UPTEC X 05 033
JUN 2005

ISSN 1401-2138

KARIN JOHANSSON

Integrated transcript and protein analysis – a bioinformatics approach.

Master's degree project

UPPSALA
UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 05 033

Date of issue 2005-06

Author

Karin Johansson

Title (English)

Integrated transcript and protein analysis – a bioinformatics approach

Title (Swedish)

Abstract

Gene expression can be measured at both the level of mRNA (transcriptomics) and that of protein (proteomics). Today, very little is known about the relationship between gene expression as determined by mRNA and protein techniques. Understanding the biological (and mathematical) correlation that exists between mRNA and protein is valuable since it may enable prediction of gene expression from one level to the other. The aim of this project was to examine the correlation between mRNA and protein in rat liver, and to categorize *when*, *how*, and *if* transcripts and proteins should be analyzed in an integrated fashion. The conclusion that could be drawn was that the mRNA-to-protein correlation was significantly different between cellular compartments and functional categories. Although it is not yet possible to predict protein expression from only microarray experiments, this study introduces an interesting field of research and provides a benchmark for future studies.

Keywords

Transcriptomics, proteomics, integrated analysis, correlation, microarrays, 2D-PAGE

Supervisor

Magnus L Andersson
AstraZeneca R&D, Mölndal

Scientific reviewer

Christopher Southan
AstraZeneca R&D, Mölndal

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **55**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Integrated transcript and protein analysis - a bioinformatics approach

Karin Johansson

Sammanfattning

Den centrala dogmen, myntad av Francis Crick fastslår att den genetiska informationen överförs från DNA via RNA till protein. Trots att vi idag har stor kännedom om de relaterade processerna transkription och translation, så är kunskapen om sambandet mellan mRNA- och proteinuttryck mycket liten. Hittills har ett fåtal studier gjorts för att undersöka den biologiska och matematiska korrelationen mellan mRNA- och proteinuttryck och flertalet av dessa analyser är gjorda på modellorganismen *Saccharomyces cerevisiae*. I det här examensarbetet undersöks för första gången sambandet mellan mRNA och protein i ett eukaryotiskt system: råttlever.

Målet med examensarbetet är att identifiera organeller och cellulära processer och funktioner för vilka mRNA-protein-korrelationen är signifikant avvikande, samt huruvida det är möjligt att förutspå proteinuttryck utifrån microarrays och liknande transkripttekniker. Baserat på tidigare studier i jäst har fem hypoteser ställts upp, vilka besvaras och undersöks i studien. Vissa kategorier såsom ‘endoplasmatiskt reticulum’, ‘golgiapparaten’ och ‘aminosyrametabolism’ uppvisar en avsevärt bättre korrelation än t.ex. ’mitokondrie’, ’cytosol’, och ’transkription’. Slutsatsen är trots allt att microarrays och andra metoder för att mäta mRNA-uttryck inte kan ersätta proteintekniker eller förutspå proteinuttrycket helt och hållet. Det är dock uppenbart att skillnaden är signifikant mellan olika organeller och cellulära processer, och i framtiden, när bl.a. mekanismerna bakom mRNA och proteindegradering blir kända, kommer troligtvis sambandet mellan mRNA och protein bli lättare att förstå. Den här studien är av explorativ karaktär och syftar till att skapa en diskussion om hur man kan utnyttja transkript- och proteintekniker på bästa sätt samt att skapa en modell för framtida korrelationsanalyser av detta slag.

Examensarbete 20 p, Molekylär Bioteknikprogrammet
Uppsala Universitet juni 2005

1 Introduction

The flow of genetic information in all cells progresses from DNA to messenger RNA to protein. To date, very little is known about the relationship between gene expression as determined by measuring mRNA and protein expression levels. In order to address this issue, researchers have tried to find mathematical correlations between mRNA and protein expression levels. Surprisingly, there is not much published work on this topic and most of the attempts to correlate mRNA and protein expression have been employed in yeast cells (1-11) and human cancers (12-20). The results have been varied, with the majority reporting a minimal or limited correlation. As mRNA is eventually translated into protein, it seems reasonable to assume that there should be some sort of correlation between the level of mRNA and that of protein. However, considering all the post-transcriptional mechanisms that occur after mRNA transcription, it is rather obvious that the correlation is not as simple as a 1:1 relationship. Nevertheless, an interesting thought is if it would be enough, in some cases, to investigate one level in order to predict the response of the other.

In the aspect of correlation, it is crucial to identify what genes are actually detectable with the techniques that exist today. Gene expression might be very hard to detect with both transcript and protein techniques as is the case for G-protein coupled receptors (GPCRs) (21), or gene expression might be detectable for only one of the approaches due to detection limitation of the technique. Several of the techniques used today are far from complete, and the development of better ones, especially for proteomics, will improve the comparisons between mRNA and protein.

The aim of this project is to investigate whether there is a significant correlation between mRNA and protein in rat liver and to categorize *when*, *how* and *if* transcripts and proteins should be analyzed in an integrated fashion. This explorative analysis should help identifying gene clusters, specific pathways or cell compartments where microarray experiments may be sufficient to predict protein expression, as opposed to those where post-transcriptional regulation has to be taken into account. In this sense, the outcome of this project can be used as a first step towards a decision support for technology choices for future disease area projects.

Based on a comprehensive literature study on what has been done previously in this area of research, a set of hypotheses regarding correlation for specific ontological categories was generated and assessed. The approach used is of bioinformatics character, merging several protein and transcript databases into a common database management system allowing for extraction of data suitable for the stated hypotheses. This study, for the first time, explores the qualitative comparison of mRNA and protein expression for a large number of genes expressed in rat liver tissue. The project was made in collaboration with AstraZeneca R&D, Mölndal. It should be noted that this field of research is continuously being updated, and articles published after January 2005 have not been included.

Before continuing, the outline of the report is introduced: *Chapter 2* presents the theoretical background and the detailed aims of this project; the mRNA-to-protein correlation approach is explained in *Chapter 3,* the hypotheses generated from literature sources are presented and investigated in *Chapter 4*, and discussed in *Chapter 5.* Eventually *Chapter 6* elaborates on future directions for these kinds of studies.

2 Background

This chapter explains the background to transcriptomics and proteomics as well as the efforts made to integrate these. Relevant transcript and protein techniques are explained and a clarification of the concept of correlation, as well as limitations to correlation analysis is given. The detailed aims of the project are thereafter elucidated.

2.1 'Omics' approaches – a new era in biology

Completion of the major milestones of the Human Genome Project (22,23) has demonstrated the impact of the "omics" revolution on modern research in life sciences. Several different terms have been coined using the suffixes '-ome' and '-omics' and a comprehensive glossary lists over 50 of them (24). The majority of the omics are supported by high-throughput technologies that generate massive amounts of information. The techniques used in molecular biological research and drug discovery have changed dramatically over the past ten years and we are now entering an attractive but complex era of data overload (25). In order to deal with this mass of information, powerful tools, as well as common "glue" is needed. Bioinformatics serves as this common glue and is an integral part of transcriptomic and proteomic research. It plays a major role in data storage, management and analysis, and can be used in drug target identification and validation (26).

Functional genomics, broadly defined as the comprehensive analysis of genes and their products, involves a comparison of gene expression between different genetic or physiological states (27). As illustrated in Figure 1, four levels of analyses are usually exploited in functional genomics: genomics, transcriptomics, proteomics, and metabolomics. The top level, genomics, first introduced by Tom Roderick at the Jackson Laboratory in 1986 (28) is defined as the discipline involved in the measurement of "the population of open reading frames" (29). Metabolomics, the bottom level, is defined as "the measurement of metabolite concentrations and fluxes in isolated cell systems or cell complexes" (30). These omics approaches will not be covered in this project; however, the fields of transcriptomics and proteomics require a thorough presentation.

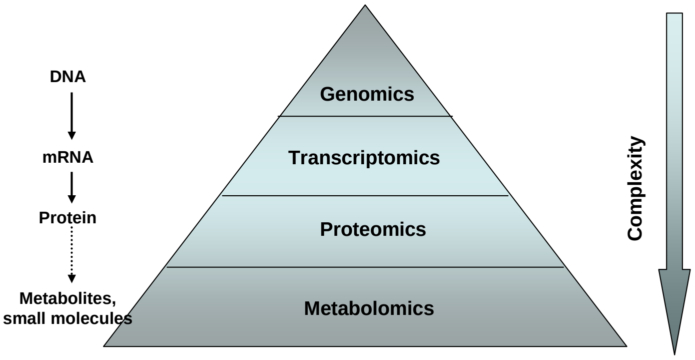


Figure 1: Levels of analyses exploited in functional genomics
Genomics refers to the measurement of the population of open reading frames. *Transcriptomics* refers to the techniques for identifying the complete set of mRNAs that are transcribed from the genome (31). *Proteomics* is defined as the investigations for identifying the expressed set of proteins that are encoded by the genome (32), and *metabolomics* is described as the measurement of metabolite concentrations and fluxes in isolated cell systems or cell complexes (30). The last three levels are all context-dependent and change with the physiological, developmental, or pathological state of living cells (6).

2.1.1 Transcriptomics

The total complement of mRNA in a cell or tissue at any given moment constitutes its transcriptome (33). Transcriptomics, originally coined by *Velculescu et al* in 1997 (31), was one of the first omics approaches to be developed (34). With recent advances including the development of differential display-PCR (35), cDNA microarray and DNA chip technology (36,37), and of serial analysis of gene expression (SAGE) (11,38), the relative abundance of transcripts can be monitored simultaneously for thousands of genes under various experimental conditions. Transcriptomics has a distinct advantage in high-throughput and moderate cost, but is not routinely set up to systematically detect changes in splice variants (39). Whereas genomics is a static repository of information, the transcriptome is dynamic, continuously changing and responding to external stimuli. Transcriptomics can generate useful information for a variety of applications, such as target validation, disease classification, and functional annotation. A full review of the transcriptomic techniques is beyond the scope of this project, however since the data used in this analysis is generated by the GeneChip® technology from Affymetrix Inc., this microarray technique needs a clarification.

2.1.1.1 The Affymetrix technology

Microarrays record the expression levels of thousands of genes in parallel (40). The application of microarrays is so broad that several variants of the technology have been developed, in general falling into two categories: high density oligonucleotide arrays (HDAs) (41) and cDNA probe arrays (42). Currently, HDAs are principally supplied by Affymetrix and most HDA-related algorithms have been developed for the company's GeneChip® products. The Affymetrix system uses millions of oligonucleotide probes, each designed to hybridize to a particular part of a transcript. Chips exist for a variety of organisms including human, mouse, yeast, Arabidopsis, and *C. elegans.* The chipset used in this analysis, the rat genome U34 set, consists of three GeneChip® arrays and monitors the expression of more than 24,000 mRNA transcripts and EST clusters from the UniGene database (43).

The chips are designed so that every transcript is represented by between 11 to 20 probes that match different parts of the 3' end of the mRNA sequence. Every probe consists of a pair of 25 residue oligonucleotides, one is a perfect match to the transcript and the other a mismatch in which the middle residue has been changed. This probe-pairing strategy helps minimize the effects of non-specific hybridization and background signal (44). Fragmented RNA is labeled with a fluorescent tag and presented to the microarray. Wherever there is a complementary probe sequence on the chip, the RNA can hybridise to it. When the array is scanned by a laser, the tagged fragments glow, producing spots with a brightness proportional to the amount of RNA that has hybridized. This way, mRNA levels for different genes can be estimated.

2.1.2 Proteomics

The word proteomics, PROTEin complement expressed by a genOME, was coined by Wilkins et al in 1994 (45). Proteome analysis is attractive because of its potential to determine biological properties that are not apparent by DNA or mRNA sequence analysis alone. Such properties include the quantity of protein expression, the subcellular location, the state of modification, and the association with ligands. Many biology textbooks explain that proteins embody the active life of cells, while nucleic acids represent only plans. Ultimately, proteins carry out most of the biological functions encoded by genes and therefore, study of the proteome is needed in order to resolve fundamental biological questions. Leigh Anderson explains the concept in a metaphorical way (46):

“There is more to paella than the recipe, more to Bach than the ink of paper, and more to a society than its codes of laws”

In each case, the implementation of a series of instructions is far more complex and interesting than one would expect from simply reading them.

The proteome is a dynamic entity thereby possessing an enormous complexity. Protein molecules have greater individual chemical variation than nucleic acids and post-translational mechanisms such as glycosylation, phoshorylation, and protein cleavage greatly influence this complexity. Hence, protein techniques are less suitable for high-throughput analyses than genomic and transcriptomic techniques (47). At present, two techniques are central for proteomic analyses: (1) two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), which can separate thousands of proteins in a few steps, and (2) mass spectrometry (MS) for the identification of proteins and their post-translational modifications. Several other protein techniques exist, e.g. 2D differential in-gel electrophoresis (2D-DIGE) (48,49), isotope-coded affinity-tag-based protein profiling (ICAT) (50), and multidimensional protein identification technology (MudPit) (51). The protein data used in this analysis comes from 2D-PAGE and mass spectrometry and these techniques will consequently be evaluated in more detail.

2.1.2.1 2D Gel electrophoresis

Since its introduction in 1975 by O´Farell and Klose (52,53), 2D-PAGE has been established as the dominant technique for proteomic analysis. This technique separates proteins in two dimensions; the first dimension according to the isoelectric point using a pH gradient and the second dimension according to the molecular mass (32). Although 2D-gel electrophoresis is promoted as a tool to detect the total cellular proteome, in actuality, it fails on several criteria. A high level of expertise is needed to obtain reproducible gels, and two-dimensional electrophoresis is generally limited to proteins that are neither too acidic, too basic, nor too hydrophobic, and that are between 10 and 200 kDa in size. Furthermore, this approach detects only those proteins that are expressed at relatively high levels and that have long half-lives (54). The technique is also limited by a problem called coverage (5). Since there are thousands of proteins present on each gel, many proteins are resolved, but are close together. Weak spots are therefore hard to visualize and can only be seen when they are well separated from strong spots.

2.1.2.2 Mass spectrometry

The mass spectrometry (MS) technique has developed strongly over the past decade (55). MS for protein identification relies on the digestion of protein samples into peptides by a sequence-specific protease such as trypsin. After the proteins are digested, the peptides are usually delivered to a mass spectrometer for analysis via chromotographic separation coupled online to electrospray ionization (LCMS for liquid chromatography mass spectrometry). Matrix-assisted laser desorption/ionization (MALDI) (56) is an alternative ionization method, but it is less frequently applied to the proteomic analysis of complex protein mixtures. The mass spectrometer measures all the peptides eluting at any given time from a chromatography column, and then selects a number of peptide ions for fragmentation. This is achieved by only allowing ions of a particular mass through a “collision cell”, where the peptide ions’ kinetic energy is increased and they collide with inert gas molecules with sufficient energy to break peptide bonds. The resulting spectrum is called a tandem or “MS/MS” spectrum and contains several adjacent fragments that spell out a partial sequence of a particular peptide. These fragments are then compared to fragments calculated from all peptides in protein databases to arrive at protein identification (57).

2.3 Transcript and protein repositories

Much of our understanding of global gene expression patterns comes from data generated from microarray experiments and the fact that researchers have deposited their raw data into the public domain. Today, mRNA expression data is easily available and several transcript databases that can be publicly accessed exist. An example is the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI), which serves as a public repository for high throughput molecular abundance experimental data, especially gene expression data (58). Since fall 2004, GEO holds over 30,000 submissions representing approximately half a billion individual molecular abundance measurements for over 100 organisms, submitted by over 600 researchers. (59). The easy availability of mRNA expression data has led to countless computational analyses of the same data sets.

Unfortunately, the availability of protein data lags far behind that of transcript data and there is no centralized proteomics database. This is certainly a result of poor distribution of raw data to the community (60). Besides the fact that protein techniques face technical challenges, researchers have done a poor job depositing their protein expression data into the public domain. Today, several 2D databases exist (61-68) but since this technique only detects highly expressed proteins, the construction of complete proteome maps is difficult, irrespective of type of organism. Moreover, it is not easy to compare proteomics results between laboratories because the technologies are and will not be standardized for quite some time. Efforts are nevertheless underway to define common reporting categories and forms (69,70). An example is the proteomics standards initiative, PSI, established as a working group of the HUman Proteome Organization (HUPO), which aims to define community standards for data representation in proteomics and to facilitate data comparison, exchange, and certification (69). Today, the best repositories of protein expression information are public databases such as SwissProt and TrEMBL, but even though they contribute to open availability of protein data, they are far from complete.

2.4 Correlating transcript and protein data

To date, there have been only a handful of efforts to estimate correlations between mRNA and protein expression levels. Most of these have been conducted in yeast using microarray and two-dimensional electrophoresis techniques. The results have been varied but in most cases, the correlation has been reported as minimal to moderate (2-4,7-9,15,17-18,71-77). For a variety of reasons, including saving time and money, it would be useful to know the extent to which mRNA expression is predictive of the corresponding protein expression. Introducing the concept of correlation, it is important to discriminate different *types* of correlation. Three different correlation types can be distinguished: (1) Quantitative vs. qualitative, (2) perturbed vs. steady-state, and (3) global vs. local correlation analysis.

2.4.1 Quantitative vs. qualitative correlation analysis

A qualitative correlation analysis refers to the level of congruence of up- or down-regulated genes in transcript and protein experiments, regardless of the abundance of mRNA or protein expression. Quantitative correlation analysis, in contrast, provides a reliable way to quantitate the amount of a given mRNA or protein in a sample even with low levels of gene expression. Many of the attempts to correlate mRNA and protein expression in yeast have been of a quantitative character, comparing mRNA and protein abundance levels (2-5,8-9,11-14,18-19,71-72,76,78-83). *Gygi et al* (1999) and *Futcher et al* (1999) were among the first to approach this area of research. They both investigated the relationship between mRNA and protein expression levels for a large number of expressed genes in cells representing the same state in *Saccharomyces Cerevisiae*. Even though they performed the same type of investigation with similar data, they reached markedly different conclusions.

Whereas *Futcher et al* stated that there is a good correlation between protein abundance and mRNA expression (5), *Gygi et al* concluded that transcript levels provide little predictive value with respect to the extent of protein expression (3). The lack of congruence between these studies is partly due to dissimilar viewpoints; while *Gygi et al* focus on the fact that the correlation of mRNA with protein abundance is far from perfect, *Futcher et al* propose that, considering the wide range of mRNA and protein abundance and the presence of other mechanisms affecting protein abundance, the correlation is quite good. In addition, they used different 2D measurement techniques and methods of statistical analysis. Two statistical methods are normally employed in correlation analyses: The Pearson product-moment correlation coefficient, r_p, and the Spearman rank correlation coefficient, r_s. These can both be calculated to discover the strength of a link between two sets of data (84). While *Gygi* used the Pearson coefficient, which requires that the variables assessed are normally distributed; *Futcher* used the non-parametric Spearman correlation coefficient.

A different research group, *Hood et al*, have extended their studies on this issue in yeast (2,4) to include other organisms such as halobacterium (75), and mouse (72). Their results support Gygi's theory that mRNA measurements are not sufficient to predict protein expression. Several other research groups (3,17-18,71,73-74) have reached the same conclusion, which validates a central idea to the systems approach to biology; that it is only through the integration of different levels of information that the system can be described comprehensively (2). In contrast, several researchers take a positive standpoint towards the concept of correlation (5,19-20,78,85-86).

Several qualitative mRNA-to-protein correlation analyses have been carried out (1,15,20,77,85-86) with the majority applied to cancer research. The relationship between gene expression measured at the mRNA level and the corresponding protein level is not well characterized in human cancer. Still, several attempts have been made to correlate mRNA and protein expression. Studies of individual genes in solid tumors have revealed a good correlation between mRNA and protein levels (87,88). On the other hand, another investigation in lung adenocarcinomas demonstrated a significant correlation in only a small subset of the genes studied (18). Conversely, *Orntoft et al* found a significant correlation in bladder cancer (19). Even though they could only compare the levels of about 40 well resolved and focused abundant proteins, it was clear that in most cases there was a good correlation between transcript and protein levels. *Kern et al* (20) observed a very good qualitative correlation between proteomic and transcriptomic methods in the diagnosis of acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). For the 39 genes examined, about 70% of the comparisons were congruent at the level of transcript and protein. Another positive qualitative correlation analysis made on human platelets (86) revealed that 69% of the secreted proteins were detectable at the mRNA level, which indicates that the transcriptome in platelets mirrors the profile of protein expression.

2.4.2 Perturbed vs steady-state correlation analysis

In physics, complex systems are frequently analyzed by what are called perturbation methods. A system, whose internal state we cannot adequately model a priori, is subjected to small perturbations in one or more input variables, and the effects on output variables are examined (89). An important issue is the extent to which the changing expression patterns of mRNAs reflect corresponding changes in their cognate proteins. *Hood et al* were the first to perform such an integrated analysis by exploring the process of galactose utilization in yeast (4). By applying perturbations to the galactose pathway, they examined the global changes in mRNA expression resulting from each perturbation. The resulting correlation coefficient was weak. In 2004, the same research group used a mouse model to demonstrate how well individual mRNA and protein levels correlate in response to multiple drug treatment over time (72). Mice were treated with three PPAR α and γ agonists daily for 1,2,3 and 7 doses. Out of 12 identified candidate genes, there was a good correlation for 8 of them. Several other perturbation-induced correlation analyses have been carried out (1-2,4,13,72,79-80,82-83,85,90) and the results of these have been varied.

In contrast, steady state mRNA-to-protein correlation analyses have also been carried out (3,5,8-9,75-76,81,86). One of the earliest steady-state correlation analyses in human looked at 19 liver proteins (76). *Anderson and Seilhamer* found a somewhat positive correlation with a Pearson coefficient of $r_p = 0.48$. This result, halfway between a perfect correlation and no correlation, has occasioned considerable comment, ranging from consternation that it is so low, to amazement that it is so high. In any way, it does provide a reasonable benchmark.

2.4.3 Global vs. local correlation analysis

The majority of the correlation analyses made on a global level (2-5,19,71-72,76,81,86,91), entailing an all-against-all mRNA and protein comparison have reported a minimal correlation. This is most likely due to a high level of noise in mRNA and protein data (6). In order to get around this noise, several researchers have looked at correlation for smaller, but broad categories, such as function, localization or structure. In 2003, *Greenbaum et al* extended the yeast studies made by *Gygi* and *Futcher*, and constructed a new merged dataset from their yeast data combined with other transcript and protein datasets, resulting in two publications on this topic (6,7). Using the resulting data, they compared mRNA expression and protein abundance globally as well as locally in terms of categories related to compartments and functional modules. They observed a better correlation for cytoplasmic and nucleolar proteins than for nuclear and mitochondrial proteins. Several local mRNA-to-protein correlation analyses have been carried out (6-12), but the majority are made by *Greenbaum*, *Gerstein* and *Jansen*. One common conclusion is that the mitochondrion is the compartment in which the correlation is the poorest (2,6-8,72). In addition, several processes associated with mitochondria also show a poor correlation (2,4,8,72). Genes belonging to the category metabolism demonstrate a good correlation (6-8) whereas modules involved in cell defense, cellular communication, and response to stimulus have a quite poor correlation (8). Further division into a more local level, correlation analysis for individual genes, has resulted in a weak correlation. In 1996, *Tew et al* compared protein and mRNA abundances for one gene product across 60 human cell lines and the resulting correlation coefficient was below 0.5 (92).

2.5 Control of gene expression - limitation to correlation analysis

Besides the technique limitations, the protein repository problem and the significant amount of error and noise in both protein and mRNA experiments mentioned in previous sections, there are several biological explanations to the lack of congruence between mRNA and protein expression. These explanations all fall under the category of gene expression control.

Expression of a gene can be controlled at many levels including transcription, mRNA splicing, mRNA stability, translation, and post-translational events such as protein stability and modification (Figure 2). There are a number of complex steps between transcription and translation, and to date these are not well characterized. All of the regulatory mechanisms that control gene expression following transcription initiation are referred to as post-transcriptional control. An important post-transcriptional control is exerted on mRNA stability, which varies considerably from one mRNA species to another. At steady state, it is the balance between the two opposing processes of the rate of synthesis and rate of degradation that determines the concentration of any mRNA (93,94). The same is true for proteins. Protein turnover is often a missing dimension in proteomic experiments (27), but should be considered when measuring protein abundance data and comparing it to the relative abundance of cognate mRNA species. In order to fully interpret abundance data from proteomic and transcriptomic experiments, it is vital to understand the contributions made by these opposing processes.

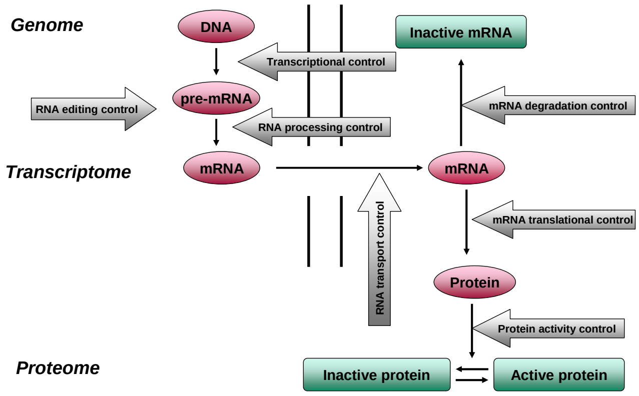


Figure 2: Control mechanisms in gene expression
A gene (DNA) is transcribed to pre-mRNA that may be edited and then processed to one mRNA or by alternative splicing to several forms of mRNAs. mRNA is transported out of the nucleus to the cytosol, where it is either degraded or translated into protein. Protein activity is controlled and proteins may be synthesized as inactive forms that are later reversibly or irreversibly activated or, alternatively, they may be synthesized as active proteins that are later inactivated (95).

2.6 Aim of the study

The aim of this study is to evaluate the *qualitative* correlation between mRNA and proteins in rat liver at *steady-state*. With a bioinformatics approach merging several databases into a common database framework, the goal is to investigate the validity of the hypotheses stated from what has been concluded by other researchers on this topic. The correlation analysis is therefore of a *local* character, looking at the correlation in different compartments, processes, and functional modules. By evaluating the concordance of mRNA and protein data, the ambition is to:

- *Categorize when, how and if transcripts and proteins should be analyzed in an integrated fashion*

- *Identify gene clusters, specific pathways or cell compartments where microarray experiments may be sufficient to predict protein expression, as opposed to those where post-transcriptional regulation has to be taken into account*

- *Offer a guideline for further comparative studies*

- *Provide a platform for the continued expansion of the rat liver proteome and transcriptome, and the incorporation of that information into clinical studies*

3 Materials and methods

This section presents the approach used to compare transcript and protein expression data in rat liver. The reasoning behind the selection of organism and tissue is discussed, and the collection, organization, and visualization of data from the databases used are described.

3.1 Selection of organism and tissue

Rat, *Rattus Norvegicus*, is used extensively as a model organism for studying normal and disease processes in human. It is the most genetically amenable mammalian system in biomedical research, and today, the knowledge of rat physiological mechanisms is extensive. Once genes are identified in rats, pathophysiological mechanisms can be revealed leading to clues to the identification of human genetic counterparts. The liver is the largest and one of the most important organs in the body. It is the body's primary detoxification center for ingested agents and the main site at which nutrients are processed for use by other cells of the body. The cells of the liver, the hepatocytes, are responsible for the synthesis, degradation, and storage of a vast number of substances; they play a central part in the carbohydrate and lipid metabolism of the body; and they secrete most of the protein found in blood plasma.(2) Today, several 2D-PAGE databases of rat liver (61) or its subcellular fractions such as mitochondria (62,96), Golgi complex (97,98), and nuclear pore complex (99) have been established.

3.2 Data collection

As mentioned in the previous section, the availability of protein expression data currently lags far behind microarray data. While mRNA expression data can be easily accessed in structured transcript databases, the field of proteomics has been slow to embrace open availability of raw data. Nevertheless, protein and transcript data from rat liver were collected from available published data sources as well as major public and AstraZeneca in-house databases.

3.2.1 Transcriptomics – Affymetrix data from Gene Logic

Gene Logic is a provider of bioinformatic databases and software tools (100). The company's GeneExpress® system is a gene expression database, where the majority of the expression data is from the Affymetrix technology. Samples from different organisms are collected from experimental studies and tissue biopsies of both normal and diseased tissue as well as tissue exposed to toxic compounds. In GeneExpress, data on more than 10 billion expression data points from over 20,000 clinically significant tissues are stored (100). Besides expression data, clinical and experimental information related to each sample are associated with each gene. Clinical data include basic donor information such as age and gender, family history, medical information, and social history such as diet information and alcohol consumption.

mRNA expression data from the Affymetrix chipset RG_U34 was extracted for rat liver transcripts from (1) normal liver tissue and (2) normal hepatocytes. Selected data attributes such as 'fragment ID' (chip ID), 'fragment name' (probeset), 'fragment warning', 'gene symbol', 'percent present', 'median and mean expression level', 'standard deviation', and 'sample count' were imported to Microsoft Excel. Percent presence refers to the proportion of samples in which a probeset was called 'present' by the Affymetrix software, and gives an indication of what transcripts are easy to detect. All the probesets which had a fragment warning, meaning that the probeset might be of low quality, were removed and the remaining data tables were imported into Microsoft Access 2000 (see Section 3.5).

3.2.2 Proteomics – published protein data vs Swissprot/TrEMBL

The data source of rat liver proteins were divided into two parts: (1) published proteomic data from analyses made by *Fountoulakis et al* (61,62) and (2) Swissprot/TrEMBL (101). The first dataset contains proteins expressed in normal (untreated) rat liver at steady-state, whereas the latter contains all proteins that have *ever* been detected in this specific tissue and organism regardless of steady-state or perturbed condition.

3.2.2.1 Published proteomics data

The biomedical literature search service PubMed and AstraZeneca's in-house service Glides were used to look for literature on rat liver proteome data. Protein data was extracted from analyses made by a Swiss research group, *Fountoulakis et al*, who have made several proteomic analyses on rat and mouse. *Fountoulakis et al* have previously studied the changes in the levels of liver proteins of mice treated with acetaminophen (64), of rats treated with carbon tetrachloride (90), as well as changes in brain proteins of rats treated with the neurotoxin kainic acid (102,103). Furthermore, they have constructed two-dimensional databases on mouse liver (63, 64) and rat brain proteins (65). Here however, the analyses of interest were the ones made in 2002 on normal, untreated rat liver at steady-state (61, 62).

Fountoulakis et al used a subcellular proteomics approach in their analysis of proteins expressed in rat liver tissue. In order to increase the chance of detecting low-copy number proteins; mitochondrial, cytosolic and total protein fractions were prepared prior to 2D gel electrophoresis and mass spectrometry analysis. Some cross-contamination due to incomplete separation of the organelles is hard to avoid with this approach (104), and therefore some cytosolic proteins were present in the mitochondrial fraction and vice-versa. 273 proteins were identified in the total and cytosolic fractions and 192 in the mitochondrial fraction. Most of these were enzymes with a broad spectrum of catalytic activities. A lot of the proteins were common to both fractions and the total amount of unique rat liver proteins was 327.

3.2.2.1.1 Protein frequency – an indication of protein abundance and easily detected proteins

During the construction of the 2D rat liver protein database, the observation was that not all species of a protein mixture were detected equally often (61,62). The often detected proteins were the major, hydrophilic and easy to solubilize components, with average pI and molecular weight values. Furthermore, they were the easily digested proteins that delivered a sufficient number of peptides, so that an identity easily could be assigned. *Fountoulakis et al* evaluated the frequency of detection in all the rat protein samples analyzed in the laboratory in order to get an idea of which proteins are hard vs. easy to detect. 110 protein samples were analyzed for the mitochondrial fractions and 60 for the total and cytosolic fractions. Thus, the frequency of detection is simultaneously an indication of how successful the identification process was for the particular protein, and points towards what proteins are considered to be high vs. low abundant. The most frequently detected proteins were heat shock proteins, in particular heat shock cognate and glucose-regulated proteins. These proteins are also the most frequently detected in other proteomes, for example mouse liver (63), human and rat brain (65). These proteins can therefore be used as markers of a successful identification process during a protein batch analysis.

3.2.2.2 Swissprot/TrEMBL

Information about the rat liver proteins in the Swissprot/TrEMBL databases was extracted using the Sequence Retrieval System (SRS). SRS is one of the most popular molecular biology database query systems and a powerful tool to manage, retrieve, and analyse protein data (101). SRS can be either publicly accessed via http://srs.ebi.ac.uk or via E-lab, the AstraZeneca portal to a huge range of bioinformatics tools, databases, and resources. In order to facilitate mapping of the proteins to Gene Catalogue, another database in the E-lab bioinformatics tool repository, the proteins were extracted were from SRS in E-lab. Searching for organism: rat, and tissue: liver, yielded 684 entries (October 2004).

3.3 Linking transcriptomics and proteomics via Gene Catalogue

The first step towards linking protein and transcript data and performing a correlation analysis of rat liver genes was to compare the transcript and protein datasets by reference to a common database, namely Gene Catalogue.

3.3.1 Gene Catalogue

Gene Catalogue is an AstraZeneca in-house repository of annotated human, mouse and rat gene, transcript and protein sequences. It is reduced-redundant and collapses duplicated sequences from a number of in-house and public databases (105). This is beneficial since all relevant information about a specific gene can be viewed without having to search several databases for the information. The Rat Gene Catalogue stores annotated rat transcripts and protein sequences and allows for the search by gene (official symbol, synonyms or gene names), ID (Gene Catalogue or Affymetrix ID, Accession nr), text, protein domain, pathway, classification or location (chromosome loci).

The way Gene Catalogue is built up is quite complex and will not be explained in detail, however Figure 3 gives an overview of the mapping pipeline. RNA from RefSeqN, EMBL, KEGG (GenesN), Incyte Full Length, GeneSeqN, and AZSeqN is selected in a species centric manner for input into Gene Catalogue. Since Gene Catalogue is designed to only show high quality data, many low quality predicted RNAs are omitted, and only RNAs which are over 100 nucleotides long and are not 'predicted' in nature are included. The input RNAs are mapped to the genome and clustered based on a common splice site to form Gene Catalogue clusters. A single entry in Gene Catalogue is represented by a cluster of sequences that map to the same region on the genome. The clusters generally consist of transcript sequences and corresponding protein variants.

RNA/genome mapping can result in different outcomes and there are five types of clusters within Gene Catalogue: (1) primary clusters, (2) secondary clusters, (3) predicted clusters, (4) freespace clusters, and (5) deleted clusters. The best mapping is labelled a 'primary' mapping and maps to the genome at more than 95% identity. Secondary clusters contain the sub-optimal RNA mappings whereas deleted clusters are clusters that existed in previous releases of Gene Catalogue but have been removed subsequently. Freespace clusters contain all the RNAs that fail to map to the genome and predicted clusters were introduced in GC release 6.0 (October 2004). These clusters are derived from Ensembl gene predictions and they represent Ensembl core genes that are not present in the Gene Catalogue system. When a search is performed in Gene Catalogue, the information for a primary cluster is arranged under six main headings:

- *Gene summary (the gene name(s) for the cluster is displayed in the title)*
- *Transcript Variants*
- *Protein Variants*
- *Literature Mining*
- *Quick Links (rapid access to key underlying databases)*
- *Sequence Tools (access of GC transcripts, proteins or SNPs in SRS).*

Gene Catalogue displays sequence relationships and mappings in a number of viewers where the information in each viewer depends on the sequence type. For example, the GeneViewer allows you to see how a gene maps to the genome, the intron/exon boundaries, splice variants, and protein products of the gene. Affymetrix probe sets have been mapped to the transcripts for each gene and the Gene Logic expression profiles for every gene can be accessed.

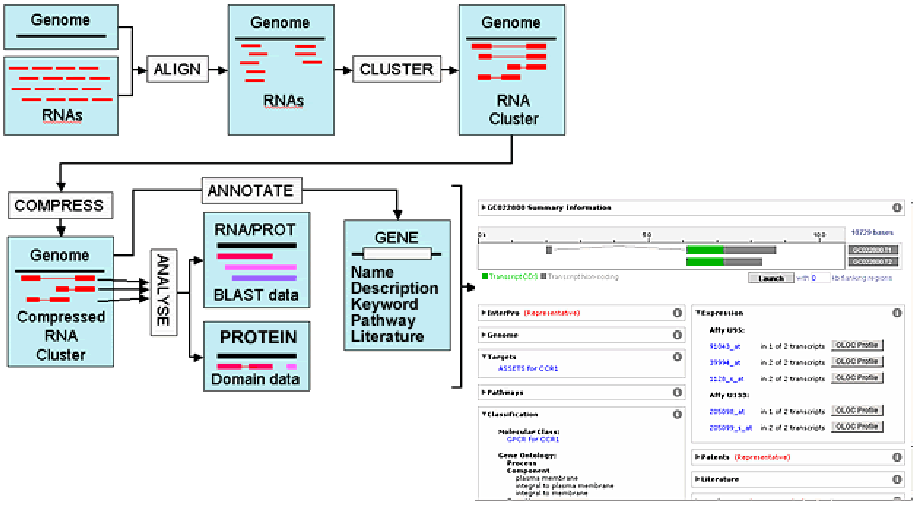


Figure 3: The Gene Catalogue mapping pipeline
RNA from RefSeqN, EMBL, KEGG (GenesN), Incyte Full Length, GeneSeqN, and AZSeqN is selected for input into Gene Catalogue. A Gene Catalogue entry is represented by a cluster of sequences that map to the same region on the genome. Sequence relationships and mappings are displayed in a number of viewers (101).

3.3.2 Connecting transcript data to protein data

To allow comparisons between mRNA and protein expression data, probesets were linked to GeneCatalogue clusters (Figure 4). Protein accession numbers from the Swissprot/TrEMBL and *Fountoulakis* protein datasets were cross-referenced with Gene Catalogue and assigned Gene Catalogue ID:s. One problem with the Gene Catalogue algorithm is that one mRNA sequence might match more than one probeset, and several probesets may define one gene. In order to find the one best probeset per gene, the criteria “highest % presence in exon or 3’UTR” was stated. This means that the matching score should be as high as possible and the probeset should preferentially be in a (1) coding exon or (2) 3’UTR. Probesets binding at the furthest 3’end of the transcript are sorted first because of a more efficient cDNA synthesis at this site and thereby a better quality of the transcript. After mapping protein accession numbers to Gene Catalogue, only 435 of the Swissprot proteins and 128 of the *Fountoulakis* proteins could be linked to a probeset.

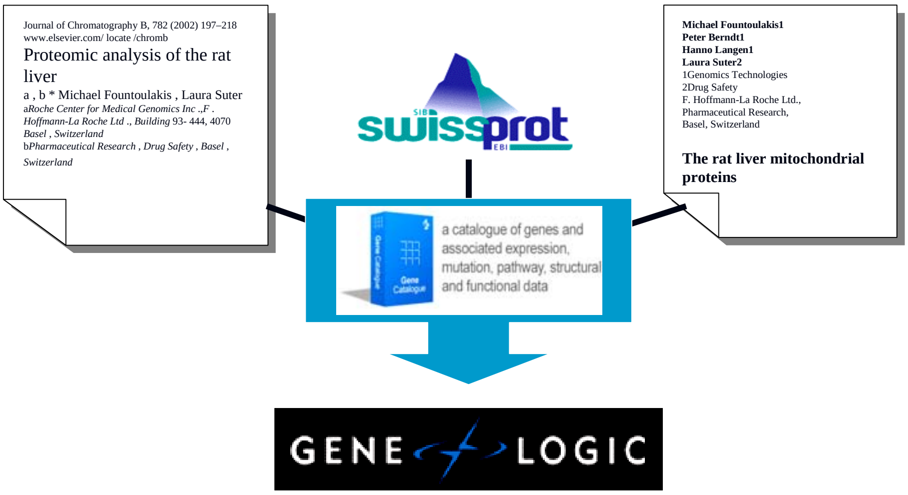

Journal of Chromatography B, 782 (2002) 197–218
www.elsevier.com/ locate /chromb
Proteomic analysis of the rat liver
a , b * Michael Fountoulakis , Laura Suter
aRoche Center for Medical Genomics Inc .,F . Hoffmann-La Roche Ltd ., Building 93- 444, 4070 Basel , Switzerland
bPharmaceutical Research , Drug Safety , Basel , Switzerland
SIB swissprot EBI
Michael Fountoulakis1
Peter Berndt1
Hanno Langen1
Laura Suter2
1Genomics Technologies
2Drug Safety
F. Hoffmann-La Roche Ltd., Pharmaceutical Research, Basel, Switzerland
The rat liver mitochondrial proteins
Gene Catalogue
a catalogue of genes and associated expression, mutation, pathway, structural and functional data
GENE LOGIC

Figure 4: Linking proteomics to transcriptomics
Transcript and protein data are compared by reference to a common database, GeneCatalogue. GeneCatalogue clusters information from multiple genomic and transcriptional sequence databases to give a single best sequence for each gene, which can be compared with diverse data formats, such as Affymetrix probesets. The illustration was made with permission from E-lab, AstraZeneca R&D.

3.4 Gene ontology – linking probesets to biological contexts

Functional assignments of the genes expressed in rat liver were obtained by the Gene Ontology (GO) hierarchy. The GO Consortium (www.geneontology.org) is composed of all the major genome projects and is a collaborative effort to address the need for uniform descriptions of gene products across different databases. Gene products are described in terms of their associated *molecular functions*, *biological processes,* and *cellular components*. The use of GO terms by several collaborating databases facilitates consistent queries across them and the controlled terms are structured so that they can be queried at different levels. The Gene Ontology tool AmiGO can be used to search for a GO term and view all gene products annotated to it, or search for a gene product and view all its associations. All the probesets were coupled to Gene Ontology codes, which allowed the determination of correlation between mRNA and protein for a GO category of interest.

3.5 Data organization

In order to manage the large amount of data and allow for a correlation analysis across GO categories suitable for the stated hypotheses, a relational database was created. This database consists of all collected protein-, transcript-, and probeset related data.

3.5.1 Construction of a relational database in Microsoft Access

All levels of probeset-related information were linked and organized in Microsoft Access 2000. Microsoft Access is a database management system that helps managing data that is stored in computer databases (106). Microsoft Access provides a good basis for the database technique called data mining. David Hand defines the concept of data mining in the journal The American Statistician, 1998 (107) as:

"*A new discipline lying at the interface of statistics, data base technology, pattern recognition, and machine learning, and concerned with secondary analysis of large data bases in order to find previously unsuspected relationships, which are of interest of value to their owners."*

By importing relevant information from Gene Logic, Gene Catalogue, and the two protein datasets into tables in Access, the relationship between transcripts and proteins in particular ontology categories could be revealed. Relationships tying the tables together were defined and queries relevant to the stated hypotheses designed. By creating 'queries' out of existing tables, a rapid search path to all necessary information is provided and unnecessary duplications eliminated. Queries allow you to answer questions about your data, to extract specific information from tables, and to change selected data in various ways. Clearly, it is a good way to manage large amounts of related data. An example of the organization and relationships between tables in the database are given in Figure 5.

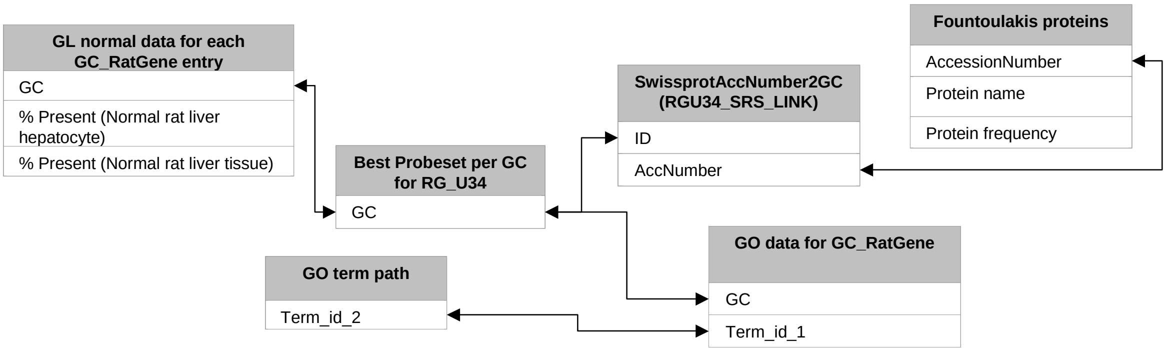

GL normal data for each GC_RatGene entry
GC
% Present (Normal rat liver hepatocyte)
% Present (Normal rat liver tissue)
Best Probeset per GC for RG_U34
GC
SwissprotAccNumber2GC (RGU34_SRS_LINK)
ID
AccNumber
Fountoulakis proteins
AccessionNumber
Protein name
Protein frequency
GO term path
Term_id_2
GO data for GC_RatGene
GC
Term_id_1

Figure 5: Designing queries out of tables in Microsoft Access
Queries provide a rapid search path to all necessary information buried in the imported tables. By linking tables through appropriate data fields, stated hypotheses can be approached in a straightforward manner.

3.5.1.1 Definition of correlation

Since the correlation analysis in this report is of qualitative nature, the correlation between transcript and protein is based on the mRNA vs protein detection simplicity/difficulty. In the *Fountoulakis* protein dataset, a measure on protein frequency indicating what proteins are easy to detect in rat liver is given. A relational protein frequency across the 110 vs 60 protein samples was calculated in order to get a number between 0 and 1, where 1 means that the protein is *always* detectable in this tissue. The relative protein frequency is consequently analogues to the Gene Logic percent present term, which indicates what transcripts are easy vs hard to detect. A correlation expression was created in Microsoft Access expression builder by simply dividing percent present by relative protein frequency, returning a correlation value between 0 and 1, where 1 corresponds to a perfect correlation. A definition of the yielded correlation value (x) was stated, and divided into separate matching score categories, (see Table 1).

Correlation value, x	**Matching score**	**Correlation**
$0.5<x<1.5$	1	Very good
$x<0.5$ and $1.5<x< 2$	2	Good
$2<x<2.5$	3	Ok
$2.5<x<10$	4	Poor
$x>10$	5	Very poor

Table 1: Definition of correlation
A correlation score was calculated and separated into matching score categories between 1 and 5, where 1 corresponds to a very good correlation and 5 a very poor.

3.5.1.2 The Swissprot/TrEMBL criteria – a surrogate frequency assessment

The proteins extracted from Swissprot/TrEMBL represent proteins that have ever been detected in normal rat liver tissue. The databases do not provide any information on protein frequency; if the protein is hard/easy to detect or if it has been detected more than once. Hence, a correlation analysis such as the one described for the *Fountoulakis* protein set is harder to make. To get around this problem, a correlation criteria or a surrogate frequency assessment, was stated:

”All the proteins that are detectable in rat liver, according to Swissprot/TrEMBL, should be at least 50% present at the transcriptional level”.

4 Results

In this chapter, the hypotheses will be presented. A short description of the background and rationale to each of the five stated hypothesis precedes each section. The majority of the hypotheses are based on previous results from mRNA-to-protein correlation analyses made on the model organism Saccharomyces cerevisiae.

4.1 Presentation of hypotheses

Five hypotheses were generated from 40 articles on this subject, which represents the majority on what has been published regarding this research topic. Researchers have reached different conclusions about the correlation between mRNA and protein, but the stated hypotheses are based on the most common conclusions. Thus it remains to be investigated whether these are suitable for the protein and transcript data in this rat liver study. The five hypotheses are:

- Hypothesis I: *"High-abundant proteins present a better mRNA-to-protein correlation than low-abundant proteins"*
- Hypothesis II: *"Transcripts that are not detectable with Affymetrix chips but are detectable at the protein level have a shorter mRNA half-life"*
- Hypothesis III: "*Cytoplasmic proteins present a better correlation with mRNA than nuclear proteins"*
- Hypothesis IV: *"The correlation between mRNA and protein is very poor in the mitochondrion"*
- Hypothesis V: "*Genes belonging to the category 'metabolism' are well correlated at the mRNA and protein levels"*

4.2 Hypothesis I: "High-abundant proteins present a better mRNA-to-protein correlation than low-abundant proteins"

According to *Gygi et al* (1999), the level of protein abundance is a factor that influences the correlation between mRNA and protein. A good mRNA-to-protein correlation was reported when the 11 most abundant proteins were examined in yeast (3). Another proteomic and microarray study of bladder cancer, made by *Celis et* al (2000) demonstrated a good correlation between transcript and protein levels among 40 well resolved, abundant proteins (84). The first hypothesis basically states that proteins that are easy to detect in rat liver should also be easy to detect at the transcript level.

4.2.1 Correlation for abundant proteins

In the *Fountoulakis* protein dataset, 28 proteins have a protein frequency above 50%. These are considered the most abundant proteins (see Table 2 below). All of these proteins fall under correlation categories 1 and 2 (Table 1 in 'Materials and Methods'), which means that they represent a 'good' to a 'very good' correlation. The average and median correlation values in liver are 1.49 and 1.47, and the transcript detectability is very high (100%). The three proteins 3-oxo-5-beta-steroid 4-dehydrogenase (3O5B_RAT), serum albumin (ALBU_RAT), and carbamoyl-phosphate synthase

(CPSM_RAT) are practically always detectable at both transcript and protein level, and the mRNA-to-protein correlation is nearly perfect (1.02, 1.11, and 1.22). For the 96 proteins with a protein frequency below 50%, only 5% (6 proteins) represent a good to a very good correlation, and for these cases, the good correlation value obtained is due to poor detectability at both transcript and protein levels (data not presented).

Protein	**%T$_{\text{hepatocyte}}$**	**%T$_{\text{liver}}$**	**%P$_{\text{liver}}$**	**Corr$_{\text{hepatocyte}}$**	**Corr$_{\text{liver}}$**
3HAO_RA	73.86	100	72	1.03	1.39
3O5B_RAT	100	99.87	82	1.22	1.22
ALBU_RAT	100	99.8	98	1.02	1.02
BUP_RAT	99.15	100	62	1.6	1.61
COMT_RAT	100	99.8	78	1.28	1.28
CPSM_RAT	99.72	99.8	90	1.11	1.11
DHAM_RAT	100	100	76	1.32	1.32
DHE3_RAT	100	100	76	1.32	1.32
DIDH_RAT	100	99.8	68	1.47	1.47
ER60_RAT	99.72	99.93	72	1.38	1.39
GR78_RAT	100	100	79	1.27	1.27
HMCM_RAT	100	99.93	60	1.67	1.67
HPPD_RAT	100	100	54	1.85	1.85
IDHC_RAT	99.72	100	68	1.47	1.47
K2C8_RAT	100	99.8	64	1.56	1.56
M2GD_RAT	91.19	100	52	1.75	1.92
METL_RAT	100	99.82	70	1.43	1.43
OTC_RAT	98.58	99.87	54	1.83	1.85
P60_MOUSE	100	100	71	1.41	1.41
PDI_RAT	100	100	60	1.67	1.67
PH4H_RAT	100	100	60	1.67	1.67
PYC_RAT	74.43	100	62	1.2	1.61
SAHH_RAT	100	100	68	1.47	1.47
SM30_RAT	76.42	100	72	1.06	1.39
SUAC_RAT	95.45	99.93	70	1.36	1.43
TERA_RAT	100	99.87	58	1.72	1.72
THIM_RAT	100	100	66	1.52	1.52
TRFE_RAT	100	100	64	1.56	1.56

Table 2: Correlation for the most abundant proteins in the *Fountoulakis* dataset
The 28 most abundant proteins detected in rat liver according to *Fountoulakis et al*. %T refers to the %presence for hepatocytes and liver tissue (transcript detectability), %P is the protein frequency (protein detectability), and Corr is the mRNA-to-protein correlation value in hepatocytes and rat liver.

For the proteins extracted from Swissprot and TrEMBL, nothing can be said about the protein frequency or abundance, since such data is not provided in these databases. Referring to the surrogate frequency assessment described in the previous chapter; that all the proteins detectable in rat liver according to Swissprot/TrEMBL should be at least 50% present at the transcript level, gives a slight hint to the correlation for this dataset. Out of the total 426 proteins detectable in rat liver, 71% fulfil this surrogate correlation criteria and out of these, 90% are very easy (transcript frequency > 80%) to detect.

4.2.1.1 Protein properties for high abundant proteins

The often detected proteins in a 2D gel analysis are usually the major, hydrophilic components, with average pI and molecular weight values. Moreover, they are easily digested and deliver a sufficient number of peptides. In order to investigate differences in protein characteristics between high and low abundant proteins in the *Fountoulakis* dataset, some critical protein properties were examined. Table 3 shows the average values of some protein characteristics that have an effect on the level of protein abundance (a more detailed table can be found in the attachment Table 1).

Protein properties	**High abundant proteins (21)**	**Low abundant proteins (25)**
Level	1.33	0.17
Spots	189	5
pI	6.5	6.9
MW	71 574	45 578
TM	0.27	0
GRAVY	-0.24	-0.41

Table 3: Average values of the protein properties for high abundant vs low abundant proteins
Level refers to the approximate percentage of the volume of spots representing a particular protein, in comparison with the total proteins present in the 2D gel (61,62)). *Spots* correspond to the number of spots which represent the particular protein and which were identified by MS in the rat liver samples (61,62). *MW* is the average molecular weight. *TM* is the number of predicted transmembrane regions according to Klein et al (108). *GRAVY* is the grand average hydrophobicity value according to Kyte-Dolittle (109).

4.3 Hypothesis II: "Transcripts that are not detectable with Affymetrix chips but are detectable at the protein level have a shorter mRNA half-life"

The second hypothesis is important but not straightforward to test. There are several factors that influence transcript detectability and mRNA turnover rates, and the detailed process of mRNA degradation is not yet clear. Here, the transcripts that are not detectable with Affymetrix chips are defined as those with a percent presence below 5%.

4.3.1 Background – results from previous transcript decay rate studies

In 2003, *Yang et al* measured the decay rates in two human cell lines and reached the conclusion that certain transcripts belonging to particular gene ontology categories decay faster than others (110). Similar results have been observed in *E. coli* and *S. cerevisiae* (111-113). *Yang et al* used an approach similar to the one used in this study; they used the GO hierarchy of biological processes and assigned mRNAs to functional classes by coupling probesets to GO codes and comparing the decay rate statistics between these classes. The decay rate determination was based on the changes in mRNA levels observed after application of the RNA polymerase inhibitor actinomycin D. One of their conclusions was that gene-regulatory transcripts had a rapid turnover whereas transcripts related to "biosynthesis" or "metabolism" had a reduced turnover. The study of *Yang et al* resulted in the discrimination of fast vs. slow decaying transcripts:

Fast-decaying transcripts:

- *Transcription and transcription factors*
- *Cell cycle regulation*
- *Apoptosis*
- *Development*

Slow decaying transcripts:

- *Biosynthesis*
- *Catabolism*
- *Carbohydrate metabolism*
- *Transport*
- *Protein metabolism*

4.3.2 Discrepancies in transcript detectability for different GO categories in the Swissprot/TrEMBL dataset

In an attempt to test the proposition by *Yang et al*.; that certain transcripts decay faster or slower than others, the detectability for the transcripts among the GO categories: process, component, and function was examined. The Swissprot/TrEMBL surrogate frequency assessment can also be used to measure the transcript detectability (and thus give an indication of the mRNA abundance) among different GO categories. In order to find out if there are some transcripts in certain GO categories that are easier to detect than others, the following method was used: the number of transcripts that fulfilled the surrogate frequency assessment ("high abundant" transcripts with a %presence above 50) were divided by the total number of transcripts for each GO category resulting in a number between 0 and 1, where 1 corresponds to a category in which all the transcripts are *always* detected. Figure 6 below illustrates the transcript detectability for different categories within the GO annotations: process, component, and function.

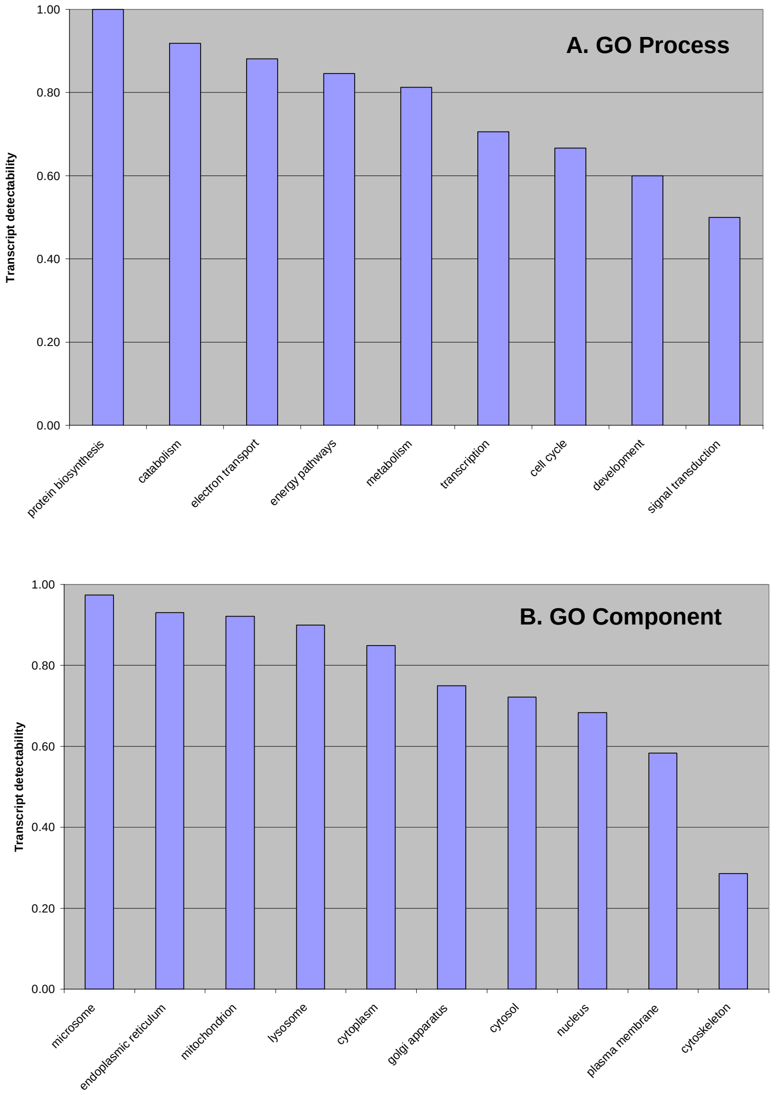
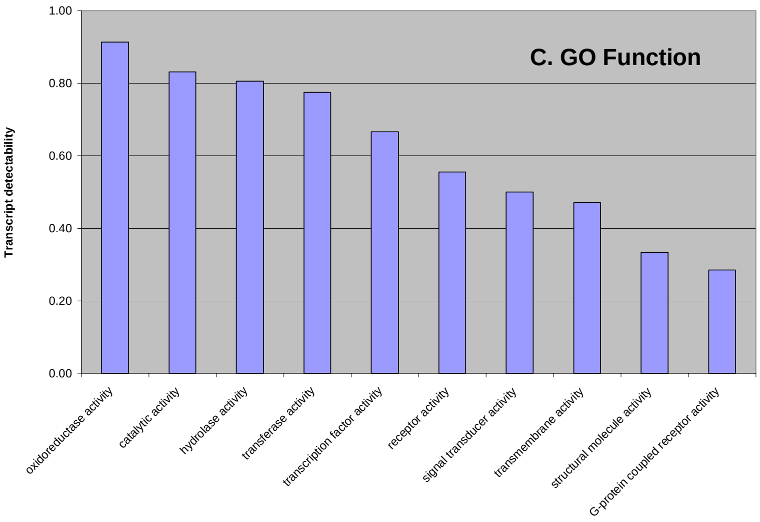


Figure 6: Transcript detectability among different GO categories

A. The difference in transcript detectability for transcript groups belonging to the *GO Process* category. Transcripts belonging to "protein biosynthesis" are much easier to detect than transcripts in the category "signal transduction"

B. The difference in transcript detectability for transcript groups belonging to the category *GO Component*. Transcripts in the microsome and endoplasmic reticulum are much more easily detected than those in the plasma membrane or cytoskeleton.



C. The difference in transcript detectability for the transcript groups belonging to the category *GO Function*. mRNAs for ‘oxidoreductase activity’ genes are easier to detect than transcripts involved in ‘G-protein coupled receptor activity’.

4.3.3 Transcripts with low detectability in rat liver

In the *Fountoulakis* dataset, three liver proteins are not detectable at the transcript level: spectrin alpha chain (SPTA2), golgi autoantigen golgin subfamily A member 2 (GOGA2) and heat shock-related 70kDa protein 2 (HSP72) (Table 4).

Protein	%T$_{hepatocyte}$	%T$_{liver}$	%P$_{liver}$	Corr$_{hepatocyte}$	Corr$_{liver}$
SPTA2_RAT	46.31	2.56	3	15.44	0.85
GOGA2_RAT	2.56	0.46	3	0.85	0.15
HSP72_RAT	0	0.33	32	0	0.01

Table 4: Proteins not detectable at the transcript level in rat liver
The three proteins SPTA2, GOGA2, and HSP_72 have a transcript detectability less than 5% in rat liver and are practically never detected at the transcript level.

In the Swissprot/TrEMBL dataset, 59 proteins are very hard to detect at the transcript level (see Table 2 in the attachments). As can be observed in Table 4, the transcript detectability for SPTA2 is higher in hepatocyte than in liver (46.31 vs. 2.56). This observation may indicate that hepatocytes are present in other organs than in the liver.

4.3.3 Transcripts with low detectability in rat hepatocytes

For hepatocytes, five proteins are not detectable in the *Fountoulakis* set: acyl-CoA dehydrogenase short/branched chain (ACDB), golgi autoantigen golgin subfamily A member 2 (GOGA2), heat shock-related 70kDa protein 2 (HSP72), creatine kinase B chain (KCRB), and propionyl-CoA carboxylase beta chain (PCCB) (Table 5). 68 proteins in the Swissprot/TrEMBL dataset are not detectable at the transcript level (refer to Table 2 in the attachments for a complete list)

Protein	%T$_{hepatocyte}$	%T$_{liver}$	%P$_{liver}$	Corr$_{hepatocyte}$	Corr$_{liver}$
ACDB_RAT	0	13.6	22	0	0.62
GOGA2_RAT	2.56	0.46	3	0.85	0.15
HSP72_RAT	0	0.33	32	0	0.01
KCRB_RAT	0.28	26.22	8	0.04	3.28
PCCB_RAT	0	16.62	14	0	1.19

Table 5: Proteins not detectable at the transcript level in rat hepatocytes
The five proteins ACDB, GOGA2, HSP72, KCRB, and PCCB have a transcript detectability less than 5% in rat hepatocytes and are practically never detectable at the transcript level.

In the table above, it is worth noting that two proteins, KCRB and PCCB, can be detected with transcript techniques in liver, but not in hepatocytes (although the transcript detectability is rather low in liver as well). This observation reflects the fact that the liver consists of other cells besides the dominant hepatocytes.

4.4 Hypothesis III: "Cytoplasmic proteins present a better correlation with mRNA than nuclear proteins"

Greenbaum, *Jansen,* and *Gerstein et al* have made several studies on the correlation between mRNA and protein in yeast (6-7,10-11). In 2003, they merged all available yeast mRNA and protein abundance datasets in order to compare the mRNA-to-protein correlation between different compartments and functional modules (6-7). They found that the cellular populations of transcripts and proteins were both enriched in cytoplasmic proteins relative to nuclear ones. In addition, they observed a good correlation in the nucleolous. The same correlation pattern was observed for processes related to these locations: whereas the correlation in the nucleus-related process 'transcription' was poor, the category 'protein biosynthesis', a process highly associated with the cytoplasm, demonstrated a better mRNA-to-protein correlation. Their results form the basis for Hypothesis III.

4.4.1 Correlation in the nucleus

In the *Fountoulakis* dataset, there are only 6 proteins related to "nucleus" (see Table 6 below), and the correlation between mRNA and protein is very poor with a median correlation value of 7.5, and only one protein falling into the correlation category 1-3. In the Swissprot/TrEMBL dataset, 61 proteins are linked to "nucleus" and 69% of these fulfil the Swissprot/TrEMBL surrogate frequency assessment. There are no proteins associated with 'nucleolous' in either of the datasets. Thus, no conclusions about the correlation here can be drawn. Proteins related to "transcription" are also depleted in the *Fountoulakis* protein data set. Only 2 transcriptional proteins are found. To see the number of mRNA-to-protein connections for the categories 'nucleus' and 'cytoplasm', as well as their relating processes, obtained from each of the protein datasets, refer to Table 7.

Protein	%$T_{hepatocyte}$	%T_{liver}	%P_{liver}	$Corr_{hepatocyte}$	$Corr_{liver}$
GTT2_RAT	99.43	100	6	16.57	16.67
PRC2_RAT	100	99.93	30	3.33	3.33
PRCT_RAT	100	100	10	10	10
PRCZ_RAT	100	100	20	5	5
PRS4_RAT	100	99.93	8	12.5	12.49
TERA_RAT	100	99.87	58	1.72	1.72

Table 6: **mRNA-to-protein correlation for 6 nuclear proteins.**
The majority of the proteins (83%) are poorly correlated with their mRNA counterpart. The average and median correlation values are 8.2 and 7.5 in rat liver. The transcript detectability for these 6 proteins is high in both hepatocyte and liver (median %T = 100%).

4.4.2 Correlation in the cytoplasm

Compared to the relatively few proteins in the nucleus category, there are a lot more proteins detected and thus, more protein-to-mRNA connections linked to the cytoplasm. Both the *Fountoulakis* and the Swissprot/TrEMBL data sets are enriched in cytoplasmic proteins; with 79 and 172 proteins respectively. Since the cytoplasm is a quite big entity and the correlation situation ought to be far from homogenous, the cytoplasm was split into the subcategories: cytosol, endoplasmic reticulum (ER), Golgi apparatus, peroxisome, lysosome, and mitochondrion. The correlation in the mitochondrion, however, will be elucidated in the next chapter. The mRNA-to-protein correlation was examined and compared on the basis on the proportion of proteins falling into correlation categories 1-3 ('very good', 'good' or 'ok' correlation), and median correlation values. The median value is used instead of the average since it is more stable and less affected by extreme values; the differences between mRNA expression and protein abundance level may be quite dramatic for individual genes.

A good correlation was obtained in the ER and Golgi compared to the cytosol. 63% of the ER proteins and 60% of the proteins in the Golgi fall into matching categories 1-3. In contrast, only 26% of the cytosolic proteins belong to these correlation categories. The median correlation values are 1.79 and 1.85 for the ER and Golgi, and 3.57 for the cytosol. For clarification, the correlation distribution in terms of correlation category and median value is illustrated in Figure 7. There were too few proteins in the categories 'peroxisome' and 'lysosome' to be able to draw conclusions, however for the two proteins in each of these categories the correlation obtained between mRNA and protein was very poor. Only 2 proteins in the *Fountoulakis* dataset and 6 in the Swissprot/TrEMBL set are related to the cytoplasmic process 'protein biosynthesis'. The correlation values for these two are 5.56 and 3.12 and the transcript detectability for the 6 proteins in the Swissprot/TrEMBL set is 100% (Table 7).

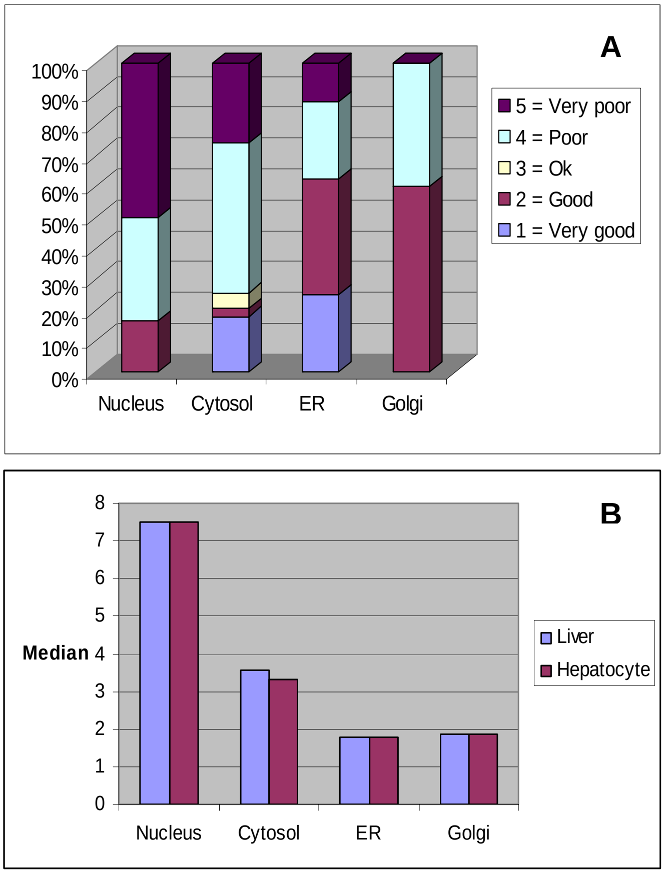


Figure 7: Correlation distribution in nucleus and cytoplasm

A: The majority of the proteins in the endoplasmic reticulum and the Golgi apparatus fall into correlation categories 1-3 (63% vs. 60%) and represent a good mRNA-to-protein correlation. The correlation is not as good for the 'nucleus' and 'cytosol' categories (17% vs. 26% in correlation categories 1-3).

B: The difference in mRNA-to-protein correlation between the cytoplasmic subcategories is reflected by the median correlation value for each compartment: nucleus: 7.5, cytosol 3.33, ER: 1.79 and Golgi: 1.85.

It is important to take into account the difference in the number of proteins associated with each compartment and its related processes. The number of proteins along with some statistical information is summarized in Table 7 below.

	Σ mRNA-to-protein connections in the Fountoulakis data set	Σ mRNA-to-protein connections in the Swissprot/TrEMBL data set	Median correlation in rat liver	% proteins in correlation categories 1-3	Swissprot/ TrEMBL surrogate frequency assessment (%)
Nucleus	6	61	7,5	17	69
Cytosol	39	19	3,57	26	74
Endoplasmic reticulum, ER	9	44	1,79	63	93
Golgi apparatus	6	16	1,85	60	75
Transcription	2	18	(2,78)	50	67
Protein biosynthesis	2	6	(4,34)	0	100

Table 7: Summary of the mRNA-to-protein relationships in the cytoplasm
There is a noteworthy difference in the number of mRNA-to-protein connections between categories in the two datasets. The correlation statistics is displayed in terms of the median correlation value and the % proteins that fall in the correlation categories 1-3. The Swissprot/TrEMBL surrogate frequency percentage reflects the detectability of the transcripts for each category.

4.5 Hypothesis IV: "The correlation between mRNA and protein is very poor in the mitochondria"

Gygi (2002), Hood (2004), *Greenbaum* and *Gerstein* (2003) and *Beyer* (2004) have all reached the conclusion that the mitochondrion is the compartment in which the correlation between mRNA and protein is the poorest. However, there are diverse opinions about the correlation in mitochondria-related processes, such as 'glycolysis', 'electron transport', and 'energy generation'. Whereas *Greenbaum*, *Gerstein,* and *Beyer* suggest a good correlation for genes belonging to category 'energy generation', *Hood et al* have obtained a poor correlation for the same category, but a very good correlation for the category 'glycolysis' (2).

4.5.1 Mitochondrial proteins in the *Fountoulakis* data set

In order to evaluate the forth hypothesis, the proteins belonging to the 'mitochondrion' in the *Fountoulakis* dataset were analyzed. There were 46 proteins in total in this category. In order to facilitate the search for possible correlation patterns within this, rather complex, compartment, the proteins were split into subcompartments; mitochondrial matrix (MM), mitochondrial inner membrane (MIM), outer mitochondrial membrane (OMM) and proteins associated with the general term 'mitochondrion'. The distribution of proteins in each of the subgroups is shown in Figure 8 below.

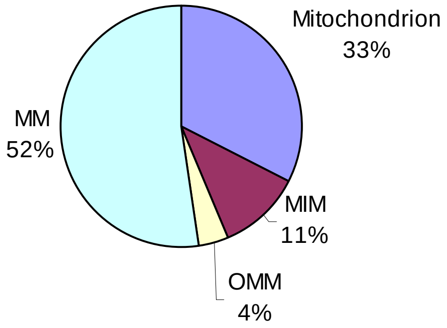


Figure 8: Distribution of mitochondrial proteins There are 46 mitochondrial proteins in total in the *Fountoulakis* dataset. 52% (24) of these are associated with the mitochondrial matrix, MM; 11% (5) with the mitochondrial inner membrane, MIM; 4% (2) are found in the outer mitochondrial membrane, OMM; and 33% (15) are associated with the general term 'mitochondrion'.

4.5.2 Correlation in the mitochondrion

The mRNA-to-protein correlation for the entity 'mitochondrion', as well as its subparts was evaluated according to the median correlation value and the proportion of proteins falling into correlation categories 1-3. The median correlation was poor in total (3.23), and in the mitochondrial subparts as well; 3.12 in the mitochondrial matrix and 5.26 for the general term 'mitochondrion'. The correlation in the mitochondrial inner membrane, MIM, was better, with a median correlation value of 2.5. 60% of the proteins in MIM fell into correlation categories 1-3, whereas 27% and 38% of the proteins in 'mitochondrion' and 'mitochondrial matrix, MM' belonged to these categories. It is important to mention the fact that a lot of the proteins in the 'mitochondrial matrix' category exhibited a good correlation value due to poor detectability at both the transcript and protein level. The correlation distribution is illustrated in Figure 8.

4.5.3 Correlation for mitochondrial-related processes

The mitochondrion is a ubiquitous organelle responsible for the energy metabolism of eukaryotic cells. It is known for housing the oxidative phosphorylation machinery as well as enzymes needed for free fatty acid metabolism and the Kreb's cycle. In addition to serving as the main intracellular source of energy, mitochondria regulate several other cellular processes such as electron transport and apoptosis. They are also the storage site for a number of soluble proteins that mediate apoptosis, including cytochrome c, certain procaspases and apoptosis-inducing factor (114). Key steps of heme biosynthesis, ketone body generation, and hormone synthesis also reside within this organelle (115).

Several mitochondrial-related processes were examined, however for 'glycolysis', 'oxidative phosphorylation', and 'apoptosis', there were to few proteins to be able to draw any conclusions. Nevertheless, the categories 'electron transport' and 'energy pathways' were studied in depth. The correlation for proteins belonging to 'electron transport' was relatively good, with 55% of the proteins belonging to correlation categories 1-3 and a median correlation value of 2.63. For the 'energy pathways' group, however the correlation was poor for the majority of the proteins (80%). The median correlation value was 3.49 for this category. The correlation and median distribution for the mitochondrion, its subparts and the compartment-related processes, 'energy pathways' and 'electron transport' is demonstrated in Figure 9.

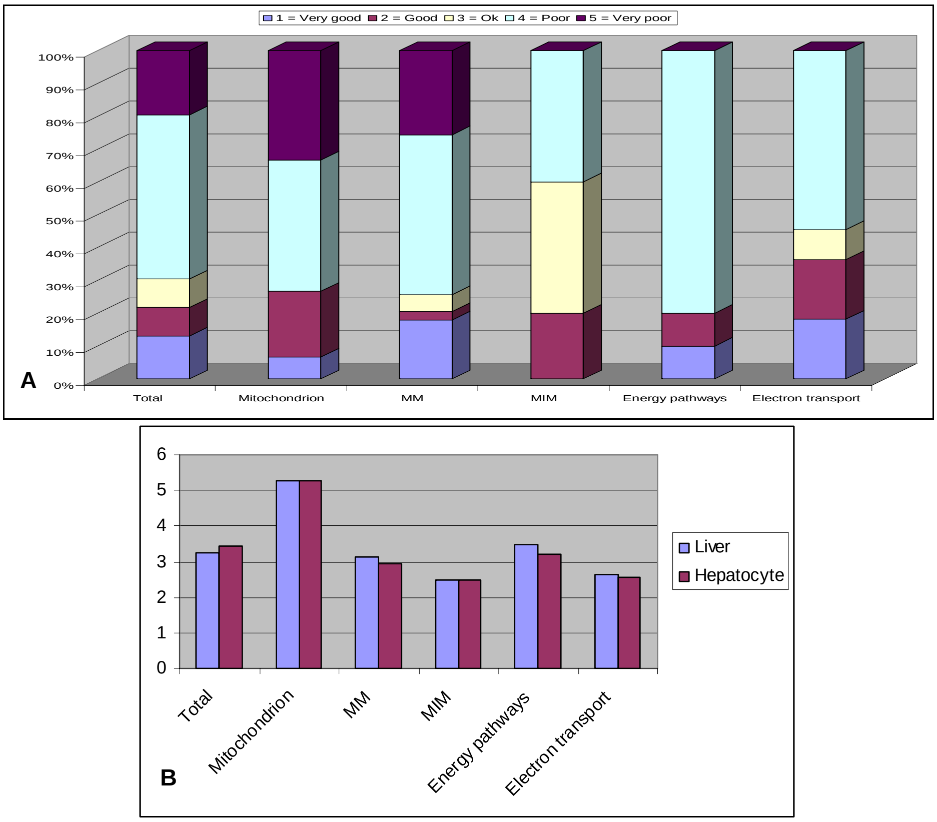


Figure 9: Correlation distribution for the mitochondrion, its subparts and related processes
A: The majority of the mitochondrial proteins have a poor mRNA-to-protein correlation and fall into correlation categories 4 and 5. However, the proteins in the mitochondrial inner membrane present a good correlation with their mRNA counterparts; 60% fall into correlation categories 1-3. The correlation in the group 'energy pathways' is poor (80% fall into correlation categories 4-5), whereas the correlation for genes in the process 'electron transport' is quite good with 55% in categories 1-3.
B: The median correlation values reflect the same general differences as the correlation category distribution among mitochondrial subcompartments and processes. The median correlation value in total is 3.23, 'mitochondrion': 5.26, MM: 3.12, MIM: 2.5, energy pathways: 3.49, and electron transport: 2.64.

4.6 Hypothesis V: "Genes belonging to the category 'metabolism' are well correlated at the mRNA and protein levels"

Greenbaum and *Gerstein* (2003) and *Beyer et al* (2004) reached the same conclusion about a good correlation for genes belonging to the category 'Metabolism'. A particularly good correlation for genes belonging to the category 'carbohydrate metabolism' was observed by *Hood et al* in 2002 (2).

4.6.1 Transcript detectability for 'metabolism' genes

Both protein datasets (*Fountoulakis* and Swissprot/TrEMBL) were enriched in proteins belonging to the category 'metabolism'. The number of mRNA-to-protein connections for this category was 85 in the *Fountoulakis* set and 274 in the Swissprot/TrEMBL set. The transcript detectability was explored for this category as a whole, and also as subcategories in terms of amino acid metabolism, carbohydrate metabolism, fatty acid metabolism, lipid metabolism, and nitrogen metabolism. The same calculations as in Section 4.3.2 and Figure 6 were made to determine the detectability for mRNAs within a certain group. The transcript detectability was high for each of the subcategories (see

Figure 10 below). Thus, genes belonging to 'metabolism' are easy to detect at the level of mRNA.

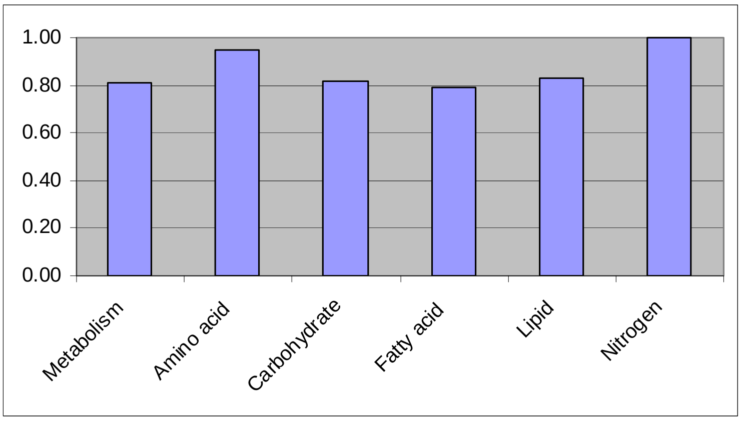


Figure 10: Transcript detectability for transcripts belonging to 'metabolism' and 'metabolism'-subcategories'
All of the genes in the metabolism categories have a transcript detectability above 80% and are very easy to detect with transcript techniques at the mRNA level. The transcripts in the category 'nitrogen metabolism' are always easily detected.

4.6.2 Correlation for genes belonging to 'metabolism' and 'metabolism sub-categories'

The mRNA-to-protein correlation was examined for the general term 'metabolism', which resulted in a median correlation value of 2.94 for the entire category. 38% of the proteins within this category belonged to correlation categories 1-3. The correlation was also explored in line with the same subcategories as mentioned above. The correlation was quite poor for 'carbohydrate- and fatty acid metabolism'; 20% vs. 36% belonged to correlation categories 1-3, and the median correlation values were 3.5 and 3.6 respectively. In contrast, the correlation between mRNA and protein in the categories 'amino acid-, lipid-, and nitrogen metabolism' was good with median correlation values ranging from 1.85 to 2.64, and the majority of the proteins belonging to correlation categories 1-3 (50-60%). The correlation distribution for each of the subcategories is illustrated in Figure 11 below and the statistical data is found in Table 8.

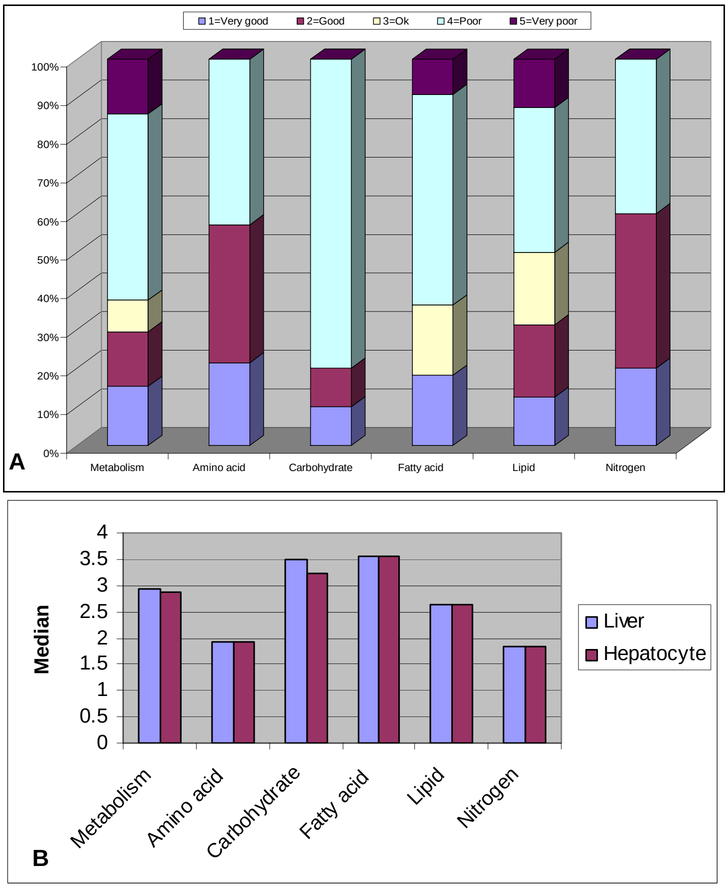


Figure 11: Correlation distribution for the GO process 'metabolism' and metabolism-subcategories
A: The difference in mRNA-to-protein correlation between the metabolism subcategories is observed by the dissimilar proportion of proteins falling under correlation categories 1-3: 'metabolism': 38%, amino acid: 57%, carbohydrate: 20%, fatty acid 36%, lipid 50% and nitrogen metabolism 60%.
B: The median correlation values are relatively high in the categories 'carbohydrate- and fatty acid metabolism' (3.49 and 3.57) compared to amino acid-, lipid-, and nitrogen metabolism (1.93, 2.64, and 1.85)

	Σ mRNA-to-protein connections in the Fountoulakis data set	Σ mRNA-to-protein connections in the Swissprot/TrEMBL data set	Median correlation in rat liver	% proteins in correlation categories 1-3	Swissprot/ TrEMBL surrogate frequency assessment (%)
Metabolism	85	274	2.94	38	81
Amino acid metabolism	14	22	1.93	57	95
Carbohydrate metabolism	10	22	3.49	20	82
Fatty acid metabolism	11	24	3.57	36	79
Lipid metabolism	16	53	2.64	50	83
Nitrogen metabolism	5	5	1.85	60	100

Table 8: Statistical information on the mRNA-to-protein relationships for the metabolism subcategories
This data summary details the number of mRNA-to-protein connections found in the two datasets, the median correlation values, the proportion of proteins that belong to correlation categories 1-3, and the Swissprot frequency assessment.

5 Discussion

This chapter comments on the results introduced in Chapter 4 and presents evaluations of the stated hypotheses. Some speculative reasons for the lack vs. presence of correlation associated with each compartment, process, and hypothesis are given. To start with, the limitations and biases in the collected data are discussed.

5.1 Materials and methods

5.1.1 Limitations given the small size of protein data

The largest complication in this analysis is the limited amount of protein data. Besides yielding fewer connections to the mRNA counterparts for statistical analysis, the small number of protein frequency measurements biased the results towards certain protein families. The 128 proteins are certainly not a random selection from the possible 684 in rat liver (101). They are, instead, skewed towards proteins that are highly expressed. The results will be more complete and definitive when larger proteomics datasets become available. However, the formalism and approach developed in this report should be relevant for future datasets. Although the protein datasets used in this study are small in comparison to the transcript data, many protein features in both the proteome and the transcriptome are dominated by the very highly expressed proteins. Under this circumstance, it should be sufficient to look at this smaller number of dominating proteins to approximately characterize the whole population.

5.1.2 Biases in transcript and protein data

mRNA expression data was extracted from the Affymetrix chipset RG_U34, an array set designed to contain a maximal representation of the majority of rat genes. Although high-throughput expression data of this kind suffer with regard to cross hybridization and saturation of probes for highly expressed genes, the major limitation lies in the proteomics data collected. Protein data was extracted from two sources: (1) published proteomic analyses made by *Fountoulakis et al*. and (2) Swissprot/TrEMBL. The fact that the *Fountoulakis* study is fairly “old” (3 years) complicates the connection with the mRNA counterparts. Some proteins may not have been detected at the time of the study, and a lot of proteins have changed names and accession numbers since then.

The Swissprot and TrEMBL databases suffer from poor distribution of protein expression data by researchers. According to Swissprot, there are 4313 proteins (January 2005) expressed in rat, and out of these 684 are expressed in rat liver. Due to the fact that 2D gels are unable to resolve membrane proteins, proteins with low abundance and extremes in pI, more proteins should be expressed in reality. Furthermore, the procedures for identification and quantification of the protein spots are subject to uncertainties (116). Human biases include differences between laboratories in sample preparation. A positive aspect of the *Fountoulakis* study was the use of a subcellular fractionation approach prior to 2D analysis, which increased the probability of detecting low-abundant proteins.

5.2 *Hypothesis I: “High-abundant proteins are better correlated to mRNA than low abundant proteins”*

5.2.1 High abundant proteins present a very good mRNA-to-protein correlation

The first hypothesis basically means that proteins that are easy to detect should also be easy to detect at the transcript level. From Table 1 in the ‘Results’ section, one can see that this is definitely the case. The most abundant proteins are very easy to detect at the transcript level and the average and median correlation values of 1.49 and 1.47 in liver further supports this conclusion. There are three proteins that are practically always detectable with both transcript and protein techniques, and that present a nearly perfect mRNA-to-protein correlation; ALBU, CPSM, and 3OB5. ALBU is a secreted serum albumin and the main protein of plasma. Its main function is the regulation of the colloidal osmotic pressure of blood. CPSM is a mitochondrial protein with tissue specificity primarily in the liver and small intestine. It is involved in the urea cycle of ureotelic animals where the enzyme plays an important role in removing excess ammonia from the cell. 3OB5, a reductase present in the cytoplasm, catalyzes the reduction of the delta (4) double bond of bile acid intermediates and steroid hormones. (101)

5.2.2 Differences in protein characteristics between high and low abundant proteins

The number of spots which represent a particular protein on a 2D gel gives an indication on the protein abundance. Thus, it is not surprising that the average spot number is about 40 times higher for the high abundant protein set (189 compared to 5). *Fountoulakis* also consider the so called “level”; the approximate percentage of the volume of the spots representing a particular protein, in comparison with the total proteins present in the gel (61,62), which simultaneously is an indication on the protein abundance. The level is significantly higher for high abundant proteins (average 1.33 compared to 0.17 for low abundant proteins). The difference in pI and molecular weight between the datasets is not as significant. The grand average hydrophobicity value (GRAVY) scores provide an image of the hydrophobicity of the whole protein. It is hard to draw conclusions about the average GRAVY score since this kind of data is missing on nearly half of the proteins. One could, however, conclude that the majority of the proteins have negative GRAVY values, regardless of abundance, which indicates that most of the proteins in the dataset are hydrophilic. In addition, there is a lack of data on the number of theoretical transmembrane segments on several proteins. Nevertheless, the majority of the proteins are not associated with any transmembrane segments.

5.3 Hypothesis II: “The transcripts that are not detectable with Affymetrix chips but are detectable at the protein level have a shorter mRNA half-life”

5.3.1 mRNA degradation –unresolved mechanistic complexity

The steady-state levels of mRNA in cells depend on the rate of transcriptional initiation and elongation, on the efficiency of splicing and termination of transcription, on the rate of export to the cytoplasm, and on the stability of the mRNA in the cytoplasm (117). The latter process has been subject to a lot of research, and many researchers have attempted to identify the mechanisms which affect the mRNA half-life. Although some underlying principles affecting the process of mRNA degradation have been established biochemically, many major questions remain unanswered about this critical biological process. It is known that the half-life of an mRNA depends on a combination of elements which stabilize or destabilize mRNA, e.g. AU-rich elements (ARE) and proteins which bind to the mRNA and directly or indirectly affect the degradation of the transcript (118). (See Table 9 below for a summary of factors influencing mRNA half-lives).

Factor	**Note**
Cap	Uncapped mRNAs are less stable. Removal of cap is an important step in the degradation pathway'
Length of PolyA tail + PolyA binding protein (PABP)	Poly(A)tail removal accelerates degradation and PABP protects mRNA in vitro from rapid decay
Destabilizing elements	AU-rich elements, ARE, confers instability of mRNAs
Premature termination codons or nonsense codons	May result in Non-mediated decay (NMD) or mRNA surveillance

Table 9: Critical factors that are known to influence the stability of mRNA
The 5'cap, the length of the poly(A) tail, poly(A)binding proteins, destabilizing elements such as AREs, and nonsense codons affect the rate at which mRNAs degrade (118).

It is known from previous analyses that transcripts that are induced rapidly are also destroyed more rapidly (119-121) suggesting that mRNAs with short half-lives respond to changes in transcription more rapidly than those that are relatively stable (Figure 12).

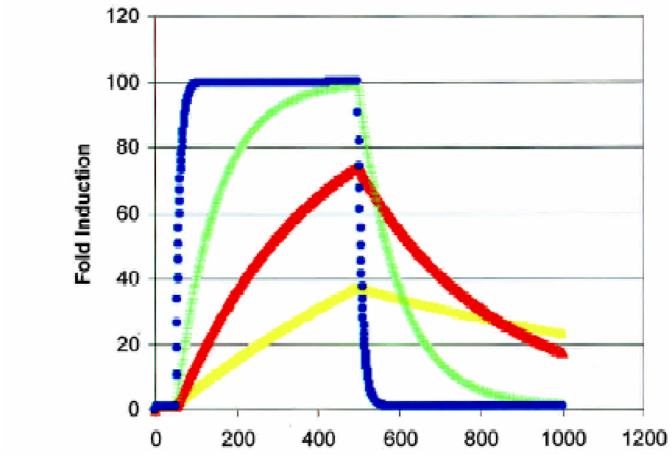


Figure 12: The rate of mRNA induction is proportional to the rate of degradation.
mRNAs that are induced rapidly are also destroyed more rapidly (blue and green), whereas mRNAs that are produced at a slow and more constant level remain longer in the cell (yellow and red). The x-axis corresponds to time (minutes). (Figure 12 is adapted from reference 110)

5.3.2 The transcripts that are hard to detect may be fast-decaying

The transcript detectability among categories within GO process, component, and function was examined and evaluated in Section 4.3.2. Referring to Figure 6 in the same section, a significant difference between gene families regarding this matter was observed. It is however, important to be aware of the fact that there are several factors influencing transcript detectability; and mRNA degradation is only one of these factors, although an important one. In line with the supposition made by *Yang et al*, that certain transcripts decay faster or slower than others, the detectability for these categories was examined in the dataset in this study. Genes belonging to the categories 'protein biosynthesis', 'catabolism', and 'metabolism' were more easily detected than those in 'transcription, 'cell cycle', 'development', and 'signal transduction'. These results are in agreement with the categories that *Yang et al* concluded to be slow vs. fast decaying.

For the categories belonging to GO component, it was noted that 'microsome', 'endoplasmic reticulum', and 'mitochondrion' related transcripts were easier to detect than those related to 'nucleus', 'plasma membrane', and 'cytoskeleton'. This may be reflected by differences in the rate of mRNA degradation, but no comparative study, such as the one made by *Yang et al* on GO process, has been made on this subject. The same goes for GO function, although remarkable differences in transcript detectability could be observed. Whereas the functional categories 'oxidoreductase activity' and 'catalytic activity' were very easy to detect, 'transmembrane activity' and 'G-protein coupled receptor activity' were a lot harder. The latter categories, however, are known to be hard to detect at both transcript and protein levels, and this is probably due to the nature of these rather than an issue of fast degrading mRNAs.

5.3.3 Transcripts that are hard to detect in rat liver and hepatocytes

There are three proteins in rat liver, and five in rat hepatocytes that are practically never detected with transcript techniques. The fact that these proteins (6 in total) are detected with protein techniques (although the protein frequency is quite low) supports the hypothesis that these transcripts are subject to fast degradation after induction. Examining the properties of the six proteins allows for speculative explanations of the reasons they may decay faster. Referring to what was mentioned in the previous section - that transcripts that require a rapid induction are also caused to degrade quickly - could explain the low mRNA expression observed for at least three of the proteins: KCRB, GOGA2 and HSP72.

KCRB is a cytoplasmic protein that catalyzes the transfer of phosphate between ATP and various phosphate acceptors. It plays a central role in energy transduction in tissues with large, fluctuating energy demands, such as skeletal muscle, heart, brain, and spermatozoa. The protein is normally expressed in the outer medulla of the kidney (101). Interestingly, it is rapidly and specifically induced by estrogen in the uterus of the immature rat. The stimulation of mRNA levels is rapid (a 7- to 10-fold increase is detected 1-3 h after estrogen administration), but transient, as levels return to near control values by 6 h (122). GOGA2, an integral membrane protein that contributes to the maintenance of the Golgi structure, is a key protein in the mitotic golgi fragmentation process (123). During the course of metaphase, anaphase, and telophase GOGA2 is phosphorylated and dephosporylated in concert with the dissociation and reassociation of p115, a vesicle-tethering protein, with Golgi membranes. In addition, GOGA2 is a specific interacting partner of the small GTPase rab1b (124). HSP72 is a stress-induced chaperone that belongs to the heat shock protein family Hsp70. It is a chaperone that stabilizes preexistent proteins against aggregation and mediates the folding of newly translated polypeptides in the cytosol as well as within organelles (101).

The poor transcript detectability for SPTA2 may be due to the fact that this protein belongs to the GO process 'cytoskeleton' (101). Referring to Figure 6 in Section 4.3.2, there is a very poor detectability for transcripts in this group (29% are detectable at the level of transcription). SPTA2 interacts with calmodulin in a calcium-dependent manner and is candidate for the calcium-dependent movement of the cytoskeleton at the membrane (101). It belongs to the spectrin family, responsible for providing support for the plasma membrane to maintain cell shape. Proteins similar to spectrin and to its associated attachment proteins are present in the cortex of most of our cells (125).

The two remaining proteins, ACDSB and PCCB are both proteins of the mitochondrial matrix, and the motive for a fast degradation of these two is indefinite. ACDSB belongs to the acyl-CoA dehydrogenase family and catalyzes the first step in mitochondrial fatty acid beta oxidation system (126), whereas PCCB is a biotin-dependent enzyme that functions in the catabolism of branched-chain amino acids, fatty acids with odd-numbered chain lengths, and other metabolites (127).

5.4 Hypothesis III: Cytoplasmic proteins present a better correlation with mRNA than nuclear proteins

5.4.1 Differences in correlation patterns between the nucleus and cytoplasm

The correlation between mRNA and protein is much better in the cytoplasm than in the nucleus, and especially good is the correlation in the endoplasmic reticulum and Golgi apparatus. There are very few nuclear proteins compared to cytoplasmic proteins; thus, the correlation comparisons between these compartments are not quite analogous. The depletion of proteins in the categories ‘transcription’ and ‘protein biosynthesis’, and hence the difficulty to correlate them with their mRNA counterparts may be explained by (1) fast decaying transcripts, and (2) ‘translation on demand’.

5.4.1.1 Short burst production of transcription factors

It is not so surprising that both the *Fountoulakis* and the Swissprot/TrEMBL protein datasets are depleted in proteins belonging to the category ‘transcription’. As already mentioned, transcription factor mRNAs are fast decaying. The fact that transcription factors are induced rapidly is crucial for their function to act as “switches”. Again, the rate of mRNA turnover not only determines the rate of disappearance but also its induction, and mRNAs with short half-lives respond to changes in transcription more rapidly than those that are relatively stable (93). Thus, the explanation to the depletion of proteins belonging to this category, and the difficulty to make an mRNA-to-protein correlation analysis, is probably due to the fact that these types of transcripts are produced only in short bursts.

5.4.1.2 Protein biosynthesis genes may be subject to ‘translation on demand’

What may seem more surprising, is the fact that proteins related to ‘protein biosynthesis’ or ‘ribosome’ are depleted in the protein data sets. According to *Yang et al*, transcripts belonging to the category ‘protein biosynthesis’ are slow-decaying, which is also observed from the transcript detectability calculation in Section 4.2.4, which resulted in an average of 100% mRNA detectability (Figure 13). In addition, the yeast transcriptome analysis made by *Velculescu et al* in 1997, showed that the transcripts of the 137 genes encoding 78 ribosomal proteins belong to the most abundant mRNAs in the yeast cell (31).

One suggestion to the depletion of proteins in the category ‘protein biosynthesis’, and consequently the difficulty to correlate these to their corresponding mRNAs, is that this category is subject to so called “translation on demand”. In 2004, *Beyer et al* identified a set of modules that are regulated via “translation on demand”, which means that the cell constitutively maintains a sufficient level of mRNA, but blocks translation until the protein is actually needed (8). Such proteins are synthesized at low levels under standard conditions, but mRNA is present at reasonable amounts to allow for a quick response to take place. They suggested that modules involved in ‘cell rescue/defense’, ‘response to stimulus’, and ‘cellular communication/signal transduction’ will have a poor mRNA-to-protein correlation due to suppressed translation. This might be the case for the ‘protein synthesis’ category as well. The formation of functional ribosomes is a highly structured and complex process requiring the interplay of a large number of molecular reactions. Ribosome formation in eukaryotic cells involves the participation of three different RNA polymerases for the production of four different rRNAs and about 80 different ribosomal proteins (128).This complex formation of ribosomes suggests that the expression of ribosomal genes are subject to tight coordinate control, and in yeast, this regulation of expression occurs almost entirely at the level of transcription (129-130).

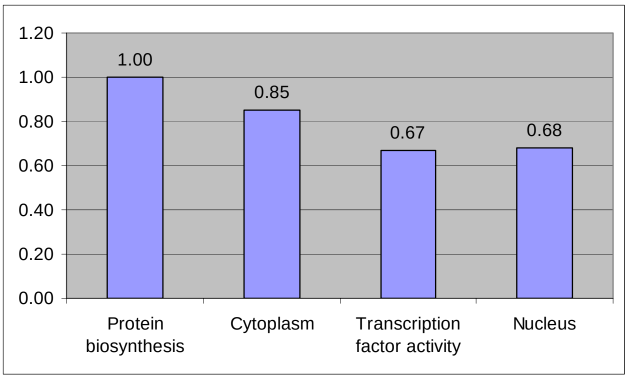


Figure 13: Transcript detectability for protein biosynthesis, cytoplasm, transcription factor activity and nucleus
Better transcript detectability is observed in the cytoplasm and for the cytoplasm-related process protein biosynthesis, than for the nucleus and transcription factor activity. The figure is a comprehended version of Figure 6 in Section 4.3.2.

5.4.2 Better mRNA-to-protein correlation in the ER and Golgi compared to the cytosol – a consequence of protective protein stabilization?

The correlation observed in the endoplasmic reticulum and the Golgi apparatus is considerably higher than in the cytosol. The majority of the genes (63% vs. 60%) have an ok to very good correlation (fall into correlation categories 1-3). Is this a coincidence or is it due to the nature of these compartments? One likely explanation is the protective actions against protein degradation that are undertaken upon protein entrance into the ER and Golgi. The cytosol is a reducing environment, where proteins are exposed to a lot of degradative enzymes and fluctuations in pH. In contrast, the ER and Golgi are relatively “safe” environments for proteins to reside. In the ER, the proteins are modified to become more stable; disulfide bonds are formed, proteins are glycosylated and proteins interact with chaperones. All these mechanisms help to stabilize the structure of the proteins in order to avoid degradation upon encounter of degradative enzymes and changes in pH. The reactions of disulfide bond formation and glycosylation are catalyzed by enzymes present in the ER lumen; enzymes that are absent in the cytosol. Besides functioning as protection, the attachment of oligosaccharides (glycosylation) on proteins serves other functions; it retains the proteins in the ER until they are properly folded and helps to guide them to the appropriate organelle by serving as a transport signal for packaging the proteins into appropriate transport vesicles (131). The retention of protein in the ER is also a result of the interaction with chaperones. Some proteins made in the ER are destined to function there. They are retained in the ER (and are returned to the ER when they escape to the Golgi apparatus) by KDEL, a carboxyl-terminal four-amino acid sequence called ER retention signal (132), which is recognized by a membrane-bound receptor protein in the ER and Golgi apparatus. Most proteins that enter the ER, however, are destined for other locations. Thus, the endoplasmic reticulum controls the quality of the proteins that it exports to the Golgi apparatus.

Upon entrance in the Golgi apparatus, the protein quality control continues. Many of the oligosaccharide groups that are added to proteins in the ER undergo further modifications in the Golgi apparatus. On some proteins, for example, complex oligosaccharide chains are created by a highly ordered process in which sugars are added and removed by a series of enzymes as the protein passes through the Golgi stack (131). Another possible explanation for the better mRNA-to-protein correlation in the Golgi deals with protein stabilization when proteins are sorted; one of the most important Golgi functions (131). Proteins that are destined for secretory vesicles have special surface properties that cause them to aggregate with one another under ionic condition. This aggregation is likely to increase the protein stability and hence, explain the better mRNA-to-protein correlation observed in this compartment.

5.5 Hypothesis IV: "The correlation between mRNA and protein is very poor in the mitochondrion"

Taken as a whole, the correlation in the mitochondrion and its subparts is poor, with the exception seen in the mitochondrial inner membrane. Although the correlation in the mitochondrion might not seem significantly poor (median correlation values in the range 3-5), the correlation distribution observed is somewhat misleading. In a lot of cases, a good correlation value is obtained due to poor detectability at both transcript and protein levels. Some possible explanations for the lack of correlation may be transient expression of proteins, rapid degradation of mitochondrial proteins, and the fact that the proteomic rat liver analyses have been carried out on rats at steady-state.

5.5.1 Mitochondrial biogenesis - a product of complex interactions between the nuclear and mitochondrial genomes

According to a recent estimate, about 1,000 polypeptides exist in mitochondria (133). Mitochondrial proteins are encoded by the mitochondrial genome and the nuclear genome. The mitochondrial DNA, mtDNA, is replicated and transcribed semi-autonomousy and is different from nuclear DNA in that it contains no introns, has no protective histones or non-histone proteins (114). Thus, it is more susceptible to damage. Another difference is the fact that mtDNA undergoes continuous replication, even in non-dividing cells, and lacks efficient repair mechanisms compared to nuclear mechanisms (134). Nuclear-encoded proteins are imported into the mitochondria where they carry out their function. Mitochondrial biogenesis requires protein targeting to four compartments: the outer membrane, the intermembrane space, the inner membrane, and the matrix. Nuclear gene products are translated in the cytosol as precursor proteins with intrinsic targeting signals. These precursor proteins interact with molecular chaperones that direct them to the import machinery of the outer membrane (Tom complex). The precursor is unfolded and transferred through the outer membrane, across the intermembrane space to the mitochondrial inner membrane translocases (Tim complex) (135). Translocation of proteins across the mitochondrial membranes is illustrated in Figure 14.

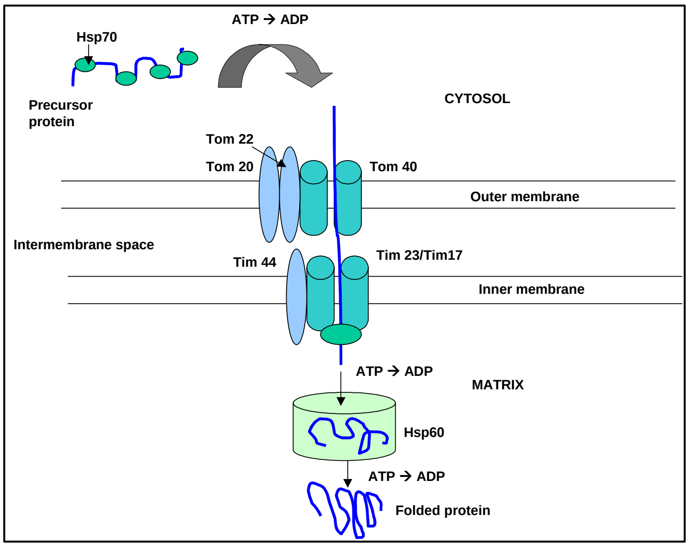


Figure 14: Protein translocation across membranes
Import of nuclear-encoded mitochondrial preproteins is mediated by translocation complexes in the outer and inner mitochondrial membranes. Initially, receptors of the TOM complex, the translocation machinery in the outer membrane, recognize precursor proteins. The TOM complex mediates insertion of precursors into the outer membrane and import of some precursors into the intermembrane space. For import of preproteins in the inner membrane and into the matrix, the TOM complex cooperates with translocases of the inner membrane, TIM complexes (135).

5.5.2 Poor correlation – a dual effect of transiently expressed proteins and rat liver at steady state

Translocation of mitochondrial precursors is an energy-dependent process that is assisted by heteromeric translocation processes in both membranes. One possible explanation for the poor mRNA-to-protein correlation observed in this organelle may be that the mitochondrial surface has a destabilizing effect on the imported nuclear-encoded precursor proteins. Another potential reason is the fact that a lot of proteins reside transiently in this organelle. Examples of transiently existing proteins are components of the process of oxidative phoshorylation; other enzymes and components of the Kreb’s cycle, proteins involved in mtDNA replication, transcription, and translation, and proteins involved in apoptosis (136). These proteins are depleted in both the *Fountoulakis* and the Swissprot/TrEMBL datasets, which suggests that these proteins are expressed transiently.

As mentioned in the introductory part, the correlation between mRNA and protein in this analysis is of a qualitative and local character at steady-state. The fact that the rat liver is examined at steady-state may also be an explanation to the poor correlation observed. Upon exercise, signaling pathways are activated, which initiate the activation of transcription factors that increase the production of mRNA from nuclear and mitochondrial DNA. Perhaps, the correlation results would have looked different if the

Fountoulakis proteome study would have been made on activated rats, with activated mitochondria.

5.5.3 The role of protein degradation in mitochondrial function and biogenesis

It has been known for a long time that mitochondria contain their own protein degradation system (137). Degradation of mitochondrial proteins serves different purposes such as: 1) protection – removal of polypeptides that are potentially harmful to the cell, 2) regulation – controlling the concentration of enzymes or regulatory proteins, and 3) metabolism – release of amino acids to be used for other purposes (137). Interestingly, half-lives are widely divergent among proteins of every mitochondrial compartment (138, 139). In 1979, *Kalnov et al* reported that one-third to one-half of the proteins synthesized in isolated yeast mithochondra was degraded with a half-life of about 35 minutes (140). Instability of a subset of mitochondrial translation products has also been observed in rat liver mitochondria (141). Thus, protein degradation and rapid turnover of imported proteins could explain the lack of mRNA-to-protein correlation observed in mitochondria.

5.5.4 Surprisingly good correlation in 'electron transport' and 'mitochondrial inner membrane'

The enzymes that carry out electron transport and oxidative phosphorylation are encoded either by the mitochondrial or the nuclear genome. These enzymes constitute about 50% of the protein content of the mitochondrial inner membrane (142), and are a major target for proteolysis (137). Turnover of subunits is necessary to prevent accumulation of single subunits and sub-complexes in the mitochondrial inner membrane, which may disturb assembly processes or change the properties of the inner membrane (143). Thus, many subunits of the mitochondrial inner membrane enzyme complexes are quickly degraded, which is quite surprising considering the relatively good mRNA-to-protein correlation found in the mitochondrial inner membrane (median 2.5) and in the category 'electron transport' (median 2.64). One possible explanation for the good correlation observed, is that proteins in electron transport as well as oxidative phosphorylation pathways are known to be high-abundant proteins (104). In addition, from the transcript detectability analysis in section 4.2.4, it can be seen that genes belonging to 'electron transport' are easily detected at the level of mRNA. Thus, as a consequence of good detectability with transcript and protein techniques, the correlation obtained is fairly good.

5.6 Hypothesis V: ”Genes belonging to the category 'metabolism' are well correlated at the mRNA and protein levels”

5.6.1 'Metabolism' genes are easily detected at the transcript level

Common to all the transcripts in the category 'metabolism' and its subcategories, is that they are easily detected. This could be observed in Figure 10, which showed transcript detectability above 80% for all the metabolism subcategories. Thus, a good correlation value is an effect of successful detection with both transcript and protein techniques. The fact that the protein and mRNA expression data are measured in the liver, which serves a crucial role in the body's metabolism, may partly explain this. Metabolism is the sum of all chemical reactions involved in maintaining the living state of the cells (95). The major metabolic reactions are those involving the breakdown of the three basic biomolecules to obtain energy: carbohydrates, lipids, and proteins (Figure 15).

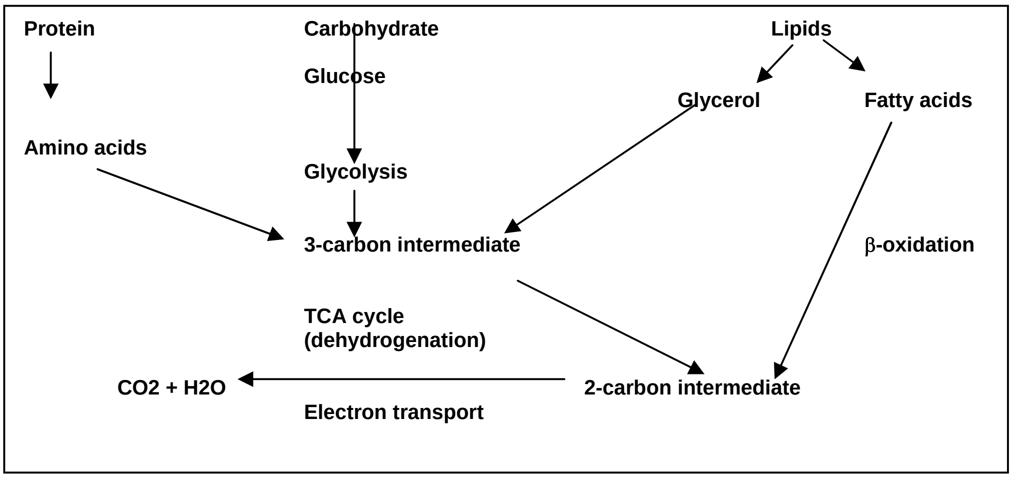


Figure 15: Outline of metabolism for the three basic biomolecules: proteins, carbohydrates, and lipids
Proteins, carbohydrates, and lipids are eventually converted to three-carbon compounds. The carbon skeletons derived enter the TCA cycle (directly or indirectly) where they are oxidized to CO2, resulting in energy production to support the life of the cell.

5.6.2 The mRNA-to-protein correlation differs among 'metabolism' subgroups

Although the transcript detectability is high for all 'metabolism' subcategories, the protein detectability differs between categories. Consequently, this has an effect on the mRNA-to-protein correlation observed for each subcategory. Some patterns can be distinguished; the correlation is better in amino acid and nitrogen metabolism compared to carbohydrate and fatty acid metabolism.

5.6.2.1 Poor correlation for carbohydrate and fatty acid metabolism

According to *Hood et al* (2002), the correlation between mRNA and protein is good for proteins belonging to the category 'carbohydrate metabolism'. In contrast, the observation in this analysis is that the correlation is poor. This may be a consequence of the fact that in this study, the protein expression was measured at steady state. Perhaps the situation would have been different if protein and mRNA expression measurements had been on active rats. The liver is the storage site for the essential carbohydrate glucose, which plays a central role in the whole body's metabolism. Besides its role in supporting CNS function, it provides signals to the endocrine system, which regulates overall metabolic activity. Glucose is stored in the form of glycogen to meet the needs during fasting or when extra fuel is needed as during exercise and intense muscle activity (131).

Fat is synthesized in the liver, but a healthy liver does not store fat. Instead, newly synthesized fat is converted to a transport form known as very low density lipoprotein, VLDL. The VLDLs are released from the liver to the blood, transporting the triglyceride for storage in various tissues of the body, primarily adipose (fat) tissue and, to a lesser extent, muscle (144). Fatty acid breakdown occurs within the mitochondrion, which may explain the poor correlation observed for this category (median correlation 3.57 and 36% in correlation categories 1-3.

5.6.2.2 Good correlation for ‘amino acid metabolism’ and ‘nitrogen metabolism’

Because of its very rich blood supply, the liver has access to circulating amino acids. Free amino acids are used for two purposes: 1) supporting the synthesis of proteins needed by the liver to maintain its own structures and processes, and 2) synthesis of additional glucose for use by other tissues (gluconeogenesis); a process that is unique to the liver (95)**.** While carbohydrate and fat can be stored by cells, there is no storage form for amino acids. They are either converted into protein or into other compounds. The liver is also central to nitrogen metabolism, since this is where the synthesis of urea occurs. The simplest way to catabolize amino acids would be to remove the amino groups directly and release the nitrogen as free ammonia (NH_3). However, since ammonia is extremely toxic, the amino group is converted to urea, which is a water-soluble non-toxic compound. Once synthesized, urea is released from the liver to the blood and transported to the kidneys where it is concentrated and released from the body in the urine (125). As already mentioned, CPSM, a protein with a nearly perfect mRNA-to-protein correlation plays an important role in removing excess ammonia from the cell.

With a median correlation of 1.93 and 1.85, and 57% vs. 60% in correlation categories 1-3, the mRNA-to-protein correlation for the categories ‘amino acid metabolism’ and ‘nitrogen metabolism’ can be regarded as good. Interestingly, the majority of the proteins within the category ‘amino acid metabolism’ belong to the family aromatic amino acids. It is somewhat hard to speculate on the reasons for the very good correlation observed for this subcategory. One possible explanation, however, may be the mode of degradation for different amino acids. One can make a distinction between glycogenic and ketogenic amino acids depending on their degradation pathways (145). The majority of all amino acids are glycogenic, which means that the carbon skeletons are degraded to Krebs cycle intermediates. This means that they can give rise to blood glucose via the gluconeogenic pathway. In contrast, ketogenic amino acids are degraded to compounds such as acetoacetate and acetyl-CoA. Aromatic amino acids are degraded into both Krebs cycle acids and to acetyl-CoA, and are thus *both* glycogenic and ketogenic.

Protein synthesis and degradation are regulated by hormones such as: insulin, IGF-1 and growth hormones. In addition, both in vivo (146) and in vitro studies conducted in cultures of isolated rat hepatocytes (147) have shown that amino acids stimulate liver protein synthesis. *Jaleel et al* (2004) identified 16 specific proteins whose synthetic rates were enhanced by increased amino acid concentration. These proteins were involved in ‘translation initiation’, ‘protein folding and modification’, and ‘transport’. It remains to be determined whether synthesis of all liver proteins or only certain specific proteins are stimulated by amino acids

6 Conclusions and future studies

This section summarizes the conclusions drawn from the hypotheses evaluations, and elaborates on future directions; what experiments could be done next to improve the mRNA-to-protein correlation outcome, how could this study be used in the future, and why is this study important for future research.

6.1 Conclusions

To start with, a brief overview on the results of each of the hypotheses is given:

Hypothesis I: *"High-abundant proteins present a better mRNA-to-protein correlation than low-abundant proteins"*

- The high abundant proteins are defined as those with a protein frequency above 50%. All of these are very well correlated with their mRNA counterparts, hence Hypothesis I is true (cannot be falsified) for the data in this study.

Hypothesis II: *"Transcripts that are not detectable with Affymetrix chips but are detectable at the protein level have a shorter mRNA half-life"*

- Some of the genes detectable at the transcript level but not at the protein level belong to transcript categories known to be rapidly induced and rapidly degraded. Others are present in organelles known to have poor transcript detectability such as the cytoskeleton. These results may indicate that the 2nd hypothesis is valid, however further investigation is needed to evaluate this concept. In addition, the process of mRNA degradation needs to be better understood.

Hypothesis III: *Cytoplasmic proteins present a better correlation with mRNA than nuclear proteins*

- Taken as a whole, the mRNA-to-protein correlation is not significantly better in the cytoplasm compared to the nucleus. However, for the endoplasmic reticulum and the golgi apparatus, a very good correlation is observed compared to the cytosol and the nucleus. Thus, Hypotesis III is true (cannot be falsified) for certain subcategories within the cytoplasm.

Hypothesis IV: *"The correlation between mRNA and protein is very poor in the mitochondrion"*

- Overall, the correlation in the mitochondrion is poor apart for the mitochondrial inner membrane and proteins associated with the category 'electron transport'. The poor correlation observed for mitochondrion validates Hypothesis IV.

Hypothesis V:*"Genes belonging to the category 'metabolism' are well correlated at the mRNA and protein levels"*

- Although the transcript detectability is high in general, the correlation between mRNA and protein differs between metabolism subcategories. The correlation is very good for nitrogen and amino acid (especially aromatic amino acid) metabolism, whereas it is poor in carbohydrate and fatty acid metabolism. Thus Hypothesis V is rejected for the whole metabolism category, but is applicable after division into subcategories.

The aim of this project was to categorize *when*, *how,* and *if* transcripts and proteins should be analyzed in an integrated fashion. From this analysis, it is apparent that some categories have a better mRNA-to-protein correlation, and that some correlation patterns can be distinguished. Integrating knowledge from genomic technologies is easier said than done. The limited predictive value (mRNA to protein) is explained partly by biological differences between the processes of transcription and translation, and partly by experimental challenges. Biological differences result from RNA splicing that is not detectable by microarrays, differential RNA and protein turnover, post-translational modifications, and proteolytic processing events. On the experimental front, challenges in experimental design and data interpretation, as well as technological limitations, contribute to some of the differences observed. Thus, although the idea is appealing, it is not yet possible to predict protein expression from only microarray experiments. This study, however, opens up an interesting research field and offers a framework for future studies of this kind.

6.2 Future studies

So far, the vast majority of the mRNA-to-protein correlation studies have been made on yeast. This study, for the first time is made on the eukaryotic organism rat, however in the future it should be interesting to expand it to humans. In order for this analysis to become valuable and useful in the future, some critical aspects need to be considered: (1) determination of protein and mRNA turnover, (2) measurement of time-related fluctuations in gene and protein expression, (3) release of new protein data, (4) standardized methods for integrating protein and mRNA data when performing concordance analyses of this kind, (5) proceed with limited case studies, and (6) finding gene sequence elements common among poor correlation categories.

6.2.1 Determination of protein and mRNA turnover

The degradation rates of both proteins and mRNAs are essential parameters in evaluating the concordance of transcript and protein techniques. Today, too little is known about what factors influence these processes. The development of approaches to determine mRNA and protein decay rates should lead to a better understanding of the relationship between transcriptome and proteome.

6.2.2 Timescales of ‘omics’ events

A confounding aspect when relating gene and protein expression data is the time-displacement that exists between gene and protein events and their end-points. For example, if the action of a stimulus is at the genetic level, it will take a certain amount of time before the associated protein synthesis occurs. In addition, the duration of the gene events and protein events may be very different (148). Figure 16 illustrates a theoretical view of the problems that can occur when correlating mRNA expression to protein abundance. To get a better understanding of the absence vs. presence of correlation between mRNA and protein, the issue of the gene and protein expression time courses requires further investigation. A future direction could be to measure these time-related fluctuations in detail, which should lead to better understanding of the correlation between mRNA and protein. Today this is in fact possible, and methodologies exist to perform such measurements, however the cost is too high for routine analyses.

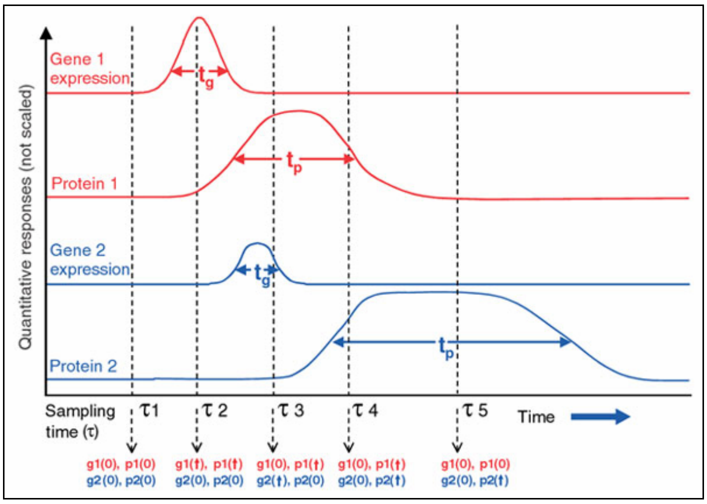


Figure 16: Time courses for two hypothetical gene-protein couples that are up-regulated following a system stimulus

After stimulation at the genetic level, it will take some time before the corresponding protein synthesis occurs. The duration of gene events (t_g) and protein events (t_p) may also be very different. Thus, correlating mRNA and protein at time τ may be misleading. (Figure 16 is adapted from reference 148)

6.2.3 Release of new protein data - the HUPO International mouse and rat proteome project, MRPP

As already mentioned, the largest complication in this analysis is the limited amount of protein data. Although the *Fountoulakis* study is the first large-scale analysis of the rat liver proteome, it is quite old and far from complete, as is the Swissprot/TrEMBL database. Consequently the results should be more complete and definitive with larger proteomics datasets. An important initiative that is likely to have a big impact in future proteomics and correlation studies is the HUPO international mouse and rat proteome project, MRPP, established by the Human proteome project, HUPO, in 2004 (149). The goal of this initiative is to map, characterize, and localize the total proteome in normal and diseased mouse and rat liver. Since the rat and mouse are the most versatile and most genetically amenable mammalian systems in biomedical research, MRPP will serve as a ‘gold standard’ for proteomic analysis of human organs and tissues. The aims are to 1) obtain a comprehensive functional map of the liver, and 2) generate a tool to accelerate the development of diagnostics and therapeutics aimed at liver diseases. MRPP is expected to be finished in 2008. Hence, it would be motivating to perform this type of mRNA-to-protein correlation analysis at that time.

6.2.4 The need for standardized methods to compare high-throughput mRNA expression data and protein abundance data

This analysis was of a qualitative character. Beforehand, this was not the actual study motive, but instead a result of the difficulty in correlating protein and mRNA abundance data. Therefore, it will be central for future research to define a common policy for making correlation analyses of this kind. Although *Greenbaum*, *Jansen,* and *Gerstein* have outlined a formalism for merging and scaling many different gene expression and protein abundance data sets into a comprehensive reference set, they also denote this as being a problem. Thus, it is essential to gather enough data sets from proteomic and transcript analyses and detail patterns in the correlation of mRNA and protein expression in order to develop strategies by which to organize the correlated expression data sets.

When obtaining a reliable and consistent mRNA-to-protein correlation platform, studies of this kind should be better accepted.

6.2.5 Proceed with limited case studies of disease tissues

A disease state is accompanied by significant or subtle changes in the expression of many genes and/or their protein products. It would be interesting to compare the up vs. down regulated genes by both transcript and protein techniques, and determine the congruence between these techniques for a certain disease state or condition. In that case it should be possible to find disease marker genes and assess the mode of detection by transcript and protein techniques, and whether any of the two could substitute for one another. From this analysis it is observed that the correlation between mRNA and protein is much better in some Gene Ontology categories concerning process or compartment, however it is not clear enough to eliminate neither transcript nor protein techniques in the detection procedure. It would be of interest to do a limited study on a human disease state, identify the "druggable" genes, and investigate whether these are better detected with either transcript or protein techniques and if one or the other could be excluded.

6.2.6 Targeted gene sequence experiments to discover the reason for lack of correlation in specific categories.

One likely reason for the lack of correlation between mRNA and protein is post-transcriptional regulation. Thus, a good idea could be to proceed with more specifically targeted gene sequence experiments. One such direction is to look for conserved gene sequence features among those genes showing remarkable differences between mRNA and protein abundance ratios. AlignACE is a publicly available computational tool which identifies potential cis-regulatory sequence motifs common to a set of gene sequences (150). *Gygi* and *Hood et al* (2002) applied AlignACE to gene sequences for protein-synthesis genes and rRNA-processing genes - genes which showed a very poor mRNA-to-protein correlation in their study (2). They found a highly conserved sequence between these genes previously described as a ribosomal processing element (2). A suggestion to improve or continue this analysis, is to apply the AlignACE tool to the genes belonging to GO categories with poor mRNA-to-protein correlation such as 'mitochondrion', 'energy generation', 'nucleus', and 'transcription'.

7 Acknowledgements

First and foremost, I would like to thank Magnus L. Andersson; my supervisor at AstraZeneca for guidance, friendship and enthusiasm throughout this project, and for giving me a lot of responsibility and flexibility working with this project. I would also like to thank Chrishopher Southan, my scientific reviewer, for always taking his time to answer my questions and for showing genuine enthusiasm and interest in this project. I want to thank everyone else at the department of Molecular Pharmacology at AstraZeneca, Mölndal for a pleasant working atmosphere, especially Lars Stålhult and Lisa Öberg, for creating an enjoyable environment within the Bioinformatics group. Eventually I would like to thank Torgny Fornstedt at Uppsala University for advice and for promptly responding to questions and emails.

8 References

1. Bro C., Nielsen J. *et al*. (2003). Transcriptional, proteomic, and metabolic responses to lithium in galactose-grown yeast cells. *The journal of biological chemistry.* **278**(34): 32141-32149.
2. Gygi S. P., Hood L. *et al.* (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Molecular and cellular proteomics.* **1**(4): 323-333.
3. Gygi S. P. Aebersold R. *et al.* (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology.* **19**(3): 1720-1730.
4. Hood L., Aebersold R. *et al.* (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science.* **292**: 929-934.
5. Futcher B., Garrels J. I. *et al.* (1999). A sampling of the yeast proteome. *Molecular and cellular biology.* **19**(11): 7357-7368.
6. Greenbaum D., Gerstein M. *et al*. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology.* **4**(117): 1-8.
7. Greenbaum D., Gerstein M., *et al.* (2002). Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics.* **18**(4): 585-596.
8. Beyer A., Hollunder J. *et al.* (2004). Post-transcriptional expression in the yeast Saccharomyces cerevisiae on a genomic scale. *Molecular and cellular proteomics*. **3**(11): 1083-1092.
9. Washburn M. P., Koller A. *et al.* (2003). Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences.* **100**(6): 3107-3112.
10. Greenbaum D., Gerstein M. *et al.* (2001). Interrelating different types of genomic data, from proteome to secretome: 'oming in on function'. *Cold spring harbor laboratory press*. **11**: 1463-1468.
11. Jansen T., Greenbaum D. *et al.* (2002). Relating whole-genome expression data with protein-protein interactions. *Cold spring harbor laboratory press.* **12**:37-46.
12. Tew K. D., Schmidt D. E. *et al.* (1996). Glutathione-associated enzymes in the human cell lines of the National Cancer Institute Drug Screening Program. *American Society for Pharmacology and Experimental Therapeutics.* **50**(1): 149-159.
13. Hibbs K., Skubitz K. M. *et al*. (2004). Differential gene expression in ovarian carcinomas: identification of potential biomarkers. *American journal of pathology.* **165**(2): 397-414.
14. Celis J. E., Orntoft T. *et al.* (2003). Integrating proteomic and functional genomic technologies in discovery-driven translational breast cancer research. *Molecular and cellular proteomics*. **2**(6): 369-377.
15. Huber M., Bahr I. *et al.* (2004). Comparison of proteomic and genomic analyses of the human breast cancer cell line T47D and the antiestrogen-resistant derivative T47D-r. *Molecular and cellular proteomics.* **3**(1): 43-55.
16. Grate L. R., Mian I. S. *et al*. (2003). Integrated analysis of transcript profiling and protein sequence data. *Mechanisms of ageing and development.* **124**: 109-114.
17. Chen S-T., Juan H-F. *et al.* (2002). Biomic study of human myeloid leukemia cells differentiation to macrophages using DNA array, proteomic, and bioinformatics analytical methods. *Electrophoresis.* **23**: 2490-2504.
18. Chen G, Beer D. G. *et al.* (2002). Discordant protein and mRNA expression in lung adenocarcinomas. *Molecular and cellular proteomics.* **1**(4): 304-313.
19. Orntoft T., Celis J. E. *et al.* (2002). Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas. *Molecular and cellular proteomics.* **1**(1): 37-45.
20. Kern W., Kohlmann A. *et al.* (2003). Correlation of protein expression and gene expression in acute leukemia. *Cytometry part B (Clinical cytometry).* **55B**:29-36.
21. Gabor Miklos G. L., Maleszka R. (2001). Protein functions and biological contexts. *Proteomics.* **1**: 169-178.

22. Lander E. S., Linton L. M. *et al*. (2001). International Human Genome Consortium. Initial sequencing and analysis of the human genome. *Nature*. **409**, 860-921.
23. Venter J. C., Adams A. D. *et al*. (2001). The sequence of the human genome. *Science* **291**: 1304-1351.
24. –Omes and –omics glossary: evolving terminology for emerging technologies: http://www.bioon.com/book/biology/genomicglossaries/omes.asp.htm.
25. Dennis C. (2002). Information overload. *Nature*. **417**(6884): 14.
26. Whittaker P. A. (2003) What is the relevance of bioinformatics to pharmacology? *Trends in Pharmacological Sciences*. **24**(8):434-9.
27. Pratt J.M.,Petty J. *et al*. (2002) Dynamics of protein turnover, a missing dimension in protein turnover. *Molecular & Cellular Proteomics*.**1**(8):579-91.
28. Mooser V., Ordovas J. M. (2003). 'Omic' approaches and lipid metabolism: are these new technologies holding their promises? *Current Opinion in Lipidology*. **14**(2):115-9.
29. Greenbaum D., Gerstein M., *et al.* (2002). Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics.* **18**(4): 585-596.
30. Watkins SM, German JB. (2002). Toward the implementation of metabolomic assessments of human health and nutrition. *Current Opinion in Biotechnology*. **13**(5):512-6.
31. Velculescu V.E, Zhang L. *et al*. (1997). Characterization of the yeast transcriptome. *Cell*. **88**(2):243-51.
32. de Hoog C. L., Mann M. (2004). Proteomics. *Annu. Rev. Genomics. Hum. Genet.* **5**: 267-93.
33. Polyak K., Riggins G. J. (2001). Gene discovery using the serial analysis of gene expression technique: implications for cancer research. *Journal of Clinical Oncology*. **19**(11):2948-58.
34. Lockhart DJ, Winzeler EA. (2000). Genomics, gene expression and DNA arrays, *Nature*. **405**(6788):827-36.
35. Liang P. and Pardec A.B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science.* **257**:967-971.
36. Lashkari D.A, DeRisi J.H. *et al*. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences USA* **94**:13057-13062.
37. Shalon D, Smith S.J. *et al*. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research.* **6**:639-645
38. Velculescu V.E, Zhang L. *et al*. (1995). Serial analysis of gene expression. *Science.* **270**:484-487.
39. Hegde P. S., White I. R. *et al*. (2003). Interplay of transcriptomics and proteomics. *Current opinion in Biotechnology*. **14**(6):647-51.
40. Zhou Y, Abagyan R. (2003). Algorithms for high-density oligonucleotide array. *Current opinion in drug discovery & development* **6**(3):339-345.
41. Lockhart D. J., Dong H. *et al*. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology.* **14**(13):1675-80.
42. Haverty P. M., Hsiao L. L. *et al*. (2004). Limited agreement among three global gene expression methods highlights the requirement for non-global validation. *Bioinformatics*. **20**(18):3431-41.
43. http://www.affymetrix.com.
44. Lipshutz R.F.S., Lockhart D.J. *et al*. (1999). High density synthetic oligonucleotide arrays *Nature Genetics*. **21**:20-24.
45. Wilkins M. R., Sanchez J. C. *et al*. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechology and Genetic Engineering Reviews*. **13**:19-50.
46. Anderson N. L., Anderson N. G. (1998). Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*. **19**: 1853-1861.
47. Honore B., Ostergaard M. *et al.* (2004). Functional genomics studied by proteomics. *Bioessays.* **26**(8): 901-915.

48. Tonge R., Shaw J. *et al.* (2001). Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics.* **1**(3): 377-396.
49. Zhou G., Zhao Y. *et al.* (2002). 2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers. *Molecular and cellular proteomics.* **1**(2): 117-124.
50. Zhou H., Aebersold R. *et al*. (2002). Quantitative protein profiling using two-dimensional gel electrophoresis, isotope-coded affinity tag labeling, and mass spectrometry. *Molecular and Cellular Proteomics*. **1**(1): 19-29.
51. Garvik B. M., Yates J. R. *et al*. (1999). Direct analysis of protein complexes using mass spectrometry. *Nature biotechnology.* **17**: 676-682.
52. Klose J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik.* **26**(3): 231-43.
53. O'Farell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *Journal of biological chemistry.* **250**: 4007-4021.
54. Gygi S. P., Aebersold R. *et al.* (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proceedings of the National Academy of Sciences.* **97**(17): 9390-9395.
55. Mann M., Hendrickson R. C. *et al*. (2001). Analysis of proteins and proteomes by mass spectrometry. *Annual Review of Biochemistry*. **70**:437-73.
56. Li Y. L., Gross M. L. (2004). Ionic-liquid matrices for quantitative analysis by MALDI-TOF mass spectrometry. *Journal of the American Society for Mass Spectrometry*. **15**(12):1833-7.
57. Aebersold R., Mann M. (2003). Mass spectrometry-based proteomics. *Nature*. **422**(6928):198-207.
58. Edgar R., Lash A. E. *et al*. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*. **30**(1):207-10.
59. Barrett T., Suzek T. O. *et al*. (2005). NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Research*. **33**: 562-6.
60. Prince J. T., M. W. Carlson. *et al.* (2004). The need for a public proteomics repository. *Nature biotechnology.* **22**(4): 471-472.
61. Fountoulakis M., Suter L. *et al.* (2002). The rat liver mitochondrial proteins. *Electrophoresis.* **23** : 311-328.
62. Fountoulakis M., Suter L. (2002). Proteomic analysis of rat liver. *Journal of chromatography B*. **782**: 197-218.
63. Fountoulakis M., Berndt. P. *et al.* (2001). Two-dimensional database of mouse liver proteins. An update. *Electrophoresis.* **22**(9): 1747-63.
64. Fountoulakis, M., Suter L. *et al* (2000). Two-dimensional database of mouse liver proteins. Changes in hepatic protein levels following treatment with acetaminophen or its non-toxic regioisomer 3-acetamidophenol. *Electrophoresis* **21**, 2148–2161.
65. Fouontoulakis M., Hardmeier R. *et al.* (1999). Rat brain proteins : two-dimensional protein database and variations in the expression level. *Electrophoresis.* **20**(18): 3572-9.
66. Mann M., Andersen J. S. *et al*. (2002). Directed proteomic analysis of the human nucleolus. *Current biology.* **12**: 1-11.
67. Washburn M. P., Yates J. R. *et al*. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology.* **19**: 242-247.
68. Ruse C. I., Kinter M. *et al.* (2004). Integrated analysis of the human cardiac transcriptome, proteome and phosphoproteome. *Proteomics.* **4**: 1505-1516.
69. Orchard S., Apweiler R. *et al*. (2003). The proteomics standards initiative. *Proteomics*, **3**(7):1374-6.
70. Taylor C. F., Kirby P. D. *et al.* (2003). A systematic approach to modeling, capturing and disseminating proteomics experimental data. *Nature biotechnology.* **21**: 247-54.
71. Gerstein M., Weissman S. M. *et al.* (2001). Genomic and proteomic analysis of the myeloid differentiation program. *Blood.* **98**(3): 513-524.
72. Hood L., Collins S. J. *et al*. (2004). Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Molecular and cellular proteomics*. **3**(10): 960-969.

73. Hatzimanikatis V., Choe L. H. *et al.* (1999). Proteomics: theoretical and experimental considerations. *Biotechnology progress.* **15**: 312-318.
74. Hatzimanikatis V. (1999). Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metabolic engineering.* **1**: 275-281.
75. Hood L., Baliga N. S. *et al.* (2002). Coordinate regulation of energy transduction modules in Halobacterium sp. analyzed by a global systems approach. *Proceedings of the National Academy of Sciences.* **99**(23): 14913-13918.
76. Anderson L., Seilhamer J. (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis.* **18**: 533-537.
77. Izziotti A., De Flora S. *et al.* (2004) Proteomic analysis as related to transcriptome data in the lung of chromium(VI)-treated rats. *International journal of oncology.* **24**: 1513-1522.
78. Cardozo A. K., Orntoft T. *et al* (2003). Gene microarray study corroborates proteomic findings in rodent islet cells. *Journal of proteome research.* **2**: 553-555.
79. Corvera S., Kahn C. R. *et al.* (2004). Role of insulin action and cell size on protein expression patterns in adipocytes. *The journal of biological chemistry.* **279**(30): 31902-31909.
80. Ruepp S. U., Tonge R. P. *et al.* (2002). Genomics and proteomics analysis of acetaminophen toxicity in mouse liver. *Toxicological sciences.* **65**: 135-150.
81. Kawamoto S., Matsumoto Y. *et al.* (1996). Expression profiles of active genes in human and mouse livers. *Gene.* **174**: 151-158.
82. Franzén B., Duvefelt K. *et al.* (2003). Gene and protein expression of human cerebral endothelial cells activated with tumor necrosis factor α. *Molecular brain research.* **115**: 130-146.
83. Gu Jang W., Soon Kim H. *et al.* (2004). Analysis of proteome and transcriptome of tumor necrosis factor α stimulated vascular smooth muscle cells with or without alpha lipoic acid. *Proteomics.* **4**: 1-11.
84. Celis J. E., Orntoft. T. F. *et al.* (2000). Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *Federation of European biochemical societies.* **480**: 2-16.
85. Kleno T. G., Kiehr B. *et al.* (2004). Combination of 'omics' data to investigate the mechanism(s) of hydrazine-induced hepatotoxicity in rats and to identify potential biomarkers. *Biomarkers.* **9**(2): 116-138.
86. Reilly D. F., Fitzgerald D. J. *et al.* (2004). Integration of proteomics and genomics in platelets: a profile of platelet proteins and platelet-specific genes. *Molecular and cellular proteomics.* **3**(2): 133-144.
87. de Cremoux P., Martin E. C. *et al.* (1999). Quantitative PCR analysis of c-erb B-2 (HER2/neu) gene amplification and comparison with p185(HER2/neu) protein expression in breast cancer drill biopsies. *International journal of cancer.* **83**: 157-161.
88. Korf B. R., Slavc I. *et al.* (1990). Myc gene amplification and expression in primary human neuroblastoma. *Cancer research.* **50**: 1459-1563.
89. Alper H., Moxley J. *et al*. (2004). Exploiting biological complexity for strain improvement through systems biology. *Nature biotechnology*. **22**: 1261-1267.
90. Fountoulakis M., Suter L. *et al.* (2002). Modulation of gene and protein expression by carbon tetrachloride in the rat liver. *Toxicology and applied pharmacology.* **183**: 71-80.
91. Ge H., Vidal M. *et al.* (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends in genetics.* **19**(10): 551-560.
92. Tew K. D., Schmidt D. E. *et al*. (1996). Glutathione-associated enzymes in the human cell lines of the National Cancer Institute Drug Screening Program. *American Society for Pharmacology and Experimental Therapeutics.* 50(1): 149-159.
93. Guhaniyogi J., Brewer G. (2001). Regulation of mRNA stability in mammalian cells. *Gene.* **265**: 11-23.
94. Ross J. (1995). mRNA stability in mammalian cells. *Microbiological reviews.* **59**(3): 423-450
95. Alberts B.,Johnson A. *et al*. Molecular biology of the cell. 4th edition. New York: Garland Publishing; 2002.
96. Lopez M. F. (2000). High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis*. **21**: 3427-3440.

97. Tayor R. S., Howell K. E. *et al*. (1997). Two-dimensional mapping of the endogenous proteins o the rat hepatocyte Golgi complex cleared of proteins in transit. *Electrophoresis.* **18**: 2601-2612.
98. Taylor R. S., Yate J. R. *et al*. (2000). Proteomics of rat liver Golgi complex: minor proteins are identified through sequential fractionation. *Electrophoresis.* **21**: 3441-3459.
99. Cronshaw J. M., Zhang W. *et al*. (2002). Proteomics analysis of the mammalian nuclear pore complex. *Journal of cell biology.* **158**: 915-927.
100. Gene Logic web page: http://www.genelogic.com
101. Swissprot/TrEMBL in E-lab – AstraZeneca sequence analysis portal.
102. Fountoulakis M., Krapfenbauer K. *et al.* (2001). Changes in the brain protein levels following administration of kainic acid. *Electrophoresis*. **22**(10): 2086-91.
103. Fountoulakis M., Krapfenbauer K. *et al.* (2001). Changes in the levels of low abundance brain proteins induced by kainic acid. *European journal of biochemistry.* **268**: 3532-3537.
104. Zeng R., Zhou H., *et al*. (2004). A high-throughput approach for subcellular proteome: identification of rat liver proteins using subcellular fractionation coupled with two-dimensional liquid chromatography tandem mass spectrometry and bioinformatics analysis. *Molecular and cellular proteomics.* **3(**5): 441-455.
105. Gene catalogue help pages in E-lab – AstraZeneca sequence analysis portal.
106. Simpson A., Robinson C. Mastering Access 2000, SYBEX;1999.
107. Hand D. (1998). Data mining: statistics and more? *The American statistician.* **53**(2): 112-118.
108. Klein P., DeLisi C. *et al*. (1985). The detection and classification of membrane-spanning proteins. *Biochimica et biophysica acta.* **815**: 468-476.
109. Kyte J., Doolittle R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular biology.* **157**: 105-132.
110. Yang E., Schroeder M. *et al.* (2003). Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Cold spring harbour laboratory press*. **13**: 1863-1872.
111. Holstege F. C., Jennings E. G. *et al*. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell.* **95**: 717-728.
112. Bernstein J. A., Cohen S. N. *et al*. (2002). Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences.* **99**: 9697-9702.
113. Wang Y., Brown P. O. *et al.* (2002). Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences.* **99**: 5860-5865.
114. Verma M., Kagan J. *et al.* (2003). Proteomic analysis of cancer-cell mitochondria. *Nature reviews cancer*. **3**(10): 789-795.
115. Mann M., Lander E. S., *et al.* (2003). Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell.* **115**: 629-640.
116. Haynes P. A., Yates J. R. (2000). Protein profiling-pitfalls and progress. *Yeast.* **17**(2):81-7.
117. Chiaromonte F., Miller W. *et al*. (2003). Gene length and proximity to neighbours affect genome-wide expression levels. *Cold spring harbour laboratory press.* **13**: 2602-2608.
118. Hilleren P., Parker R. (1999). Mechanisms of mRNA surveillance in eukaryotes. *Annual review of genetics.* **33**: 229-260.
119. Ross J. (1995). mRNA stability in mammalian cells. *Microbiology Reviews.* **59**: 423-450.
120. Schiavi S. C., Chen C. Y. *et al*. (1994). Multiple elements in the c-fos protein-coding region facilitate deadenylation and decay by a mechanism coupled to translation. *Journal of biological chemistry*. **269**: 3441-3448.
121. Paulding W. R:, Czyzyk-Krzeska M. F. (2000). Hypoxia-induced regulation of mRNA stability. *Advances in experimental medicine and biology.* **475**: 111-121.
122. Kumar S. A., Pentecost B. T. *et al.* (1990). Estrogen regulation of creatine kinase-B in the rat uterus. *Molecular endocrinology.* **4**: 1000-1010.
123. Warren G., Lowe M. *et al.* (2000). The Mitotic Phosphorylation Cycle of the cis-Golgi Matrix Protein GM130. *The journal of cell biology*. **149**(2): 341-356.
124. Bayer M., Peters R*. et al.* (2001).The Golgi matrix protein GM130: a specific interacting partner of the small GTPase rab1b. *The EMBO journal*. **2**(4): 336-341.

125. Lodish H., Lawrence S. *et al*. Molecular cell biology. 4th edition. New York: W. H Freeman & Co; 2000.
126. Arden K. C., Fu K. *et al*. (1995).Localization of short/branched chain acyl-CoA dehydrogenase (ACADSB) to human chromosome 10. *Genomics.* **25**(3): 743-745.
127. Desviat L. R., Gravel R.A. *et al* .(2001). Structure of the PCCA gene and identification of novel mutations in propionic acidemia. *Journal of inheritable metabolic diseases.* **24:**56.
128. Planta R. J. (1997). Regulation of ribosome synthesis in yeast. *Yeast.* **13**: 1505-1518.
129. Kef D. R., Warner J. R. (1981). Coordinate control of synthesis of ribosomal ribonucleic acid and ribosomal proteins during nutritional shift up in *Saccharomyces cerevisiae. Molecular and cellular biology.* **1**: 1007-1015.
130. Mager W. H., Planta R. J. (1991). Coordinate expression of ribosomal protein genes in yeast as a function of cellular growth rate. *Molecular and cellular biochemistry.* **104:** 181-187.
131. Cooper G. M. (2000). The cell – a molecular approach. 2nd edition. Sunderland (MA): Sinauer Associates, Inc; 2000Munro S., Pelham H. R. (1987). A C-terminal signal prevents secretion of luminal ER proteins. *Cell*. **48**: 899-907.
132. Bu G., Rennke S. *et al*. (1997). ERD2 proteins mediate ER retention of the HNEL signal of LRP's receptor-associated protein (RAP). *Journal of cellular sciences.* **110**: 65-73.
133. Lopez M. F., Melov S. (2002). Applied proteomics: mitochondrial proteins and effect on function. *Circulation research.* **90**: 380-389.
134. Ref. Dai J., Zhang L. *et al.* (2005). A comparative proteomic strategy for subcellular proteome research: Icat approach coupled with bioinformatics prediction to ascertain rat liver mitochondrial proteins and indication of mitochondrial localization for catalase. *Molecular & Cellular Proteomics.* **4**: 12-34.
135. Taylor R. D., Pfanner N. (2004). The protein import and assembly machinery of the mitochondrial outer membrane. *Biochimica et biophysica acta.* **1658**: 3743.
136. Huckriede A., Heikema A. *et al.* (1996). Transient expression of a mitochondrial precursor protein. A new approach to study mitochondrial protein import in cells of higher eukaryotes. *European journal of biochemistry*. **237**(1): 288-294.
137. Grivell L. A., Rep M. (1996). The role of protein degradation in mitochondrial function and biogenesis. *Current genetics.* **30**(5): 367-380.
138. Desautels M. (1986). Mitochondrial proteolysis. (In: Fiskum G (ed) Mitochondrial physiology and pathology. Van Nosrand Reinhold, New York, p 40-65.
139. Hare J. F. (1990). Mechanisms of membrane protein turnover. *Biochimica et biophysica acta.* **1031**: 71-90.
140. Kalnov S. L., Novikova L. A. *et al.* (1979). Proteolysis of the products of mitochondrial protein synthesis in yeast mitochondria and sub-mitochondrial particles. *Biochem J.* **182**: 195-202.
141. Wheeldon L W., Bof M., *et al.* (1974). Stable and labile products of mitochondrial protein synthesis in vitro. *European journal of biochemistry.* **46**: 189-199.
142. Hafeti Y. (1985). The mitochondrial electron transport and oxidative phosphorylation system. *Annual reviews in biochemistry.* **54**: 1015-1069.
143. Manoil C., Traxler B. (1995). Membrane protein assembly: genetic, evolutionary and medical perspectives. *Annual reviews in genetics.* **29**: 131-150.
144. Barter P. J., Connor W. E. (1975). The transport of triglyceride in the high-density lipoproteins of human plasma. *Journal of laboratory and clinical medicine.* **85**(2): 260-272.
145. http://www.bmb.leeds.ac.uk/illingworth/metabol/amino.htm.
146. Jaleel A., Skreekumaran Nair K. (2004). Identification of multiple proteins whose synthetic rates are enhanced by high amino acid levels in rat hepatocytes. *American journal of physiology endocrinology and metabolism.* **286**: E950-E957.
147. Kimball S. R., Jefferson L. S. *et al.* (1996). Translational and pretranslational regulation of protein synthesis by amino avid availability in primary cultures of rat hepatocytes. *International journal of biochemistry and cell biology*. **28**: 285-294.
148. Nicholson J. K., Wilson I. D. *et al*. (2004). The challenges of modeling mammalian complexity. *Nature biotechnology.* **22**(10): 1268-1274.
149. HUPO, the Human Proteome Organization: http://www.hupo.org.

150. AlignACE, Aligns Nucleic Acid Conserved Elements: http://www.psc.edu/general/software/packages/alignace/alignace.html.

9 Attachments

Table 1: Protein characteristics for easy vs. hard detecting proteins

Proteins that are easy to detect								
3O5B_RAT	P31210	82	0.38	157	-	-	6.61	37639
ALBU_RAT	P02770	98	0.71	625	-0.39	1	6.44	70669
CPSM_RAT	P07756	90	8.87	651	-0.12	1	6.75	165672
GR78_RAT	P06761	79	1.02	203			4.9	72473
COMT_RAT	P22734	78	0.46	142	0.04	1	5.41	290876
DHAM_RAT	P11884	76	0.57	202	-0.14	0	7.02	56965
DHE3_RAT	P10860	76	4.96	237			8.04	61731
3HAO_RA	P46953	72	0.18	88			5.71	32846
ER60_RAT	P11598	72	0.61	261			6.14	57043
SM30_RAT	Q03336	72	0.97	151	-0.31	0	5.32	33937
P60_MOUSE	P19226	71	2.64	145			6.02	61088
METL_RAT	P13444	70	0.57	143	-0.18	0	5.83	44240
SUAC_RAT	P50237	70	0.13	83	-0.65	0	6.53	35854
DIDH_RAT	P23457	68	0.03	72	-0.34	0	7.07	37517
SAHH_RAT	P10760	68	0.24	85	-0.05	0	6.51	47889
IDHC_RAT	P41562	68	0.2	76	-0.04	0	6.99	47046
THIM_RAT	P13437	66	4.31	170	-0.04	0	7.92	42243
TRFE_RAT	P12346	64	0.22	125	-0.25	1	7.12	78538
K2C8_RAT	Q10758	64	0.23	138	-0.62	0	5.83	53854
PYC_RAT	P52873	62	0.36	136	-0.17	0	6.7	130348
BUP_RAT	Q03248	62	0.2	68	-0.34	0	6.92	44584
	Average		1.33	189	-0.24	0.27	6.47	71574
	Median		0.46	143	-0.21		6.53	53854
Proteins that are hard to detect								
Protein	**Acc. nr**	**%P$_{liver}$**	**Level**	**Spots**	**GRAVY**	**TM**	**pI**	**MW**
CATB_RAT	P00787	2	0.02	1			5.42	38357
DECR_RAT	Q64591	2	0.35	1			9.73	36465
FABL_RAT	P02692	2	0.35	1	-0.43	0	8.39	14320
SSDH_RAT	P51650	2	-	1	-0.02	0	6.76	52668
GM13_RAT	Q62839	3	0.04	5			4.57	103964
MPPB_RAT	Q03346	3	0.01	4			6.81	55024
SPCN_RAT	P16086	3	0.02	1			5.7	118671
CATD_RAT	P24268	4	-	4			7.07	45165
ECHB_RAT	Q60587	4	0.47	5			10.3	51666
GLO2_RAT	O35952	4	-	3	-0.37	0	6.95	29162
NLTP_RAT	P11915	4	0.08	4			7.02	59516
SUO2_RAT	P52845	4	-	2	-0.53	0	5.61	35683
SUO3_RAT	P49889	4	-	3	-0.55	0	5.61	35734
THIL_RAT	P17764	5	0.67	8			9.17	45008
ATPO_RAT	Q06647	6	0.04	6	-0.02	0	10.8	23439
GRP1_RAT	P97576	6	0.01	9			8.5	24509
GTT2_RAT	P30713	6	0.02	5	-0.01	0	7.98	27461
ACTB_RAT	P02570	8	1.13	1			5.24	42051
APT_RAT	P36972	8		5	0.11	0	6.52	19761
GDIB_RAT	P50399	8	-	6	-	-	5.69	51165
KCRB_RAT	P07335	8	0.01	8	-0.47	0	5.36	42970
PRS4_RAT	P49014	8	-	6	-	-	6.14	49324
SUO1_RAT	P52844	8	-	4	-0.55	0	6	35827
MTE1_RAT	O55171	9	0.87	15			7.85	49954
RINI_RAT	P29315	9	0.01	10			4.47	51583
	Average		0.31	4.72	-0.23	0	6.95	45578
	Median		0.04	4	-0.24	0	6.76	42970

Table 2: Transcripts hard to detect in rat liver vs. hepatocytes and both liver and hepatocytes

Rat liver			Hepatocytes			Rat liver and hepatocytes		
Protein	**%T$_{\text{hepataocyte}}$**	**%T$_{\text{liver}}$**	**Protein**	**%T$_{\text{hepataocyte}}$**	**%T$_{\text{liver}}$**	**Protein**	**%T$_{\text{hepataocyte}}$**	**%T$_{\text{liver}}$**
K6PL_RAT	61.7	3.75	THBG_RAT	4.55	30.2	NMZ1_RAT	2.56	0.13
PEX3_RAT	0.4	3.51	ZP1_RAT	4.26	11.5	DNR2_RAT	2.4	0.35
SPCN_RAT	46.3	2.56	V1AR_RAT	3.69	99.3	M3K8_RAT	2.27	1.05
CP1B_RAT	10.8	2.43	LPP_RAT	2.84	98.2	A1A3_RAT	2.27	0.39
TTF1_RAT	11.4	2.3	NMZ1_RAT	2.56	0.13	NOS2_RAT	2.27	0.07
B2AR_RAT	7.1	1.97	ATS1_RAT	2.4	81.7	MYCB_RAT	1.99	0.07
CTE1_RAT	0	1.84	DNR2_RAT	2.4	0.35	PTHY_RAT	1.7	0.85
TF_RAT	25.9	1.77	M3K8_RAT	2.27	1.05	IP3L_RAT	0.57	0.59
CTPT_RAT	58.8	1.25	NOS2_RAT	2.27	0.07	PEX3_RAT	0.4	3.51
SP1_RAT	6.0	1.18	CP7A_RAT	2.27	84.4	STP1_RAT	0.28	0.46
M3K8_RAT	2.27	1.05	SC31_RAT	2.27	9.59	CTE1_RAT	0	1.84
MAFB_RAT	0	0.99	A1A3_RAT	2.27	0.39	MAFB_RAT	0	0.99
PRIO_RAT	92.3	0.85	MYCB_RAT	1.99	0.07	HS72_RAT	0	0.33
PTHY_RAT	1.7	0.85	PTHY_RAT	1.7	0.85	PAP1_RAT	0	0.33
IP3L_RAT	0.57	0.59	HNF6_RAT	1.7	65.9	MERL_RAT	0	0.2
STP1_RAT	0.28	0.46	HN3A_RAT	1.7	73.9	SOC3_RAT	0	0.2
NP14_RAT	57.1	0.39	GPDM_RAT	1.7	8.8	GPV_RAT	0	0.13
A1A3_RAT	2.27	0.39	CEBA_RAT	1.7	55.7	SOMA_RAT	0	0.13
DNR2_RAT	2.40	0.35	IBP3_RAT	1.42	99.5	DPG1_RAT	0	0.09
HS72_RAT	0	0.33	COA1_RAT	1.42	9.46	ACH3_RAT	0	0.07
PAP1_RAT	0	0.33	IBP2_RAT	1.14	16.0	AT7B_RAT	0	0.07
MERL_RAT	0	0.2	ATHA_RAT	1.14	17.9	CRF_RAT	0	0.07
SOC3_RAT	0	0.2	HGF_RAT	1.14	25.0	CYA5_RAT	0	0.07
NMZ1_RAT	2.56	0.13	RXRA_RAT	1.14	17.7	ERBP_RAT	0	0.07
GPV_RAT	0	0.13	NT3_RAT	0.85	37.1	IAPP_RAT	0	0.07
SOMA_RAT	0	0.13	IP3L_RAT	0.57	0.59	IL8B_RAT	0	0.07
DPG1_RAT	0	0.09	PEX3_RAT	0.4	3.51	CYGB_RAT	0	0
AAK1_RAT	94.9	0.07	STP1_RAT	0.28	0.46	DPO2_RAT	0	0
JUND_RAT	20.5	0.07	SOC3_RAT	0	0.2	DPOD_RAT	0	0
NOS2_RAT	2.27	0.07	ERBP_RAT	0	0.07	FAK2_RAT	0	0
MYCB_RAT	1.99	0.07	DRNG_RAT	0	68.1	GAT1_RAT	0	0
ACH3_RAT	0	0.07	DPOD_RAT	0	0	GFAP_RAT	0	0
AT7B_RAT	0	0.07	DPO2_RAT	0	0	HH1R_RAT	0	0
CRF_RAT	0	0.07	DPG1_RAT	0	0.09	HSB2_RAT	0	0
CYA5_RAT	0	0.07	CYGB_RAT	0	0	INB_RAT	0	0
ERBP_RAT	0	0.07	CYA5_RAT	0	0.07	KCN3_RAT	0	0
IAPP_RAT	0	0.07	FAK2_RAT	0	0	MY1A_RAT	0	0
IL8B_RAT	0	0.07	CRF_RAT	0	0.07	N107_RAT	0	0
GS28_RAT	22.4	0	PSPC_RAT	0	0	NER2_RAT	0	0
CYGB_RAT	0	0	SOMA_RAT	0	0.13	NHR2_RAT	0	0
DPO2_RAT	0	0	STB2_RAT	0	0	OXYR_RAT	0	0
DPOD_RAT	0	0	CAO3_RAT	0	9	P2Y4_RAT	0	0
FAK2_RAT	0	0	TR16_RAT	0	0	PI5B_RAT	0	0
GAT1_RAT	0	0	AT7B_RAT	0	0.07	PSPC_RAT	0	0
GFAP_RAT	0	0	ACH3_RAT	0	0.07	STB2_RAT	0	0
HH1R_RAT	0	0	ACDB_RAT	0	13.6	TR16_RAT	0	0
HSB2_RAT	0	0	CTE1_RAT	0	1.84			
INB_RAT	0	0	IAPP_RAT	0	0.07			
KCN3_RAT	0	0	NER2_RAT	0	0			
MY1A_RAT	0	0	MY1A_RAT	0	0			
N107_RAT	0	0	MERL_RAT	0	0.2			
NER2_RAT	0	0	MAFB_RAT	0	0.99			
NHR2_RAT	0	0	NHR2_RAT	0	0			
OXYR_RAT	0	0	KCN3_RAT	0	0			
P2Y4_RAT	0	0	INB_RAT	0	0			
PI5B_RAT	0	0	OXYR_RAT	0	0			
PSPC_RAT	0	0	GAT1_RAT	0	0			
STB2_RAT	0	0	PAP1_RAT	0	0.33			
TR16_RAT	0	0	IL8B_RAT	0	0.07			
			HSB2_RAT	0	0			

			HS72_RAT	0	0.33			
			PI5B_RAT	0	0			
			HH1R_RAT	0	0			
			PRLR_RAT	0	57.8			
			GPV_RAT	0	0.13			
			N107_RAT	0	0			
			GFAP_RAT	0	0			
			P2Y4_RAT	0	0			