

Post-transcriptional Expression Regulation in the Yeast *Saccharomyces cerevisiae* on a Genomic Scale*

Andreas Beyer^{‡§}, Jens Hollunder[‡], Heinz-Peter Nasheuer[¶], and Thomas Wilhelm[‡]

Based on large-scale data for the yeast *Saccharomyces cerevisiae* (protein and mRNA abundance, translational status, transcript length), we investigate the relation of transcription, translation, and protein turnover on a genome-wide scale. We elucidate variations between different spatial cell compartments and functional modules by comparing protein-to-mRNA ratios, translational activity, and a novel descriptor for protein-specific degradation (protein half-life descriptor). This analysis helps to understand the cell's strategy to use transcriptional and post-transcriptional regulation mechanisms for managing protein levels. For instance, it is possible to identify modules that are subject to suppressed translation under normal conditions ("translation on demand"). In order to reduce inconsistencies between the datasets, we compiled a new reference mRNA abundance dataset and we present a novel approach to correct large microarray signals for a saturation bias. Accounting for ribosome density based on transcript length rather than ORF length improves the correlation of observed protein levels to translational activity. We discuss potential causes for the deviations of these correlations. Finally, we introduce a quantitative descriptor for protein degradation (protein half-life descriptor) and compare it to measured half-lives. The study demonstrates significant post-transcriptional control of protein levels for a number of different compartments and functional modules, which is missed when exclusively focusing on transcript levels. *Molecular & Cellular Proteomics* 3:1083–1092, 2004.

Recent publication of high-throughput data of the yeast *Saccharomyces cerevisiae* (1–3) opens the possibility to analyze the relationship between protein abundance, mRNA levels, and translational status on a genome-wide scale. Often mRNA abundance is used as a surrogate for protein amounts. Most studies employing cDNA microarrays assume that a high transcription of an ORF correlates with a high abundance of the corresponding protein. Previous studies either could

not find a correlation between protein and mRNA abundance (4) or the correlation was only weak (5–8). Greenbaum and coworkers (7) discuss three potential reasons for the lack of a perfect correlation between mRNA and protein levels: i) translational regulation, ii) difference of *in vivo* protein half-lives, and iii) the significant amount of experimental error including differences with respect to the experimental conditions. Understanding post-transcriptional regulation is crucial for correctly interpreting gene expression data. A full understanding of cell responses to external stimuli includes both transcription and translation regulation (6, 9, 10). It is important to distinguish processes regulating the overall translation (such as the total number or activity of available ribosomes) from protein-specific mechanisms of translation regulation (11–13). In addition to these translation-related mechanisms, selective degradation of proteins (protein turnover) regulates the cellular protein levels and affects the observed correlation between protein and mRNA abundance (14).

In this article, we explore the premise that protein levels are mainly determined by the corresponding mRNA levels, and we show to what extent translational regulation and selective degradation obliterate a perfect correlation between mRNA and protein abundance. Our focus lies on post-transcriptional regulation of protein amounts measured under standard log-growth conditions. For different compartments and functional modules we investigate to what extent protein levels are determined by the three factors mRNA concentration, translation rate (ribosome density and ribosome occupancy), and protein specific degradation. With "compartment" we always denote spatial subcellular structures, while "module" applies to functionally related genes and proteins. The value of analyzing protein-mRNA correlations for different functional modules and pathways has been noted previously (7–9). We demonstrate that the quality of protein-mRNA correlations varies among different cellular compartments and functional modules, and we quantify the contribution of post-transcriptional steps, including protein turnover, to the observed expression regulation of proteins. In addition, this study constitutes an example of how large-scale transcriptomics and proteomics data can be combined to gain new insights into cellular regulation.

From the [‡]Theoretical Systems Biology, Institute of Molecular Biotechnology, 07745 Jena, Germany; and [¶]Department of Biochemistry, National University of Ireland, Galway, Ireland

Received, July 28, 2004, and in revised form, August 19, 2004

Published, MCP Papers in Press, August 23, 2004, DOI 10.1074/mcp.M400099-MCP200

MATERIALS AND METHODS

Reference Data for mRNA Abundance—We obtained a reference mRNA abundance value for each ORF by combining 36 microarray datasets taken from the literature (supplemental material). All selected experiments are performed with wild-type yeast strains, grown in YPD medium under log-growth conditions without any stressors or chemical agents. These experiments generally served as control experiments. In order to obtain the reference data set the following steps were taken:

1. Microarray signal normalization according to Bolstad *et al.* (15)
2. Select median signal of each ORF as microarray-based standard signal (MSS)¹
3. Saturation correction

In order to correct the MSS for saturation effects, we compared the MSS values to values obtained by serial analysis of gene expression (SAGE) (16). Based on that comparison, we derived a correction for MSS larger than 13 copies per cell (see below).

More details on the data processing are given in the supplemental material. The complete mRNA reference dataset with the median, arithmetic mean, and different error estimates can be obtained from our web site (www.imb-jena.de/tsb/yeast_proteome).

Saturation Correction—Microarray measurements tend to underestimate high levels due to saturation during hybridization (5, 17). SAGE data on the other hand are inaccurate at low mRNA concentrations. In accordance with previous work (5, 7), we used SAGE measurements (16) to adjust the microarray measurements in the upper range. Plotting the two datasets against each other reveals a systematically increasing deviation between the datasets (Fig. S1, supplemental material). Average expression levels start to deviate significantly above 13 molecules per cell. Hence, the correction $y = 0.053 \cdot \text{MSS}^{2.098}$ (y is the corrected mRNA signal) is applied to all MSS above 13 mRNA molecules per cell. This correction is based on a regression of the deviation for values larger than 13. Our approach of adjusting microarray data has the advantage of using a signal-dependent correction. This way of adjusting the two datasets to each other is similar to state-of-the-art normalization methods used for microarray analysis (15). Its main advantage is that large mRNA values (and all properties derived from it such as the protein-to-mRNA abundance ratio (PRR)) get more realistic. Finally, the consideration of SAGE measurements allows us to include three additional mRNA expression values from SAGE for which no microarray measurement is available.

Grouping and Correlation Analysis—Proteins have been grouped according to their localization and function. Annotations to modules and compartments were done following the MIPS classification (mips.gsf.de, ftp files from March 2004), using only the most general first level annotation. Protein groups are compared on the basis of median values. Significance of correlation is measured with the Spearman rank correlation coefficient r_s throughout. No correlation has been calculated if less than 10 data points are available for the regression. If not stated differently the r_s mentioned in the text are statistically significant (1% confidence level).

Protein Half-life Descriptor (PHD)—By combining the models from (7, 18, 19) we set up the following differential equation:

$$\frac{d[P_i]}{dt} = k_p \cdot k_{transl,i} \cdot [mRNA_i] - k_{d,i} \cdot [P_i] \quad (\text{Eq. 1})$$

where $[P_i]$ and $[mRNA_i]$ are the protein and mRNA concentrations of the i th ORF; $k_{transl,i}$ is the product of ribosome density and ribosome

occupancy (fraction of mRNA bound to ribosomes) of the i th ORF (19); k_p is a genome wide translation constant (essentially it quantifies the speed of elongation) (19) and $k_{d,i}$ is an ORF-specific destruction rate. Because average protein levels are constant at steady state ($dP/dt = 0$), we can calculate the half-life descriptor as $PHD_i = k_p/k_{d,i} = [P_i]/([mRNA_i] \cdot k_{transl,i})$. The half-life of a protein is $\ln(2)/k_{d,i}$, thus the PHD is proportional to the *in vivo* half-life.

At the current stage, we refrain from expressing the PHD as a half-life in units of time, because the uncertainty of the underlying measurements is still too large to warrant its interpretation as an actual half-life. However, current data allow a classification of proteins into those with high and low stability. As the quality of protein abundance measurements improves also quantitative interpretation will become feasible.

RESULTS

In order to obtain a reference mRNA abundance for each ORF, we have compiled a set of 36 independent mRNA abundance measurements originating from different research groups. All measurements were performed using oligonucleotide microarrays and the same medium (YPD). This ensures high consistency of the data, while the selection from different research groups minimizes possible biases. Not all ORFs were measured in all studies, but the dataset contains at least 30 independent measurements for >6,000 ORFs. We applied a signal-dependent correction to high mRNA values in order to account for saturation effects (5, 17). This correction yields larger, more realistic absolute values in the upper range of the microarray data set (c.f. “Materials and Methods”). The resulting dataset is characterized by low noise: 94% of the abundance values have a coefficient of variation (CV) less than 1 and 99% of them have a CV below 2.5.

Protein abundance data are taken from Refs. 2 and 7. The protein abundances are much more uncertain than the reference mRNA levels, because they are based on fewer measurements and because the measurement techniques are less mature. To reduce also the error of protein abundances, we calculated the average of protein levels from Refs. 2 and 7 whenever possible (1,669 ORFs are contained in both datasets). Protein *versus* mRNA correlations are most significant when using the averaged protein levels (see supplemental material), suggesting that the averaged protein concentrations are in fact characterized by reduced noise. The availability of two measurements for some of the genes allows to get at least some idea of the uncertainty of the protein abundance values and properties derived from them. See the supplemental material for a detailed comparison of the available datasets.

The efficiency of translation can be measured by ribosome density on the mRNAs and the fraction of mRNA bound to ribosomes (ribosome occupancy) (1, 19–21). Hence, the observed protein levels should be better explained if in addition to mRNA abundance also ribosome density is taken into account. Previous calculations of ribosome density were based on ORF length (1, 20, 21). Using data from Refs. 1, 3, and 21, we calculated two different ribosome densities: Either

¹ The abbreviations used are: MSS, microarray-based standard signal; PHD, protein half-life descriptor; PRR, protein-to-mRNA abundance ratio; SAGE, serial analysis of gene expression.

TABLE I

Correlation of protein and mRNA abundance versus ORF length, transcript length, UTR length, ribosome occupancy, and ribosome density

Spearman rank correlation coefficients are shown. All correlations except for mRNA abundance versus ORF length are significant ($p < 0.01$, $n > 3600$). Abundance values are from the reference dataset as outlined in the main text.

	ORF length	Transcript length ^a	UTR length ^b	Ribosome occupancy ^c	Ribosome density 1 ^d	Ribosome density 2 ^e
mRNA abundance	0.03	-0.18	-0.23	0.38	0.13	0.17
Protein abundance	-0.10	-0.11	-0.16	0.39	0.33	0.37

^a Taken from Ref. 3.

^b Length of UTR = transcript length – ORF length.

^c Average ribosome occupancy from Ref. 1 and 21.

^d Ribosome density based on ORF length. Averages from Refs. 1 and 21.

^e Ribosome density based on transcript length. Numbers of ribosome per transcript from Refs. 1 and 21, transcript length from Ref. 3.

the number of ribosomes gets divided by the ORF length (Ribosome Density 1) or by the transcript length (Ribosome Density 2). In both cases, the average number of ribosomes from Refs. 1 and 21 was used. We find that correlations between protein levels versus ribosome density are strongest when using Ribosome Density 2 (Table I), suggesting that ribosome densities calculated on the basis of transcript length better describe translational efficiency.

Protein and mRNA Abundance as Indicators of Suppressed Translation—The genome-wide arithmetic mean protein abundance is 9,400 molecules per cell; the arithmetic mean mRNA abundance after applying our correction algorithm is 3.9 mRNA molecules per cell. However, we prefer to use median values for comparing compartments and functional modules, because the median is less affected by extreme values. Comparison of the available datasets shows that the median is much more stable against variations between the studies. The corresponding median values for the entire cell are 2,800 protein and 0.7 mRNA copies per cell, which are substantially below the arithmetic mean values. Median protein and mRNA levels vary significantly between compartments and functional modules (Fig. 1). The compartment and module-specific medians have to be interpreted with care: for instance, one might expect that the compartment “cytoskeleton” should have large average mRNA and protein levels. However, although some of those proteins are highly expressed they present only a small fraction of all gene products in that compartment.

The solid lines in Fig. 1 indicate the median values for the whole cell (*i.e.* the median of all ORFs taken together). Compartments or modules above or right of these lines in Fig. 1, *a* and *b* have relatively high protein or mRNA levels. While the majority of compartments and modules in Fig. 1, *a* and *b* lies in the bottom left or upper right quadrant, only few compartments reside in the bottom right quadrant. Interestingly, only the functional module “transposable elements/viral & plasmid proteins” lies in the upper left quadrant of Fig. 1*b*. Proteins in this part of the figure would be present in large amounts, although only a few mRNA molecules exist. On the other hand, there are compartments with relatively high mRNA but

only low protein levels. Such genes are efficiently transcribed, but either translation is suppressed or the translation products are rapidly degraded. In either case, the cell has the option to establish higher protein concentrations without having to transcribe additional mRNAs.

Protein-to-mRNA Ratio (PRR)—If there would be no post-transcriptional regulation of protein levels, the PRR were the same for all proteins. Thus, varying PRRs are indicative of post-transcriptional regulation.

The median PRR of all ORFs is 2,500 protein molecules per mRNA molecule, and the median values of modules and compartments vary by a factor of two around this cell-wide median (Fig. 1, *c* and *d*). The median PRR is smallest in the compartments “extracellular proteins” and “cell wall” and it is largest in the lipid particles. Among the functional modules, low PRRs appear in the module “protein synthesis” and the largest occur in the “energy” module. Large PRRs (*i.e.* efficient translation) of energy-related proteins is plausible, because many of these proteins are needed throughout the cell cycle and under all environmental conditions.

The PRR is similar to the “enrichment” proposed previously by Greenbaum *et al.* (22), who were using a smaller set of protein data. They came to similar conclusions with respect to enrichment in most modules. A significant difference between Greenbaum’s and our studies occurs only in case of “protein synthesis” where we find a PRR below the global average, while Greenbaum and colleagues observe an enrichment for these ORFs. The protein abundances used by Greenbaum and coauthors were obtained by gel-based techniques, which are known to have a bias toward proteins with longer half-lives and higher abundances (4, 5). Thus, proteins with low PRRs likely have been missed in previous studies. In addition, the two measures (median PRR, enrichment) are not identical, and differences with respect to growth conditions cannot be ruled out as a potential cause of this inconsistency.

Protein Abundance Is Weakly Correlated to mRNA Abundance and Translational Activity—The number of proteins synthesized per unit time depends on the number of mRNA molecules coding for this protein and on the respective translation rate. Therefore, we define “translational activity” as the

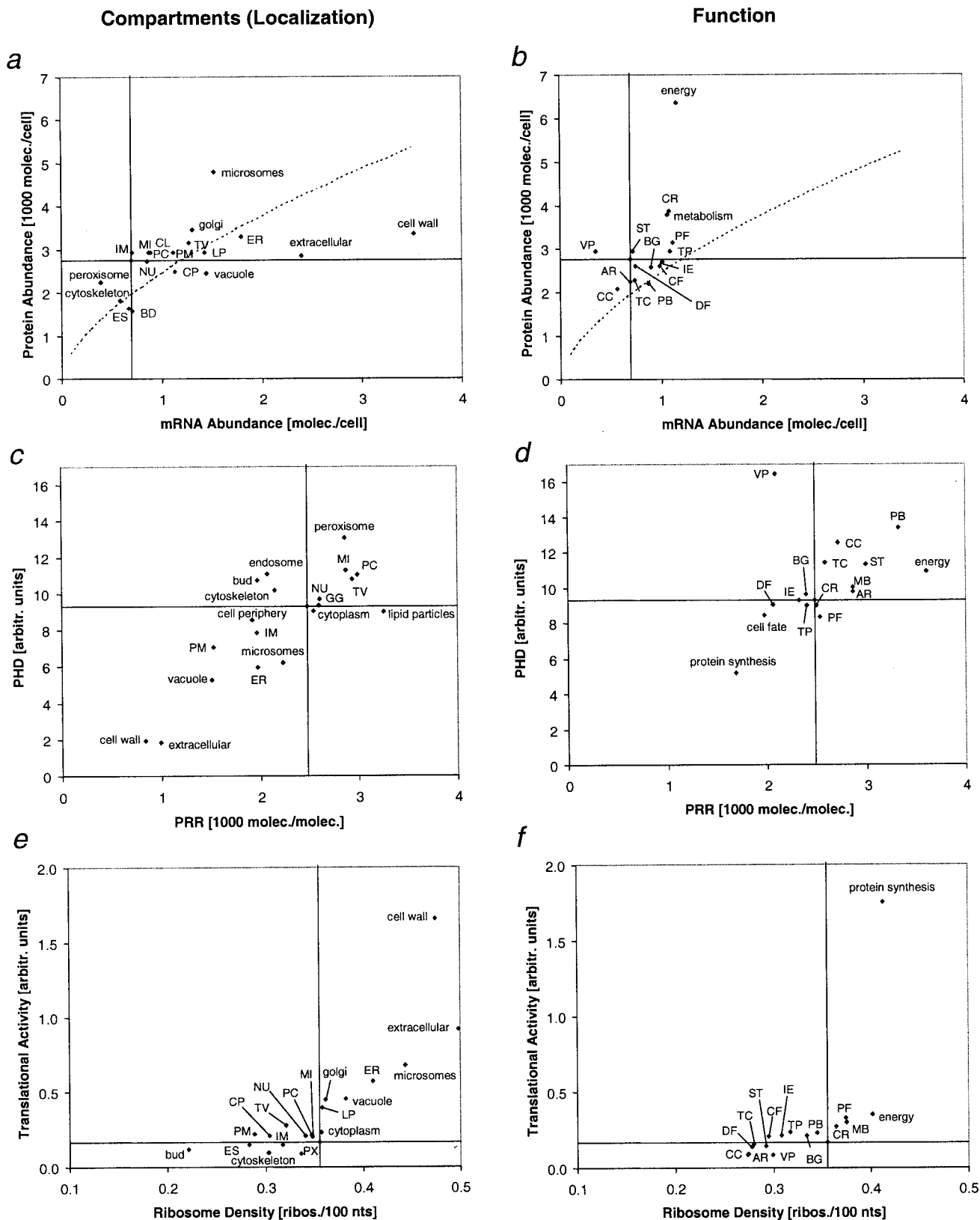


FIG. 1. Median values of protein properties grouped by localization (cell compartments, *left*) and function (*right*). *a* and *b*, protein levels versus mRNA abundance; *c* and *d*, PHD versus PRR; *e* and *f*, translational activity versus ribosome density. PRR is the number of proteins per mRNA molecule; PHD is PRR divided by translation rate; translation rate is the product of ribosome density and ribosome occupancy; ribosome

product of mRNA abundance and translation rate (*i.e.* ribosome density times ribosome occupancy (1, 19)) (Figs. 1 and 2). If there is little post-transcriptional regulation for most proteins of a certain compartment or functional module, protein levels will be correlated to the corresponding mRNA concentrations. A significantly improved correlation when additionally accounting for the translation rate indicates strong translational regulation.

We calculated Spearman rank correlation coefficients (r_s) of protein abundances *versus* mRNA levels and *versus* translational activities for the different protein groups (Fig. 2). It has been shown that protein and mRNA abundance data are not normally distributed (2, 5), therefore the Spearman rank correlation coefficient is more suitable than the Pearson correlation coefficient. For the whole cell the r_s are 0.580 and 0.596 for the protein-mRNA and protein-translational activity correlation, respectively. While the protein-mRNA correlation in most spatial compartments is relatively weak (six compartments have a r_s below 0.4), functional modules generally exhibit stronger correlations. It is plausible that expression regulation is more strongly synchronized within functionally homogeneous modules. The modules “metabolism,” “energy,” and “protein synthesis” exhibit the strongest correlation between mRNA and protein levels, suggesting that these modules are substantially regulated at the transcriptional level.

In agreement with previous observations (20, 21), ribosome density and ribosome occupancy are positively correlated with mRNA abundance (Table I). Thus, our analysis suggests a general tendency to increase mRNA levels and ribosome density in concert (“homodirectional changes” (20)). The different ribosome densities discussed above also yield two variants of translational activity. We find slightly improved correlations with mRNA and protein abundance when dividing the number of ribosomes per mRNA by the transcript length rather than the ORF length (Table I, supplemental material Fig. S4). A long transcript length compared with the ORF length is indicative for regulatory elements in the UTR (3). Such UTRs may be populated by ribosomes, *e.g.* if they contain upstream ORFs (12). Hence, regulatory elements on the UTR may reduce the effective ribosome density, which is partly being accounted for by using the transcript length instead of the ORF length. In the remainder we restrict our analysis to ribosome densities and translational activities based on transcript length.

Interestingly, there is no consistent improvement of the correlations when using translational activity instead of mRNA abundance (Fig. 2). In case of the functional modules, the correlation *versus* translational activity is mostly the same or it is slightly better than the correlation *versus* mRNA levels. There is, however, a strong, significant improvement of the correlation for the module “protein activity regulation,” indicating that translational control of protein amounts is highly important for these proteins.

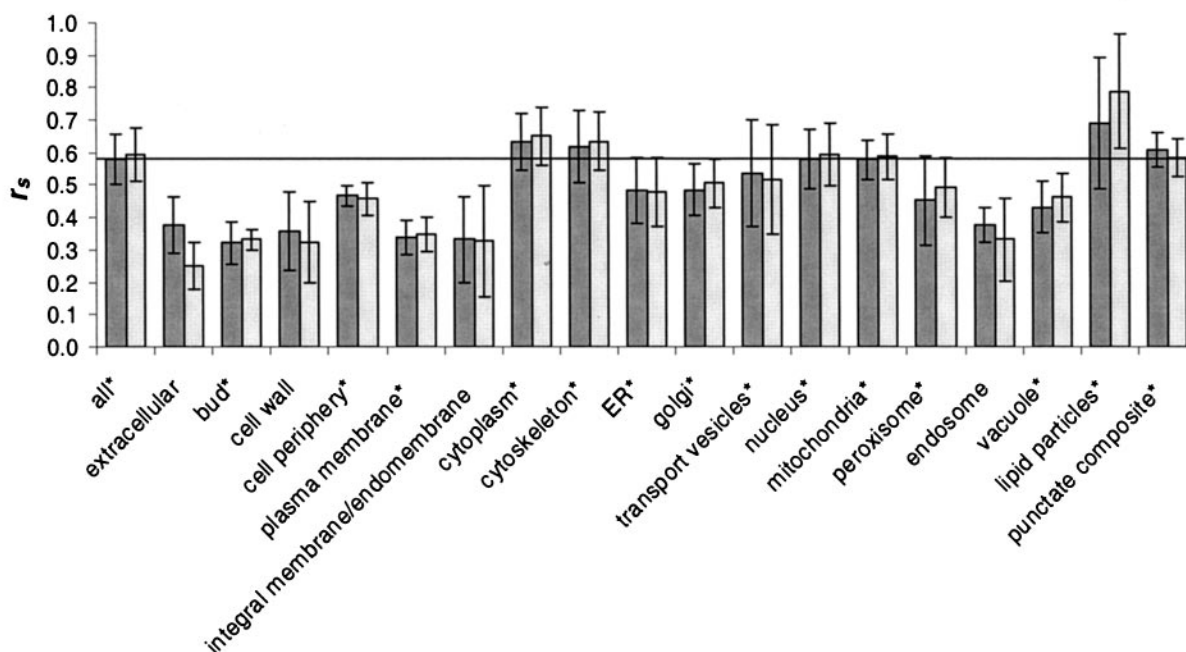
Evidence for Translation on Demand—When environmental signals require a quick cellular response, protein expression regulation via altering transcription may be too slow for urgently needed proteins. In such situations the cell constitutively maintains a sufficient level of mRNA, but blocks translation until the protein is actually needed (*e.g.* *GCN4* regulation (23)). Most of such proteins will be synthesized at low levels under standard conditions (*i.e.* without stressors), while mRNA should be present at reasonable amounts to allow for “translation on demand.” Translation on demand has been suggested for the yeast proteins *GCN4*, *HAC1*, and *ICY2* (23–25). By analyzing the correlations, mRNA levels, and ribosome densities we confirm this notion (Table II) and we identify new candidate genes that are potentially subject to translation on demand.

Because protein amounts of such ORFs depend on translation rather than transcription, an improved correlation is expected if ribosome densities are taken into account. The module “protein activity regulation” is a good example in that respect (Fig. 2*b*). Proteins of this module comprise regulatory proteins (such as GTPases or GDP/GTP exchange factors) that are needed at temporally varying amounts. Further evidence for translation on demand in this module can be gained by looking at the median mRNA levels and ribosome densities (Fig. 1, *b* and *f*). The module has a median mRNA abundance close to the cell average, but a low median ribosome density (0.27 ribosomes per 100 nts), resulting in a very low translational activity (median value 0.09). Thus, there is potential for a significant enhancement of translation in response to environmental signals.

Other modules involved in fast response to environmental stimuli (“cellular communication/signal transduction,” “cell rescue/defense/virulence,” and “interaction with cellular environment”) show similar patterns with respect to correlations, mRNA levels, and ribosome densities (Figs. 1 and 2), suggest-

density is ribosomes per transcript length; translational activity is the product of mRNA concentration and translation rate. *Solid lines* indicate median values for the whole cell (*i.e.* all available ORFs); *dashed lines* in *a* and *b* are power-law regressions of protein *versus* mRNA levels for all ORFs (exponent = 0.6). A power-law regression gives the best fit of the data. Deviations from the regression are due to variable post-transcriptional control or due to noisy data. The module “protein synthesis” is not shown in *b*; its median protein and mRNA abundance are 9,500 and 4.6 molecules per cell, respectively. Two-letter code for protein groups are as follows. Compartments: *BD*, bud; *CL*, cytoplasm; *CP*, cell periphery; *ER*, endoplasmic reticulum; *ES*, endosome; *GG*, Golgi; *IM*, integral membrane/endomembrane; *LP*, lipid particles; *MI*, mitochondria; *NU*, nucleus; *PC*, punctate composite; *PM*, plasma membrane; *PX*, peroxisome; *TV*, transport vesicle. Functional modules: *AR*, protein activity regulation; *BG*, biogenesis; *CC*, cell cycle/DNA processing; *CF*, cell fate; *CR*, cell rescue/defense/virulence; *DF*, differentiation; *IE*, interaction with cellular environment; *MB*, metabolism; *PB*, protein with binding function; *PF*, protein fate; *ST*, cellular communication/signal transduction; *TC*, transcription; *TP*, transport; *VP*, transposable elements/viral and plasmid proteins.

a (Localization)



b (Function)

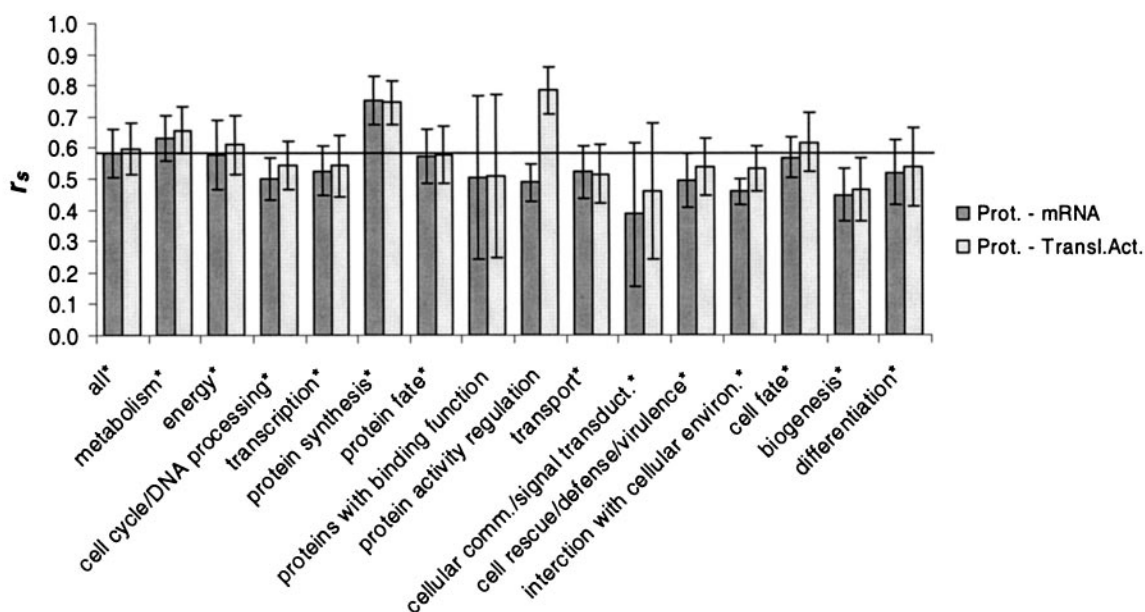


FIG. 2. **Correlation of protein abundance versus mRNA levels (dark gray) and versus translational activity (light gray).** Translational activity is the product of mRNA abundance, ribosome occupancy, and ribosome density, while ribosome density is the number of ribosomes per transcript length. Correlation is measured with the Spearman rank correlation coefficient (r_s) for (a) spatial cell compartments and (b) functional modules. Horizontal line indicates the correlation versus mRNA abundance for all available ORFs (“all,” first bar on the left). Protein groups with significant correlations ($p < 0.01$) are indicated by an asterisk (*) next to their name. Module “protein activity regulation” has no significant correlation with respect to mRNA abundance, but correlation versus translational activity is significant. Error bars were determined as follows: Correlations were separately calculated for the protein abundance data from Refs. 2 and 7. The error bars show the maximum deviation of the two r_s from the reference correlation. Variations among the different mRNA datasets were substantially smaller than for the two protein datasets.

TABLE II
Translation on demand for selected ORFs

Translation on demand has been suggested for the yeast proteins *GCN4*, *HAC1*, *ICY2*, and *CPA1* (23–25, 32). *GCN4* has an exceptionally high mRNA abundance but a very low ribosome density. Protein abundance has not been measured. The other proteins have low PRRs and also low ribosome densities, which is both indicative for translation on demand. The available data thus confirm the notion that these four proteins are subject to translation on demand. For comparison, we have included the median values for the whole cell and the data for *CPA2*, which likely is not regulated at the translational level (32). In contrast to the other proteins, *CPA2* has a very high PRR and a comparably high ribosome density, confirming constitutive expression of *CPA2*.

	mRNA abundance	Protein abundance	PRR	Ribosome density 2
Genome-wide median	0.7	2,800	2,500	0.36
<i>GCN4</i>	26	NA	NA	0.09
<i>HAC1</i>	5.2	9,000	1,700	0.26
<i>ICY2</i>	1.5	450	300	0.28
<i>CPA1</i>	4.8	4,900	1,000	0.23
<i>CPA2</i>	1.1	13,200	11,600	0.37

ing that a sub-set of the ORFs is regulated via translation on demand.

Protein Degradation Significantly Affects Protein-mRNA Ratios—Although protein levels in some modules are better explained by taking into account translation rates and mRNA abundance together, there remains a large amount of scatter. While this scatter must partly be attributed to uncertainty and variability of the measurements, also regulated protein turnover will be causative. We calculated a protein half-life descriptor (PHD, see “Materials and Methods”) for about 4,000 proteins. The PHDs are provided as supplemental data and they can be downloaded from our web site.

The PHD quantifies the deviation from a perfect relationship between observed protein abundance and translational activity. Assuming that Equation 1 is a valid approximation of the real kinetic and neglecting noise in the data, the PHD values are proportional to the *in vivo* half-lives of the proteins (see “Materials and Methods”). Hence, small PHDs correspond to short half-lives and large PHDs to long half-lives. The PHD values lie between 0.04 (Rpl21p) and 5,000 (Pck1p). This is a range over 5 orders of magnitude, which is not unrealistic given that *in vivo* half-lives vary from a few seconds up to many days (26). However, more than 95% of the PHDs are between 0.1 and 100, *i.e.* by far the most PHDs lie within just 3 orders of magnitude.

Because the PHDs are based on five measured properties (protein and mRNA abundance, ribosome density and occupancy, transcript length), we expect large uncertainty of the calculated PHDs. If, for instance, a gene was differently expressed during the mRNA and protein abundance measurements, the PHD derived from these values could substantially over- or underestimate the true value.² However, a qualitative

agreement between *in vivo* half-lives and the PHDs should be achievable for many proteins. Based on protein abundance data from Refs. 2 and 7 and on ribosome densities/occupancies from Refs. 1 and 21, we can estimate the uncertainty of PHDs for 1,554 proteins that are contained in all four datasets. Here we define the PHD uncertainty range as the deviation of the maximum and minimum PHD based on all possible parameter combinations from the four datasets. PHDs vary by less than a factor of 2 for 186 proteins and the PHDs of 453 proteins vary by more than a factor of 10. Thus, about 30% of the PHDs deviate by more than one order of magnitude.

Table III shows calculated PHDs along with measured protein half-lives taken from the literature. The table shows that large PHDs often correlate with long half-lives. As a rough rule we can conclude that proteins with half-lives in the range of a few minutes up to an hour usually have PHDs below 3. Relative differences between turnover rates of related proteins are often well reflected by the PHD (P1p *versus* P2p, Hmg1p *versus* Hmg2p, Rad51p *versus* Rad52p). Fig. 1, *c* and *d* show median PHDs for different compartments and functional modules. High PRRs are unlikely to coincide with low PHDs (bottom right quadrant). A low PHD means that the respective protein has a low stability, which renders high PRRs unlikely. The module “protein synthesis” has a median PHD significantly below the cell average and at the same time a very good protein-mRNA correlation (Fig. 2*b*), suggesting that both transcription and turnover of ribosomal proteins are strongly regulated (3, 8, 9). Our findings confirm previous suggestions that ribosomal protein amounts are regulated via degradation of excess proteins (6, 27–29).

DISCUSSION

Understanding all steps of protein expression regulation is important for a full elucidation of a cell’s response to environmental signals. The availability of genome-wide data of mRNA levels, translational status, and protein abundances in yeast allows us to perform an integrated analysis of post-transcriptional expression regulation in a whole cell. The mRNA levels used here are based on a large number of independent meas-

² In addition, the affinity tagging employed in Ref. 2 could potentially alter the *in vivo* half-lives of some proteins. According to Ref. 2, the large majority of proteins that are known to be short-lived is rapidly degraded also after tagging, “indicating that the tag is not inhibiting their proteolysis” (citation from the supplemental material for Ref. 2). However, such analysis does not exclude the possibility that the half-life of stable proteins are shortened or even increased.

TABLE III

Measured protein half-lives from the literature and calculated PHDs

For comparison to other proteins, some PHDs are shown even if no measured half-life is published (e.g. the GRX family).

Protein	Std-Name	PHD ^a	Measured half-life ^b
Far1p	YJL157c	0.32	~25 min
Gic2p	YDR309c	2.5	<30 min
Glc7p	YER133w	7.6	>3 h
Grx3p	YDR098c	4.9	
Grx4p	YER174c	15	
Grx5p	YPL059w	11	~4 h
Hmg1p	YML075c	5.4	>4 h
Hmg2p	YLR450w	0.93	50–60 min
Met30p	YIL046w	1.5	<20 min
P1βp	YDL130w	0.13	<15 min
P2αp	YOL039w	4.7	
P2βp	YDR382w	7.5	~5 h
Pap1p	YKR002w	18	~14 h
Rad51p	YER095w	9.5	>2 h
Rad52p	YML032c	7.8	15 min
Rpl40Ap	YIL148w	1.2	~2 h
Rpl40Bp	YKR094c	1.9	~2 h
Tat2p	YOL020w	2.6	>90 min

^a PHD is the number of proteins divided by the product of mRNA concentration, ribosome occupancy, and ribosome density for each ORF. Large PHD values correspond to long predicted *in vivo* half-lives.

^b References are given in the supplemental material.

urements. To our knowledge we have compiled the largest reference dataset of yeast transcript levels at vegetative growth published so far, which ensures high fidelity of the mRNA levels used. In contrast to the mRNA data, there currently exists only one (almost) genome-wide study of protein levels in yeast (2), and therefore protein abundance data are more uncertain than mRNA levels.

We compared protein abundances from Ref. 2 with previous studies, which have used different techniques and which were performed on a smaller scale (7, 22). Although there is no obvious bias in the data (e.g. when plotting the two datasets against each other), the measurements sometimes deviate by orders of magnitude (Fig. S2, supplemental material). A comparison of ribosomal proteins reveals the uncertainty inherent to the available protein abundance data. In our reference dataset, which is the average of the data from Refs. 2 and 7, protein concentrations of ribosomal proteins range from 3,000 to 300,000 molecules per cell. When looking at the protein datasets from Refs. 2 and 7 separately, the ranges are 450–600,000 and 500–100,000, respectively. Whatever dataset is chosen, all ranges deviate significantly from a 1:1 stoichiometry assumed for ribosomal subunits.³ This certainly is of concern when looking at individual proteins. However, aver-

³ Despite of experimental errors, this deviation may also be due to additional functions of the proteins. For instance, RPL40 also encodes a ubiquitin protein. Similarly, other ribosomal proteins may have additional functions, which would partly explain the scatter of protein abundances.

age results for compartments or modules are more stable against unbiased noise. A comparison of the module- and compartment-specific correlations (Fig. S3, Table S1, supplemental material) shows that the results with respect to differences between the protein groups are largely independent of the dataset chosen. For example, the finding that the median PHD of ribosomal proteins is comparably low is independent of the range of the values and the conclusion does not depend on the protein abundance dataset used. That means, even if we use only the protein abundances from Ref. 7, which are gel-based measurements, we find a median PHD significantly below the cell average (cell-wide median PHD, 6.4; median PHD for module “protein synthesis,” 1.7). Although the current protein abundance data hardly allow quantitative prediction, their improvement in the future will yield substantially more detailed insights and even quantitative conclusions can be drawn.

A rising number of studies is looking at ribosome density as a measure for translational efficiency (1, 20, 21). It is therefore important to clarify whether ORF length or transcript length is more relevant for the translational efficiency. This study provides statistical evidence that ribosome densities based on transcript length are the better descriptor of translational efficiency, possibly because transcript length correlates to the presence of regulatory motives in UTRs (11, 30). A longer UTR has a higher probability of containing such a regulatory element (3), which might explain the weak though statistically significant negative correlation between protein abundance and UTR length (Table I).

The correlation of protein to mRNA levels in individual compartments is often weaker than for the cell-wide average. This finding supports the assumption that there is no general correlation between mRNA and protein abundance, and it is consistent with previous studies analyzing specific biochemical pathways (9, 21). However, a more pronounced correlation of the two properties can be observed for certain functional modules (Fig. 2b). This is most likely an effect of co-expression and common translation regulation of these genes (8, 20, 21).

The fact that protein abundance in some cases is better correlated to translational activity than to mRNA copy numbers alone supports the hypothesis that regulation at the translational level can at least partly be described by ribosome occupancy and ribosome density (20, 21). However, in some cases this assumption may be wrong. Ribosomes could bind to mRNA without actively translating the message (e.g. *HAC1* (25)). As long as the number of ribosomes binding to such transcripts is low, conclusions are not significantly affected. In case of *HAC1*, the ribosome density is comparably low (0.26 per 100 nt) and even under normal conditions a small amount of Hac1p is detected (Table II, this is most likely the unspliced, noninduced form Hac1p^u (25)). Thus, in this example, a low ribosome density corresponds with a low protein abundance, which is in agreement with our assumptions. Future work should more in detail investigate the blocking of translation during the elongation step. If this mechanism

of translation regulation turns out to be relevant for a significant number of proteins, this could have severe implications also for previous studies that rest on the same assumptions as this work (20, 21). As long as only a small number of proteins is regulated in this way, our general conclusions based on the analysis of protein groups would not change.

In general, the protein abundance is only slightly better explained by translational activity than by mRNA abundance. Thus, the remaining scatter underlines the importance of other post-transcriptional control mechanisms. The large variability of the PHDs documents the importance of turnover for protein level regulation (13, 26). The PHD values calculated for all available proteins vary over 5 orders of magnitude. Even if we assume that, for instance, two orders of magnitude were due to noisy data, there would be a remaining variability of 3 orders of magnitude. If the large scatter is not completely random, protein turnover may be similarly important for protein abundance regulation than translational control, at least during vegetative growth.

Exploring the available data allows to identify compartments that are subject to translation on demand (e.g. “signal transduction”) or that are regulated via protein turnover (such as “cell wall”). Combinations could also be observed such as for ribosomal proteins, for which the data suggest joint transcriptional and post-transcriptional regulation. In agreement with previous findings (8, 9, 21), this study implies a significant primary response to environmental changes at the translational level, which remains undiscovered in the exclusive analysis of mRNA levels. In case of the module “protein activity regulation,” the relevance of translational control is particularly apparent, which should trigger more detailed analyses of the expression regulation of this group of proteins.

Finally, we demonstrate the possibility to calculate a PHD, which relates the steady-state protein level to the synthesis rate (18). We had to combine measurements from different laboratories, where growth conditions might not always be identical. In order to improve the precision of PHDs, independent experimental verification of protein levels is particularly important. Recently, other means of measuring protein turnover rates at a larger scale have been suggested (14, 18, 31), but up to now no half-life dataset of the size presented here has been published.

Our analysis helps to identify compartments where microarray experiments might be sufficient to predict protein level regulation as opposed to those where post-transcriptional regulation has to be taken into account. Future work should more precisely identify conditions under which a good correlation between gene transcription and protein abundance can be expected. Having this information available is crucial for correctly interpreting gene expression data, such as those obtained from microarray experiments.

Acknowledgments—We thank an anonymous reviewer for helpful comments.

* This work has been funded by the Federal Ministry of Education and Research, Germany. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ The on-line version of this manuscript (available at <http://www.mcponline.org>) contains supplemental material.

§ To whom correspondence should be addressed: Institute of Molecular Biotechnology, Beutenbergstr. 11, D-07745 Jena, Germany. Tel.: 49-3641-65-6331; Fax: 49-3641-65-6191; E-mail: beyer@imb-jena.de.

REFERENCES

- Arava, Y., Wang, Y., Storey J. D., Long Liu, C., Brown, P. O., and Herschlag, D. (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3889–3894
- Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O’Shea, E. K., and Weissman, J. S. (2003) Global analysis of protein expression in yeast. *Nature* **425**, 737–741
- Hurowitz, E. H., and Brown, P. O. (2003) Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae*. *Genome Biol.* **5**, R2
- Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730
- Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S., and Garrels, J. I. (1999) A sampling of the yeast proteome. *Mol. Cell. Biol.* **19**, 7357–7368
- Ideker T., Thorsson, V., Ranish, J. A., Christman, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934
- Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **4**, 117.1–117.8
- Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, G., Plouffe, D., Deciu, C., Winzeler, E., and Yates, III, J. R. (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
- Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., and Aebersold, R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **1**, 323–333
- Jansen, R., Greenbaum, D., and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12**, 37–46
- Sonenberg, N., and Dever, T. E. (2003) Eukaryotic translation factors and regulators. *Curr. Opin. Struct. Biol.* **13**, 56–63
- Vilela, C., and McCarthy, E. G. (2003) Regulation of fungal gene expression via short open reading frames in the mRNA 5’ untranslated region. *Mol. Microbiol.* **49**, 859–867
- Komar A. A., Lesnik, T., Cullin, C., Merrik, W. C., Trachsel, H., and Altmann, M. (2003) Internal initiation drives the synthesis of Ure2 protein lacking the prion domain and affects [URE3] propagation in yeast cells. *EMBO J.* **22**, 1199–1209
- Pratt, J. M., Petty, J., Riba-Garcia, I., Robertson, D. H. L., Gaskell, S. J., Oliver, S. G., and Beynon, R. J. (2002) Dynamics of protein turnover, a missing dimension in proteomics. *Mol. Cell. Proteomics* **1**, 579–591
- Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance an bias. *Bioinformatics* **19**, 185–193
- Velculescu V. E., Zhang, L., Zhou, W., Vogelstein J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997) Characterization of the yeast transcriptome. *Cell* **88**, 243–251
- Hekstra, D., Taussig, A. R., Magnasco, M., and Naef, F. (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res.* **31**, 1962–1968
- Gerner, C., Vejda, S., Gelbmann, D., Bayer, E., Gotzmann, J., Schulte-Hermann, R., and Mikulits, W. (2002) Concomitant determination of absolute values of cellular protein amounts, synthesis rates, and turnover rates by quantitative proteome profiling. *Mol. Cell. Proteomics* **1**, 528–537

19. Fraser, H. B., Hirsch, A. E., Giaever, G., Kumm, J., and Eisen, M. B. (2004) Noise minimization in eukaryotic gene expression. *PLoS Biol.* **2**, E137
20. Preiss, T., Baron-Benhamou, J., Ansorge, W., and Hentze, M. W. (2003) Homodirectional changes in transcriptome composition and mRNA translation induced by rapamycin and heat shock. *Nat. Struct. Biol.* **10**, 1039–1047
21. MacKay, V. L., Li, X., Flory, M. R., Turcott, E., Law, G. L., Serikawa, K. A., Xu, X. L., Lee, H., Goodlett, D. R., Aebersold, R., Zhao, L. P., and Morris, D. R. (2004) Gene expression in yeast responding to mating pheromone: Analysis by high-resolution translation state analysis and quantitative proteomics. *Mol. Cell. Proteomics* **3**, 478–489
22. Greenbaum, D., Jansen, R., and Gerstein, M. (2002) Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* **18**, 585–596
23. Hinnebusch, A. G., and Natarajan, K. (2002) Gcn4, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. *Eukar. Cell* **1**, 22–32
24. Beilharz, T. H., and Preiss, T. (2004) Translational profiling: the genome-wide measure of the nascent proteome. *Brief. Func. Gen. Prot.* **3**, 103–111
25. Kuhn, K. M., DeRisi, J. L., Brown, P. O., and Sarnow P. (2001) Global and specific translational regulation in the genomic response of *Saccharomyces cerevisiae* to rapid transfer from a fermentable to a nonfermentable carbon source. *Mol. Cell. Biol.* **21**, 916–927
26. Varshavsky, A. (1996) The N-end rule: Functions, mysteries, uses. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 12142–12149
27. Planta, R. J. (1997) Regulation of ribosome synthesis in yeast. *Yeast* **13**, 1505–1518
28. Nomura, M. (1999) Regulation of ribosome biosynthesis in *Escherichia coli* and *Saccharomyces cerevisiae*: Diversity and common principles. *J. Bacteriol.* **181**, 6857–6864
29. Nusspaumer, G., Remacha, M., and Ballesta, J. P. G. (2000) Phosphorylation and N-terminal region of yeast ribosomal protein P1 mediate its degradation, which is prevented by protein P2. *EMBO J.* **19**, 6075–6084
30. Rajkowitzsch, L., Vilela, C., Berthelot, K., Ramirez, C. V., and McCarthy, J. E. (2004) Reinitiation and recycling are distinct processes occurring downstream of translation termination in yeast. *J. Mol. Biol.* **335**, 71–85
31. Dantuma, N. P., Lindsten, K., Glas, R., Lelie, M., and Masucci, M. G. (2000) Short-lived green fluorescent proteins for quantifying ubiquitin/proteasome-dependent proteolysis in living cells. *Nat. Biotechnol.* **18**, 538–543
32. Messenguy, F., Feller, A., Crabeel, M., and Pierard, A. (1983) Control-mechanisms acting at the transcriptional and post-transcriptional levels are involved in the synthesis of the arginine pathway carbamoylphosphate synthase of yeast. *EMBO J.* **2**, 1249–1254