Data merging for integrated microarray and proteomic analysis

Katrina M. Waters, Joel G. Pounds and Brian D. Thrall

Advance Access publication date 10 May 2006

Abstract
The functioning of even a simple biological system is much more complicated than the sum of its genes, proteins and metabolites. A premise of systems biology is that molecular profiling will facilitate the discovery and characterization of important disease pathways. However, as multiple levels of effector pathway regulation appear to be the norm rather than the exception, a significant challenge presented by high-throughput genomics and proteomics technologies is the extraction of the biological implications of complex data. Thus, integration of heterogeneous types of data generated from diverse global technology platforms represents the first challenge in developing the necessary foundational databases needed for predictive modelling of cell and tissue responses. Given the apparent difficulty in defining the correspondence between gene expression and protein abundance measured in several systems to date, how do we make sense of these data and design the next experiment? In this review, we highlight current approaches and challenges associated with integration and analysis of heterogeneous data sets, focusing on global analysis obtained from high-throughput technologies.

Keywords: proteomics; microarray; data integration

INTRODUCTION

Gene expression profiling is rapidly becoming accepted as the state-of-the-art approach for investigating genome–wide changes in gene expression. Microarray and serial analysis of gene expression (SAGE) approaches have evolved substantially during the past decade and continue to offer significant advantages over conventional approaches that investigate the regulation of a small subset of genes [1]. These profiling technologies have an important strategic advantage over most means of measuring gene expression, as they do not require selection of the important genes in advance. As such, microarrays and SAGE provide an unbiased approach to identifying genes whose regulation is changed during environmental perturbation of the cell, tissue or organism [2].

Complementary protein profiling on a global scale is a more diverse, rapidly evolving and expanding field. Proteomics has come to encompass many technologies and approaches to study protein changes in time and space on a global basis. As protein concentration is an important variable with respect to enzyme activity, proteomic data connects genomics to the physical chemistry of the cell. Despite explosive growth in both academic and commercial efforts, concrete technical capabilities are far from adequate to realize the promise of this field. Yet, accurate and sensitive identification of proteins, their form and abundance are essential to defining the relationship between gene regulation, protein interactions and signalling networks in a cellular system. A description of the various transcript and protein-profiling technologies is outside the scope
of this review; the reader is directed to recent, focused reviews of proteomic technologies and approaches [3–6].

By correlating changes in gene expression and protein abundance with changes in cell and tissue function, it should be possible to derive insight into a broad range of biological processes. However, growing numbers of studies indicate that the use of gene expression patterns is insufficient to predict abundance of proteins, as additional post-transcriptional mechanisms, including translation, post-translational modifications and degradation, also modulate the steady-state level of a protein present in a given cell or tissue [7–10]. The expectations for concordance between transcript or protein abundance and the statistical/bioinformatics approaches required are defined by the type and limitations of transcript and proteomic information obtained by application of these diverse technologies, along with the strengths and limitations of the experimental design and the objectives of the experiment. Furthermore, the tools and approaches for merging of gene expression and protein abundance data sets into a comprehensive reference set prior to integrated analysis remain a limiting factor. Nonetheless, it is expected that by integrating genomic and proteomic data sets along with improved annotation, functional correlation can be derived across biological pathways for these disparate data [11]. From a practical perspective, the facile, routine merging and integration of high-density, heterogeneous data sets taken from different laboratories or across experimental platforms is essential for high-throughput data validation, prioritization of follow-on experiments and effective use of the data to illuminate complex biological systems.

Many sources of variability associated with the global measurement of transcripts and proteins, such as differences in sensitivity, dynamic range, ambiguity in identification, etc., may contribute to the potential discordance between mRNA and protein abundances. It is also well-recognized that cells regulate gene expression and protein abundance separately, and that a single gene does not usually translate into a single protein. Protein abundance depends not only on transcription rates of the gene but also on additional control mechanisms, such as mRNA stability, translational regulation and protein degradation. Moreover, the activity of proteins can be altered through a variety of post-translational modifications or proteolytic cleavage. These issues are outside the scope of this review and hence are only briefly mentioned. In this review, we focus on the steps involved in merging and integration of proteomic and transcriptomic data, by providing examples of challenges associated with working across databases, requiring manual curation of data sets and making high-throughput comparisons extremely time-consuming. In particular, we discuss the inconsistency of databases that cross-reference gene and protein identifiers and sources of biological annotation, highlighting commercial and freely available tools needed for such analyses. Finally, we review the results of recent studies that have examined the correlation between mRNA and protein abundance in eukaryotic systems and the emerging trends observed from these efforts that illustrate how our understanding of the complexity of gene regulation at a global scale is still limited.

**EXPERIMENTAL DESIGN CONSIDERATIONS**

Mapping the quantitative relationships between thousands of genes and proteins is extraordinarily complex, yet global understanding of these relationships should provide new insights into cellular function and dysfunction. ‘Concordance’ is less a mathematical or statistical definition than a subjective conclusion based on the expectations of the investigator in the context of a particular experiment. In addition to the well-described uncertainties in their identification and quantification, transcripts and proteins have quite different half-lives, 0.1–10 h versus 0.5–500 h. Moreover, protein synthesis and maximal abundance is delayed relative to the mRNA expression/abundance. Thus, the concordance of transcriptomic–proteomic data is dependent, in part, on measurement times and whether the biological system is in steady state or perturbed by a stimulus. Concordance may be operationally defined by the investigator as direction (both mRNA and protein increased or decreased in concert), amplitude (consistent magnitude of change), temporal (appropriate change of levels in time) and/or functional (consistent modification of a functional pathway). As both transcriptomic and proteomic measurements become more sensitive, cheaper and more global, it will be possible to conduct more insightful experiments that define the temporal relationships between gene expression and protein abundances.
POTENTIAL SOURCES OF ERROR IN mRNA AND PROTEIN ABUNDANCE

The sources of quantitative and qualitative variability that contribute to discordance between mRNA and protein abundances are multiple. In part, error associated with different microarray and proteomic measurement platforms contributes to the variation observed. Analytical variability associated with technical replicates in microarray platforms is becoming well-characterized, with the major source of experimental variability arising from hybridization errors, including fluctuations in target molecule binding and cross-hybridization [12, 13]. This analytical noise is strongly dependent on the expression level, generally decreasing with increasing mRNA abundance [12]. Thus, at medium-to-high mRNA levels, expression changes of 1.5-fold can be statistically significant in well-controlled microarray experiments. In contrast, the sources of technical variability inherent in emerging global proteomic technologies are only beginning to be understood. While 2D-polyacrylamide gel electrophoresis (PAGE)-based approaches have provided a wealth of new detail into biological systems, limitations in the sensitivity, dynamic range and resolution of co-migrating proteins inherent to this technology introduce bias towards identification of the most abundant proteins [14, 15]. Liquid chromatography–mass spectrometry (LC–MS) technologies have significantly improved upon some of these limitations, providing a much greater dynamic range and reducing biases associated with isoelectric point and co-migration found with 2D-PAGE [16–18]. Introduction of stable isotope post-labelling strategies, such as isotope-coded affinity tags (ICAT) [19], and quantitative cysteinyl-peptide enrichment [20] continue to improve the quantitative aspects of global proteomics [21–23]. Studies using model protein (standard) mixtures indicate that technical reproducibility of these quantitation strategies can be quite good, with analytical variabilities of <15% achievable from the point of trypsin digestion to MS quantitation [20]. Furthermore, coefficients of variation for technical replicates in LC–MS-based proteomic analyses can be comparable with the reproducibility of current microarray technologies [24]. These reports suggest that technical error associated with carefully controlled LC–MS experiments is insufficient to explain the discordance frequently observed between microarray and proteomic analyses. However, measurements of analytical variability do not account for variability associated with sample collection, preparation and fractionation, and while this topic has received little attention, it is an area where significant advancements in process control can likely be made.

DATA MERGING AND INTEGRATION

The first fundamental step in integrated analysis of microarray and proteomics data is to merge the data by cross-referencing the sequence identifiers. While this merging might appear to be rather straightforward, anyone who has attempted to integrate two microarray data sets from different platforms will agree that it is non-trivial. Each manufacturer or research group has its favourite set of identifiers, and sequence databases have frequent annotation updates. Also, many microarray platforms have multiple targets or ‘probe-sets’ for each gene, which must be averaged or dealt with statistically, as well as variable coverage of splice isoforms. The inclusion of protein identifications adds another level of complexity to the data-merging process. As proteomic technologies continue to improve in sensitivity, it is possible to measure hundreds of proteins per sample; thus, bioinformatics tools are necessary to automate the annotation-merging process and integration approaches. As discussed in subsequent sections, the non-uniformity of sequence databases and limitations in the bioinformatics tools available for cross-platform data merging are major contributors to the challenges associated with combined analysis of microarray and proteomic data.

Disparity among databases

Several DNA sequence databases are organized under the International Nucleotide Sequence Database Collaboration (http://www.insdc.org), which includes the National Center for Biotechnology Information (NCBI) sequence database (GenBank), as well as the sequence databases of the European Molecular Biology Laboratory (EMBL) and the DNA Databank of Japan. This collaboration seeks to use unified taxonomy and vocabulary terms for efficient cross-referencing across all databases. Similarly, several protein sequences databases are organized within the Universal Protein Resource (UniProt) [25], which is a compilation of the Swiss Protein Databank (Swiss-PROT), the translation of DNA sequences in...
EMBL and the Protein Information Resource. Although these efforts seek to create a single identifier for all genes or proteins, a one-gene-to-one-protein correlation is not inherent to biology, making the task of integrating sequence databases challenging. This can be particularly problematic in the case of splice variants, which are not easily discernable by most microarray platforms and peptide-based proteomic platforms.

As a result of the large amount of data continually submitted by scientists all over the world, most databases are not manually curated for accuracy. One problem, then, is inconsistent or incorrect annotation, as well as redundant sequences, present in public sequence databases. NCBI maintains the Reference Sequence project (RefSeq), a non-redundant, curated reference subset of the GenBank data, and EntrezGene, which provides a curated, gene-specific database with functional information, protein interactions, key citations and external database links [26]. Another key database is Unigene (also maintained by NCBI), which is generated from species-specific clustered nucleotide sequences that overlap with high-percent sequence identity [27]. This database is more comprehensive in its coverage of the genomes; whenever new sequences are added, the clusters are recalculated, resulting in some sequences being moved to a new cluster and redundant identifiers (IDs) removed completely. Thus, in merging genomic and proteomic data sets using Unigene IDs, it is important to ensure that the database build dates are the same.

A similar sequence integration effort, called the International Protein Index (IPI) [28], for the protein community is undertaken by the European Bioinformatics Institute for human, mouse and rat proteomes. IPI contains protein sequence data taken from UniProt, Ensembl and RefSeq databases, which are combined to create protein sequence sets for each species with a low level of redundancy. However, this method of sequence clustering has the same caveats as the Unigene database, so cross-referencing data sets across different IPI release dates can be difficult.

**Approaches to database merging**

There are many publicly available tools to cross-reference gene and protein identifiers. Most tools, however, were designed for only a particular organism or only allow the user to search one identifier at a time. A few of these tools provide for batch look-up and retrieval of sequence identifiers, which is a requirement for large data sets. DAVID is a web-accessible program that will rapidly annotate any list of gene identifiers or Affymetrix probe IDs with corresponding gene symbols, RefSeq and Unigene IDs for human, mouse, rat or fly genomes [29]. SOURCE [30] and MatchMiner [31] are comparable web-accessible tools that allow the user to cross-reference large lists of gene identifiers for human, mouse and rat genomes. Panther is a web-accessible tool provided by Applied Biosystems (http://www.pantherdb.org/) that cross-references gene identifiers as well as translating protein identifiers into genes. Unlike many of the tools mentioned above, the UniProt archive has a batch retrieval tool that translates IPI identifiers into RefSeq gene IDs, although the coverage is often incomplete (http://www.pir.uniprot.org/search/batch_AR.shtml).

Few papers analysing combined microarray and proteomics data have described in detail the steps required for merging large heterogeneous data sets. In a study comparing the presence or absence of proteins and their corresponding mRNA in platelets, McRedmond et al. [32] used a straightforward approach to merge two large data sets using the Unigene database. Affymetrix cross-reference tables were used to populate approximately 85% (over 10 000 genes) of their data set with Unigene identifiers. They then used gene names, GenBank IDs and sequence descriptions to identify the corresponding Unigene identifiers for the remaining 2000 probe-sets. When replicate identifiers were consolidated, they found that 25% of the Affymetrix probe-sets were redundant according to the Unigene identifier. To merge the microarray data with their proteomics data set, SwissProt and GenBank protein accession numbers were used to manually retrieve corresponding Unigene identifiers. This workflow was successful in merging 69% of their secreted protein list with corresponding data on the microarray platform, as well as other published microarray data sets, to identify platelet-specific proteins that had not been previously described.

The disparity between the multiple databases by use of different gene and protein identifiers creates a situation where currently there is no single bioinformatics tool sufficient to perform data merging across any two different microarray or proteomic platforms in an automated manner. To illustrate the complexity of this task, we attempted to merge three in-house data sets, obtained from our recent studies.
of epidermal growth factor receptor signalling in human mammary epithelial cells. mRNA expression changes were analysed using microarray platforms from Affymetrix (54 630 oligonucleotides) as well as from Nimblegen Systems (38 108 oligonucleotides). Proteomics data in this analysis were obtained by LC–MS analyses [20, 33–35], using the ‘accurate mass and time tag’ approach [18]. To date, these proteomic measurements have resulted in identification of 7348 proteins expressed in human mammary epithelial cells. Each of these technology platforms, while widely used, relies on the use of different sequence databases and identifiers. As shown in Figure 1, Affymetrix provides a combination of GenBank accession numbers, whereas Nimblegen uses primarily RefSeq accession numbers and gene symbols. The mass tag proteomic approach relies on the IPI database for protein sequence identifiers. Thus, we merged mRNA and proteins across the three platforms using the existing identifiers and available retrieval tools. As shown in Figure 1, Affymetrix provides a combination of GenBank accession numbers, whereas Nimblegen uses primarily RefSeq accession numbers and gene symbols. The mass tag proteomic approach relies on the IPI database for protein sequence identifiers. Thus, we merged mRNA and proteins across the three platforms using the existing identifiers and available retrieval tools. We were only able to achieve 40% overlap between the two microarray sets using perl scripts to parse existing identifiers. Additional gene symbols from the Affymetrix accession numbers were retrieved using MatchMiner, but this only marginally increased the overlap between the microarray data sets. In addition, MatchMiner created redundancy in the merged list because all splice variants are represented by a single gene symbol. We then used a cross-reference table from the IPI database to populate our proteomics data set with gene accession numbers, and after again retrieving gene symbols using MatchMiner, we achieved 40% overlap of our proteomics data with both microarray data sets (Figure 3). This left us with approximately 3000 proteins for which we could compare gene and protein expression.

**Figure 1:** Examples of gene and protein sequence identifier annotation used by different microarray and proteomics platforms.

**Figure 2:** Schematic representation of data-merging workflow for microarray (Affymetrix and Nimblegen) and proteomic (LC–MS) data.

**Figure 3:** Venn diagram showing the overlap between sequence identifiers obtained from two microarray data sets and one proteomics data set using existing identifiers provided by the platforms along with automated retrieval of gene symbols using MatchMiner. The percentages indicate the fraction of gene (oligo) or protein sequences that were not cross-referenced. The numbers in the circles do not add up to the total for each platform because of the redundancy of probe-sets for each gene on the microarrays (which is platform-specific).
protein expression patterns across experiments. However, in the end, nearly 60% of the proteins were not matched to sequence identifiers in the microarray platforms.

This example illustrates how merging large heterogenous data in an automated fashion becomes increasingly difficult as the volume of data increases. A major source of this difficulty was attributed to inconsistencies in the gene and protein annotations used among different databases. In the absence of complete standardization of international sequence databases, one solution to this problem is the development of computational tools that automate the data-merging process, using routinely updated annotation tables and merging scripts that are not reliant on human curation.

**INTRODUCING BIOLOGICAL PROCESS ANNOTATION**

Given that high-throughput genomics and proteomics methods identify different parts of a biological system, ultimate interpretation of the results requires representation of the data in the context of pathways or functional processes. Incorporating biological function information provides a more complete picture of the data set, complementing and extending the information provided by the experimental platform. It is important to remember, however, that even the best characterized organisms, functional annotation is usually incomplete and exists for only a fraction of the genes. Electronic databases of biological pathways are often even more limited in coverage or have been curated for a single organism.

The Gene Ontology (GO) database is a controlled vocabulary that describes the roles of genes and proteins in all organisms [36]. GO is comprised of three independent ontologies: biological process, molecular function and cellular component. There are several freely available tools for batch retrieval of GO annotation for a list of genes, such as GoMiner [37], FatiGO [38], EASE [39], FuncAssociate [40] and OntoMiner [41]. Many of these tools contain algorithms that provide statistical analysis of ontology terms representing a gene list and have been used to compare data sets across multiple gene expression platforms [42] and across microarray and proteomics platforms [8].

The Kyoto Encyclopedia of Genes and Genomes (KEGG) contains a comprehensive collection of databases for genes, pathways and ligands for several eukaryotic organisms and over 200 species of bacteria (http://www.genome.ad.jp/kegg/). KEGG maintains web-accessible tools for the retrieval of pathways and the annotation of gene lists. However, KEGG uses unique identifiers distinct from the gene and protein sequence databases discussed above. The Gene Map Annotator and Pathway Profiler (GenMAPP) program is another freely available, stand-alone application for viewing and analysing gene expression data in the context of biological pathways [43]. MAPPFinder is an accessory tool that incorporates GO annotations to identify over-represented GO terms in a data set, which can then be visualized on GenMAPP graphical files [44].

While using functional annotation to describe a data set is attractive, limiting factors include the unreliability of available annotation databases and the wide variability of information provided by such data sources. For example, Hoerndli et al. [45] compared three gene expression data sets describing similar experiments in neuroblastoma cells (Affymetrix), amygdala (SAGE) and whole brain (Affymetrix) using GO, GenMAPP and KEGG. The representative biological pathways that differentiated the data sets across the three databases were inconsistent. For example, when the differentially expressed neuroblastoma genes were compared with differentially expressed genes within the amygdala, the common GenMAPP pathways were electron transport chain and ribosomal proteins, and the common KEGG pathways were oxidative phosphorylation and translation. In contrast, when the neuroblastoma genes were compared with the differentially expressed genes from the brain, common GO biological process terms included cell cycle, phosphate metabolism and DNA metabolism, whereas common KEGG pathways included neurodegenerative disorders, sorting and degradation and proteasome.

There are several factors that may contribute to inconsistent outcomes, such as the fact that the KEGG database has fewer annotated genes than GenMAPP and GO. There is a lack of consistency in vocabulary for functional annotation across databases and across species. Other sources of functional annotation include databases of protein–protein interactions, transcription-factor-binding sites, protein functional motifs, etc., which will provide more insight as these data sources become more complete across multiple species.
In addition, a few commercial resources exist for building signalling networks from data sets, such as MetaCore from GeneGo (www.genego.com), Ingenuity Pathways Knowledgebase (www.ingenuity.com) and PathArt from Jubilant Biosystems (http://jubilantbiosys.com/pd.htm). These tools are based upon proprietary databases of curated metabolic pathways and individually modelled signalling relationships between proteins, genes, complexes, cells, tissues, drugs and diseases, assembled from peer-reviewed journal articles.

COMPARING mRNA AND PROTEIN ABUNDANCE
Studies that have compared quantitative profiles of both gene transcript and protein abundance at a global scale are still quite limited in number (Table 1), yet the general picture emerging from these efforts is that protein abundance patterns cannot be reliably inferred from mRNA levels. In general, two fundamental experimental designs have been employed in these studies: (i) comparison of expression profiles across many genes and proteins within a single cell type or tissue in response to stimulus challenge and (ii) comparison of a set of expression profiles across a range of cell types. No consistent correlation has been reported from these efforts, with results ranging between no statistical association to relatively strong correlation.

Initial studies that have evaluated the concordance between genomic and proteomic data took advantage of yeast model systems, where the relationship between the genetic elements that control cell cycle and metabolism are considerably more defined than in mammalian systems. For instance, Gygi et al. [7] compared steady-state protein levels measured by 2D-PAGE with mRNA levels calculated from SAGE frequency tables and reported a Pearson correlation coefficient of 0.93 for a set of 106 genes. However, this good correlation was significantly biased by a small number of genes with highly abundant mRNA expression levels. When corrected to remove this bias, the correlation was reduced to 0.35. Subsequent studies in Saccharomyces using cDNA microarrays and a quantitative LC–MS/MS proteomics approach identified 30 proteins whose abundance is associated with switching to the galactose utilization pathway [8]. Half of these proteins are regulated post-transcriptionally, with no significant corresponding change in mRNA abundance. Among the total 289 proteins identified in the study by Ideker et al. [8], only a moderate correlation with mRNA abundance was observed (R = 0.61). These results are in general agreement with those of Griffin et al. [9], who compared the relative abundance of 245 unique proteins to mRNA changes, reporting a Spearman rank correlation value of 0.21. Apparent discrepancies between mRNA and protein abundance observed in these studies of yeast metabolism are consistent with the results of a genome-wide epitope tagging study, which found that the abundance of many essential transcription

Table 1: Selected quantitative studies comparing mRNA and protein abundance

<table>
<thead>
<tr>
<th>Author (Ref.)</th>
<th>System</th>
<th>State</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gygi et al. [7]</td>
<td>Yeast</td>
<td>Steady state</td>
<td>$r = 0.935$ including highly abundant proteins; $r = 0.356$ if excluded</td>
</tr>
<tr>
<td>Ideker et al. [8]</td>
<td>Yeast</td>
<td>Perturbed</td>
<td>$r = 0.61$, however concordance was high for metabolic and respiratory pathways</td>
</tr>
<tr>
<td>Chen et al. [47]</td>
<td>Lung adenocarcinoma</td>
<td>Steady state</td>
<td>$r = 0.2445$ for signature genes; $= -0.025$ for all</td>
</tr>
<tr>
<td>Griffin et al. [9]</td>
<td>Yeast</td>
<td>Perturbed</td>
<td>$r = 0.21$ for all, however concordance was high for glycolysis and galactose utilization pathways</td>
</tr>
<tr>
<td>Nishizuka et al. [61]</td>
<td>NCI-60 cell lines</td>
<td>Steady state</td>
<td>$r = 0.52$ for cDNA arrays/protein and $r = 0.40$ for oligo arrays/protein</td>
</tr>
<tr>
<td>Tian et al. [48]</td>
<td>Liver tissue &amp; hematopoietic cell lines</td>
<td>Perturbed and Steady state</td>
<td>$r = 0.54$ (liver-perturbed) and $r = 0.59$ (cell line-steady-state)</td>
</tr>
<tr>
<td>Verhoeckx et al. [66]</td>
<td>U937 cell line</td>
<td>Perturbed</td>
<td>$r = 0.86$ including highly abundant proteins; $r = 0.49$ if excluded</td>
</tr>
<tr>
<td>Cox et al. [62]</td>
<td>Mouse Lung</td>
<td>Steady state</td>
<td>$r = 0.18$ for all data, concordance greater for structural genes</td>
</tr>
</tbody>
</table>
factors in yeast are not easily detected at the mRNA level [46].

Studies integrating mRNA and protein abundance profiles in mammalian systems also suggest that more than half of the discordance observed between mRNA and protein abundance might be attributed to post-transcriptional regulatory mechanisms [47, 48]. For instance, comparison of mRNA and protein levels for 98 genes across 76 neoplastic and normal lung tissues using 2D-PAGE and oligonucleotide microarrays resulted in concordance in expression for only 17% of the genes [47]. This low level of concordance is similar to that observed in a comparative analysis of mRNA and protein abundance conducted in myeloid precursor cells [48]. Among 150 proteins identified as altered using ICAT-based methods, 76% of the genes and proteins showed changes in the same direction (increased or decreased), although many of these changes were not robust enough to be considered statistically significant. When only statistically significant changes in abundance were considered, 79% of the changes occurred at either the mRNA or protein level alone (but not both). The conclusion from this study is that, at most, 40% of the changes in protein abundance can be attributed to differential mRNA expression.

The apparent low level of concordance between mRNA and protein abundance emerging from many studies clearly indicates that analysis at a single hierarchical level of biological regulation does not always provide an adequate descriptor of the biological response. Caution should be used when extrapolating these results globally, since the majority of studies conducted to date integrated relatively small data sets, predominantly using a single-time-point study design. Few published studies provide comparison across two or more time points. While it is expected that the concordance between mRNA and protein abundance improves when temporal delays for post-transcriptional processing are considered, the picture from these studies is still complex, probably explaining less than 50% of the discrepancy observed between mRNA and protein levels [48, 49]. Because of this complexity, approaches that couple computational modelling of the temporal relationships between transcription rates with other cellular processes that dictate steady-state protein abundance will ultimately need to be employed in concert with microarray and proteomic analyses to attain a better understanding of these interrelationships.

Mathematical models of the intrinsic noise in prokaryotic gene regulation indicates stochastic fluctuations and cell–cell variations in gene regulation can have significant impacts on the steady-state relationships between mRNA and protein abundance across time [50,51]. These analytical models suggest the ‘two-step’ process of transcription and translation provides the flexibility needed for an organism to independently regulate the average concentration of a protein across a cell population and the distribution of protein concentrations within the population [51]. Such cell–cell variation is not readily identified in typical microarray and proteomic studies, where the measurements usually represent steady-state averages across a cell population.

**POST-TRANSCRIPTONAL SOURCES OF DISCORDANCE**

Much of the apparent discordance observed between mRNA and protein abundance is certainly indicative of diverse biological regulatory mechanisms and should not be unexpected (Figure 4).

---

**Figure 4:** Diverse post-transcriptional regulatory mechanisms account for much of the apparent discordance between mRNA and protein abundance. The central circle represents where the majority of proteins or mRNA abundances lie in a typical experiment (unchanged). Common post-transcriptional mechanisms of regulation include rapid mRNA degradation and translational control mechanisms (quadrant A), as well as targeted protein degradation and stabilization of mRNA through ARE-binding proteins (quadrant D).
despite the error attributed to a specific measurement platform. There are many known biological mechanisms that describe how mRNA expression and protein abundance are regulated separately in the cell. Beyond rates of mRNA synthesis, processes such as nuclear export, splicing, and mRNA stability are well-described regulatory mechanisms that ultimately influence protein abundance [52, 53]. For instance, regulation of steady-state mRNA levels through factors that bind to 3’ untranslated adenosine uracil (AU)-rich elements (ARE) and retard mRNA degradation are now well-documented [52, 54, 55]. Regulation of mRNA stability by ARE-binding proteins can result in differences in the half-lives of individual mRNAs by orders of magnitude, and as such is an important strategy for controlling the mRNA abundance of many rapidly responsive genes, such as transcriptional regulators and cytokines [56]. Similarly, targeted degradation of mRNA by altering mRNA stability would eventually result in the decrease of protein abundance, dictated by the half-life of the protein. Whether this affected the degree of concordance between mRNA and protein abundance would depend on the sampling times used in the experimental design.

Control of the translational initiation is also emerging as a broad mechanism for regulation of protein abundance [57]. Analytical models indicate that translational efficiency is a major contributor to noise in gene regulatory systems, particularly for highly expressed proteins [50]. The importance of translational control is further emphasized in oocyte development, where the first few rounds of cell division occur in the near absence of transcription, regulated through coordinated control of protein translation and degradation. Moreover, based on analysis of differential recruitment between free ribonucleoprotein particles and ribosome-bound mRNAs, it was estimated that more than 10% of the genes modulated upon antigen activation of T-cells are regulated by translational control [58]. In addition, a study evaluating inhibition of initiation of protein translation using rapamycin [59] suggests that ~7% of cellular mRNAs are still translated even in the context of a general down-regulation of protein synthesis, including a large number of kinases and DNA-binding proteins. Once translated, multiple mechanisms can modulate protein stability. Stabilization of the protein through proper folding into its native conformation often requires the interaction between chaperones and other factors that can vary under different cellular conditions. Furthermore, targeted degradation of proteins, particularly through ubiquitin-dependent proteolysis, is one of the most highly regulated processes in the cell, functioning to tightly and selectively control protein abundance [60].

### Concordance Among Functional Classes

The ultimate goal of integrated analysis of microarray and proteomic studies extends beyond that provided by general correlation analyses. Meaningful trends in the data may be extracted through interpretation of the data in the context of biological pathways. Whether there are generalizable relationships that describe the concordance between mRNA and protein levels for specific functional classes of proteins remains an open question, yet some recent studies suggest this may be the case. For instance, genomic and proteomic profiling of 52 gene–protein pairs across the NCI-60 cancer cell lines revealed that cell-structure-related proteins invariably showed a higher correlation between mRNA and protein levels than did non-structure-related proteins [61]. In part, the greater correlation with structural proteins observed in this study could potentially reflect a higher level of protein abundance for some proteins in the cancer cells (i.e. ERB-B2) rather than a reflection of structural function per se. However, others have noted greater concordance between mRNA and protein for structural genes as well. In a recent example of an analysis of mouse lung development, Cox et al. [62] used a regression-based approach to compare the slopes of protein and mRNA expression as a function of time with a merged set of over 800 mRNA–protein pairs. The calculation of slope (expression ratio versus time) as a temporal discriminator permits binning of genes that show positive or negative concordance between mRNA and protein changes from those that are discordant. Using this approach, genes showing co-regulation at the mRNA and protein level were enriched for structural molecules, whereas negatively co-regulated groups were enriched for products
involved in regulating gene expression, including transcription factors. Broadly speaking, these conclusions are in general agreement with results obtained in yeast, where genes associated with protein synthesis (ribosomal proteins, elongation factors) tended to show discordance between mRNA and protein abundance [9].

**SUMMARY AND FUTURE NEEDS**

The rapidly evolving field of proteomics, combined with microarray and SAGE technologies, is providing an unprecedented opportunity for interrogating the stoichiometric and temporal relationships between changing transcript and protein levels. These emerging technologies have also provided fuel for major advances in the field of computational biology to address both the statistical needs for analysis of global expression profile data, as well as the need to integrate and store metadata sources in a query-compatible environment. Continued improvements in database consistency and annotation, along with advances in bioinformatics tools for querying across databases, will be needed to catalogue the stoichiometric and temporal relationships between mRNA and protein abundance changes. To date, the integration and interpretation of global transcript–protein analyses has been frustrating to perform, and the degree of discordance difficult to interpret at the systems level. The next few years should bring a shift from the focus on steady state to characterization of the dynamics of transcript and protein regulation in perturbed states. These dynamics are essential to understand the relationship between transcript or protein abundance and, ultimately, biological function. Indeed, a pulse-labelling method for genome-wide analysis of mRNA synthesis and decay rates was recently demonstrated [63]. Similarly, the enhanced sensitivity and dynamic range of LC–MS instrumentation, with stable-isotope-labelling strategies employed with time-course experimental designs that permit distinguishing between protein synthesis and degradation rate changes [64, 65]. As data describing the dynamics of the transcriptome and proteome become more accessible, the need for innovative tools that facilitate data integration will become essential for developing the foundational databases necessary for systems-level models of cell and tissue function.

---

**Key Points**

- Mapping the quantitative relationships between thousands of gene and proteins is extraordinarily complex, yet global understanding of these relationships should provide insight into cellular function and dysfunction.
- The concordance between mRNA and protein abundances is dependent, in part, on experimental measurement times and whether the biological system is in steady state or perturbed by a stimulus.
- The ultimate interpretation of genomic and proteomic results requires the representation of the data in the context of biological pathways or functional processes.
- Integration of heterogeneous data from disparate technology platforms represents a significant challenge because of the inconsistencies in gene and protein annotations used among different databases, and because the bioinformatics tools necessary for automated merging of microarray and proteomic data sets are limited.
- Advances in bioinformatics tools, continued improvements in database consistency and annotation, and a shift from steady-state to dynamical measurements at global scales will become essential for developing the foundational databases necessary for systems-level models of cell and tissue function.

---

**Acknowledgements**

Portions of this work were supported through the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL) and the NIH National Center for Research Resources (RR018522). PNNL is operated by Battelle for the US Department of Energy under Contract DE-AC06–76RLO 1830.

**References**


