

A gene's mRNA level measured by microarrays or RT-PCR does not necessarily predict its protein level: R^2 (or R_s) for plots of mRNA versus protein is ≤ 0.4

by Nancy Kendrick, Kendrick Labs Inc, Madison, WI.

The 21st century began with a most magnificent success: publication of the human genome sequence of 2.85 billion base pairs. After two preliminary papers by HGC (International Human Genome Sequencing Consortium) and Celera in 2001, the finalized publication in 2004 (HGC) provided 99.7% of the sequence with only 1 error/100,000 bases. [1], www.genome.gov/12513430 This colossal feat has influenced all subsequent biomedical research.[2] A multitude of high-throughput tools have since been developed to expand and compare genomes and mRNA. Genome sequences of >250 eukaryotes and > 4000 bacteria and viruses have been published; BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) is providing a giddy amount of information, especially with regard to evolution.

But although RNA was the primordial molecule, proteins are the stuff of life now. The >200 cell types in the human body have the same DNA yet they take different shapes and functions because of varying protein composition. Unfortunately, the latter is hard to figure out. The human genome codes for only 20,500 proteins [3], but alternative splicing [4] and post-translational modifications create many variants of physiological importance.

Since the 2004 milestone, a great deal of biomedical research has been published that assumes that proteins, difficult to measure, are proportional to the levels of corresponding mRNA, easy to measure. Unfortunately, this intuitive assumption is *wrong*. *Less than 40% of cellular protein levels can be predicted from mRNA measurements.*

What is the evidence? Since 2004, at least 4 groups have published data on the global correlation between protein and mRNA concentration in mammalian cells. Their correlation results are either expressed in terms of the coefficient of determination (R^2) or the Spearman correlation coefficient (R_s). R^2 gives the fraction of the variability in Y that can be explained by the variability in X through their linear relationship. R_s is similar, but is recommended for use with data that are skewed or have outliers. [5]. Results from the four groups, summarized below, are in surprisingly good agreement.

1. Tian et. al. [6] in 2004 measured steady state levels of protein (ICAT/MS) and mRNA (Agilent microarrays) in multipotent mouse EML cells and their differentiated progeny MPRO cells. The abundance ratios of 425 proteins were mapped to the corresponding mRNA expression levels. Signature genes (150) were identified that showed significant changes at either protein and/or mRNA level between the two cell types. Of these, 29 (19%) showed good correlation between mRNA and protein levels, 67 (45%) showed significant changes at the mRNA but not the protein level, and 52 genes (35%) showed significant changes at the protein but not the mRNA level. Two genes (1%) showed opposite expression patterns of mRNA and protein. The correlation coefficient R between mRNA and protein was 0.64 ($R^2 = 0.41$) for the signature genes and 0.59 ($R^2 = 0.35$) for all genes examined. Interestingly, the c-kit receptor kinase protein and mRNA expression varied 7-fold and 9-fold respectively between the two cell lines. However the c-kit *ligand* protein, aka stem cell factor, had a 5-fold higher level in the EML cells but no change in mRNA levels. Nine mitochondrial proteins showed significantly lower protein levels in the MPRO cells but higher or similar levels of mRNAs compared to those in the EML cells. Five out of six mRNA processing genes showed a negative correlation between mRNA and protein levels.

2. Vogel et. al. [7] in 2010 pointed out that transcription, mRNA decay, translation, and protein degradation are key processes determining steady state protein concentrations. They measured protein using shotgun MS and mRNA levels using microarrays for 1025 genes from human medulloblastoma cells harvested at steady state logarithmic growth. A high-confidence dataset of 512 genes was chosen for close examination. A plot of protein versus mRNA level for these proteins gave an R^2 value of 0.29; R_s value was 0.46. Furthermore, about 200 sequence features (coding sequence lengths, nucleotide frequencies and properties etc) were checked for correlation with protein abundance. The authors concluded that the processes of translation and protein degradation are at least as important as mRNA transcription and stability to steady-state protein abundance.
3. Lundberg et. al. [8] in 2010 performed a global analysis of mRNA and relative protein abundance in 3 human cancer cell lines: a brain glioblastoma (U-251MG) ; epidermoid squamous cell carcinoma (A-431); and bone osteosarcoma (U-2 OS). To measure mRNA, they used digital RNA seq, a method more specific and sensitive than microarrays,. A total of 15,538 transcripts were detected in the three cell lines of which 11,575 (74.5%) were common to all three. Only 559, 572 and 922 transcripts were unique to the brain, epidermoid and bone cell lines respectively.

To measure protein levels, they used a triple-SILAC (stable isotope labeling by amino acids in cell culture) method in which the three cell lines were grown with amino acids with different isotopes and then analyzed by mass spectrometry. In this method, triple peak patterns are generated for each protein; identification of one of the peptides automatically yields the identity of the other two. A total of 5456 proteins were quantified. Of these, 5333 (97.7%) were found in all three cell lines. About 65% of the detected proteins have similar expression levels (less than 2-fold differences) in the three cell lines while about one third had differential expression. Only 3, 27, and 34 proteins were found to be unique to the brain, epidermoid, and bone cell lines respectively.

While this paper focused on the differences between the cell lines, Figure S8 in supplemental data showed the 3 correlation plots between mRNA transcript and protein levels. The Spearman correlation coefficients were 0.42, 0.42 and 0.43 respectively for bone osteosarcoma ($n = 5210$), epidermoid squamous cell carcinoma ($n = 5158$), and brain glioblastoma cells ($n = 5,197$).

4. Going further, Schwanhausser, et. al. [9] in 2011 performed a definitive global analysis of protein versus mRNA levels in NIH3T3 mouse fibroblasts. This German group used metabolic pulse labeling (SILAC) followed by mass spectrometry to measure protein turnover, and 4-thiouridine to measure mRNA turnover in exponentially growing, non-synchronized cells. They identified 6445 unique proteins of which 5279 were quantified by at least three peptide ratios. In parallel, newly synthesized RNA was pulse labeled for 2 h with 4-thiouridine. RNA samples were fractionated, analyzed by mRNA sequencing, and quantified. Both protein and mRNA half-lives were calculated. Proteins were, on average 5 times more stable (median half-life 46 h) than mRNA (9h). There was no correlation between mRNA and protein half-lives ($R^2 = 0.02$).

These authors calculated absolute cellular mRNA copy number based on number of sequencing reads in the unfractionated sample. They calculated absolute protein copy numbers from the mass spectrometry data, by summing peak intensities of all peptides matching to a specific

protein. To avoid bias, they restricted their analysis to 5028 genes that were identified at both the mRNA and protein levels. In this subset, proteins were on average, 900 times more abundant than corresponding mRNAs and their concentrations spanned 5 orders of magnitude. Despite the huge spread, log-log plot of mRNA versus protein showed a R^2 value of 0.41. Removal of less reproducible data points did not improve the correlation, but rather brought R^2 to about 0.3 (Supplemental Figure S7).

The authors looked at protein and mRNA stability of different functional sets using gene ontology (Figure 5).

- Many housekeeping genes such as those coding for ribosomal, glycolytic and TCA cycle proteins, showed stable mRNAs and stable proteins.
- Chromatin modifying enzymes, transcription factors, and genes with cell-cycle-specific functions tended to have unstable mRNA and unstable protein.
- RNA-processing proteins tended to have stable mRNAs and unstable proteins.
- Genes for secreted proteins, kinases, proteases, and integrin-mediated pathways tended to have stable mRNA and unstable protein.

Thus, since 2004, at least 4 groups using increasingly sophisticated methods have independently determined the following: only about 40 percent of protein levels in cultured mammalian cells are explained by mRNA levels. One group, Schwanhausser, et. al. [9] presented evidence suggesting that mRNA is probably a good surrogate for protein levels for housekeeping genes such as ribosomal proteins, glycolytic enzymes and TCA cycle proteins (stable mRNA and protein), but probably a poorer surrogate for kinases, proteases, secreted proteins and transcription factors that are targets for cancer drugs (stable mRNA, unstable protein). mRNA probably correlate poorly with protein levels for transcription factors (unstable mRNA, unstable protein).

While the studies above are in agreement about the general relationship between mRNA and protein, they were limited to cultured cells. But cultured cancer cells are almost certainly not representative of differentiated cells in the human body. *In vivo*, cells are under the influence of constant paracrine and exocrine signaling required for tissue differentiation. Post-translational modifications of proteins such as tyrosine phosphorylation are known to be a key mechanism in signal transduction. Matthias Mann's group recently found that a remarkable 70% of HeLa cell proteins are phosphorylated sometime during the cell cycle [10]. Six cases where the correlations between mRNA and protein levels was studied in human patients are described below.

5. Taquet, et. al. [11] studied mRNA and protein expression of somatostatin receptor 5 (SSR5) and chemokine(C-C motif)receptor 7 (CCR7) in ten Crohn's disease patients and healthy controls using Real-Time PCR and IHC. Peripheral blood mononuclear cells (PBMCs) were separated using density gradient centrifugation; mucosal biopsies were obtained during colonoscopy. A significant increase in mRNA expression, 417 +/- 71 times ($P < 0.05$), was observed for STTR5 in Crohn's disease versus control patients in PBMCs. However, no increase in protein expression was detected. On the other hand, CCR7 mRNA and protein expression in mucosal biopsies were both 10-fold increased in Crohn's disease compared to controls. Data from mRNA alone would have been misleading for STTR5.
6. Dickson et. al. [12] used in situ hybridization with an anti-sense ^{33}P -labeled cRNA probe to quantify mRNA and immunohistochemistry (IHC) to determine protein expression of the JAG1

gene in breast cancer. The JAG1 protein activates the Notch signaling cascade that plays a role in cell differentiation and division. This group had previously showed that high levels of JAG1 or NOTCH1 mRNA were correlated with poor prognosis of breast cancer [13]. While the correlation between high JAG1 and poor prognosis held, this group found only 65% agreement between mRNA and protein levels. Patients with tumors expressing high JAG1 protein, high mRNA or both, had a 10-year survival of 31%, 19% and 11% respectively. Thus the IHC protein measurement alone was not as good predictor as the mRNA alone. But the best prediction occurred when protein and mRNA measurements agreed.

7. Stark et. al. [14] compared the mRNA (RT-PCR) and protein levels (IHC) of apoptosis regulating genes p53, BCL-2 and BAX in breast cancer primary tumors and brain metastases. The mRNA level of p53 was significantly lower in brain metastases than primary tumors but protein levels were only slightly lower (not significant). BCL-2 mRNA and protein expression were in good agreement; both significantly lower in brain metastases. BAX mRNA and protein levels were clearly discordant; mRNA levels were down in metastases while protein levels were higher.
8. CD20 antigen, a glycosylated phosphorylated protein expressed on the surface of B cells, is the target of the therapeutic monoclonal antibody rituximab. Sarro et al. [15] quantified CD20 mRNA using RT-PCR and protein levels using quantitative immunoblots in chronic lymphocytic leukemia (CLL) cells from patients. They found that CD20 protein was decreased by about 60% in CLL cells versus healthy donors. However, CD20 mRNA levels were normal or near-normal in CLL cells and did not correlate with protein levels. The protein decrease is likely the reason for the lower effectiveness of rituximab against CLL compared to other B cell malignancies.
9. Matrix metalloproteinases (MMP) degrade extracellular matrix and have been implicated in tumor invasion and metastasis. Lichtinghagen et al. [16] measured MMP-2 and MMP-9 levels along with tissue inhibitor of metalloproteinases (TIMP-1) in 17 patients with prostate cancer. They measured mRNA using RT-PCR and protein levels using quantitative zymography (with calibration curve) in paired benign and malignant tissue samples. Zymography showed that the proforms of MMP-2 and MMP-9 predominate in the prostate; activated forms below the main bands were not detected. MMP-9 protein levels were higher in cancer tissue than in normal. However, there was no significant correlation between the mRNA and protein expression of MMP-2, MMP-9 and TIMP-1 in either cancerous or noncancerous tissue. Values for R_s (Spearman rank correlation coefficient) for plots of protein versus mRNA from 34 tissue samples were 0.109 for MMP-2, 0.059 for MMP-9 and 0.077 for TIMP-1.
10. Shebl et. al. [17] measured the secreted protein and cellular mRNA levels of 20 cytokines produced by peripheral blood mononuclear cells (PBMCs) from 26 women: 19 vaccinated with an GPV-16 VLP vaccine versus 7 placebo. The cells were collected pre- and 2 month post-vaccination and cryopreserved until use. They were cultured for 72 hours in single wells in-vitro with the VLP influenza A virus or baculovirus insect cell lysate (BAC) control. Supernatants and cell pellets were obtained from the same well. The cell free supernatants were tested in duplicate for cytokines using the 22-plex assay developed by Linco Research. Remaining supernatants and cell pellets were used for total RNA extractions and analyzed with Affymetrix Human Genome Focus Assay. Plots of protein versus mRNA were generated and Spearman's correlation coefficients determined. Of the 20 cytokines analyzed, one, IFN-gamma, showed a very strong correlation with R_s of 0.90. Three cytokines (MIP1A, IP10 and TNF-alpha) showed

correlations of 0.6 or higher. Five showed modest correlations of 0.4 to 0.59 (MCP1, IL-2, GM-CSF, IL-5 and RANTES). The remaining 11 cytokines showed weak or negative correlations. One conclusion of these authors was "Investigators who use currently available expression array tools should be careful not to assume that mRNA expression changes identified by expression studies would necessarily reflect similar changes in corresponding protein levels."

The conclusion from the ten examples listed above seems inescapable: mRNA levels cannot be used as surrogates for corresponding protein levels without verification.

References

1. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
2. Lander, E.S., *Initial impact of the sequencing of the human genome.* Nature, 2011. **470**(7333): p. 187-97.
3. Clamp, M., et al., *Distinguishing protein-coding and noncoding genes in the human genome.* Proc Natl Acad Sci U S A, 2007. **104**(49): p. 19428-33.
4. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
5. Zou, K.H., K. Tuncali, and S.G. Silverman, *Correlation and simple linear regression.* Radiology, 2003. **227**(3): p. 617-22.
6. Tian, Q., et al., *Integrated genomic and proteomic analyses of gene expression in Mammalian cells.* Mol Cell Proteomics, 2004. **3**(10): p. 960-9.
7. Vogel, C., et al., *Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line.* Mol Syst Biol, 2010. **6**: p. 400.
8. Lundberg, E., et al., *Defining the transcriptome and proteome in three functionally different human cell lines.* Mol Syst Biol, 2010. **6**: p. 450.
9. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.
10. Olsen, J.V., et al., *Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis.* Science signaling, 2010. **3**(104): p. ra3.
11. Taquet, N., et al., *Differential between protein and mRNA expression of CCR7 and SSTR5 receptors in Crohn's disease patients.* Mediators Inflamm, 2009. **2009**: p. 1-10.
12. Dickson, B.C., et al., *High-level JAG1 mRNA and protein predict poor outcome in breast cancer.* Mod Pathol, 2007. **20**(6): p. 685-93.
13. Reedijk, M., et al., *High-level coexpression of JAG1 and NOTCH1 is observed in human breast cancer and is associated with poor overall survival.* Cancer Res, 2005. **65**(18): p. 8530-7.
14. Stark, A.M., et al., *Reduced mRNA and protein expression of BCL-2 versus decreased mRNA and increased protein expression of BAX in breast cancer brain metastases: a real-time PCR and immunohistochemical evaluation.* Neurol Res, 2006. **28**(8): p. 787-93.
15. Sarro, S.M., et al., *Quantification of CD20 mRNA and protein levels in chronic lymphocytic leukemia suggests a post-transcriptional defect.* Leuk Res, 2010. **34**(12): p. 1670-3.
16. Lichtinghagen, R., et al., *Different mRNA and protein expression of matrix metalloproteinases 2 and 9 and tissue inhibitor of metalloproteinases 1 in benign and malignant prostate tissue.* Eur Urol, 2002. **42**(4): p. 398-406.
17. Shebl, F.M., et al., *Comparison of mRNA and protein measures of cytokines following vaccination with human papillomavirus-16 L1 virus-like particles.* Cancer Epidemiol Biomarkers Prev, 2010. **19**(4): p. 978-81.