.
Gerstein, M., J. Lin, et al. (2000). Protein folds in the worm genome. *Pac Symp Biocomput*: 30-41.
Goffeau, A., Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996). Life with 6000 genes. *Science* **274** 5287.
Gygi, S.P. , G. L. Corthals et al (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci USA* **97** 9390–9395.
Gygi, S. P., B. Rist, et al. (2000). Measuring gene expression by quantitative proteome analysis [In Process Citation]. *Curr Opin Biotechnol* **11**(4): 396-401.
Gygi, S. P., B. Rist, et al. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**(10): 994-9.
Gygi, S. P., Y. Rochon, et al. (1999). Correlation between protein and mRNA abundance in yeast.

Molecular Cell Biology **19**(3): 1720-30.
Harry, J. L., M. R. Wilkins, et al. (2000). Proteomics: Capacity versus utility. *Electrophoreisis* **21**(6): 1071-1081.
Hatzimanikatis, V., L. H. Choe, et al. (1999). Proteomics: theoretical and experimental considerations. *Biotechnol Prog* **15**(3): 312-8.
Haynes, P.A. and Yates, J.R. (2000) Proteome profiling-pitfalls and progress. *Yeast* **17**: 81-87.
Hegyi, H. and M. Gerstein (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* **288**(1): 147-64.
Holstege, F. C., E. G. Jennings, et al. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**(5): 717-728.
Ishii M., S.Hashimoto, et al. (2000). Direct Comparison of GeneChip and SAGE on the Quantitative Accuracy in Transcript Profiling Analysis. Genomics **68**(2):136-143.
Ito, T., K. Tashiro, et al. (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**(3): 1143-1147.
Jackson, R. J. and M. Wickens (1997). Translational controls impinging on the 5'-untranslated region and initiation factor proteins. *Curr Opin Genet Dev* **7**(2): 233-41.
Jacobs Anderson, J. S. and R. Parker (2000). Computational identification of cis-acting elements affecting post- transcriptional control of gene expression in Saccharomyces cerevisiae. *Nucleic Acids Res* **28**(7): 1604-17.
Jaeger, J. A., D. H. Turner, et al. (1990). Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol* **183**: 281-306.
Jansen, R. and M. Gerstein (2000). Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res* **28**(6): 1481-8.
Jelinsky, S. A. and L. D. Samson (1999). Global response of Saccharomyces cerevisiae to an alkylating agent. *Proc Natl Acad Sci U S A* **96**(4): 1486-91.
Jones, D. T. (1998). Do transmembrane protein superfolds exist? *FEBS Lett* **423**(3): 281-5.
Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* **287**(4): 797-815.
Kidd, D et al. (2001) Profiling serine hydrolase activities in complex proteomes. Biochemistry **40**(13):4005-4015.
Klaff, P., D. Riesner, et al. (1996). RNA structure and the regulation of gene expression. *Plant Mol Biol* **32**(1-2): 89-106.
Klose, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**(3): 231-43.
Krogh, A. et al (2001).Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**(3):567-580.
Lin, J. and M. Gerstein (2000). Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res* **10**(6): 808-18.
Lindahl, L. and A. Hinnebusch (1992). Diversity of mechanisms in the regulation of translation in prokaryotes and lower eukaryotes. *Curr Opin Genet Dev* **2**(5): 720-6.
Lipshutz, R. F. S., Gingeras TR, Lockhart DJ (1999). High density synthetic oligonucleotide arrays. *Nature Genetics* **21**(1): 20-24.

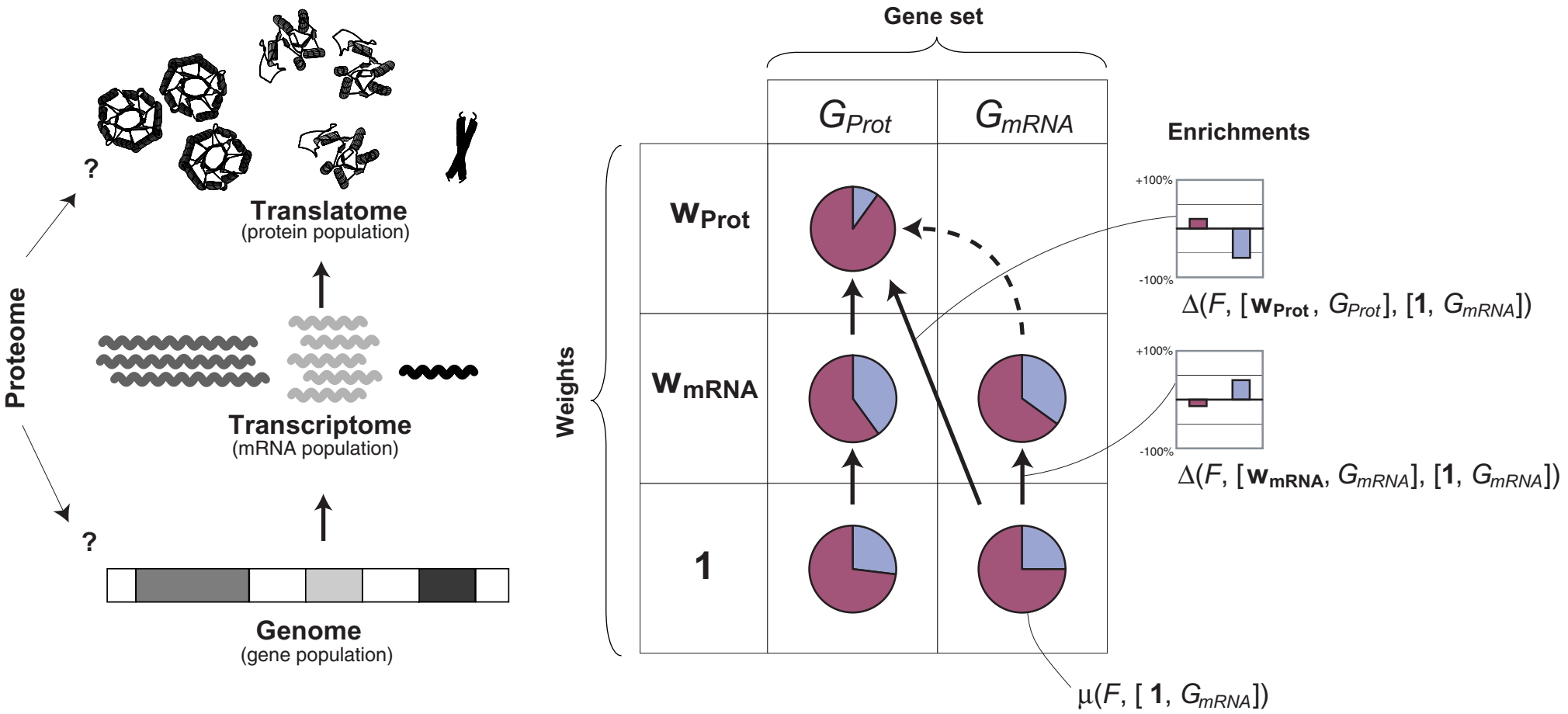
Lopez, M. F. (2000). Better approaches to finding the needle in a haystack: Optimizing proteome analysis through automation. *Electrophoreisis* **21**(6): 1082-1093.
Luban, J. and S. P. Goff (1995). The yeast two-hybrid system for studying protein-protein interactions. *Current Opinions in Biotechnology* **6**(1): 59-64.
MacBeath, G. and S. L. Schreiber (2000). Printing proteins as microarrays for high-throughput function determination. *Science* **289**(5485): 1760-3.
Matton, D. P., P. Constabel, et al. (1990). Alcohol dehydrogenase gene expression in potato following elicitor and stress treatment. *Plant Mol Biol* **14**(5): 775-83.
McCarthy, J. E. (1998). Posttranscriptional control of gene expression in yeast. *Microbiol Mol Biol Rev* **62**(4): 1492-553.
Mewes, H. W., D. Frishman, et al. (2000). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **28**(1): 27-40.
Millar, A. A., M. R. Olive, et al. (1994). The expression and anaerobic induction of alcohol dehydrogenase in cotton. *Biochem Genet* **32**(7-8): 279-300.
Molloy, M. P. (2000). Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients. *Anal Biochem* **280**(1): 1-10.
Morris, D. R. and A. P. Geballe (2000). Upstream open reading frames as regulators of mRNA translation. *Mol Cell Biol* **20**(23): 8635-42.
Nauchitel, V. V. and R. L. Somorjai (1994). Spatial and free energy distribution patterns of amino acid residues in water soluble proteins. *Biophysical Chemistry* **51**(2-3): 327-336.
Nelson, R. W., D. Nedelkov, et al. (2000). Biosensor chip mass spectrometry: a chip-based proteomics approach [In Process Citation]. *Electrophoresis* **21**(6): 1155-63.
O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**(10): 4007-21.
Pandey, A. and M. Mann (2000). Proteomics to study genes and genomes. *Nature* **405**(6788): 837-46.
Qi, S. Y., A. Moir, et al. (1996). Proteome of Salmonella typhimurium SL1344: identification of novel abundant cell envelope proteins and assignment to a two-dimensional reference map. *J Bacteriol* **178**(16): 5032-8.
Ross-Macdonald, P., P. S. Coelho, et al. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**(6760): 413-418.
Rost, B., R. Casadio, et al. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci* **4**(3): 521-33.
Roth, F. P., J. D. Hughes, et al. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat BIOTECHNOL* **16**(10): 939-45.
Rubin, G. M., M. D. Yandell, et al. (2000). Comparative genomics of the eukaryotes. *Science* **287**(5461): 2204-15.
Sali, A. (1999). Functional Links between Proteins.
Nature **402**(23): 25-26.
Santoni, V., M. Molloy, et al. (2000). Membrane proteins and proteomics: un amour impossible? *Electrophoreisis* **21**(6): 1054-1070.
Schena, M., D. Shalon, et al. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray [see comments]. *Science* **270**(5235): 467-70.
Searls, D. B. (2000). Using bioinformatics in gene and drug discovery. *Drug Discovery Today* **5**(4): 135-143.

Senes, A., M. Gerstein, et al. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol* **296**(3): 921-36.
Shapiro, L. and T. Harris (2000). Finding function through structural genomics. *Current Opinions in Biotechnology* **11**(1): 31-35.
Sharp, P. M. and W. H. Li (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**(3): 1281-95.
Sherlock, G. (2000). Analysis of large-scale gene expression data. *Curr Opin Immunol* **12**(2): 201-5.
Shevchenko, A., O. N. Jensen, et al. (1996). Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci U S A* **93**(25): 14440-5.
Smith, R. D. (2000). Probing proteomes-seeing the whole picture? [In Process Citation]. *Nat Biotechnol* **18**(10): 1041-2.
Tatusov, R. L., E. V. Koonin, et al. (1997). A genomic perspective on protein families. *Science* **278**(5338): 631-7.
Tekaia, F., A. Lazcano, et al. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Res* **9**(6): 550-7.
Vilela, C., B. Linz, et al. (1998). The yeast transcription factor genes YAP1 and YAP2 are subject to differential control at the levels of both translation and mRNA stability. *Nucleic Acids Res* **26**(5): 1150-9.
Vilela, C., C. V. Ramirez, et al. (1999). Post-termination ribosome interactions with the 5'UTR modulate yeast mRNA stability. *Embo J* **18**(11): 3139-52.
Wallin, E. and G. von Heijne (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* **7**(4): 1029-38.
Washburn, M. P., D. Wolters, et al. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**(3): 242-7.
Washburn, M. P. and J. R. Yates, 3rd (2000). Analysis of the microbial proteome. *Curr Opin Microbiol* **3**(3): 292-7.
Winzeler, E. A., D. D. Shoemaker, et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**(5429): 901-6.
Wittes, J. and H. P. Friedman (1999). Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data [editorial; comment]. *J Natl Cancer Inst* **91**(5): 400-1.
Wolf, Y. I., S. E. Brenner, et al. (1999). Distribution of protein folds in the three superkingdoms of life. *Genome Res* **9**(1): 17-26.
Young, K. H. (1998). Yeast two-hybrid: so many interactions, (in) so little time... *Biol Reprod* **58**(2): 302-311.
Zhang, M. Q. (1999). Large-scale gene expression data analysis: a new challenge to computational biologists [published erratum appears in Genome Res 1999 Nov;9(11):1156]. *Genome Res* **9**(8): 681-8.
Zhu, H., J. F. Klemic, et al. (2000). Analysis of yeast protein kinases using protein chips. *Nat Genet* **26**(3): 283-9.
Zuker, M. (2000). Calculating nucleic acid secondary structure. *Curr Opin Struct Biol* **10**(3): 303-10.

	Data set	Description	Size [ORFs]	Reference	URL [http://]
mRNA expression	Young	Gene chip profiles yeast cells with mutations that affect transcription	5455	Holstege et al., Cell 1998. 95(5):717-28	www.wi.mit.edu/young/expression
	Church	Gene chip profiles of yeast cells under four different conditions	6263	Roth et al, Nat Biotech 1998. 16(10):939-45	atlas.med.harvard.edu/roth
	Samson	Comparing gene chip profiles for yeast cells subjected to alkylating agent	6090	Jelinsky et al., PNAS 1998. 96:1486-1491	www.hsph.harvard.edu/geneexpression
	SAGE	Yeast cells during vegetative growth	3778	Velculescu et al., Cell 1997. 88(2):484-7	www.sagenet.org/yeast/yeastintro.htm
	Reference expression	Scaling and integrating the mRNA expression set into one data source	6249	-	bioinfo.mbb.yale.edu/expression
Protein abundance	2-DE #1	Measurement of yeast protein abundance by two-dimensional (2D) gel electrophoresis and mass spectrometry	156	Gygi et al., Mol Cell Biol 1999. 19(3):1720-30	depts.washington.edu/~ruedilab/aebersold.html
	2-DE #2	Similar to 2-DE set #1	71	Futcher et al., Mol Cell Biol 1999. 19(11):7357-68	-
	Transposon	Large-scale fusions of yeast genes with *lacZ* by transposon insertion	1410	Ross-Macdonald et al., Nature 1999. 402(6760):413-8	ycmi.med.yale.edu/ygac/triples.htm
	Reference abundance	Scaling and integrating the 2-DE data sets into one data source	181	-	bioinfo.mbb.yale.edu/expression
Annotation	Annotated Localization	Subcellular localizations of yeast proteins	2133 (6280)	Drawid A., Gerstein M. J Mol Biol. 2000, 301(4):1059-75	bioinfo.mbb.yale.edu/genome/localize
	Transmem-brane segments	Predicted transmembrane and soluble proteins in yeast	2710 (6280)	Gerstein, M., Proteins 1998. 33(4): 518-34	bioinfo.mbb.yale.edu/genome/new/update.html
	MIPS functions	Functional categories for yeast ORFs	3519 (6194)	Mewes et al., NAR 2000. 28(1):27-40	www.mips.biochem.mpg.de/ proj/yeast/catalogues/funcat
	GOR secondary structure	Predicted secondary structure for yeast ORFs	6280	Gerstein, M., Proteins 1998. 33(4): 518-34	bioinfo.mbb.yale.edu/genome/browser/db/SC/gorss.fa

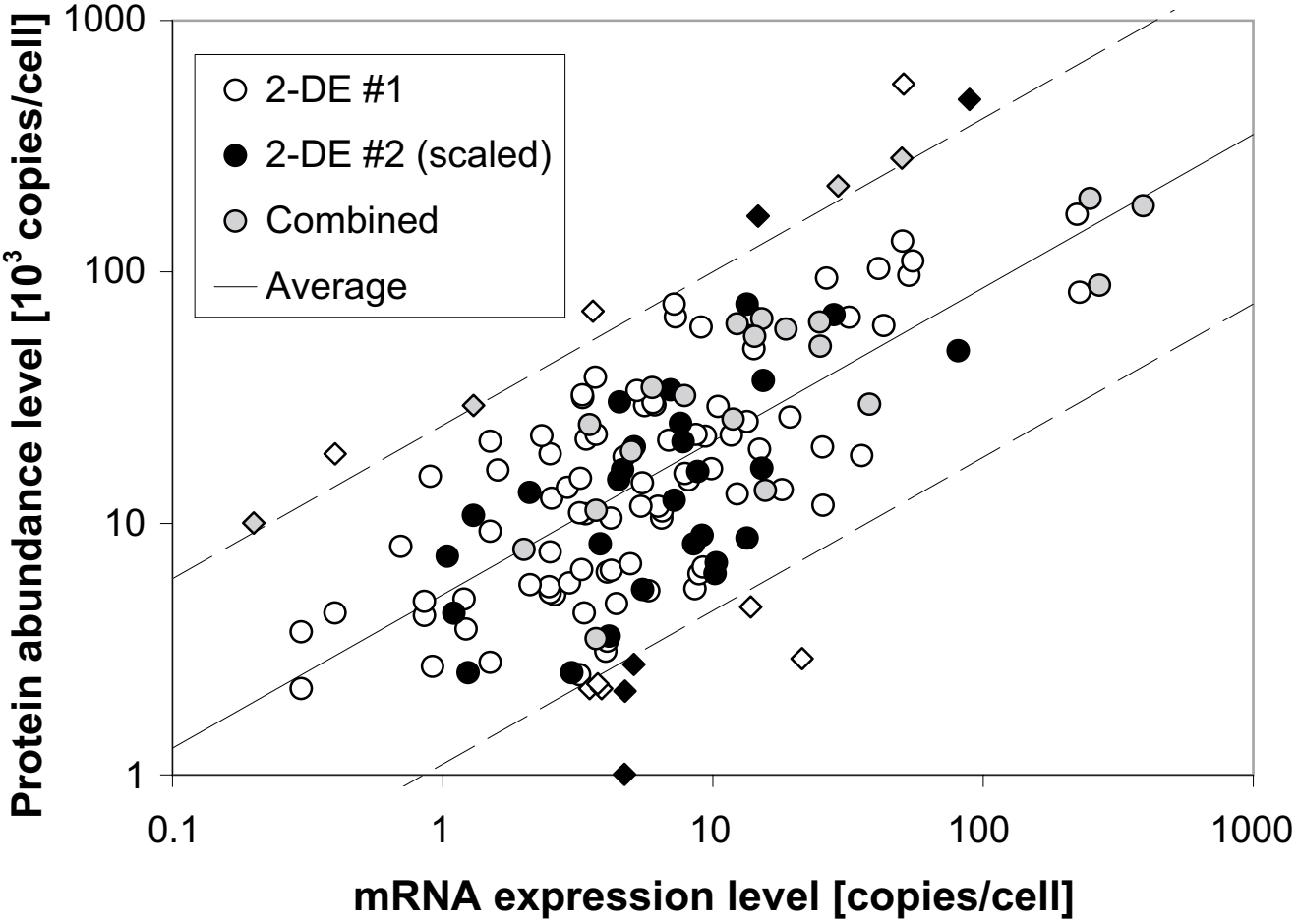
Greenbaum et al -Table 1

Greenbaum et al - Figure 1



Greenbaum et al - Figure 2

a

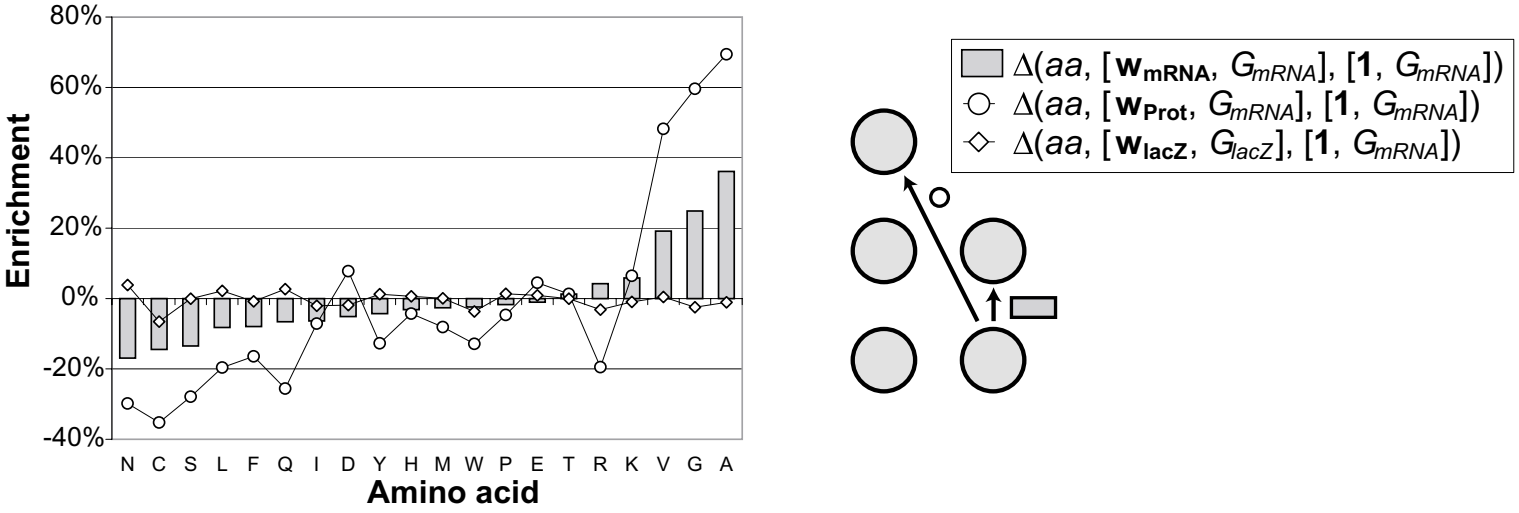


b

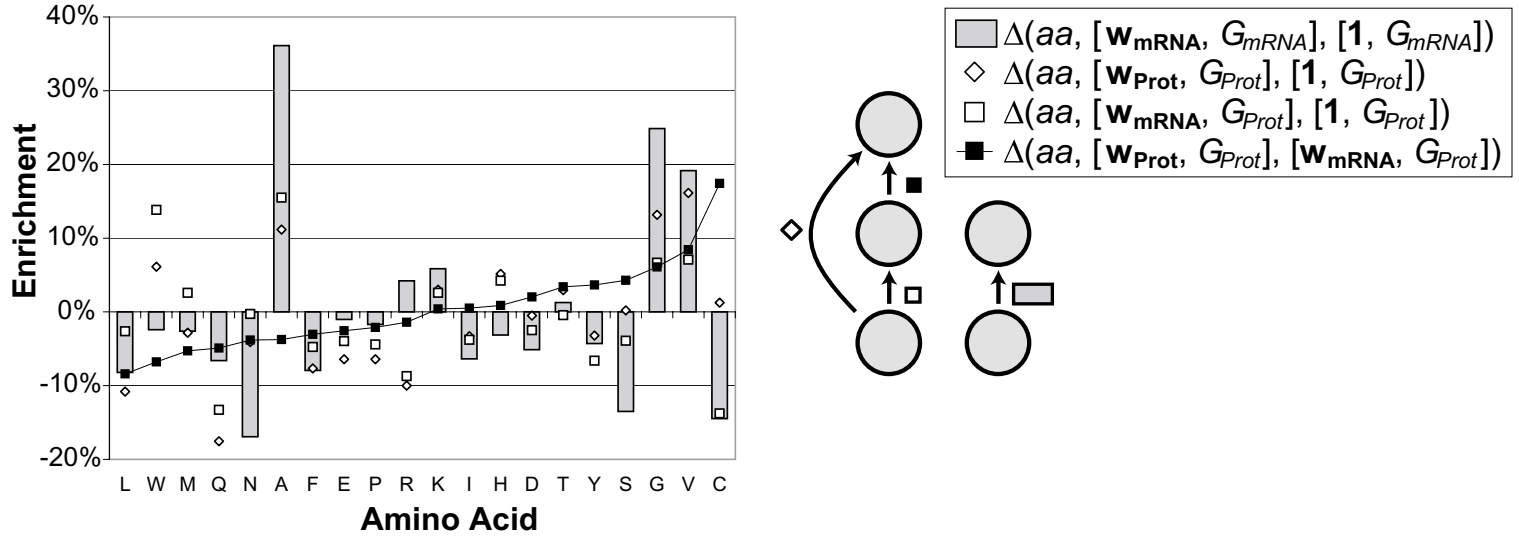
	ORF	FUNCTION	MIPS
above trendline	YBR118W	translation elongation factor eEF1 alpha-A chain	5,30
	YER065C	Isocitrate Lyase	1,2, 30
	YMR303C	Alcohol dehydrogenase II	1, 2, 30
	YOL086C	Alcohol dehydrogenase I	1, 2, 30
	YJR009C	Glyceraldehyde-3-phosphate dehydrogenase 2	1, 2, 30
	YGR192C	Glyceraldehyde-3-phosphate dehydrogenase 3	1, 2, 30
	YJR104C	Copper-zinc superoxide dismutase	11,30
	YML054C	lactate dehydrogenase cytochrome b2	1,2,30
	YJL052W	glyceraldehyde-3-phosphate dehydrogenase 1	1,2,30
below trendline	YKR059W	Translation initiation factor	5,30
	YML008C	S-adenosyl-methionine delta-24-sterol-c-methyltransferase	1,30
	YFL022C	Phenylalanine-- tRNA Ligase beta chain	5,30
	YJL008C	Component of chaperonin-containing T-complex	6,30
	YPL160W	leucine--tRNA ligase	5,30
	YOR361C	translation initiation factor eIF3 subunit	3,5,30
	YCL030C	phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase/histidinol dehydrogenase	1
	YNL209W	heat shock protein of HSP70 family	5,30

Greenbaum et al - Figure 3

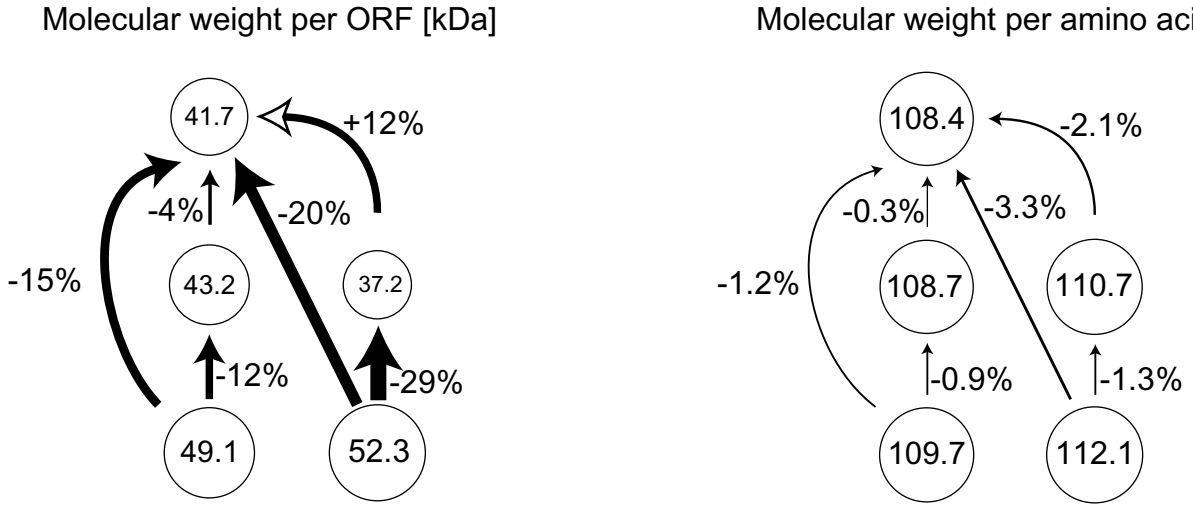
a



b

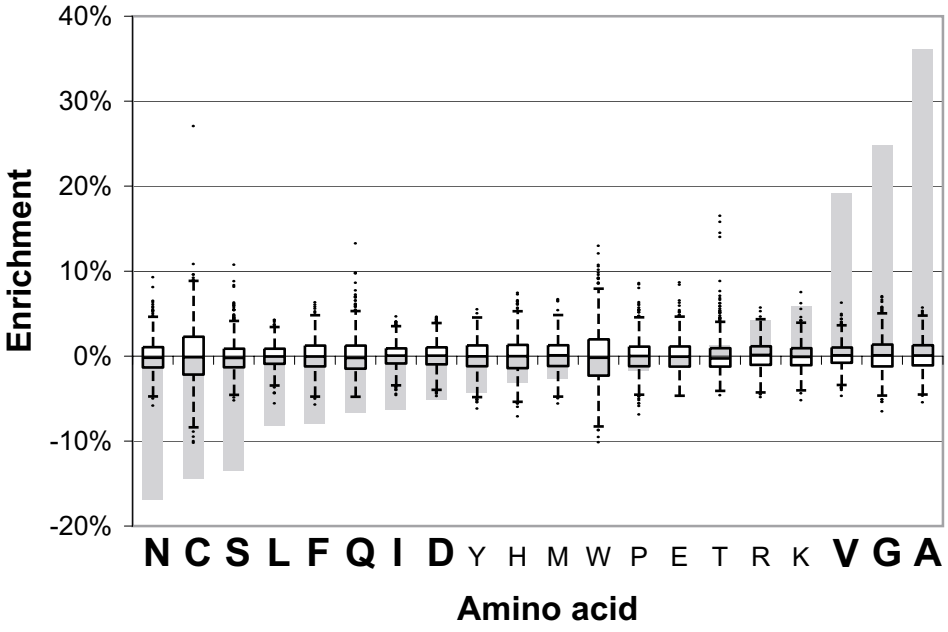


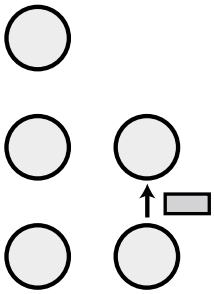
c



Greenbaum et al - Figure 3

d





Greenbaum et al - Figure 4