Determinants of Protein Abundance and Translation Efficiency in *S. cerevisiae*

Tamir Tuller[1,2*], Martin Kupiec[2], Eytan Ruppin[1,3*]

1 School of Computer Science, Tel Aviv University, Tel Aviv, Israel, **2** Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv, Israel, **3** School of Medicine, Tel Aviv University, Tel Aviv, Israel

The translation efficiency of most *Saccharomyces cerevisiae* genes remains fairly constant across poor and rich growth media. This observation has led us to revisit the available data and to examine the potential utility of a protein abundance predictor in reinterpreting existing mRNA expression data. Our predictor is based on large-scale data of mRNA levels, the tRNA adaptation index, and the evolutionary rate. It attains a correlation of 0.76 with experimentally determined protein abundance levels on unseen data and successfully cross-predicts protein abundance levels in another yeast species (*Schizosaccharomyces pombe*). The predicted abundance levels of proteins in known *S. cerevisiae* complexes, and of interacting proteins, are significantly more coherent than their corresponding mRNA expression levels. Analysis of gene expression measurement experiments using the predicted protein abundance levels yields new insights that are not readily discernable when clustering the corresponding mRNA expression levels. Comparing protein abundance levels across poor and rich media, we find a general trend for homeostatic regulation where transcription and translation change in a reciprocal manner. This phenomenon is more prominent near origins of replications. Our analysis shows that in parallel to the adaptation occurring at the tRNA level via the codon bias, proteins do undergo a complementary adaptation at the amino acid level to further increase their abundance.

Citation: Tuller T, Kupiec M, Ruppin E (2007) Determinants of protein abundance and translation efficiency in *S. cerevisiae*. PLoS Comput Biol 3(12): e248. doi:10.1371/journal.pcbi.0030248

Introduction

DNA microarrays are now commonly used to measure the expression levels of large numbers of genes simultaneously [1]. Since proteins are the direct mediators of cellular processes, the abundance level of each protein is likely to be a better indicator of the cellular state than its corresponding mRNA expression level. However, genome-wide technologies to detect protein abundance are still lagging behind those that measure mRNA, and only few studies that measure protein abundance on a large scale are currently available [2–6].

The relationship between mRNA and protein abundance levels has been studied by several groups. Genes with similar mRNA levels may have very different protein abundance levels [7]. Yet, the correlation between protein and mRNA abundance after a log-transform was shown to be quite high [8]. A more recent study, combining three technologies for measuring mRNA expression, has yielded correlation levels of about 0.7 with protein abundance [9]. Several studies have aimed at correlating protein abundance to various other features of proteins, such as their codon bias, molecular weight, stop codon identity, and more [3,4,10,11] These investigations and other previous proteomic studies [12–14] were usually based on small- to medium-scale measurements.

The current study revisits these issues and presents a comprehensive investigation of the relationship between factors that influence protein abundance and the associated protein levels. We begin by constructing a predictor for protein abundance levels, which, in contrast to previous studies, is tested and validated on unseen data (see Methods). To this end, we rely on two large-scale protein abundance datasets [2,5]. Overall, to our knowledge this is the first time that the whole body of data currently available is collated and analyzed to this aim, and we obtain a predictor with a correlation of 0.76 with experimentally determined abundance levels. Applying the resulting predictor to pertaining mRNA expression data testifies to its utility. Our analysis provides new key insights concerning the regulation of translation efficiency and its evolution.

Results

Genome-wide studies have measured mRNA and protein levels in the yeast *Saccharomyces cerevisiae* growing either in rich medium (yeast extract, peptone, and dextrose [YEPD]) or on poor, defined medium (synthetic dextrose [SD]) [2,3,5]. When protein abundance is compared to the corresponding mRNA levels in a given medium, the translation efficiency (TE), i.e., the ratio between protein abundance and mRNA levels, exhibits a large variability among genes (spanning across six orders of magnitude; Figure 1A and 1B). However, when the TEs of a given protein are compared across the two different growth conditions, notably very little variation is observed

Editor: Mark B. Gerstein, Yale University, United States of America

Received May 14, 2007; **Accepted** October 30, 2007; **Published** December 21, 2007

Abbreviations: ARS, autonomously replicating sequence; CAI, codon adaptation index; ER, evolutionary rate; GO, Gene Ontology; ORF, open reading frame; RTE, relative translation efficiency; SD, synthetic dextrose; SVM, support vector machine; tAI, tRNA adaptation index; TE, translation efficiency; YEPD, yeast extract, peptone, and dextrose

* To whom correspondence should be addressed. E-mail: tamirtul@post.tau.ac.il (TT); ruppin@post.tau.ac.il (ER)

Author Summary

DNA microarrays measuring gene expression levels have been a mainstay of systems biology research, but since proteins are more direct mediators of cellular processes, protein abundance levels are likely to be a better indicator of the cellular state. However, as proteomic measurements are still lagging behind gene expression measurements, there has been considerable effort in recent years to study the correlations between gene expression (and a plethora of protein characteristics) and protein abundance. Addressing this challenge, the current study is one of the first to introduce a predictor for protein abundance levels that is tested and validated on unseen data using all currently available large-scale proteomic data. The utility of this predictor is shown via a comprehensive set of tests and applications, including improved functional coherency of complexes and interacting proteins, better fit with gene phenotypic data, cross-species prediction of protein abundance, and most importantly, the reinterpretation of existing gene expression microarray data. Finally, our revisit and analysis of the existing large-scale proteomic data reveals new key insights concerning the regulation of translation efficiency and its evolution. Overall, a solid protein abundance prediction tool is invaluable for advancing our understanding of cellular processes; this study presents a further step in this direction.

(Figure 1C): the ratios between the TEs of most proteins in the two conditions are close to 1, with >90% of the proteins showing a ratio between 0.5 and 2. This observation, albeit currently limited to the two types of media for which genome-wide data are available, suggests that the efficiency of translation per mRNA molecule of many genes may be largely invariable under different conditions. This fairly constant TE of yeast genes has motivated us to create a large-scale predictor of protein abundance, with the aim of studying its utility for inferring protein abundance levels across different conditions.

The simplest predictor we studied is a linear one based on mRNA expression levels. Training this predictor on a randomly selected subset of the full complement of yeast mRNA and protein levels yields a Spearman rank correlation coefficient of $r_s = 0.55$ on held-out test data (the protein abundance was from [2] and mRNA levels were from [15]; see Methods). To improve the prediction accuracy, we examined the potential utility of combining 32 additional protein attributes into a multivariable linear predictor, some of which have been previously shown to have predictive value (Table S1). A greedy feature selection algorithm identified two useful protein attributes, while the inclusion of all other features resulted in a marginal and insignificant improvement in the performance of the linear, mRNA-based predictor. Performing the prediction by a support vector machine (SVM) using a variety of nonlinear kernels did not improve the results (Methods).

The two protein features yielding a significant improvement in prediction accuracy were the tRNA adaptation index (tAI) [16,17], and the evolutionary rate (ER) [18,19]. tAI is based on the synonymous codon usage bias and gene copy number of different tRNAs and is related to the codon adaptation index (CAI) [16,17]. ER measures the rate of evolution of a protein by comparing its orthologs across related species [18,19]. These two features have been shown previously to be correlated with protein abundance levels

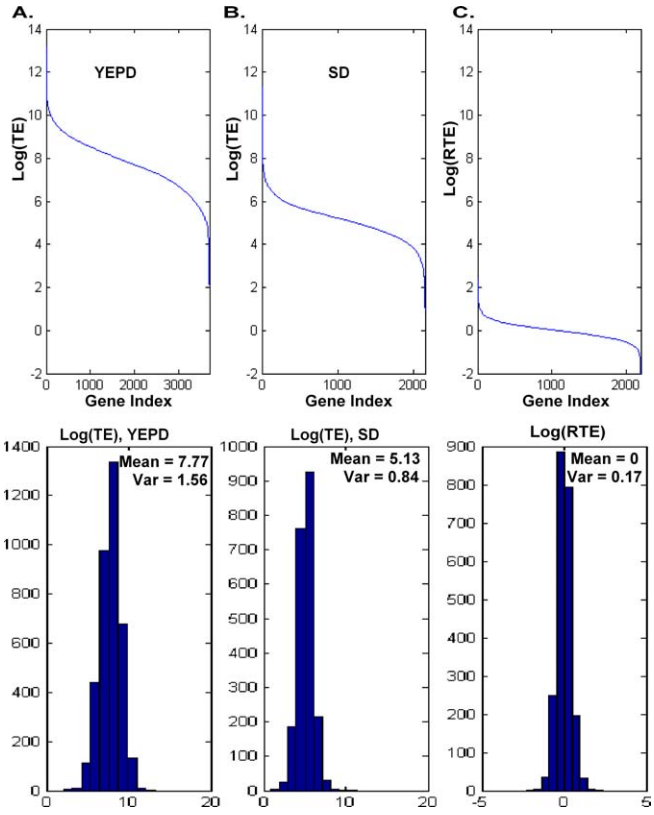

Figure 1. Distribution of TE and RTE in *S. cerevisiae*

(A) Top: *S. cerevisiae* genes sorted by their TE (log scale) in YEPD (rich) medium. A large variability of TE values (more than six orders of magnitude) is observed. Bottom: histogram, mean, and variance of TE in YEPD.

(B) Top: *S. cerevisiae* genes sorted by their TE (log scale) in SD (poor) medium. A similar large variability of TE values is seen. Bottom: histogram, mean, and variance of TE in SD.

(C) Top: *S. cerevisiae* genes sorted by the log-ratio of their TEs [RTE = $(p_{SD}/m_{SD})/(p_{YEPD}/m_{YEPD})$] in SD versus YEPD (log scale). A total of 91% of the genes have an RTE value between 0.5 and 2. Bottom: histogram, mean, and variance of RTE.

doi:10.1371/journal.pcbi.0030248.g001

[18,20]. Combining tAI with mRNA levels increases the prediction accuracy from the levels of $r_s = 0.55$ obtained using mRNA levels alone to a Spearman rank correlation coefficient of $r_s = 0.61$ on the same dataset as above. Adding evolutionary rate values increases the correlation to 0.63. The incremental improvement of consecutively adding these two features to the basic linear regression protein abundance predictor is statistically significant (Figure 2 and Methods).

Large-scale measurements of mRNA and protein levels tend to be noisy. Thus, in the (yet rare) cases where several independent measurements of mRNA and protein levels at the same conditions are available, they can be used to reduce potential individual measurement biases by pooling them together [9] (the correlation between two proteomic datasets generated by two different techniques and in different labs are between $r_s = 0.6$ and $r_s = 0.8$; see Text S1). We thus averaged mRNA and protein abundance results obtained with different technologies (see Methods for the description of the pertaining datasets used to this end). This results in a further notable improvement of prediction accuracy ($r_s = 0.76$; Figure 2), suggesting that a considerable fraction of the variability in the datasets is due to experimental measurement errors (the

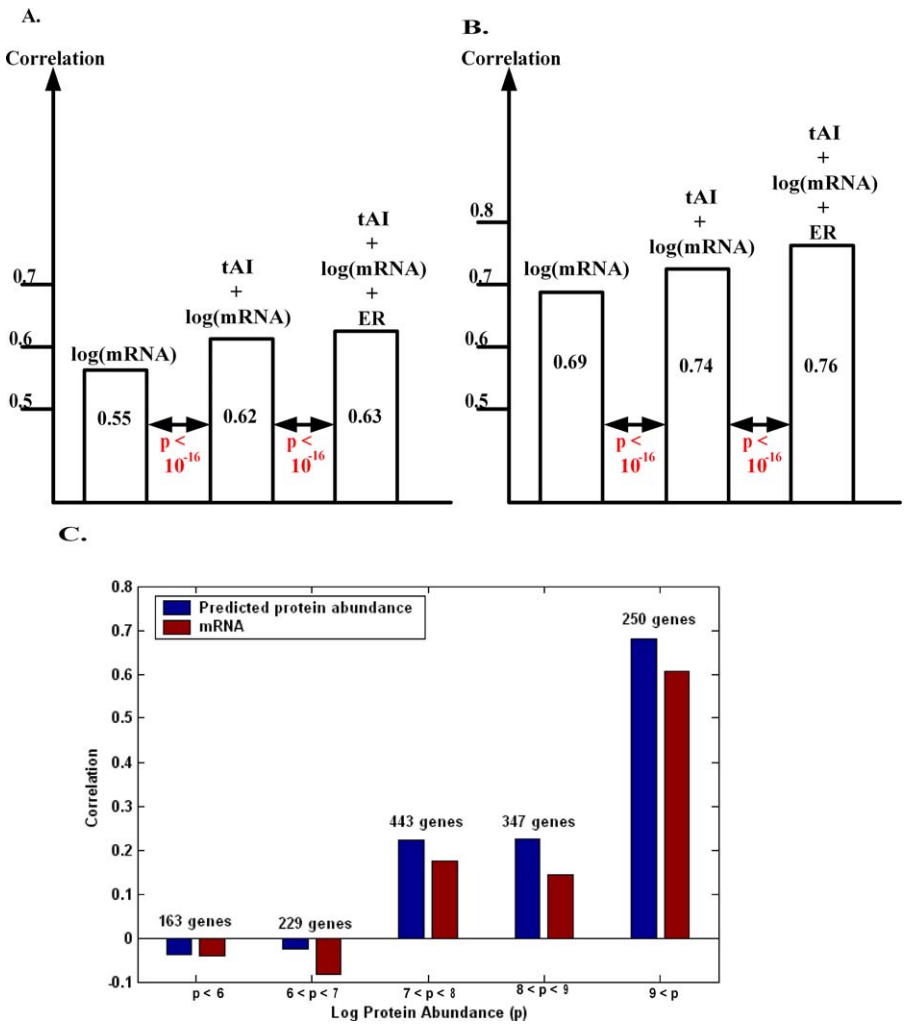

Figure 2. Performances of the Linear Predictor of (log) Protein Abundance

(A) The accuracy of various linear predictors of (log) protein abundance, measured by the Spearman rank correlation coefficient over a held-out test set, using a single data source of protein abundance [2] and mRNA levels [15]. ER values are from [19], and tAI data are taken from [20]. The numbers below the arrows denote the *t*-test *p*-values for checking the null hypothesis that the predictor with the new added feature has identical performance to its predecessor (see Methods). The final predictor for protein abundance (PA) is $\log(PA) = 3.97 + 0.4 \times \log(mRNA) + 10.34 \times tAI - 3.35 \times ER$.

(B) Accuracy of various linear predictors, in the case where protein and mRNA levels are generated by averaging measurements from at least two data sources. The final predictor for protein abundance obtained in this case is $\log(PA) = 3.47 + 0.63 \times \log(mRNA) + 10.89 \times tAI - 2.923 \times ER$.

(C) The Spearman correlations (*y*-axis) of predicted protein abundance (mRNA) with measured protein abundance levels, binned at various levels of protein abundance *p* (*x*-axis, natural log). All the correlations are higher and significant in the case of predicted protein abundance ($p < 2 \times 10^{-5}$), except for the lowest bin $\log(p) < 7$.

doi:10.1371/journal.pcbi.0030248.g002

improvement of the correlations observed upon averaging can also be due to the blurring of the effects of different posttranscriptional regulation processes taking place in the different conditions in which the measurements were done [temperature, strains, media, technique], but since we averaged over relatively similar conditions, we expect this effect to be relatively minor). In the following investigations reported in this paper, multiple independent measurements at the same conditions were not available, and the results reported are hence without pooling and averaging the data.

Examining the performance of our YEPD-trained predictor on a new unseen dataset of 238 genes whose protein abundance levels were measured under very different conditions (exposure to pheromone [13]) resulted in a high correlation of $r_s = 0.69$. The correlation between mRNA levels solely and protein abundance levels was 0.62, in comparison. The standard deviation of 1,000 cross-validation runs of the predictor was 0.016, and the improvement compared to mRNA-based prediction was significant, with $p < 10^{-16}$. Further information on the predictors' performance on specific Gene Ontology (GO) annotation gene sets is provided in Table S2. This table also shows that the predictor improves the prediction of protein abundance levels (compared to mRNA levels) in 92% of the GO annotation categories. Our predictor obtains higher correlations with protein abundance levels than using mRNA alone across numerous ranges of protein abundance; however, this correlation is not statistically significant in the lowest protein abundance range (Figure 2C).

Using our multivariate linear predictor, expression of genes whose products are members of the same complex (according to SGD [21]) exhibits significantly higher coher-

Table 1. The Spearman Rank Correlation Coefficients and Partial Spearman Correlations between mRNA, Protein Abundance, and Predicted Protein Abundance Levels in YEPD and SD for Gene Pairs That Are Part of the Same Complex

Description	Correlation	*p*-Value	Partial Correlation	*p*-Value
YEPD, mRNA level	0.1378	6.31×10^{-10}	—	—
YEPD, protein abundance	0.1908	$<10^{-16}$	0.1296	1.2×10^{-8}
YEPD, predicted protein abundance	0.1897	$<10^{-16}$	0.1	4.5×10^{-5}
SD, mRNA level	0.0863	5.5×10^{-4}	—	—
SD, protein abundance	0.1758	1.4×10^{-12}	0.1321	3.8×10^{-7}
SD, predicted protein abundance	0.1487	2.3×10^{-9}	0.0957	3.6×10^{-4}

Measuring the coherency of expression levels of proteins that are part of the same complex or are interacting neighbors in the protein network. Methods and Text S8 include a detailed description of how these values were computed.

doi:10.1371/journal.pcbi.0030248.t001

ency than when calculated from their corresponding mRNA levels. Table 1 displays the pertaining Spearman rank correlation coefficients for pairs of genes that are part of the same complex. For the cases of experimentally determined and predicted protein abundance levels, we also computed the partial correlations after controlling for the effect of mRNA expression levels (Methods). A similar, but weaker trend is also observed when examining the abundance coherency of protein pairs that exhibit a protein–protein interaction (Text S2). These results indicate that our prediction approach is likely to be more appropriate for proteins in large macromolecular complexes than for proteins involved in signaling and transcriptional control, since the latter are heavily posttranslationally modified. This notion is further supported by noting that in the highest protein abundance bin (Figure 2C), there are 26 genes that are related to the "Ribosome" GO category, providing a hyper-geometric enrichment of $p < 4.2 \times 10^{-4}$.

Given the observation that the TE of most proteins is fairly similar across the two different conditions analyzed, we examined the utility of the protein abundance predictor in interpreting the results of two yeast mRNA gene expression datasets, obtained under a variety of environmental conditions (see Text S3). The first dataset investigated the yeast response to low-shear modeled microgravity. It included 12 different conditions (six under low-shear and six controls) [22]. To analyze this dataset, we clustered and bi-clustered the genes in the microarray data in accordance with the mRNA expression patterns, in a conventional manner. In parallel, we used our predictor to generate predicted protein abundance levels from the expression levels, and repeated the clustering and bi-clustering process on the resulting protein abundance data. We then compared the resulting cluster sets with respect to their functional enrichment in GO annotations (Methods). We performed a similar analysis on a gene expression dataset consisting of 36 timepoints taken from yeast cells growing in continuous, nutrient-limited conditions [23] (the first dataset includes gene expression measurements of a system that is close to equilibrium, while the second includes gene expression measurements of a system in a transient state; see Text S4).

Table 2 shows that the use of the predicted protein abundance values in these datasets results in a significant increase in the percentage of clusters that exhibit enrichment for specific GO terms (for comparison, random predictors significantly deteriorate the clustering enrichment scores; see Text S5). In the case of Sheehan's data [22], the protein abundance predictor improved both the separation and the homogeneity. In the case of Tu's data [23], the homogeneity improved while the separation score deteriorated (Table 2). A closer analysis provides evidence for the advantage of using the predictor: in the first dataset, a new bi-cluster is detected (cluster 4) in the protein abundance analysis that does not appear in the mRNA level analysis. This bi-cluster spans over 11 of the 12 conditions and is enriched with many GO annotations (mainly related to metabolism; Table S4). Similarly, in the second dataset, cluster 7 in the predicted protein abundance analysis is a novel group that does not appear when analyzing mRNA levels. This cluster shows a striking periodic expression that coincides with the respiratory bursts observed under continuous nutrient-limited conditions [23]. Thus, using predicted protein abundance levels, a simple conventional clustering method suffices to reveal novel central clusters that were not apparent in the

Table 2. The Percentage of GO-Enriched Clusters and the Percentage of GO-Enriched Bi-Clusters Obtained by Analyzing mRNA Levels or Predicted Protein Abundance Levels in Two Gene Expression Datasets, and the Total Homogeneity and Separation Scores for the Clustering Results

Dataset	Percentage Enriched Clusters	Homogeneity	Separation	Percentage Enriched Bi-Clusters
Sheehan et al. [22] mRNA	57	0.77	−0.58	100
Sheehan et al. [22] predicted protein abundance	85	0.78	−0.6	100
Tu et al. [23] mRNA	85	0.68	−0.28	100
Tu et al. [23]predicted protein abundance	100	0.71	−0.23	100

doi:10.1371/journal.pcbi.0030248.t002

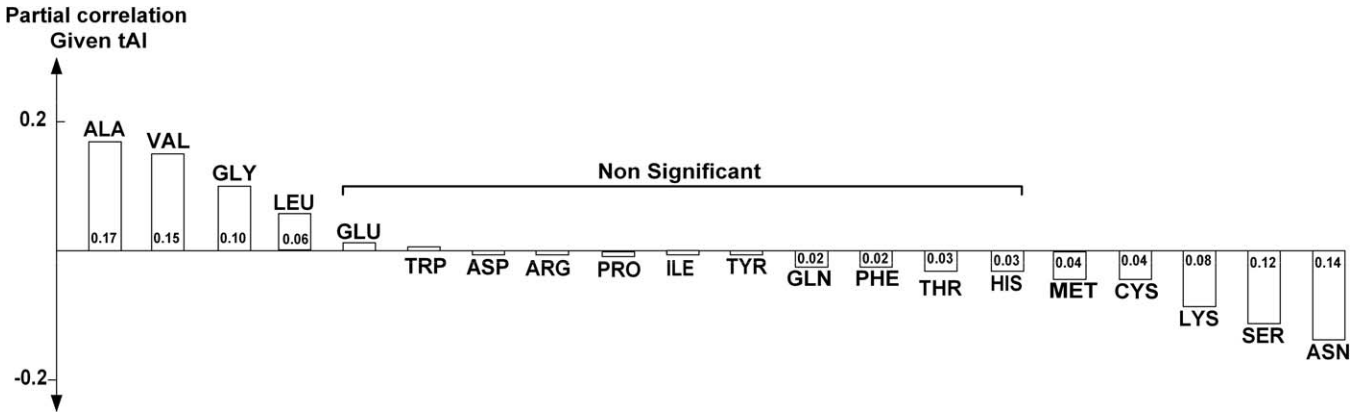

Figure 3. Partial Correlations between the Frequencies of Amino Acids Composing a Protein and Its Abundance Level (after Controlling for the Effect of tAI)

doi:10.1371/journal.pcbi.0030248.g003

original study at the mRNA expression level. Tables S3, S4, S5, and S6 provide a detailed analysis (list of clusters, bi-clusters, and GO enrichments) for the two datasets.

We used our protein abundance predictor to reanalyze the intriguing results reported by [24], showing that only a very small fraction of the genes whose expression is significantly elevated under a specific condition actually cause a significant decrease in fitness when deleted. Overall, we find that the fraction of expressed genes that lead to a significant reduction in fitness when deleted is 2-fold to 3-fold higher than the corresponding fraction reported using mRNA levels (e.g., 2.9% versus 0.76% in the case of yeast cells responding to 1.5 M sorbitol, and 13.2% versus 6.4% in the case of 1 M NaCl). Although the absolute fraction of genes accounted for still remains small, the relative increase observed by using the predictor is substantial.

Finally, we tested our predictor's ability to correctly estimate protein abundance levels from mRNA expression data in a different organism, *Schizosaccharomyces pombe*. To this end, we used mRNA and protein data from a recent genome-wide study that reported a Spearman rank correlation coefficient of 0.61 between the two measurements [25]. Focusing on a subset of *S. pombe* genes that have an ortholog in *S. cerevisiae,* the Spearman rank correlation of the predicted protein levels with actual protein abundance measurements was 0.675. Notably, for the same subset of genes, the Spearman rank correlation between the protein abundance and mRNA levels of *S. pombe* was only 0.629 (and the rank correlation between the mRNA levels of the two organisms was 0.48). These results are quite remarkable, since the predictor used to predict protein abundance in *S. pombe* was based on the ER and tAI values of the corresponding orthologs in *S. cerevisiae*.

Like previous studies [4,26], we have also found a significant correlation between the abundance of a particular protein and the frequency of certain amino acids composing it, the most prominent being alanine and valine (positive correlation), and serine and aspargine (negative correlation; Figure S1). This observation has been previously attributed to the different values of the tAI (or the CAI) of these amino acids, which can modulate translation efficiency [16,17]. However, we find that even after controlling for the effect of their different tAIs, the frequency of these amino acids remains significantly correlated with protein abundance, and their frequency at abundant proteins remains highly significant (see partial correlations reported in Figure 3, and similar results after controlling for CAI in Figure S2). The Spearman rank correlation of amino acid frequencies and protein abundance remains significant even after additionally controlling for the effect of mRNA expression levels (Table S7). This finding suggests that in parallel to the adaptation occurring at the tRNA level via the codon bias [27,28], proteins do undergo a complementary adaptation at the amino acid level via amino acid substitution to further increase their protein abundance. The small, neutral, and nonpolar amino acid alanine is probably ideally suited for this putative substitute role, given its known neutral effect on protein stability [29]. Both alanine and valine are present at relatively high concentrations within the yeast cell, and their corresponding acyl-tRNA synthases are also expressed at high levels (Table S8), aiding in their efficient incorporation during transcription (however, adding frequencies of amino acids to our predictor did not improve its performance significantly; see Text S6).

The recent direct measurement of absolute protein levels under two distinct growth conditions [5] enabled us to compare the ratio between the translation efficiency observed in cells grown on poor medium versus the one observed in rich medium, i.e., the relative TE (RTE; $(p/m)_{SD}/(p/m)_{YEPD}$). There is a significant negative correlation (–0.213; $p < 10^{-50}$) between the RTE and the change in transcription levels between the two growth conditions. Even when focusing only on genes that change their protein abundance between the two conditions in a considerable manner (protein abundance ratio > 1.4 or $< 1/1.4$), the resulting negative correlation remains significant ($r = -0.08$; $p = 0.018$). This may suggest that there is a global homeostasis between transcription and translation, with a tendency to increase translation when transcription decreases, and vice versa. The average RTE is 1.091 (about half the genes, 1,072 out of 2,204, have RTE > 1). Since the relative decrease of the ribosomal protein abundance ($p_{SD}/p_{YEPD} = 0.88$) is higher than the total relative decrease of mRNA levels ($m_{SD}/m_{YEPD} = 0.98$), the number of ribosomes per mRNA is lower in SD. Thus, the findings of average RTE > 1 are probably due to lower protein degradation rates or other causes of higher trans-

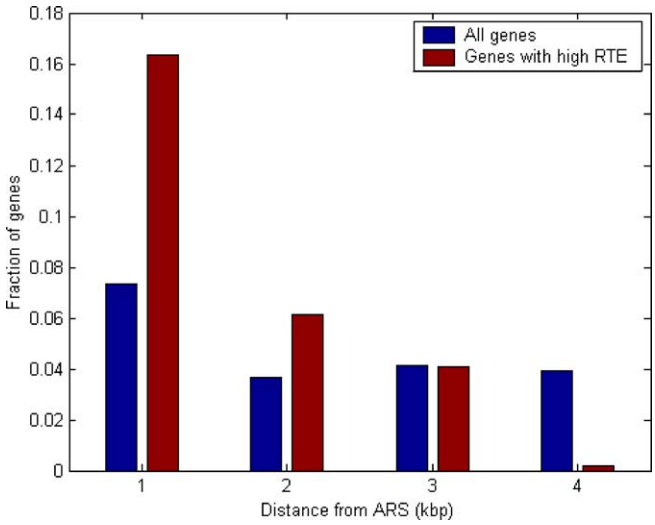

Figure 4. The Distribution of Genes with High RTEs at Different Distances from Origins of Replication

The distribution of genes with high RTE (RTE $>$ 2.5), and distribution of all genes at different distances from origins of replication. The number of genes with high RTE is 49; the total number of genes studied is 2,200. The number of genes with high RTE that are located within 1 kbp from an ARS is statistically significant using a hyper-geometric text ($p < 0.05$).
doi:10.1371/journal.pcbi.0030248.g004

lation rates in SD, rather than increased ribosomes per mRNA levels (Figure S3 depicts the mean RTE levels of different GO annotation groups; Text S7 displays the variance in protein abundance levels in the two growth conditions).

While the large majority of the genes have RTE levels ranging between 0.5 and 2 (Figure 1B), two sets have extreme RTE values, one with RTE $>$ 2.5 (48 genes), and the other with RTE $<$ 0.45 (65 genes; Tables S9 and S10). The distribution of mRNA and protein abundance levels of genes within each of these groups is similar to that of the rest of the genes (see Figure S4A and S4B), and extreme ratios of protein abundance or mRNA levels do not necessarily imply extreme RTE values (see Figure S4C). Interestingly, our predictor obtains more significant improvement in the correlations with actual protein abundance levels on genes with extreme RTEs (see Figure S4D). In contrast to the inverse (homeostatic) relation observed in general, the set with extremely high RTE also exhibits extremely high m_{SD}/m_{YEPD} ratios (an average mRNA ratio of 5.35, 14 times the general average). This indicates that the extreme RTE values reflect the fact that the cell is making a concerted effort to maintain their protein abundance levels at the extreme levels needed. By the same token, the mean mRNA ratio for the set with extremely low RTE is 0.36, somewhat below the total average.

The group of genes exhibiting extremely high RTE levels is enriched for mitochondrial genes (21/48 are mitochondrial genes; chi-square $p = 10^{-16}$), with many of these genes being related to mitochondrial biosynthesis and metabolism. Thus, the increase in the level of mitochondrial proteins, reflecting the need for higher-yield energy production in poor growth conditions, is achieved mainly by boosting translation efficiency. Interestingly, the high RTE group is also enriched with genes that map very close to origins of replication (autonomously replicating sequence [ARS]), including four genes abutting at the origin of replication (out of a total of 24 genes with a similar location in the yeast genome, providing a chi-square $p = 1.1 \times 10^{-6}$), and twice the expected number of genes located within 1 kbp from an ARS ($p < 0.05$; see Figure 4). A possible explanation for this intriguing connection is that the replication machinery, when binding to origins of replication, attenuates transcription, either by steric hindrance or by competition for DNA binding [30]. This interference is then compensated in turn by higher translation efficiency and a more flexible regulation of translation, as reflected by its high RTE levels. Indeed, the average m_{SD} / m_{YEPD} ratios of genes that have extremely high RTE and that are less than 1 kb from an ARS is only 0.8. One putative mechanism that may underlie this intriguing phenomenon is that certain proteins that participate in replication and transcription (e.g., Rap1 and Abs1) could be incorporated into the mRNA, exported from the nucleus, and differentially affect the rate of translation at the ribosome. Similar mechanisms have been suggested for the activity of proteins such as Yra1, Sub2, and the THO complex, which affect transcription, splicing efficiency, and nuclear export [31].

Discussion

The availability of whole-genome measurements of protein abundance provides a unique opportunity to analyze the forces that affect protein translation and abundance. Combining several protein features yields a predictor of protein abundance that can serve as a useful tool for analyzing gene expression measurements. Our results indicate that highly expressed proteins undergo adaptation at the amino acid level, and that proximity to an origin of replication enhances the efficiency of translation.

Translation efficiency is determined by invariant, condition-independent factors such as the amino acid and codon composition of the protein and the availability of the different tRNAs. It is also modulated by dynamic factors such as ribosome occupancy and ribosome density (determining the total number of ribosomes per mRNA), which are dependent on environmental clues [10]. Assuming that TE is constant to a first approximation for most genes (as its levels across poor and rich media testifies), this study has focused on the first group of factors, and has shown the utility of such a predictor in interpreting biological data. We anticipate that as information gradually accumulates concerning the second group of factors, more accurate protein abundance predictors will emerge that can incorporate information on posttranscriptional regulation [32–34]. Recent work has suggested that transcription factors and signaling genes tend to be posttranscriptionally regulated [32]. Indeed, a large proportion of the genes with extreme RTE levels belong to these two categories (see Tables S9 and S13). However, not all genes regulated at the posttranscriptional level exhibit extreme RTE values: a recent genome-wide study in yeast has identified 16 genes with extreme TE levels, presumably regulated posttranscriptionally [9]. Examination of the RTE levels of these genes reveals that only one has extreme RTE levels (*MET6,* with RTE = 0.47); the rest have RTE levels between 0.93 and 1.38 (see Table S13). Finally, protein degradation and turnover are obviously important modulators of protein abundance, and should be considered in future predictors as pertaining reliable data accumulates.

That said, it is interesting and encouraging to see how far one can go in predicting protein abundance levels even without this information.

An important corollary of our work is that gene expression results obtained with DNA microarray technology may in some cases be misleading. For example, Tables S11 and S12 include a subset of genes that exhibit inversely correlated regulatory trends at the transcription versus the translation level. An increase in mRNA expression levels of a particular gene does not necessarily mean a higher level of its protein. The corresponding protein abundance could not be differentially expressed or could even be differentially expressed but in the opposite direction. As Tables S11 and S12 include about 5% of the yeast genes, this type of error may be nonnegligible at times. Our predictor cannot solve this problem; its solution will probably require much larger biological datasets than those currently available.

We demonstrated that our predictor (which is based on *S. cerevisiae*) can be used to successfully predict protein abundance levels in a different organism (*S. pombe*), which has an evolutionary distance of 350–1,000 million y from *S. cerevisiae* [35]. It will be interesting to examine the effect that evolutionary distance may have on determining the "transferability" of protein predictors across species. However, answers to this question will need to wait until protein abundance data of additional organisms becomes available.

Building on the existing large-scale protein abundance data, this study has shown that a predictor of protein abundance levels can improve the interpretation of gene expression measurements and provide new insights into the regulation and evolution of protein translation. The utility of such a tool should be further enhanced as our understanding of the determinants affecting protein abundance and translation improves and the pertaining data continues to accumulate.

Methods

Generating a predictor of protein abundance. For training the predictors, we used all the genes whose required features (mRNA measurements, protein abundance, ER, tAI) were available. The series of linear predictors studied were generated using a linear regressor and using the following cross-validation procedure: (1) randomly choose 80% of the genes (training set) and use them for generating a linear predictor; (2) use the resulting predictor for predicting the protein abundance of the remaining 20% of the genes (test set); and (3) for the genes in the test set, calculate the Spearman rank correlation coefficient between the predicted and experimentally measured protein abundance values.

This cross-validation procedure is repeated 10^5 times, and the mean of the Spearman rank correlation coefficient (computed in step 3) is the predictor accuracy reported in the main text.

As reported in the main text, we generated a sequence of linear predictors of protein abundance, each time adding the most informative feature in a greedy manner. During this process, we checked if the resulting incremental improvement in prediction performance is statistically significant by performing a *t*-test, comparing the distribution of Spearman rank correlation coefficients obtained by each predictor over the 10^5 cross-validation runs. Note that in the case of a multivariate linear predictor, this cross-validation procedure may lead to similar prediction accuracy values as those obtained by training a multivariate regressor on the whole dataset. However, in the general scope of nonlinear predictors investigated in this study, the cross-validation prediction scenario used is conceptually different from a multivariate regression, and the results obtained significantly differ.

Going beyond a linear predictor, we used two implementations of SVMs, SVM-light [36] and Partek (Partek Software, http://www.partek.com), and examined radial, polynomial, and sigmoid kernels. The initial set of features included all the 32 features described in Table S1, and we also examined various forward and backward algorithms for feature selection. Quite surprisingly, none of these SVM predictors gave a significant increase in prediction performance compared to the best linear predictor reported upon in the main text. In constructing the predictors we used the following data sources.

Protein abundance and mRNA expression data. We analyzed four protein abundance datasets: (1) a dataset generated by merging (with the appropriate normalization) protein abundance data from numerous small-scale datasets [3]; (2) a large-scale measurement of protein abundance in yeast (normal log phase) [2]; and (3) protein abundance large-scale measurements by [5] in two different growth media conditions (YEPD and SD). We analyzed two major mRNA expression datasets: (1) one generated by combining 36 microarray datasets (wild-type yeast grown in YEPD without any stress) [10]; and (2) an mRNA measurement of wild-type yeast grown in YEPD [21].

The dataset of [5] also includes the ratio (but not the absolute values) between the mRNA levels in the two conditions (SD and YEPD), m_{SD}/m_{YEPD}. This information, combined with the protein abundance measurements in these two conditions, enabled us to compute the RTEs across growth conditions. Combined with the absolute mRNA measurements from [2], it was used to calculate the absolute mRNA levels in SD.

For computing mean protein abundance levels in constructing the pooled-data predictor, we averaged at least two of three measurements reported in [2,5,8]. For computing mean mRNA abundance levels to this construction, we averaged at least two of three measurements reported in [21,37,38]. The averaging was done following the procedure described in [9].

Sources of additional data. Protein half-life measurements were obtained from Belle et al. [39]. The protein properties examined in the construction of the protein abundance predictor (properties 1–28 in Table S1) were obtained from the *Saccharomyces* genome database [21]. The tAI data were downloaded from [20]. Evolutionary rates of proteins were taken from Wall et al. [19]. The mRNA gene expression data, protein abundance data, and list of 447 relevant orthologous genes needed for testing the predictor performance on *S. pombe* were from [25]. Relative protein abundance and mRNA levels after exposure to pheromone were downloaded from [13].

Clustering, bi-clustering, and GO enrichment analysis of mRNA and predicted protein abundance levels. We used two mRNA gene expression datasets that were generated by the same technology as that used for training the predictor. The two datasets are measurements by affymetrix GeneChip, and were downloaded from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gds). The first dataset includes the 12 samples from [22]. The second dataset includes the 36 samples from [23]. Clustering and bi-clustering was performed by using the Expander program [40]. We used CLICK for clustering and SAMBA for bi-clustering. Gene enrichment was computed using the GO categories of [21] (by computing the hyper-geometric probability of seeing at least x number of genes out of the total n genes in the cluster/bi-cluster annotated to a particular GO term, given the proportion of genes in the whole genome that are annotated to that GO term), examining the three ontologies of molecular function, biological process, and cellular components. The resulting enrichments were filtered by false discovery rate (FDR) to correct for multiple testing [41].

Measuring the coherency of expression levels of proteins that are part of the same complex or are interacting neighbors in the protein network. Protein complex data were downloaded from [21]. We measured coherency of mRNA levels, protein abundance, and predicted protein abundance of genes that are part of the same complex (in SD and YEPD) by the following steps: (1) we listed all pairs of genes in the dataset which are both comembers in one of the complexes; (2) for each case (mRNA levels, protein abundance, and predicted protein abundance), we generated two vectors, u and v, such that $u(i)$ and $v(i)$ denote a pair of proteins that are part of the same complex; we calculated the Spearman rank correlation coefficient between the two vectors (u and v); and we compared the resulting correlation to the correlations between pairs of vectors with the same length that include measurements of randomly selected pairs of genes.

For predicting protein abundance, we used a predictor that was trained on a different dataset (i.e., the predictor used for YEPD was trained on the SD measurements and vice versa; training the predictor on the same dataset gives an even better result, so we wanted to demonstrate that the results are significantly good even if the trained set and the test set are different.). The computation of the

pertaining partial correlations and their associated p-values are described in Text S8.

For computing the coherency of expression/abundance of neighboring proteins in the protein interaction network, we used the yeast protein interaction network from the work of [42].

We used a similar procedure to that used to compute the complexes' coherency, but this time u and v are composed of protein pairs that are adjacent in the protein interaction network.

Comparing mRNA expression profiling and fitness profiling. For comparing the number of genes that exhibits both an increase in expression levels (mRNA levels and predicted protein abundance) and a significant decrease in fitness when adding NaCl or sorbitol, we used the mRNA levels from [43] and fitness profiling from [24]. For each of the two cases (mRNA levels and predicted protein abundance), we used five measurements of expression levels and four measurements of fitness. We focused on the set of genes for which we had all the predictor's features. In the case of fitness profiling, a gene was considered "significant" if it had significant value (as defined in [24]) in at least one of the four fitness measurements. In both cases of protein abundance or mRNA expression levels, a gene was considered significant if it exhibited a log ratio of at least 0.25 in one of the five measurements.

Supporting Information

Figure S1. Variables That Have Significant Correlation and Partial Correlation with Protein Abundance, TE, and RTE

(A) Variables that have significant correlation with protein abundance, TE, and RTE.
(B) Variables with significant correlation with protein abundance given mRNA, given CAI, and given mRNA and CAI. The full names and the description of each variable appear in Table S1. The correlation with amino acid distribution at the C and N terminus was substantially less significant than the general correlations of amino acid distribution (it was not significant for most of the amino acids).

Found at doi:10.1371/journal.pcbi.0030248.sg001 (82 KB DOC).

Figure S2. Partial Correlations of Amino Acid Frequencies and Protein Abundance after Removing the Effect of CAI

Found at doi:10.1371/journal.pcbi.0030248.sg002 (53 KB DOC).

Figure S3. The Average RTE of GO Annotation Groups

The average RTE of each GO annotation group for the three ontologies (molecular function, cellular component, and biological process).

Found at doi:10.1371/journal.pcbi.0030248.sg003 (71 KB DOC).

Figure S4. mRNA Levels, Protein Abundance, mRNA Ratio, Protein Abundance Ratio, and Correlation with Protein Abundance of mRNA and Predicted Protein Abundance of Genes with Extreme RTE

(A) mRNA levels and protein abundance of genes with RTE $>$ 2.5 (blue), RTE $<$ 0.45 (red), and the rest of the genes (yellow) in YEPD.
(B) mRNA levels and protein abundance of genes with RTE $>$ 2.5 (blue), RTE $<$ 0.45 (red), and the rest of the genes (yellow) in SD.
(C) mRNA ratio (m_{SD}/m_{YEPD}) levels and protein abundance ratio (p_{SD}/p_{YEPD}) of genes with RTE $>$ 2.5 (blue), RTE $<$ 0.45 (red), and the rest of the genes (yellow).
(D) Correlation with protein abundance of mRNA and predicted protein abundance for genes with modest RTE ($0.5 < \text{RTE} < 2$), and for genes with extreme RTE (RTE $<$ 0.5 and RTE $>$ 2). The correlation increase after implementing the predictor is more significant for the group with extreme RTE.

Found at doi:10.1371/journal.pcbi.0030248.sg004 (86 KB DOC).

Table S1. Protein Features Used in the Study

Abbreviation and full description of all the protein features that were used in our study. We also checked the frequency of amino acids at the N and C terminus of the protein.

Found at doi:10.1371/journal.pcbi.0030248.st001 (52 KB DOC).

Table S2. The Correlation of the Predicted Protein Abundance of the Predictor with Real Protein Abundance, mRNA, tAI, and ER for Each GO Annotation Group Separately, and the Performances When Inferring a Different Predictor for Each Cellular Component GO

(A–C) The correlation of the predicted protein abundance of our predictor with real protein abundance, mRNA, tAI, and ER for each GO annotation group separately. The last column includes the correlation of mRNA level with protein abundance for each GO group (blue, cases where the predictor improved the correlation with protein abundance; red, cases where the mRNA level has higher correlation with protein abundance).
(A) The results for the cellular component GO annotation groups.
(B) The results for the biological process GO annotation groups.
(C) The results for the molecular function GO annotation groups.
(D) The performances (correlation of predicted and real protein abundance) when inferring a different predictor for each cellular component GO annotation group. The average performances in this case are not better than the original predictor (one predictor for all the GO groups).

Found at doi:10.1371/journal.pcbi.0030248.st002 (209 KB DOC).

Table S3. Clustering (Sheet 1) and Bi-Clustering (Sheet 2) of the mRNA Gene Expression, from the Work of Sheehan et al.

The list of genes in each cluster and bi-cluster is depicted together with the GO enrichment categories (for each of the ontologies: process, function, and component) of each cluster/bi-cluster. The score of each bi-cluster (by Expander) is depicted near the name of the bi-cluster (as mentioned by the authors of the pertaining Expander software used there, these scores are good only for comparing bi-clusters with the same size). The mean pattern of each bi-cluster and the index of conditions that are related to it (x-axis) appear near each bi-cluster.

Found at doi:10.1371/journal.pcbi.0030248.st003 (1.2 MB XLS).

Table S4. Clustering (Sheet 1) and Bi-Clustering (Sheet 2) of the Predicted Protein Abundance from the work of Sheehan et al.

The list of genes in each cluster and bi-cluster is depicted together with the GO enrichment categories (for each of the ontologies: process, function, and component) of each cluster/bi-cluster. The score of each bi-cluster (by Expander) is depicted near the name of the bi-cluster (as mentioned by the authors of the pertaining Expander software used there, these scores are good only for comparing bi-clusters with the same size). The mean pattern of each bi-cluster and the index of conditions that are related to it (x-axis) appear near each bi-cluster.

Found at doi:10.1371/journal.pcbi.0030248.st004 (1.6 MB XLS).

Table S5. Clustering (Sheet 1) and Bi-Clustering (Sheet 2) of the mRNA Gene Expression from the Work of Tu et al.

The list of genes in each cluster and bi-cluster is depicted together with the GO enrichment categories (for each of the ontologies: process, function, and component) of each cluster/bi-cluster. The score of each bi-cluster (by Expander) is depicted near the name of the bi-cluster (as mentioned by the authors of the pertaining Expander software used there, these scores are good only for comparing bi-clusters with the same size). The mean pattern of each bi-cluster and the index of conditions that are related to it (x-axis) appear near each bi-cluster.

Found at doi:10.1371/journal.pcbi.0030248.st005 (3 MB XLS).

Table S6. Clustering (Sheet 1) and Bi-Clustering (Sheet 2) of the Predicted Protein Abundance from the Work of Tu et al.

The list of genes in each cluster and bi-cluster is depicted together with the GO enrichment categories (for each of the ontologies: process, function, and component) of each cluster/bi-cluster. The score of each bi-cluster (by Expander) is depicted near the name of the bi-cluster (as mentioned by the authors of the pertaining Expander software used there, these scores are good only for comparing bi-clusters with the same size). The mean pattern of each bi-cluster and the index of conditions that are related to it (x-axis) appear near each bi-cluster.

Found at doi:10.1371/journal.pcbi.0030248.st006 (2.1 MB XLS).

Table S7. Partial Correlations of Amino Acid Frequencies and Protein Abundance for All the Genes and for Genes with Low mRNA Levels and High Protein Abundance

(A) Partial correlations of amino acid frequencies and protein abundance for all genes. The correlations for the amino acids alanine and valine are significant and positive, and the correlations for asparagine and serine are significant and negative.
(B) Partial correlations of the frequencies of amino acids and protein abundance for genes with low mRNA levels (lower 20%) and high protein abundance (top 20%). The correlations for the amino acids alanine and valine are positive but not significant (due to the low number of genes).

Found at doi:10.1371/journal.pcbi.0030248.st007 (13 KB XLS).

Table S8. Protein Abundance of the Various tRNA Synthetases and the Stechiometry of the Different Amino Acids

Protein abundance of the various tRNA synthetases and the stechiometry of the different amino acids (downloaded from the work of Förster et al. [44]). Alanine and valine tRNA synthetases have high levels of protein abundance, and the amino acids exhibit a high concentration in the yeast cell. These factors also make the translation of alanine and valine more efficient. Data that do not appear in our dataset are denoted by *###*.

Found at doi:10.1371/journal.pcbi.0030248.st008 (15 KB XLS).

Table S9. Genes with RTE > 2.5

Table includes the open reading frame (ORF), name, RTE, and description of each gene. Genes that are related with regulation are marked in blue. GO enrichments according to SGD for this group of genes appear below.

Found at doi:10.1371/journal.pcbi.0030248.st009 (27 KB XLS).

Table S10. Genes with RTE < 0.45

Table includes the ORF, name, RTE, and description of each gene. Genes that are related to regulation are marked in blue. GO enrichments according to SGD for this group of genes appear below.

Found at doi:10.1371/journal.pcbi.0030248.st010 (29 KB XLS).

Table S11. Subset of Genes That Exhibit Counteracting Regulatory Trends at the Transcriptional versus the Translational Levels (RTE $<$ 1/1.5 and $m_{SD}/m_{YEPD} > 1.5$)

Subset of genes that exhibit counteracting regulatory trends at the transcriptional versus the translational levels. Each gene in the set has RTE $<$ 1/1.5 and $m_{SD}/m_{YEPD} > 1.5$. For each gene, the table includes its ORF ID, name, RTE, and the ratio between the mRNA levels in SD and YEPD.

Found at doi:10.1371/journal.pcbi.0030248.st011 (11 KB XLS).

Table S12. Subset of Genes That Exhibit Counteracting Regulatory Trends at the Transcriptional Versus the Translational Levels (RTE $>$ 1/1.5 and $m_{SD}/m_{YEPD} < 1.5$)

Subset of genes that exhibit counteracting regulatory trends at the transcriptional versus the translational levels. Each gene in the set has RTE $>$ 1/1.5 and $m_{SD}/m_{YEPD} < 1.5$. For each gene, the table includes its ORF ID, name, RTE, and the ratio between the mRNA levels.

Found at doi:10.1371/journal.pcbi.0030248.st012 (32 KB XLS).

Table S13. RTE of Genes with Extreme TE

(A) The RTE of the genes that were reported by Lu et al. as genes with high TE. The table includes the name, ORF, and RTE of each gene. (B) The RTE of the 14 genes with extreme TE; in this case, the TE was calculated using the protein abundance of Ghaemmaghami et al. [2] and the mRNA levels of Holstege et al. [15] The table includes the name, ORF ID, RTE, TE, and TE rank (among all the genes) of each gene.

Found at doi:10.1371/journal.pcbi.0030248.st013 (12 KB XLS).

Text S1. Correlation Between Independent Measurements of Protein Abundance

Found at doi:10.1371/journal.pcbi.0030248.sd001 (24 KB DOC).

Text S2. Correlation Between mRNA Levels, Protein Abundance, and Predicted Protein Abundance between Interacting Proteins

Found at doi:10.1371/journal.pcbi.0030248.sd002 (25 KB DOC).

Text S3. Clustering and Bi-Clustering Predicted Protein Abundance

Found at doi:10.1371/journal.pcbi.0030248.sd003 (24 KB DOC).

Text S4. The Analysis of Steady-State and Transient Gene Expression Datasets

Found at doi:10.1371/journal.pcbi.0030248.sd004 (21 KB DOC).

Text S5. Clustering the Protein Abundance Levels Obtained from Random Predictors of Protein Abundance

Found at doi:10.1371/journal.pcbi.0030248.sd005 (25 KB DOC).

Text S6. Nonsignificant Improvement of the Predictor when Adding Amino Acid Frequencies

Found at doi:10.1371/journal.pcbi.0030248.sd006 (20 KB DOC).

Text S7. Variance in Protein Abundance for the Two Sets with Extreme RTEs

Found at doi:10.1371/journal.pcbi.0030248.sd007 (24 KB DOC).

Text S8. Supplementary Methods

Found at doi:10.1371/journal.pcbi.0030248.sd008 (25 KB DOC).

Acknowledgments

We would like to thank Elhanan Borenstein, Tomer Shlomi, and Roded Sharan for helpful discussions.

Author contributions. TT, MK, and ER conceived and designed the experiments. TT analyzed the data. TT, MK, and ER wrote the paper.

Funding. TT was supported by the Edmond J. Safra Bioinformatics program at Tel Aviv University. This research was supported by grants from the Israel Research Fund and the Israeli Ministry of Science and Technology to MK, and by grants from the Yishayahu Horowitz Center for Complexity Science, the Israeli Science Foundation (ISF), the German-Israeli Foundation for scientific research and development (GIF), and the Tauber fund to ER.

Competing interests. The authors have declared that no competing interests exist.

References

1. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. (2003) Arrayexpress—A public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 31: 68–71.
2. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737–741.
3. Greenbaum D, Colangelo C, Williams K, Gerstein M (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol 4: 1–8.
4. Greenbaum D, Jansen R, Gerstein M (2002) Analysis of mRNA expression and protein abundance data: An approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. Bioinformatics 18: 585–596.
5. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, et al. (2006) Single-cell proteomic analysis of *S. Cerevisiae* reveals the architecture of biological noise. Nature 441: 840–846.
6. Zhu H, Klemic J, Chang S, Bertone P, Casamayor A, et al. (2000) Analysis of yeast protein kinases using protein chips. Nature Gen 26: 283–289.
7. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol. Cell Biol 19: 1720–1730.
8. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI (1999) A sampling of the yeast proteome. Mol Cell Biol 19: 7357–7368.
9. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2006) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nature Biotech 25: 117–124.
10. Beyer A, Hollunder J, Nasheuer HP, Wilhelm T (2004) Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. Mol Cell Prot 3: 1083–1092.
11. Lithwick G, Margalit H (2003) Hierarchy of sequence-dependent features associated with prokaryotic translation. Genome Res 13: 2665–2673.
12. Kolkman A, Daran-Lapujade P, Fullaondo A, Olsthoorn MMA, Pronk JT, et al. (2006) Proteome analysis of yeast response to various nutrient limitations. Mol Syst Biol 2: 1–16.
13. MacKay VL, Li X, Flory MR, Turcott E, Law GL, et al. (2004) Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: Response of yeast to mating pheromone. Mol Cell Prot 3: 478–489.
14. Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, et al. (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A 100: 3107–3112.
15. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95: 717–728.
16. Reis M, Wernisch L, Savva R (2003) Unexpected correlation between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* k-12 genome. Nucleic Acid Res 31: 6976–6985.
17. Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: A test for translational selection. Nucleic Acids Res 32: 5036–5044.
18. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23: 327–337.
19. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, et al. (2005) Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A 102: 5483–5488.
20. Man O, Pilpel Y (2007) Differential translation efficiency of orthologous

genes is involved in phenotypic divergence of yeast species. Nat Genet 39: 415–421.
21. (2006) Saccharomyces genome database. Available: http://www.yeastgenome.org. Accessed 9 November 2007.
22. Sheehan KB, McInnerney K, Purevdorj-Gage B, Hyman LE (2007) Yeast genomic expression patterns in response to low-shear modeled microgravity. BMC Genomics 8: 1–12.
23. Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. Science 310: 1152–1158.
24. 24 . Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418: 387–391.
25. Schmidt MW, Houseman A, Lvanov AR, Wolf DA (2007) Comparative proteomic and transcriptomic profiling of fission yeast *Schizosaccharomyces Pombe*. Mol Sys Biol 3: 1–12.
26. Akashi H (2003) Translational selection and yeast proteome evolution. Genetics 164: 1291–1303.
27. Gouy M, Gautier C (1982) Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acid Res 10: 7055–7074.
28. Jansen R, Bussemaker HJ, Gerstein M (2003) Revisiting the codon adaptation index from a whole-genome perspective: Analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res 31: 2242–2251.
29. Kortemme T, Kim DE, Baker D (2004) Computational alanine scanning of protein-protein interfaces. Sci STKE 219: 1–12.
30. Deshpande AM, Newlon CS (1996) DNA replication fork pause sites dependent on transcription. Science 272: 1030–1033.
31. Jimeno S, Luna R, Garcia-Rubio M, Aguilera A (2006) Tho1, a novel hnRNP, and Sub2 provide alternative pathways for mRNP biogenesis in yeast THO mutants. Mol Cell Biol 26: 4387–4398.
32. Brockmann R, Beyer A, Heinisch JJ, Wilhelm T (2007) Posttranscriptional expression regulation: What determines translation rates. PLoS Comput Biol 3: 531–539.
33. Gebauer F, Hentze MW (2004) Molecular mechanisms of translational control. Nat Rev Mol Cell Biol 5: 827–835.
34. Holcik M, Sonenberg N (2005) Translational control in stress and apoptosis. Nat Rev Cell Biol 6: 318–327.
35. Berbee M, Taylor J (2001) Systematics and evolution. In: McLaughlin D, McLaughlin E, Lemke P, editors. The Mycota. Volume VIIB. Berlin: Springer. pp. 229–245.
36. Joachims T (2002) Learning to classify text using support vector machines. Boston: Kluwer Academic Publishers. 205 p.
37. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, et al. (2002) Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A 99: 5860–5865.
38. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, et al. (1997) Characterization of the yeast transcriptome. Cell 88: 243–251.
39. Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK (2006) Quantification of protein half-lives in the budding yeast proteome. Proc Natl Acad Sci U S A 103: 13004–13009.
40. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, et al. (2005) Expander—An integrative program suite for microarray data analysis. BMC Bioinformatics 6: 1–12.
41. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—A practical and powerful approach to multiple testing. J R Stat Soc B Mat 57: 289–300.
42. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, et al. (2005) Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A 102: 1974–1979.
43. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, et al. (2001) Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell 12: 323–337.
44. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Res 13: 244–253.