

## Comparing protein abundance and mRNA expression levels on a genomic scale

Dov Greenbaum\*, Christopher Colangelo<sup>†‡</sup>, Kenneth Williams<sup>†‡</sup> and Mark Gerstein<sup>†§</sup>

Addresses: \*Department of Genetics, <sup>†</sup>Department of Molecular Biophysics and Biochemistry, <sup>‡</sup>HHMI Biopolymer Laboratory and W. M. Keck Foundation Biotechnology Resource Laboratory, and <sup>§</sup>Department of Computer Science, Yale University, New Haven, CT 06520-8114, USA.

Correspondence: Mark Gerstein. E-mail: Mark.Gerstein@yale.edu. Kenneth Williams. E-mail: Kenneth.Williams@yale.edu

Published: 29 August 2003

*Genome Biology* 2003, 4:117

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2003/4/9/117>

© 2003 BioMed Central Ltd

### Abstract

Attempts to correlate protein abundance with mRNA expression levels have had variable success. We review the results of these comparisons, focusing on yeast. In the process, we survey experimental techniques for determining protein abundance, principally two-dimensional gel electrophoresis and mass-spectrometry. We also merge many of the available yeast protein-abundance datasets, using the resulting larger 'meta-dataset' to find correlations between protein and mRNA expression, both globally and within smaller categories.

Although some of the underlying technology for quantifying protein abundance was introduced almost thirty years ago [1,2], there has recently been a significant increase in the development of new tools. Concurrently, tools for analyzing mRNA expression are becoming more mainstream. The quantification of both of these molecular populations is not an exercise in redundancy; measurements taken from mRNA and protein levels are complementary and both are necessary for a complete understanding of how the cell works [3]. Additionally, as mRNA is eventually translated into protein, one might assume that there should be some sort of correlation between the level of mRNA and that of protein. Alternatively, there may not be any significant correlation, which, in itself, is an informative conclusion.

The two commonly used high-throughput methods for measuring mRNA expression, microarrays and Affymetrix chips, have both been extensively reviewed elsewhere [4-6]. There are also two basic methods for determining protein abundance; either based on two-dimensional electrophoresis or on mass-spectrometric methods (Table 1). We provide a brief review of these technologies and recent efforts to determine

correlations between quantified protein abundances and mRNA expression.

### Methods for determining protein levels Two-dimensional electrophoresis

Determining relative protein expression levels by conventional two-dimensional electrophoresis requires isoelectric focusing, SDS-polyacrylamide gel electrophoresis, staining, fixing, densitometry, and careful matching of the same spots on two or more gels. Differentially expressed spots are then excised and enzymatically digested, and the resulting peptides are identified using mass spectrometry. An attractive aspect of this approach is the low capital equipment cost, but a high level of expertise is needed to obtain reproducible gels, and two-dimensional electrophoresis is generally limited to proteins that are neither too acidic, too basic, nor too hydrophobic, and that are between 10 and 200 kDa in size, so that they are reliably separated on gels. Additionally, this approach detects only those proteins that are expressed at relatively high levels and that have long half-lives [7,8]. In one study using 40  $\mu$ g yeast lysate, the average protein

**Table 1****Overview of selected protein profiling technologies**

Technology	Type of labeling required	Ability to detect many post-translational modifications	Biomolecules that are optimally quantified	Approximate dynamic range (and reference)	Number of proteins/spots quantified (and reference)
Two-dimensional gel electrophoresis	Silver staining	Yes	Naturally occurring forms of proteins larger than 10 kDa	10 [9]	1,500 [8]
Differential two-dimensional fluorescence gel electrophoresis (DIGE)	<i>In vitro</i> with Cy2, Cy3 or CY5 fluorophores at primary amines	Yes	Naturally occurring forms of proteins larger than 10 kDa	10,000 [9]	1,100 [51]
SELDI- or MALDI-MS disease biomarker discovery	None	Yes	Naturally occurring forms of proteins smaller than 10 kDa	25	Not applicable
Isotope-coded affinity tag (ICAT) - LC/MS	<i>In vitro</i> with H <sup>1</sup> /D or C <sup>12</sup> /C <sup>13</sup> ICAT reagent at cysteine	No	Cysteine-containing tryptic peptides from digests of protein extracts	10,000*	496 [18]
N <sup>14</sup> /N <sup>15</sup> - LC/MS	<i>In vivo</i> at nitrogens in amino acids	Yes	Tryptic peptides from digests of protein extracts	10,000 [19]	872 [20]

\*Assumed to be similar to that for multidimensional protein identification. Abbreviations: SELDI-MS, surface-enhanced laser desorption ionization mass spectrometry; MALDI-MS, matrix-assisted laser desorption ionization mass spectrometry; LC/MS, liquid chromatography and mass spectrometry.

abundance detected was 51,200 copies per cell, with no proteins detected with abundances less than 1,000 copies per cell [8]. Given that 1,500 spots were resolved on a 1.0 pH unit gel [8], several gels covering different pH ranges would be needed to resolve a whole cell lysate. Given these limitations, conventional two-dimensional electrophoresis technology has limited potential for large-scale proteome analysis [8].

Two-dimensional fluorescence-difference gel electrophoresis (DIGE) utilizes mass- and charge-matched, spectrally resolvable fluorescent dyes (such as Cy3 and Cy5) to label two different protein samples *in vitro* prior to two-dimensional electrophoresis. Its main advantage over conventional two-dimensional electrophoresis is that both the control and the experimental sample are run in a single polyacrylamide gel. The samples are then imaged separately but can be perfectly overlaid without any 'warping' of the gels. This substantially raises the confidence with which protein changes between samples can be detected and quantified. Changes in the relative level of expression of a protein may be detected that are as little as 1.2-fold for large-volume spots [9]. Because detection is based on fluorescence, DIGE has a large dynamic range of about 10,000, which permits differential expression analysis of proteins that are present at relatively low copy number [9]. The limit of detection of DIGE for quantifying protein expression ratios is between 0.25 and 0.95 ng protein, which is similar to that for silver staining [9,10]. In a recent study [11], the relative levels of expression of approximately 1,050 protein spots were compared in 250,000 laser-dissected normal versus esophageal carcinoma cells. This analysis identified 58 spots that were

up-regulated by more than three-fold and 107 that were down-regulated by more than three-fold in cancer cells.

### Mass spectrometric approaches

#### Disease biomarker discovery

Current approaches to discovering protein or peptide markers of disease involve batch chromatography, matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) and statistical analysis of large numbers of disease versus normal serum or other biological samples. Most recent studies have relied on surface-enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF-MS) [12,13]. The SELDI approach [13] involves using a gold-coated chip with eight or sixteen 2 mm spots that are modified with chromatographic surfaces (for example anionic, cationic, hydrophobic, and so on). After spotting a few microliters of serum, any contaminants and salt are removed by washing with water, and the target is dried by adding a MALDI matrix solution, such as  $\alpha$ -cyano-4-hydroxy-cinnamic acid. In a study by Petricoin *et al.* [14] SELDI-MS analysis of serum from 50 control and 50 case samples from patients with ovarian cancer resulted in identifying five peptide biomarkers that ranged in size from 534 to 2,465 Da. The pattern formed by these markers was then used to correctly classify all 50 ovarian cancer samples in a masked set of serum samples from 116 patients who included 50 patients with ovarian cancer and 66 unaffected women. Similar promising results have been reported in studies of serum samples from breast and prostate cancer patients [12,15]. In a recent study [16], which compared the relative ability of several different statistical approaches to classify samples based on MS data, the disease biomarker approach

was extended to a conventional MALDI-MS platform. Although powerful, the disease biomarker approach does not provide accurate relative amounts of the control versus experimental biomarker, only the relative intensity difference.

#### *Isotope-coded affinity-tag-based protein profiling*

While both MALDI-MS-based disease biomarker discovery and DIGE comparatively profile the naturally occurring forms of peptides and proteins, isotope-coded affinity-tag (ICAT) analysis profiles the relative amounts of cysteine-containing peptides derived from tryptic digests of protein extracts. Because only a single tryptic peptide is needed to quantify the expression of the corresponding parent protein, the ICAT reagent utilizes a thiol protein-reactive group that attaches both a biotin tag and either nine  $^{12}\text{C}$  (light) or nine  $^{13}\text{C}$  (heavy) atoms to each cysteine residue. Following derivatization of the control protein extract with [ $^{12}\text{C}$ ]-ICAT reagent and the experimental extract with [ $^{13}\text{C}$ ]-ICAT reagent, the pooled samples are subjected to trypsin digestion followed by both cation and avidin chromatography. Liquid chromatography and tandem mass spectrometry (LC/MS/MS) is then used to identify ICAT peptide pairs and to quantify the relative  $^{12}\text{C}/^{13}\text{C}$  ratios. It is important to note that the ICAT approach provides the relative expression ratios of individual proteins under two conditions; it does not provide absolute protein concentrations, nor does it provide the ratio of the concentration of one protein relative to another in a single condition. A nice feature of this approach is that the *in vitro* incorporation of a stable isotope into one of the two samples being compared obviates the need to separately analyze the control and experimental samples by MS. Although a tryptic digest of a whole-cell human protein extract might produce more than 500,000 peptides, less than 100,000 of these might be expected to contain cysteine, but based on a search of the SwissProt database [17], less than 5% of human proteins lack cysteine and would therefore be missed (that is, more than 95% of proteins would include at least one cysteine-containing peptide).

ICAT results are analogous to those obtained by the use of two different fluorescent dyes in DNA microarray analysis of mRNA levels or DIGE analysis of protein expression. The largest number of proteins profiled so far using this approach with a single sample are the 491 proteins contained in microsomal fractions of naive and *in vitro* differentiated human myeloid leukemia cells [18].

#### *Multidimensional protein identification technology*

Multidimensional protein identification technology (MudPit) is similar to ICAT in that it utilizes cation-exchange prefractionation followed by reverse-phase (RP) high-performance liquid chromatography (HPLC) separation and MS/MS analysis [19]. In contrast to the ICAT approach, however, MudPit analyzes the entire mixture of typically digested proteins and utilizes tandemly coupled

(cation-exchange followed by reverse-phase) columns. A specific subset of peptides is eluted from the cation-exchange column, using a step gradient of increasing salt concentration, onto the front of the RP column. Peptides are then eluted from the RP column and enter the mass spectrometer for analysis. After the RP gradient is complete, the next step of the salt gradient releases another subset of peptides from the cation-exchange column onto the RP column, and the process repeats itself. Using this approach on the yeast proteome, Wolters *et al.* [19] identified 5,540 unique peptides from 1,484 proteins and demonstrated a dynamic range of detection of 10,000-fold. This method has been extended to comparative protein profiling by using *in vivo*  $^{14}\text{N}/^{15}\text{N}$  metabolic labeling [20,21]. Washburn *et al.* [20] used *Saccharomyces cerevisiae* grown in both  $^{14}\text{N}$ - and  $^{15}\text{N}$ -containing minimal media, and 2,264 peptides and 872 proteins were uniquely identified. Also, accurate  $^{14}\text{N}/^{15}\text{N}$  quantitation was determined for each peptide with an average standard deviation of 30%.

#### **Comparison of mRNA and protein levels**

Even with the significant developments in the technologies used to quantify protein abundance over the past couple of years, protein identification and quantification still lags behind the high-throughput experimental techniques used to determine mRNA expression levels. Yet, while mRNA expression values have shown their usefulness in a broad range of applications, including the diagnosis and classification of cancers [22,23], these results are almost certainly only correlative, rather than causative; in the end it is most probably the concentration of proteins and their interactions that are the true causative forces in the cell, and it is the corresponding protein quantities that we ought to be studying.

Primarily because of a limited ability to measure protein abundances, researchers have tried to find correlations between mRNA and the limited protein expression data, in the hope that they could determine protein abundance levels from the more copious and technically easier mRNA experiments. Alternatively, if there is definitively no correlation between mRNA and protein data, both quantities could be used as independent sources of information for use in machine-learning algorithms, for example, to predict protein interactions. To date, there have been only a handful of efforts to find correlations between mRNA and protein expression levels, most notably in human cancers and yeast cells; for the most part, they have reported only minimal and/or limited correlations.

One of the earliest analyses of correlation looked at 19 proteins in the human liver. Anderson and Seilhamer [24] found a somewhat positive correlation of 0.48. Another limited analysis, of the three genes *MMP-2*, *MMP-9* and *TIMP-1* in human prostate cancers, showed no significant relationship [25]. An additional cancer study [26] showed a

significant correlation in only a small subset of the proteins studied. Conversely, Orntoft *et al.* [27] found highly significant correlations in human carcinomas when looking at changes in mRNA and protein expression levels.

### Protein and mRNA correlations in yeast

Many of the present efforts at correlating mRNA and protein expression have been conducted in yeast using two-dimensional electrophoresis techniques. In particular, Gygi *et al.* [7] found that even similar mRNA expression levels could be accompanied by a wide range (up to 20-fold difference) of protein abundance levels, and *vice versa*. These results contrast with those of Futcher *et al.* [28], who found relatively high correlations ( $r = 0.76$ ) after transforming the data to normal distributions. In a previous analysis [29], we merged the data from both of these datasets (referred to as 2DE-1 [7] and 2DE-2 [28]), comparing the resulting new larger protein abundance set ('merged data-set 1') with a comprehensive mRNA expression dataset. The mRNA expression reference set was constructed through iteratively combining, in a non-trivial fashion, three sets that used Affymetrix chips and a SAGE dataset [29]. Using these reference datasets, we were able to do an all-against-all comparison of mRNA and protein expression levels, in addition to a number of analyses comparing protein and mRNA expression using smaller, but broad categories [29,30].

Given the difficult, laborious, and limiting nature of two-dimensional electrophoresis analysis, many of the newer protein abundance determinations have been done using MudPit and derivative technologies. Washburn *et al.* [31] used MudPit to analyze and detect 1,484 arbitrary proteins: they were able to detect a somewhat random sampling of proteins independent of abundance, localization, size or hydrophobicity (we refer to this dataset as MudPit-1). In a further experiment, the authors, comparing expression ratios for both proteins and mRNA levels, found that although they could not find correlations for individual loci, they could find overall correlations when looking at pathways and complexes of proteins that functioned together [21]. Peng *et al.* [32] analyzed 1,504 yeast proteins with a false-positive rate - misidentification of a protein - of less than 1% (we refer to this dataset as MudPit-2). In their analysis [32], they contrasted their methodology with that of Washburn *et al.* [31] with which there was significant overlap of proteins.

### A new merged dataset

Expanding upon our previous merged dataset, we constructed a new merged dataset (merged data set-2) using the two two-dimensional electrophoresis and two MudPit datasets described above. Succinctly (more information is available on our website at [33]), we transformed each of the protein-abundance datasets into more quantitative data by fitting each protein dataset individually onto the reference mRNA expression dataset. The MudPit-1 dataset was also fitted onto the more finely grained MudPit-2 dataset. Each

of the new, fitted datasets was then inversely transformed back into protein space. These derived protein datasets were then combined into a larger reference dataset; when we had more than one abundance value for an open reading frame (ORF), we chose the value from the dataset according to a prescribed quality ranking (see Figure 1). The resulting set contained protein abundance information for approximately 2,000 ORFs. (One caveat with the MudPit data: while quantitative analysis can be subsequently done on the results of MudPit experiments, MudPit data alone are only semi-quantitative, in that the number of peptides determined is relative to the actual protein abundance within the cell [31]. Some may therefore argue that MudPit alone is not optimal for a comparison with mRNA data. Nevertheless, we feel that our methodical merging process creates a quantitative and representative dataset that can be compared with the mRNA expression data.) Using the resulting data we could compare mRNA expression and protein abundance globally (Figure 1a) as well as looking at smaller, broad categories, such as function or localization (see Figure 1b,c). In particular, we show that some localization categories - for example, the nucleolus - have significantly higher correlations than the global correlation. Other localizations may present less of a correlation between mRNA and protein data - for example, the mitochondria - possibly reflecting the heterogeneous nature and function of the latter organelle. In terms of MIPS functional categories [34,35], we show that although some categories, such as cell rescue, show a lower correlation than the whole merged set, other functional categories, such as cell cycle, show a significant increase in correlation. Logically, this increased correlation reflects the co-regulated nature of the proteins in this functional category.

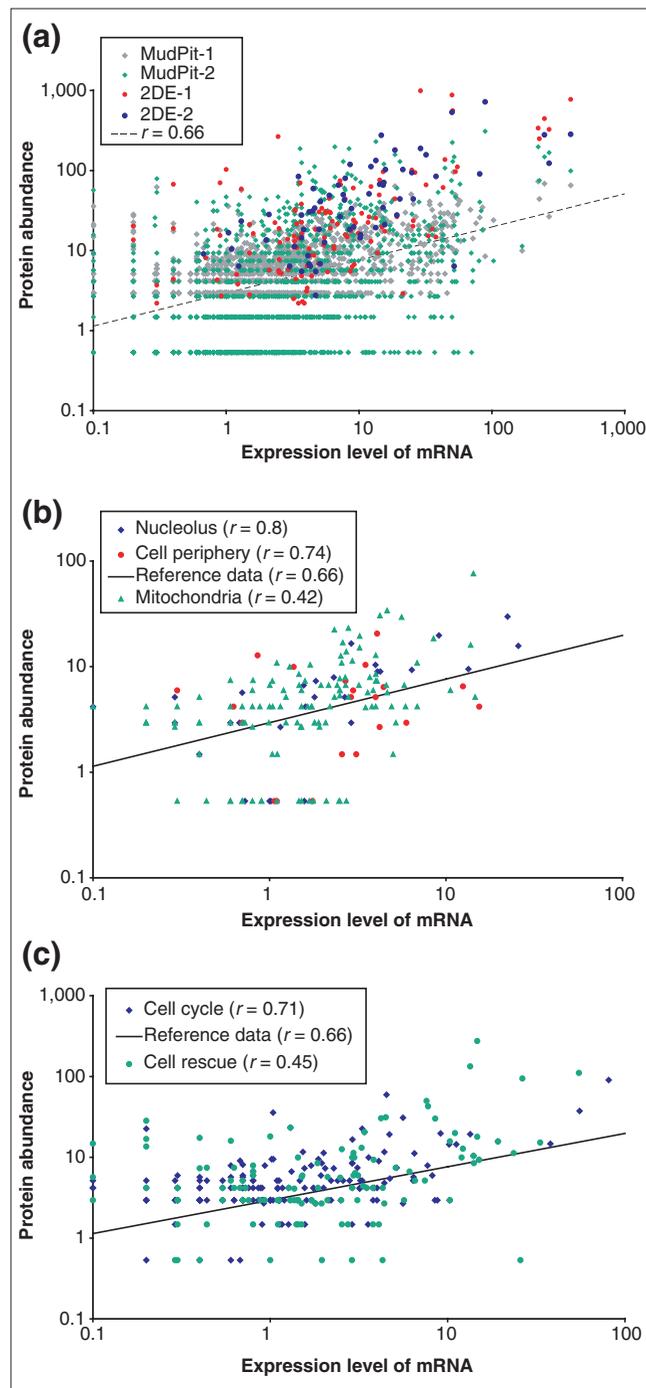
### Reasons for the absence of correlation

There are presumably at least three reasons for the poor correlations generally reported in the literature between the level of mRNA and the level of protein, and these may not be mutually exclusive. First, there are many complicated and varied post-transcriptional mechanisms involved in turning mRNA into protein that are not yet sufficiently well defined to be able to compute protein concentrations from mRNA; second, proteins may differ substantially in their *in vivo* half lives; and/or third, there is a significant amount of error and noise in both protein and mRNA experiments that limit our ability to get a clear picture [36,37].

Examining the first option - that there are a number of complex steps between transcription and translation - we looked at correlations between mRNA and protein abundance for those ORFs that had varied or steady levels of mRNA expression over the course of the cell cycle [38]. To normalize for the varied degrees of expression for different ORFs, we took the standard deviation divided by the average expression level as representative of the variation of each ORF over the course of the yeast cell cycle (Figure 2). Broadly speaking, the cell can control the levels of protein at

the transcriptional level and/or at the translational level. Logically, we would assume that those ORFs that show a large degree of variation in their expression are controlled at the transcriptional level - the variability of the mRNA expression is indicative of the cell controlling mRNA expression at different points of the cell cycle to achieve the resulting and desired protein levels. Thus we would expect, and we found, a high degree of correlation ( $r = 0.89$ ) between the reference mRNA and protein levels for these particular ORFs; the cell has already put significant energy into dictating

the final level of protein through tightly controlling the mRNA expression, and we assume that there would then be minimal control at the protein level. In contrast, those genes that show minimal variation in their mRNA expression throughout the cell cycle are more likely to have little or no correlation with the final protein level; the cell would be controlling these ORFs at the translational and/or post-translational level, with the mRNA levels being somewhat independent of the final protein concentration. And indeed, we found only minimal correlation between protein and mRNA expression for these ORFs ( $r = 0.2$ ).

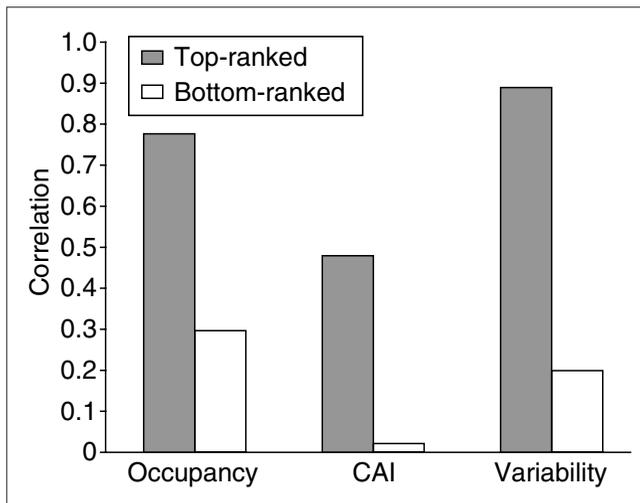


Furthermore, we found that those ORFs that have higher than average levels of ribosomal occupancy - that is that a large percentage of their cellular mRNA concentration is associated with ribosomes (being translated) - have well correlated mRNA and protein expression levels (Figure 2). These cases probably represent a situation wherein the cell, having significantly controlled the mRNA expression to produce a specific level of protein, will probably not also employ mechanisms to control the translation. Alternatively, those proteins that have very low occupancy rates have uncorrelated mRNA and protein expression; thus, given that the cell has not tightly controlled the mRNA expression for this ORF, it will dictate the resulting protein levels through rigorous controls of its translation (that is, through tight limits on occupancy) [39].

A second option for a general lack of correlation between mRNA and protein abundance may be that proteins have very different half-lives as the result of varied protein synthesis and degradation. Protein turnover can vary significantly depending on a number of different conditions [40]; the cell can control

**Figure 1**  
 Comparison of mRNA expression and protein abundance. **(a)** A plot comparing our mRNA reference expression set [29] with our newly compiled protein abundance dataset. The mRNA axis is in copies per cell; the protein axis is in thousand copies per cell. The protein dataset is the result of iteratively fitting two MudPit datasets (MudPit-1 [32] and MudPit-2 [31]) and two two-dimensional electrophoresis datasets (2DE-1 [7] and 2DE-2 [28]). Given the semi-quantitative nature of the MudPit data [31], we transformed the data into a more quantitative set by fitting each set individually onto our reference mRNA expression dataset. In addition, we fit the MudPit-1 dataset onto the more finely-grained MudPit-2 dataset. Each of the datasets was then moved back into 'protein space' using an inverse transformation derived from the 2DE-1 set, as this set has the most precise values. These datasets were then combined into the new reference abundance dataset. In cases in which there were overlapping values for a given ORF we used the dataset in accord with the following ordering: 2DE-1, 2DE-2, MudPit-2, MudPit-1. The resulting reference protein abundance dataset ( $N = 2044$ ) had a correlation of 0.66 with the mRNA reference dataset. **(b,c)** Additionally, we show that when looking at specific subsets (subcellular localization [52] or functional groups [34,35]) we can find both higher and lower correlations amongst these groups. The lower correlations are generally reflective of a more heterogeneous category. This analysis indicates that while correlations may be weak when looking at the global data, we tend to find higher correlations when looking at smaller well-defined subsets of ORFs. Further analysis is available at [33].

comment  
 reviews  
 reports  
 deposited research  
 refereed research  
 interactions  
 information



**Figure 2**

The differences in correlation between mRNA and protein expression values using novel categories. We see significant differences when looking at the highest and lowest ranking of groups of ORFs in the following categories: occupancy, CAI (codon adaptation index) value [45-47] and variability. Occupancy refers to the percentage of transcripts associated with ribosomes; we compared the correlation between the top 100 ORFs and the bottom 100 in terms of occupancy ( $r = 0.78$  versus 0.30). For the CAI, we compared the correlation between mRNA and protein for those ORFs with the highest CAI and those with the lowest ( $r = 0.48$  versus 0.02). Variability refers to the normalized standard deviation (that is, the standard deviation divided by the average expression level) for all ORFs in the cell-cycle expression dataset of Cho *et al.* [38]. Here, we compared the correlations between protein abundance and mRNA expression for the most variable compared with the least variable proteins ( $r = 0.89$  versus 0.20). We found significant differences between the correlations of mRNA and protein levels for the top and bottom ranking populations for each of the comparisons.

the rates of degradation or synthesis for a given protein, and there is significant heterogeneity even within proteins that have similar functions [41]. Recent efforts have been made to computationally measure these rates [42].

Simplistically, it can be presumed that the change in a protein's concentration over time will be equal to the rate of translation minus the rate of degradation. By analogy to concepts in chemical kinetics, we can approximate this equation:  $dP(i,t)/dt = SE(i,t) - DP(i,t)$ , where  $P$  is protein abundance  $i$  at time  $t$ ,  $E$  is the mRNA expression level of protein  $P$ ,  $S$  is a general rate of protein synthesis per mRNA, and  $D$  is a general rate of protein degradation per protein [43]. Additionally there are some experimental methods that can also be used to measure turnover and the translational control of protein levels [41-44].

Given the degenerate nature of the genetic code, there are many synonymous codons (codons that translate into the same amino acid). As the cell is biased in its usage of synonymous codons - that is, the usage of a subset of codons results in a higher level of mRNA expression, possibly as a result of

differing cellular tRNA levels [45] - the codon adaptation index (CAI), a measurement of codon usage, can be used to predict the expression of a gene [46] (we recently calculated new parameters for this model, with some improvement in predictive strength [47]). It is thought that the CAI will correlate differently with mRNA levels than with protein abundance levels due, in part, to protein turnover rates [48]. Ranking the ORFs in terms of their CAI value, we found that although those ORFs that ranked the highest in terms of CAI did not show a very strong correlation between mRNA and protein levels, they nevertheless showed a significantly higher correlation than ORFs that were ranked as having the lower CAI values ( $r = 0.48$  versus 0.02). The low correlations reflect the fact that the CAI will correlate differently for protein and mRNA values because of the additional cellular controls on protein translation, namely the effect of protein turnover rates. Nevertheless, the sizable difference in correlations between the two groups of ORFs with high- and low-ranking CAI values (Figure 2) shows that there is some relationship between mRNA and protein values, possibly indicating that highly expressed genes tend to result in a more correlated level of protein abundance than lower expressed ones.

Correlations have been found between the mRNA expression levels of different protein subunits within protein complexes [49]. This implies that there should be, in general, a correlation between mRNA and protein abundance, as these subunits provide a special case as they have to be available in stoichiometric amounts of proteins for the complexes to function. Thus, we believe that a major limitation to finding correlations is the degree of natural and manufactured systematic noise in mRNA and protein expression experiments. There is a continued effort to both describe and reduce this noise [50]. Meanwhile, in an attempt to get around the noise one could look at broad categories of proteins - for example, groups defined by function, structure, or localization - such that the background noise is cancelled out to some degree [29].

Although proteomics is still in its infancy, given the pace of technological advancement in protein quantification, mRNA expression analysis and noise reduction, more comprehensive correlation studies will soon be feasible. This will allow for more robust analyses of the relationship between mRNA expression and protein abundance values. Finally, to be fully able to understand the relationship between mRNA and protein abundances, the dynamic processes involved in protein synthesis and degradation have to be better understood; is the protein level changing because of a change in the rate of protein synthesis, or mRNA, or protein turnover? These questions need to be looked into further before we can appreciate in full the relationship between mRNA and protein abundance levels.

### Acknowledgements

This project was funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28186.

## References

- O'Farrell PH: **High resolution two-dimensional electrophoresis of proteins.** *J Biol Chem* 1975, **250**:4007-4021.
- Klose J: **Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals.** *Human-genetik* 1975, **26**:231-243.
- Hatzimanikatis V, Choe LH, Lee KH: **Proteomics: theoretical and experimental considerations.** *Biotechnol Prog* 1999, **15**:312-318.
- Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW: **Microarrays: biotechnology's discovery platform for functional genomics.** *Trends Biotechnol* 1998, **16**:301-306.
- McGall GH, Christians FC: **High-density genechip oligonucleotide probe arrays.** *Adv Biochem Eng Biotechnol* 2002, **77**:21-42.
- Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21**:33-37.
- Gygi SP, Rochon Y, Franz A, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19**:1720-1730.
- Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R: **Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology.** *Proc Natl Acad Sci USA* 2000, **97**:9390-9395.
- Tonge R, Shaw J, Middleton B, Rowlinson R, Rayner S, Young J, Pognan F, Hawkins E, Currie I, Davison M: **Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology.** *Proteomics* 2001, **1**:377-396.
- Gharbi S, Gaffney P, Yang A, Zvelebil MJ, Cramer R, Waterfield MD, Timms JF: **Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system.** *Mol Cell Proteomics* 2002, **1**:91-98.
- Zhou G, Li H, DeCamp D, Chen S, Shu H, Gong Y, Flaig M, Gillespie JW, Hu N, Taylor PR, et al.: **2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers.** *Mol Cell Proteomics* 2002, **1**:117-124.
- Adam BL, Vlahou A, Semmes OJ, Wright GL Jr.: **Proteomic approaches to biomarker discovery in prostate and bladder cancers.** *Proteomics* 2001, **1**:1264-1270.
- Issaq HJ, Veenstra TD, Conrads TP, Felschow D: **The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification.** *Biochem Biophys Res Commun* 2002, **292**:587-592.
- Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, et al.: **Use of proteomic patterns in serum to identify ovarian cancer.** *Lancet* 2002, **359**:572-577.
- Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW: **Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer.** *Clin Chem* 2002, **48**:1296-1304.
- Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H: **Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data.** *Bioinformatics*, in press.
- SwissProt [<http://www.expasy.ch/sprot/>]
- Han DK, Eng J, Zhou H, Aebersold R: **Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry.** *Nat Biotechnol* 2001, **19**:946-951.
- Wolters, DA Washburn MP, Yates JR 3rd: **An automated multidimensional protein identification technology for shotgun proteomics.** *Anal Chem* 2001, **73**:5683-5690.
- Washburn MP, Ulaszek R, Deciu C, Schieltz DM, Yates JR 3rd: **Analysis of quantitative proteomic data generated via multidimensional protein identification technology.** *Anal Chem* 2002, **74**:1650-1657.
- Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates JR 3rd: **Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2003, **100**:3107-3112.
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al.: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
- Macgregor PF, Squire JA: **Application of microarrays to the analysis of gene expression in cancer.** *Clin Chem* 2002, **48**:1170-1177.
- Anderson L, Seilhamer J: **A comparison of selected mRNA and protein abundances in human liver.** *Electrophoresis* 1997, **18**:533-537.
- Lichtinghagen R, Musholt PB, Lein M, Romer A, Rudolph B, Kristiansen G, Hauptmann S, Schnorr D, Loening SA, Jung K: **Different mRNA and protein expression of matrix metalloproteinases 2 and 9 and tissue inhibitor of metalloproteinases 1 in benign and malignant prostate tissue.** *Eur Urol* 2002, **42**:398-406.
- Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, Kardias SL, Giordano TJ, Iannettoni MD, Orringer MB, Hanash SM, et al.: **Discordant protein and mRNA expression in lung adenocarcinomas.** *Mol Cell Proteomics* 2002, **1**:304-313.
- Orntoft TF, Thykjaer T, Waldman FM, Wolf H, Celis JE: **Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas.** *Mol Cell Proteomics* 2002, **1**:37-45.
- Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI: **A sampling of the yeast proteome.** *Mol Cell Biol* 1999, **19**:7357-7368.
- Greenbaum D, Jansen R, Gerstein M: **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* 2002, **18**:585-596.
- Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M: **Interrelating different types of genomic data, from proteome to secretome: 'oming in on function.** *Genome Res* 2001, **11**:1463-1468.
- Washburn MP, Wolters D, Yates JR 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology.** *Nat Biotechnol* 2001, **19**:242-247.
- Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP: **Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome.** *J Proteome Res* 2003, **2**:43-50.
- Gerstein Lab - Supplementary data tables [<http://bioinfo.mbb.yale.edu/expression/prot-v-mrna/>]
- MIPS database [<http://mips.gsf.de/>]
- Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkötter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2002, **30**:31-34.
- Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17**:509-519.
- Szallasi Z: **Genetic network analysis in light of massively parallel biological data acquisition.** *Pac Symp Biocomput* 1999, 5-16.
- Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, et al.: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2**:65-73.
- Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D: **Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2003, **100**:3889-3894.
- Glickman MH, Ciechanover A: **The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction.** *Physiol Rev* 2002, **82**:373-428.
- Pratt JM, Petty J, Riba-Garcia I, Robertson DH, Gaskell SJ, Oliver SG, Beynon RJ: **Dynamics of protein turnover, a missing dimension in proteomics.** *Mol Cell Proteomics* 2002, **1**:579-591.
- Lian Z, Kluger Y, Greenbaum DS, Tuck D, Gerstein M, Berliner N, Weissman SM, Newburger PE: **Genomic and proteomic analysis of the myeloid differentiation program: global analysis of gene expression during induced differentiation in the MPRO cell line.** *Blood* 2002, **100**:3209-3220.
- Gerner C, Vejda S, Gelbmann D, Bayer E, Gotzmann J, Schulte-Hermann R, Mikulits W: **Concomitant determination of absolute values of cellular protein amounts, synthesis rates, and turnover rates by quantitative proteome profiling.** *Mol Cell Proteomics* 2002, **1**:528-537.
- Serikawa KA, Xu XL, MacKay VL, Law GL, Zong Q, Zhao LP, Bumgarner R, Morris DR: **The transcriptome and its translation during recovery from cell cycle arrest in *Saccharomyces cerevisiae*.** *Mol Cell Proteomics* 2003, **2**:191-204.
- Bennetzen JL, Hall BD: **Codon selection in yeast.** *J Biol Chem* 1982, **257**:3026-3031.
- Sharp PM, Li WH: **The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15**:1281-1295.

47. Jansen R, Bussemaker HJ, Gerstein M: **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* 2003, **31**:2242-2251.
48. Coghlan A, Wolfe KH: **Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*.** *Yeast* 2000, **16**:1131-1145.
49. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions.** *Genome Res* 2002, **12**:37-46.
50. Qian J, Kluger Y, Yu H, Gerstein M: **Identification and correction of spurious spatial correlations in microarray data.** *Biotechniques* 2003, **35**:42-44.
51. Yan, JX, Devenish, AT, Wait, R, Stone, T, Lewis, S, Fowler, S: **Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*.** *Proteomics* 2002, **2**:1682-98.
52. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, et al.: **Subcellular localization of the yeast proteome.** *Genes Dev* 2002, **16**:707-719.