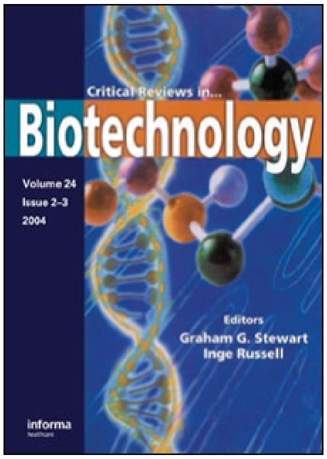

Critical Reviews in Biotechnology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400865

Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications

Lei Nie [ab]; Gang Wu [c]; David E. Culley [d]; Johannes C. M. Scholten [d]; Weiwen Zhang [d]

[a] Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington, DC, USA
[b] Department of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA
[c] Department of Biological Sciences, University of Maryland at Baltimore County, Baltimore, MD, USA
[d] Microbiology Department, Pacific Northwest National Laboratory, Richland, WA, USA

Online Publication Date: 01 April 2007
To cite this Article: Nie, Lei, Wu, Gang, Culley, David E., Scholten, Johannes C. M. and Zhang, Weiwen (2007) 'Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications', Critical Reviews in Biotechnology, 27:2, 63 - 75
To link to this article: DOI: 10.1080/07388550701334212
URL: http://dx.doi.org/10.1080/07388550701334212

Critical Reviews in Biotechnology, 27:63–75, 2007
Copyright © Informa Healthcare
ISSN: 0738-8551 print / 1549-7801 online
DOI: 10.1080/07388550701334212

informa
healthcare

Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications

Lei Nie
Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington DC, USA, and
Department of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA

Gang Wu
Department of Biological Sciences, University of Maryland at Baltimore County, Baltimore, MD, USA

David E. Culley, Johannes C. M. Scholten, and Weiwen Zhang
Microbiology Department, Pacific Northwest National Laboratory, Richland, WA, USA

ABSTRACT Recent advances in high-throughput technologies enable quantitative monitoring of the abundance of various biological molecules and allow determination of their variation between biological states on a genomic scale. Two popular platforms are DNA microarrays that measure messenger RNA transcript levels, and gel-free proteomic analyses that quantify protein abundance. Obviously, no single approach can fully unravel the complexities of fundamental biology and it is equally clear that integrative analysis of multiple levels of gene expression would be valuable in this endeavor. However, most integrative transcriptomic and proteomic studies have thus far either failed to find a correlation or only observed a weak correlation. In addition to various biological factors, it is suggested that the poor correlation could be quite possibly due to the inadequacy of available statistical tools to compensate for biases in the data collection methodologies. To address this issue, attempts have recently been made to systematically investigate the correlation patterns between transcriptomic and proteomic datasets, and to develop sophisticated statistical tools to improve the chances of capturing a relationship. The goal of these efforts is to enhance understanding of the relationship between transcriptomes and proteomes so that integrative analyses may be utilized to reveal new biological insights that are not accessible through one-dimensional datasets. In this review, we outline some of the challenges associated with integrative analyses and present some preliminary statistical solutions. In addition, some new applications of integrated transcriptomic and proteomic analysis to the investigation of post-transcriptional regulation are also discussed.

KEYWORDS transcriptomics, proteomics, integration, statistical

The research described in this paper was conducted under the Laboratory Directed Research and Development Program at the Pacific Northwest National Laboratory (PNNL), a multiprogram national laboratory operated by Battelle for the US Department of Energy under Contract DE-AC05-76RLO1830.

Address correspondence to Weiwen Zhang, Ph.D., Microbiology Department, Pacific Northwest National Laboratory, 902 Battelle Blvd., P.O. Box 999, Mail Stop P7-50, Richland, WA 99352, USA. E-mail: Weiwen.Zhang@pnl.gov

1. INTRODUCTION

The past decade has witnessed a paradigm shift in biological research. In contrast to traditional qualitative studies of the function of individual genes, biologists can now quantitatively investigate complex cellular processes at the systems level. The ability to monitor the abundance of various biological molecules and measure their variation between biological states on a genomic scale depends on recent advances in high-throughput technologies.

This has stimulated the rapid growth of computational biology and systems biology. Technologies to measure mRNA expression level, such as DNA microarray and Serial Analysis of Gene Expression (SAGE), can establish a global transcriptome profile once the genome sequence is available.[39] (For reviewing technical details of different array technologies, please refer to recently published reviews[35,36,71]). However, emerging evidence suggests that transcriptome profiling is necessary, but insufficient for comprehensive delineation of biological systems.[34] For example, transcript levels detected in mRNA profiling clearly do not reflect all regulatory processes in the cell, as post-transcriptional processes altering the amount of active proteins, such as synthesis, processing and modification of proteins, are not considered. Therefore, in addition to monitoring gene expression at the transcriptional level, large-scale analysis of the proteome is also important for the understanding of the cellular, metabolic and regulatory networks in living organisms. Proteome-based expression analysis is generally performed by two-dimensional-gel electrophoresis, in which proteins are separated according to their isoelectric point and mass. These approaches are generally labor-intensive. Furthermore, for technical reasons the current two-dimensional-gel based analyses mainly focus on a cytoplasmic subset of the cellular proteome in a limited range of molecular weights and isoelectric points. In most cases, only a few dozen proteins were identified by subsequent mass spectrometry analyses.[8,17,49,78,88] Fortunately, recent progress in mass spectrometry-based technology allows us to perform large-scale characterization of the proteome.[1,23] More specifically, high-performance liquid chromatographic (HPLC) fractionation of protein tryptic digests, followed by automated tandem mass spectrometry (MS/MS) to identify these peptide fragments, allows identification of several hundred proteins simultaneously from a single cellular extract.[23] Integrative analyses of genome-wide mRNA and protein expression patterns have enabled researchers to unravel global regulatory mechanisms and complex metabolic networks in living organisms.[2,37,60]

2. IS THERE ANY MEANINGFUL CORRELATION BETWEEN TRANSCRIPTOME AND PROTEOME?

There is a growing body of literature that describes methods for integrating and comparing transcriptomic and proteomic datasets, including microarray-derived gene expression patterns, the composition of protein complexes, and protein-protein interaction networks.[1,60] However, a fundamental and pressing question is how closely cellular protein abundance corresponds to mRNA concentrations. Although one would hypothesize that the correlation between mRNA expression levels and protein abundance will be strong based on the central dogma of molecular genetics, the support from early experimental data is not immediately apparent. Anderson and Seilhamer found a positive correlation of 0.48 on 19 proteins in the human liver,[4] while another limited analysis of the three genes *MMP-2, MMP-9* and *TIMP-1* in human prostate cancers showed no significant correlation.[51] An additional study on lung adenocarcinomas showed a significant correlation in a small subset of the proteins studied; 28 of the 165 protein spots (17%) or 21 of 98 genes (21.4%) had a statistically significant correlation between protein and mRNA expression ($r > 0.2445$; $p < 0.05$); among all 165 proteins the correlation coefficient values (r) ranged from -0.467 to 0.442 at individual gene level.[19] Conversely, Orntoft *et al.* found highly significant correlations in human carcinomas between mRNA and protein expression levels.[66] More recently, a group of researchers correlated the expression patterns of mitochondrial proteins in mammalian tissue with public microarray data, using a simpler present/absent test for concordance.[60] A positive score was assigned when a similar preferential tissue pattern was detected for both the corresponding mRNA and protein. By this scheme, 426 of 569 detected gene products were found to be concordant. In a study with the yeast *Saccharomyces cerevisiae*, 150 protein spots were identified by capillary liquid chromatography-tandem mass spectrometry (LC-MS/MS) and quantified by metabolic labeling and scintillation counting. Correlation between the protein abundance and their corresponding mRNA expression levels that were calculated from serial analysis of gene expression (SAGE) analysis was found to be only 0.356.[34] In addition, the study found that, for some genes, while the mRNA levels were of the same value the protein levels varied by more than 20-fold. Conversely, invariant steady-state levels of certain proteins were observed with respective mRNA transcript levels that varied by as much as 30-fold.[34] One criticism raised for the correlation studies of protein abundance and mRNA expression levels, however, is that an obvious bias in the data selection may greatly affect the research finding, wherein the average mRNA abundance of the detectable proteins was found to be nearly 5-fold higher than for all annotated mitochondrial genes, suggesting only high-abundance gene products strongly correlate.[60] The general consensus nowadays is that the correlation between transcriptomes and proteomes across large datasets was typically modest, and that mRNA levels could not be consistently relied upon to predict protein abundance.[20,23,34,60]

In *S. cerevisiae*, for example, three potential reasons for the apparently weak correlation between mRNA and protein expression levels were proposed. Two of these possibilities have a biological origin: i) translational regulation, or ii) differences in protein and mRNA half-lives *in vivo*. However, one proposed explanation for the weak correspondence between mRNA and protein levels in these studies is more of a methodological problem based on significant levels of experimental error, including differences with respect to the experimental conditions being compared.[12,31,32] This explanation was supported by a recent study designed to estimate the relative contribution of decay to steady-state transcript levels; the study has shown that more than 70% of transcripts showed no change in the steady-state levels following the treatments, suggesting the abundance of only a small portion of transcripts is significantly affected by the decay.[44] While it has been widely accepted that cellular protein levels should thus depend more directly on mRNA abundance in prokaryotes where transcription and translation are more tightly coupled than in eukaryotes,[30] one recent study in eukaryotes suggested that the discrepancy between protein and mRNA expression is most likely a result of the biology of gene expression rather than the measurement errors.[83] Using DNA microarrays together with quantitative proteomic techniques (ICAT reagents, two-dimensional DIGE, and MS), the researchers evaluated the correlation of mRNA and protein levels in two hematopoietic cell lines representing distinct stages of myeloid differentiation, as well as in the livers of mice treated for different periods of time with three different peroxisome proliferative activated receptor agonists. The results showed that the differential expression of mRNAs can be used to explain at most 40% ($r^2 = 0.62$) of the differential expression of proteins both in steady-state cell lines and under dynamic process of drug perturbations. To evaluate the effect of measurement errors, mRNA and protein error models and Monte Carlo simulation were applied to examine to what extent one can corrupt a perfect correlation solely by noise in the measurements. The results showed that the moderate correlation coefficient between mRNAs and proteins cannot be attributed solely to noise in the data and is more likely a reflection of the underlying biological mechanisms.[83] However, in order to further clarify the question, several flaws within these early correlation studies need to be addressed first. For example, in some comparisons the measurements of mRNA and protein abundance were not carried out using identical samples.[23,34,60] As such, variations from cell culture conditions and sample preparation further confound the correlation analysis between mRNA and protein abundance.[34] In addition, only a small set of protein abundance data (generally less than 200 mRNA/protein pairs), known to be biased towards highly expressed proteins, was used in the analyses.[34] Improvements on these issues have been found critical to achieve better measurements of mRNA–protein correlation; for example, integrated analysis of correlation between ratio changes of 289 proteins and their corresponding genes is 0.61 ($P < 1.3 \times 10^{20}$) based on a integrated analysis in yeast.[41] In a comparative study with pathogenic staphylococcal biofilm, 258 non-redundant proteins whose abundances were significantly changed in *Staphylococcus aureus* biofilm when compared with planktonic cells were identified. Clustering analysis of protein abundance with RNA expression data showed a good qualitative agreement between mRNA and protein abundance, with exception of only four proteins whose transcriptomic and proteomic data showed different trends.[69] The similar improvement was also observed in our recent study with *Desulfovibrio vulgaris*, in which mRNA expression level and protein abundance were determined for a pair of samples prepared in parallel to minimize the variations across experiments. Correlation between more than 400 proteins and their corresponding genes expressions was computed. The results showed that the correlation is about 0.45–0.53, indicating the presence of a reasonably high correlation.[63] Furthermore, it is unclear how much of an impact that technical limitations, such as the detection sensitivity and reproducibility of protein or mRNA measurements, have on the correlation results, and whether by using sophisticated statistical methods we can improve the detection of the correlation. In the past two years, significant efforts have been directed towards addressing some of these fundamental questions related to protein and mRNA correlations. In the remaining parts of this review, we will describe some of the progress being made in this area, with a particular focus on the statistical challenges associated with transcriptomic and proteomic data integration. In addition, several new applications of integrated transcriptomic and proteomic analysis will also be discussed.

3. CHALLENGES AND SOLUTIONS IN STUDYING TRANSCRIPTOME AND PROTEOME CORRELATION

3.1 Data Transformation and Normalization

Appropriate transformation of transcriptomic and proteomic datasets is a prerequisite for capturing the true correlation. For example, protein and mRNA abundance data has been shown to be non-normally distributed.[27,29] Therefore, without performing a transformation to normalize the protein and mRNA abundance data, the *Pearson* correlation can not be appropriately applied.[12] Fortunately, in most cases, simple logarithmic transformation will produce a close approximation to a normal distribution for transcriptomic or proteomic data, making it suitable for *Pearson* correlation analysis. However, to ensure normality, the Box-Cox transformation[14] of the original data or of the log-transformed data could also be utilized as a valid tool to stabilize variances.[27,47] In the case of semi-quantitative proteomic abundance data, such as peptide hits,[28] a square root transformation may be more appropriate to stabilize the variances.[68] The square root transformation is often used to stabilize the variance when the variance of a variable is proportional to the mean.[15] As a matter of fact, both logarithmic and square root transformation can be viewed as special cases of the Box-Cox transformation.

$$\tau(x) = \frac{x^{\lambda} - 1}{\lambda}, \quad \text{if } \lambda \neq 0; \quad \tau(x) = \log(x) \text{ if } \lambda = 0 \quad (1)$$

When $\lambda = 0$, this is actually the logarithm transformation; when $\lambda = 1/2$, it is equivalent to the square root transformation. The Box-Cox transformation is basically a power transformation, but can be done in such a way as to make it continuous with the parameter λ at $\lambda = 0$. Typically through a computer program, e.g. R program (http://www.r-project.org/), one can automatically select the best Box-Cox transformation by automatically estimating the best value for the parameter $\widehat{\lambda}$ and then using $\tau(x, \widehat{\lambda})$ to normalize the data. Although little work has been done on data normalization of proteomic data, the importance of transformation for the mRNA expression data has been discussed in various previous studies.[25,40,57]

It has been argued that measurements of mRNA expression and proteomic abundance of a gene may be influenced by the length of the transcript or protein. Thus, normalization of transcriptomic or proteomic data by transcript/protein length would be a necessary step prior to correlation analysis. This is especially true for data obtained from cDNA microarray in which cDNA fragments of different lengths are used as probes. Even with oligonucleotide microarrays in which equal length of probes were used, it has been suggested that in general the correlation between transcriptome and proteome datasets was improved after size normalization (Table 1).[63,64,93] For example, in a bacterial *D. vulgaris* dataset the *Pearson* correlation between un-normalized proteomic abundance and mRNA was 0.45–0.53 in various growth conditions. After size normalization the correlation was increased to 0.58–0.66. The differences between un-normalized and normalized correlation are

TABLE 1 **Correlations of transcriptomic and proteomic data: effects of normalization**[a]

Conditions	*Pearson*	*Spearman*	*Kendall*
Un-normalized datasets[b]			
D. vulgaris & LE	0.53	0.45	0.32
D. vulgaris & FE	0.45	0.39	0.27
D. vulgaris & LS	0.50	0.46	0.32
S. cerevisiae[d]	0.62	0.58	0.42
Proteomics dataset normalized only[c]			
D. vulgaris & LE	0.61	0.57	0.40
D. vulgaris & FE	0.54	0.54	0.37
D. vulgaris & LS	0.62	0.62	0.45
S. cerevisiae[d]	0.70	0.67	0.49
Transcriptomic and proteomic dataset normalized[d]			
D. vulgaris & LE	0.66	0.64	0.45
D. vulgaris & FE	0.58	0.58	0.40
D. vulgaris & LS	0.65	0.66	0.46
S. cerevisiae[d]	0.73	0.71	0.53

[a]The *Pearson*, *Spearman*, and *Kendall*'s correlation coefficients were computed for the correlation of mRNA expression and proteomic abundance levels. In all cases, logarithmic transformation was applied after the size normalization. b) No normalization was applied. c): Proteomic data normalized by protein size (number of amino acids); d: microarray data was normalized by gene length (bp) and proteomic data normalized by protein size (number of amino acids). LE: *D. vulgaris* grown on Lactate-based medium at Exponential phase; FE: *D. vulgaris* grown on Formate-based medium at Exponential phase; LS: *D. vulgaris* grown on Lactate-based medium at Stationary phase; d): The mRNA and protein abundance data was obtained from reference 12 (Beyer et al., 2004. *Mol. Cell. Proteomics.* 3: 1083–1092).

statistically significant (p-value < 0.0003). Similar results were also obtained when using mRNA and protein abundance data of *Saccharomyces cerevisiae* for the correlation study (Table 1).[12] Consistently, Munoz *et al.*[62] suggested that the gene length was a significant factor contributing to biases in measured gene expression rates in *Caenorhabditis elegans*.

Protein abundance can be measured by liquid chromatography-tandem mass spectrometry (LC-MS/MS) using either peptide hits or by isotopic signal after isotope labeling.[16,28] Both methods would inevitably introduce amino acid composition dependence, *e.g.*, lysine- and arginine-rich proteins produce more tryptic peptides. In addition, ionization efficiency and fragmentation properties are also amino-acid sequence dependent. However, the effects of the amino acid composition on transcriptome-proteome correlation had not previously been fully examined. As an initial step to address this issue, Nie *et al.*[64] recently introduced an "effective peptide number" normalization method for the peptide-hit based abundance data in the *D. vulgaris* proteome. The "effective peptide number" is derived as follows: First, each protein was analyzed for the possible cleavage sites by trypsin using a computational method (*e.g.* the PeptideCutter program (http://ca.expasy.org/tools/peptidecutter/) and the number of peptides of all sizes after trypsin cleavage was determined. Second, the number of peptides whose molecular mass falls into the mass (m/z) detection range of mass spectrometry was summed as the correction factor for the protein under scrutiny.[64] Finally, the "effective peptide number" of a protein is computed as the peptide hits of that protein divided by its correction factor. Surprisingly, although normalization of the data by digestion sites has indeed improved the overall correlation between mRNA and protein abundance in *D. vulgaris* datasets, only a small difference was found when compared to normalization by gene length, suggesting that the two normalization methods may be overlapped in removing the bias.[64] Nevertheless, the idea of normalizing protein abundance data by digestion sites is worthy of more rigorous evaluation using quantitative proteomics data from other species.

3.2 Effects of Measurement Errors on the mRNA-Protein Correlation

Measurement errors in microarray and proteomic analysis have long been considered to have the potential to be a major cause of poor correlation between mRNA concentration and protein abundance.[23] In addition to various experimental efforts to improve accuracy of transcriptomic and proteomics measurements, statistical methods were also employed. For example, in a study using human serum samples an analytic approach based on a two-component error model was proposed for LC-MS proteomic data.[70] This model proposed to divide the measurements errors into sample preparation errors and analytical equipment errors.[3,83] Their simulation studies indicated that the correlation between mRNA and protein can be easily corrupted by the noise if not analyzed properly.[83] However, to the best of our knowledge, no experimental study has been reported regarding the effects of the measurement errors on mRNA-protein correlation. One statistical method was recently proposed to deal with the measurement error in the correlation analysis, in which correlation was found dependent on the ability to identify significant changes in m RNA expression ratio. The high significant correlation (*Pearson* 0.8) was only observed for the group of genes with high average confidence values (λ) of 40.[36] With the goal of stimulating further investigations, preliminary data from our group are presented below to serve as an example of the problems and potential solutions to methodological and statistical biases facing the application of correlation analyses in integrative analysis of multiple levels of gene expression. With multiple regression analysis, Zhang *et al.*[92,93] determined that in one pair of corresponding transcriptomic and proteomic datasets collected from *D. vulgaris*, measurement errors in mRNA quantifications explained about 9–22% of the variations in mRNA-protein correlation, whereas errors in quantifying protein abundance contributed up to 34–44% of the variations in mRNA-protein correlation.[63] More interestingly, measurement errors contributed significantly to mRNA-protein correlation even though transcriptomic data appears to have good reproducibility among replicates (*Pearson* correlation coefficient 0.96 to 0.99 for transcriptomic replicates, and 0.87 to 0.88 for proteomic replicates.[63]) These results suggest that correlation between mRNA and protein abundances is very sensitive to measurement errors and, therefore, this phenomenon requires further attention.

One direct impact of measurement errors is the under-estimation of the correlation between mRNA and protein abundance. This can be easily proven using the following reasoning: Assume that the *observed* protein abundance, after appropriate data transformation, for the ith gene, y_i, consists of the *real* protein abundance level p_i and measurement error ε_{ip}, *i.e.*, $y_i = p_i + \varepsilon_{ip}$. Likewise, the *observed* mRNA abundance, after transformation, for the ith gene, x_i, consists of the *real* mRNA abundance level m_i and measurement error ε_{im}, *i.e.*, $x_i = m_i + \varepsilon_{im}$. Assume that the transformed mRNA and protein abundance and their measurement errors will follow a normal distribution, *i.e.*, $p \sim N(\mu_p, \sigma_p^2)$, $\varepsilon_p \sim N(0, \sigma_{p\varepsilon}^2)$, $m \sim N(\mu_m, \sigma_m^2)$, and $\varepsilon_m \sim N(0, \sigma_{m\varepsilon}^2)$, where μ is the mean and σ^2 is the variance (var). Because measurement errors are random and independent of the real abundances of mRNAs or proteins, the covariance (cov) between the real abundances of mRNA and proteins will be the same as that between the observed (*i.e.* measured) abundances of mRNA and proteins. That is,

$$\mathrm{cov}(y, x) = \mathrm{cov}(p, m) \tag{2}$$

where we used the facts cov(p, ε_m) = 0, cov(m, ε_p) = 0, and cov(ε_m, ε_p) = 0 based on the independence of measurement errors and abundance levels. Thus the *Pearson* correlation coefficient ρ_{yx} between *observed* abundances of mRNA and protein is,

$$\mathrm{corr}(y, x) = \frac{\mathrm{cov}(y, x)}{\sqrt{\mathrm{var}(y)\ \mathrm{var}(x)}} = \frac{\mathrm{cov}(p, m)}{\sqrt{\mathrm{var}(y)\ \mathrm{var}(x)}} \tag{3}$$

Note that this is different from the actual correlation between the *real* abundances of mRNA and protein (equation 4),

$$\mathrm{corr}(p, m) = \frac{\mathrm{cov}(p, m)}{\sqrt{\mathrm{var}(p)\ \mathrm{var}(m)}} \tag{4}$$

Thus, when measurement errors are not zero, the variance of observed mRNA and protein abundances [var(y) and var(x)] will be greater than that of the real values [var(p) and var(m)]. Hence, ρ_{yx} will be less than ρ_{pm}. Put another way, the correlation between observed abundances of mRNA and protein is increasingly under-estimated with an increasing magnitude of measurement errors.

In practice, the variability of measurement errors, namely $\sigma_{m\varepsilon}^2$ and $\sigma_{p\varepsilon}^2$, may be much smaller than the variability of the measurements σ_m^2 and σ_p^2. In other words, the within-gene variability may be much smaller than the between-gene variability. Thus the difference between ρ_{yx} and ρ_{pm} could be negligible. However, due to limitations in the assay technology, this is not the case for quantification of mRNA and protein molecules, especially for protein abundance. For instance, it has been observed that in *D. vulgaris* the variability of measurement error was 10–25% of the variability of the measurement. Therefore, this high level of measurement error could not be ignored under this circumstance.

In theory, it is possible to adjust the correlation between observed mRNA and protein abundances with an appropriate scaling factor to approximate the actual correlation coefficient. For instance, from equations (3) and (4) we are able to derive the following equation,

$$\mathrm{corr}(p, m) = \mathrm{corr}(y, x)\frac{\sqrt{\mathrm{var}(y)\mathrm{var}(x)}}{\sqrt{\mathrm{var}(p)\mathrm{var}(m)}} \tag{5}$$

Since var(y) = var(p)+ var (ε_p) and var(x) = var(m)+ var (ε_m), equation (4) can be transformed to,

$$\mathrm{corr}(p, m) = \mathrm{corr}(y, x) \sqrt{(1 + \mathrm{var}(\varepsilon_p)/\mathrm{var}(p)) \times (1 + \mathrm{var}(\varepsilon_m)/\mathrm{var}(m))} \tag{6}$$

Equation (6) is identical to the following,

$$\mathrm{corr}(p, m) = \mathrm{corr}(y, x)\sqrt{(1 + \sigma_{p\varepsilon}^2/\sigma_p^2) \times (1 + \sigma_{m\varepsilon}^2/\sigma_m^2)} \tag{7}$$

As a result, the correlation between the *real* abundances of mRNA and protein is inflated by a factor of $\sqrt{(1 + \sigma_{p\varepsilon}^2/\sigma_p^2) \times (1 + \sigma_{m\varepsilon}^2/\sigma_m^2)}$, a value greater than 1, from the *observed* correlations. This method of compensating for experimental error was tested using the transcriptomic and proteomic data of *D. vulgaris* described previously.[92,93] Although the correlation coefficients were only slightly increased if experimental errors were taken into consideration (Table 2), as a method it demonstrates how the measurement error could shrink the real correlation and how measurement error can be corrected with statistical methods. In addition, it should be pointed out that the correlation correction approach

TABLE 2 Measurement error-adjusted correlations of mRNA and protein abundance*

Conditions	*Pearson*	*Spearman*	*Kendall*
Transcriptomic and proteomic dataset normalized			
LE	0.68	0.66	0.47
FE	0.61	0.61	0.43
LS	0.68	0.69	0.48

*LE: *D. vulgaris* grown on Lactate-based medium at Exponential phase; FE: *D. vulgaris* grown on Formate-based medium at Exponential phase; LS: *D. vulgaris* grown on Lactate-based medium at Stationary phase.

we proposed here, which assumed that measurement errors are random and independent of abundances (in the log scale), can only remove random noises. Further development of the method is still needed to deal with the distortions due to systematic and non-independent noises.

3.3 Compensating for Missing Proteomic Data

While microarray analysis generally produces data on transcript levels for most of the genes in a genome, proteomic datasets are often incomplete due to the imperfect identification of coding sequences within a genome and the limited sensitivity of current peptide detection technologies.[90] Typically current technologies allow detection of only several hundreds or less of proteins.[41,74,75,93] Several approaches, such as *k* nearest neighbor,[68] least square method,[13,86] and local least squares method, have been proposed to deal with the missing data issue in microarray datasets.[45] As the studies found,[11] *k* nearest neighbor method requires a sufficient percent of data to be present before imputation of any gene of similar expression can be performed. It was recommended that any genes or arrays with a very large fraction (*e.g.* 25%) of missing values be excluded from the data set. Similar restrictions on missing fraction were also suggested for the least square or local least square approaches.[13,45] With the large fraction of data missing in the typical proteomic dataset, most of the approaches fail to provide reasonable estimation for the missing data, and more sophisticated methods need to be developed to address the issue. Some methods have been adapted from the estimation of missing values in gene expression data to overcome this problem and estimate the missing values by using the available measurements from other proteins, such as the *k* nearest neighbor method being applied to Difference Gel Electrophoresis data.[42] Another of the recent efforts was to integrate the GO (Gene Ontology) information into the data imputation; this approach could enhance the imputation even when the missing fraction is large.[85]

In most of the previous comparisons of mRNA and protein abundance,[2,41,48,60,61,89] undetected proteins were simply assigned a value of "zero" and excluded from the correlation analysis. This simplification could seriously affect the interpretation of the relationship between mRNA and protein abundance. For instance, current technologies for proteomic analysis tend to be biased towards detection of relatively abundant proteins. Correlation patterns between mRNA and protein abundances in these highly abundant genes cannot be generally extended to all protein-coding genes in the genome since correlation patterns may be different for lowly abundant genes. Hence, improved methods of coping with missing protein abundance values are necessary for integrative analysis of transcriptomic and proteomic datasets. To address this need, Nie *et al.*[65] proposed a novel statistical model in a recent study, the zero-inflated *Poisson* regression model,[48,56] to deal with missing protein abundance data. The study utilized a set of transcriptomic data that included mRNA abundance information for all 3507 genes in the *D. vulgaris* genome, and a set of semi-quantitative LC-MS/MS proteomic data that included abundance information for 600∼700 proteins.[65,92,93] In developing the Zero-inflated *Poisson* (ZIP) regression model we treated the identification of proteins through peptide hits as rare events and modeled peptide hits as a *Poisson* distribution with the mean (λ). The *Poisson* regression model offers a valid framework while it provides no explanation for the fact that ∼83% genes have zero protein abundance. Nie *et al.*[65] ascribed the high percentage of proteins with zero abundance to technical limitations.[54] Therefore, a zero-inflated *Poisson* regression model[48] was proposed to analyze the data. In this model, we assumed that $100 \times p\%$ $(0 < p < 1)$ of the genes with proteomic abundance level of 0 could be unexpressed genes or expressed genes that were undetected due to the technical limitations. Thus, the proteomic abundance (y) was distributed as a mixture of 0's with probability p and a *Poisson* regression distribution with probability $(1 - p)$. The key improvement in analysis provided by this model is that the undetected proteins are also taken into consideration, thus allowing an estimation of protein expression even when the proteins were experimentally undetectable due to technical limitations. In addition, the predicted values can be used to correct the measured protein abundance level for experimentally detected proteins by considering their mRNA levels. The zero-inflated *Poisson* regression model was constructed independently with three sets of microarray and proteomics data obtained from *D. vulgaris* and verified by demonstrating that the coefficients of variation (CV) of estimated protein abundance values within operons are indeed smaller than CV for random groups of proteins.[92,93] In addition, the predicted protein abundance from this novel model has helped provide new insights into the expression of some genes in *D. vulgaris*. The advantage of applying the ZIP model was also demonstrated by its prediction for the expression of *c*-type cytochromes, which have been reported highly expressed in *D. vulgaris* and are involved in electron transport processes for sulfate reduction.[92] However, protein abundance data for *c*-type cytochromes are not accessed by standard LC-MS/MS approach because they undergo a complex post-translational maturation process involving covalent attachment of heme groups, and this modification can change the charge state in the gas phase and cause atypical fragmentation of peptides, resulting in loss of detection.[7,91] By integrating microarray and proteomic data using the ZIP model, the cytochrome *c*3 encoded by DVU3171 was predicted to be present in all conditions and to have significant expression at the exponential phase in both lactate- and formate-based media, which is consistent with the previous observation that DVU3171 is the primary electron acceptor for periplasmic hydrogen oxidation.[38,65]

3.4 Non-Linearity of Correlations at the Whole Genome Scale

One important conclusion from early integrated transcriptomic and proteomic studies in both prokaryotic and eukaryotic systems is that correlations may be different in different groups of genes and, therefore, that correlations may not follow a uniform pattern at the whole genome scale.[12,63] First of all, different functional groups appear to exhibit different extents of correlation. In a recent transcriptomic and proteomic study with *S. cerevisiae*, *Spearman* rank correlation coefficients were calculated for protein *versus* mRNA abundance in the different protein functional groups.[12] The results showed that, although the differences in mRNA-protein correlation between various functional groups were generally not very significant, several functional groups, such as “metabolism,” “energy” and “protein synthesis,” exhibited stronger correlations between mRNA and protein levels than other groups.[12] This conclusion is strengthened by another integrated transcriptomic and proteomic study of inorganic phosphate-induced pre-osteoblast cells, which showed that, in spite of weak overall correlation between mRNA and protein expression levels, 40 osteoblast relevant proteins (such as periostin, osteoglycin, and TIMP-1) showed a much higher correlation with the corresponding mRNA level.[9] In prokaryotic systems it has also been observed that there was generally no clear correlation between mRNA and protein abundance levels across various functional categories. However, in *D. vulgaris*, a more pronounced correlation was observed for “central intermediary metabolism,” “energy metabolism” and “transport and binding proteins” when *D. vulgaris* was grown with lactate or formate as electron donors.[63]

A second factor to be considered when interpreting correlation data is that, even within the same functional group of genes, the strength of the correlation can vary depending on the time of sampling and on growth conditions. For example, in the previously mentioned integrated study of inorganic phosphate-induced pre-osteoblast cells,[9] the correlation within certain pathway/function groupings was high at multiple time points, indicating that a time-persistent coordination between transcription and translation was maintained during a biological process.[22] In addition, these results also demonstrated that the correlation between mRNA/protein pairs in pre-osteoblast cells was not consistent during the time course following inorganic phosphate induction, with the 21-hour mRNA profile showing the highest correlation to the proteomic data.[22] To determine if this holds true in prokaryotic systems, we compared the correlations between proteomic abundance and mRNA expression levels of *D. vulgaris* between two different conditions via a multiple regression model with first order interaction of the main effects:[46]

$$y_{cj} = \beta_0 + \beta_1 x_{cj} + \beta_2 c + \beta_3 x_{cj} c + \varepsilon_{cj} \tag{8}$$

where y_{cj} and x_j are the transformed and normalized protein abundance and mRNA abundance for the *j*th gene under growth condition *c*. *c* equals 1 or -1, indicating two different conditions. If the interaction term $\beta_3$is significantly different from 0, *i.e.*, there is significant interaction between mRNA level and growth condition, then the correlation between the transcriptomic and proteomics datasets in these two conditions are also significantly different. The results showed that the correlation between mRNA and protein abundances is indeed significantly different between some growth conditions in *D. vulgaris* (data not shown).

Finally, non-uniform correlation between different genes could also be due to variations in gene regulation and post-transcriptional processes, such as translational control or differences between mRNA or protein degradation rates. For example, the comparison between the proteomic expression direction of *Staphylococcus aureus* and the corresponding transcriptomic expression direction according to the functional classes (according to the COG database) shows important similarities for most categories. However, two functional classes (COG O and P, corresponding, respectively, to post-translational modification, protein turnover, chaperones and inorganic ion transport) show opposed quantification trends.[75] Several other studies showed that factors related to translation, such as codon usage, could significantly affect mRNA-protein correlation.[52,58,59] Although it is not the major focus of this review, effects of codon usage on translation and mRNA-protein correlation could be found in several extensive reviews.[5,43,77] Taken together, these results suggest that there exist correlations at both the global and individual gene level. The low correlation at the global level does not necessarily exclude the possibility that some individual mRNA/protein pairs will have high correlation.[12,22] In addition to the overall correlation at a global level, we also need to identify the correlation patterns for those genes that do not follow the general trends. Only by doing this, can we achieve a better understanding of the mRNA/protein relationship in cells.

4. NEW APPLICATIONS OF INTEGRATED ANALYSES

One of the major challenges in integrative analysis of transcriptomic and proteomic datasets is to facilitate extraction of new knowledge that is not accessible through any single-dimensional analysis. In addition to the classic application of transcriptomic and proteomic datasets to identify or cross-validate differentially regulated genes and proteins associated with a given condition or biological state, high-throughput transcriptomic and proteomic data can also be applied to analyze translational status or post-transcriptional expression regulation on a genome scale. One recent study used protein and mRNA abundance, translational status and transcript length data from the yeast *S. cerevisiae* to investigate the genome-wide relationship between transcription, translation, and protein turnover.[12] In this study, variations in correlations were observed 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). The results showed that the protein abundance is explained only slightly better by translational activity than by mRNA abundance, suggesting the importance of other post-transcriptional control mechanisms. In addition, the integrative analysis approach used in that study allowed the identification of functional compartments that are subject to translation on demand (*e.g.* "signal transduction") or that are regulated via protein turnover (such as "cell wall"). Joint transcriptional and post-transcriptional regulation could also be observed for categories, such as ribosomal proteins. In agreement with previous findings,[10,33,89] this study implies a significant primary response to environmental changes at the translational level, an observation that was not apparent using mRNA level analyses exclusively.

Another application of exploring correlations between transcriptomic and proteomic datasets is to gain insights into sequence-dependent features that affect the stage from the gene to protein expression, namely translation. Protein levels in the cell are affected by many factors that operate at the various stages of gene expression, including transcription and translation. However, many of the studies published so far have focused on the transcriptional stage and on features that affect mRNA abundance. This is largely due to the conventionally held opinion that protein levels depend to a large extent on the corresponding mRNA levels. This has led to the conclusion that it is therefore sufficient to study the transcription stage comprehensively and extrapolate to the protein abundance level.[50,53,76,81,84] However, these opinions have recently been brought into question by the poor correlation between transcriptomic and proteomic data, suggesting that protein expression in the cell may be affected by various factors at the translational stage as well. The efficiency of protein biosynthesis and accumulation depends on many factors. First, initial anchoring of ribosomes onto the mRNA depends on complementary binding of the Shine-Dalgarno (SD) sequence $\sim$10 bases upstream of the start codon.[79] Second, non-random use of synonymous codons in the coding region of highly expressed *Escherichia coli* genes indicates that sequences further downstream of the start codon could be of importance for translational efficiency.[21,26,55,80,82] Third, translational efficiency also depends on the availability of various amino acids. Evidence for natural selection of amino acid usage to enhance translational efficiency has been found in the proteomes of *E. coli* and *Bacillus subtilis*.[6] Fourth, translational termination depends upon the attachment of a release factor (RF) in the place of a tRNA in the ribosomal complex.[72] Moreover, studies showed that nucleotide distribution around the stop codons, especially the base following the stop codon, is related to translation termination efficiency.[18,67] Translational initiation has previously been suggested to be the rate-limiting step when compared with the elongation and termination stages of protein biosynthesis.[72,73,87] To test this idea, Nie *et al.*[63–65] used multiple regression analysis of whole-genome mRNA expression and LC-MS/MS proteome abundance data collected from *D. vulgaris* grown in three conditions to gain insights into how the mRNA-protein correlation may be affected by various sequence features related to translation efficiency.[92,93] Surprisingly, this multiple regression analysis suggests that the mRNA-protein correlation is affected primarily by the factors important during the elongation stage, *i.e.*, codon usage and amino acid composition (5.3–15.7% and 5.8–11.9% of the total variation of mRNA-protein correlation, respectively). In contrast, factors related to translation initiation and termination, such as stop codon context and the Shine-Dalgarno sequence, appear to be less important (3.7–5.1% and 1.9–3.8%, respectively).[64] This result is consistent with the conclusion by Lithwick and Marglit,[52] who found that codon bias had the greatest influence on protein expression levels. This idea is further supported by a recent study that estimated that codon bias accelerates translation in *E. coli* by up to 60% in comparison to microbes with very little codon bias.[24] Altogether, sequence features contributed 15.2–26.2% of the total variation of mRNA-protein correlation. This study constitutes an example of how large-scale transcriptomic and proteomic data, along with sequence-level information, can be integrated to gain new insights into regulation of cellular processes.

5. CONCLUSIONS

It is obvious that the ability to conduct transcriptome and proteome correlation analyses would represent an additional and novel means to generate discrete and testable biological hypotheses from large-scale high throughput datasets. For example, a strong correlation between transcriptomic or proteomic data can serve as confirmation for the discovery of an induced response to a treatment, and the lack of a strong correlation can help detect experimental errors or suggest the possibility of a biological uncoupling between the corresponding levels of the respective mRNA and protein species.[54] However, this type of analyses can be misleading if fundamental questions regarding correlation patterns between mRNA expression and protein abundance are not understood well enough and if appropriate statistical methods are not available. The elimination of known procedural issues and/or statistical biases using well developed statistical tools will maximize the chance of finding a correlation when there is one, and/or give more confidence in the conclusion that there really is no a correlation if one has not been found. This review has focused on several statistical challenges in the integrative analysis of high-throughput transcriptomic and proteomic data. Specifically, we have discussed importance of data normalization, the effects of measurement errors, missing values and non-uniform correlation patterns among different groups of genes, and have presented some results from preliminary efforts in developing statistical protocols for studying the correlation between transcriptomic and proteomic datasets. We have also described several applications of integrative analyses of transcriptomic and proteomic data. Finally, we would like to emphasize the complexity of determining a correlation and that most of the preliminary solutions presented here based on our work on *D. vulagris* should be treated as tentative and need to be more rigorously examined and/or challenged.

REFERENCES

[1] Aebersold, R., and Mann, M. 2003. Mass spectrometry-based proteomics. *Nature*. 422: 198–207.

[2] Alter, O., and Golub, G.H. 2004. Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proc. Natl. Acad. Sci. USA*. 101: 16577–16582.

[3] Anderle, M., Roy, S., Lin, H., Becker, C., and Joho, K. 2004. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics*. 20: 3575–3582.

[4] Anderson, L., and Seilhamer, J. 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*. 18: 533–537.

[5] Akashi, H. 1997. Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene*. 205: 269–278.

[6] Akashi, H., and Gojobori, T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci. (USA)*. 99: 3695–3700.

[7] Aubert, C., Leroy, G., Bianco, P., Forest, E., Bruschi, M., and Dolla, A. 1998. Characterization of the cytochromes C from *Desulfovibrio desulfuricans* G201. *Biochem. Biophys. Res. Commun.* 242: 213–218.

[8] Basler, M., Linhartova, I., Halada, P., Novotna, J., Bezouskova, S., Osicka, R., Weiser, J., Vohradsky, J., and Sebo, P. 2006. The iron-regulated transcriptome and proteome of *Neisseria meningitidis* serogroup C. *Proteomics*. 6: 6194–6206.

[9] Beck, G. R., Jr., and Knecht, N. 2003. Osteopontin regulation by inorganic phosphate is ERK1/2-, protein kinase C-, and proteasome-dependent. *J. Biol. Chem.* 278: 41921–41929.

[10] Berg, O. G., and Kurland, C. G. 1997. Growth rate-optimised tRNA abundance and codon usage. *J. Mol. Biol.* 270: 544–550.

[11] Berrar, D. P., Dubitzky, W., and Granzow, M. 2003. Missing value estimation, Kluwer Academic publishers, New York

[12] Beyer, A., Hollunder, J., Nasheuer, H. P., and Wilhelm, T. 2004. Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol. Cell. Proteomics.* 3: 1083–1092.

[13] Bø, T. H., Disvik, D., and Jonassen, I. 2004. LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res*. 32: e34.

[14] Box, G. E. P., and Cox, D. R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26: 211–252.

[15] Breen, E. J., Hopwood, F. G., Williams, K. L., and Wilkins, M. R. 2000. Automatic poisson peak harvesting for high throughput protein identification. *Electrophoresis.* 21: 2243–2251.

[16] Bronstrup, M. 2004. Absolute quantification strategies in proteomics based on mass spectrometry. *Expert Rev. Proteomics*. 1: 503–512.

[17] Brotz-Oesterhelt, H., Bandow, J. E., and Labischinski, H. 2005. Bacterial proteomics and its role in antibacterial drug discovery. *Mass Spectrom. Rev.* 24: 549–565.

[18] Brown, C. M., Stockwell, P. A., Trotman, C. N., and Tate, W. P. 1990. Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. *Nucleic Acids Res.* 18: 6339–6345.

[19] Chen, G., Gharib, T. G., Huang, C. C., Taylor, J. M., Misek, D. E., Kardia, S. L., Giordano, T. J., Iannettoni, M. D., Orringer, M. B., Hanash, S. M., and Beer, D. G. 2002. Discordant protein and mRNA expression in lung adenocarcinomas. *Mol. Cell Proteomics*. 1: 304–313.

[20] Chen, G., Wang, H., Gharib, T. G., Huang, C. C., Thomas, D. G., Shedden, K. A., Kuick, R., Taylor, J. M., Kardia, S. L., Misek, D. E., Giordano, T. J., Iannettoni, M. D., Orringer, M. B., Hanash, S. M., and Beer, D. G. 2003. Overexpression of oncoprotein 18 correlates with poor differentiation in lung adenocarcinomas. *Mol. Cell Proteomics.* 2: 107–116.

[21] Collins, R. F., Roberts, M., and Phoenix, D. A. 1995. Codon bias in *Escherichia coli* may modulate translation initiation. *Biochem. Soc. Trans.* 23: 76

[22] Conrads, K. A., Yi, M., Simpson, K. A., Lucas, D. A. Camalier, C. E., Yu, L. R., Veenstra, T. D., Stephens, R. M., Conrads, T. P., and Beck, G. R. Jr. 2005. A combined proteome and microarray investigation of inorganic phosphate-induced pre-osteoblast cells. *Mol. Cell. Proteomics.* 4: 1284–1296.

[23] Cox, B., Kislinger, T., and Emili, A. 2005. Integrating gene and protein expression data: pattern analysis and profile mining. *Methods.* 35: 303–314.

[24] Dethlefsen, L., and Schmidt, T. M. 2005. Differences in codon bias cannot explain differences in translational power among microbes. *BMC Bioinformatics*. 6: 3.

[25] Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. 2002. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics.* 18: S105–S110.

[26] Faxen, M., Plumbridge, J., Isaksson, L. A. 1991. Codon choice and potential complementarity between mRNA downstream of the initiation codon and bases 1471-1480 in 16S ribosomal RNA affects expression of glnS. *Nucleic Acid Res.* 19: 5247–5251.

[27] 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.

[28] Gao, J., Opiteck, G. J., Friedrichs, M. S., Dongre, A. R., and Hefta, S.A. 2003. Changes in the protein expression of yeast as a function of carbon source. *J. Proteome Res.* 2: 643–649.

[29] 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.

[30] Gowrishankar, J., and Harinarayanan, R. 2004. Why is transcription coupled to translation in bacteria? *Mol. Microbiol.* 54: 598–603.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Freiberg, C., and Brunner, N. A. 2002. Genome-wide mRNA profiling: impact on compound evaluation and target identification in anti-bacterial research. *Targets*. 1: 20–29.

[36] Hack, C. J. 2004. Integrated transcriptome and proteome data: the challenges ahead. *Brief. Funct. Genomic Proteomic.* 3: 212–219.

[37] Hegde, P. S., White, I. R., and Debouck, C. 2003. Interplay of transcriptomics and proteomics. *Curr. Opin. Biotechnol.* 14: 647–651.

[38] Heidelberg, J. F., Seshadri, R., Haveman, S. A., Hemme, C. L., Paulsen, I. T., Kolonay, J. F., Eisen, J. A., Ward, N., Methe, B., Brinkac, L. M., Daugherty, S. C., Deboy, R. T., Dodson, R. J., Durkin, A. S., Madupu, R., Nelson, W. C., Sullivan, S. A., Fouts, D., Haft, D. H., Selengut, J., Peterson, J. D., Davidsen, T. M., Zafar, N., Zhou, L., Radune, D., Dimitrov, G., Hance, M., Tran, K., Khouri, H., Gill, J., Utterback, T. R., Feldblyum, T. V., Wall, J. D., Voordouw, G., and Fraser, C. M. 2004. The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. *Nat. Biotechnol.* 22: 554–559.

[39] Horak, C. E., and Snyder, M. 2002. Global analysis of gene expression in yeast. *Funct. Integr. Genomics*. 2: 171–180.

[40] Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*. 1: 1–9.

[41] Ideker, T., Thorsson, V., Ranish, J.A., Christmas, 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.

[42] Jung, K., Gannoun, A., Sitek, B., Meyer, H.E., Stäuhler, K., and Urfer, W. 2005. Analysis of dynamic protein expression data. *REVSTAT-Statistical J.* 3: 99–111.

[43] Kane, J. F. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr. Opin. Biotechnol.* 6: 494–500.

[44] Khodursky, A. B., and Bernstein, JA. 2003. Life after transcription—revisiting the fate of messenger RNA. *Trends Genet.* 19: 113–115.

[45] Kim, H., Golub, G. H., and Park, H. 2005. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*. 21: 187–198.

[46] Kleinbaum, D. G., Kupper, L. L., Muller, K. E., and Nizam, A. 1998. *Applied Regression Analysis and Other Multivariate Methods*. Duxbury Press, Pacific Grove, CA.

[47] Labbe, A., and Wormald, H. 2005. On normality, ethnicity, and missing values in quantitative trait locus mapping. *BMC Genet.* 6 *Suppl*: 1S52.

[48] Lambert, D. 1992. Zero-inflated *Poisson* regression, with an application to defects in manufacturing. *Technometrics*. 34: 1–14.

[49] Lee, J. H., Lee, D. E., Lee, B. U., and Kim, H. S. 2003. Global analyses of transcriptomes and proteomes of a parent strain and an L-threonine-overproducing mutant strain. *J. Bacteriol.* 185: 5442–5451.

[50] Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T. Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*. 298: 799–804.

[51] Lichtinghagen, R., Musholt, P. B., Lein, M., Romer, A., Rudolph, B., Kristiansen, G., Hauptmann, S., Schnorr, D., Loening, S. A., and Jung, K. 2002. 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.* 42: 398–406.

[52] Lithwick, G., and Margalit, H. 2003. Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res.* 13: 2665–2673.

[53] 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 analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. *Mol. Cell. Proteomics* 3: 478–489.

[54] Maziarz, M., Chung, C., Drucker, D. J., and Emili, A. 2005. Integrating global proteomic and genomic expression profiles generated from islet alpha cells: opportunities and challenges to deriving reliable biological inferences. *Mol. Cell Proteomics.* 4: 458–474.

[55] McCarthy, J. E. G., and Brimacombe, R. 1994. Prokaryotic translation: the interactive pathway leading to initiation. *Trends Genet.* 10: 402–407.

[56] McCullagh, P., and Nelder, J. A. 1989. *Generalized Linear Models*, Chapman and Hall, London.

[57] McLachlan, G. J., Do, K. A., and Ambroise, C. 2004. Analyzing microarray gene expression data. John Wiley & Sons, Inc., Hoboken, NJ.

[58] Mehra, A., Lee, K. H., and Hatzimanikatis, V. 2003. Insights into the relation between mRNA and protein expression patterns: I. Theoretical considerations. *Biotechnol. Bioeng.* 84: 822–833.

[59] Mehra, A., and Hatzimanikatis, V. 2006. An algorithmic framework for genome-wide modeling and analysis of translation networks. *Biophys. J.* 90: 1136–1146.

[60] Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M., Wisniewski, J. R., Stahl, E., Bolouri, M. S., Ray, H. N., Sihag, S., Kamal, M., Patterson, N., Lander, E. S., and Mann, M. 2003. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* 115: 629–640.

[61] Mootha, V. K., Lepage, P., Miller, K., Bunkenborg, J., Reich, M., Hjerrild, M., Delmonte, T., Villeneuve, A., Sladek, R., Xu, F., Mitchell, G. A., Morin, C., Mann, M., Hudson, T. J., Robinson, B., Rioux, J. D., and Lander, E. S. 2003. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc. Natl. Acad. Sci. (USA)*. 100: 605–610.

[62] Munoz, E. T., Bogarad, L. D., and Deem, M. W. 2004. Microarray and EST database estimates of mRNA expression levels differ: the protein length versus expression curve for *C. elegans*. *BMC Genomics*. 5: 30.

[63] Nie, L. Wu, G., Zhang, W. 2006. Correlation between mRNA and protein abundance in *Desulfovibrio vulgaris*: a multiple regression to identify sources of variations. *Biochem. Biophys. Res. Commun.* 339: 603–610.

[64] Nie, L., Wu, G., and Zhang, W. 2006. Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: a quantitative analysis. *Genetics*. 174: 2229–2243.

[65] Nie, L. Wu, G., Brockman, F. J., and Zhang, W. 2006. Integrated analysis of transcriptomic and proteomic data of *Desulfovibrio vulgaris*: zero-inflated *Poisson* regression models to predict abundance of undetected proteins. *Bioinformatics*. 22: 1641–1647.

[66] Orntoft, T. F., Thykjaer, T., Waldman, F. M., Wolf, H., and Celis, J. E. 2002. 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.* 1: 37–45.

[67] Poole, E. S., Brown, C. M., and Tate, W. P. 1995. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in *Escherichia coli*. *EMBO J.* 14: 151–158.

[68] Purohit, P. V., Rocke, D. M., Viant, M. R., and Woodruff, D. L. 2004. Discrimination models using variance-stabilizing transformation of metabolomic NMR data. *OMICS*, 8: 118–130.

[69] Resch, A., Leicht, S., Saric, M., Pasztor, L., Jakob, A., Gotz, F., and Nordheim, A. 2006. Comparative proteome analysis of *Staphylococcus aureus* biofilm and planktonic cells and correlation with transcriptome profiling. *Proteomics*. 6: 1867–1877.

[70] Resing, K. A., and Ahn, N. G. 2005. Proteomics strategies for protein identification. *FEBS Lett.* 579: 885–889.

[71] Rhodius, V. A., and LaRossa, R. A. 2003. Uses and pitfalls of microarrays for studying transcriptional regulation. *Curr. Opin. Microbiol.* 6: 114–119.

[72] Rocha, E. P., Danchin, A., and Viari, A. 1999. Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis. *Nucleic Acids Res.* 27: 3567–3576.

[73] Romby, P., and Springer, M. 2003. Bacterial translational control at atomic resolution. *Trends Genet.* 19: 155–161.

[74] Scherl, A., Francois, P., Bento, M., Deshusses, J. M., Charbonnier, Y., Converset, V., Huyghe, A., Walter, N., Hoogland, C., Appel, R. D., Sanchez, J. C., Zimmermann-Ivol, C.G., Corthals, G.L., Hochstrasser, D.F., and Schrenzel, J. 2005. Correlation of proteomic and transcriptomic profiles of *Staphylococcus aureus* during the post-exponential phase of growth. *J. Microbiol. Methods*. 60: 247–257.

[75] Scherl, A., Francois, P., Charbonnier, Y., Deshusses, J. M., Koessler, T., Huyghe, A., Bento, M., Stahl-Zeng, J., Fischer, A., Masselot, A., Vaezzadeh, A., Galle, F., Renzoni, A., Vaudaux, P., Lew, D., Zimmermann-Ivol, C. G., Binz, P. A., Sanchez, J. C., Hochstrasser, D. F., and Schrenzel, J. 2006. Exploring glycopeptide-resistance in *Staphylococcus aureus*: a combined proteomics and transcriptomics approach for the identification of resistance-related markers. *BMC Genomics*. 7: 296.

[76] Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M., Richmond, C. S., Blattner, F. R., Lockhart, D. J., and Church, G. M. 2000. RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.* 18: 1262–1268.

[77] Sharp, P. M., and Matassi, G. 1994. Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* 4: 851–860.

[78] Shimizu, T., Shima, K., Yoshino, K., Yonezawa, K., Shimizu, T., and Hayashi, H. 2002. Proteome and transcriptome analysis of the virulence genes regulated by the VirR/VirS system in *Clostridium perfringens.J. Bacteriol.* 184: 2587–2594.

[79] Shine, J., and Dalgarno, L. 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. (USA)*. 71: 1342–1346.

[80] Sorensen, M. A., Kurland, C. G., Pedersen, S. 1989. Codon usage determines translation rate in*Escherichia coli*. *J. Mol. Biol.* 207: 365–377.

[81] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*. 9: 3273–3297.

[82] Stenstrom, C. M., Jin, H., Major, L. L., Tate, W. P., and Isaksson L. A. 2001. Codon bias at the 3′-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene*, 263: 273–284.

[83] Tian, Q., Stepaniants, S. B., Mao, M., Weng, L., Feetham, M. C., Doyle, M. J., Yi, E. C., Dai, H., Thorsson, V., Eng, J., Goodlett, D., Berger, J. P., Gunter, B., Linseley, P. S., Stoughton, R. B., Aebersold, R., Collins, S. J., Hanlon, W. A., and Hood, L. E. 2004. Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Mol. Cell Proteomics.* 3: 960–969.

[84] Tjaden, B., Saxena, R. M., Stolyar, S., Haynor, D. R., Kolker, E., and Rosenow, C. 2002. Transcriptome analysis of *Escherichia coli*

using high-density oligonucleotide probe arrays.*Nucleic Acids Res.* 30: 3732–3738.

[85] Tuikkala, J., Elo, L., Nevalainen, O. S., and Aittokallio, T. 2006. Improving missing value estimation in microarray data with gene ontology. *Bioinformatics*. 22: 566–572.

[86] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 17: 520–525.

[87] Vellanoweth, R. L., and Rabinowitz, J. C. 1992. The influence of ribosome-binding-site elements on translational efficiency in *Bacillus subtilis* and *Escherichia coli in vivo*. *Mol. Microbiol.* 6: 1105–1114.

[88] Wang, D., Jensen, R., Gendeh, G., Williams, K., and Pallavicini, M. G. 2004. Proteome and transcriptome analysis of retinoic acid-induced differentiation of human acute promyelocytic leukemia cells, NB4. *J. Proteome Res.* 3: 627–635.

[89] Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, G., Plouffe, D., Deciu, C., Winzeler, E., Yates, J. R. 3rd. 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.

[90] Wilkins, M. R., Appel, R. D., Van Eyk, J. E., Chung, M. C., Gorg, A., Hecker, M., Huber, L. A., Langen, H., Link, A. J., Paik, Y. K., Patterson, S. D., Pennington, S. R., Rabilloud, T., Simpson, R. J., Weiss, W., and Dunn, M. J. 2006. Guidelines for the next 10 years of proteomics. *Proteomics*. 6: 4–8.

[91] Yu, X.L., Wojciechowski, M., and Fenselau, C. 1993. Assessment of metals in reconstituted metallothioneins by electrospray mass-spectrometry. *Anal. Chem.* 65: 1355–1359.

[92] Zhang, W., Culley, D. E., Scholten, J. C., Hogan, M. Vitiritti, and L., Brockman, F. J. 2006. Global transcriptomic analysis of *Desulfovibrio vulgaris* on different electron donors. *Antonie Van Leeuwenhoek*. 89: 221–237.

[93] Zhang, W., Gritsenko, M. A., Moore, R. J., Culley, D. E., Nie, L., Petritis, K., Strittmatter, E. F., Camp, D. G. II, Smith, R. D., and Brockman, F. J. 2006. A proteomic view of *Desulfovibrio vulgaris* metabolism as determined by liquid chromatography coupled with tandem mass spectrometry. *Proteomics*. 6: 4286–4299.