



Review

Correlation of mRNA and protein in complex biological samples

Tobias Maier*, Marc Güell, Luis Serrano

Center for Genomic Regulation, Systems Biology Unit, Dr. Aiguader 88, 08003 Barcelona, Spain

ARTICLE INFO

Article history:

Received 31 August 2009

Revised 9 October 2009

Accepted 14 October 2009

Available online 20 October 2009

Edited by Stefan Hohmann

Keywords:

mRNA

Protein

Quantitative

Systems biology

Genomics

Proteomics

ABSTRACT

The correlation between mRNA and protein abundances in the cell has been reported to be notoriously poor. Recent technological advances in the quantitative analysis of mRNA and protein species in complex samples allow the detailed analysis of this pathway at the center of biological systems. We give an overview of available methods for the identification and quantification of free and ribosome-bound mRNA, protein abundances and individual protein turnover rates. We review available literature on the correlation of mRNA and protein abundances and discuss biological and technical parameters influencing the correlation of these central biological molecules.

© 2009 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

1. Introduction

The acquisition and interpretation of large quantitative data-sets across multiple samples is a major challenge for systems biology. Systems biology aims for the discovery and understanding of newly emerging properties arising from a global, systemic angle of analysis. Ultimately, systems biology intends to reproduce biological systems in terms of mathematical models and simulations, deriving biologically meaningful conclusions there from. Systems biology crucially depends on the accurate quantitative description of complex biological samples. It makes use of large-scale data-sets often derived from genomics and proteomics analyses and requires determination of concentrations, binding constants and interactions. For some time many groups have used data regarding changes in gene expression as evidence of consequent changes in protein expression. However, this is not always necessarily the case.

The central dogma of molecular biology deals with the transfer of information from DNA via mRNA to proteins. It is mechanistically very well understood how genes get transcribed, mRNA gets processed and sequentially translated into amino acid chains at the ribosome and subsequently fold into functional proteins. However, most reports on mRNA and protein abundances find only a weak correlation between the respective abundances of these two classes of biological molecules. Several biological factors were identified which influence this correlation, but also methodological constraints play a role when comparing mRNA to protein levels.

The increasing amount of available large-scale high-quality quantitative data-sets for both mRNA and proteins resulting from genomics and proteomics experiments should allow the accurate analysis of this physiological phenomenon observed for years but not readily accessible experimentally until now. Methods for the global identification, quantification and analysis of the transcriptome and the translome in biological samples have rapidly improved over the last years, but still demand a high level of expertise both experimentally and computationally.

In course of this review, we will present an overview over available techniques to quantify mRNA and proteins in complex biological samples. We review available mRNA–protein correlation studies and comment in detail on technical and biological factors influencing the quantitative relation between mRNA and protein.

2. Methodology for the quantitative analysis of mRNA and proteins in complex samples

2.1. mRNA

Transcript and protein tracking systems have offered an unprecedented view of inner workings of cells. Mechanisms that regulate the synthesis, processing, transport, translation and stability of mRNAs are critical points in cell function. The ability to measure RNAs is providing a more accurate view of gene expression. Classical methods such as Northern blotting and RT-PCR allowed the steady state measurement of selected transcripts. Recent methods provide a wider scope of possibilities. Next-generation sequencing methods provide with the mapping and quantification of several

* Corresponding author.

E-mail address: tobias.maier@crg.es (T. Maier).

thousands of transcripts in single experiments. Although microarrays and Affymetrix chips are still the most used tools to measure mRNA levels [1,2], they are progressively being replaced by deep sequencing technologies. On the other hand, accurate description of concrete pathways can be achieved in vivo at single cell resolution.

2.1.1. High-throughput measurements

The high throughput of next-generation sequencing technology, rapidly producing huge numbers of short sequencing reads, made possible the analysis of a complex sample containing a mixture of different transcripts. However, expression microarrays are still probably the most widely used methodology for transcriptome analysis. They consist of a large number of clusters of molecules of DNA called probes distributed in rows on a surface. One probe or subgroups of them are designed to hybridize with a concrete transcript. Microarrays permit to monitor the expression of transcripts with complementarity to the probes. They can be used to measure absolute transcript concentrations whenever the probe binds to its target specifically [3]. For estimating transcript concentrations a proper calibration is required. Although having been extensively used they rely in sequence-specific probe hybridization, suffer from background and cross-hybridization problems [4,5], dye-based detection issues and design constraints that seriously limit the detection of RNA splice patterns or unknown genes. More recently, whole-genome tiling arrays have been also used for measuring gene expression or new gene discovery. These arrays differ from microarrays in the nature of the design. Probes are targeted to cover the entire genome and not only the annotated transcripts. These specific arrays can discover new exons but still conserve the rest of microarrays limitations. In contrast, tag-based methods like SAGE (Serial Analysis of Gene Expression) measure absolute abundances and are not limited to array content. They can provide precise digital gene expression levels but only a portion of the transcript is analyzed making splice isoforms not distinguishable.

Recently, the possibility to use next-generation sequencing methods for mapping and quantifying transcriptomes has been reported [6]. Novel sequencing techniques provide high speed and throughput. RNA is transformed to cDNA and amplified. Determination of sequence data can be achieved with a massive sequence parallelization using different techniques [7]. Sequences produced using RNA-seq [8] are aligned to the genome and the different transcripts are quantified in reads per kilobase of exon model per million mapped reads (RPKM) [6]. The RPKM measure is related to the molar concentration of the transcript in the starting sample which may be used for comparison of transcript levels within and between samples. With the use of spike in RNA, absolute transcript levels can be calculated. A 40-million-read transcriptome data-set provides reliable measurement of a single transcript per cell in human cell lines. Owing to the massive number of sequences, low-abundance RNAs can be detected. With this technique, transcripts are characterized through their sequence avoiding the limitation to known transcript products. One of the most attractive advantages of deep sequencing methods is the ever decreasing cost. The different advances in the field have led to an exponential reduction in cost per base. An inconvenience is the cost of acquiring the necessary equipment. However, when considering the price per sequenced base, it is already in most of the cases, much less expensive than microarray-based methods.

2.1.2. Single cell in vivo measurements

Individual native RNA molecules can be visualized in vivo. By labeling RNAs with stem loop repeats from the MS2 bacteriophage, which will bind to a GFP-tagged MS2-coat-protein product construct, one can examine the dynamics of specific RNAs in single

cells [9]. This ability to measure has provided important details of how transcription is carried out in pulses [10–12]. The MS2-based technology can be embedded in a system that allow all the components involved in gene expression (DNA, RNA, protein) to be visualized in real time [13]. This system combines lac-repressor-operator and MS2-coat-protein-translational operator interaction units and tetracycline response elements. It allows an in vivo direct tracking of gene expression but it is limited to a subset of transcripts since it needs specific labeling for each of them.

2.2. Protein

2.2.1. 2D SDS-PAGE coupled to mass spectrometry (MS)

Traditionally, individual protein expression levels in complex samples are analyzed by two-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis (2D SDS-PAGE). The methodology involves an isoelectric focusing step along an immobilized pH gradient and subsequent separation of the proteins by SDS-PAGE. Obtaining reproducible gels by two-dimensional gel electrophoresis is technically demanding and often complicates the comparison of gels from different samples prepared under similar conditions. Very basic and acidic proteins, very large and very small proteins, as well as low abundant species are often hard if not impossible to analyze by 2D SDS-PAGE [14,15].

Different visualization techniques exist for SDS-PAGE separated proteins. Classical protein staining methods are either not sensitive enough (Coomassie blue) or not quantitative over a wide range of spot intensities (silver staining) to directly infer protein abundance levels from the protein spots [16,17]. Alternative staining methods with fluorescent dyes overcome several problems. Difference Gel Electrophoresis (DIGE) is more sensitive and allows the quantitative analysis of protein spots on 2D gels over a wide dynamic range [18,19]. Because several samples can be separated on the same gel, the direct comparison between different samples is greatly facilitated.

Separation of proteins by 2D SDS-PAGE and quantification of the spots using suitable dyes merely requires their subsequent identification by mass spectrometry to assign quantitative information to individual proteins. Small gel pieces corresponding to individual gel spots are excised with suitable tools and the separated proteins are digested with trypsin and identified either by peptide mass fingerprinting or tandem MS.

2.2.2. Quantitative MS technologies

In recent years, mass spectrometry techniques for the identification and quantification of complex biological samples have advanced tremendously [20,21]. Sample complexity is reduced by separating digested protein mixtures by liquid chromatography prior to on-line electrospray ionization and subsequent mass spectrometric analysis of the isolated peptides. Several mass spectrometer designs exist, each with specific advantages for mass accuracy, dynamic range, resolving power and sensitivity [22]. Hybrid instruments, such as the LTQ-Orbitrap machines combine high mass accuracy, high resolving power and a high dynamic range [23]. Selected reaction monitoring (SRM) on a hybrid triple quadrupole/ion trap mass spectrometer recently allowed the identification and quantification of low-abundance proteins in a whole cell lysate of *Saccharomyces cerevisiae* [24].

Mass spectrometry on protein samples is intrinsically not quantitative. However, several methods have been developed to overcome this problem. They basically fall into two groups: quantification using stable isotope labeling and label free quantification. Stable isotope labeling of peptides can be achieved either by metabolic labeling of intact proteins during cell culture (e.g. SILAC [25]) or by chemical labeling of peptides after tryptic digestion (e.g. ICAT, iTRAQ [26,27]). The SILAC method has been used to quantify

very complex samples, such as the yeast proteome and mouse embryonic stem cells [28,29].

Label-free quantification of proteins relies on information obtained directly from mass spectrometer read outs. A simple method involves counting and comparing the number of fragment ion spectra of a given peptide. This method is based on the finding that the MS/MS sampling rates for particular peptides are directly related to the abundance of a peptide represented by its precursor ion in the sample. Typically the 1–5 most intense precursor ions per protein are used for quantification [30,31]. A slightly different approach is the protein abundance index (PAI) [32]. Here the number of identified peptides in an MS/MS run are set in relation to the number of theoretically observable peptides for each protein. Proteins can also be quantified by measuring and comparing peptide precursor ion signal intensities integrated over their full retention time [33]. This latter method relies on the alignment of chromatograms and often requires very specific software. In contrast to the stable isotope, label-free quantification relies more heavily on technical replicates.

While the aforementioned methods give mostly relative quantitative information between two, or several, protein samples, recently approaches were developed allowing for the absolute quantification of entire proteomes. The absolute quantification (AQUA) of proteins by mass spectrometry involves spiking known amounts of synthesized, stable isotope labeled peptides mimicking expected tryptic peptides into the digested protein sample [34]. The labeled peptides serve as an internal standard. They behave identical as their native counter parts in liquid chromatography prior to MS, since they have the same chemical properties. However, the heavy isotope label, leads to a mass shift which can be detected by mass spectrometry. Integration of respective peak areas and comparison between labeled and unlabeled isoforms allows the absolute quantification of the corresponding protein. Combining the AQUA technique with label free quantification approaches allows the absolute quantification of proteins in complex mixtures [35]. A recent article reports the first example of the absolute quantification of an entire proteome in the model organism *Leptospira interrogans*, making use of this combinatorial approach [36].

2.2.3. Alternatives for measuring protein abundance

Alternative, non MS-based methods exist to for the quantification of individual proteins in complex samples. They involve in vivo tagging of proteins with detection markers. Namely the TAP-tag (tandem affinity purification) and fusion of cellular proteins with the green fluorescent protein (GFP) lead to global measurements of protein abundance in cells [37,38]. TAP-tagged proteins are analyzed by Western blotting and GFP fusion proteins can be quantified by flow cytometry, even on the single cell level. Both approaches correlate well ($r^2 = 0.80$) [38]. Again, relating the relative or abstract quantities derived from these techniques to suitable standards with known concentrations allows the absolute quantification of all identified species in a given sample. For example quantitative Western blotting with purified proteins as standards can be used [38].

However, tag-based quantification of proteomes is very work-intensive and interference of the tag with protein expression, stability and function cannot be excluded. Also, perhaps more importantly, genome-wide tagging approaches are only applicable to organisms with established protocols for genetic manipulation in a high-throughput format.

3. Available data on mRNA protein correlation

Published studies on the correlation of mRNA and corresponding protein levels in complex samples are not very abundant. They

focus mostly on yeast species, involving sub-sets of the genome and proteome accessible by the respective experimental techniques. Also few studies on mRNA–protein correlations in bacteria and mammalian cells have been published. It has been shown that protein and mRNA abundances are not following a normal distribution [38,39]. Therefore the Spearman rank coefficient (r_s) is more suitable than the Pearson correlation coefficient (r_p) to describe the correlation between mRNA and protein levels. However, both correlation methods give similar results and they are used equivalently in available publications on mRNA–protein correlations. A summary of available publications on the correlation of mRNA and protein abundances in complex samples is given in Table 1.

Gygi et al. [40] compared mRNA and protein levels for 106 yeast proteins. The authors relied on serial analysis of gene expression (SAGE) tables for mRNA levels. Protein spots were excised from 2D gels and quantified metabolically labeling the cells with ^{35}S -methionine and scintillation counting. They found a 20–30-fold difference in expression levels and overall weak correlation. High expression level was indicative of better correlation values. Their analysis contained a systematic experimental error for the protein identification and quantification. Only proteins with high codon bias and long predicted half-life were present in the analyzed data-set.

Futcher et al. [39] followed a similar methodology, albeit finding a better correlation between mRNA and protein abundance ($r_p = 0.76$, $r_s = 0.74$). Their identified proteins also had a high codon bias, when compared with the theoretical distribution of the yeast genome. Their results indicated a good correlation between the Codon Adaptation Index (CAI) and protein abundance. Protein turnover data showed that, similar to [40] only proteins with long half-lives were accessible by the employed methodology involving 2D SDS–PAGE.

Greenbaum et al. [41] integrated these and other studies on protein abundance in yeast. Their reference data-set containing 2044 proteins showed a good correlation between mRNA and protein levels ($r = 0.66$). The authors looked at the correlation of mRNA and protein abundance for genes with either variable or steady levels of mRNA along the yeast cell cycle. They found that varying mRNA levels correlated well with corresponding protein levels ($r_p = 0.89$), whereas genes expressed at steady levels had fluctuating protein levels along the cell cycle ($r_p = 0.2$). The authors also showed that high ribosomal occupancy (that is the fraction of mRNA bound to ribosomes at any given time) for mRNA resulted in a better correlation with protein levels, as compared to uncorrelated mRNA and protein levels when respective mRNA species showed low ribosomal occupancy.

A study using ICAT reagents for relative quantification of proteins monitoring fold changes revealed varying degrees of correlation between mRNA and protein levels in yeast in response to

Table 1

Overview of mRNA–protein correlation studies in different organisms. r_p : Pearson correlation coefficient and r_s : Spearman rank coefficient. Data-set size refers to the number of mRNA–protein pairs used for the determination of respective correlation coefficients.

Organism	r_p	r_s	Data-set size	Reference
<i>Saccharomyces cerevisiae</i>	0.36	n.d.	73	[40]
<i>Saccharomyces cerevisiae</i>	0.76	0.74	148	[39]
<i>Mus musculus</i>	0.59	n.d.	425	[46]
<i>Saccharomyces cerevisiae</i>	n.d.	0.45	678	[43]
<i>Desulfovibrio vulgaris</i>	0.50	n.d.	703	[45]
<i>Escherichia coli</i>	0.57	0.50	1103	[32]
<i>Schizosaccharomyces pombe</i>	0.58	0.61	1367	[44]
<i>Saccharomyces cerevisiae</i>	0.66	n.d.	2044	[41]
<i>Saccharomyces cerevisiae</i>	n.d.	0.57	4251	[38]

n.d.: not determined.

carbon source perturbations [42]. A relative quantification approach with metabolically labeled yeast samples lead to the identification and quantification of 678 yeast proteins. The correlation with microarray data from identical samples was reported as weakly positive ($r_s = 0.45$) [43]. Label-free quantification of the *Schizosaccharomyces pombe* proteome by spectral counting allowed the comparison of 1367 mRNA–protein pairs. Determined correlation coefficients were $r_s = 0.61$ and $r_p = 0.58$, respectively [44].

Correlation of protein abundances determined by immunostaining for 4251 individually TAP-tagged yeast strains with published microarray data was similarly shown to be weakly positive ($r_s = 0.57$) [38]. In a correlation study between mRNA and protein abundance in the bacterium *Desulfovibrio vulgaris* [45] similar correlation coefficients were observed, indicating that the reported weakly positive correlations are not limited to yeast cells but also extend to prokaryotic organisms. Ishihama et al. used the Codon Adaptation Index (CAI, see below) as a measure for gene expression and found correlation coefficients of $r_s = 0.50$ and $r_p = 0.57$, respectively, in *Escherichia coli* [32]. A study in two hematopoietic mouse cell lines revealed similar correlation coefficients for mammalian cells ($r_p = 0.59$, 425 mRNA–protein pairs) [46].

A software package called PARE (Protein Abundance and mRNA Expression) has been published which allows the rapid assessment of mRNA–protein correlation for complex samples. Data-sets for yeast, rat, mouse and *Halobacterium* are provided, but own quantitative results can be uploaded for analysis [47].

The available body of literature shows that the correlation of mRNA and protein levels in complex samples is far from perfect. This finding points towards complex and diverse regulatory mechanisms responsible for the observed differences in the quantitative relation between transcriptome and proteome. Several reasons can be causative for the apparent poor correlation: (1) post-transcriptional parameters; (2) post-translational parameters; and (3) noise and experimental error. Specific mechanisms from these topics are discussed in the following paragraphs.

4. Parameters influencing mRNA–protein correlation

Transcription and translation are far from having a linear and simple relationship. Different mechanisms involving *cis*-acting and *trans*-acting mechanisms generate a big repertoire of systems that enhance or repress the synthesis of proteins from a certain copy number of mRNA molecules. Different events may uncouple transcription and translation continuously or under certain conditions.

4.1. RNA secondary structure and Shine Dalgarno sequence differences

The physical transcript properties modify translation efficiency at different levels. Prokaryotic protein coding transcripts may have a specific ribosomal binding site upstream of the start codon. It is complementary to 3' end of the 16S rRNA and it is called Shine Dalgarno (SD) sequence. Rate of translation is dependent on the SD sequence. Transcripts with a weak SD (not perfect complementarity to rRNA) sequence are translated with lower efficiency. This effect already adds some complexity, since not all RNAs are translated into proteins equivalently.

In addition, other condition dependent features of the mRNA such as RNA structure may change under certain conditions translation efficiency. Base-paired structures in the mRNA can selectively sequester and expose the ribosome binding site. For instance, secondary and tertiary interactions involved in mRNA folding are sensitive to temperature. Temperature dependent structural changes may distinguish translationally active and inactive mRNA structures. A well studied example is *rpoH* mRNA in

E. coli. It acts as a thermostat, at low temperatures *rpoH* mRNA is stably folded in a conformation that prevents 30S subunit association but upon heat shock induction it unfolds enough to be translated [48]. Conformational changes in the mRNA can be also induced by small metabolites. For example, translation of the mRNA encoding the cobalamin-transport protein (*btuB*) in *E. coli* is repressed at high coenzyme B12 concentration [49].

4.2. Regulatory proteins

Regulatory proteins and sRNAs can act as translational modulators. For instance, in *E. coli*, genes encoding R-proteins (*rpl*, *rps*) are spread in several operons. A regulatory R-protein represses translation of some cistrons by binding to its own mRNA at a region contiguous with the SD sequence [10]. The target RNA sites are similar to their corresponding binding sites in the rRNA but with a lower affinity. Only when all rRNA is assembled into ribosomes, R-proteins bind to their mRNA so as to stop translation.

4.3. Regulatory sRNAs

It is clear now that sRNAs have a key regulatory role in regulating gene expression in prokaryotes. They are known to influence the evolution and stability of mRNAs but they can also affect translation efficiency. Different mechanisms have been described. Some promote ribosome binding [50] to the target mRNA whereas others block translation [50]. Quantitative analysis of the mechanisms revealed sRNA higher capacity to filter noise from input signals and the ability to respond fast to big input signals to modify specific protein concentrations [51]. In many cases, sRNAs introduce complexity in the mRNA protein relationship. In *E. coli*, translation inhibition of galactokinase (third cistron of the *lac* operon) is carried out by a sRNA called spot 42. It binds to the SD sequence in *galk* RNA inhibiting translation [52]. Also examples of translation activation have been reported. For example, in *Staphylococcus aureus* a sRNA called RNAIII controls the expression of toxin genes (*hla*). The 5' UTR of *hla* mRNA containing the SD sequence forms a secondary structure by base-pairing with an upstream *cis*-acting sequence and therefore blocking translation. RNA III binds with such a *cis*-acting element avoiding translation inhibition [53].

4.4. Codon bias and codon adaptation index

In many organisms, synonymous codons (i.e. coding for the same amino acid) are used with different frequencies. This phenomenon is called codon bias. A sophisticated measure for the codon bias is the codon adaptation index (CAI) [54]. It is based on a test set of highly expressed genes and considers the relative adaptiveness of each gene. The CAI for an individual gene *g* is defined as:

$$CAI_g = \prod_{i=1}^N w_i^{1/N},$$

where *N* is the number of codons in gene *g* and w_i is the relative adaptiveness of codon *i* [55].

It has been shown that a large codon bias correlates with highly expressed genes [56,57] and proteins [55,58]. The codon bias is believed to be mainly a mechanism by which the cell can maximize translation efficiency. Determining the codon bias distribution for identified proteins from organism-wide proteomics studies and comparing it to the calculated bias distribution of the respective genome therefore gives information on the experimental coverage of low abundant proteins [40]. A comparative study showed that the codon bias has a higher influence on mRNA–protein correlation than sequences upstream of the translation initiation site (Shine

Dalgarno or Kozak sequence) [59]. Multi regression analysis showed that the specific codon usage contributes to 8.9% of the total variation in mRNA–protein correlation [60].

4.5. Ribosomal density and ribosome occupancy

The translational efficiency is the number of completed protein molecules produced per mRNA and time [61]. One measure for translational efficiency is the ribosomal density (i.e. the number of ribosomes per transcriptional unit) or the ribosome occupancy, which denotes the specific enrichment of individual mRNA species on ribosomes. Different translation efficiencies for mRNA molecules directly influence the mRNA–protein correlation. Ribosome occupancy can be determined by obtaining polysome profiles by sucrose gradient centrifugation [61], affinity tag purification [62] and subsequent microarray analysis and/or Northern blotting [1] as well as by deep sequencing of ribosome protected mRNA fragments [63] (Fig. 1).

Translation efficiency is highly diverse both quantitatively and qualitatively and no linear relationship between mRNA abundance and individual protein synthesis rate can be assumed [61]. Results on exponentially growing yeast cells showed that mRNA species

can have different numbers of ribosomes attached, indicating higher translation rates with more ribosomes attached [1]. It is estimated that roughly one third of expressed genes are translationally regulated [63]. A quantitative comparison of ribosome attached mRNA to corresponding free mRNA fragments revealed a roughly 100-fold range of translation efficiency, additionally to a translationally inactive fraction of identified transcripts [63]. The fraction of ribosome-associated mRNA molecules, indicating actively translated species is thought to be a better predictor of protein abundance than mRNA abundances alone. Indeed, Ingolia et al. showed that numbers for ribosome-associated mRNA correlated significantly better than mRNA abundance alone with absolute protein levels [63].

In yeast, transcripts with high occupancy correlate better with their corresponding protein abundance [41] and very long transcripts showed on average a lower ribosomal density than shorter mRNAs [1,61]. Ribosomes are approximately three fold enriched on the first 30–40 nucleotides of protein coding sequence [63]. This finding can be attributed to either a slow translational start or to premature translation termination. A recent study in yeast showed a highly correlated response to cellular stress between actual transcripts and ribosome-associated mRNA [64].

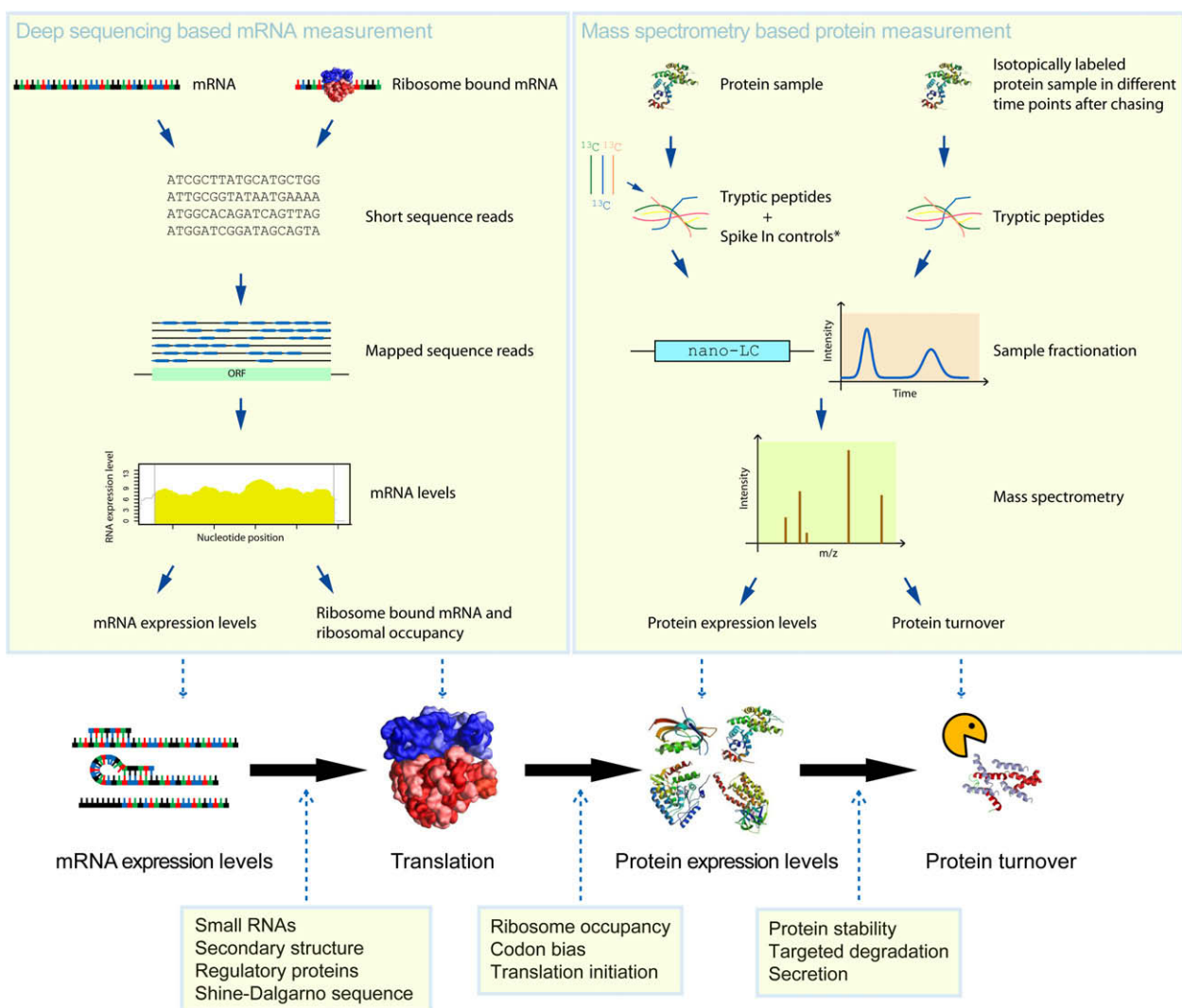


Fig. 1. Central pathway of molecular biology, experimental techniques to measure mRNA and protein at various stages and parameters influencing the mRNA–protein correlation. Sequencing approaches are used to measure free and ribosome-bound mRNA. Mass spectrometry based methods allow the identification and quantification of cellular proteins and their individual turnover rates.

Translation efficiency significantly alters the mRNA–protein correlation at least for a subset of genes [1,63]. A multiple regression analysis of different biological factors influencing the mRNA–protein correlation in yeast estimated the contribution of ribosome occupancy and ribosomal density to be around 5% of the total variation [60].

4.6. Protein half-lives

The major post-translational factor influencing mRNA–protein correlation is the individual half-life of proteins. The cellular lifetime of a protein depends on several factors: intrinsic protein stability, the first amino-terminal amino acid (N-end rule), post-translational processing, such as phosphorylation, and ubiquitination, and on the localization of the respective protein. Protein half-lives are highly variable, ranging from a few seconds to several days. A prediction algorithm based on protein and mRNA abundance, ribosome density and occupancy and transcript length provided theoretical half-lives ranging over five orders of magnitude [65]. Correlation analysis of various parameters showed that protein turnover, estimated by this protein half-life descriptor is the most important biological parameter influencing mRNA–protein correlation [60]. A somewhat simpler and probably less accurate method to estimate individual protein half-lives can be applied by following the N-end rule [66], based on the finding that the dominant feature governing the turnover of a specific protein is a destabilizing N-terminal residue. The N-end rule and associated mechanisms have been shown to be active in *E. coli*, *S. cerevisiae* and mammalian cells. Although prokaryotes and eukaryotes use different strategies for degradation of N-end rule substrates, recent findings indicated that they share common principles of substrate recognition [67].

Protein half-lives can be determined by different labeling strategies. Pulse-chase experiments involving metabolic labeling of cells with radioactive methionine have traditionally been used to measure global protein turnover. Identification of labeled proteins from 2D gels allows for the determination of individual protein half-lives [68]. However, application of the N-end rule to protein samples separated by 2D gels and identified by mass spectrometry revealed a bias for detected proteins to have high turnover times [40]. Concurrent with the recent advances in mass spectrometry, stable isotope labeling approaches have been developed to determine individual protein turnover rates [68,69] (Fig. 1). A corresponding approach has recently been used to determine the turnover rates of nearly 600 proteins from a human cell line [70]. Further advances in these mass spectrometry based approaches should yield proteome-wide data on individual protein turnover rates. Benchmarking and integration of this type of data with current prediction algorithms is necessary to obtain a more accurate picture of the biological factors influencing mRNA–protein correlation.

4.7. Other biological factors influencing mRNA–protein correlation

Several other factors have regulatory function for gene expression and protein synthesis. Absolute ribosome abundance globally regulates translation efficiency, but has not individual impact on the mRNA–protein correlation. The specific amino acid usage for protein synthesis has been reported to contribute to 7–8% of the overall variation in *Desulfovibrio* and yeast [60,71]. Minor effects are attributed translation initiation, start codon, stop codon and stop codon context [59,71].

Other, perhaps more important features to be considered and possibly responsible for significant outliers are untranslated RNA species, as well as secreted proteins escaping identification by mass spectrometry. In eukaryotic cells mRNA distribution and

sequestration to compartments such as the nucleus also influences the translation rate.

4.8. Experimental error and noise

Gene expression is composed by several stochastic steps involving species acting at very low concentration. These properties make it especially difficult to monitor. Some microarrays based experiments have been used successfully in different systems to measure down to the level of two copies per cell in yeast [72]. Similar results have been found for other systems and platforms [73,74]. Although RNA-seq has improved the signal to noise ratio, it is not completely free of ambiguity. However, methods for transcript quantification are reproducible and sensitive. Some methods offer reliable quantifications in a broad range of concentrations spanning four orders of magnitude [6]. Also it provides a better accuracy than microarrays. A 40-million-read transcriptome data-set provides reliable measurement of a single transcript per cell in human cell lines [6]. A comparison with five array platforms demonstrates a better resolution, reproducibility and robustness of a deep sequencing based method [75].

Noise is a prevalent feature of biological systems. For example, studying protein levels on a single cell level by GFP tagging of individual ORFs revealed strong protein specific differences regarding abundance and function [37]. Besides noise on a single cell level and the biological factors described above, experimental error is significantly contributing to the observed variation of the correlation of mRNA and proteins. Nie et al. [45] analyzed global mRNA and protein levels in the bacterium *D. vulgaris*. Their multiple regression analysis showed that the analytical variation in protein abundance contributed to 34–44% of the total variation observed. However, this value cannot be generalized to other analyses; especially recent advances in quantitative mass spectrometry techniques should significantly yield lower error rates for transcriptome measurements. For example, Silva et al. reported an error of $\pm 15\%$ in an absolute quantification study using the average MS signal intensities for the three most abundant peptides per protein as quantitative measure and subsequent normalization to isotopically labeled peptides as internal standards [35]. For a proteome-level comparison of two cell lines, Pan et al. reported a very high accuracy and reproducibility for the quantification of 4000 proteins. The Pearson correlation coefficient between biological replicates was $r_p = 0.95$ [76].

The total error for quantitative characterization of complex samples using mass spectrometry is difficult to evaluate. Several experimental and technical sources, as well as data processing issues contribute. In elaborate large-scale MS studies typically few biological replicates are studied, mainly because data analysis is very time-consuming but also for reasons of machine availability. Sample processing and quantitative recovery of digested peptides can be a source of experimental error. Differences in liquid chromatography separation, peptide ionization and equipment variation complicate the comparison between two different MS-setups [77].

MS data analysis itself requires several steps, such as baseline adjustments, data normalization and filtering, peak detection and quantification, the application of error models and statistics analysis [78]. Key to the error evaluation of complex proteomics data-sets are threshold criteria for peptide identifications. Raising the required identification scores reduces the number of false positive identifications, while at the same time increasing the false negative rate [79]. Researchers are well aware of this signal to noise problem and other error sources, such as sequence data base bias, when interpreting mass spectrometric data and sophisticated experimental and statistical methods exist to evaluate and minimize the technical error due to MS analysis [80,81].

Recently significance analysis of microarray data (SAM) has been adapted to quantitative proteomic data. SAM assigns a significance value, a false positive and false negative rate for differential expression of individual proteins. Such information is not readily available by conventional *t*-test or fold change test alone [82]. A large variety of MS data analysis software packages exist, also comprising statistical tools for data analysis (for an overview see [81,83–85]).

5. Concluding remarks and outlook

The correlation between mRNA and protein abundance depends on various biological and technical factors. To quantify their respective influence, multiple regression analysis has proven useful [45,60]. While some of these factors can be directly determined from available sequence data, such as RNA secondary structure, codon bias and amino acid usage, others need to be determined experimentally, such as mRNA abundance, ribosome occupancy, protein abundance and turnover. Fig. 1 gives an overview of the central pathway of molecular biology and schematically shows current approaches to determine these factors experimentally.

It is difficult to evaluate on a global level which biological factor, translation efficiency or protein half-life, most prominently influences the correlation between mRNA and protein abundances. In yeast, one third of all transcripts appear to be regulated translationally with roughly a 100-fold range of translation efficiency [63]. Individual protein half-lives range from several seconds to tens of hours [70], a more than 1000-fold range. Hence protein turnover is probably influencing the correlation between mRNA and protein abundances to a greater degree. Studies on individual protein turnover rates on an organism-wide level appear within reach with current technical advances in mass spectrometry applying labeled isotope techniques.

In general, the recent advances in mass spectrometry-based protein quantification techniques with low experimental error and high accuracy in quantification, and quantitative deep sequencing methods for mRNA should help in future analyses to give a more accurate picture of the organism-wide correlation between mRNA and protein and the influence of respective modifying parameters. Analyses of samples from the same biological system under various conditions, such as response to cellular stresses or analysis at different stages of the cell cycle should yield specific physiological information derived from the -omics analysis of complex biological samples.

Future challenges lie in data base development and the computational integration of -omics data-sets from different sources, allowing for example the seamless mapping of identified proteins to corresponding mRNA data [86]. Also the correct assessment of technical and experimental errors is crucial for the meaningful interpretation of integrative studies, such as on the correlation of mRNA and protein abundances. Solid statistical methods embedded in a larger framework need to be established, to allow meaningful conclusions from comparative studies on large data-sets.

Systems biology crucially depends on the availability of large-scale high-quality quantitative data-sets. Therefore this thriving new field in biology is intimately related to the advancement of -omics methods. The example of mRNA-protein correlation shows how accurate data-sets help to identify and quantify biological factors influencing this central pathway.

Acknowledgments

The authors would like to thank Henrik Molina for critically reading the manuscript. T.M. is funded by an EMBO Long-Term Fellowship and M.G. by a FPU Fellowship from the Spanish Education

Ministry. The work is partly funded by a European Research Council and the Consolider Programme of the Spanish Education Ministry.

References

- Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O. and Herschlag, D. (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA 100, 3889–3894.
- McGall, G.H. and Christians, F.C. (2002) High-density genechip oligonucleotide probe arrays. Adv. Biochem. Eng. Biotechnol. 77, 21–42.
- Draghici, S., Khatri, P., Eklund, A.C. and Szallasi, Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 22, 101–109.
- Eklund, A.C., Turner, L.R., Chen, P., Jensen, R.V., deFeo, G., Kopf-Sill, A.R. and Szallasi, Z. (2006) Replacing cRNA targets with cDNA reduces microarray cross-hybridization. Nat. Biotechnol. 24, 1071–1073.
- Okoniewski, M.J. and Miller, C.J. (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinform. 7, 276.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628.
- Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. Nat. Biotechnol. 25, 195–203.
- Ramakrishnan, S.R. et al. (2009) Integrating shotgun proteomics and mRNA expression data to improve protein identification. Bioinformatics 25, 1397–1403.
- Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H. and Long, R.M. (1998) Localization of ASH1 mRNA particles in living yeast. Mol. Cell 2, 437–445.
- Golding, I., Paulsson, J., Zawilski, S.M. and Cox, E.C. (2005) Real-time kinetics of gene activity in individual bacteria. Cell 123, 1025–1036.
- Chubb, J.R., Treck, T., Shenoy, S.M. and Singer, R.H. (2006) Transcriptional pulsing of a developmental gene. Curr. Biol. 16, 1018–1025.
- Le, T.T., Harlepp, S., Guet, C.C., Dittmar, K., Emonet, T., Pan, T. and Cluzel, P. (2005) Real-time RNA profiling within a single bacterium. Proc. Natl. Acad. Sci. USA 102, 9160–9164.
- Janicki, S.M. et al. (2004) From silencing to gene expression: real-time analysis in single cells. Cell 116, 683–698.
- Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc. Natl. Acad. Sci. USA 97, 9390–9395.
- Gorg, A., Weiss, W. and Dunn, M.J. (2004) Current two-dimensional electrophoresis technology for proteomics. Proteomics 4, 3665–3685.
- Miller, I., Crawford, J. and Gianazza, E. (2006) Protein stains for proteomic applications: which, when, why? Proteomics 6, 5385–5408.
- Patton, W.F. and Beechem, J.M. (2002) Rainbow's end: the quest for multiplexed fluorescence quantitative analysis in proteomics. Curr. Opin. Chem. Biol. 6, 63–69.
- Larbi, N.B. and Jefferies, C. (2009) 2D-DIGE: comparative proteomics of cellular signalling pathways. Methods Mol. Biol. 517, 105–132.
- Timms, J.F. and Cramer, R. (2008) Difference gel electrophoresis. Proteomics 8, 4886–4897.
- Ong, S.E. and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. Nat. Chem. Biol. 1, 252–262.
- Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422, 198–207.
- Domon, B. and Aebersold, R. (2006) Mass spectrometry and protein analysis. Science 312, 212–217.
- Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K. and Horning, S. (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. Anal. Chem. 78, 2113–2120.
- Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B. and Aebersold, R. (2009) Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. Cell.
- Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A. and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell Proteomics 1, 376–386.
- Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 17, 994–999.
- Ross, P.L. et al. (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol. Cell Proteomics 3, 1154–1169.
- de Godoy, L.M., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Frohlich, F., Walther, T.C. and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455, 1251–1254.
- Graumann, J. et al. (2008) Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantification of mouse embryonic stem cells to a depth of 5111 proteins. Mol. Cell Proteomics 7, 672–683.

- [30] Liu, H., Sadygov, R.G. and Yates 3rd, J.R. (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 76, 4193–4201.
- [31] Allet, N. et al. (2004) In vitro and in silico processes to identify differentially expressed proteins. *Proteomics* 4, 2333–2351.
- [32] Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J. and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of detected peptides per protein. *Mol. Cell Proteomics* 4, 1265–1272.
- [33] Bondarenko, P.V., Chelius, D. and Shaler, T.A. (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal. Chem.* 74, 4741–4749.
- [34] Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. and Gygi, S.P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. USA* 100, 6940–6945.
- [35] Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P. and Geromanos, S.J. (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell Proteomics* 5, 144–156.
- [36] Malmstrom, J., Beck, M., Schmidt, A., Lange, V., Deutsch, E.W. and Aebersold, R. (2009) Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. *Nature* 460, 762–765.
- [37] Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L. and Weissman, J.S. (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 441, 840–846.
- [38] 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.
- [39] Fletcher, B., Latter, G.L., Monardo, P., McLaughlin, C.S. and Garrels, J.I. (1999) A sampling of the yeast proteome. *Mol. Cell Biol.* 19, 7357–7368.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] Washburn, M.P., Koller, A., Oshiro, G., Ulaszek, R.R., Plouffe, D., Decui, C., Winzler, E. and Yates 3rd, J.R. (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* 100, 3107–3112.
- [44] Schmidt, M.W., Houseman, A., Ivanov, A.R. and Wolf, D.A. (2007) Comparative proteomic and transcriptomic profiling of the fission yeast *Schizosaccharomyces pombe*. *Mol. Syst. Biol.* 3, 79.
- [45] Nie, L., Wu, G. and 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.
- [46] Tian, Q. et al. (2004) Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Mol. Cell Proteomics* 3, 960–969.
- [47] Yu, E.Z., Burba, A.E. and Gerstein, M. (2007) PARE: a tool for comparing protein abundance and mRNA expression data. *BMC Bioinform.* 8, 309.
- [48] Grossman, A.D., Zhou, Y.N., Gross, C., Heilig, J., Christie, G.E. and Calendar, R. (1985) Mutations in the *rpoH* (*htpR*) gene of *Escherichia coli* K-12 phenotypically suppress a temperature-sensitive mutant defective in the sigma 70 subunit of RNA polymerase. *J. Bacteriol.* 161, 939–943.
- [49] Nahvi, A., Barrick, J.E. and Breaker, R.R. (2004) Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. *Nucl. Acids Res.* 32, 143–150.
- [50] Gottesman, S. (2004) The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.* 58, 303–328.
- [51] Mehta, P., Goyal, S. and Wingreen, N.S. (2008) A quantitative comparison of RNA-based and protein-based gene regulation. *Mol. Syst. Biol.* 4, 221.
- [52] Ikemura, T. and Dahlberg, J.E. (1973) Small ribonucleic acids of *Escherichia coli*. II. Noncoordinate accumulation during stringent control. *J. Biol. Chem.* 248, 5033–5041.
- [53] Morfeldt, E., Taylor, D., von Gabain, A. and Arvidson, S. (1995) Activation of alpha-toxin translation in *Staphylococcus aureus* by the trans-encoded antisense RNA, RNAlII. *EMBO J.* 14, 4569–4577.
- [54] Sharp, P.M. and Li, W.H. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* 4, 222–230.
- [55] Friberg, M., von Rohr, P. and Gonnet, G. (2004) Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in *Saccharomyces cerevisiae*. *Yeast* 21, 1083–1093.
- [56] Jansen, R., Bussemaker, H.J. and Gerstein, M. (2003) Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucl. Acids Res.* 31, 2242–2251.
- [57] Kliman, R.M., Irving, N. and Santiago, M. (2003) Selection conflicts, gene expression, and codon usage trends in yeast. *J. Mol. Evol.* 57, 98–109.
- [58] Wu, G., Nie, L. and Freeland, S.J. (2007) The effects of differential gene expression on coding sequence features: analysis by one-way ANOVA. *Biochem. Biophys. Res. Commun.* 358, 1108–1113.
- [59] Lithwick, G. and Margalit, H. (2003) Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res.* 13, 2665–2673.
- [60] Wu, G., Nie, L. and Zhang, W. (2008) Integrative analyses of posttranscriptional regulation in the yeast *Saccharomyces cerevisiae* using transcriptomic and proteomic data. *Curr. Microbiol.* 57, 18–22.
- [61] MacKay, V.L. et al. (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.
- [62] Inada, T., Winstall, E., Tarun Jr., S.Z., Yates 3rd, J.R., Schieltz, D. and Sachs, A.B. (2002) One-step affinity purification of the yeast ribosome and its associated proteins and mRNAs. *RNA* 8, 948–958.
- [63] Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. and Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223.
- [64] Halbeisen, R.E. and Gerber, A.P. (2009) Stress-dependent coordination of transcriptome and translateome in yeast. *PLoS Biol.* 7, e105.
- [65] 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.
- [66] Varshavsky, A. (1992) The N-end rule. *Cell* 69, 725–735.
- [67] Mogk, A., Schmidt, R. and Bukau, B. (2007) The N-end rule pathway for regulated proteolysis: prokaryotic and eukaryotic strategies. *Trends Cell Biol.* 17, 165–172.
- [68] Pratt, J.M., Petty, J., Riba-Garcia, I., Robertson, D.H., Gaskell, S.J., Oliver, S.G. and Beynon, R.J. (2002) Dynamics of protein turnover, a missing dimension in proteomics. *Mol. Cell Proteomics* 1, 579–591.
- [69] Cargile, B.J., Bundy, J.L., Grunden, A.M. and Stephenson Jr., J.L. (2004) Synthesis/degradation ratio mass spectrometry for measuring relative dynamic protein turnover. *Anal. Chem.* 76, 86–97.
- [70] Doherty, M.K., Hammond, D.E., Clague, M.J., Gaskell, S.J. and Beynon, R.J. (2009) Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. *J. Proteome Res.* 8, 104–112.
- [71] 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.
- [72] Holland, M.J. (2002) Transcript abundance in yeast varies over six orders of magnitude. *J. Biol. Chem.* 277, 14363–14366.
- [73] Kane, M.D., Jatke, T.A., Stumpf, C.R., Lu, J., Thomas, J.D. and Madore, S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucl. Acids Res.* 28, 4552–4557.
- [74] Czechowski, T., Bari, R.P., Stitt, M., Scheible, W.R. and Udvardi, M.K. (2004) Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* 38, 366–379.
- [75] t Hoen, P.A. et al. (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucl. Acids Res.* 36, e141.
- [76] Pan, C., Kumar, C., Bohl, S., Klingmueller, U. and Mann, M. (2009) Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. *Mol. Cell Proteomics* 8, 443–450.
- [77] 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.
- [78] Listgarten, J. and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell Proteomics* 4, 419–434.
- [79] Steen, H. and Mann, M. (2004) The ABC’s (and XYZ’s) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* 5, 699–711.
- [80] Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. and Gygi, S.P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* 2, 43–50.
- [81] Nesvizhskii, A.I., Vitek, O. and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 4, 787–797.
- [82] Roxas, B.A. and Li, Q. (2008) Significance analysis of microarray for relative quantitation of LC/MS data in proteomics. *BMC Bioinform.* 9, 187.
- [83] Mueller, L.N., Brusniak, M.Y., Mani, D.R. and Aebersold, R. (2008) An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.* 7, 51–61.
- [84] Lau, K.W., Jones, A.R., Swainston, N., Siepen, J.A. and Hubbard, S.J. (2007) Capture and analysis of quantitative proteomic data. *Proteomics* 7, 2787–2799.
- [85] Panchaud, A., Affolter, M., Moreillon, P. and Kussmann, M. (2008) Experimental and computational approaches to quantitative proteomics: status quo and outlook. *J. Proteomics* 11, 19–33.
- [86] Kumar, C. and Mann, M. (2009) Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett.* 583, 1703–1712.