UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

Electronics and Computer Science

Modelling at the Transcriptome - Proteome Interface

by

Yawwani P. Gunawardana

Thesis for the degree of Doctor of Philosophy

November 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING
Electronics and Computer Science

Doctor of Philosophy

MODELLING AT THE TRANSCRIPTOME - PROTEOME INTERFACE

by Yawwani P. Gunawardana

In high-throughput experimental biology, it is widely acknowledged that mRNA expression levels and the corresponding protein abundances are jointly analysed to observe the relationship between these two *omic* measurements. While some experiments have shown a good correlation between transcriptome and proteome for some species under different conditions, such correlation values are not universal due to post-transcriptional and post-translational regulations. Thus, bridging the gap between transcriptome and proteome measurements allow us to uncover useful biological insights of the above regulations which are important to study on protein generation process and several disease conditions. We develop a data-driven predictor using transcriptome layer properties as proxies to protein abundance and employ the model in a novel manner to detect post-translationally regulated proteins, hypothesizing that model failures (outlier proteins) occur due to protein stability disruption by post-translational modifications (PTMs). Three outlier detection techniques were employed with our protein abundance predictor to detect post-translationally regulated protein. Those are; (1) simple linear regression model which detects outliers by looking at the predicted and the measured protein scatter plot, (2) Outlier Rejecting Regression (ORR) model, a novel mathematical formulation which returns user-specific fraction of the data as outliers by solving a non-convex optimization problem using Difference of Convex functions Algorithm (DCA) and (3) Quantile Regression (QR) which employs an asymmetric loss model to detect outliers only with negative losses for the first time in *omic* world. Proteins extracted as outliers using above techniques confirmed our hypothesis on post-translational regulation (PTR) by providing high statistical confidence for functional annotations and pathway information. Therefore, this data-driven framework can be used as a reliable technique for biologists to reduce laboratory experimental workspace in detecting post-translationally regulated proteins.

We also perform a thorough inference analysis on most commonly used high-throughput microarray and RNA-Seq measurements using several machine learning inference techniques to observe whether their high numerical precision provides additional information about the gene with respect to the binary representation of gene switch on/off status. We perform this analysis at the transcriptome level and as well at the proteome level as an extended experimental setting of our PTR detection framework. These analyses suggest that binarized mRNA concentrations, which are measured using high-throughput RNA-Seq and microarray technologies are sufficient to perform accurate machine learning inferences similar to continuous measurements, not only at the transcriptome level but also at the proteome level to predict protein abundance and to detect protein with post-translation regulation to a high confidence level.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Yawwani P. Gunawardana , declare that the thesis entitled *Modelling at the Transcriptome - Proteome Interface* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:
  - Y.Gunawardana, M.Niranjan (2013), Bridging the Gap Between Transcriptome and Proteome Measurements Identifies Post Translationally Regulated Genes, *Bioinformatics*, btt537.
  - Y.Gunawardana, S. Fujiwara, A. Takeda, J. Woo, C. H. Woelk, M.Niranjan (2014), Outlier-Detection at the Transcriptome-Proteome Interface, *Bioinformatics*, btv182.
  - Y.Gunawardana, S.Tuna, C. H. Woelk, M.Niranjan (In Preparation), Numerical Precision in Transcritome Representation is Illusory, *Nature Methods*

- Presentation and Poster (Peer Reviewed Extended Abstract) - Y.Gunawardana, M.Niranjan (2013), Bridging the Gap Between Transcriptome and Proteome Measurements Identifies Post Translationally Regulated Genes, *Seventh International Workshop on Machine Learning in Systems Biology*,Berlin,Germany,July 19-21, 2013.
- Presentation and Poster (Peer Reviewed Extended Abstract) and presented as a poster - Y.Gunawardana, S. Fujiwara, A. Takeda, C. Woelk, M.Niranjan (2014), Outlier-Detecting Support Vector Regression for Modelling at the Transcriptome-Proteome Interface, *Eighth International Workshop on Machine Learning in Systems Biology*,Strasbourg,France,September 6-7, 2014.
- Presentation - Y.Gunawardana, S.Tuna & M.Niranjan Precision in Transcriptome-based Inference RNA-Seq Vs Microarray *Fourth Next Generation Sequencing Symposium*, Southampton General Hospital, Southampton, UK, May 23, 2014.
- Poster - Y.Gunawardana, M.Niranjan (2013), Modelling at Transcriptome and Proteome Measurements Identifies Post Translationally Regulated Genes, *Functional Genomics & Systems Biology 2013*,Wellcome Trust Conference Centre,Cambridge, UK,November 21-23, 2013.

Signed:.....

Date:.....

Acknowledgements

Foremost, I would like to express my sincere gratitude to my first supervisor, Prof Mahesan Niranjan for his expertise, enthusiasm, patience, motivation and encouragement. His guidance helped me in every aspect of my PhD study and also to build up my career as an excellent researcher.

I'm very fortunate to get regular advice and support from my second supervisor, Dr Christopher Woelk and his research group at the Southampton General Hospital. His knowledge in genomics and biology helped me to take this research to a higher level in a biological perspective.

I would also like to thank Dr Akiko Takeda for collaborating with my research and supporting me with a mathematical background to enhance quality of this research.

This research was carried out in the Vision, Control and Learning (VLC) group at the University of Southampton. Many thanks to my colleagues (in alphabetical order): Bariboon Deeka (Boon), Jianhao Xiong, Karnkamon Maneenil (Por), Thabiso Maupong, Xiaoru Sun and Dr.Xin Lui. All these friends gave their constant support, advice and good humor throughout my PhD study.

A special thanks goes to Gayan Dilhara Jayakody, my loving husband, without him my effort would have been worth nothing. His love, support, patient inspires me to overcome all the obstacle in life and achieve goals with success.

My heartiest gratitude goes to my parents, Susil and Aselin Gunawardana for their continuous emotional and moral support.

Finally, I would like to acknowledge my financial sponsor the School of Electronics and Computer Science of University of Southampton.

Chapter 1

Introduction

1.1 Background

Computational analysis of high-throughput *omic* measurements has played a major role in systems biology research over the last decade or so. Advanced measuring techniques coupled with strong archiving methods have revolutionized the way of uncovering biological insights, more at a system level than at a component level. Terabytes of metabolomics, transcriptomic and proteomic experimental data are archived for computational inferences. Transcriptome expression measurements made with cost effective microarray technology being the most dominant *omic* measurement type with respect to the other two. Most of the previous studies are based on simply looking at the correlation between mRNA measurements and corresponding protein measurements (Gygi et al., 1999; Futcher et al., 1999; Beyer et al., 2004; Wu et al., 2008) and report varying levels of correlation. However, some authors have developed data-driven models such as classification (Pancaldi and Bähler, 2011; Muppirala et al., 2011), clustering (Eisen et al., 1998; Heard et al., 2005) and probabilistic approaches (Rogers et al., 2008; Kannan et al., 2007) to investigate the relationship between these two properties. Tuller et al., 2007 construct a machine learning based protein abundance predictor which is a different approach to previous studies. They use several properties which are related to translation process including mRNA abundance, and train a linear regression to predict protein concentrations. Greedy feature selection algorithm selects mRNA concentration, tRNA adaptation index (tAI) and evolutionary rate (ER) as the most dominant features for their predictor. In fact, Tuller et al., 2007 achieve a correlation of 0.76 between the true and the relevant predicted protein concentrations. However, it is difficult to demonstrate the relationship between transcriptome-proteome data due to post-transcriptional and post-translational regulations. During these processes mRNA (post-transcription) and protein (post-translation) stability get disrupted due to enzymatic and structural

modifications, hence we believe that data-driven model between transcriptome and proteome interface can be used to extract information about these post-transcriptional and post-translational regulations. Taking inspiration from Tuller et al., 2007's study, we also develop a protein abundance predictor, but using an extended feature space. In contrast to previous work (Tuller et al., 2007), we employ sparse inducing LASSO technique and select five features to predict protein abundance more accurately ($R^2 = 0.86$); those are mRNA, tAI, codon bias, ribosome density and occupancy. Further, we expand the data-driven approach to detect post-translationally regulated proteins by considering model failures (outliers) occur due to protein stability disruption by post-translational modifications (PTMs). Thus, here we introduce a computational framework which can be considered as a reliable technique for experimental biologists to reduce the laboratory workspace in detecting post-translationally regulated proteins.

1.2 Motivation and Hypothesis

Post-translational modifications are important to study on different diseases. Protein chemical structure changes occur after translation process effects on several physiological diseases such as Alzheimer's disease (AD), rheumatoid arthritis and Parkinson's disease (PD) (Gong et al., 2005; Oueslati et al., 2010). Furthermore, PTMs are used as candidates for biomarker discovery for many cancer types. Phosphorylation of protein B/Akt kinase enhances the effectiveness of the drug to suppress tumor growth (Gulmann et al., 2005). Similarly, glycosylation was used to discover biomarker CA125 for ovarian carcinoma caner (Wong et al., 2003) and glycomics profiling was employed to obtain serum biomarker for hepatocellular carcinoma (Block et al., 2005). Additionally, these PTMs are also being used as therapeutic interventions in cancer treatment. Autophosphorylation target on Tyrosine Kinase inhibitor as a treatment for lung cancer (Lynch et al., 2004; Paez et al., 2004). Scientists use very complex and expensive mass spectrometry (MS) technique to detect post-translationally modified proteins. However, this process is not straightforward due to several technical reasons. Firstly, stoichiometry of PTR is relatively low with respect to protein peptide pool (Wei and Li, 2009). Therefore it is difficult to detect these changes by considering low abundance peptide peaks. Next, PTMs are largely heterogeneous among different modification types and also within a single modification. For example phosphorylation has several different phosphorylated forms (Jensen, 2004; Yoon and Seger, 2006). Thus, unknown prior knowledge of which PTMs are going to detect will enforce difficulties to identify variabilities inside a single modification. Further, most of these analytical methods cannot detect low abundant minor sites of PTMs due to technical difficulties (Chandramouli and Qian, 2009; Yoon and Seger, 2006). Though there are several PTM detection methods have developed with MS technique, still this task remains as a technical challenge (Arnott et al., 2003). Therefore it is important to develop a computational model to detect post-translationally

regulated (modified) proteins. Here we extract sub set of proteins which are likely to be post-translationally regulated. In fact, during our finer level functional annotation test, we explicitly detect proteins with phosphorylation, acetylation and ubiquitination. Thus, our framework provides prior knowledge of the PTMs which are likely to be detected by the MS experiment on these sub set of proteins and also reduces the experimental time and cost due to the fewer number of testing samples. Additionally, since we are targeting on a sub set of proteins, experimentalists will be able to detect minor sites by amplifying the low abundances peptide ratios. This motivated us to develop a data-driven computational framework to detect proteins which endure post-translation regulation.

In this research we take a novel approach to detect post-translationally regulated proteins by developing a protein abundance predictor and looking at the model failures (also known as outliers) which give large errors between actual (measured) concentrations and the predictions. *We hypothesise that these model failures occurred due to protein stability disruption caused by post-translational modifications during post-translation regulation process.* Therefore, outlier proteins with measured abundance lower than predicted are likely candidates of post-translationally regulated proteins. Several studies have shown that protein degradation can be triggered by post-translational modifications (Levine, 1983; Callis, 1995). In fact, protein stability can be disrupted by attaching new substitutions to the amino acid. Post-translational modifications such as phosphorylation and acetylation can act as proxies for such mutations by attaching to specific local sites which increase the susceptibility of the protein to proteinase action to catalyse protein degradation (Hood et al., 1977; Holzer and Heinrich, 1980). Martinez et al., 2003 showed that ABCA1-PEST sequence phosphorylation regulates ABCA1 calpain degradation. They performed several *in-vitro* experiments to show that phosphorylation with PEST motifs regulate ABCA1 protein degradation and reduce the overall protein expression level. Firstly, a flag-tagged PEST deletion mutant (which removes the degradation process) was expressed in HEK293 cells (ABCA1delPEST) to compare the cell surface expression levels with wild-type ABCA1 proteins. This experiment showed that wild-type ABCA1 gave $3.9 \pm 0.4\%$ mean fold reduction of the protein concentration with respect to protein degradation inactive PEST deleted ABCA1delPEST protein (Wang et al., 2003). Next, a wild-type ABCA1 protein expression level was compared with phosphorylation sites mutated ABCA1 proteins. These mutations impair protein degradation activity in ABCA1 proteins. Therefore, the phosphorylation site mutated ABCA1 expression levels gave higher expression levels compared to wild-type ABCA1 proteins (mean fold $3.4 \pm 0.3\%$ with MutAAAA in Thr-1286 site and $3.3 \pm 0.3\%$ with MutASSA in Thr-1305 site). Thus, these experiments suggest that post-translational modifications such as phosphorylation and acetylation catalyse the protein degradation and reduce total protein expression levels. However, other regulations occur post-translationally, such

as localization, hydrophobicity and enzymatic activities cannot be detected using this data-driven approach.

1.3 Contributions

1.3.1 Modelling Transcriptome-Proteome Measurements & Detecting Post-translationally Regulated Proteins

In the first part of this thesis, we develop a linear predictor using the best five features out of 37 transcriptomic properties including mRNA abundance of yeast (*Saccharomyces cerevisiae*), which are selected by the sparsity inducing lasso (L_1 norm regularization) technique. We then look for the systematic errors made by the predictor by hypothesizing that those mRNA and protein pairs which have large errors where the predicted protein abundance is lower than the actual measurement are likely candidates for post-translational regulation. This experiment follows the fact that input features of our predictor do not contain any information for post-translational regulation. We confirme our hypothesis by performing a functional annotation check on these outlier proteins and showing that they are highly enriched with post-translationally regulated proteins with a high statistical confidence. Thus, here we introduce a data-driven machine learning approach to reduce the laboratory experimental workspace to detect post-translationally regulated protein.

1.3.2 Outlier Detection at the Transcriptome-Proteome Interface

Secondly, we introduce two regression models to extract outliers more systemically at the transcriptome-proteome interface to prove our initial hypothesis on post-translational regulation. (1) Outlier Rejecting Regression (ORR) model, devised through a collaboration with Dr Akiko Takeda and Shuhei Fujiwara, allowing to specify a fraction (percentage) of data as outliers before performing the weight optimization in regression problem to obtain robust outliers. (2) Quantile Regression (QR) has the asymmetric loss model property to extract proteins with only negative losses where the predicted abundance is lower than the actual measurement. These outliers confirmed our initial hypothesis on post-translational regulation by providing high confidence levels for functional annotations and biological evidence such as PTR related gene ontologies and pathways.

1.3.3 Numerical Precision in Transcriptome-based Inference & Coherence with Protein Prediction

Finally, we explore machine learning inference capabilities between high (continuous) and low (binarized) numerical precision of microarray and RNA-Seq high-throughput transcriptome measurements. Here we perform this task (1) using only transcriptome measurements to explore quantitative analysis and (2) using our machine learning framework to observe the PTR detection capability as a qualitative analysis. This is an extended experimental setting of the PTR detection framework under different transcriptomic inputs. Previous authors have looked into binarized microarray data inference capabilities at the transcriptome level (Tuna and Niranjan, 2009, 2010). However, as a novel approach, here we also incorporate very recently developed RNA-Seq technology measurements and compare these two high-throughput measuring techniques (microarray and RNA-Seq) not only at the transcriptomic level but also at the proteome level. These experiments suggest that, at both transcriptome and proteome levels, RNA-Seq and microarray data perform similarly under high and low numerical precision where binary data is sufficient to perform quantitative analysis. However, the PTR detection framework showed that RNA-Seq binary data was able to capture more qualitative information on post-translation regulation compared to microarray binary measurements.

1.4 Thesis Organization

This report is organized as follows. Chapter 2 presents a literature review which includes an overview of the central dogma of molecular biology, the importance of detecting post-translationally regulated proteins, machine learning inference techniques and the joint analysis of transcriptome and proteome data. Development of protein abundance predictor and proving hypothesis of post-translation regulation by analysing model failures or outliers are described in Chapter 3. Chapter 4 introduces two novel formulations to systemically detect outliers at the transcriptome-proteome interface and shows the over representation of post-translationally regulated proteins as model failures using the three types of regression models (including the simple linear regression model in Chapter 3). Chapter 5 provides a thorough analysis of numerical precision in RNA-Seq and microarray transcriptome measurements using continuous and binarized data and explore the PTR detection capability by modelling at the transcriptome-proteome interface. Finally, conclusions and future work are presented in Chapter 6.

1.5 Publications

The following publications, presentations and posters are based on contributions made during my PhD research:

Publications

- ⋆ Y. Gunawardana, M. Niranjan (2013), Bridging the Gap Between Transcriptome and Proteome Measurements Identifies Post-translationally Regulated Genes, *Bioinformatics*, btt537.
- ⋆ Y. Gunawardana, S. Fujiwara, A. Takeda, J. Woo, C. Woelk, M.Niranjan (2014), Outlier-Detection at the Transcriptome-Proteome Interface, *Bioinformatics*, btv182.
- ⋆ Y. Gunawardana, S.Tuna, C. Woelk, M.Niranjan (In Preparation), Numerical Precision in Transcriptome Representation is Illusory, *Nature Methods*
- ⋆ H. Johnson, C. White, Y. Gunawardana, B. Oliver, C. Woelk, S. Garbis (Under Revision), Quality-Weighted Statistics Improves Differentially Expressed Protein Determination by Isobaric Tag Quantitation, *Journal of Proteome Research*

Presentations & Posters

- ⋆ Presentation and Poster (Peer Reviewed Extended Abstract) - Y. Gunawardana & M. Niranjan (2013), Bridging the Gap Between Transcriptome and Proteome Measurements Identifies Post Translationally Regulated Genes, *Seventh International Workshop on Machine Learning in Systems Biology*, Berlin, Germany, July 19-21, 2013.
- ⋆ Presentation and Poster (Peer Reviewed Extended Abstract) - Y. Gunawardana, S. Fujiwara, A. Takeda, C. Woelk & M. Niranjan (2014), Outlier-Detecting Support Vector Regression for Modelling at the Transcriptome-Proteome Interface, *Eighth International Workshop on Machine Learning in Systems Biology*, Strasbourg, France, September 6-7, 2014.
- ⋆ Presentation - Y. Gunawardana, S. Tuna & M. Niranjan Precision in Transcriptome-based Inference RNA-Seq Vs Microarray *Fourth Next Generation Sequencing Symposium*, Southampton General Hospital, Southampton, UK, May 23, 2014.
- ⋆ Poster - Y. Gunawardana, M. Niranjan (2013), Modelling at Transcriptome and Proteome Measurements Identifies Post-translationally Regulated Genes, *Functional Genomics & Systems Biology 2013*, Wellcome Trust Conference Centre, Cambridge, UK, November 21-23, 2013.

- ⋆ Poster - A. Heinson, C. Denman, Y. Gunawardana, B. Moekser, M. Niranjan, C. Woelk (2013), Bacterial Vaccine Design Using Reverse Vaccinology, *23th Annual International Society for Computational Biology (ISMB) 2015*, Dublin, Ireland, UK, July 10-14, 2015.

Chapter 2

Literature Review

This chapter provides a comprehensive review on biological, computational and machine learning concepts we employed in this thesis to develop our hypothesis and to perform experiments to confirm it. We divided these background material into four main sections.

1. Central Dogma of Molecular Biology
2. Importance of Post-translational Regulation
3. Machine Learning Inference of High-throughput Omic Measurements
4. Joint Analysis of Transcriptome and Proteome Data

The first two sections are focused on biological information of protein generation process and explain why it is important to develop a computational approach to detect post-translationally regulated proteins at the transcriptome-proteome interface upon which our main hypothesis is built. We then provide details on machine learning inference techniques to analyse high-throughput transcriptomic measurements under high and low numerical precision. Finally, we review previous work related to integrated analysis of mRNA and protein data.

2.1 Central Dogma of Molecular Biology

This section describes the most important system in all living organisms, also known as the central dogma of molecular biology, the flow of passing genetic information to generate proteins. Francis Crick first stated this model in 1958 and re-stated it again in 1970 in Nature publication (Crick et al., 1970). This system describes on protein synthesis process inside living cells. There are three major classes of biopolymers; DNA,

RNA and protein. Further, new studies have found microRNA and sRNA as supporting biopolymers in central dogma. This framework illustrates the transmission of genetic instructions from DNA to generate proteins (Figure 2.1).

- **DNA**: Deoxyribonucleic acid (DNA) is a molecule which contains genetic information used for the development and functioning of all living organisms. These are mostly located in nucleus but small amounts can be found in mitochondria region as well. This is a long polymer made of repeating nucleotide units (Saenger, 1984; Butler, 2005). Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) are the four nucleotides in a DNA sequence. These nucleotides pair with each other (A with T and C with G) creating double-strand helices (Berg et al., 2002);

- **RNA**: Ribonucleic acid (RNA) is generated by DNA during the transcription regulation. The main function of RNA is to transfer the genetic code from nucleus to ribosomes to generate proteins. RNA helps DNA to pass on genetic information without leaving the nucleus (Berg et al., 2002);

- **mRNA, rRNA and tRNA**: The first type of RNA is known as messenger RNA (mRNA) and it carries information from DNA to ribosomal RNAs (rRNAs). mRNA is the starting point of the translation process and the information is transferred to protein by transfer RNA (tRNA) (Mattick, 2001; Mattick and Gagen, 2001; Berg et al., 2002);

- **Proteins**: These perform a vast range of functions within living organisms. They are large biological molecules with one or more chains of amino acids. Some of the functions carried out by catalysing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another. They fold in a specific three-dimensional structure which determines their functional activities. Amino acid sequences in protein consist of genetic information which is transferred by DNA (Kent, 2009; Lodish et al., 2000).

- **microRNA**: microRNAs (miRNAs) are small non-coding RNA molecules with around 22 nucleotides in length. These biopolymers are regulated by directly binding to 3'UTR region in mRNAs (Bartel, 2004). miRNAs inhibit the translation or mRNA degradation processes which are important in protein generation.

- **sRNA**: These are small non-coding RNAs with 50-250 nucleotides and found in bacteria with high level structure and stem loops. Microarray, RNA-Seq, Northern blotting techniques are employed to detect sRNA molecules. These sRNAs can either bind to the mRNAs and regulate the gene expressions or bind to proteins and modify their functionalities (Vogel and Wagner, 2007).

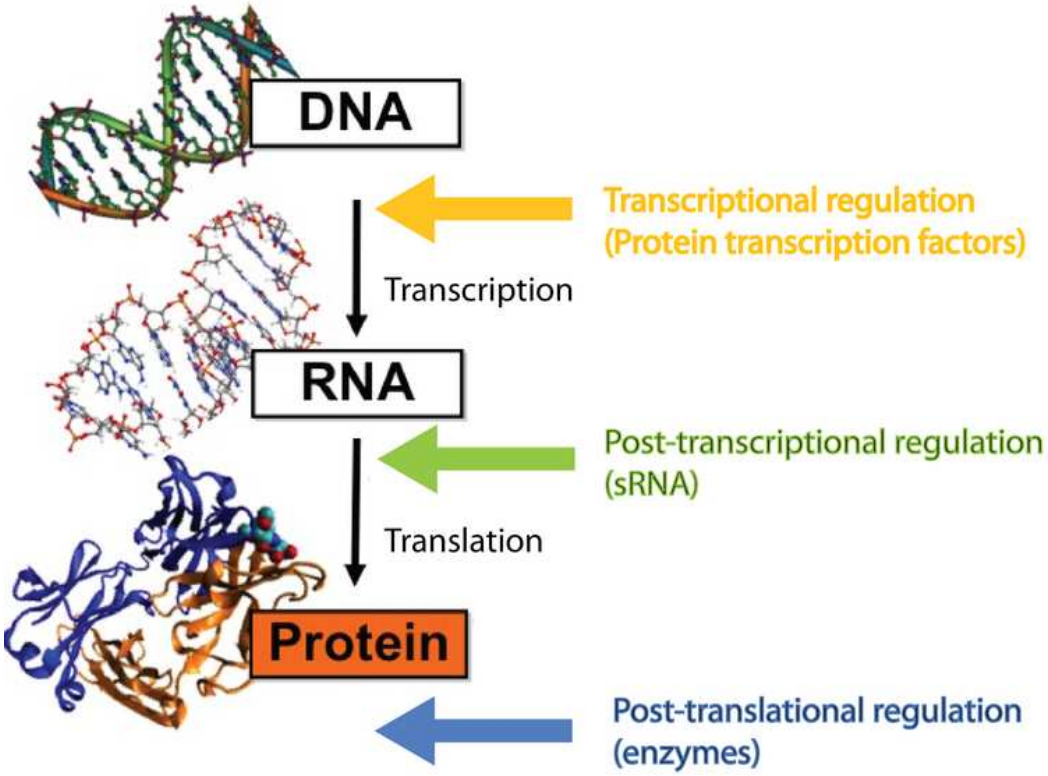


Figure 2.1: Central Dogma of Molecular Biology. Different levels of protein generation and important regulations are highlighted. sRNA indicates small RNAs with mRNA targets (taken from Silencing, 2011)

The central dogma of molecular biology consists of two main stages; namely transcription and translation, which allow DNA to create useful proteins. However, there can be some changes after each of these main processes. Post-transcriptional and post-translational changes happen after transcription and translation processes respectively.

2.1.1 Transcription

During transcription regulation, DNA passes its genetic information to mRNA which is a gene data transportation from the nucleus to cytoplasm. The DNA strand engages with specific RNA polymers and transcription factors to generate mRNA. This step can include 5' cap, a poly-A tail and splicing. Alternative splicing also helps to generate a range of proteins using a single mRNA. Following are the three main steps in the transcription process.

→ **RNA polymerase binds to DNA**: RNA polymerase binds at the promoter region of DNA and specific nucleotide sequences provide information about the

beginning and the end of the transcription regulation. Transcription factor (TF) proteins are directly involved with promoter region initiation;

→ **Elongation**: TFs unwind the DNA and RNA polymerase makes a copy of a single strand of DNA into a single stranded RNA polymer called mRNA. The template strand of DNA which makes the copy is known as the antisense stand and the other one which does not transcribe is known as the sense strand;

→ **Termination**: The RNA polymer moves along the DNA antisense strand until it reaches the terminator region and releases the mRNA polymer. After releasing the mRNA polymer, the DNA antisense strand again binds with the sense strand to create original DNA.

Levine, 2003 used several genome sequences to analyse the correlation between physiological and behavioural complexities with gene expression data by looking at transcription regulation. This study suggests that organism diversity occurs due to transcription regulation of gene expression. In fact, different complexes are required to regulate *cis*-DNA elements in a tissue specific and temporal manner. A single core promoter can be detected by TFIID complexes and at the same time it can also interact with co-factor complexes to vary the gene expression patterns. Different computational approaches such as clustering of *cis* regulatory elements have been used to identify novel enhancers in *Drosophila* genome (Markstein et al., 2002; Rajewsky et al., 2002). Several studies have also shown that RNA polymerase II enzymes in *Drosophila* and mammal genes pause at the promoter-proximal sites. Therefore the this regulation can occur at different speeds in different organisms. There are several phases in the transcription process such as initiation of transcription, elongation and termination. It is important to understand how regulation works for a particular gene to understand its underlying biological mechanism. Furthermore, transcriptionally inactivated promoters perform unusually and are generally viewed as exceptions (Core and Lis, 2008; Gariglio et al., 1981).

2.1.2 Post-transcription

Post-transcription regulation occurs between the transcription and translation processes. The changes that happen to mRNA after transcription are known as post-translational regulation. Modern experiments show that multiple mRNAs are co-regulated by one or more sequence-specific RNA-binding proteins. These are associated with RNA splicing, stability, localization and translation. Scientists are currently investigating post-transcription regulation properties using transcriptome and proteome data to further explore information on immune system, stress response, and disease related properties (Keene, 2007; Wu et al., 2008).

Alternative splicing is an important post-transcriptional regulation which occurs in eukaryotic genomes such as human. The rearrangement of exons and introns of a single mRNA generating several isoforms is known as alternative splicing mechanism. The complexity of human protein abundance prediction lies in this alternative splicing process. High-throughput sequencing showed that the human genome has far few genes compared to the number of human proteins. Alternative splicing enhanced the protein generation process and produce more than 150,000 proteins by using only 32,000 genes. Yeakley et al., 2002 performed a microarray and fiber optic study on gene-to-protein cycle. They observed that isoforms which are generated by alternative splicing have short bits of genetic material to produce variety of proteins. These unique genetic signatures reveal which portion of the gene is activated while producing different proteins. Therefore, alternative splicing rules out one-to-one mapping between genes to proteins in the human genome.

Microarray transcriptome measuring technique combined with chromatin immunopreciptitations can be used to reveal global network of transcriptional control in a variety of organisms and physiological conditions (Luscombe et al., 2004a; Barrera et al., 2006). DNA and its interactions with transcription factors and mRNA and its association with RNA-binding proteins are important for the regulation of gene expression at the post-transcriptional level. Most of the recent studies have focused on large scale system analysis of mRNA-protein interactions and dynamics (Wang et al., 2012b). Theses have employed microarray based approaches to study different processes of the genome-wide scale such as mRNA, RNA-binding proteins, mRNA stability, ribosomes and translational efficiency (Brockmann et al., 2007; Brazma et al., 2001; Wang et al., 2002). Nevertheless, these large-scale approaches are especially useful to uncover the importance of post-transcriptional regulatory mechanisms. In this research we are interested in post-transcription properties which can influence protein production in the translation process such as mRNA half-life, alternative splicing, miRNA etc.

2.1.3 Translation

This is the main functional phase of protein generation by incorporating mRNA and other transcription and post-transcription properties. mRNA carries coded genetic information to ribosomes and this information is read by the ribosomes as codon triplets, normally beginning with AUG. Transfer RNAs (tRNAs) bring matching codons to ribosome, completing amino acid sequence. Finally when the ribosome reaches the stop codons such as UAA, UGA or UAG, it releases the nascent polypeptide chain as a mature protein.

Effects such as cell exposure to stress or changing conditions as hypoxia, heat shock or change in carbon source on global mRNA specific translation regulation have been studied with the aid of translational profiling (Grolleau et al., 2002; Preiss and WHentze, 2003; Thomas and Johannes, 2007). Qin et al., 2007 used a high-resolution translation profiling approach to analyse mRNA translational control for early *Drosophila* embryogenesis. They measured ribosomal density and ribosome occupancy over 10,000 transcripts during the first ten hours (in 2-hour intervals) after egg laying and observed a variety of translational profiles. This indicates that there are multiple mechanisms modulating the transcript-specific translation. During our study we model this translation process computationally, using mRNA and other transcriptomic properties to predict protein abundance using yeast data.

2.1.4 Post-translation

Changes occur in some proteins after the translation process that increase or decrease protein production are known as post-translation regulation. This regulation can occur either by enzymatic events such as post-translational modifications or structural changes such a proteolysis. Post-translational regulation changes the properties of a protein by proteolytic cleavage or by addition of a modifying group to one or more amino acids. This will also effect on activities of the protein such as localization, turnover and interaction with other proteins (Mann and Jensen, 2003a). Though there are several regulations that can occur post-translationally, our study is focused on regulation where protein stability is disrupted by post-translation modifications and it starts to degrade faster.

Recent experiments have discovered that proteomic data is vastly more complex than genome data. It is estimated that the human genome consists of around 32,000 genes (Modrek and Lee, 2002), and over 1 million proteins can be found in human proteome (Jensen, 2004). Therefore a single gene encoded with many proteins makes proteomic data more complex (Figure 2.2). Post-translation modifications increase the complexity from the genome level to the proteome level by introducing new modifications to proteins after the translation process. Alternative splicing is another important regulation which improves the production of different proteins. Therefore the complexity of proteome is highly related to both transcriptome and genome data.

In fact, it is estimated that 5% of the proteome is comprised of enzymes that undergo more than two-hundred types of PTMs (Walsh et al., 2006). PTMs can occur at several stages after the translation process. Many proteins are modified soon after the translation process to mediate proper protein folding and stability. Others occur after folding and those modifications influence the biological activities of the proteins. These are covalently linked to tags which impact on protein degradation. Additionally, proteins

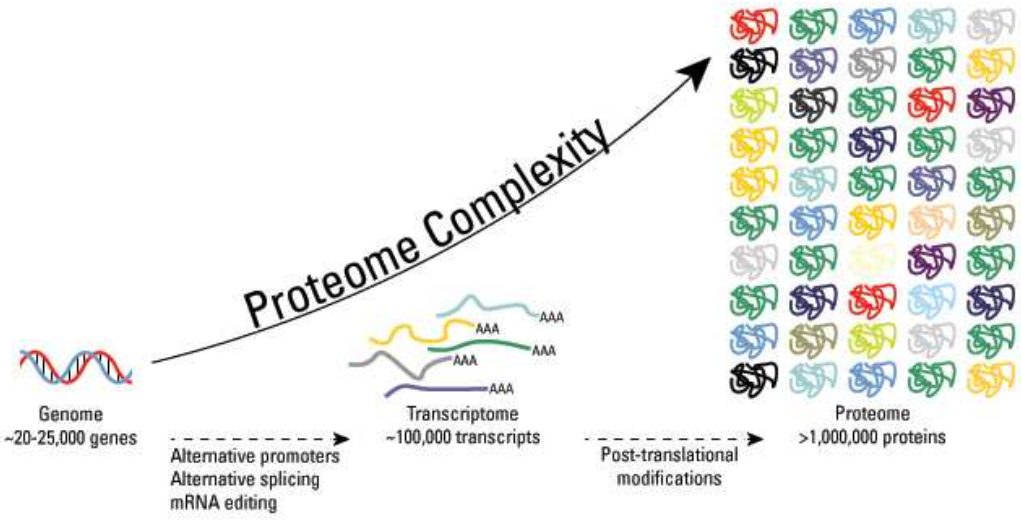


Figure 2.2: Post-translational modifications increase the proteomic diversity. Transcription process increases the number of transcriptomes relative to genome, and PTMs exponentially increases the complexity of proteome relative to both transcriptome and genome (Products, 2013)

are often modified by a combination of post-translational cleavage and attachment of functional groups.

2.1.4.1 Different Types of Post-translational Modifications (PTMs)

Different types of post-translational modifications are explained below. These post-translational modifications are used as functional annotations to detect post-translationally regulated proteins in later chapters (Section 3.4).

Phosphorylation

The addition of phosphate groups to the three main amino acids, serine, threonine and tyrosine in eukaryotic cells is the main function of phosphorylation. Nuclephilic (-OH) group in these amino acids attacks the terminal phosphate group on adenosine triphosphate (ATP), resulting in the attachment of a phosphate group to the amino acid chain. A large amount of free energy is released by the broken phosphate-phosphate bond in ATP. Phosphorylation is the most common PTM and it transmits signals throughout the cell. This can be observed in bacterial proteins and also one-third of the proteins in the human proteome (Cohen, 2000).

Glycosylation

The attachment of glucose moieties to protein is known as glycosylation and it gives greater proteomic diversity than other PTMs. There are several glycosidic linkages such as, N-, O-, C-linked glycosylation, glypiation (attachment of a GPI anchor) and phosphoglycosylation. All these modifications help critical functions of biosynthetic-secretory pathways in endoplasmic reticulum (ER) and golgi apparatus. Glycosylated proteins (glycoproteins) can be found in all living organisms including eukaryotes, eubacteria and archaea (Lechner and Wieland, 1989; P., 1997). This post-translation modification is known to be the most complex modification due to its involvement in a large number of enzymatic steps (Walsh, 2006).

Ubiquitination

Ubiquitination is an enzymatic process where the carboxylic acid of the terminal glycine in the activated ubiquitin forms an amide bond to the epsilon amine of lysine. 'Ubiquitin' is a small regulatory protein which directs proteins for recycling and other functions. Activation of ubiquitin is the first step of ubiquitination. Afterwards, the activated ubiquitin is transferred to the active cysteine site of an ubiquitin-conjugating enzyme. Finally the ubiquitination cascade creates an isopeptide bond between a lysine of the target protein and the C-terminal glycine of ubiquitin. Polyubiquitinated proteins are recognised by the 26s proteasome that catalyses the degradation of proteins.

Methylation

This post-translational modification increases hydrophobicity and neutralises negative amino acid charges by transferring one-carbon methyl group to nitrogen or oxygen, and introducing N-methylation and O-methylation. S-adenosyl methionine (SAM) is the primary methyl group and is considered as one of the most used substrates in enzymatic reactions after ATP (Walsh, 2006). In addition, N-methylation is irreversible while O-methylation can be reversed. This is important in numerous cellular processes such as, embryonic development, genomic imprinting, X-chromosome inactivation, and preservation of chromosome stability. A review by Ehrlich, 2002 suggested that errors in methylation could give rise to various diseases including cancer. Hypermethylation is considered as a biomarker for cancer.

N-Acetylation

Transferring an acetyl group to the nitrogen is considered as the N-acetylation and this can be occurred in both reversible and irreversible mechanisms. $80-90\%$ of eukaryotic

proteins are acetylated by replacing amino acids with an acetyl group which involves the cleavage of N-terminal methionine by methionine aminopeptidase (MAP) (Walsh, 2006). This was first identified in histones and cytoplasmic proteins, where acetylation seems to play a major role in cell biology rather than transcription regulation (Glozak et al., 2005). Furthermore, phosphorylation, ubiquitination and methylation can change the biological function of acetylated proteins (Yang and Seto, 2008).

S-Nitrosylation

This is a useful post-translational modification to stabilise proteins, regulate gene expression and act as nitric oxide (NO) donors. S-nitosylation is a reversible PTM and half-lives of the S-nitrothiols (SNOs) are very short. The attachment of NO with free cysteine residues is the main action in this process. Specific cysteine residues undergo S-nitrosylation; therefore it is not a random event.

Proteolysis

Amino acid sequences break their bonds to fold in a stable manner. This process is known as proteolysis. This is thermodynamically favourable and not reversible. Thus, this process is tightly regulated to avoid uncontrolled proteolysis from temporal and spatial mechanisms. The main advantage of proteolysis is that it converts an inactive or non-functional protein into an active proteins post-translationally. This may also be involved in the removal of signal peptides and/or N-terminal methionine.

2.1.4.2 Protein Degradation by Post-translation Regulation (PTR)

Proteins are continuously being synthesised and degraded in all living organisms. However, some proteins degrade faster due to post-translational regulation. Covalent modifications occur as PTMs serve as markers for the protein degradation process (Stadtman, 1990). Callis, 1995 studied plant proteins to investigate the relationship between post-translational modifications and regulated degradation. They observed that multiple levels of post-translational modifications regulate protein degradation of a single species. In addition, Stadtman, 1990 found that phosphorylation post-translation modification catalyses the protein degradation process. This study also claimed that most of the post-translationally modified proteins have a common structural feature that serves as a recognition signal for protein degradation. Early experiments by Levine, 1983; Holzer and Heinrich, 1980 also explained how oxidative modifications act as markers for intracellular proteolytic degradation.

Proteasome is a common process for degrading unneeded or damaged proteins by proteolysis or by enzymatic reactions such as post-translational modifications. These proteasomes can be found in the cytosol and nucleus of eukaryotic cells. Proteins which are degraded by proteasome are tagged with a multimer 76 amino acid polypeptide ubiquitin by ubiquitination post-translation modification. This chemical reaction is known as 26s proteasome and after the degradation of the tagged protein, ubiquitin monomer is released and can be reused with another protein. The review of Hartmann-Petersen and Gordon, 2004 discusses several proteins which interact with 26s proteasome and degrade faster. Rpn10/Pusl/S5a binds with multi-ubiquitin chains and interacts with proteasome via its N-terminus to catalyse the degradation process (Hofmann and Falquet, 2001). Dsk2/Dph1, Rad23/Rhp23 and Ddi1 are also degraded faster following ubiquitination during the 26s proteasome process (Hofmann and Bucher, 1996). Martinez et al., 2003 showed that ABCA1-PEST sequence phosphorylation regulates ABCA1 calpain degradation. Several *in-vitro* experiments were carried out to show that phosphorylation with PEST motifs regulated ABCA1 protein degradation and reduces the overall protein expression level. Firstly, they compared protein expression levels of a PEST deleted mutation called ABCA1delPEST and a wild-type ABCA1 protein in HEK293 cells. Here the PEST deletion mutation damages the protein degradation attribute, therefore the wild-type ABCA1 gave $3.9 \pm 0.4\%$ mean fold reduction of the protein concentration with respect to ABCA1delPEST protein (Wang et al., 2003). Secondly, a wild-type ABCA1 protein expression level was compared with phosphorylation sites mutated ABCA1 proteins. These mutations impair protein degradation activity in ABCA1 proteins. Therefore, the phosphorylation site mutated ABCA1 expression levels gave higher expression levels compared to wild-type ABCA1 proteins (mean fold $3.4 \pm 0.3\%$ with MutAAAA in Thr-1286 site and $3.3 \pm 0.3\%$ with MutASSA in Thr-1305 site). Thus, these experiments suggest that post-translational modifications such as phosphorylation and acetylation catalyse protein degradation and reduce total protein expression levels.

2.2 Importance of Post-translational Regulation

As described in the introduction, modelling at the transcriptome-proteome interface allow us to develop a novel machine learning framework to detect post-translationally regulated proteins by exploring the model failures. Here we describe the importance of post-translationally regulated proteins and significance of developing a computational approach to detect post-translationally regulated proteins. Chemical and structural changes occurring post-translationally have implications on several pathological and physiological processes. Gene expression regulation, differentiation of epithelial terminals and apoptosis are some physiological processes and Alzheimer's disease (AD), rheumatoid arthritis and Parkinson's disease (PD) are some pathological diseases (Gong et al., 2005; Oueslati et al., 2010). Citrullination is a process in which an amino acid

is converted into an organic compound called citrulline. Peptidylarginine deiminases (PADs) replace the primary =NH (ketamine) group with a =O group (ketone). These protein modifications are normally related to diseases in which the immune system attacks citrullinated proteins. Citrulline proteins are generated as a result of post-translational modifications. Therefore post-translational modifications play a major role in finding a cure for the above mentioned diseases.

PTMs are also useful in the study of severe cancers in the human body. Bode and Dong, 2004's review says that the mutations which disrupt $p53$ function, occur in approximately half of all human cancer cases. These mutations are caused by a multitude of covalent post-translational modifications, including phosphorylation, acetylation, methylation and ubiquitination. Therefore PTMs in $p53$ protein contribute to tumourigenesis.

Furthermore, proteins with PTMs are used as candidates for biomarker discovery for several cancer types. Phosphoproteomics provides information on tumour growth signalling pathways where the clinician can make rational decisions in prognosis that drives the treatment strategy. Gulmann et al., 2005's study discovered that the phosphorylation of protein B/Akt kinase improves the efficacy of drug treatment of recurrent tumours. Similarly, Glycomics also provides a fertile class of molecular structures to perform as biomarkers for cancer. Ovarian carcinoma biomarker CA125 was discovered using structural properties of glycosylation (Wong et al., 2003) and glycomics profiling can be used to obtain serum biomarker for hepatocellular carcinoma by isolating N-linked oligosaccharides (Block et al., 2005). Table 2.1 is taken from a review by Krueger and Srivastava, 2006 which shows proteins with PTMs that are used as cancer biomarkers.

Table 2.1: Example of post-translational modifications of proteins being used as cancer biomarkers (Krueger and Srivastava, 2006).

PTMs	Protein Localization		
	Nuclear	Cytosolic, intracellular organs	Plasma membrane
Phosphorylation	$pRBs$, $p53$, histones, HDACs, $STAT-3$	PTEN, Akt, MAP kinases, death related protein kinase, cyclindependent kinases, GP73	EGFRs, PDGFR, ILK, osteopontin
Glycosylation			CD44; galectins; CA125, CA19-9; MUC4, osteopontin; prostate antigen;
Ubiquitination	$p53$, NF-kB, HDACs	Inhibitor of apoptosis proteins	
Prenylation		Ras, Rho, Braf	G-protein- receptors
Methylation	Histones, DNA polymerase		
Acetylation	$p53$, GATA TF histones, HDACs, NF-kB		

Post-translationally modified proteins have also been used in therapeutic interventions in cancer treatment. PTM perturbation can slow the growth of cancer cells, if the cancer causing pathways are required or involved in post-translational regulation (Krueger and Srivastava, 2006). HDAC inhibitors are involved in cancer progression and histone acetylation and deacetylation can control the cell growth (Benson et al., 2006; Marks et al., 2004). Similarly, autophosphorylation targeting tyrosine kinase inhibitors are used to treat non-small-cell lung cancer (Lynch et al., 2004; Paez et al., 2004). These examples illustrate the importance of detecting post-translational regulation to study on different diseases.

Mass spectrometry is the standard way of detecting post-translational modified proteins (Mann and Jensen, 2003a). This relies on the mass alternation of tryptic peptides where chemical modifications of the amino acid chains can be detected. However, this process is very complex, expensive and time consuming (Yates et al., 2009; Mann and Jensen, 2003a) and high quality protein samples are required. Therefore, it is important to develop a computational approach cut down the experimental workspace needed to detect post-translationally regulated proteins.

2.3 Machine Learning Inference of High-throughtput Omic Measurements

Transcriptome data analysis using microarray and RNA-Seq technologies have rapidly improved over the last decade. Brazma et al., 2001 proposed a microarray data representation standard called Minim Information About Microarray Experiment (MIAME). This reporting technique archives microarray data with minimum information, which will facilitate the establishment of proper transcriptome databases and public repositories. Functional classification of genes along signalling pathways was carried out by Brown et al., 2000a using transcriptome measurements. Molecular classification of critical diseases such as cancer (Golub et al., 1999a) and subspace projection of transcriptome expression data (Zheng-Bradley et al., 2010; Brunet et al., 2004) are some successful applications of microarray data analysis. Genomic analysis using regulatory networks is another important area involving high-throughput transcriptome measurements (Liao et al., 2003; Luscombe et al., 2004b; Sanguinetti et al., 2006). Furthermore, some other important properties such as mRNA decay (Wang et al., 2002), translation efficiency (Washburn et al., 2003; Arava et al., 2003) and transcription factor binding locations along the genome (Harbison et al., 2004) were measured and analysed to discover interesting biological properties. However, transcriptome data itself is strong enough to provide an approximate picture of cellular functions. Important biological phenomena such as dynamical cellular function and differential spatio-temporal behaviours can be revealed by using extensive mathematical and computational models. Chen et al., 2004

have developed a mathematical model based on biochemical rate equations using for budding yeast cell data. This model is largely successful in explaining the phenotypes of mutants. Parameter estimation of heat shock response (Liu and Niranjan, 2012), heat beat modelling (Zhang et al., 2000) and measuring robustness of circadian oscillations with respect to molecular noise (Gonze et al., 2002) are some examples for complex mathematical models. Scientists have also looked into spatial selectivity in morphogenesis using early *Drosophila* embryo (Houchmandzadeh et al., 2002; Liu and Niranjan, 2011) to understand complex biological properties.

Transcriptome and proteome abundance data and their turn-over rates of mammalian organisms were measured in a parallel manner for thousands of genes by Schwanhäusser et al., 2011. They measured these properties for more than 5000 mammalian. Schwanhäusser et al., 2011 used pulse labelling with amino acid and 4sU to quantify mRNA and protein expression levels and turnover in a parallel manner. A population of exponentially growing NIH3T3 mouse fibroblasts were employed with this experiment. Newly synthesized RNA abundances in 2h with 4SU were measured with parallel to protein levels. These RNA samples were divided into newly synthesized and pre-existing fractions. mRNA half-lives were calculated using the ratios between newly synthesized RNA/total RNA and pre-existing RNA/total RNA. By using these measurements, they observed that the mRNA levels can explain around 40% of the variability in protein levels. Most recently, Schwanhäusser et al., 2013 have analysed synthesis and degradation processes jointly to determine the responsiveness of the cellular proteome. These different transcriptome and proteome data have been used with different data-driven models to uncover important biological mechanisms (Pancaldi and Bähler, 2011; Muppirala et al., 2011; Rogers et al., 2008; Kannan et al., 2007).

2.3.1 Microarray and RNA-Seq Transcriptome Measurements

Here we discuss the most commonly used high-throughput microarray and RNA-Seq transcriptome measuring techniques in detail. In Chapter 5, a thorough machine learning analysis is carried out to observe the importance of the high numerical precision of these transcriptome measurements.

2.3.1.1 Microarrays

Microarray transcriptome measurements provide a vast amount of information about different cell types (Kai et al., 2005) and tissues (Chan et al., 2009). Microarray technology has been used in many biological studies including gene expression changes during a development time course (Graveley et al., 2011; Arbeitman et al., 2002), classification of different diseases (Golub et al., 1999b; Tuna and Niranjan, 2010) and discovery of

new phenotypes in different species (Zhang et al., 2007). Initially the microarray chip was designed as a set of short oligonucleotides to represent genomic DNA. In a modern microarray, we can find patches of short oligonucleotide probes which are complementary to investigating transcripts. Thus, prior knowledge of transcriptome is important to develop a microarrays (*i.e.* normally sequence data or open reading frames are incorporated to develop the probes in microarray (Malone and Oliver, 2011)). There are four main steps to measure mRNA abundance using microarray technology.

1. **Sample Preparation**: mRNA is extracted from the total RNA with ployA tail and converted into cDNA by reverse transcription and random priming. Fluorescent dye is added to label the sample. For example cancer and normal samples can use red and green colour fluorescent dye respectively.
2. **Hybridization Procedure**: Next, the samples are added to microarray chip where complementary probes of transcriptome are employed and hybridization process is carried out. Hybridization allows two cDNA strands from different sources to be paired with each other.
3. **Acquisition**: The microarray chip is scanned by a laser to obtain the gene expression levels by looking at the colour intensities.
4. **Data Process**: Finally, the image file is processed into a text file by converting colour intensities into gene expression values.

Figure 2.3 shows an example involving measurements of gene expression levels of male and female fly heads of *D.pseudoobscura* using the microarray technique (Malone and Oliver, 2011). There are both advantages and disadvantages of this technique. The main advantages are that it is fast, user friendly, capable of measuring high-throughput transcriptome data, and most importantly, it is a low cost technique compared to other transcriptome measuring techniques. Disadvantages include the need to have prior knowledge of the transcripts to design the microarray probes, meaning it is unable to detect new exon junctions and isoforms in genomes (Reimers, 2010; Shi et al., 2006).

2.3.1.2 RNA-Seq

RNA-Seq is a recently developed powerful transcriptome measuring technique. This uses massive parallel deep sequencing of RNA molecules to detect gene expression levels with high sensitivity and accuracy (Wang et al., 2009; Mortazavi et al., 2008; Fu et al., 2009; Marioni et al., 2008). Illumina HiSeq 2500, High-Sequencing and Applied Bio-system's SOLID are examples of fast instruments used to perform this deep sequencing process.

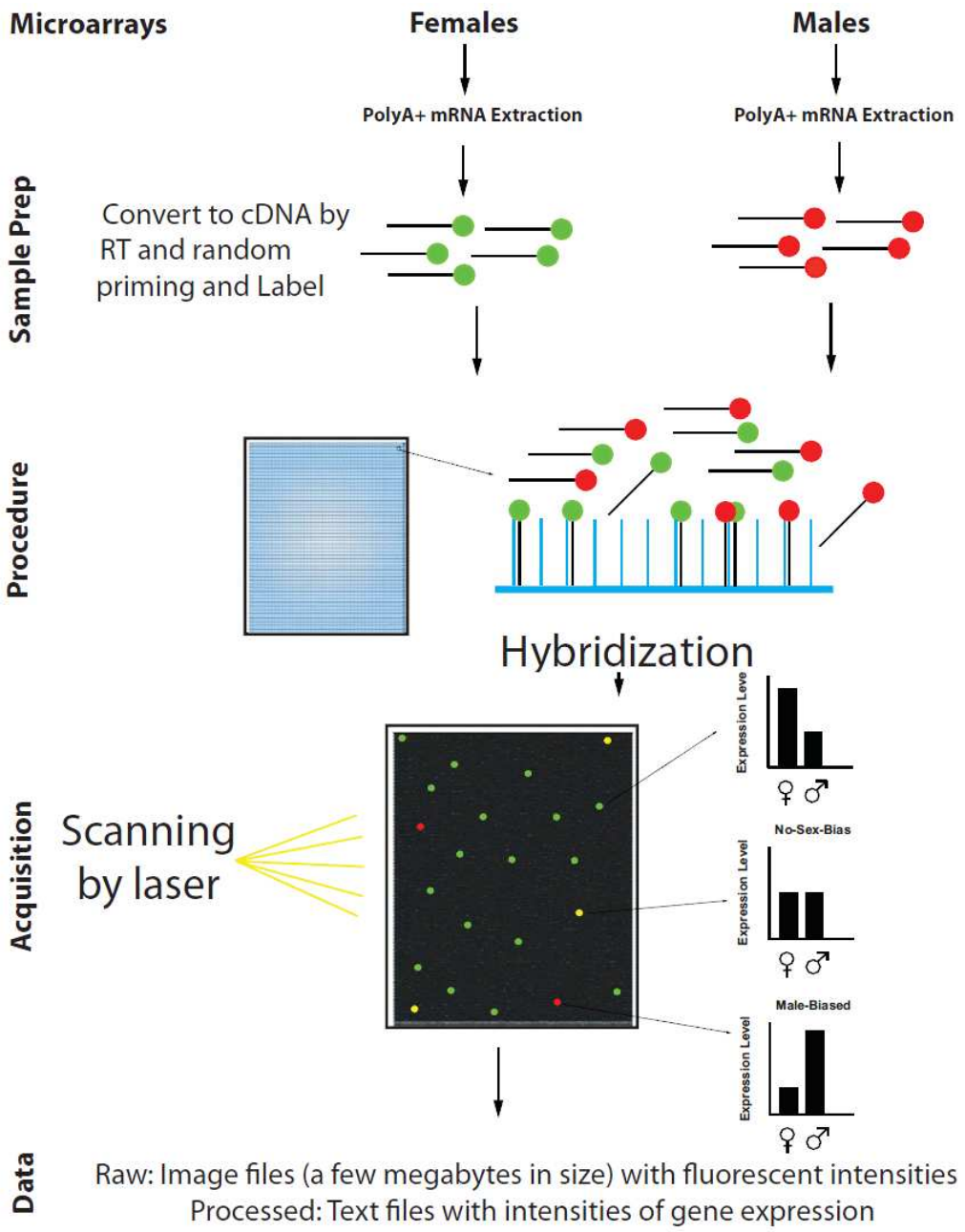


Figure 2.3: Microarray transcriptome measuring work flow taken from Malone and Oliver, 2011: This shows the four main steps in the microarray measuring process for male and female fly heads of *D.pseudoobscura*. Dominant gene expression levels for each probe are indicated in red/green or yellow colour.

The main difference between microarray and RNA-Seq is that the latter uses direct sequencing instead of hybridization to capture transcripts of interest. Here, the transcript reads are mapped to the reference genome and the mapped read counts are used to obtain the gene expression levels. There are several advantages of using RNA-Seq compared to microarray technology. Since RNA-Seq directly involves gene sequences, it has

the capability to detect novel exon junctions, RNA- editing events and other properties of the gene structure without any prior knowledge needed. In fact, unlike microarray technology, RNA-Seq method does not need prior knowledge of the gene structure to measure the gene expression levels (Wang et al., 2009; Malone and Oliver, 2011). RNA-Seq can also be used to detect gene expression levels for species, whose full genome sequence information is not available (Malone and Oliver, 2011), whereas microarray is limited due to the sequence divergence of different species. Another interesting advantage of RNA-Seq is that it has the potential to quantify expression levels of different isoforms from the same transcript generated by the alternative splicing process (Richard et al., 2010; Trapnell et al., 2010). In fact, high-throughput genome sequencing allows the detection of new isoforms which have not been detected before. Thus, RNA-Seq is a more powerful tool for obtaining qualitative properties of the transcriptome with respect to microarray technology. However, Malone and Oliver, 2011's study showed that both microarray and RNA-Seq quantitative values are highly correlated and these measurements provide a consistent story with respect to quantitative properties. Figure 2.4 shows that both microarray and RNA-Seq measurements are highly correlated for *D.pseudoobscura* female ($R^2 \approx 0.9$) and male ($R^2 \approx 0.9$) fly head data (Malone and Oliver, 2011).

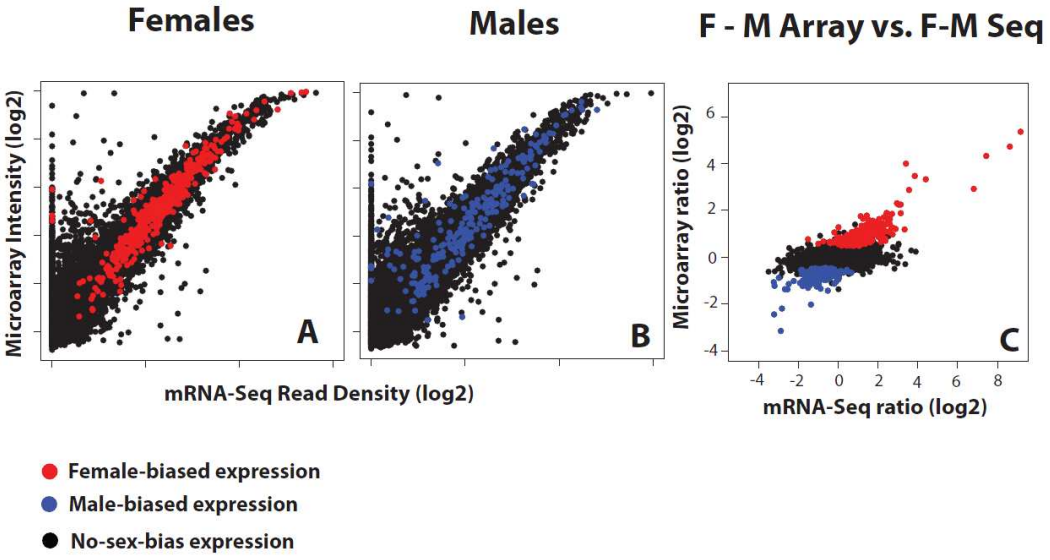


Figure 2.4: Comparison of microarray and RNA-Seq gene expressions of *D.pseudoobscura* by Malone and Oliver, 2011. Both female (A) and male (B) gene expressions are highly correlated, but not the combined (C) no sex bias gene expressions

However, there are some practical disadvantages of the RNA-Seq technique. The main disadvantage is that it is very expensive compared to microarray technology. For example, a 12-plex array costs less than $100 per sample whereas RNA-Seq process costs around $1000 per sample (Malone and Oliver, 2011). Moreover, RNA-Seq requires a more in depth sample sequencing process. If the gene is highly expressed a small amount

of sequencing is sufficient, otherwise it needs a great deal of sequencing of samples to obtain many reads to give an accurate measure (Graveley et al., 2011). Finally, the data storage space needed is high for RNA-Seq measurements. Microarray image files only use around 30MB file space, but RNA-Seq deep sequencing data use more than 600MB to store all the sequencing data and the read file. Thus, RNA-Seq requires a large storage space to store important sequencing data.

Figure 2.5 shows the main steps of the RNA-Seq measuring procedure using the same *D.pseudoobscura* male and female fly head data by Malone and Oliver, 2011. This process is more complex compared to microarray technology. The following are the four main stages of the RNA-Seq measuring process.

1. **Sample Preparation**: The tissue sample is collected and mRNA with polyA tails are isolated for the measuring process. Next, the mRNA is fragmented using alkaline hydrolysis and random hexamer primers are employed to generate reverse transcribed double stranded cDNAs. Finally, oligonucleotide adaptors are added to the ends of the cDNAs .

2. **Procedure**: Fragments are then injected into a flow cell. This consists of a glass slide containing lawns of complementary oligonucleotide adaptors so that the fragments can bind. Afterwards, the isothermal bridge amplification process is used to amplify the fragments and generate clusters of DNA clones.

3. **Acquisition**: In this step, DNA strands are synthesised in a cycle. In each iteration, sequencing reagents are added to the flow cell and the bound base is detected using a laser.

4. **Data Processing**: Finally, FASTQ files are obtained at the acquisition process with millions of sequences mapped to the reference genome and the gene expression levels are computed by counting the number of mapped reads.

2.3.2 Precision Based Inference of Transcriptome Measurements

According to the literature RNA-Seq is considered as a more sensitive and accurate transcriptome measuring technique (Wang et al., 2009; Malone and Oliver, 2011; Fu et al., 2009; Marioni et al., 2008). Importantly, it also has high qualitative properties compared to the microarray method including being able to detect transcriptome properties like exon junctions, different isoforms etc. Thus, in our work we would like to investigate whether these two techniques perform differently in a quantitative environment. In fact, we perform a machine learning inference of the RNA-Seq and microarray techniques using high (continuous) and low (binary) precision measurements. There are only a few mRNAs found in a single cell. Therefore, high-throughput transcriptome measuring

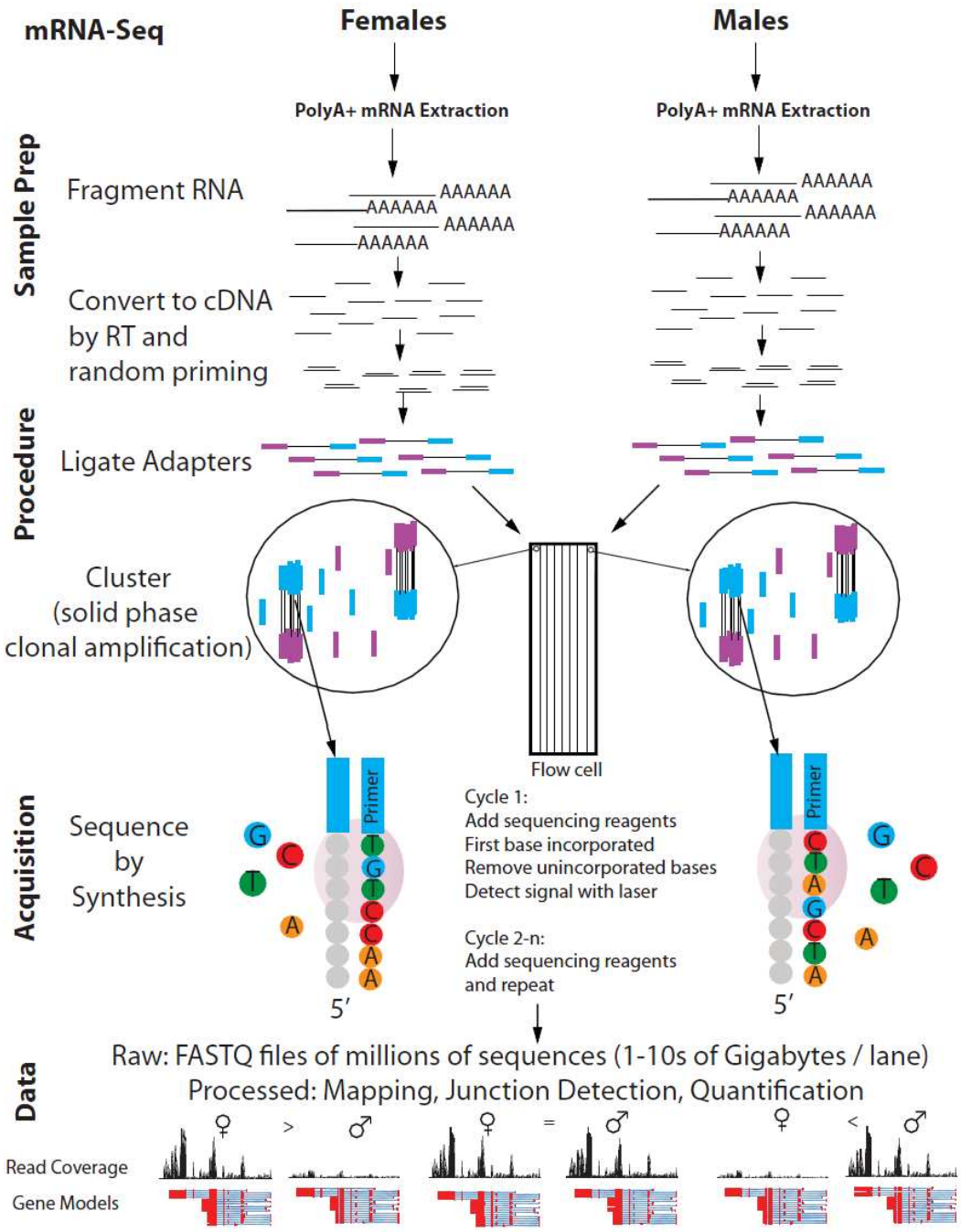


Figure 2.5: RNA-Seq transcriptome measuring work flow by Malone and Oliver, 2011: Gene expression quantification of male and female fly heads of *D.pseudoobscura*. Read counts mapped to a scaled region of genome are considered as the index of gene expression levels.

techniques such as microarray and RNA-Seq, amplify the number of mRNA with respect to the cell population to obtain more accurate relative abundances (Nygaard et al., 2003; Ozsolak and Milos, 2011). This amplification process generates high numerical precision

in mRNA measurements. Tuna and Niranjan, 2010 showed that high precision microarray does not provide additional information with respect to binary interpretation of gene switched on/off status. Tuna and Niranjan, 2009 also showed that binary microarray data has the capability to perform machine learning classification tasks with high accuracy. Since, the microarray data have been experimented previously, in our work we mainly focus on RNA-Seq measurements. We perform a quantitative analysis using RNA-Seq and microarray measurements at the transcriptome level and a qualitative analysis using our PTR detection framework at the proteome level. In Chapter 5, we discuss this experimental setup and results in detail.

2.3.3 Machine Learning Inference Techniques

The two main categories of machine learning inference techniques are known as supervised and unsupervised learning. Supervised learning uses target information of the data to perform inferences, *e.g.* class labels should be given to train the learning algorithm. In contrast, unsupervised learning methods do not need any target information as they understand the data by looking at similar patterns and grouping them into clusters (Cristianini and Shawe-Taylor, 2000; Rogers and Girolami, 2012b). These supervised and unsupervised techniques are used in different areas such as computer vision (Esposito and Malerba, 2001; Cipolla et al., 2012), artificial intelligence (Anderson et al., 1986), finger print analysis (Wilson et al., 1994) etc. However, our main focus lies on making inferences on transcriptome data using gene expressions. Therefore in this section, we describe the algorithms which are mainly used for gene expression analysis. The state-of-the-art classification approach in gene expression analysis uses Support Vector Machines (SVM) (Brown et al., 2000b; Statnikov et al., 2008). However, in our study we also used the K-Nearest Neighbour (KNN) method to compare the classification performance with SVM (Dudani, 1976; Singh et al., 2002). We also used two main clustering techniques to analyse transcriptome measurements. Those are K-means (MacQueen et al., 1967; Causton et al., 2009) and Spectral clustering (Shi and Malik, 2000; Higham et al., 2007) learning algorithms.

2.3.3.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) learning algorithm was introduced by Vapnik (1998). SVM was initially used to classify linearly separable data. However, kernel SVM was later introduced to deal with non-linearly separable data. Here we explain linear, kernel and one class SVM learning algorithms. One class SVM algorithm is a popular algorithm to detect outliers and we will discuss about this approach as we are interested in detecting outliers in our main hypothesis.

Linear SVM

This learning method performs well with linearly separable data sets. Figure 2.6 shows an example of two-dimensional linearly separable data. SVM selects optimal hyperplane to separate classes by maximizing the margin between the closest data points of each class.

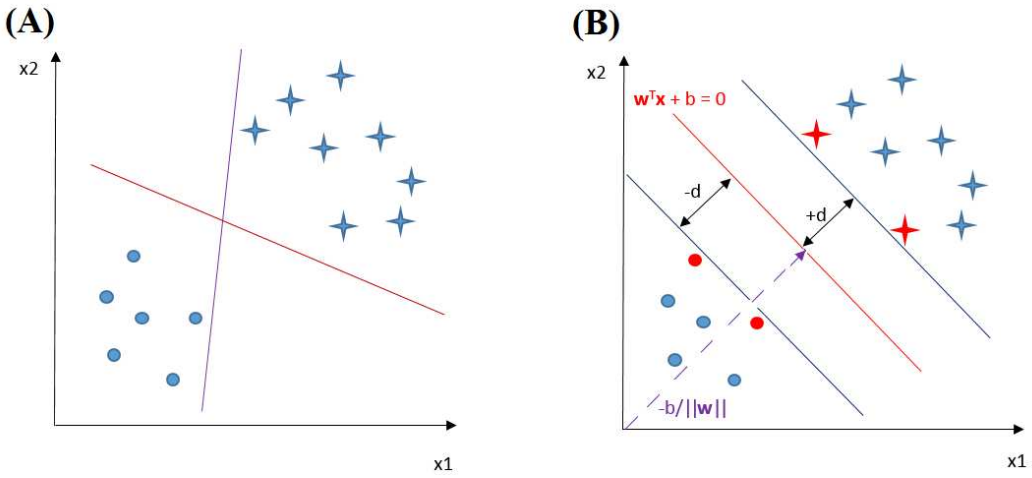


Figure 2.6: Example of Linearly Separable Data: (A) Two possible hyperplanes to linearly separate data, (B) Red line demonstrates the optimal hyperplane with maximum margin from two classes (red data points are the support vectors of the two classes)

Suppose $\boldsymbol{x_i}$ data points can be divided into two classes $y_i = \{-1, +1\}$ and the hyperplane can be defined as below,

$$y(\boldsymbol{x_i}) = \boldsymbol{w}^\top \boldsymbol{x_i} + b, \forall i \tag{2.1}$$

where $\boldsymbol{w}$ weight vector and b bias are learnt by the data. The perpendicular distance from the origin to the hyper plane is defined as $|b|/\|\boldsymbol{w}\|$. $+d$ and $-d$ define the distance for the optimal margin from the support vectors of the two classes. Training data should satisfy the following constraints to obtain the optimal margin with maximum distance d.

$$\boldsymbol{w}^\top \boldsymbol{x_i} + b \geq +1, \text{ for } y_i = +1, \tag{2.2}$$

$$\boldsymbol{w}^\top \boldsymbol{x_i} + b \leq -1, \text{ for } y_i = -1. \tag{2.3}$$

We can combine these inequalities and define the constraint as below:

$$y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) - 1 \geq 0, \ \ \forall i. \tag{2.4}$$

For data points which satisfy Equation (2.2) on the hyperplane $h1 : \boldsymbol{w}^\top \boldsymbol{x_i} + b = +1$ with normal $\boldsymbol{w}$ have the perpendicular distance from the origin as $|1 - b|/\|\boldsymbol{w}\|$ and for data points satisfying Equation (2.3) with hyperplane $h2 : \boldsymbol{w}^\top \boldsymbol{x_i} + b = -1$ have $|-1 - b|/\|\boldsymbol{w}\|$ distance from the origin. Therefore, the margin between the two classes is $\frac{|1-b|}{\|\boldsymbol{w}\|} - \frac{|-1-b|}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|}$. In order to find the two hyperplanes with maximum margin using Equation (2.4) for all data points can be obtained by minimizing $\|\boldsymbol{w}\|$ as below:

$$\begin{aligned} \min \ & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \ & y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) - 1 \geq 0, \quad \forall i. \end{aligned} \tag{2.5}$$

Lagrange multiplier is used to obtain the solution for the above optimization problem. This is a strategy to find local minima or maxima of a function subject to equality constraints.

For example assume following optimization problem,

$$\min \qquad f(x, y) \tag{2.6}$$

$$\text{s.t.} \qquad g(x, y) = 0 \tag{2.7}$$

Both f and g functions should get partial derivatives. Thus, a new variable λ called Lagrange multiplier is introduced.

$$L(x, y) = f(x, y) + \lambda . g(x, y), \tag{2.8}$$

This will yield necessary conditions for optimality in constrained problems (Vapnyarskii, 2012; Lasdon, 2013).

Thus, Equation (2.5) can be written as below using a Lagrange multiplier:

$$L(\boldsymbol{w}, b, \lambda) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \lambda_i [y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) - 1] \tag{2.9}$$

Derivatives of L with respect to $\boldsymbol{w}$ and b will produce the solution to the optimization problem:

$$\frac{\partial_{L(\boldsymbol{w},b,\lambda)}}{\partial \boldsymbol{w}} = \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{x}_i = 0 \tag{2.10}$$

$$\frac{\partial_{L(\boldsymbol{w},b,\lambda)}}{\partial_b} = \boldsymbol{w} - \sum_{i=1}^{n} \lambda_i y_i = 0 \tag{2.11}$$

Soft Margin SVM

In some cases, training data cannot be separated without any error. In order to overcome this issue, Cortes and Vapnik (1995a) suggested the soft margin SVM technique, which separates the training data with minimum error. In this method, a set of variables ϵ_i are introduced to allow the possibility to violate the constraint which is given by Equation (2.4). Therefore, we can re-write Equation (2.4) as below:

$$y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) \geq 1 - \epsilon_i, \quad \forall i \tag{2.12}$$

by relaxing the separation constraint, where any large value for ϵ_i will be able to satisfy the constraint. Thus, to penalized the effect of large ϵ_i, the constraint is multiplied by a constant C to have a trade-off with training error and margin maximization, which can be written as below:

$$\begin{aligned} \min \ & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \epsilon_i^p \\ \text{s.t.} \ & y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) \geq 1 - \epsilon_i; \epsilon_i \geq 0, \quad \forall i \end{aligned} \tag{2.13}$$

where a small value for C maximizes the training error and minimizes the margin and a large value for C minimizes the training error and maximizes the margin. Therefore, penalty parameter C needs to be tuned to obtain the best value. Cross-validation can be used to tune hyper-parameter C.

Lagrange multiplier is used to obtain solution for this optimization problem. Equation (2.13) can be written as below;

$$L(\boldsymbol{w}, b, \epsilon, \lambda, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \epsilon_i - \sum_{i=1}^{n} \lambda_i[y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) - 1 + \epsilon_i] - \sum \alpha\epsilon_i \quad (2.14)$$

The solutions for the Equation (2.14) can be obtained by taking derivatives of L with respect to $\boldsymbol{w}$, b and ϵ:

$$\frac{\partial_{L(\boldsymbol{w},b,\epsilon,\lambda,\alpha)}}{\partial_{\boldsymbol{w}}} = \boldsymbol{w} - \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{x}_i = 0 \quad (2.15)$$

$$\frac{\partial_{L(\boldsymbol{w},b,\epsilon,\lambda,\alpha)}}{\partial_b} = \boldsymbol{w} - \sum_{i=1}^{n} \lambda_i y_i = 0 \quad (2.16)$$

$$\frac{\partial_{L(\boldsymbol{w},b,\epsilon,\lambda,\alpha)}}{\partial_\epsilon} = C - \lambda_i - \alpha_i = 0 \quad (2.17)$$

where $C \geq \lambda_i \geq 0$

2.3.3.2 K-Nearest Neighbours (KNN)

This is a very popular classification technique due to its simplicity and good performance with empirical data. It does not assume any parametric form to obtain the decision boundary and rather the distances between input objects are considered (Dudani, 1976; Singh et al., 2002; Rogers and Girolami, 2012a). Figure 2.7 represents a binary class problem to be solved using KNN method where $K = 3$. The red stars A and B are the testing data objects and 3 closest neighbouring (because $K = 3$) data objects are used to determine the classes of these testing data points. In fact, the majority class from the neighbours is assigned to the testing data object. Thus, data point A will be assigned to the class circle and data point B will be assigned to the class square.

The main drawback of the method occurs when K is an even number and an equal number of samples are found from each class as neighbours. One solution is to assign data objects to a random class from neighbouring classes. However, this is not a good solution if we are testing the data object using more than one iteration. Therefore, it is better to use odd numbers for K in binary classifications. We can also determine the value of K using cross validation technique. Following are some distance measuring

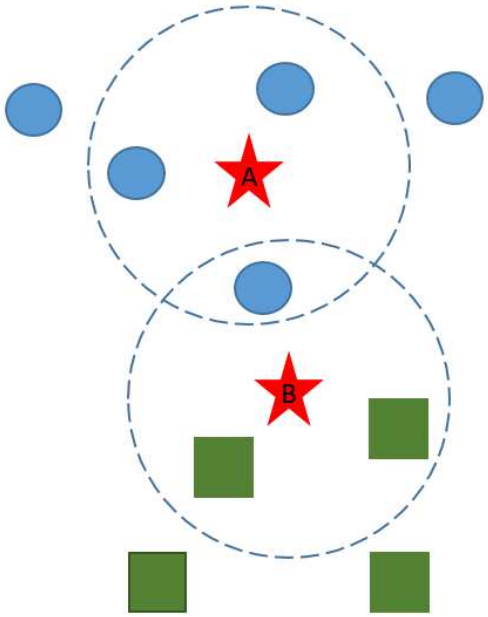


Figure 2.7: Example for K=3 Nearest Neighbour classification: Data point **A** falls to the majority class circle and point **B** classifies to the class square.

functions:

Euclidean Distance

$$d1 = \sqrt{\sum_i (x_i - y_i)^2}, \quad \forall i \tag{2.18}$$

Manhattan Distance

$$d2 = \sum_i |x_i - y_i|, \quad \forall i \tag{2.19}$$

Minkowski Distance

$$d3 = \left[\sum_i (|x_i - y_i|)^p \right]^{1/p}, \quad \forall i \tag{2.20}$$

We can use either of the above distance measuring functions to perform the classification.

2.3.3.3 Gaussian Mixture Model (GMM)

Gaussian Mixture Model is a combination of M Gaussian (normal) distributions where each distribution has its own mean and standard deviation $N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$. Equation (2.21) represents a GMM with M components:

$$p(\mathbf{x}) = \sum_{k=1}^{M} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) \tag{2.21}$$

where π_k satisfies $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{M} \pi_k = 1$ is the mixing coefficient. Figure 2.8 shows a two component GMM example.

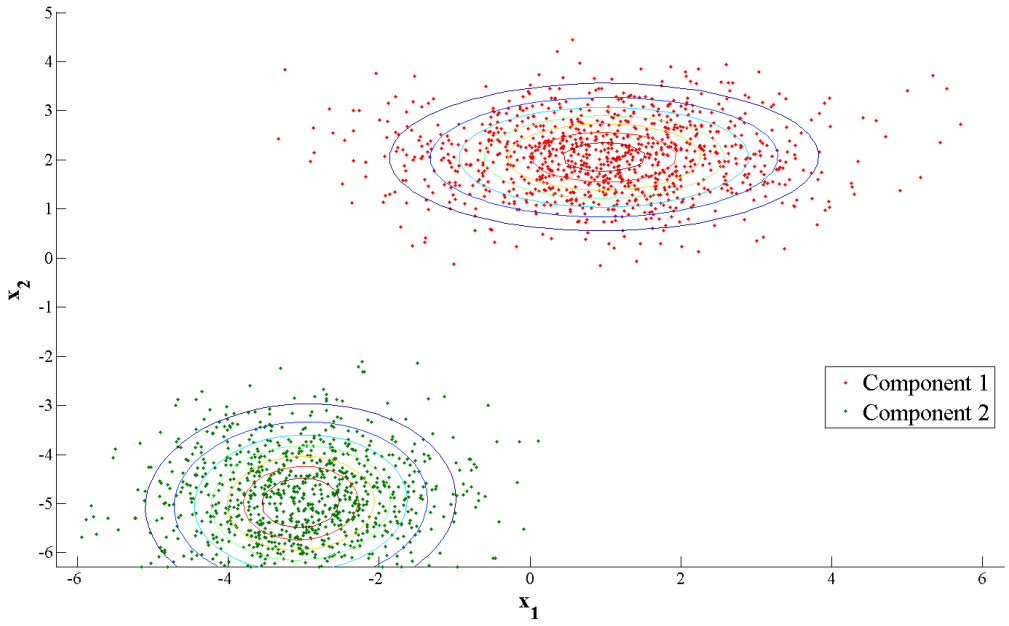


Figure 2.8: Mixture of two Gaussian distributions

GMM is used to model data in a high dimensional space by maximizing the likelihood function with respect to the model parameters ($\boldsymbol{\mu}_k$, $\boldsymbol{\sigma}_k$ and π_k). Equation (2.22) shows the log likelihood function of GMM;

$$L = \log p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{p=1}^{P} \log \sum_{k=1}^{M} \pi_k N(\mathbf{x}_p|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) \tag{2.22}$$

Expectation-Maximization (EM) algorithm (Dempster et al., 1977a) is used to maximize the log likelihood function to estimate unknown parameters. It is difficult to obtain

optimum parameters with a summation inside the logarithm. Thus, EM algorithm derives a lower bound for this likelihood where a function of parameters ($\boldsymbol{x}$, $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$) are always lower than or equal to L. This method maximizes the lower bound instead of L. Jensen's inequality is used to obtain the lower bound: the log of the expected value of a function $f(z)$ is always greater than or equal to the expected value of the log $f(z)$ (Equation (2.23)).

$$\log \boldsymbol{E}_{p(z)}\{f(z)\} \geq \boldsymbol{E}_{p(z)}\{\log\ f(z)\} \tag{2.23}$$

There are four main steps in applying EM algorithm to GMM (Dempster et al., 1977a). Those are;

1. Initialize distribution parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ and obtain the log likelihood.

2. Expectation (E) Step : Calculate the posterior probabilities of $P(z|\mathbf{x})$ where z represents the latent variable indicating probability of $\mathbf{x}$ belongs to which component. Current parameter values are used to calculate the posterior probability:

$$\psi(z_{pi}) = \frac{\pi_i N(\mathbf{x}_p|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)}{\sum_{i=1}^{M} \pi_i N(\mathbf{x}_p|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)} \tag{2.24}$$

3. Maximization (M) Step: Re-calculate the parameter values by maximizing the likelihood function. Obtain the derivatives of log $p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ and estimate the new parameters $\boldsymbol{\mu}^{new}$ and $\boldsymbol{\sigma}^{new}$. Use a Lagrange multiplier and maximize log $p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ with respect to π (π has the condition of $\sum \pi = 1$) to estimate the new $\boldsymbol{\pi}^{new}$.

$$\boldsymbol{\mu}_k^{new} = \frac{1}{Q_k} \sum_{p=1}^{P} \psi(z_{pk}) \mathbf{x}_p \tag{2.25}$$

$$\boldsymbol{\sigma}_k^{new} = \frac{1}{Q_k} \sum_{p=1}^{P} \psi(z_{pk})(\mathbf{x}_p - \mu_k^{new})(\mathbf{x}_p - \mu_k^{new})^T \tag{2.26}$$

$$\pi_k^{new} = \frac{Q_k}{p} \tag{2.27}$$

where

$$Q_k = \sum_{p=1}^{P} \psi(z_{pk}) \tag{2.28}$$

4. Calculate the log likelihood with new parameter values and check for convergence. If the convergence condition is not satisfied return to step 2.

Additionally, it is important to determine the number of components per GMM to best represent true distribution of data. There are few model selection techniques to over-come this problem. Akaike Information Criterion (AIC) (Akaike, 1998) and Bayes Information Criterion (BIC) (Schwarz et al., 1978) are the most common model selection techniques in GMM. These model selection techniques reward the goodness of the fit, but also penalise the model complexity.

Suppose k is the number of parameters to be estimated and L is the likelihood function. AIC value of the model can be obtained as below;

$$AIC = 2k - 2.\ln(L) \tag{2.29}$$

Similarly BIC value can be obtained using following function.

$$BIC = k.\ln(n) - 2.\ln(L) \tag{2.30}$$

where n is the number of samples in the observation data. Here the asymptotic result is derived under the assumption that data distribution belongs to an exponentially family.

2.3.3.4 K-means Clustering

Clustering is an unsupervised learning model which does not use any prior knowledge (class labels) to divide data into groups. This is widely used with gene expression analysis (MacQueen et al., 1967).The noise and variability of transcriptome measurements reduce the accuracy of clustering results for gene expressions.

K indicates the number of clusters specified by the user and a random centroid will be assigned for the each cluster. Euclidean distance metric is used to assign each data point to its closest centroid with the minimum distance. Next, with the new clusters, K new centroids will be assigned and the process is repeated until there is no change in the centroids. However, the whole clustering process will depend on the initial centroids (MacQueen et al., 1967).

2.3.3.5 Spectral Clustering

This is also an unsupervised learning method which does not need any prior knowledge of data to perform clustering task. Spectral clustering was introduced by Shi and Malik, 2000 in the field of image processing. Later, several authors used this approach for computational biological problems including gene expression analysis (Higham et al., 2007; Tritchler et al., 2005; Xing and Karp, 2001). This technique uses eigenvectors of the pairwise similarity matrix of the data to partition them into relevant groups. Negative exponential of Euclidean distance function (Equation (2.31)) is the most widely used similarity matrix in spectral clustering.

$$\boldsymbol{A}(i,j) = \exp(-\frac{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2}{\sigma^2}) \tag{2.31}$$

Main tool in spectral clustering is the Laplacian matrix. The field of studying these matrices are known as spectral graph theory (Chung, 1997). There are normalized and unnormalized Laplacian matrices. However, here we consider the normalized Laplacian matrix which is mainly used in clustering algorithms. Symmetric normalized Laplacian matrix is defined as below;

$$\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2} = I - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \tag{2.32}$$

where $\boldsymbol{D}(i,i) = \sum_j \boldsymbol{A}(I,j)$ and $\boldsymbol{W}$ stands for the weight matrix with $w_{ij} = w_{ji} \geq 0$ (Chung, 1997).

Normalized Laplacians satisfy following properties:

1. For every $f \in \mathbb{R}^n$, $f'\boldsymbol{L}f = \frac{1}{2}\sum_{i,j=1}^{2} w_{ij}\left(\frac{f_i}{\sqrt{d_i}} - \frac{f_i}{\sqrt{d_j}}\right)^2$ where indicator vector $\mathbf{1}_A = (f_1, \ldots, f_n)'$ and $d_1, \ldots, d_n$ are degrees of diagonal matrix $\boldsymbol{D}$
2. 0 is an eigenvalure of $\boldsymbol{L}$ with eigenvector $\boldsymbol{D}^{1/2}\mathbf{1}$
3. $\boldsymbol{L}$ is positive semi-definite and have n non-negative real value eigen values $0 = \lambda_1 \geq \cdots \geq \lambda_n$.

Following four steps are involved in the spectral clustering learning algorithm which invloves similarity matrix and Laplacian matrix (Shi and Malik, 2000):

1. Similarity matrix $\boldsymbol{A}(i,j)$ is calculated for gene i and j using Equation (2.31)

2. Compute Laplacian matrix using similarity matrix $\boldsymbol{A}$:

$$\boldsymbol{L} = \boldsymbol{D}^{-1/2} \times \boldsymbol{A} \times \boldsymbol{D}^{-1/2} \tag{2.33}$$

   where $\boldsymbol{D}(i,i) = \sum_j \boldsymbol{A}(I,j)$

3. Obtain the generalized eigenvalue decomposition of $\boldsymbol{L}$:

$$(\boldsymbol{D} - \boldsymbol{L})y_i = \lambda_i \boldsymbol{D} y_i \tag{2.34}$$

4. Obtain the eigenvector which gives the second smallest eigenvalue (smallest eigenvalue of $\boldsymbol{L}$ can be 0)

This reduces the disassociation between the clusters and improves the association within the clusters themselves to obtain an accurate clustering analysis. There are several methods to select the most suitable eigenvector. Xing and Karp, 2001 used leukaemia data set from Golub et al., 1999b's study and clustered microarray data into the AML and ALL cancer subtypes using the method introduced by Shi and Malik, 2000. However, Higham et al., 2007 used a different level of eigenvectors and clustered microarray data into two or more classes. They observed that second eigenvector is more suitable for binary class problems. Additionally, Tritchler et al., 2005 changed the similarity matrix approach and used a covariance matrix on a leukaemia data set and was able to cluster the AML and ALL cancer subtypes more accurately.

2.3.4 Proteomics Techniques

This section provides an overview of proteomics measuring techniques. We describe their applications, challenges and strengths.

Proteomics is the study of charactering and measuring proteins in a cell or organism. This can be used to measure qualitative and quantitative properties such as protein profiles, protein-protein interactions, compare two or more protein samples and importantly to detect post-translationally modified proteins. However, processing and analysis protein data is very complex and involves multiple steps (Chandramouli and Qian, 2009; Kearney and Thibault, 2003). In fact identifying post-translational modifications requires prior knowledge of the type and the modification and time consuming (Jensen, 2004; Mann and Jensen, 2003b). Thus, it is important to build a computational model to cut down the experimental time and cost to detect post-translationally regulated proteins. We believe our data-driven approach provides a solution for this critical problem.

There are three types of proteomics categories and those are expression, structural and functional proteomics. Expression proteomics analysis protein abundances in large

scale to identify differentially expressed proteins in two or more samples. This category is capable of detecting proteins which are useful in drug or biomarker discovery. Two Dimensional gel Electrophoresis (2-DE) and Fluorescence 2D Differential gel electrophoresis (2D-DIGE) are some examples for expression proteomics category. Structural proteomics involves in analysing protein structures in large scale to identify main proteins to a particular sample based on the protein structure. Isotope-Coded Affinity Tag (ICAT), Isobaric Tag for Relative and Absolute Quantitation (iTRAQ) and Stable Isotope labelling with Amino Acids in Cell Culture (SILAC) are common to structural proteomics. Finally, functional proteomics analyse biological functions of unknown proteins and cellular mechanisms at the molecular level. Matrix-Assisted Laser Desorption/Ionization (MALDI) and MALDi with a tandem Quadrupole/Time-of-Fight (MALDI-QqTOF) mass spectrometers are used in functional proteomics analysis. Table 2.2 shows some common proteomics technologies and their strengths and limitations obtained from the proteomics review by Chandramouli and Qian, 2009. This shows that not all techniques have the capability of detecting PTMs.

Table 2.2: Proteomic techniques, their applications, strengths and limitations by Chandramouli and Qian, 2009.

Technology	Application	Strengths	Limitations
2DE	Separate Proteins Profiling quantitative expressions	Relative quantitative PTM information	Poor separation of acidics and low abundance
DIGE	Separate Proteins Profiling quantitative expressions	Relative quantitative PTM information High sensitivity Reduction of intergel variability	Proteins without lysine cannot be labeled, Expensive, Requires special visualization tools
ICTA	Chemical isotope labeling for quantitative proteomics	Reproducible ans sensitive Detect low expressions	Acidic proteins are not detectable
iTRAQ	Isobaric tagging of peptides	Multiplex several samples Relative quantitative High-throughput	Increase sample complexity Require fractionation of peptides before MS
MS	Primary tool for protein identification and characterization	High sensitivity and specificity, High-throughput, Qualitative and quantitative PTM information	No individual method to identify all the proteins, Not sensitive to weak spots

Following are descriptions on few proteomics technologies.

2DE

2DE is a key technique in purifying proteins and obtaining protein expression levels from complex samples based on their isoelectric points and molecular weights. Following are

the main steps in this process;

1. Sample solubilisation: Proteins need to be processes and solubilized in isoelectric focusing (IEF) compatible reagents such as urea.
2. Isoelectric Focusing (IEF): When proteins are solubized in the reagent, electric field is used to push the proteins through the acrylamide gel which incorporates a pH gradient. Thus, proteins move until it reaches the isoelectric point (pl). pl is the pH with no net charge.
3. SDS Electrophoresis: Next proteins are solubilzed again in sodium dodecyl sulphate (SDS) and are separated by the molecular weights on an orthogonal axis (2D). This technique aligns these proteins along two axes: isoelectric point vs. molecular weight.
4. Protein detection and image analysis: Finally, mass spectrometry (MS) is used to measure the protein abundances, post-translational modifications and other properties based using mass to charge ratios.

Isotope-Coded Affinity Tag (ICTA)

This is a gel-free technique and more reproducible than 2DE technique. ICTA is mostly used chemical isotope labeling technique, which employs with MS technology (Shiio and Aebersold, 2006). The chemical reagents consists of three main elements; a thiol reactive group, a biotin segment and isotopically coded linker (Chandramouli and Qian, 2009). This mixture of labeled proteins are then digested by trypsin and separated by liquid chromatographic (LC) separation. Tandem mass spectrometry (MS/MS) is used to identify peptides and relative abundances are measured using LC peaks (Shiio and Aebersold, 2006)

Isobaric Tag for Relative and Absolute Quantification (iTRAQ)

This is a well-known proteomics technique to measure relative and absolute protein abundances. iTRAQ has several advantages such as the ability use multiplexing up to four different samples in one MS based experiment, labeling process can be performed after cell or tissue lysis and simple analysis procedure in iTRAQ increases the precision and accuracy of the measurement. This also contains several stages to measure protein concentration. First the protein samples are digested using an enzyme such as typsine to generate protein peptides. These peptide mixtures are then label with a different iTRAQ reagent and combine them into one sample mixture. Finally, LC-MS/MS technique is used to quantify peptides by comparing the intensities of reporter ion signals.

Figure 2.9 shows the workflow of iTRAQ. The main disadvantage of iTRAQ technique is that this requires a powerful fractionation method due to enzymatic digestion process (Chandramouli and Qian, 2009).

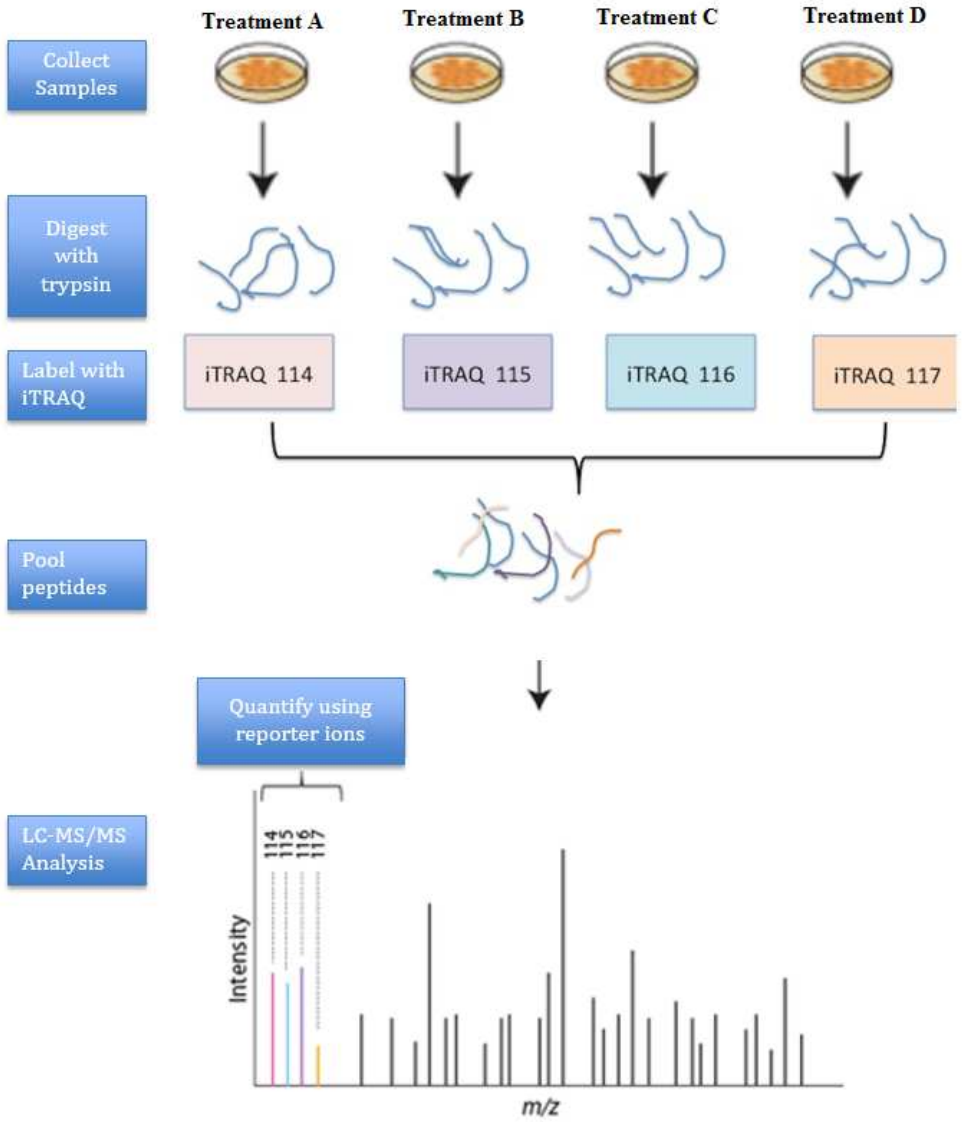


Figure 2.9: iTRAQ technique work flow

Mass Spectrometry (MS)

Mass spectrometry is the primary technique in identification of proteins, regardless the protein separation using gel-based or gel free method. MS has evolved drastically over the last decade and it consists of three main components; ion source, mass analyser and ion detection system (Chandramouli and Qian, 2009). These components are used in main four stages in MS analysis (Figure 2.10 taken from Clark, 2015);

1. Protein iInization: Molecules are ionized by knocking one or more electrons off to obtain positive ion gas.
2. Acceleration: Ions are accelerated to provide same kinetic energy.
3. Deflection: Ions are deflected using a magnetic field based on their mass to charge ratio
4. Detection: Lighter ions will be detected electrically.

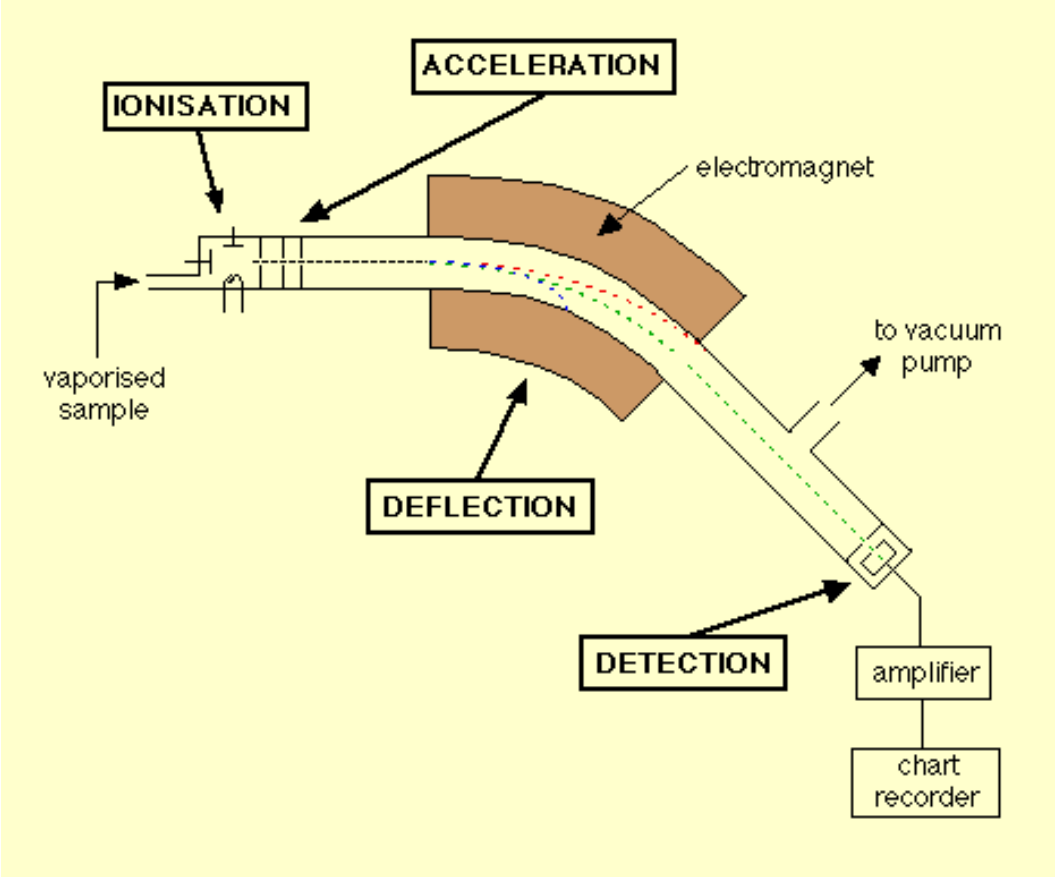


Figure 2.10: Mass Spectrometry work flow taken from Clark, 2015

Simple MS only measures mass, however tandem mass spectrometry (MS/MS) are employed to determine amino acid sequences (Dubey and Grover, 2001). MS/MS combines two different MS separations steps where the trypsin-digested peptides are fragmented after the liquid phase separation. Liquid chromatography (LC) mass spectrometry employs an analytical chemistry technique that combines physical separation ability of liquid chromatography or HPLC. This enhances the sensitivity and selectivity of mass spectrometry technique. Prior knowledge on protein sites and specific modifications are used to detect post-translational modifications. However, MS might not be able to detect all the new sights and post-translational modifications. A combination of techniques will be able to reveal PTMs, but a comprehensive proteomics is still not

feasible. These multiple enrichment techniques are more complex and time consuming (Jensen, 2004). Therefore, our computational model comes in handy to detect post-translationally modified before testing them on high-throughput proteomics techniques such as mass spectrometry.

2.4 Joint Analysis of Transcriptome and Proteome

We now turn into integrated analysis of transcriptome and proteome measurements. First, we discuss previous work related to correlation analysis between these two *omic* data. We then analyse several data-driven modelling approaches performed by previous authors to uncover useful biological information regarding cell regulation. These background material provide guidance for the development of a robust data-driven model at the transcriptome-proteome interface to detect post-translationally regulated proteins as outliers.

2.4.1 Correlation of Transcriptome and Proteome Data

Most of the previous authors have looked into the correlation between mRNA expression data and the corresponding protein expressions to understand their relationship (Gygi et al., 1999; Futcher et al., 1999; Greenbaum et al., 2003; Beyer et al., 2004). However, there are some other biological properties such as codon bias, ribosomal properties and protein half-life that help to understand the relationship between transcriptome and proteome measurements. These properties allow for more accurate predictions.

- **Codon Bias and Codon Adaptation Index (CAI)**: There are different frequencies for synonymous codons (*i.e* coding for the same amino acid) and this aspect is known as codon bias. Codon adaptation index (CAI) is a proper measure of codon bias (Maier et al., 2009);
- **Ribosome Density and Ribosome Occupancy**: These are the main components to determine translational efficiency (the number of proteins translated per mRNA over time) (Greenbaum et al., 2003). Ribosome occupancy is the fraction of mRNA molecules attached to at least one ribosome and ribosome density is the number of ribosomes active with mRNAs for a unit transcript length (Brockmann et al., 2007; Arava et al., 2003);
- **Protein Half-Life**: This is a crucial factor in mRNA-protein correlation. The cellular life time of the protein depends on several aspects. Post-translational processing, intrinsic protein stability and first or terminating amino acids are some

of the main properties. N-end rule is a simple and accurate method to measure protein half-life (Varshavsky, 1997).

Gygi et al., 1999 used *Saccharomyces cerevisiae* (yeast) mRNA and protein data to observe the relationship between these measurements. However, the correlation between these data is insufficient to predict protein expression levels by simply using mRNA measurements. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) and serial analysis of gene expression (SAGE) techniques were employed to measure protein and mRNA levels, respectively. Pearson correlation for 106 genes by only using mRNA and protein levels gave 0.935. However, this number is highly biased by a small number of genes with very large mRNA and protein expression levels. 69% of the data genes (73 of 106) had very small mRNA levels with less than 10 copies per cell and these genes gave only a 0.356 correlation coefficient. This study also incorporated codon bias and protein half-life measurements for deeper analysis. Codon bias data was extract from the Yeast Protein Database (YPD) (Hodges et al., 1998) and N-end rule was employed to calculate protein half-lives. These properties showed that post-translational mechanisms such as protein half-life controls the correlation between mRNA and protein levels in mammalian cells, where no predictive correlation can be found. Another interesting observation is that codon bias did not appear as a predictor either to protein or mRNA because the distribution patterns of mRNA and protein data with respect to codon bias are highly variable. Therefore, this study shows that transcriptome data itself provides very little information for predicting protein expression data.

However, the sampling of the yeast proteome in the study by Futcher et al., 1999 gave a good correlation between mRNA and protein abundances ($R^2 = 0.76$). SAGE and 2D gel techniques were employed to measure mRNA and protein abundances, respectively. Codon adaptation index data was extracted from YPD spread sheets (Hodges et al., 1998). Gygi et al., 1999 used Pearson product-moment correlation coefficients, which is a parametric statistical method and requires bivariate normal distribution. Thus, Pearson correlation is more appropriate if mRNA and protein data are normally distributed. However, mRNA and protein expressions are far from normal distribution. Therefore, in this study, Futcher et al., 1999 used Spearman rank coefficient correlation which is a non-parametric statistics. CAI and protein abundance also gave $R^2 = 0.80$ Spearman correlation which is similar to mRNA-protein correlation. Futcher et al., 1999 also found that protein turnover was insignificant for protein expression prediction. This experimental setting discovered some proteins with post-translational modifications (phosphorylation) by running 2D gel with labelled cells. Examining the behaviour in differential centrifugation experiments with 2D gels provided a global view of the yeast proteome. Thus, different statistical analysis and differences in data resulted in different conclusions being made by these two authors (Futcher et al., 1999 and Gygi et al., 1999).

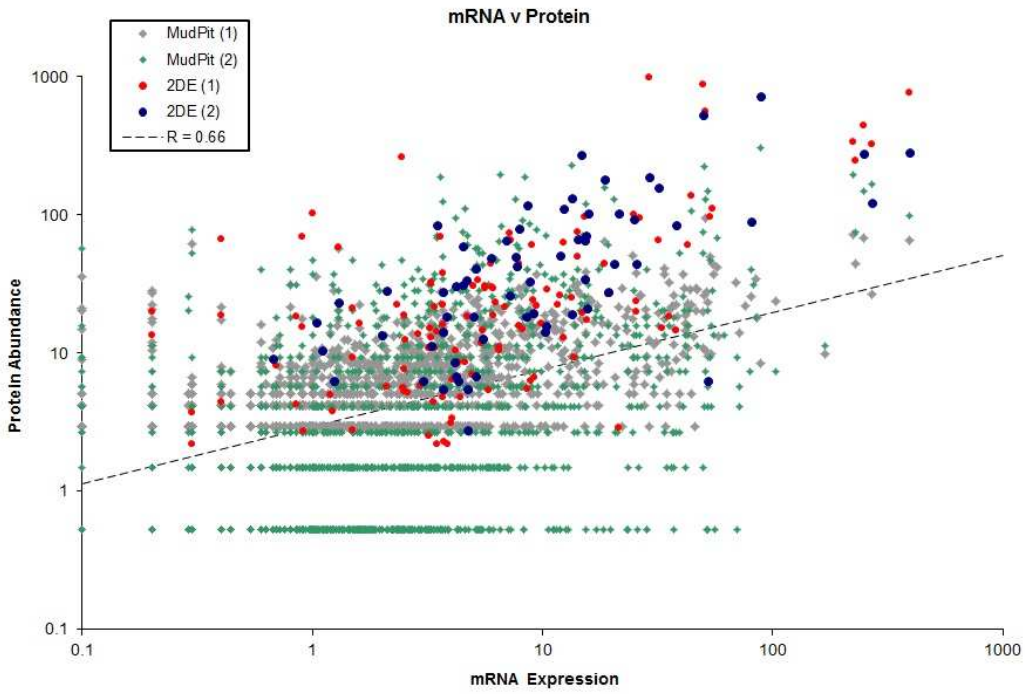


Figure 2.11: Correlation of mRNA and protein data by Greenbaum et al., 2003. This plot represents the correlation of the mRNA data and their newly compiled protein abundance data. mRNA axis is in copies per cell and the protein axis is in thousand copies per cell.

A review of mRNA and protein abundance correlations focusing on the yeast genome was carried out by Greenbaum et al., 2003. This study provides details about the methods used for determining protein levels: *eg.* 2D electrophoresis, mass spectrometry approach including isotope-coded affinity-tag bases protein profiling and multi-dimensional protein identification technology. Here they merged mRNA and protein data from Gygi et al., 1999 and Futcher et al., 1999. Figure 2.11 shows the correlation between the compiled mRNA and protein data (0.66). They observed that presumably there are three main reasons for the absence of the correlation between mRNA and protein data. Those are (1) complicated and varied post-transcriptional mechanisms, (2) differences in protein half-life measurements *in vivo* and (3) noise errors in both expression data. In order to represent the variation of genes along the yeast cell cycle, the standard deviation was divided by the average expression level. The mechanism of the protein degradation process was explained using transcriptome and proteome expression data;

$$dP(i,t)/dt = SE(i,t) - DP(i,t)$$

where P is the protein concentration i at time t, and the mRNA expression level for protein P is represented as E. The general protein synthesis rate per mRNA is shown by S and finally the general rate of protein degradation per protein is expressed as $\mathbf{D}$ (Greenbaum et al., 2003). This model represents that the change in protein abundance

over the time is equal to the total translation minus protein degradation rate. Therefore, increasing protein degradation rate by post-translational modifications, reduces the total protein concentration levels of post-translationally regulated proteins. Further, Greenbaum et al., 2003 also investigated the effects of codon adaptation index (CAI) with mRNA and protein correlation. However, genes with high level of CAI, did not show a very good correlation between mRNA and protein expressions.

Beyer et al., 2004 used large scale data of protein and mRNA abundance, translation status and transcript length of yeast (*Saccharomyces cerevisiae*) genome to investigate the relationship between transcription, translation and protein-turnover on a genome-wide scale. Several data sets were combined together to generate the final large scale data set. This study demonstrates that mRNA concentration, translation rate (which is determined by ribosome density and occupancy) and protein degradation are the three factors determining protein concentration. Therefore, mRNA abundance and translation rate related properties are important in developing a protein abundance predictor. These data also showed that the correlation between protein and mRNA ($R^2 = 0.58$) varies among different cellular compartments and functional modules. In addition, they have explained a novel approach to correct large microarray signals for a saturated bias and combined them with proteomic data to gain new insights of the cellular regulation. Further, ribosome density was identified as a good property to measure translation efficiency. In fact, transcript length with ribosome density is a better descriptor for translation efficiency with respect to the ORF length.

These data samples were also used by Wu et al., 2008 and mRNA and protein half-life information was also included to expand this analysis. A multiple regression was carried out with the sequence derived features to observe their relevance with mRNA-protein correlation (Nie et al., 2006).

$$y_i = \alpha + mRNA_i \times \beta + \sum_{j=1}^{m} \beta_j x_{ij}$$

where x_{ij} refers to the jth sequence feature such as codon usage and m is the number of features. However, they did not use this regression approach to predict protein abundance similar to Tuller et al., 2007 or our study. The main purpose of this approach is to observe the relationship between post-transcriptional and other biological properties with the mRNA-protein correlation. Protein half-life was observed as an important regulatory factor which contributed 16.9% of the mRNA-protein correlation. However, mRNA half-lives only contributed 0.2% for the mRNA-protein correlation. These results are also consistent with a similar work for *D.vulgaris* by Nie et al., 2006. Further, codon usage (8.9%) and amino acid usage (7.7%) also showed a great contribution for

protein translation. Therefore, these features can be used to develop a robust protein abundance predictor.

Investigating Biological Mechanisms Using mRNA and Protein Correlation

Here we discuss the important biological mechanisms identified by join analysis of mRNA and protein expression data.

In the study of complementary profiling of gene expression at transcriptome and proteome levels in *Saccharomyces cerevisiae*, Griffin et al., 2002 used an integrated genomic and proteomic approach to investigate the effects of carbon source perturbation on steady-state gene expressions growing on galactose or ethanol. This experiment showed that the correlation between mRNA and protein expression levels varies rapidly with the above changes. A non-parametric correlation analysis using Spearman rank correlation method gave a 0.21 correlation between mRNA and protein ratios which is lower than simple mRNA-protein correlations reported by both Gygi et al., 1999 and Futcher et al., 1999. This study demonstrates that protein expression levels significantly increase with the galactose carbon source, but do not significantly change with mRNA abundance. Therefore, joint analysis provides a better picture of the underlying mechanism. Five essential proteins were identified with respect to galactose to glucose conversion, namely; Gal1p, Gal2p, Gal7p, Gal15p and Gal10p. Genes and their quantities involved in the metabolism of ethanol through respiration have also been recognized. Another interesting result was the discovery of key regulatory genes involved in both tricarboxylate and the glyoxylated cycles. All these results were obtained by examining the correlation between mRNA and protein data.

Washburn et al., 2003 studied complex clustering of correlated transcriptome and proteome data of *Saccharomyces cerevisiae* to understand the biochemical properties of protein pathways. According to this study, the Spearman rank correlation between mRNA and protein expression was weakly positive (0.45 for 678 loci). In order to interpret the correlation of the data set, authors used a loci analysis with clustering on mRNA and protein data based on protein pathway and protein complexes. At the pathway level, mRNA and protein expressions were highly correlated (0.99) not only on a loci by loci basis but also as a whole pathway. However, for aromatic amino acid and hitidine biosynthetic pathways, mRNA and protein data only correlated at the pathway level. Thus, these results showed that there are pathway and sub pathway levels at the transcriptional interface. Protein complex clustering method identified biological components such as amino acid and hitidine biosynthetic pathways of *Saccharomyces cerevisiae*, which are not detectable by the oligonucleotide array analysis of mRNA expression. Further, several protein complexes such as Holeenzymes, SPT and GTPases were over-expressed in the cell culture and mRNA and protein expression data did not

correlate with these complexes. This indicates that there is a post-transcriptional control of protein expression functions at the protein complex level or sub complex level. Thus, this study encourages to use clustering technique in transcriptome/proteoem inferences.

Correlation between mRNA and protein expression data in *Desulfovibrio vulgaris* was studied by Nie et al., 2006. Whole genome microarray data and LC-MS/MS protein expression data for this organism were studied under three different conditions. Multiple regression approach was used to investigate the correlations between mRNA and protein data with sequence derived features such as codon bias (similar to the study by Wu et al., 2008). However, the experimental results showed that mRNA data alone only explains $20 - 28\%$ of the total variation of proteomic data suggesting that mRNA data alone cannot use to determine protein concentration. Authors also explained the three aspects to improve the potential of the current model. Those were: (1) improving the semi-quantitative measurements of protein concentration; (2) using more accurate RNA and protein stability data and (3) measuring mRNA decay rate since the expression levels were measured. Considering the correlation of mRNA and protein data on functional categories, Nie et al., 2006 observed that central intermediary metabolism, energy metabolism, and transport and binding protein categories have more pronounced correlations.

A label free MS based protein concentration technique was used by Ning et al., 2012 to measure brain stem and liver tissue protein levels of mice. Afterwards, a joint analysis of mRNA and protein data was carried out to investigate the correlation between these two properties. This is the first attempt to compare RNA sequence and microarray mRNA data with label-free protein data. In order to obtain biological properties, DAVID (Da W. H. et al., 2008) GO enrichment analysis was used. 75% of MitoCarta genes were annotated as mitochondrial in GO terms. In addition, genes annotated as ribosomal did not show a good correlation between mRNA and protein data. Ribosomal genes undergo several post-translational activities such as phosphorylation, acetylation and methylation. Therefore protein expression levels change through protein degradation due to the above post-translational modifications reducing the correlation between mRNA-protein measurements. They also looked for the correlation coefficient and average abundance of genes for each GO category and observed a noticeable difference in correlation between genes in different categories. Finally, the authors claimed that the results of this experiment can be useful to develop a robust computational pipeline for gene and label-free protein data in future studies.

2.4.2 Data-Driven Models for Transcriptome and Proteome Data

We now look at data-driven approaches used in previous studies to jointly analyse transcriptome and proteome measurements which will be highly relevant with our study to

develop a novel data-driven framework at the transcriptome-proteome interface.

2.4.2.1 Classification Approach

Support Vector Machine (SVM) (Cortes and Vapnik, 1995b) and Random Forest (RF) (Breiman, 2001) machine learning techniques have been mainly used for transcriptome and proteome classification problems. Pancaldi and Bähler, 2011 carried out an experiment to characterize and predict protein-mRNA interactions in the yeast genome. They investigated how well RNA binding protein (RBP) and RNA interactions can be predicted using RBP and other features of mRNA rather than using sequence motifs. More than 100 different features of mRNA and protein data were used for this experiment. For example mRNA half-life, GO annotations, secondary structure of proteins, relative abundance of amino acid, codon bias etc. Prediction RBPs using mRNA data was performed as a binary classification where the interaction can either be present or absent. SVM and RF predictors were used to discover the relationship. Pancaldi and Bähler, 2011 obtained 70% accuracy in 2-fold cross validation using RF and 68% using a SVM classifier. They also reported that the 5-fold and leave-one-out method gave similar results. Despite the high accuracy with the known targets, the prediction of uncharacterized RBPs remains challenging because of the limited experimental data available. However, a more complete data set with a wider range of RBPs and a strong feature selection process for SVM approach will enhance the power of their predictions. Comparing the correlation of **human** transcriptome and proteome data is a hugely encouraging for future studies. Even though their comparison was not quite successful, this can be a starting point to model the interaction between mRNA and protein data.

Muppirala et al., 2011 also performed a classification prediction to predict mRNA-protein interactions similar to Pancaldi and Bähler, 2011's study by only using sequence information. SVM and RF were used as the main predictors. They predicted whether or not mRNA and protein pairs interact by giving RNA and protein sequences as predictor inputs. Muppirala et al., 2011 also compared their results with Pancaldi and Bähler, 2011's results. RF and SVM classifiers achieved 68% and 61% accuracies respectively. These prediction accuracies are lower than those recored in the study by Pancaldi and Bähler, 2011. However, Muppirala et al., 2011 achieved 78% for RF and 65% for SVM classifiers by using all $13,243$ mRNA-protein pairs, where Pancaldi and Bähler, 2011 used only 5166 due to missing features. Thus, the technique used by Muppirala et al., 2011 can do a reasonably good job of predicting mRNA and protein interactions by only using sequence information.

Support Vector machine classifiers have been used in many RNA-protein interaction studies. Wang et al., 2012b used LLE (Local Linear Embedding) algorithm (Roweis and Saul, 2000). This is a fast non-linear dimensionality reduction algorithm, used to

project high dimensional feature space to lower dimension space. This algorithm was able to reduce the feature space from 440 to 80. Further, a binary SVM classifier was also used to predict the interactions between RNA and protein pairs. LLE-SVM model and simple binary SVM gave accuracies of 95.8% and 90.5% respectively. Thus, the LLE feature extraction method improves the accuracy of the SVM predictor.

2.4.2.2 Clustering Approach

Several authors have performed microarray expression data analysis using clustering technique (Eisen et al., 1998; Heard et al., 2005). Eisen et al., 1998 used cDNA microarray of budding yeast (*Saccharomyces cerevisiae*) and human data. They observed that gene expression data with similar functions are grouped together in both organisms. Heard et al., 2005's study was based on co-clustering technique to investigate on the immune defense response to multiple experimental challenges using a time series cDNA microarray data of *A. gambiae* mosquito. The joint analysis of mRNA and protein data using the clustering approach is an important milestone in system biology (Rogers et al., 2008).

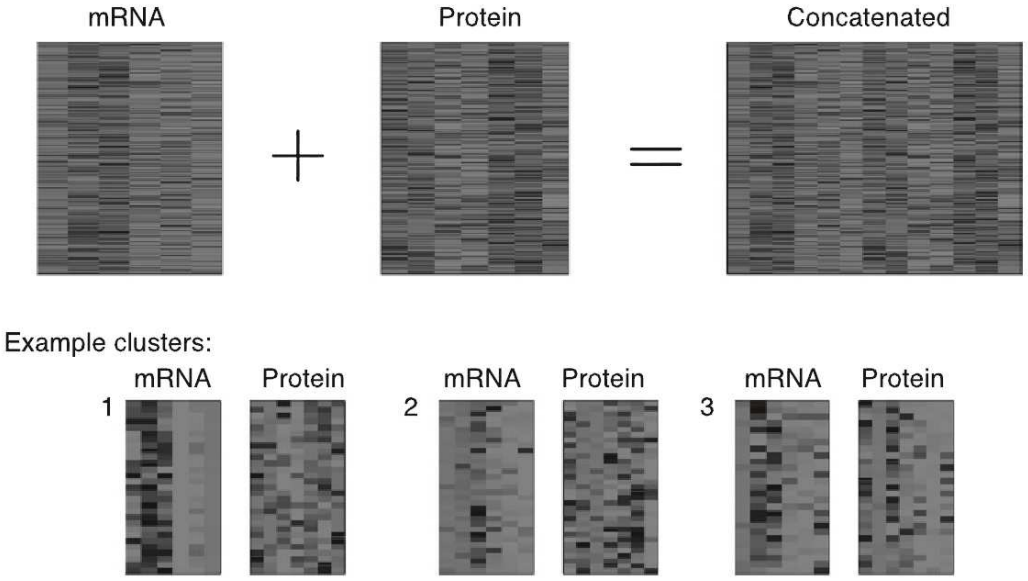


Figure 2.12: An example of a concatenated clustering: mRNA and protein data from Rogers et al., 2008. The top row shows the two data sets and in each data set rows represent genes and columns represent time-points. The bottom row represents three clusters obtained from the concatenated cluster analysis (Rogers, 2011)

Rogers et al., 2008 used a couple mixture cluster model to reveal important biological insights using mRNA and protein data which were collected along a time series from a

human breast epithelial cell line (HMEC). Concatenation is the simplest clustering approach for two real-value transcriptome and proteome data sets. However, this method is rather inflexible and doubles the size of feature space without increasing the number of data points and therefore increases the overall complexity (Figure 2.12). The second method is to analyse two data sets independently, however this approach will remove important relationships from the data. Therefore, Rogers et al., 2008 employed a probabilistic clustering model that couples mRNA and protein data in a sensible and flexible manner. They assumed that there are two separate mixture models for both mRNA and protein data. If mRNA data has k components and protein data has j components the prior distribution of the two data sets will be $p(k, j)$. If k and j are independent then $p(k|j) = p(k)p(j)$. At the other extreme if they have a one-to-one relationship, the joint distribution that would be $p(k|j) = p(k)\delta_{kj}$, where $\delta_{kj} = 1$ if $k = j$ and 0 if otherwise. However in Rogers et al., 2008's model, $p(k, j)$ was considered as a parameter to be inferred by the model. Assuming $\mathbf{X}$ is the mixture model for mRNA and $\mathbf{Y}$ be the mixture model for protein data. Δ is the used to represent some parameters required to define the distributions. Therefore, the data distribution is as follows;

$$p(\mathbf{X}, \mathbf{Y}|\Delta) = \prod_{g=1}^{G} \sum_{k=1}^{K} \sum_{j=1}^{J} p(\mathbf{x}_g|\Delta_k^x) p(\mathbf{y}_g|\Delta_j^x)$$

Expectation-maximization (EM) algorithm (Dempster et al., 1977b) was employed to infer the unknown parameter values for this distribution. These results showed a very complex relationship between mRNA and protein data. Gene ontology analysis shows a high correlation between mRNA-protein data is limited to a few molecular properties such as cell adhesion complexes, ribosomes, protein folding and TCP-1 chaperonin. However, they were able to come up with three main conclusions. Firstly, the correlation with mRNA and protein data was generally low. Secondly, their results showed that the correlation is very limited for mammalian data. Finally, they claimed that mRNA and protein data evolve independently unless there are strong selection factors present in favour of gene transcription and protein translation.

2.4.2.3 Bayesian Method

Probabilistic modelling using Bayesian networks is another significant data-driven approach to model mRNA and protein data. This method can be used to infer useful biological properties while understanding the relationship between mRNA and protein data.

Kannan et al., 2007 used microarray mRNA measurements and mass spectrometry protein measurements of laboratory mouse (*Mus musculus*) to model the relationship between transcriptome and proteome measurements. Here the authors mentioned three main problems in the use of other techniques to investigate the relationship between

the above measurements. Firstly, searching for correlations will only give a global summary while gene-by gene correlation provides more information (Mootha et al., 2003). Secondly, most systems consider that the noise of the data is distributed as Gaussian. However, in their approach they used hidden variables to deal with the non-Gaussian noise. Thirdly, most models use multiple sources (Gygi et al., 1999; Greenbaum et al., 2003) to obtain data while in this method they extracted mRNA and protein data from the same source. Kannan et al., 2007 overcome these problems by using the following methods. They used a gene-by-gene based analysis by introducing a probabilistic model which uses a Bernoulli switch variable (s). They also introduced a hidden variable (τ) with a Poisson distribution over the observed peptide counts to represent the noise of the data. The total data set contained mRNA and protein data for six main organs, namely; brain, heart, kidney, liver, lung and placenta. These were measured under the same conditions. A Bayesian network for mRNA and protein abundance is represented in Figure 2.13.

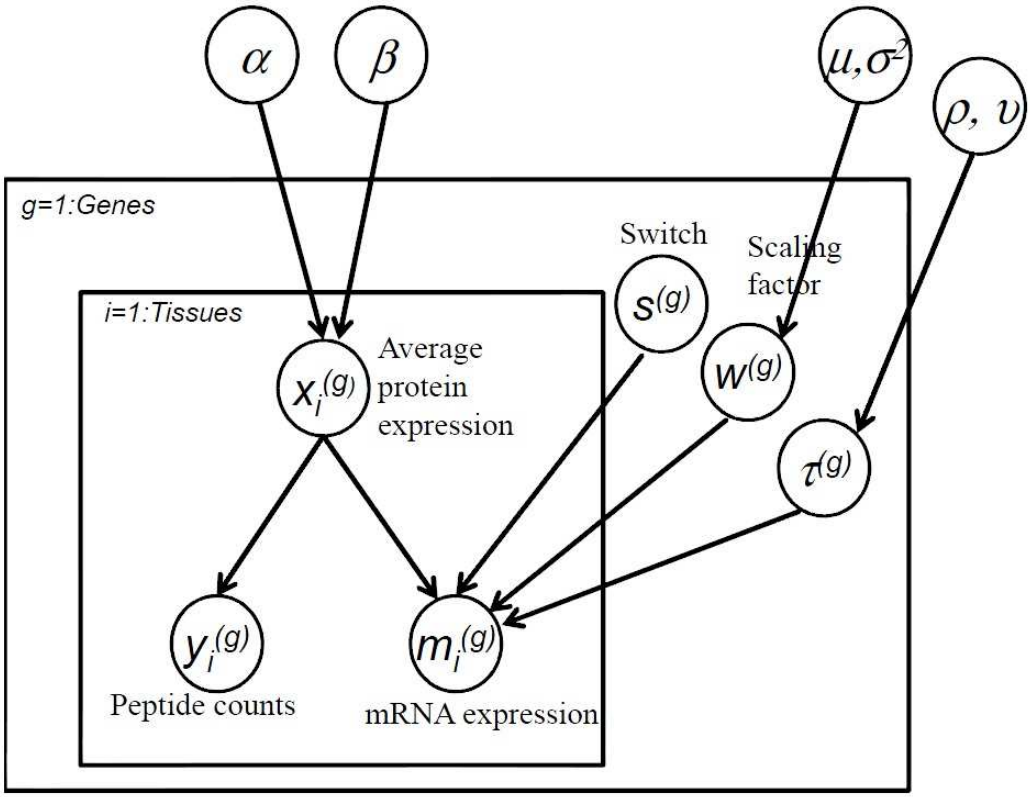


Figure 2.13: This is the Bayesian network taken from Kannan et al., 2007 which represents the relationship between peptide counts measuring protein expression and microarray mRNA expression levels. Inner rectangle represents a single gene g and all T tissues of gene g shares the same s,w and τ variables.

Let **m** and **y** are mRNA and protein abundances respectively. Average peptide count given the peptide count is modelled using Poisson distribution and the rate parameter **x** is modelled using a Gamma distribution. Prior distribution of switch variable is shown

as $P(s)$. If the switch variable is $s = 1$ then the noise is modelled as a linear function of average peptide counts, given by $m_i = wx_i + noise$ and the noise assumed to be Gaussian with mean = 0 and variance τ. Even though the added noise to the model is assumed as Gaussian, other random variables have different distributions. Therefore, the resulting model is far from Gaussian distribution. The joint distribution over all the variables would be:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{m}, \theta, s) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})p(\mathbf{m}|\mathbf{x}, s, \theta)p(\theta)P(s)$$

With $\theta = w, \tau$.

The objective of this model is to obtain the relationship between mRNA data and protein peptide count data. Therefore, in this study they chose to model $p(\mathbf{m}|\mathbf{y})$ because learning and inference is more straight forward. Parameters were learned by maximizing the probability $p(\mathbf{m}|\mathbf{y})$. In order to check the linear relationship between two measurements, $P(s = 1|\mathbf{m}, \mathbf{y})$ was calculated for each gene.

$$p(\{\mathbf{m}^{(g)}|\mathbf{y}^{(g)}\}) \approx \int_\theta \prod_{g=1}^{G} \int_{\mathbf{x}} \sum_s P(s^{(g)})p(\mathbf{x}^{(g)}|\mathbf{y}^{(g)})p(\theta)p(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)}, \theta)$$

Sample average was employed to approximate true expectation. The strength of the relationship between mRNA and protein pair was given by the $P(s|\mathbf{m}, \mathbf{y})$ probability.

$$P(s|\mathbf{m}, \mathbf{y}) = \frac{\int_{\mathbf{x}} p(\mathbf{m}|s, \mathbf{x})p(\mathbf{x}|\mathbf{y})P(s)}{\sum_s \int \mathbf{x}p(\mathbf{m}|s, \mathbf{x})p(\mathbf{x}|\mathbf{y})P(s)}$$

Kannan et al., 2007 performed a permutation test and re-learned the model to obtain the scores for the linear relationships. The data was then partitioned into three groups of biological interest: inliers, borderline and outliers. Genes that have a score in the range of $P(s = 1|\mathbf{m}, \mathbf{y}) \leq 0.33$ and a p-value within 0.05 were considered as outliers. 503 genes were detected as outliers and several of these were blood-borne factors in liver, lung and placenta. Outlier genes were enriched with GO annotations such as embryogenesis and transport. Inliers genes were obtained by considering the score in the range of $P(s = 1|\mathbf{m}, \mathbf{y}) \geq 0.66$ and a p-value within 0.05. These inliers were significantly enriched with cell adhesion and central nervous system GO annotations. The rest of the genes were taken under borderline and these genes were enriched with the functional annotations such as mitochondrion and skeletal development anomalies. Kannan et al., 2007 also compared their results with maximum likelihood (ML) version of the proposed method and standard linear regression (LR) and found Bayesian model achieves mappings with higher statistical significance compared to the other two.

2.4.2.4 Protein Abundance Predictor

Tuller et al., 2007's research was a significant remark in bridging the gap between transcriptome and proteome measurements. They developed a machine learning based predictor of protein concentrations, which takes a different approach to most of the previous research. In order to obtain the best features for the predictor, Greedy feature selection technique was used with 32 transcriptome and proteome properties of *Saccharomyces cerevisiae* (Table 2.3). In fact, mRNA, tRNA adaptation index (tAI) and evolutionary rate (ER) were selected as the most significant feature to develop the predictor. Feature selection algorithm byTuller et al., 2007 is shown below:

Algorithm 1 Forward Greedy Feature Selection

1: Initialization Feature set $F^k = \emptyset$ at $k = 0$ and with n no of selections
2: **while** $k \neq n$ **do**
3: find the best feature j to add to F^k with most significant cost reduction
4: $k++$ and $F^k = F^{k-1} \cup \{j\}$
5: **end while**

A linear predictor as developed using the above three properties (see Appendix A for linear predictor). Two main data sets were used to find the correlation between predicted and measured protein abundances. For both data sets, prediction accuracy of the predictor increased by adding one features at a time (Figure 2.14). By this process Tuller et al., 2007 achieved a correlation of 0.76 for averaged data and of 0.63 for data obtained by separate data sources. This is relatively higher than that found in previous work (Gygi et al., 1999; Futcher et al., 1999). They also used a SVM non-linear predictor and observed that there is no improvement in the prediction. GO enrichment analysis was carried out to investigate on the biological properties of the predicted proteins. The results indicate that their predictor is more appropriate for proteins in large macro-molecule are complexes. Further, they also used *Schizosaccharomyces pombe* mRNA and protein expression for the predictor. Corresponding orthologs in *Saccharomyces cerevisiae* were used obtain ER and tAI values. However, *S.pombe* obtained a correlation of 0.675 for predicted a measured protein abundance, where the correlation between mRNA and protein data was only 0.629. Comparing protein concentration in both rich and poor media also represented a general trend for homeostatic regulation.

However, Tuller et al., 2007 only forcused on developing a predictor, whereas in our research we wanted to go further and investigate the model failures with large errors between actual measurements and predictions to identify post-translationally regulated proteins.

Table 2.3: Abbreviation and full description of all the features used in Tuller et al., 2007's Study

Index	Abbreviation	Full Description
1	MW	Molecular weight
2	PI	Net charge of protein in aqueous solution
3	CAI	Codon Adaptation Index
4	PL	Protein length
5	CB	Codon bias
6	ALA	Frequency of the amino acid Alanine in the protein
7	ARG	Frequency of the amino acid Arginine in the protein
8	ASN	Frequency of the amino acid Asparagine in the protein
9	ASP	Frequency of the amino acid Aspartic acid in the protein
10	CYS	Frequency of the amino acid Cysteine in the protein
11	GLN	Frequency of the amino acid Glutamine in the protein
12	GLU	Frequency of the amino acid Glutamic acid in the protein
13	GLY	Frequency of the amino acid Glycine in the protein
14	HIS	Frequency of the amino acid Histidine in the protein
15	ILE	Frequency of the amino acid Isoleucine in the protein
16	LEU	Frequency of the amino acid Leucine in the protein
17	LYS	Frequency of the amino acid Lysine in the protein
18	MET	Frequency of the amino acid Methionine in the protein
19	PHE	Frequency of the amino acid Phenylalanine in the protein
20	PRO	Frequency of the amino acid Proline in the protein
21	SER	Frequency of the amino acid Serine in the protein
22	THR	Frequency of the amino acid Threonine in the protein
23	TRP	Frequency of the amino acid Tryptophan in the protein
24	TYR	Frequency of the amino acid Tyrosine in the protein
25	VAL	Frequency of the amino acid Valine in the protein
26	FOP	Frequency of optimal codons
27	GRAV	Gravy, hydropathicity of Protein
28	AROM	Aromaticity (Frequency of aromatic amino acids: Phe, Tyr, Trp)
29	HL	Protein Half life
30	ER	Evolutionary rate
31	TE	Translation efficiency
32	tAI	tRNA adaptation index

2.5 Summary

In summary, extensive high-throughput *omic* data are now available and there is a huge demand for computational methods to cut down the data space for laboratory experiments. Proteomic data getting complex compared to transcriptome data due to post-translational modifications. These post-translational modifications are very important in studies of different diseases such as cancer. Moreover, PTMs can be used to determine different biological processes. Further, some post-translation modifications disrupt protein stability and causing them

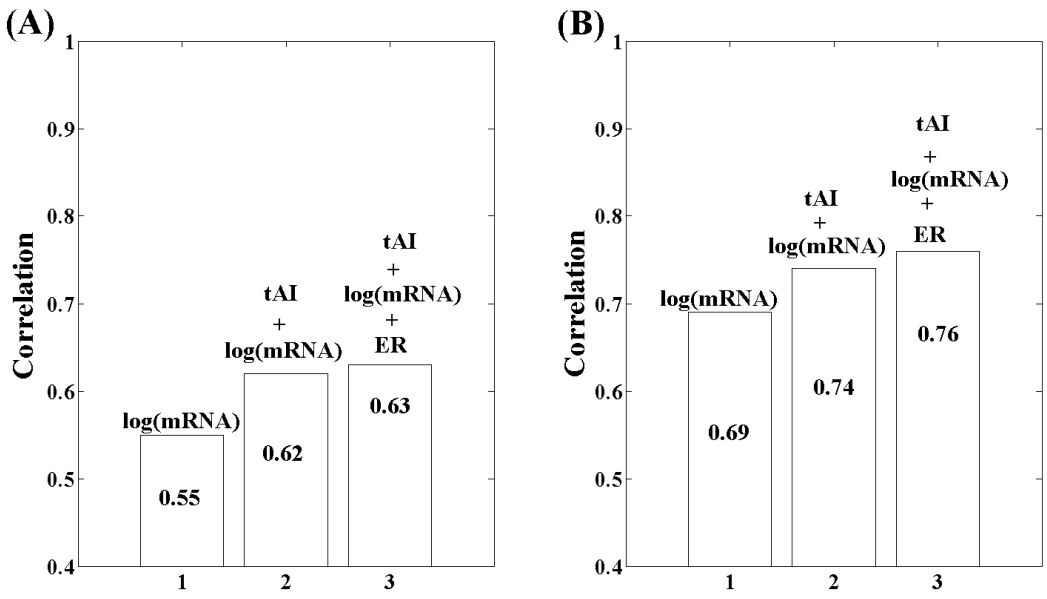


Figure 2.14: Accuracy variation of linear predictor by Tuller et al., 2007. (A) Test set was generated using separate data sources for all the features. (B) Averaged at least two data sources to generate the test data.

to degrade faster. Therefore protein concentrations will be lower due to these post-translational modifications. We also observed that most of the experiments are designed to simply look for the correlation between transcriptome and proteome measurements. Several authors have employed data-drive approaches such as clustering, classification, Bayesian and regression to extract useful biological information from the above measurements. However, Tuller et al., 2007's regression approach is different from others where they predict the protein abundance using mRNA and other translation related features. Taking inspiration from Tuller et al., 2007's study, we develop a data-driven framework which can detect post-translationally regulated proteins by looking at the model failures or outliers of a protein abundance predictor.

Chapter 3

Modelling Transcriptome-Proteome Measurements & Identifying Post-translationally Regulated Proteins

In this chapter, we develop a regression model at the transcriptome and proteome interface and detect post-translationally regulated proteins by looking at the model failures (outliers). Firstly, we consider mRNA and other transcriptomic properties as proxies to predict protein abundance and employ L_1 norm sparse solution to select the best set of features to develop the global regression predictor. Secondly, we look into the outliers of the regression model assuming that post-translational regulation primarily act by disrupting protein stability where the measured abundance is lower than predicted. Finally we carry out a functional annotation check to prove our hypothesis using statistical evidence. This also motivates the effective outlier detection at the transcriptome-proteome interface in later chapters.

3.1 Data Preparation

In order to use a machine learning approach to model at the transcriptome proteome interface, a rich set of input and output data samples are important. Yeast bacteria, which is also known as *Saccharomyces cerevisiae* under exponentially

growth conditions was selected as the main organism to obtain data for our experiment because this is the most well studied genome in transcriptomic and proteomic world. We combined 37 input variables at the transcriptome layer using ORF and gene names to develop the protein abundance predictor (Figure 3.2(A)). mRNA and protein abundances were downloaded from Greenbaum et al., 2003 and PaxDb (Wang et al., 2012a) respectively. Translation rate related features such as proteins per second, ribosomal occupancy, ribosome density, relative translation rate and gene length were obtained from Greenbaum et al., 2003. Further, mRNA half-lives Miller et al. (2011), tRNA adaptation index (tAI) (Man and Pilpel, 2007) and evolutionary rate (ER) (Wall et al., 2005) data sets were also employed as input variables. Table 3.1 shows 28 sequence derived transcriptome properties which were obtained from SGD database (Cherry et al., 2012). Here, we employed gene length, proteins per second, relative translation rate, mRNA half-lives, ribosomal density and occupancy as new features, with respect to the previous work by Tuller et al., 2007. We obtained 2000 samples with all the 37 input features.

We explored pairwise scatter plots to verify the range over which data was available and filtered out some of the data. It is important to limit the range of the data distribution for robust modelling. The length property of transcripts had the range between 157 to 14733 base pair(bp), where only 50 genes had length greater than 5000 bp. Hence, the genes which are longer than 5000 bp were removed to reduce the skewness of the data distribution (Figure 3.1(A)). We also observed that, at low levels of mRNA expression, the same mRNA values lead to very different protein expression levels. Such low expressions of mRNAs have not been measured reliably using hybridization microarrays, as they are the result of molecules that are available in very low copy numbers per cell and are pooled and amplified before hybridization. Therefore, 323 genes were filtered from the data set at the value of -1.0 from the natural log mRNA expression level (Figure 3.1(B)) in order to remove the measuring errors from the data set.

3.2 Feature Selection Using Sparse Regression (LASSO)

Feature selection is considered as the key step towards solving any practical classification problem. The two main kinds of feature selection methods are, those based on probabilistic separable measures applied to the data and those based on the error rate of a classifier as a design criterion (Lovell et al., 1998). It is beneficial to reduce feature space to avoid *curse of dimensionality* problem, where the classification accuracy increases with the number of features up to a certain

Table 3.1: Twenty-Eight Sequence Properites

Index	Abbreviation	Full Description
1	MW	Molecular weight
2	PI	Net charge of protein in aqueous solution
3	CAI	Codon Adaptation Index
4	PL	Protein length
5	CB	Codon bias
6	ALA	Frequency of the amino acid Alanine in the protein
7	ARG	Frequency of the amino acid Arginine in the protein
8	ASN	Frequency of the amino acid Asparagine in the protein
9	ASP	Frequency of the amino acid Aspartic acid in the protein
10	CYS	Frequency of the amino acid Cysteine in the protein
11	GLN	Frequency of the amino acid Glutamine in the protein
12	GLU	Frequency of the amino acid Glutamic acid in the protein
13	GLY	Frequency of the amino acid Glycine in the protein
14	HIS	Frequency of the amino acid Histidine in the protein
15	ILE	Frequency of the amino acid Isoleucine in the protein
16	LEU	Frequency of the amino acid Leucine in the protein
17	LYS	Frequency of the amino acid Lysine in the protein
18	MET	Frequency of the amino acid Methionine in the protein
19	PHE	Frequency of the amino acid Phenylalanine in the protein
20	PRO	Frequency of the amino acid Proline in the protein
21	SER	Frequency of the amino acid Serine in the protein
22	THR	Frequency of the amino acid Threonine in the protein
23	TRP	Frequency of the amino acid Tryptophan in the protein
24	TYR	Frequency of the amino acid Tyrosine in the protein
25	VAL	Frequency of the amino acid Valine in the protein
26	FOP	Frequency of optimal codons
27	GRAV	Gravy, hydropathicity of Protein
28	AROM	Aromaticity (Frequency of aromatic amino acids: Phe, Tyr, Trp)

stage but if you increase the feature space further the accuracy starts to degrade. Thus, selecting the best subset of features from a large feature space is important to explain useful aspects of the problem domain. Greedy search with series of linear predictors was used in previous work (Tuller et al., 2007), introducing mRNA abundance, tAI and evolutionary rate as the main features. However, this method is particularly weak with correlated features and also difficult to prove the correctness of the feature set (Hall, 1999). Therefore, sparse regression with L_1 norm regularization (Equation (3.1)), also known as LASSO (Tibshirani, 1994) was selected as an alternative strategy, which is interestingly popular in machine learning literature to select most dominant features (Lu et al., 2011; Wu et al., 2009; Park and Casella, 2008). Figure 3.2 illustrates the feature selection process using L_1 norm regularization technique. LASSO can be written as;

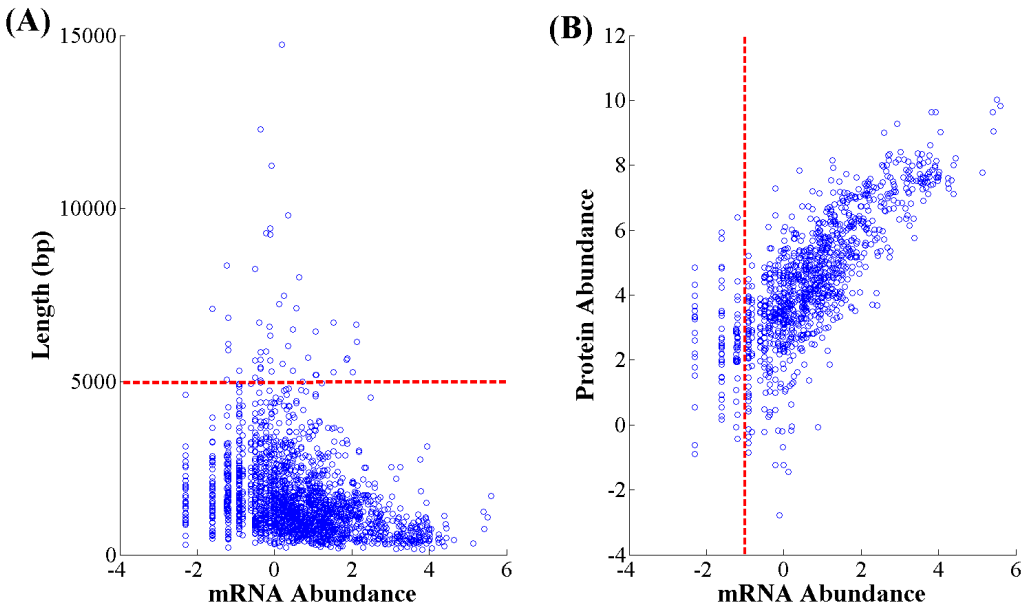


Figure 3.1: Data Filtering: Some of the data was filtered by studying the distribution of mRNA/protein species. **(A)** genes with lengths longer than $5000\,kb$ and **(B)** those with log mRNA expressions lower than -1.0 were eliminated from analysis.

$$\min_{\boldsymbol{w},b} \quad \{y - (\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)\}^2 + \lambda \|\boldsymbol{w}\|_1 \tag{3.1}$$

where $\boldsymbol{x}$ represents the input matrixof covariates, y represents the response vector and $\boldsymbol{w}$ is the unknown weight vector which minimizes the loss function. λ determines the amount of regularization happens to the loss function and this value need to be decided by sample validation. If the λ is too small all features will end up with high weights and if the λ value is too large no features will be detected as significant. This method is used to identify most suitable features from over-determined systems, where in this case we need to find best features out of 37 variables without over-fitting the data (Figure 3.2(A)). LASSO generates weights for all the features which minimizes the loss function and produces the maximum sparse regression. Features with the highest weights (absolute weight values) are considered as the most dominant features (Figure 3.2(B)). We used the `CVX` package within a `MATLAB` environment to obtain solutions for sparse regression.

Regularization

Fifty samples, which contains 500 genes per sample were selected using with and without replacement bootstrap sampling method. Each of these 50 samples were tested using 20 values for λ parameter between 0 to 1000, *i.e.* creating a total of 1000 validation trials (50 samples x 20 values for λ). In this feature selection process we had two main problems to address on; one is to select the best value for parameter λ from the above mentioned range and then to select the best set of features from the 1000 sets of highly weighted features.

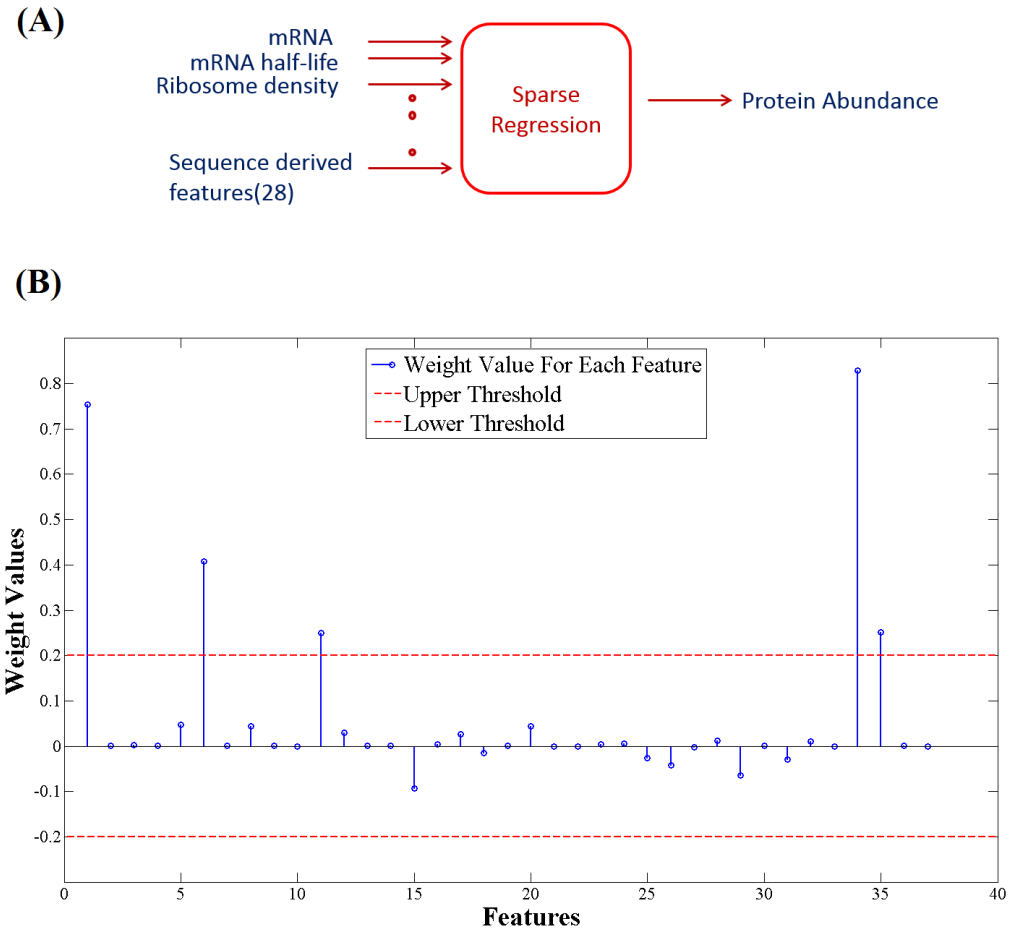


Figure 3.2: L_1 norm regularization: **(A)** 37 transcriptomic input properties were used as proxies for protein abundance. **(B)** Best set of features were obtained by selecting the non-zeros weights after thresholding. Weights between red dashed lines (thresholding points) are considered as zero.

Sparse regression with L_1 norm regularization was carried out for all the samples, producing a weight vector for each sample. By plotting the weights with respect to their features as shown in Figure 3.2(B), we observed that majority of the features resulted very low weight values and few of the features had very high

weight values. Therefore, we set a threshold to the weight vectors on -0.2 and 0.2 and removed the features which had very low weights between the threshold values. Thus, highly weighted transcriptomic properties (features) were selected as most dominant features for each test trial.

Setting the λ parameter with an appropriate value is crucial in sparse regression. Therefore, we looked into the average number of retained features as a function of λ which is shown Figure 3.3(A) to understand the variation of the feature extraction with the λ parameter. This figure shows that the number of dominant features selected using different λ values do not reduce linearly or monotonically. In fact there is a stable region over three order of magnitude of λ between 0.001 and 1. We note that set of five features get selected during the stable region, suggesting that there is an important aspect of the data set with five dominant features. In order to confirm our observation, we constructed a random data set (using same mean and variance values to the original data set) and looked into the variation of average number of features selection with respect to λ. Red dashed line in Figure 3.3(A) shows that the number of dominant features selected by randomly generated data reduces monotonically along the λ value distribution (no stable region).

Figure 3.3(B) shows the frequencies of the six sets of features (containing 6, 5, 5, 6, 4 and 4 features in each set) which are repeatedly identified more than 5 times with 500 bootstrap samples along the stable region. From these six, **set 3** turned out to be significant with highest frequency and it contained following five features; mRNA abundance, tRNA adaptation index (tAI), codon bias, ribosome density and occupancy. Thus, the aspect of selecting five features along the stable region is clarified by these results. However, our best features are slightly different from Tuller et al., 2007's three features; mRNA, tAI and evolutionary rate (ER). Sparse regression identifies ribosome density, ribosomal occupancy and codon bias as relevant features, whereas Tuller et al., 2007 did not use translation efficiency related properties, which are important to predict protein abundance (Greenbaum et al., 2003). Ribosome occupancy is the fraction of mRNA molecules attached to at least one ribosome and ribosome density is the number of ribosomes active with mRNA with a unit transcript length (Brockmann et al., 2007; Arava et al., 2003). Frequency of occurring synonymous codons in coding sequence is known as codon bias, which is an important factor for protein synthesis to produce efficient amino acid sequences (Brockmann et al., 2007; Tuller et al., 2010). tAI predicts the level of adaptation of amino acids to the coding sequence relative to the cells tRNA pool (Man et al., 2006). This gives the ratio between gene copy number (GCN)

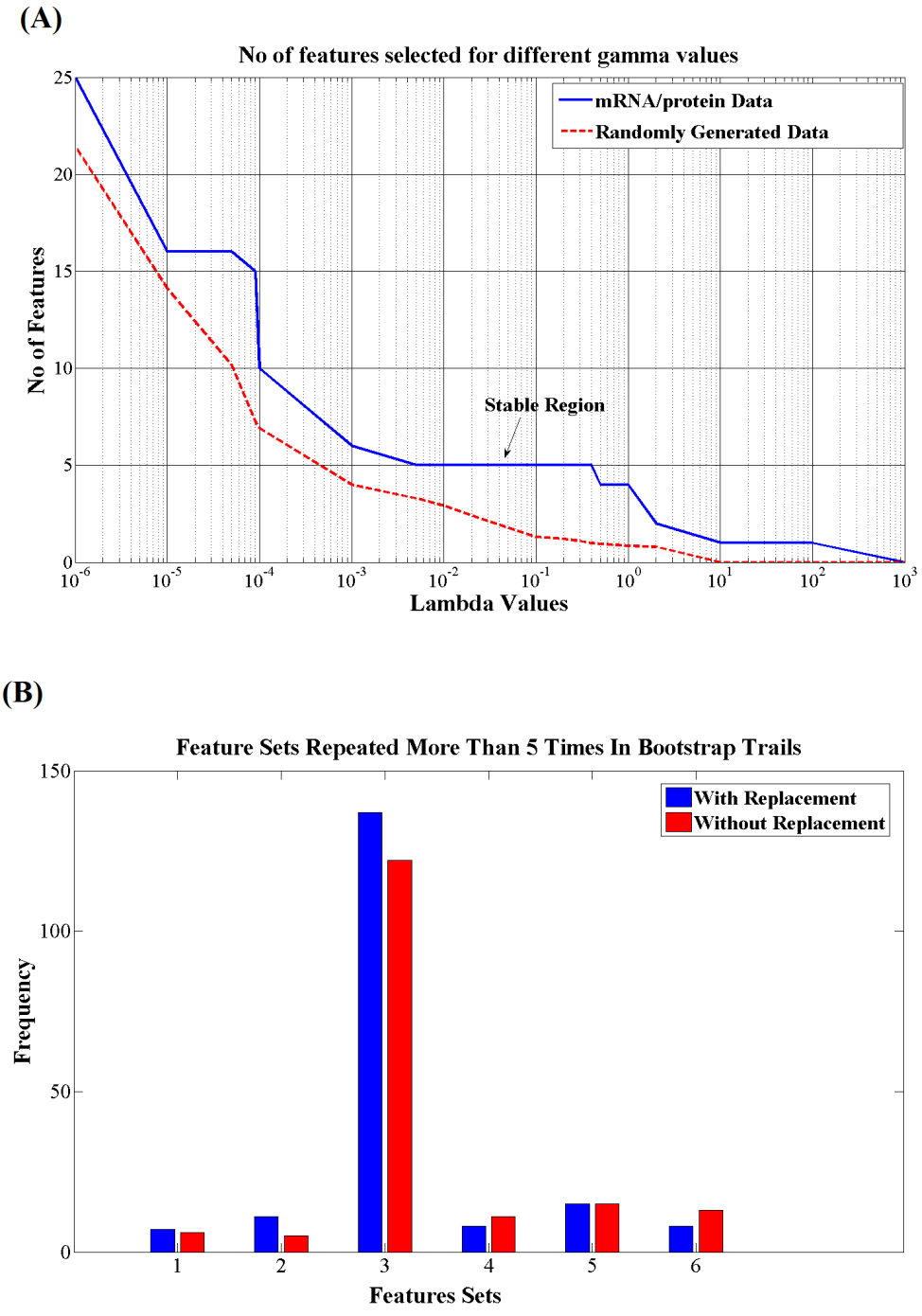


Figure 3.3: Feature Selection: **(A)** Average number of selected features as a function of λ regularization term; which have a stable region over 3 orders of magnitude of λ (0.001 and 1). **(B)** Identifying the best set of features (set three) using the most frequent features sets repeated more than 5 times in bootstrap trials over the stable region of λ. Set three contained mRNA abundance, codon bias, tAI, ribosome density and occupancy.

of the corresponding tRNA with codon k and the maximum GCN for that amino acid (Brockmann et al., 2007):

$$\mathbf{tAI} = \frac{GCN_k}{max\{GCN_k\}}$$

tAI directly associate with translation efficiency in human proteome (Waldman et al., 2010). Thus, features selected by sparse regression are directly associated with protein generation process.

3.3 Development of Protein Abundance Predictor

With the five best features selected by LASSO technique were employed with a linear regression model (Equation (3.2)) to predict the protein abundance more accurately. Linear regression is considered as the starting approach in data-driven modelling scenarios. Suppose we have a set of m samples $\{(\boldsymbol{x_i}, y_i)\}_{i=1,\ldots,m}$ where $\boldsymbol{x_i} \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ are the inputs and targets respectively. In this regression model, our main objective is to predict y as $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ with the minimum squared loss. Leave-one-out cross validation was employed to obtain the average prediction accuracy.

$$\min_{\boldsymbol{w},b} \quad \{y - (\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)\}^2 \tag{3.2}$$

Our best five features (mRNA, occupancy, ribosome density, tAI and codon bias) gave a regression of $R^2 = 0.86$ between the predictions and the targets, while previous work (Tuller et al., 2007) features (mRNA, tAI and evolutionary rate) gave only $R^2 = 0.80$. Combining all 37 features gave $R^2 = 0.80$, which is lower than our best five feature accuracy. Therefore, with various tests on unseen (leave-one-out cross validated) data, the five features selected by sparse regression gave the highest protein abundance prediction accuracy.

Afterwards, neural network machine learning technique was employed to observe non-linear prediction accuracy. Neural net curve fitting application in `MATLAB` was used to develop the predictor. Stochastic gradient descent method was used to optimize the parameters and 10 neurons were used as the hidden layers (Bishop, 1995). 50% of the data set was employed as training data and the rest were partitioned in to two groups for validation and testing purposes, which contained 25% from the total data set in each partition, *i.e.* validation data was used to

measure the network generalization, and to halt training when generalization stops improving. Ensuring that all genes in the data set were subjected to a test group, this process was carried out four times while swapping the test data with training and validation data.

Figure 3.4 shows the regression comparison between linear and non-linear predictors. Neural net non-linear predictor gave $R^2 = 0.82$ regression with our five features and $R^2 = 0.79$ for the Tuller et al., 2007's three features. However, the prediction accuracy dropped drastically to $R^2 = 0.69$, by including all 37 features. Tuller et al., 2007 also had a similar observation where the non-linear kernel SVM model did not improve the prediction accuracy. However, with both predictors our five features outperformed in predicting protein abundance more accurately.

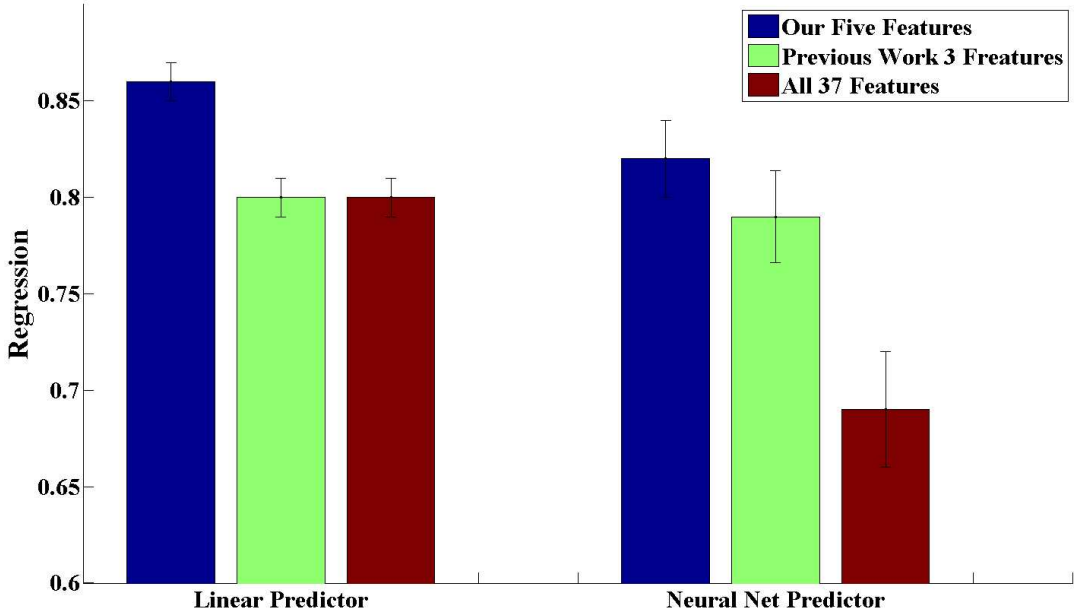


Figure 3.4: Regression comparison between linear and non-linear (neural net) predictors using unseen (cross-validated) data. Our five features gave high accuracies with both predictions. However, there is no advantage of using a non-linear predictor in this task.

Additionally, ER was not selected as an important feature by our sparse regression feature selection. Therefore, we examined the correlation from our predictor by adding our five features progressively and finally including ER as a sixth feature to observe the effects of ER. Figure 3.5 shows that, adding our five features improved the prediction accuracy monotonically (this is true for any order of feature adding), but including ER as the sixth feature reduced the accuracy to $R^2 = 0.80$. Further, ER and protein concentration only gave a correlation of $R^2 = -0.46$. Thus, ER is not a good feature for this task and Moreira et al., 2002 mentioned that ER is just an empirical observation as a feature for the protein abundance predictor.

However, in this research we are more interested in identifying post-translationally regulated proteins by looking at the model failures or outliers of a protein predictor, rather than improving the prediction accuracy.

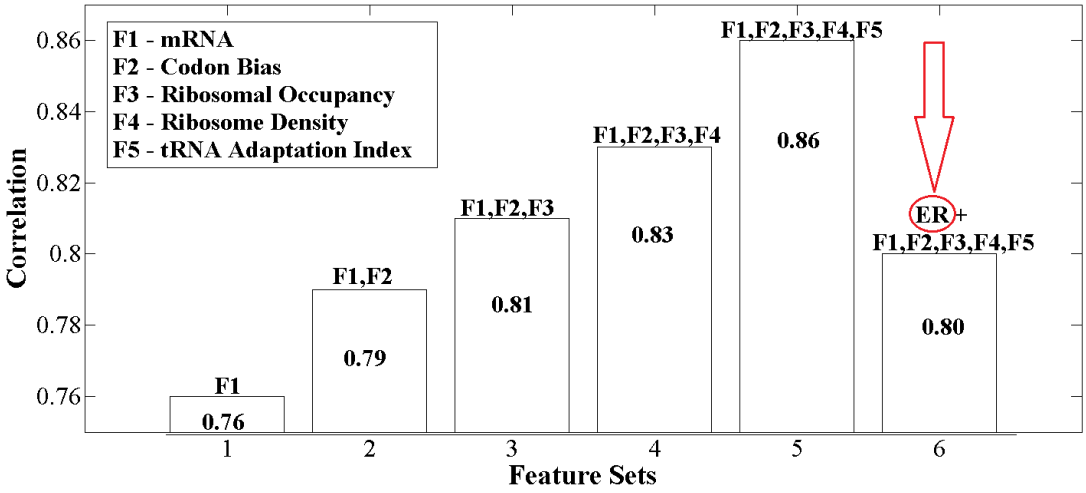


Figure 3.5: Adding our five features (mRNA abundance, tAI, codon bias, ribosome occupancy and density) improved accuracy monotonically in each step. However, adding ER as the sixth features to the linear predictor reduced the overall accuracy.

3.4 Identifying Proteins with PTR as Outliers

As we explained in the Introduction, post-translational regulation (PTR) disrupts protein stability, therefore proteins with PTR degrade faster than we actually predict from a global regression. Based on the above fact, we developed our main hypothesis as follows; model failures or outliers of a protein abundance predictor those having large errors between the actual measurement (P)and the predicted protein levels ($\hat{P}$) (*i.e.* measured abundance lower than the predicted - $P < \hat{P}$) are more likely to be post-translationally regulated.

Here we use a simple technique to detect outliers from the regression model. We produced a regression plot between predicted protein abundance ($\hat{P}$) versus measured protein abundance (P) and obtained the 50 proteins that are lying further away from the regression line as outliers (2.5% outlier cut-off region) as shown in Figure 3.6. Here we use 50 proteins as the benchmark number of outlier proteins to test PTR as it is a small number of samples to be tested in a biological experimental setting. Afterwards, we carried out functional annotation checks at two levels (coarse and finer levels) to confirm that these outlier proteins are enriched with post-translationally regulated proteins. (1) At the coarse level, we used `UniProt`

database (Magrane and Consortium, 2011) which is cross referred by the PaxDb (Wang et al., 2012a) where we obtained the protein abundance data and checked for post-translational modification (PTM) keywords as the primary requirement to detect PTR from the outlier proteins. (2) At the finer level, we coupled PTM keywords with protein stability determinant motif information as a stronger indicator of post-translational regulation (*i.e.* Phosphorylation with PEST motifs, Acetylation with N-termini segments and Ubiquitination with D/KEN Box motifs). `Epestfind` database (Rice et al., 2000) and `NetAcet 1.0` database (Kiemer et al., 2005) were employed to detect PEST motifs and N-termini segments respectively. D and KEN box motifs related to ubiquitination were detected using `GPS-ARM 1.0` toolkit (Liu et al., 2012). In order to obtain statistical significance of the outliers proteins with respect to post-translation regulation, 1000 random samples (with sample size 50) were employed as a computationally exhausting sampling process.

3.4.1 Results and Discussion

Figure 3.6 shows the regression plot of the measured (P) and predicted protein concentrations ($\hat{P}$). Outliers in this figure are points that are lying further away from the regression line (shown as solid line). Fifty proteins were selected as the outliers from the 2.5% cut-off boundary and 48 of them were detected in upper outlier section with $P < \hat{P}$ (negative losses), where our interest focus lies.

3.4.1.1 Level 1 : Coarse Level PTM Analysis

Forty-two from 48 outliers with $P < \hat{P}$ had PTM keywords and they are showed in Table 3.2. No proteins were detected with PTM keywords in the lower region ($P > \hat{P}$) outliers, thus this region does not contain any significant post-translational regulation.

The detected outlier set had six main PTM key words and their functional descriptions according to `UniProt` database Magrane and Consortium (2011) are given below:

- **Phosphoprotein**, the most frequently noted annotation, a process that attaches either a single phosphate group, or a complex molecule, such as 5'phospho-DNA, through a phosphate group;
- **Glycoprotein** (also known as Glycosylated proteins) containing one or more covalently linked carbohydrates of various types;

Table 3.2: Coarse level check - PTM keywords identified with 50 outliers (cut-off at 2.5%) using `UniProt` database Magrane and Consortium (2011)

ORF Name	Gene Name	PTMs
YJL129C	TRK1	Phosphoprotein, Glycoprotein
YBR038W	CHS2	Phosphoprotein, Glycoprotein
YDL093W	PMT5	Glycoprotein
YDL217C	TIM22	x
YFL029C	CAK1	Phosphoprotein
YHR031C	RRM3	Phosphoprotein
YJR124C	YJR124C	Phosphoprotein
YDL048C	STP4	Phosphoprotein
YGL159W	YGL159W	x
YDR006C	SOK1	Phosphoprotein
YIL169C	YIL169C	Glycoprotein
YDL222C	FMP45	Phosphoprotein, Glycoprotein
YDL130W	RPP1B	Phosphoprotein, Acetylation
YCR010C	ADY2	Phosphoprotein
YHR141C	RPL42B	Methylation
YBR106W	PHO88	Phosphoprotein
YAR075W	YAR075W	Phosphoprotein
YHR094C	HXT1	Phosphoprotein, Glycoprotein
YDR342C	HXT7	Isopeptide b., Phosphoprotein, Ubl con.
YBR1317	RPS9B	Phosphoprotein
YJL177W	RPL17B	Phosphoprotein
YGR282C	BGL2	Glycoprotein
YBL0613	RPS8A	Phosphoprotein
YDR225W	HTA1	Isopeptide b., Phosphoprotein, Acetylation, Ubl conj.
YEL027W	VMA3	x
YKR059W	TIF1	Phosphoprotein, Acetylation
YGL030W	YGL030W	Phosphoprotein
YIL148W	RPL40A	Isopeptide b., Ubl con., Phosphoprotein
YBR010W	HHT1	Methylation, Phosphoprotein, Acetylation
YHR021C	RPS27B	Phosphoprotein
YGR034W	RPL26B	Phosphoprotein
YER102W	RPS8B	Phosphoprotein
YDL083C	RPS16B	Acetylation, Phosphoprotein
YDR064W	RPS13	Phosphoprotein
YCR031C	RPS14A	Acetylation, Phosphoprotein
YDL081C	RPP1A	Acetylation, Phosphoprotein
YEL034W	HYP2	Acetylation, Phosphoprotein
YDR447C	RPS17B	Phosphoprotein
YER117W	RPL23B	Methylation, Acetylation, Phosphoprotein
YKL180W	RPL17A	Phosphoprotein
YKL056C	TMA19	x
YKL152C	GPM1	Phosphoprotein
YLR044C	PDC1	Acetylation, Phosphoprotein
YCR012W	PGK1	Acetylation, Phosphoprotein
YGL123W	RPS2	Acetylation, Phosphoprotein
YDR382W	RPP2B	Phosphoprotein
YGR148C	RPL24B	Phosphoprotein
YDL014W	NOP1	Phosphoprotein, Methylation
YDL080C	THI3	x (lower outlier region)
YER070W	RNR1	x (lower outlier region)

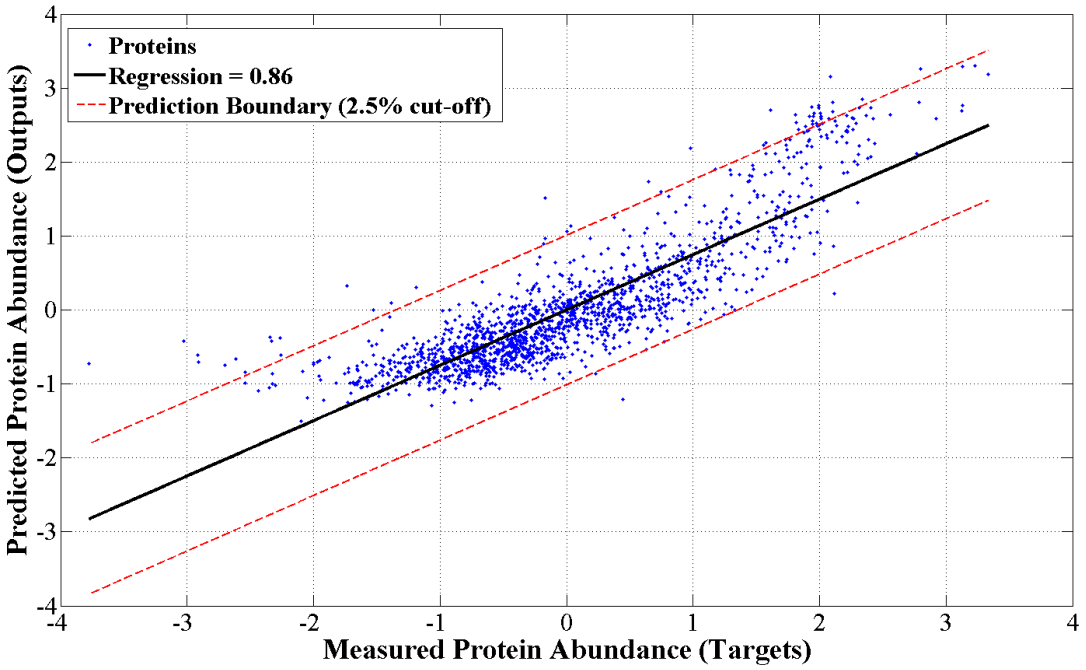


Figure 3.6: Detecting outliers using protein abundance predictor. Black solid line shows the linear regression of $R^2 = 0.86$ between the true and the predicted concentrations. Red dashed lines represent the 2.5% cut-off boundaries of the data set. Fifty proteins which are lying further away from the regression (solid) line were selected as outliers (beyond cut-off boundaries).

- **Acetylation**, the modification by the attachment of at least one acetyl group (normally at the N-terminus);
- **Methylation**, is post-transalationally modified by the attachment of at least one methyl group.
- **Isopeptide bonds**, involving the side chain of one or two amino acid residues. These facilitate catalysis by enzymes leading to the formation of dimers and other complexes.
- **Ubiquitin conjugation**, in which proteins attach with at least one ubiquitin-like modifier such as SUMO, APG12, URM1 or RUB1.

Above PTMs are largely associated with protein stability. Hood et al., 1977 have shown the interest in determinant of protein turnover and degradation. As an example, Hofmann et al., 2001 suggest that the stability of the protein is affected by the phosphorylation and acetylation by using their study of the $p53$ regulation activity. Review of post-translational modifications by Nalivaeva and Turner, 2001 imply that PTMs such as N-linked (Acetylation) and glycosylphosphatidylinositol (GPI) in protein stability. They also mentioned that members of the ubiquitin

family can be implicated in protein turnover by post-translational modifications. Amino acids substitutions due to mutations can act as a marker for protein degradation (Stadtman, 1990). Therefore localized post-translational modifications, such as methylation, can be equivalent to site-specific amino acid substitutions, affecting stability and the degradation rates of proteins. Thus, we can explain the over-representation of PTM annotated proteins in upper region outliers as an effect of fast protein degradation by the post-translational regulation, which reduce the actual measurement compared to the global regression prediction.

Statistical confidence level of outlier sample was measured using normal cumulative distribution function, using 1000 random samples of size 50 with mean and standard deviation of 34.286 and 3.576 respectively. Figure 3.7 shows the distribution of random samples. Using these random samples, the claim of over-representation of post-translational regulations among the selected outlier set was made at significance of $p \leq 0.02$. However, biological research uses 0.05 as the level of significance to accept the hypothesis (McDonald, 2009). Therefore, our initial hypothesis is true with $p \leq 0.05$, which explains that most of the proteins selected as outliers undergo with post-translational regulation.

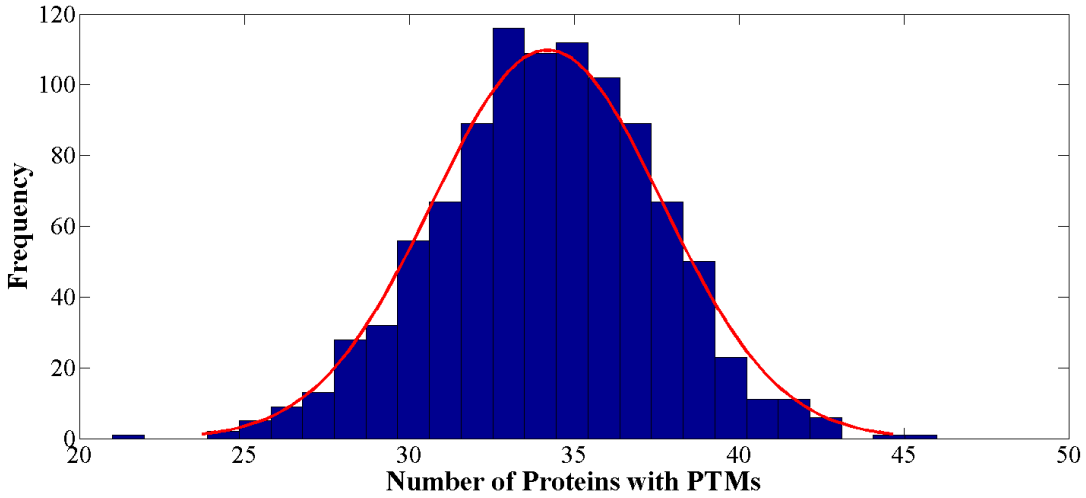


Figure 3.7: Histogram of PTM proteins identified within random subsets of 50. The distribution has a mean standard deviation of 34.286 and 3.576 respectively.

We also obtained 50 outliers from the Tuller et al., 2007 three feature predictor to compare the PTR detection ability with our five feature protein abundance predictor. Interestingly, only 37 proteins were identified as with PTM keywords providing a p-value of 0.22. This confidence level is far lower than the confidence level ($p \leq 0.02$) produced by our predictor and also the hypothesis accepting confidence boundary ($p \leq 0.05$). Therefore, despite the close prediction accuracies between the two predictors, PTM detection ability of our five feature predictor is

significantly higher than the Tuller et al., 2007's predictor. We also used mRNA and protein direct mapping and looked into PTM keywords on 50 outliers of the mRNA-protein scatter plot. Only 35 proteins were found with PTM keywords providing a p-value of 0.42, which is a lower confidence level compared to both our five feature and Tuller et al., 2007's three feature predictors. Thus, developing a protein abundance predictor using mRNA and other properties improves the ability of detecting post-translationally regulated proteins as outliers. Figure 3.8 shows the statistical confidence levels of the above three different outlier sets.

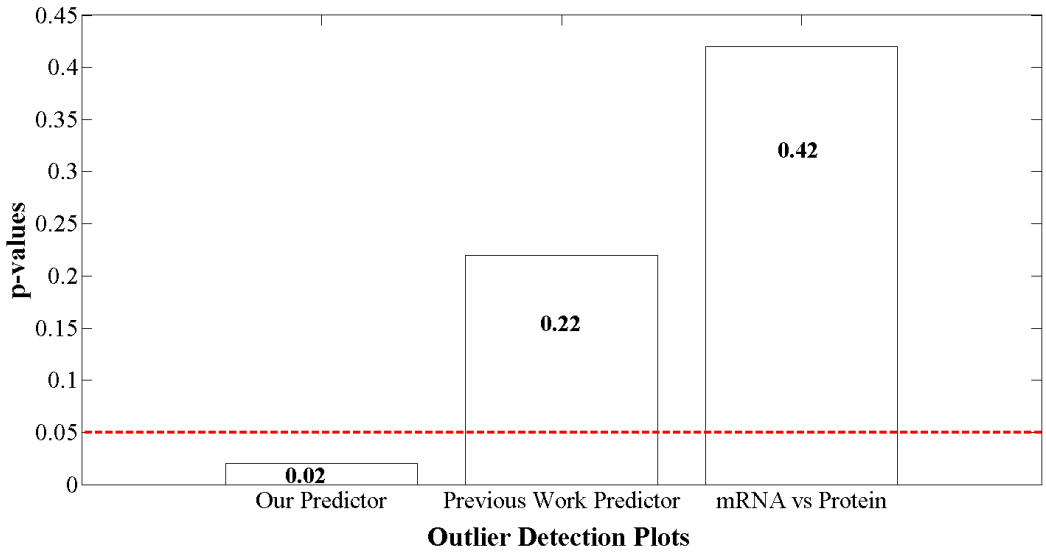


Figure 3.8: p-Values of 50 outlier samples of three different outlier detection scatter plots. Red dashed line represents the hypothesis acceptance boundary. This graph emphasizes on the following facts. (1) Our five features predictor is more capable of detecting outliers compared to previous work (Tuller et al., 2007). (2) Outliers of protein abundance predictors improve the ability of predicting PTM compared to raw mRNA and protein data scatter plot.

Further investigating on the model failures of the predicted versus measured protein abundance plot, we selected outliers by changing the cut-offs to retain 1% and 5% of the data set and carried out the same experiment. Table 3.3 shows the numbers of outliers and their level of confidence to detect post-translationally regulated proteins. Confidence level of 0.01 and 0.17 was detected for 1% and 5% cut-off boundaries respectively. Hence, when the cut-off boundary gets closer to the regression line (*i.e.* cut-off percentage increases), the number of post-translationally regulated proteins found are decreased. Thus, proteins with larger errors (further away from regression line) are more likely to be post-translationally regulated. Moreover, our predictor is capable to defeat Tuller et al., 2007's predictor and raw mRNA and protein data not only at 2.5% cut-off boundary but also at the 5% cut-off boundary with 100 outlier data.

Table 3.3: Confidence levels indicating how well the outlier subset identifies post-translationally regulated proteins, at different cut-off levels. 1000 random trials were used to obtain the p-values

Cut-off Level	Coarse Level		Finer Level	
	PTR Proteins	Confidence Level	PTR Proteins	Confidence Level
1.0% (20)	19	$p \leq 0.01$	16	$p \leq 3.05 \times 10^{-11}$
2.5% (50)	42	$p \leq 0.02$	37	$p \leq 2.11 \times 10^{-10}$
5.0% (100)	73	$p \leq 0.17$	46	$p \leq 0.001$

3.4.1.2 Level 2 : Finer Level PTM Analysis

In this functional annotation check, we combined PTM keywords with protein stability determinant motif information to provide a stronger indicator of post-translational regulation of our initial 50 outliers. Several studies show that phosphorylation of PEST motif sequences in flexible areas intensify the degradation process (García-Alai et al., 2006; Marchal et al., 1998). Similarly N-terminus segments in acetylation such as N-actelylation directly engage in the process of protein degradation (Solomon and Goldberg, 1998; Hwang et al., 2010). Ubiquitnation itself is a strong indicator for protein degradation. Further, D and Ken Box motifs signal the Anaphase Promoting Complex (APC) machinery which accelerates the degradation process by ubiquitination post-translational modification (Pfleger and Kirschner, 2000; Burton and Solomon, 2001). Addtionally, as we described in Section , there are *in-vitro* experiments showing that post-translation regulation with motif information catalyses the protein degradation and reduce the overall protein expression levels (Martinez et al., 2003; Wang et al., 2003). ABC1 calpain degradation is regulated by phosphorylation with PEST motif sequence and Martinez et al., 2003 observed a $3.9 \pm 0.4\%$ mean fold reduction of the protein concentration of ABC1 wild-type protein with respect to the PEST deleted ABC1 protein. Similarly, Wang et al., 2003 also showed that wild-type ABC1 protein gave lower protein expression levels compared to phosphorylation site mutated ABC1 proteins *e.g.* protein expression level reduction of mean fold $3.4 \pm 0.3\%$ with MutAAAA in Thr-g1286 site and $3.3 \pm 0.3\%$ with MutASSA in Thr-1305 site.

Table 3.4 shows the 50 outliers and their respective finer level functional annotations. Thirty-seven proteins out of 50 had PTM with motif information and the corresponding statistical confidence level with respect to the distribution of 1000 random samples (mean = 16.244 and standard deviation = 3.3232) is $p \leq 2.11 \times 10^{-10}$. We also obtained the annotation confidence levels for Tuller et al. (2007)'s features set and simply mRNA-protein scatter plot outliers and

Table 3.4: Finer level check - PTM + Motif keywords detected with 50 outliers

ORF Name	Gene Name	Phosphorylation + PEST	Acetylation + N-terminus	Ubiquitnation + D/Ken box
YJL129C	TRK1	✓	x	x
YBR038W	CHS2	✓	x	x
YDL093W	PMT5	x	x	x
YDL217C	TIM22	x	x	x
YFL029C	CAK1	✓	x	x
YHR031C	RRM3	✓	x	x
YJR124C	YJR124C	✓	x	x
YDL048C	STP4	✓	x	x
YGL159W	YGL159W	x	x	x
YDR006C	SOK1	✓	x	x
YIL169C	YIL169C	x	x	x
YDL222C	FMP45	✓	x	x
YDL130W	RPP1B	✓	✓	x
YCR010C	ADY2	✓	x	x
YHR141C	RPL42B	x	x	x
YBR106W	PHO88	✓	x	x
YAR075W	YAR075W	✓	x	x
YHR094C	HXT1	✓	x	x
YDR342C	HXT7	✓	x	✓
YBR1317	RPS9B	✓	x	x
YJL177W	RPL17B	x	x	x
YGR282C	BGL2	x	x	x
YBL0613	RPS8A	✓	x	x
YDR225W	HTA1	✓	✓	✓
YEL027W	VMA3	x	x	x
YKR059W	TIF1	✓	✓	x
YGL030W	YGL030W	✓	x	x
YIL148W	RPL40A	x	x	✓
YBR010W	HHT1	x	x	x
YHR021C	RPS27B	✓	x	x
YGR034W	RPL26B	✓	x	x
YER102W	RPS8B	✓	x	x
YDL083C	RPS16B	✓	x	x
YDR064W	RPS13	✓	x	x
YCR031C	RPS14A	✓	x	x
YDL081C	RPP1A	✓	✓	x
YEL034W	HYP2		✓	x
YDR447C	RPS17B	✓	x	x
YER117W	RPL23B	✓	✓	x
YKL180W	RPL17A	x	x	x
YKL056C	TMA19	x	x	x
YKL152C	GPM1	✓	x	x
YLR044C	PDC1	x	✓	x
YCR012W	PGK1	✓	✓	x
YGL123W	RPS2	x	✓	x
YDR382W	RPP2B	✓	x	x
YGR148C	RPL24B	✓	x	x
YDL014W	NOP1	✓	x	x
YDL080C	THI3	x	x	x
YER070W	RNR1	x	x	x

those two outlier sets with size 50 gave p-values of 0.0017 (26 out of 50 proteins) and 0.042 (22 out of proteins) of confidence levels respectively.

Table 3.3 shows the finer level annotation detection at the 1% and 5% cut-off levels. These results are similar to the coarse level check where the number of PTR detection as outlier decreases with the increment of cut-off percentage (getting closer to the regression line).

We observed that finer level annotation check provides a higher confidence level to support our hypothesis. Thus, by considering all cases we can conclude that the proteins identify as outliers are more likely to be post-translationally regulated. In fact, our five feature predictor outperforms in detecting post-translationally regulated proteins as outliers in both coarse level and finer level annotation checks.

Moreover, we also included finer level PTR annotation information as a binary input with our five features performed the regression to observe the protein abundance prediction accuracy knowing that the proteins are going to be post-translationally regulated or not. Interestingly, linear regression model gave $R^2 = 0.90$ regression (Figure 3.9) which is higher than our five feature predictor ($R^2 = 0.86$). In fact, adding PTR information as an input feature minimizes the overall error between the predicted and the measured abundance. Therefore, this result re-confirms that the outlier proteins with large errors between predicted and measured of the five feature predictor are likely candidates of post-translational regulation.

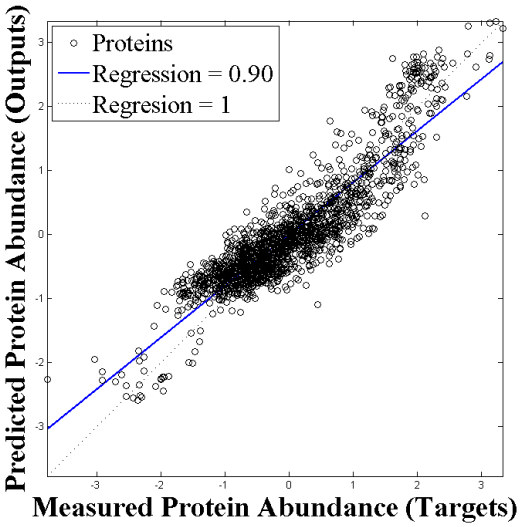


Figure 3.9: Including post-translational regulation information as the sixth feature improved the prediction accuracy of the linear predictor to $R^2 = 0.90$. Thus, outliers (with large errors) of the five feature regression model occurred due to post-translational regulation.

3.4.2 Gene Ontology (GO) Analysis

We subjected the initial 50 outlier set to a GO enrichment analysis using GOEAST web toolkit (Zheng and Wang, 2008) and found 37 GO annotations among the outlier proteins. Four from this 37 were common to more than 30 proteins, those were $GO : 0044444$, $GO : 0009058$, $GO : 1901576$ and $GO : 0032991$. Table 3.5 shows the list of annotations in detail. Interestingly, we also observed that 14 GO terms are related to ribosomal proteins. Ribosomal proteins undergo several post-translational activities such as N-termini methylation, N-termini acetylation, phosphorylation and methionine removal (Carroll et al., 2008). Therefore, ribosomal proteins provide evidence that the outlier proteins are post-translationally regulated. As we found 23 ribosomal proteins from our 50 outliers, we wanted to evaluate the effect of ribosomal proteins in our methodology. Red circles in Figure 3.10 shows the distribution of ribosomal proteins in the total data set. We can find few ribosomal proteins with high expression levels, however, their distribution was not significant with respect to the total data distribution. We repeated the entire experiment after removing the 155 ribosomal proteins from the data set and checked for the over-representation of post-translational regulation in the new 50 outliers obtained by 3% cut-off boundary. At the coarse level annotation check we observed 42 proteins with PTM keywords providing a confidence level of $p \leq 0.02$ and at the finer level 36 proteins gave $p \leq 1.38 \times 10^{-09}$ level of confidence confirming that ribosomal proteins did not unduly influence our methodology.

Further, we also carried out a pathway analysis using PANTHER (Thomas et al., 2003) web tool. Table 3.6 shows the dominant pathways discovered by the pathway analysis. Interestingly, we observed $p53$ and $p53$ pathway feedback loop pathways as dominant components in our analysis and Shin et al., 2013 showed that protein degradation process by post-translational regulation enables $p53$ regulation. Therefore, pathway analysis results re-confirmed our hypothesis by providing biological evidence that our outlier proteins are enriched with post-translational regulation.

3.4.3 Analysis of Protein Half-Life

We also investigated whether the ability of protein abundance prediction has a systematic behaviour with protein half-life because proteins with rapid degradation speeds will not be qualified. Thus, we observed the absolute and squared loss of the prediction against the protein half-lives obtained by Belle et al., 2006 which is shown in Figure 3.11, and found no meaningful correlation. Only 26 proteins (out

Table 3.5: GO Enrichment Analysis Results. Ont. stands for Ontology and those are Cellular Component (CP), Biological Component (BP) and Molecular Functions

GO ID	Ont.	Term	# Outliers	# Genome	p-Value
$GO:0044444$	CP	cytoplasmic part	36	2988	0.0127
$GO:0009058$	BP	biosynthetic process	32	2025	0.000187
$GO:1901576$	BP	organic substance biosynthetic	32	1997	0.000139
$GO:0032991$	CP	macromolecular complex	30	2137	7.53×10^{-03}
$GO:0009059$	BP	macromolecule biosynthetic process	28	1519	9.27×10^{-05}
$GO:0034645$	BP	cellular macromolecule	27	1497	0.000256
$GO:0044249$	BP	cellular biosynthetic process	27	1958	0.036
$GO:0043228$	CP	non-membrane-bounded organelle	27	1296	1.40×10^{-05}
$GO:0043232$	CP	intracellular non-membrane organelle	27	1296	1.40×10^{-05}
$GO:0019538$	BP	protein metabolic process	25	1623	0.0127
$GO:0044267$	BP	cellular protein metabolic	25	1517	4.36×10^{-03}
$GO:0010467$	BP	gene expression	25	1763	0.0458
$GO:0005829$	CP	cytosol	25	712	1.79×10^{-09}
$GO:0005840$	CP	ribosome	24	358	7.42×10^{-15}
$GO:0030529$	CP	ribonucleoprotein complex	24	2.098	2.97×10^{-08}
$GO:0006412$	BP	translation	23	670	2.83×10^{-08}
$GO:0022626$	CP	cytosolic ribosome	21	174	2.69×10^{-17}
$GO:0044445$	CP	cytosolic part	21	242	7.42×10^{-15}
$GO:0003735$	MF	structural constituent of ribosome	20	227	3.46×10^{-14}
$GO:0005198$	MF	structural molecule activity	20	372	3.60×10^{-10}
$GO:0044391$	CP	ribosomal subunit	20	240	8.71×10^{-14}
$GO:0002181$	BP	cytoplasmic translation	19	172	7.42×10^{-15}
$GO:0015934$	CP	large ribosomal subunit	11	141	1.95×10^{-06}
$GO:0022625$	CP	cytosolic large ribosomal subunit	11	93	2.83×10^{-08}
$GO:0042254$	BP	ribosome biogenesis	11	419	0.036
$GO:0006364$	BP	rRNA processing	9	318	0.0746
$GO:0016072$	BP	rRNA metabolic process	9	330	0.0951
$GO:0015935$	CP	small ribosomal subunit	9	99	1.40×10^{-05}
$GO:0022627$	CP	cytosolic small ribosomal subunit	9	64	3.22×10^{-07}
$GO:0042274$	BP	ribosomal subunit biogenesis	8	128	0.00105
$GO:0000462$	BP	maturation of SSU-rRNA	7	98	0.00177
$GO:0030684$	CP	preribosome	7	148	0.0195
$GO:0030686$	CP	90S preribosome	6	91	0.0113
$GO:0032040$	CP	small-subunit processome	4	49	0.0705

of 50 outliers) had protein half-life data and they did not show any systematic behaviour.

We also employed GMM appraoch to model our transcriptome-proteome measurements and observed that GMM provides a low confidence in detecting PTMs as outliers. See Appendix B for further details. Therefore linear regression outperforms in detecting post-translationally regulated proteins as outliers.

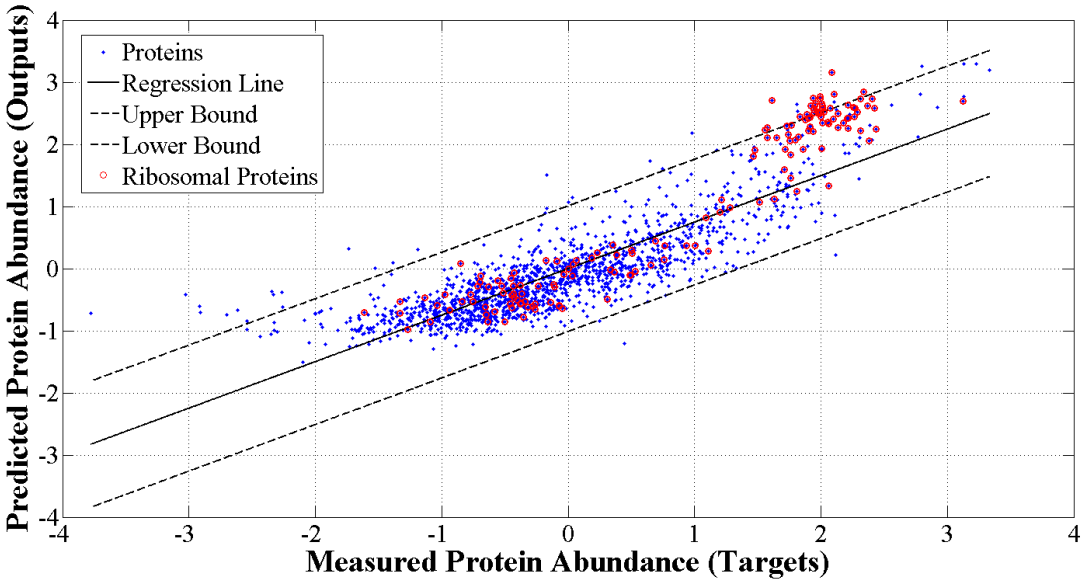


Figure 3.10: Distribution of ribosomal proteins. Red circles represents the ribosomal proteins among the data set. There were 155 ribosomal proteins in the total data set and 23 were fallen into the 50 outlier set.

Table 3.6: Pathway analysis results for 50 outliers

Pathway Accession	Pathway Name	No of Gene Components	No of Subfamilies
*P*00059	*p*53 pathway	70	103
*P*04398	*p*53 pathway feedback loops	32	64
*P*02738	De novo purine biosynthesis	23	68
*P*00017	DNA replication	18	49
*P*00024	Glycolysis	10	40
*P*02739	De novo pyrimidine deoxyribonucleotide biosynthesis	10	26
*P*02748	Isoleucine biosynthesis	5	31
*P*02785	Valine biosynthesis	4	29

3.5 Summary

In this chapter, we developed a machine learning predictor to predict protein concentration using five transcriptomic properties obtained by the LASSO feature selection and investigated on the model failures which are likely candidates of post-translational regulation. Outlier proteins, over-represented functional annotations related to post-translational regulation with high statistical confidence level. Here we consider protein stability disruption as the primary requirement for post-translation regulation and other modifications such as localization, hydrophobicity and enzymatic activities cannot be detected by this approach.

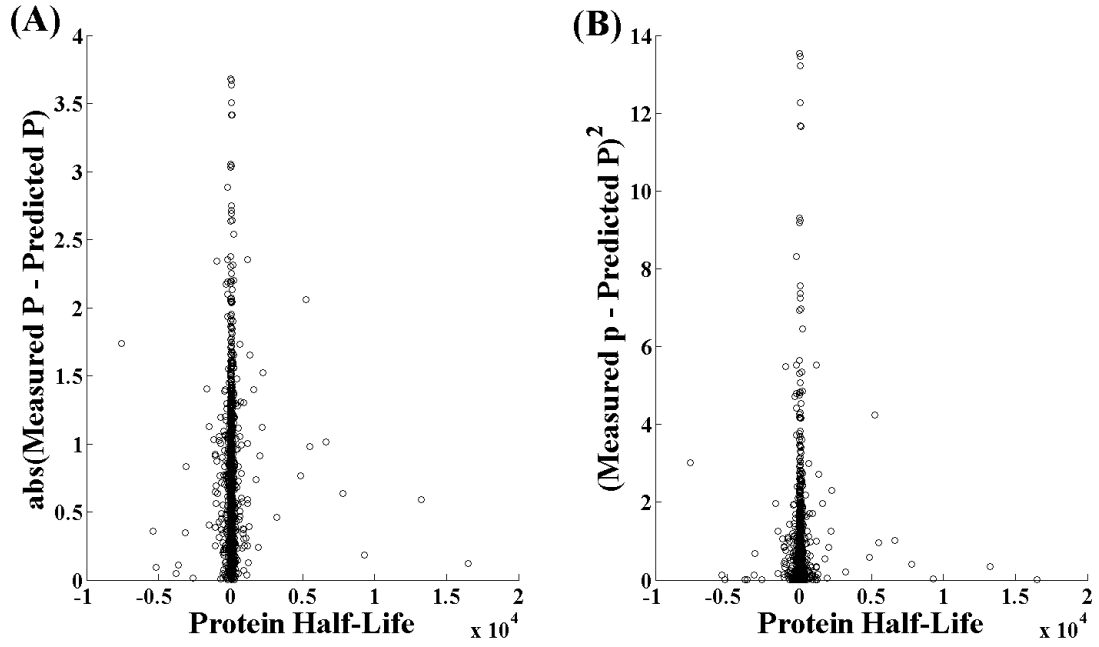


Figure 3.11: **(A)** Absolute error values of the predicted protein abundance versus protein half-lives of the relevant proteins. and **(B)** Squared error values of the predicted protein abundance versus protein half-lives of the relevant proteins.

Previous authors mostly focused on finding the correlation between mRNA and protein levels. However, in this approach we build a regression model at the transcriptome-proteome interface and use the model failures to extract useful information on post-translational regulation. This is a good example of using machine learning effectively in computational biology. Unlike computer vision or voice recognition where the performance is measured by the classification accuracy, in biology we need to cut down the experimental work space to confirm important biological properties. Thus, with this framework can reduce experimental workspace for biologists to detect post-translationally regulated proteins.

Chapter 4

Outlier Detection at the Transcriptome-Proteome Interface

In previous chapter (Chapter 3), we demonstrated that outliers detected at the transcriptome-proteome interface are likely candidates for post-translational regulation. However in Chapter 3 we used a very simple approach to obtain outliers from a linear regression model, by taking the proteins that are lying further away from the regression line (Figure 3.6). Here, we introduce two formulations of deriving a protein abundance predictor to extract outliers systematically from a regression model. First, we formulate novel Outlier Rejecting Regression (ORR) model which has the capability to obtain a proportion of the data as robust outliers by truncating or clipping the loss function, where the regression problem becomes non-convex (Xu et al., 2006). Thus, we use Difference of Convex functions Algorithm (DCA) and an alternative ad-hoc variant of optimization strategy to solve this non-convex problem. As the second method, we use Quantile Regression (QR) (Koenker, 2005) approach which allows asymmetric conditional loss models to extract outliers only with negative losses or positive losses. Based on our hypothesis (Chapter 3), we are more interested with the negative loss outliers where the measure abundance is lower than the predicted ($P < \hat{P}$). Therefore, quantile regression technique is more suitable with our data-driven framework. We believe that these two new methods are much neater ways of selecting outliers. In order to remove the ambiguity between the three models, we numbered them as below.

- Model 0 - Simple Linear Regression used in Chapter 3

- Model 1 - Outlier Rejecting Regression (ORR)
- Model 2 - Quantile Regression (QR)

With all above regression models, we use the five transcriptomic protperties obtained by LASSO features selection as main inputs (*i.e.* mRNA abundance, tRNA adaptation index (tAI), codon bias, ribosome density and occupancy).

4.1 Outlier Rejecting Regression (ORR) - Model 1

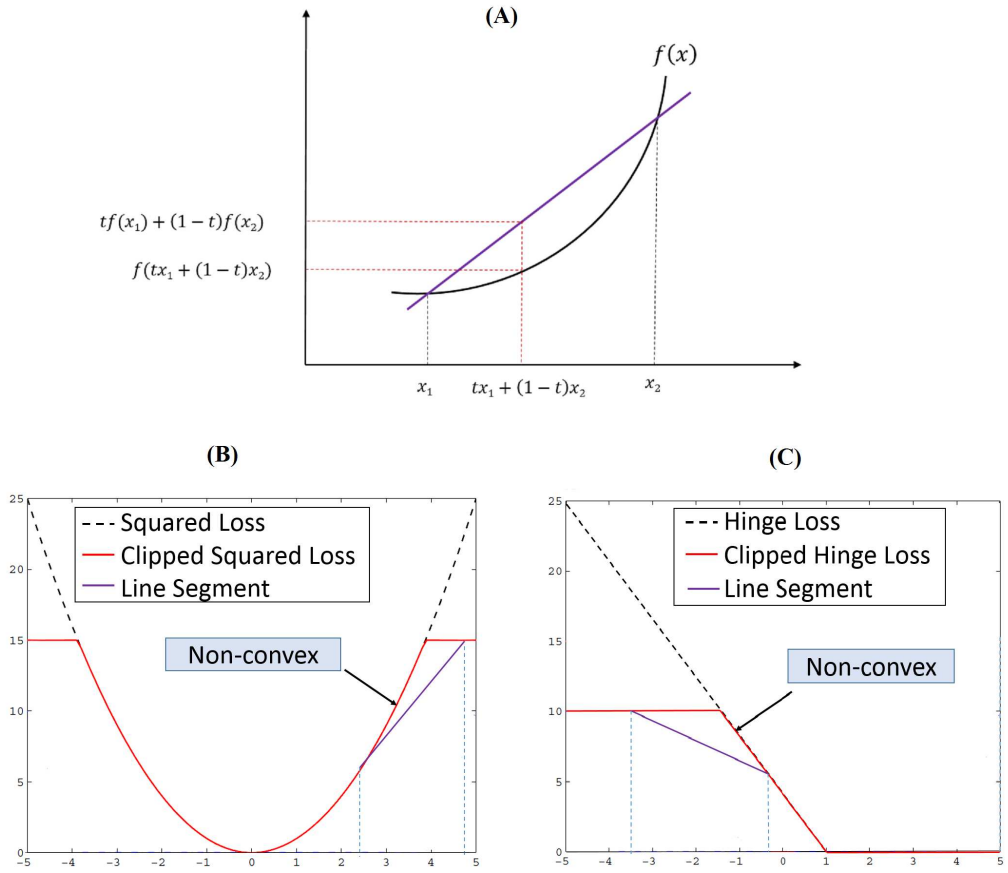


Figure 4.1: Convex and non-convex functions. **(A)** is an example of a convex function. **(B)** Squared loss and **(C)** hinge loss functions before (black dashed line) and after (red solid line) loss clipping. Purple colour lines segments in **(B)** and **(C)** represent the violation of convex definition by loss clipping.

4.1.1 Clipped Loss Functions

In regression problems, robustness of the outliers are determined by the loss function. A convex loss model is largely sensitive to outlier samples and a single

outlier can vary the total regression (Yu et al., 2010; Huber, 2011). Therefore, loss clipping or truncation is used to extract robust outliers from convex loss models. Following is the the definition of a convex function;

A function $f(\boldsymbol{x})$ is called convex, if a line segment draws between any two points in the function lies above the function graph in a Euclidean space. Mathematical definition of a convex function is given below;

f is called convex if:

$$\forall x_1, x_2 \in X, \forall t \in [0,1] : f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

Figure 4.1(A) represents an example of a convex function. In fact, Figure 4.1(B) and (C) black dashed lines show real world squared and hinge convex loss models respectively.

However, loss clipping converts convex functions in to non-convex functions. Figure 4.1(B) and Figure 4.1(C) red solid lines show the clipped loss functions of squared and hinge loss models respectively and they violate the convex definition. Therefore, clipped squared loss and clipped hinge loss are non-convex functions.

Loss Clipping in ORR Model

Suppose we have a set of m samples $\{(\boldsymbol{x_i}, y_i)\}_{i=1,\ldots,m}$ where $\boldsymbol{x_i} \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. Our main objective is to predict y as $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ with smallest error. We define the *clipped* loss function as below:

$$\ell_U(\boldsymbol{x}, y; \boldsymbol{w}, b) := \min\{U, \ell(\boldsymbol{x}, y; \boldsymbol{w}, b)\}$$

where a hyper parameter $U > 0$ denotes the clipping position. Figure 4.2 illustrates an example of truncated (clipped) squared loss model with clipping position 15.

Here we use L_2 regularized loss function (Equation (4.1)) which is also known as ridge regression with the ORR model.

$$\min_{\boldsymbol{w},b} \quad \{y - (\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)\}^2 + \lambda \|\boldsymbol{w}\|^2, \tag{4.1}$$

where $\lambda > 0$.

This is similar to the ordinary least square loss function (Equation (3.2)), but with an additional regularization term (L_2) which helps to penalize the data overfitting and model complexity. Ridge regression is almost a standard approach with

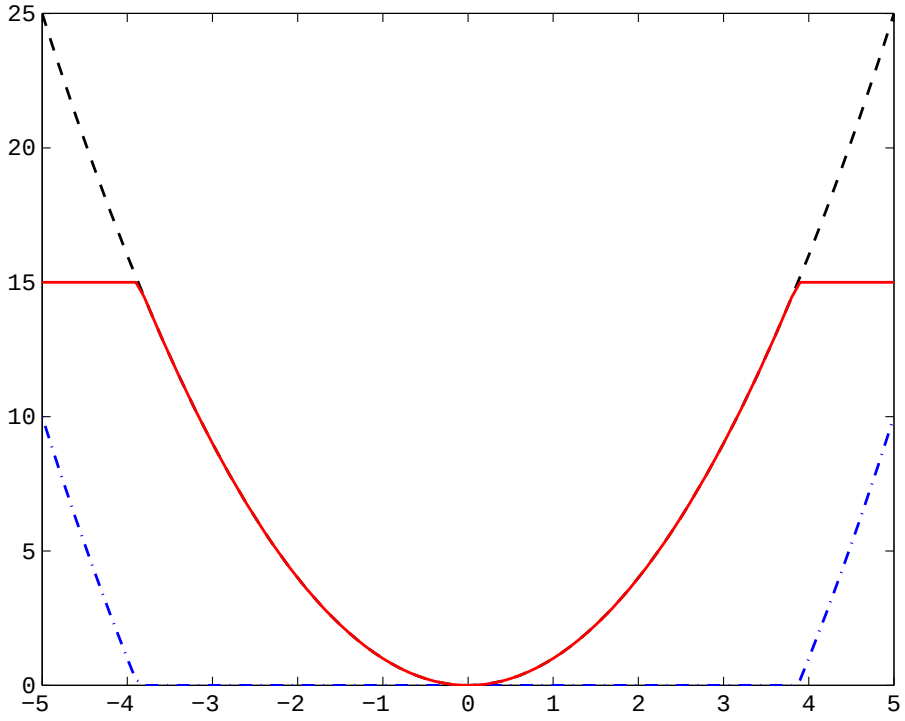


Figure 4.2: Example of truncated squared loss: $\ell_U(\boldsymbol{x}, y; \boldsymbol{w}, b)$ with $U = 15$ for the squared loss function $\ell(\boldsymbol{x}, y; \boldsymbol{w}, b)$.

its roots in Tikhonov regularization, Bayesian methods with zero mean Gaussian prior and incorporating a ridge to achieve numerical stability in matrix analysis. However, any loss function, such as ordinary least square, episilon-insensitive, hinge loss or logistic loss can be used with the ORR model. Following is the clipped loss function for the ridge regression model:

$$\min_{\boldsymbol{w},b} \quad \lambda\|\boldsymbol{w}\|^2 + \sum_i \ell_U(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b), \tag{4.2}$$

where $U > 0$ and $\lambda > 0$ are hyper parameters.

Controlling U is difficult with respect to the number of samples we need to extract as outliers at the end of the optimization. Therefore, we introduce a new parameter $\mu \in [0, 1)$ where the user can define the number of outlier samples needed as a ratio from the total data samples. Reformulated model using μ instead of U is given below;

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\eta}} \quad & \frac{1}{(1-\mu)m}\sum_i \eta_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) + \lambda\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & \textstyle\sum_i(1-\eta_i) \le \mu m, \quad 0 \le \eta_i \le 1, \ \forall i, \end{aligned} \tag{4.3}$$

where $\mu \in [0, 1)$ and $\lambda \in (0, \infty)$ are hyper parameters.

Here we believe that the specifying a fraction of data as outlier using our prior knowledge is easier than using a clipping position U. See Appendix C for the relationship between Equation (4.2) and (4.3). Note that $\sum_i(1-\eta_i) = \mu m$ holds at the optimal solution, where the samples $(\boldsymbol{x_i}, y_i)$ with $\eta_i^* = 0$ is considered as outliers for small $\mu > 0$. However, this optimization problem is non-convex and finding a global solution for a non-convex problem is very difficult.

4.1.2 Difference of Convex Functions Algorithm (DCA)

Difference of convex Functions algorithm (DCA) is a mathematical approach to obtain plausible solutions for non-convex optimization problems (Pham Dinh and Le Thi, 1997; Collobert et al., 2006). Therefore, we use DCA to solve our non-convex ORR regression model (Equation (4.3)). Appendix C shows step by step derivation of the clipped loss model using DCA. In order to solve Equation (4.3), DCA updates the $\boldsymbol{\eta}$ and $(\boldsymbol{w}, b)$ in each iteration alternatively. Selection of $\boldsymbol{\eta^k}$ in kth iteration is follows:

$$\begin{aligned} \boldsymbol{\eta^k} \in \quad & \text{argmax}_{\boldsymbol{\eta}} \textstyle\sum_i(1-\eta_i)\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}_k, b_k) \\ \text{s.t.} \quad & \textstyle\sum_i(1-\eta_i) = \mu m, \quad 0 \le \eta_i \le 1. \end{aligned} \tag{4.4}$$

We obtain the $\boldsymbol{\eta^k}$ by sorting the squared loss and assigning the 0 to largest μm samples. Therefore, $(\boldsymbol{w_{k+1}}, b_{k+1})$ is computed using η^k as a solution to the following subproblem:

$$h(\boldsymbol{w}_{k+1}, b_{k+1}) := \min_{\boldsymbol{w},b} \frac{1}{(1-\mu)m}\left[\sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) - \mu m(\langle \boldsymbol{g_w}, \boldsymbol{w}\rangle + g_b b)\right] + \lambda\|\boldsymbol{w}\|^2, \tag{4.5}$$

where

$$\boldsymbol{g_w} = \frac{1}{\mu m}\sum_i (1-\eta_i^k)\nabla_w \ell(\boldsymbol{x_i}, y_i; \boldsymbol{w}^k, b^k), \tag{4.6}$$

$$\boldsymbol{g_b} = \frac{1}{\mu m}\sum_i (1-\eta_i^k)\nabla_b \ell(\boldsymbol{x_i}, y_i; \boldsymbol{w}^k, b^k). \tag{4.7}$$

$\nabla_w \ell(.)$ and $\nabla_b \ell(.)$ denote subgradients of a nonsmooth loss function $\ell(.)$ with respect to $\boldsymbol{w}$ and b.

Algorithm 2 shows the pseudo code for the implementation of Outlier Rejecting Regression (ORR) model using DCA. `CVX` package in `MATLAB` was used as the development environment. Algorithm 2 generates a sequence of $(\boldsymbol{w}^k, b^k)$, which has following good convergence properties: in each iteration the objective value of Equation (4.3) decreases and every limit point is a critical point which satisfies a necessary condition of local minima in Equation (4.3).

Algorithm 2 DCA for Outlier Rejecting Regression

Require: Initial $(\boldsymbol{w^0}, b^0)$; hyper-parameters $\mu \in [0, 1)$ and $\lambda \in (0, \infty)$.
  $k \leftarrow 0$.
  repeat
    Obtain $\boldsymbol{\eta^k}$ from Equation (4.4) by sorting $\ell(\boldsymbol{x_i}, y_i; \boldsymbol{w}^k, b^k), \forall i$.
    Computer $(\boldsymbol{g_w}, g_b)$ using $\boldsymbol{\eta^k}$ as shown in Equation (4.6) and 4.7.
    $(\boldsymbol{w^{k+1}}, b^{k+1}) \leftarrow$ a solution of subproblem in Equation (4.5).
    $k \leftarrow k + 1$.
  until convergence.

4.1.3 Alternative Heuristic Implementation of DCA in ORR

We also developed an alternative heuristic `MATLAB` implementation of DCA in ORR model as shown in Algorithm 3. Equation (4.8) was used (without subgradients) as the main optimization problem in this implementation and set the λ and μ values similar to Algorithm 2 ($\lambda = 0.01$ and $\mu = 0.975$). We observed that both implementations select the same set of proteins as outliers, providing identical results.

Though these two algorithms produce similar results, heuristic Algorithm 3 is more intuitive and easier compared to Algorithm 2. The only difference between these two algorithms is the step where the subproblem is solved.

$$\min_{\boldsymbol{w},b,\boldsymbol{\eta}} \quad \frac{1}{(1-\mu)m} \sum_i \eta_i^k \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) + \lambda \|\boldsymbol{w}\|^2. \tag{4.8}$$

Algorithm 3 Alternative Heuristic Implementation of Outlier Rejecting Regression

Require: Initial $(\boldsymbol{w^0}, b^0)$; hyper-parameters $\mu \in [0, 1)$ and $\lambda \in (0, \infty)$.
  $k \leftarrow 0$.
  repeat
    Obtain $\boldsymbol{\eta^k}$ from Equation (4.4) by sorting $\ell(\boldsymbol{x_i}, y_i; \boldsymbol{w}^k, b^k), \forall i$.
    $(\boldsymbol{w^{k+1}}, b^{k+1}) \leftarrow$ a solution of subproblem in Equation (4.8).
    $k \leftarrow k + 1$.
  until convergence.

See Appendix C for `MATLAB` implementation scripts of Algorithm 2 (function ORR1) and 3 (function ORR2).

4.1.4 ORR Convergence Speed

We employed four data sets with different dimensionalities to compare the convergence speeds of Algorithm 2 and 3 in ORR model. Those are,

1. Transcriptome-proteome data from Chapter 3 (n≈ 2000,d=5),
2. Boston Housing data (n≈ 500,d=14) : Consists of housing values in suburb area of Boston (Harrison and Rubinfeld, 1978),
3. Concrete Compressive Strength data (n≈ 2000,d=9) : This data set comprises of important attributes to determine concert compressive strength in a civil engineering problem domain (Yeh, 1998),
4. KEGG Metabolic Network data (n≈ 50000,d=22): KEGG metabolic pathway details can be used to model directed relational or reaction networks between pathways (Shannon et al., 2003).

Last three data sets were downloaded from UCI Machine Learning Repository (Bache and Lichman, 2013). Figure 4.3 shows the convergence speed of two algorithms with the number of iterations. In all cases, Algorithm 3 converges faster than Algorithm 2, providing the same error. Though convergence speed is not a dominant factor with our regression problem (transcriptome-proteome data), Algorithm 3 well be a better solution for large regression problems (data sets with very high dimensionality).

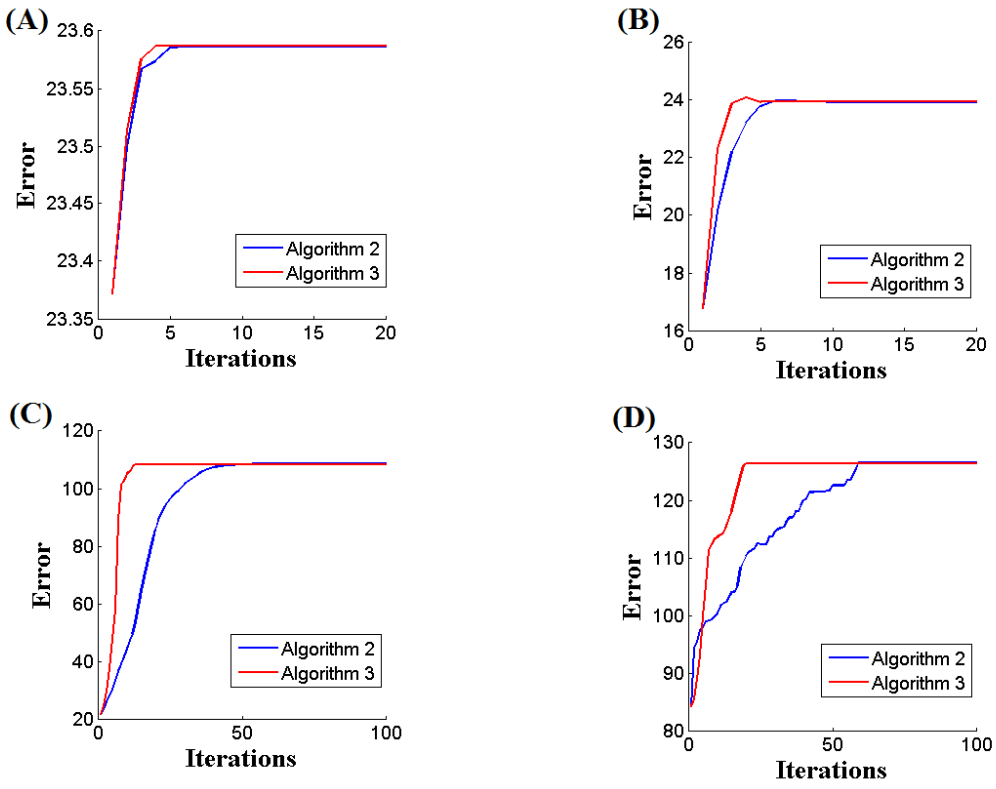


Figure 4.3: Blue and red lines represent the convergence of Algorithm 2 and Algorithm 3 respectively. (A) represents the transcriptome-proteome data set (n≈ 2000,d=5) from Chapter 3 and (B) Boston Housing data (n≈ 500,d=14), (C) Concrete Compressive Strength data (n≈ 2000,d=9) and (D) KEGG Metabolic Network data (n≈ 50000,d=22) were downloaded from UCI Machine Learning Repository (Bache and Lichman, 2013). In all cases, Algorithm 3 converges faster than Algorithm 2.

4.2 Quantile Regression (QR) - Model 2

Quantile regression is desirable if we are interested in conditional losses, where the user can define the outliers to be selected either with positive or negative losses. The main advantage of this technique over the ordinary least square (OLS) technique (Equation (3.2)), is that the outlier detection is more robust with respect to the response measurements. In fact, OLS method is inefficient if the errors are non-normal. However, QR provides a wide description of the data by looking at impact of a covariate on the entire distribution of outputs, not merely the conditional mean (Koenker and Geling, 2001; Koenker, 2005).

QR has been widely used in many real world applications such as economics (Hendricks and Koenker, 1992; Koenker, 2005), medicine (Cole and Green, 1992; Heagerty and Pepe, 1999) and survival analysis (Koenker and Geling, 2001) to detect outliers in an asymmetric problem domain. Here we use quantile regression for the first time with joint *omic* measurements, to detect outliers with negative losses (measured abundance lower than predicted - $P < \hat{P}$) as likely candidate for post-translational regulation.

Different weight will be assigned to negative and positive losses, where $y-(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)$, is considered as the loss function. Parameter $\tau \in (0, 1)$ is used to define the quantile of interest and $\tau = 0.5$ represents symmetric error with conditional median (Koenker, 2005). Quantile loss is given as $\rho_\tau(y - (\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b))$ where

$$\rho_\tau(z) = \begin{cases} \tau \cdot (z) & \text{if } (z) \geq 0, \\ -(1-\tau) \cdot (z) & \text{otherwise.} \end{cases} \tag{4.9}$$

The outliers of our interest (*i.e.* in our case $(y-(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)) < 0$) can be extracted by setting the required τ value in Equation (4.10),

$$\min_{\boldsymbol{w},b} \quad \sum_i \{\rho_\tau(y_i - (\langle \boldsymbol{w}, \boldsymbol{x_i} \rangle + b))\} \tag{4.10}$$

As shown in Equation (4.11), linear programming in `MATLAB` environment was employed to solve quantile regression problem.

$$\begin{aligned} \min_{u,v,\boldsymbol{w},b} \quad & \sum_i^m [\tau u_i + (1-\tau) v_i] \\ \text{s.t.} \quad & y_i - f(\boldsymbol{x_i}) = u_i - v_i, \ \forall i \\ & u_i \geq 0, \ v_i \leq 0 \end{aligned} \tag{4.11}$$

4.3 Validating ORR and QR Models

Hawkins et al. (1984)'s synthetic data set was employed to validate ORR model (Model 1). This is a popular data set, specially used to detect outliers with symmetric loss (both positive and negative losses) regression problem (Rousseeuw and Leroy, 1987; Colin, 2002; Atkinson and Riani, 2000; Hadi, 1992). This data set contains 75 samples with three input features and first 14 samples are considered as outliers. Outlier samples are divided into two groups; group 1 - sample 1 to 10

with positive losses and group 2 - sample 11 to 14 with negative losses. However, a classical regression outlier detection techniques can only detect sample 12, 13 and 14 as outliers (mainly group 2) (Rousseeuw and Leroy, 1987). We set μ to 0.1867 to obtain the 14 outliers and λ and ϵ were set to 0.01 and 0.00001 respectively. Interestingly, all 14 outliers were detected (observations $1-14$) and they were also clustered into two groups. Figure 4.4(A) shows the detected outliers *i.e.* observations $1-10$ (group 1) in circles and observations $11-14$ (group 2) in crosses.

Boston Housing Data (n$\approx$ 500,d=13) from UCI Machine Learning Repository (Harrison and Rubinfeld, 1978; Bache and Lichman, 2013) was used to detect one-side outliers (either with positive or negative losses) using QR model (Model 2). Here, majority of the outliers were found with positive losses by simply using a linear regression model (*i.e.* we assumed data with larger error as outliers). Therefore, by setting $\tau = 0.96$, we obtained the most dominant 20 outliers only with positive losses using QR model. Figure 4.4(B) shows the 20 data points detected as outliers only with positive losses. Thus, QR model is capable of detecting one-side outliers.

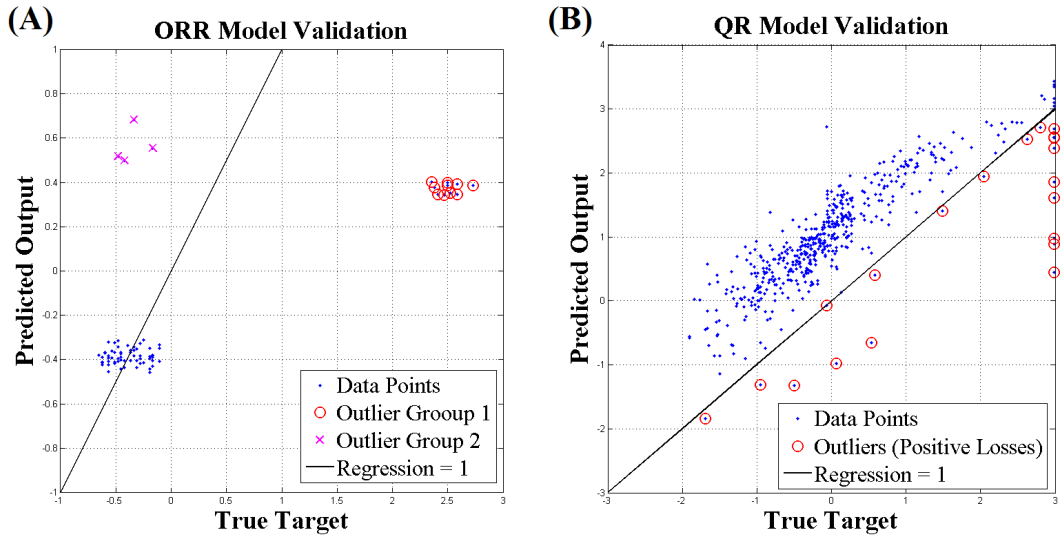


Figure 4.4: Validating outlier detection by ORR and QR models. (A) ORR model validation was carried out using Hawkins et al. (1984)'s synthetic data set. Circles represent the group 1 (observations $1-10$) and crosses represent the group 2 (observations $11-14$). (B) QR with an asymmetric loss model was validated using Boston Housing data in UCI Machine Learning repository (Harrison and Rubinfeld, 1978; Bache and Lichman, 2013) data set. Red circles represent the most dominant 20 outliers (only with positive losses) detected by setting τ parameter to 0.96.

4.4 Results

Five best features at the transcriptome level (*i.e.* mRNA abundance, ribosome occupancy, ribosome density, tAI and codon bias) of *Saccharomyces cerevisiae* were employed as the inputs for the regression models and the respective protein abundances were used as the outputs. With linear regression model (Model 0) in Chapter 3, we used a simple empirical approach to detect outliers by selecting the proteins that are lying further away from the regression line. However, here we use more mathematical or systematic ways of detecting outlier using ORR and QR models. Further, with Model 0, we used 2.5% as the cut-off boundary and obtained 50 samples as the bench mark number of outlier proteins to prove our hypothesis. Therefore, with new ORR and QR models also we obtained 50 samples as outliers by setting $\mu = 0.975$ and $\tau = 0.025$ respectively. Figure 4.5 shows the three regression plots (Model 0, 1 and 2) with the detected outlier points as red circles. We observed that, all three regression models gave a good level of protein abundance prediction, *i.e.* $R^2 = 0.86$ for simple linear regression (Model 0), $R^2 = 0.86$ for ORR model (Model 1) and $R^2 = 0.85$ for QR model(Model 2). Further, we also compared the prediction outputs of ORR and QR models. Figure 4.6 shows that these two new regression approaches produce highly correlated ($R^2 = 0.97$) outputs.

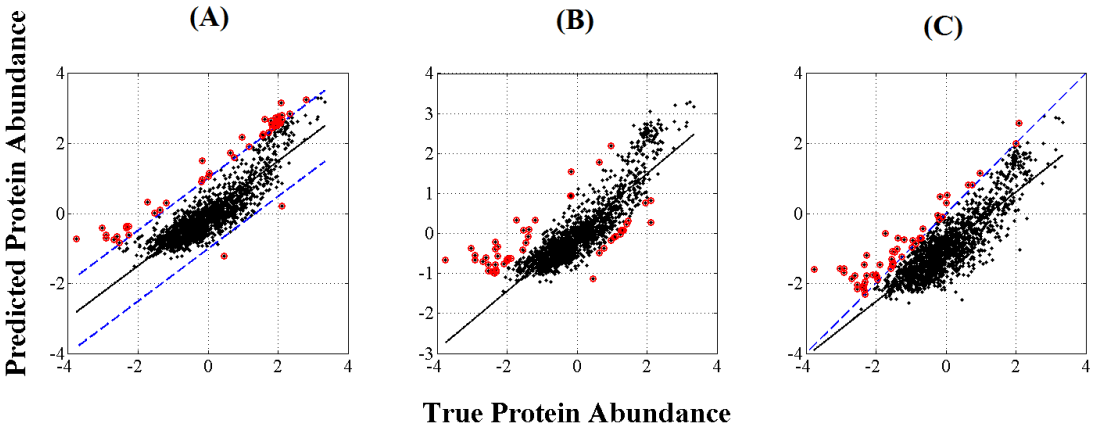


Figure 4.5: Outlier detection by three regression models: solid lines represent the regression of each model and outliers are shown in red colour circles. (A) Linear Regression (Model 0) in Chapter 3 - blue dashed line shows the 2.5% cut-off boundary where the proteins found far away from the regression line were obtained as outliers (solid line $R^2 = 0.86$) (B) Outlier Rejecting Regression (Model 1) - selects the least accurate 50 outliers using symmetric squared loss (solid line $R^2 = 0.86$) and (C) Quntitle Regression (Model 2) - blue dashed line represents the $R^2 = 1$ line where $y = f(\boldsymbol{x})$. However, this model selects the proteins only with negative errors where $y - f(\boldsymbol{x}) < 0$ (solid line $R^2 = 0.85$).

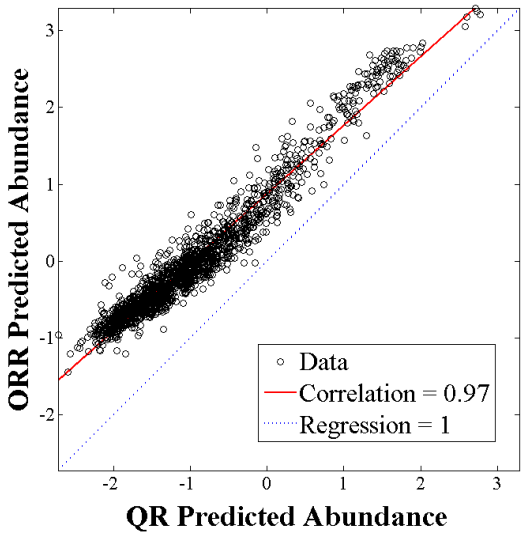


Figure 4.6: Comparison of ORR and QR output predictions. These two models produce highly correlation of $R^2 = 0.97$ outputs, showing the agreement in model fitting, but they identify different data as outliers (see Figure 4.7) due to the difference between imposed loss functions.

4.4.1 PTR Detection in Outlier Proteins

Similar to the previous Chapter (Section 3.4), functional annotation check was performed at the coarse level and finer level, to provide evidence of post-translation regulation of these outlier proteins. Table 4.1 shows the statistical confidence levels obtained by the outlier sets (size 50) of the three regression models at the coarse and finer level functional annotation checks.

Table 4.1: Function annotation check results of the three set of outliers (each set contains 50 proteins). 1000 random samples were used obtain the p-values

Regression Model	Coarse Level		Finer Level	
	No of genes	Confidence Level	No of genes	Confidence Level
Model 0	44	$p \leq 0.020$	37	$p \leq 2.11 \times 10^{-10}$
Model 1	40	$p \leq 0.048$	35	$p \leq 8.31 \times 10^{-09}$
Model 2	45	$p \leq 9.89 \times 10^{-04}$	38	$p \leq 2.94 \times 10^{-11}$

Outlier Rejecting Regression (ORR) PTR Detection

Figure 4.5(B) shows the ORR model outliers as red circles in the predicted versus true protein concentrations scatter plot. λ value was determined by cross-validation and set to 0.01. Forty proteins were found with PTM key word at the coarse level providing a $p \leq 0.048$. Thirty three proteins were detected at the finer level annotation check (PTMs+motif) with a high confidence level of

$p \leq 8.31 \times 10^{-09}$. Though these confidence levels are lower than our previous model (Model 0) outliers, these p-values can also be considered as high confidence levels with respect to the confidence level threshold $p < 0.05$ of accepting a hypothesis in biomedical research (McDonald, 2009). Thus, the outliers detected by ORR model are also highly enriched with post-translational regulation.

Quantile Regression (QR) PTR Detection

By setting the τ value to 0.025, we obtained 50 proteins with negative errors ($P < \hat{P}$) from the upper region of the regression plot (*i.e.* see the Figure 4.5(C)). According to this method, all the other proteins had positive error values ($P > \hat{P}$). Thus, these results are more appropriate with respect to our main hypothesis (*i.e.* proteins with $P < \hat{P}$ are more likely to be post-translationally regulated). We carried out both coarse and finer levels annotation check with the new outlier set. Forty-five proteins were detected at the coarse level check providing $p \leq 9.89 \times 10^{-04}$ and with finer level annotation check 38 proteins found with PTMs + motifs information giving p-value $< 2.94 \times 10^{-11}$.

All three regression models have high level of confidence ($p < 0.05$) to support our hypothesis at both coarse level and finer level. Note that quantile regression model gives the highest confidence level to detect post-translationally regulated proteins as outliers by excluding the false positives.

4.4.2 Biological Insights of Outlier Proteins

Figure 4.7 shows the overlap of the outlier proteins between the three regression models in a Venn diagram. We found 92 proteins as the union of the Venn diagram and 17 as the outlier intersection of the three models. Appendix D shows the 92 outlier proteins detected by the three regression models with their corresponding outlier detection techniques. Here, we use gene enrichment analysis and protein-protein interaction networks to extract more biological interpretations of these outlier proteins.

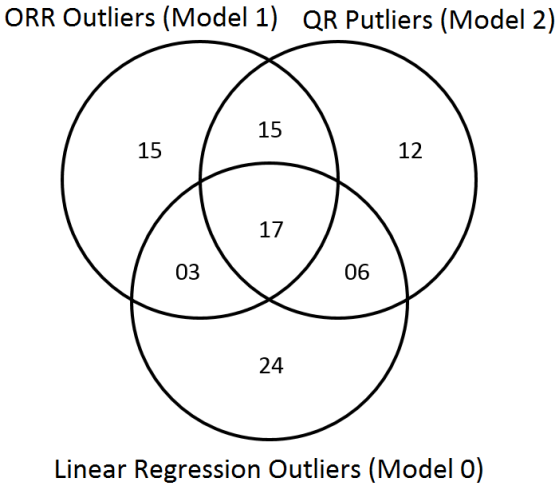


Figure 4.7: Distribution of outlier proteins between the three regression models in a Venn diagram.

4.4.2.1 Gene Enrichment Analysis

Gene Ontology (GO) and pathway analyzes were carried out using BiNGO 2.44 (a plug-in for Cytoscape) (Maere et al., 2005) and PANTHER (Thomas et al., 2003) analysis tools respectively. WebGestalt web tool (Zhang et al., 2005) was also employed to enhance the analysis results using multiple large databases.

Table 4.2 shows the most dominant GO terms and their respective p-values found with the union set of outliers (92 proteins). Additionally, Figure 4.8 illustrates these annotations in a hierarchical structure under three main categories; (A) Biological Process, (B) Cellular Components and (C) Molecular Functions. Note that ribosomal GO terms were frequent in all three categories. Further, KEGG and Wikipathway results obtained by WebGestalt web tool (Table 4.3) also showed several keywords related to ribosomal proteins. As mentioned in Chapter 3, ribosomal proteins react with several post-translational regulation processes such as phosphorylation, N-terminal acetylation, N-terminal methylation and removal of methionine (Carroll et al., 2008). However, in previous chapter we showed that ribosomal protein did not unduly influence the model failures of the global regression model.

Pathway analysis was carried out using consensus (intersections) and union set of outliers. Interestingly, with all the combinations (considering two models at a time and all three), $p53$ pathway and $p53$ pathway feedback loop significantly over-represented in these outlier proteins. Shin et al. (2013)'s study showed that protein degradation process of post-translational regulation enables $p53$ regulations in a robust manner. In fact, during DNA damage conditions, $p53$ levels are down

Table 4.2: Most dominant (over-represented) GO keywords by BiNGO Analysis

GO-ID	Description	No (%) of Genes	Corrected p
	Biological Process		
$GO:0009058$	biosynthetic process	51(56.6%)	3.76×10^{-03}
$GO:0006725$	cellular aromatic compound process	8 (8.8%)	4.42×10^{-02}
$GO:0044249$	cellular biosynthetic process	47(52.2%)	3.68×10^{-03}
$GO:0030490$	maturation of SSU-rRNA	8(8.8%)	4.80×10^{-03}
$GO:0000462$	maturation of SSU-rRNA from tricistronic rRNA	8(8.8%)	3.76×10^{-03}
$GO:0015146$	pentose transmembrane transporter activity	2(2.2%)	4.42×10^{-02}
$GO:0010608$	post-transcriptional regulation	12(13.3%)	3.00×10^{-03}
$GO:0032268$	cellular protein metabolic process	12(13.3%)	3.76×10^{-03}
$GO:0051246$	protein metabolic process	12(13.3%)	2.82×10^{-02}
$GO:0006417$	regulation of translation	12(13.3%)	1.60×10^{-03}
$GO:0000028$	ribosomal small subunit assembly	3(3.3%)	4.89×10^{-02}
$GO:0016072$	rRNA metabolic process	11(12.2%)	4.62×10^{-02}
$GO:0006364$	rRNA processing	11(12.2%)	3.96×10^{-02}
$GO:0070181$	SSU rRNA binding	2(2.2%)	4.42×10^{-02}
$GO:0006412$	translation	28(31.1%)	7.16×10^{-03}
	Cellular Component		
$GO:0032991$	macromolecular complex	42(46.6%)	1.92×10^{-02}
$GO:0043228$	non-membrane-bounded organelle	31(34.4%)	3.89×10^{-03}
$GO:0030529$	ribonucleoprotein complex	24(26.6%)	7.73×10^{-05}
$GO:0043232$	intracellular non-membrane-bounded organelle	31(34.4%)	3.89×10^{-03}
$GO:0044444$	cytoplasmic part	58(64.4%)	1.74×10^{-02}
$GO:0005840$	ribosome	22(24.4%)	9.84×10^{-09}
$GO:0005829$	cytosol	29(32.2%)	6.09×10^{-06}
$GO:0044445$	cytosolic part	23(25.5%)	1.41×10^{-11}
$GO:0022626$	cytosolic ribosome	21(23.3%)	1.41×10^{-11}
$GO:0030686$	90S preribosome	6(6.6%)	4.42×10^{-02}
$GO:0033279$	ribosomal subunit	20(22.2%)	1.68×10^{-08}
$GO:0015934$	large ribosomal subunit	11(12.2%)	3.99×10^{-04}
$GO:0015935$	small ribosomal subunit	9(10.0%)	9.14×10^{-04}
$GO:0022625$	cytosolic large ribosomal subunit	11(12.2%)	1.06×10^{-05}
$GO:0022627$	cytosolic small ribosomal subunit	9(10.0%)	3.47×10^{-05}
	Molecular Function		
$GO:0005198$	structural molecule activity	21(23.3%)	8.58×10^{-06}
$GO:0003735$	structural constituent of ribosome	20(22.2%)	1.68×10^{-08}
$GO:0016829$	lyase activity	7(7.7%)	2.29×10^{-02}
$GO:0016882$	cyclo-ligase activity	2(2.2%)	2.82×10^{-02}
$GO:0015146$	pentose transmembrane transporter activity	2(2.2%)	4.42×10^{-02}
$GO:0070181$	SSU rRNA binding	2(2.2%)	4.42×10^{-02}

Third column represents the number and the percentage of genes from the outlier sets

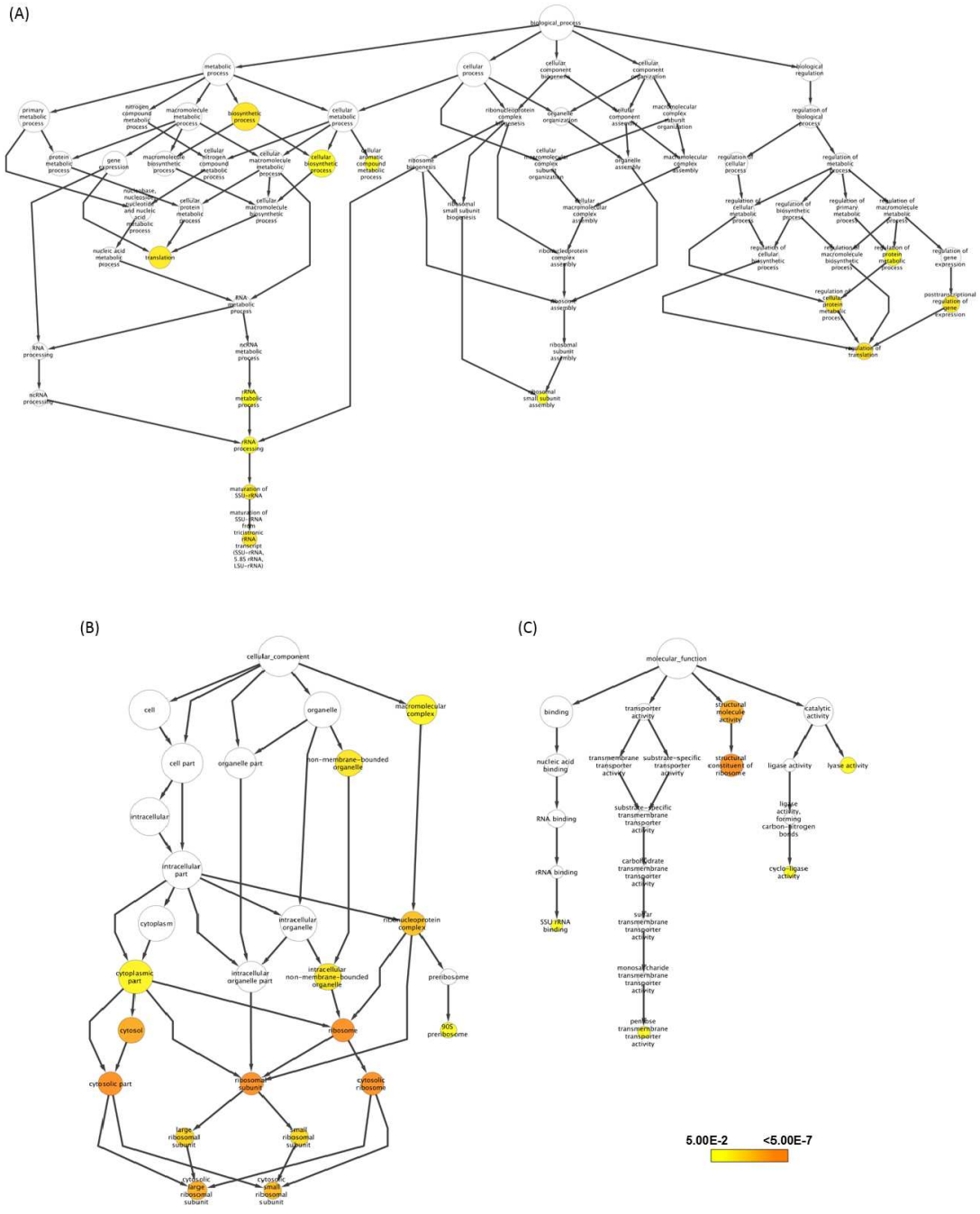


Figure 4.8: Hierarchical structure of the significantly over-represented GO terms using union outlier proteins (N=92): **(A)** Biological Process **(B)** Cellular Component and **(C)** Molecular Function are the three main categories. Each node represents a GO term and branches divide into smaller and more specific categories from top to bottom. Size of the node demonstrates the number of genes related to each GO term. Level of statistical significance associated with each GO term is illustrated as pseudo-colour where the red being most significant.

Table 4.3: Gene Enrichment Analysis by WebGestalt Tool

ID	Description	Data Source	No(%) of Genes	Corrected p
3010	ribosome	KEGG	19 (13.0%)	5.70×10^{-13}
$WP210$	cytoplasmic ribosomal proteins	Wikipathways	17(11.1%)	1.08×10^{-12}
$WP224$	aerobic glycerol catabolism	Wikipathways	3(12.7%)	1.05×10^{-02}
$WP178$	isoleucine degradation	Wikipathways	2(20.6%)	1.77×10^{-02}
$WP112$	carbon metabolism	Wikipathways	5(4.4%)	1.92×10^{-02}
$WP340$	glucose fermentation	Wikipathways	3(7.4%)	2.04×10^{-02}
$WP515$	glycolysis and gluconeogenesis	Wikipathways	3(5.4%)	4.13×10^{-02}
$WP253$	glycolysis	Wikipathways	2(8.0%)	4.22×10^{-02}
$WP416$	histidine, purine & pyrimidine	Wikipathways	3(4.6%)	4.22×10^{-02}
$WP390$	serine-isocitrate lyase	Wikipathways	2(8.5%)	4.22×10^{-02}

regulated by disrupting the protein stability of the proteins as a post-translational regulation (Shin et al., 2013; Šmardová et al., 2005). Hence, over-representation of $p53$ related pathways with the outlier proteins re-confirmed our initial hypothesis with another biological explanation (*i.e.* protein degradation process by post-translational regulation enhance $p53$ regulation).

Table 4.4 shows the enrichment of functional annotations among the common outlier genes. All combinations provide high confidence levels for detecting post-translationally regulated proteins ($p < 0.05$). We observed that, ORR (Model 1) and QR (Model 2) models have the highest number of common genes and these are found in the upper region ($P < \hat{P}$) of the regression plot which, as noted earlier, is our region of interest for this problem. In the following section we further discuss the gene enrichment of the common genes by taking two models at a time and finally all the models together.

Table 4.4: Finer level PTM annotation check for three outlier sets. 1000 random trials were used in each case.

Outlier Sets	No of Common Genes	Annotation Check
Model 2 and Model 1	20	$p \leq 2.61 \times 10^{-05}$
Model 3 and Model 1	23	$p \leq 1.74 \times 10^{-08}$
Model 2 and Model 3	32	$p \leq 4.94 \times 10^{-11}$
All three outlier sets	17	$p \leq 1.04 \times 10^{-08}$

(i) Linear Regression (Model 0) and Outlier Rejecting Regression (Model 1) Outliers

GO enrichment analysis showed that 20 common genes over-represent in catalytic activities and metabolic biological processes. Six main pathways were detected

by the PANTHER classification system including $p53$ pathway and $p53$ pathway feedback loop. We also subjected the non-common genes between these two sets into GO and pathway analysis. Those were enriched with 15 ribosomal properties and four pathways such as; valine biosynthesis, isoleucine biosynthesis, DNA replication and glycosis.

(ii) Linear Regression (Model 0) and Quantile Regression (Model 2) Outliers

Figure 4.7 shows 23 proteins that are common to linear regression and quantile regression models. Thirteen proteins have catalytic molecular functions and 14 proteins are enriched with metabolic GO biological process. Similar to the previous comparison (linear regression and ORR model outliers), these consensus genes also showed $p53$ pathway and $p53$ feedback loop as dominant pathways. Further, non-common genes in linear regression showed ribosomal properties.

(iii) Outlier Rejecting Regression (Model 1) and Quantile Regression (Model 2) Outliers

Thirty-two genes were common between these two outlier sets. Catalytic molecular functions were found under GO analysis and $p53$ pathway, $p53$ pathways feedback loop and DNA replication were detected as the main pathways in pathway analysis.

(iv) All Three Outlier Sets (Model 0, 1 and 2)

Figure 4.9 shows the GO enrichment and pathway analysis results of the 17 genes as pie charts. With 17 common genes to all three outlier sets, we observed that several biological process and molecular function GO terms. Note that $p53$ pathway and $p53$ pathway feedback loop are the only two pathways we found with these common genes. This confirmed that the duo combination results are accurate (*i.e.* all duo combinations comprised with the two $p53$ pathways).

4.4.2.2 Protein-Protein Interaction Networks

Detailed physical protein-protein interaction network was generated BioGRID database (Stark et al., 2006) using union set of outliers (Figure 4.10). We observed

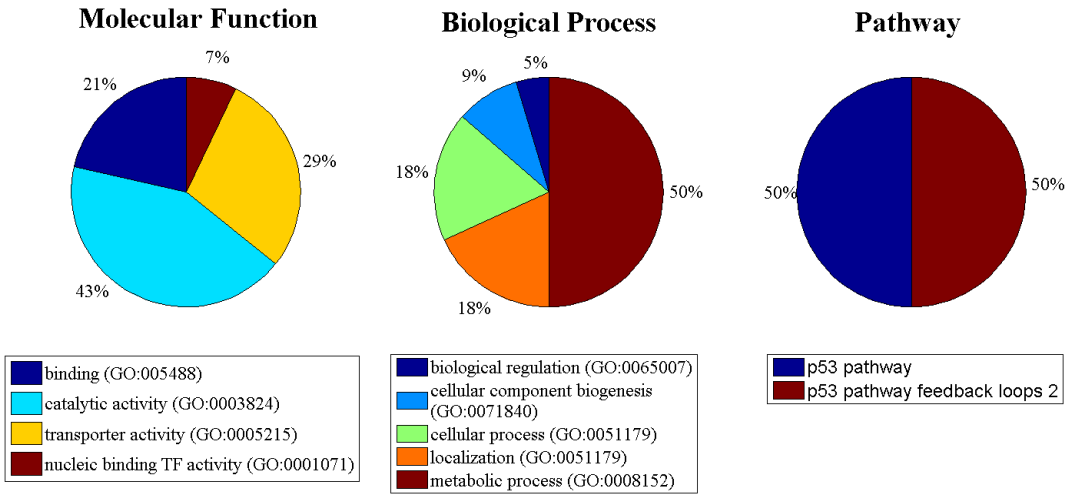


Figure 4.9: GO and Pathway anaylsis for the 17 gene common to all three outlier sets

two main clusters among the physical interactions. Large cluster consists of ribosomal subunits such as RPS13, RRP2B, RPS9B, RPS14A, RPS16B, RPL17A and RPL17B, mostly detected by Model 0. As we explained earlier, ribosomal proteins are active in translational and post-translational regulations (Carroll et al., 2008; Warner, 1999). The small cluster with hub PHO88, have proteins that are involved in phosphate ion transportation and protein maturation which occur due to phosphorylation post-translational regulation (Burnett and Kennedy, 1954).

Additionally, we also employed GeneMANIA web tool (Warde-Farley et al., 2010) to uncover further biological relationships using several other protein-protein interaction networks. Figure 4.11 shows the co-expression, genetic, predicted and physical networks obtained by the GeneMANIA web tool. We observed that co-expression network gives the largest coverage among the outlier proteins. Hence, we employed random samples with size 92 (same as union outlier set) and generated the interaction networks to observe the significance of the co-expression network generated using our outlier proteins. However, all the random samples showed a high coverage with the co-expression network and Figure 4.12 shows two examples of co-expression networks produced using random samples. Therefore, co-expression results of the union outlier set is not significant with respect to random samples. The reason for this observation is, GeneMANIA web tool employs transcript level gene expression data but not proteomic measurements where the post-translation regulation happen. Thus, the co-expression network is not suitable to measure significance of proteomic level properties such as post-translation regulation. We also observed that both BioGRID and GeneMANIA

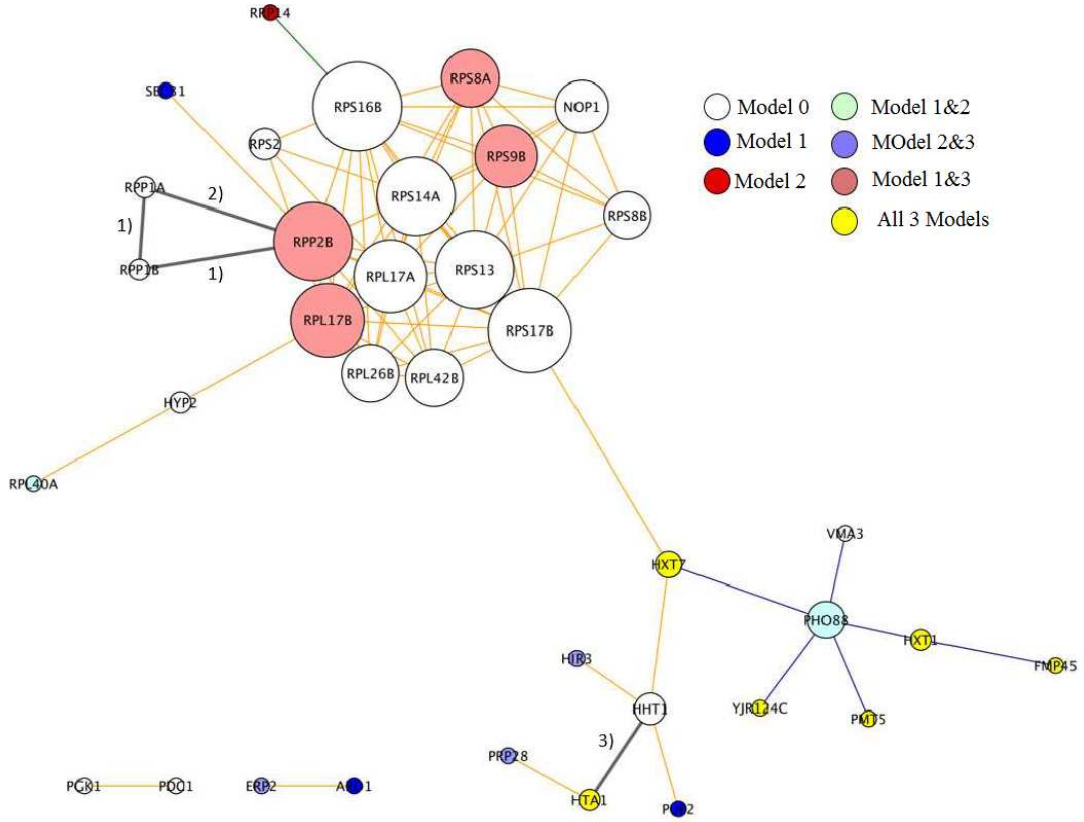


Figure 4.10: BioGRID physical interaction network constructed using union outlier set (N=92): node size and colour represent the number of interactions made by a particular protein and the regression model used to detect outlier protein respectively. Edge colour demonstrates experimental setup used to define the interactions and those are: yellow-affinity capture- MS, green two hybrid, blue-PCA (protein fragment complementation assay) and gray for the interactions defined by two or more experimental settings. Edge number representation as follows: 1) affinity capture by western, two-hybrid; 2) affinity capture by MS, affinity capture by western, two-hybrid, reconstituted complex; 3) affinity capture by MS, affinity capture by western, reconstituted complex respectively and co-purification.

tools produced similar structure physical networks, where ribosomal proteins tend to cluster together.

4.5 Discussion

All three regression models showed high confidence levels for detecting post-translationally regulated proteins as outliers. However, our main hypothesis relies on the protein abundance difference between measured and the predicted proteins *i.e.* PTR effects on the protein stability thus, proteins with PTR start to degrade faster. In

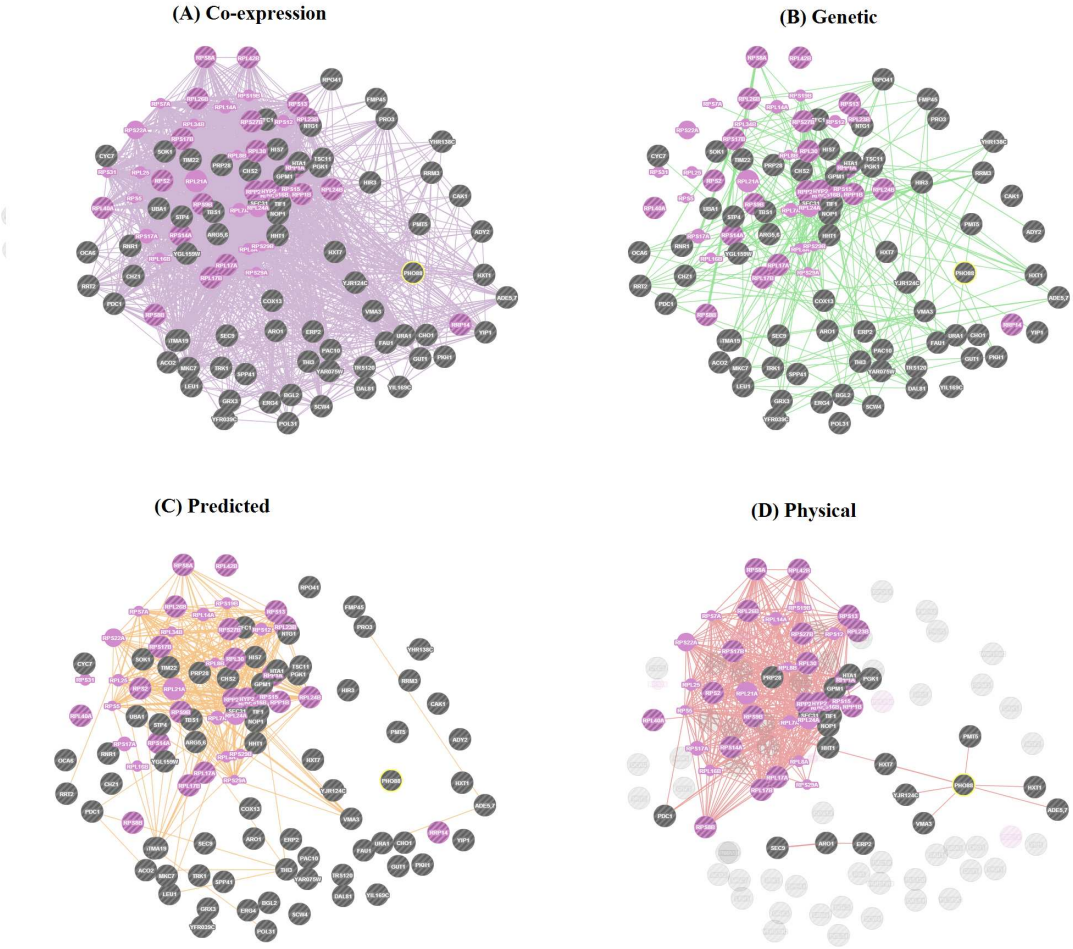


Figure 4.11: Protein-protein interaction networks obtained by GeneMANIA web tool for union outlier set (N=92): nodes and edges represent the outlier proteins and their interactions respectively. Pink colour nodes represents ribosomal proteins (A) Co-expression network with purple edges which gave the highest coverage of interactions (58.33%). (B) Green colour edges represents genetic interactions with 18.58% of coverage. (C) Predicted interaction network represents predicted relationships with other organisms such as rat, worm, human (11.06% of coverage) and finally (D) shows the physical interaction network which is similar to BioGRID physical interaction network.

fact, proteins with predicted protein abundance ($\hat{P}$) greater than measured protein abundance (P) are more likely to be post-translationally regulated. Linear regression (Model 0) and ORR (Model 1) select outliers with both negative and positive losses ($P < \hat{P}$ and $P > \hat{P}$). Thus, these two techniques have a disadvantage of adding false positive to the results. However, linear regression model from previous chapter selects only two proteins from the lower region and those two were not annotated as post-translationally regulated. Therefore, linear regression has less impact on detecting $P > \hat{P}$ proteins as outliers on the final outcome. In

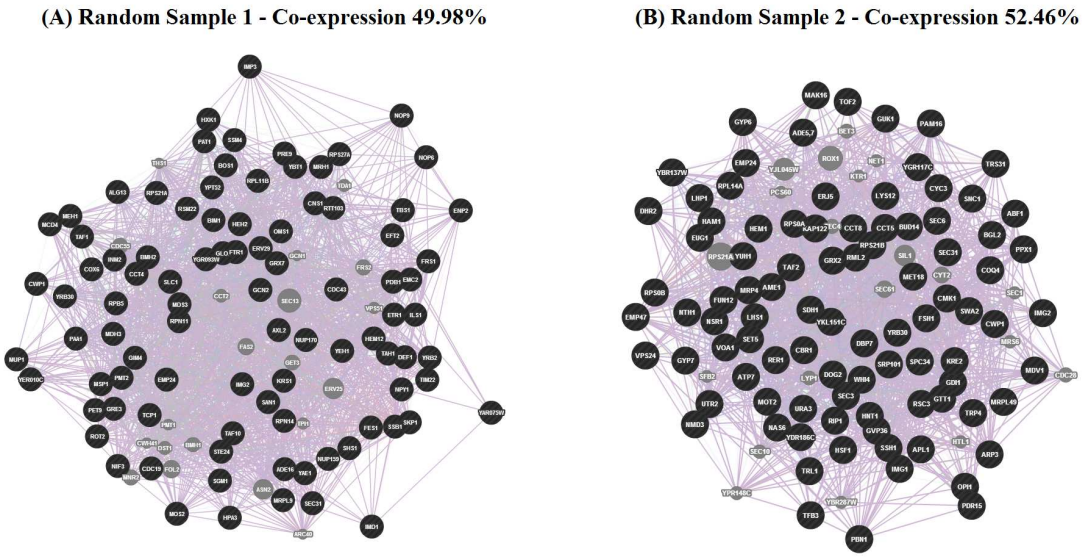


Figure 4.12: Co-expression networks obtained using random gene samples with sample size 92 (similar to the union outlier set): **(A)** and **(B)** are two examples of co-expression networks generated using random samples. All the random samples gave high co-expression network coverage which is similar to our union outlier set. Therefore, outlier co-expression network results are not significant. GeneMANIA uses transcript level gene expressions to generate co-expression networks so that it is not capable of detecting proteome level relationships as post-translational regulation.

contrast, ORR model selects 15 proteins from the lower region of the regression line as outliers and seven of them did not show post-translational regulation. This caused low confidence level of detecting post-translationally regulated proteins as outliers with respect to the other two models. In order to overcome this problem, QR model (Model 2) was introduced. This model allows us to use asymmetric error and select outliers either with negative or positive losses. Hence, according to our problem domain we selected outliers only from with negative losses (upper region in a regression plot) and tested the over-representation of post-translationally regulated proteins as outliers. Interestingly, these outliers showed the highest confidence level of detecting post-translationally regulated proteins. Thus, QR model is the best technique to detect post-translationally regulated proteins as outliers using a global regression approach. Gene ontology analysis on outlier proteins also provided evidence for post-translational regulation and pathway analysis further supported our hypothesis by over-representing $p53$ pathways which are directly associated with the protein degradation process of post-translational regulation.

4.6 Summary

In this chapter, we presented two novel outlier detection techniques at the transcriptome-proteome interface and we compare our results with the linear regression approach used in previous chapter (Chapter 3). ORR model selects a certain fraction of the data as outliers while performing the global regression and QR model minimizes the error to be one sided to detect outliers only with negative losses. All outlier sets show high confidence levels in functional annotation checks providing evidence for post-translational regulation. In fact, quantile regression (Model 2) gave the highest confidence level suggesting it as the best method to detect post-translationally regulated proteins from a global regression due to the asymmetric conditional loss property. Thus, these data-driven approaches can be used to cut down the experimental work space to detect proteins with post-translational regulation.

Chapter 5

Numerical Precision in Transcriptome-based Inference & Coherence with Protein Prediction

High-throughput microarray and RNA-Seq measurements are widely used in analysis of transcriptome-proteome interface and these are measured to very high numerical precision (*i.e.* mRNA concentrations are measured to several decimal places *e.g.* 3.698567190672298). However, these high numerical precision occur due to mRNA amplification process in these high-throughput measuring techniques to obtain more accurate relative abundances (Figure 5.1) (Nygaard et al., 2003; Ozsolak and Milos, 2011). Therefore, we investigate whether these high numerical precision carry more additional information with respect to low precision binarized data. In fact, in this chapter, we compare machine learning inferences between these two transcriptomic measurements at the low (binary measurement as 0 or 1) and high (continuous measurement as 5.786861236001966) numerical precision. Firstly, we consider transcriptomic layer and perform classification, clustering, time series analysis and cross platform analysis using only mRNA concentrations. Secondly, we develop a protein abundance predictor based on the five feature regression model in Chapter 3 to investigate protein prediction capabilities between these two techniques. Here, we also explore the variability of the prediction accuracies with respect to continuous and binary mRNA concentrations. Further, we use quantile regression outlier detection technique which is considered as the best technique among the techniques we used in Chapter 4 to detect post-translationally regulated proteins and compare PTR detection capability

between microarray and RNA-Seq techniques under low and high numerical precision. This is an extended experimental setting of our PTR detection framework under different transcriptomic inputs.

5.1 Numerical Precision in Microarray and RNA-Seq Measurements

Microarray and RNA-Seq are the most commonly used high-throughput transcriptome measuring techniques and Section 2.3.1 in Literature Review chapter gives an detailed description of these two measuring technique. However, both these techniques amplify the number of mRNA strands found in the actual cell to high number of copies to reduce the measuring errors and improve the accuracy of the relative abundance (Nygaard et al., 2003; Ozsolak and Milos, 2011). Figure 5.1 shows the process of amplifying actual mRNA strands per cell into high number of copies. Therefore, it is important to investigate whether these high numerical precision obtained by mRNA amplification carry additional information with respect to a binary representation of mRNA concentration providing the gene status as switch on or off. We believe that comparing of machine learning inferences between continuous and binarized data will allow us to uncover the mystery behind the numerical precision. In fact, if the machine learning performance drops by converting continuous data to binary, we can justify that the high numerical precision carries more additional information rather than gene on/off signal.

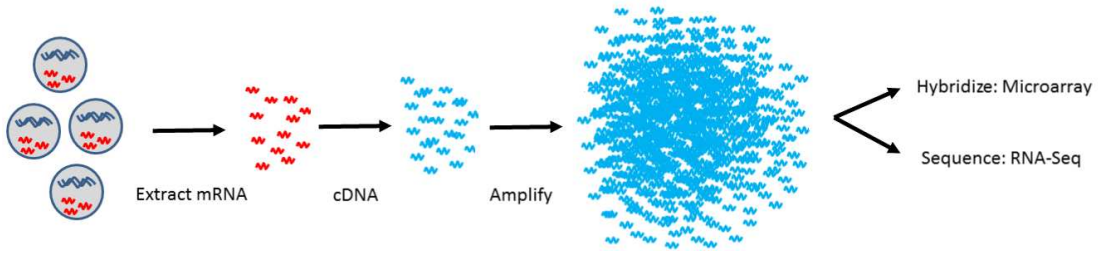


Figure 5.1: Amplification of mRNA copy numbers in transcriptome measuring techniques: both microarray and RNA-Seq measuring techniques amplify the number of mRNA copies found in a cell to a high number of copies to improve the accuracy of relative abundance.

5.2 Transcriptomic Inferences

In this section, we only use transcriptomic data (mRNA abundance) and perform different machine learning inferences using continuous (high numerical precision)

and binary (low numerical precision data) measurements to observe the information loss under different binarization techniques. As we discussed in the Literature Review (Chapter 2), RNA-Seq technique is more powerful to detect qualitative attributes such as identifying novel exon junctions, RNA-editing events and new isoforms with respect to microarray technology (Wang et al., 2009; Mortazavi et al., 2008; Fu et al., 2009; Marioni et al., 2008). Tuna and Niranjan, 2009's study showed that microarray numerical precision carries no additional information and binary representation of the data is sufficient to classify cancer and normal patients. Shmulevich and Zhang, 2002 had a similar perspective on microarray measurements and showed that binarized mRNA expression levels can solve classification problems accurately by considering the Hamming distance between signatures as distance metric. Friedman et al., 2000 used a probabilistic Bayesian network to observe the gene interactions. They employed a multinomial model with quantized gene expressions under three levels. Those are under-expressed (-1), normal (0) and over-expressed (1). This model was able to detect casual relationships and interactions among genes other than the correlation values. Thus, as a novel approach we explore the behaviour of the RNA-Seq measurements in binary world and explore its impact on classification and clustering accuracies. Also we compare RNA-Seq inference capability with microarray data to analyze the quantitative performances between these two techniques. In fact, our ultimate goal is to observe numerical precision inferences of these two high-throughput transcriptome measuring techniques.

5.2.1 Feature Selection

Our first approach was to select the most dominant genes (best features) for classification, clustering and other analysis purposes. This is an important process to overcome over-fitting problem in machine learning techniques. Thus, we used the Fisher score feature selection method similar to Golub et al., 1999b's study to obtain the most dominant features. In fact, genes were ranked according to a univariate metric and selected the highest ranking genes as the best features. The scoring reflects the discriminative power of each feature. Fisher score for gene g is given below;

$$\boldsymbol{F} = \text{abs}\left|\frac{\mu(1)_g - \mu(2)_g}{\sigma(1)_g + \sigma(2)_g}\right| \forall \in g \tag{5.1}$$

where $\mu(i)_g$ and $\sigma(i)_g$ are the mean and standard deviation of gene g in class i and $i \in \{1, 2\}$. In all cases we used 200 features as the maximum no of features.

5.2.2 Binarization Techniques

After selecting the best features (genes), four different types of binarization techniques were used to convert high numerical precision continuous data to low numerical precision binary data. Those are,

1. **Global Mean Binarization (B1)** - Obtained the mean of the total data set and gene expression below the mean considered as 0 and the rest as 1. (if $X_i \geq$ mean, then X_i =1 or else X_i =0 ; *i.e* X_i denotes a continuous mRNA abundance).

2. **Gene by Gene Mean Binarization (B2)** - Figure 5.7 shows distribution of gene expression of the best gene selected from the Fisher Score technique for bladder cancer using both RNA-Seq and microarray techniques. This figure illustrates that a single gene provides sufficient information to do a better classification. Thus, we obtained the best 200 genes from the feature selection and binarization carried out for each gene separately. Mean expression value was computed for each gene and samples below mean were converted to 0 and the rest converted to 1.

3. **Global GMM Threshold Binarization (B3)** - Tuna and Niranjan, 2009's study used Gaussian Mixture Model (GMM) technique to obtain the threshold value to binarize data. This approach was motivated by Zhou et al., 2003's study where two component GMM was employed to model highly expressed and not expressed genes in a sample. One component corresponds to highly expressed genes and the other for not expressed genes. Figure 5.2 shows an example of fitting a two GMM fitting model with to a microarray gene. Data distribution of a the best genes after the feature selection (Figure 5.7) also showed that the two component mixture model is more suitable to classify cancer and normal samples. Thus, the threshold value was obtained by fitting a two centres GMM, similar to Zhou et al., 2003. Mean (μ) and standard deviation (σ) values were used to obtain the threshold value as given below;

$$\boldsymbol{Th} = \frac{\mu(1) + \sigma(1) + \mu(2) - \sigma(2)}{2} \tag{5.2}$$

(Zhou et al., 2003) where $\mu(1)$ and $\sigma(1)$ belong to class 1 (cancer) and $\mu(2)$ and $\sigma(2)$ belong to class 2 (normal). In Gobal GMM method we use total data set to fit the mixture model. Gene expressions less than the threshold

value considered as 0 and the rest considered as 1. `Netlab` package in a `MATLAB` environment was employed to fit data into Gaussian Mixture Models.

4. **Gene by Gene GMM Threshold Binarization (B4)** - We employed the same two center GMM model gene by gene for the best 200 genes (*i.e* each gene obtained a threshold value for the binarization).

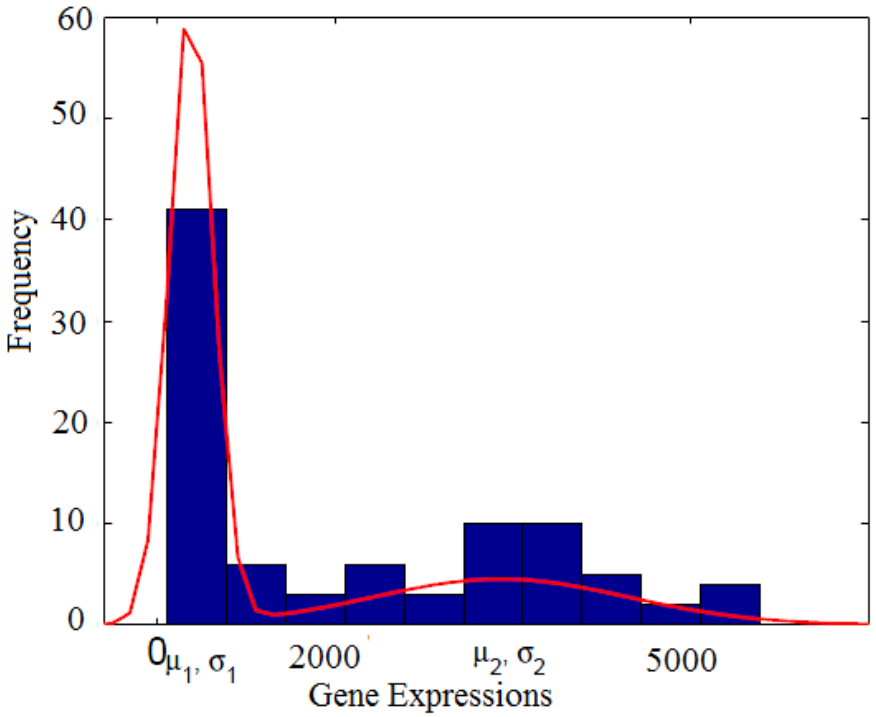


Figure 5.2: Example of fitting a two component GMM model with a microarray gene

5.2.3 Classification and Clustering

Continuous and binary mRNA measurements from microarray and RNA-Seq techniques were compared using classification and clustering machine learning approaches. K-Nearest neighbour (KNN) algorithm (Dudani, 1976) and Support Vector Machine (SVM) (Vapnik, 1998) linear classifiers were employed for classification purposes and K-Means clustering (MacQueen et al., 1967) and Spectral clustering (Shi and Malik, 2000) techniques were used to obtain clustering results. With both clustering and classification methods, area under the curve (AUC) was used to obtain the accuracy in 200 bootstrap iterations. AUC is the best technique to measure accuracies when the number of samples are not balanced between the two or more classes Chawla (2005). Hierarchical clustering was also employed to

generate dendrograms of cluster classes to compare clustering accuracies (Eisen et al., 1998).

5.2.3.1 Data sets

Several data sets were employed to compare microarray and RNA-Seq measurements at the transcriptome layer. Table 5.1 shows microarray and RNA-Seq the data sets used for classification and clustering experiments. All RNA-Seq data sets were downloaded from TCGA Cancer Genome Atlas public data portal under National Cancer Institute (http://cancergenome.nih.gov/dataportal/) as available on Feb, 2014 [TCGA].

Table 5.1: Microarray and RNA-Seq data sets used for anlysis purposes

Measuring Technique	Cancer Type	Reference	No. of Genes	Samples (Can/Nor)
Microarray	Lung adenocarcinoma	Bhattacharjee et al., 2001	12600	203 (186/17)
	Ovarian	Bonome et al., 2008	22283	196 (186/10)
	Soft-tissue sarcoma	Barretina et al., 2010	22283	158 (125/37)
	Head & Neck squamous carcinoma	Estilo et al., 2009	12625	59 (31/28)
	Colon adenocarcinoma	Alon et al., 1999	2000	62 (40/22)
	Bladder	Dyrskjøt et al., 2004	2000	60 (51/09)
	Stomach	D'Errico et al., 2009	54675	69 (38/28)
RNA-Seq	Breast invasive carcinoma	TCGA (UNCC)	20532	195 (127/11)
	Bladder urothelial carcinoma	TCGA (UNCC)	20532	67 (56/11)
	Lung adenocarcinoma cancer	TCGA (UNCC)	20532	162 (125/37)
	Stomach adenocarcinoma	TCGA (MSGSC)	22346	271 (238/33)
	Liver hepatocellular carcinoma	TCGA (UNCC)	20532	25 (16/9)
	Head and Neck squamous cell carcinoma	TCGA (UNCC)	20532	294 (263/31)

UNCC stands for University of North Carolina Cancer Center and MSGSC stands for Micheal Smith Genome Science Centre.

5.2.3.2 Results

Figure 5.3 shows the continuous and binarized classification results using SVM classifier for RNA-Seq and microarray stomach cancer transcriptome data. Here, we compare the continuous classification results with all four types of binarization techniques *i.e.* (B1) Global Mean Binarization, (B2) Gene by Gene Mean Binarization, (B3) Global GMM Threshold Binarization and (B4) Gene by Gene GMM Threshold Binarization in Section 5.2.2. We observed that both measurements (microarray and RNA-Seq) perform similarly in continuous and binary classification and gene based binarization techniques (B2 and B4) perform better than the

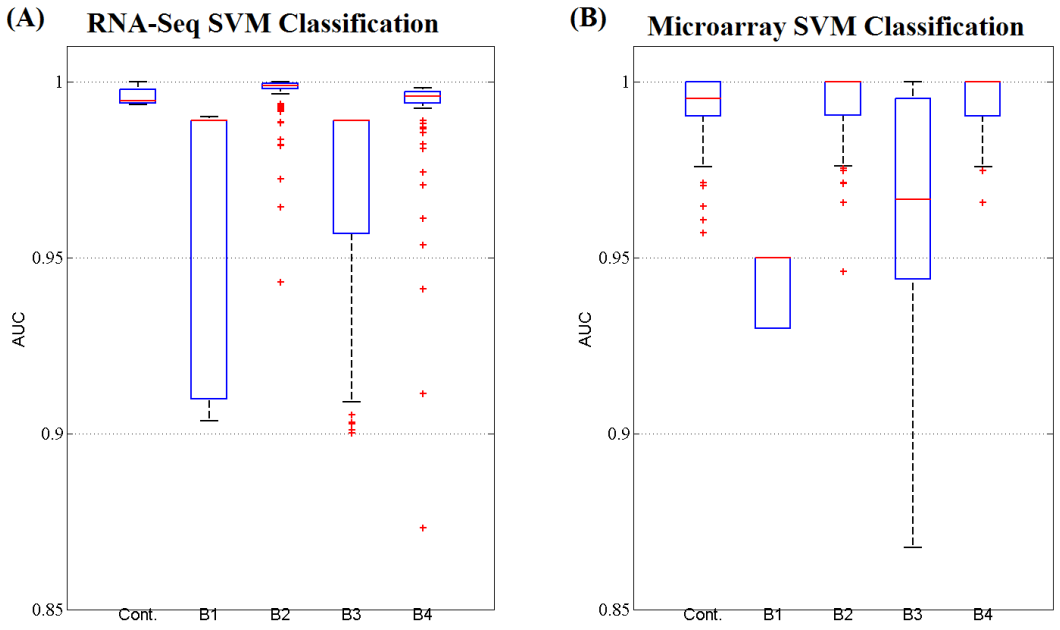


Figure 5.3: SVM classification performance on stomach cancer using continuous and binarized data (using 100 top genes in each experiment). (A) shows the RNA-Seq classification performance and (B) represents the microarray classification accuracies. 200 bootstrap cross validation trails were carried out to obatin the accuracy variation of each experiment and the four types of binarization techniques are: (B1) Global Mean Binarization, (B2) Gene by Gene Mean Binarization, (B3) Global GMM Threshold Binarization and (B4) Gene by Gene GMM Threshold Binarization

global mean techniques (B1 and B3). Therefore, precision reduction did not influence the RNA-Seq and microarray classification performance. We also performed a supervised machine learning inferences using classification and clustering techniques and obtained AUC accuracies. Table 5.2 shows the classification and clustering performance of RNA-Seq measurements and Table 5.3 shows the microarray analysis results. In all cases, both microarray and RNA-Seq measurements perform similarly, confirming that binarization does not affect either classification or clustering performances. Thus, high precision of transcriptome measurements (continuous data) obtained by RNA-Seq and microarray techniques do not carry any additional information, where the binary data is sufficient to perform a better classification and clustering analysis. Here, we also observed that B4 (gene by gene GMM threshold) binarization technique performs better compared to other three binarization techniques. Therefore with the latter experiments, we will mainly consider B4 binarization technique to convert continuous mRNA abundance data to binary representation. Hierarchical clustering was also carried out with continuous and binarized (using B4) data from both RNA-Seq and microarray data sets. Figure 5.4 and Figure 5.5 show hierarchical clustering results for bladder cancer

using RNA-seq and microarray data respectively. See Appendix E for hierarchical clustering results of all other cancer data sets. These results also confirmed binarization does not effect on the clustering analysis of RNA-Seq and microarray transcriptome measurements.

Table 5.2: Classification and clustering accuracies of RNA-Seq cancer data. Accuracies were calculated for 200 bootstrap sampling using 200 best genes in each experiment. Cont. stands for Continuous Data and B1, B2, B3, B4 stand for global mean, gene by gene mean, global GMM and gene by gene GMM binarization techniques respectively. (+/-) means (no of cancer patients/no of normal patients)

Cancer	Data Type	KNN AUC	SVM AUC	K-Means AUC	Spectral AUC
Breast	Cont. Data	0.99 ± 0.01	1 ± 0	1 ± 0	1 ± 0
	B1 Data	0.87 ± 0.07	0.95 ± 0.02	0.95 ± 0.02	0.94 ± 0.03
	B2 Data	0.97 ± 0.01	1 ± 0	1 ± 0	0.99 ± 0.01
(+127/-68)	B3 Data	0.98 ± 0.01	$0.99 \pm 4.3 \times 10^{-4}$	0.96 ± 0.03	0.94 ± 0.03
	B4 Data	0.99 ± 0.01	$0.99 \pm 2.8 \times 10^{-4}$	$0.99 \pm 5.6 \times 10^{-16}$	0.99 ± 0.01
Bladder	Cont. Data	0.98 ± 0.01	1 ± 0	1 ± 0	$0.99 \pm 1.3 \times 10^{-15}$
	B1 Data	0.93 ± 0.07	0.98 ± 0.02	0.98 ± 0.02	0.99 ± 0.01
	B2 Data	0.96 ± 0.02	0.99 ± 0.01	0.99 ± 0.01	0.94 ± 0.06
(+56/-11)	B3 Data	0.99 ± 0.01	1 ± 0	1 ± 0	0.98 ± 0.01
	B4 Data	1 ± 0	1 ± 0	1 ± 0	0.89 ± 0.11
Lung	Cont. Data	0.99 ± 0.01	1 ± 0	1 ± 0	0.92 ± 0.09
	B1 Data	0.85 ± 0.14	0.86 ± 0.11	0.86 ± 0.11	0.96 ± 0.04
	B2 Data	0.99 ± 0.01	1 ± 0	1 ± 0	0.98 ± 0.11
(+125/-37)	B3 Data	0.99 ± 0.01	1 ± 0	0.85 ± 0.05	0.91 ± 0.04
	B4 Data	1 ± 0	1 ± 0	0.87 ± 0.22	1 ± 0
Stomach	Cont. Data	0.95 ± 0.03	0.99 ± 0.01	0.99 ± 0.01	0.96 ± 0.03
	B1 Data	0.85 ± 0.06	0.97 ± 0.03	0.95 ± 0.02	0.91 ± 0.01
	B2 Data	0.95 ± 0.03	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.11
(+238/-33)	B3 Data	0.94 ± 0.02	0.98 ± 0.01	0.90 ± 0.09	0.86 ± 0.01
	B4 Data	0.95 ± 0.04	0.99 ± 0.01	$0.99 \pm 7.0 \times 10^{-4}$	0.93 ± 0.09
Liver	Cont. Data	0.99 ± 0.01	1 ± 0	1 ± 0	1 ± 0
	B1 Data	0.98 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	1 ± 0
	B2 Data	1 ± 0	1 ± 0	1 ± 0	1 ± 0
(+16/-9)	B3 Data	0.95 ± 0.04	0.99 ± 0.01	1 ± 0	1 ± 0
	B4 Data	1 ± 0	1 ± 0	1 ± 0	1 ± 0
Head & Neck	Cont. Data	1 ± 0	1 ± 0	1 ± 0	0.91 ± 0.09
	B1 Data	0.91 ± 0.05	0.98 ± 0.02	0.98 ± 0.02	1 ± 0
	B2 Data	0.98 ± 0.01	$1 \pm 5.7 \times 10^{-17}$	$1 \pm 1.57 \times 10^{-17}$	1 ± 0
(+263/-31)	B3 Data	0.98 ± 0.01	0.99 ± 0.01	0.92 ± 0.07	0.88 ± 0.12
	B4 Data	1 ± 0	1 ± 0	1 ± 0	1 ± 0

We then extended our experiments by changing the number of genes used to generate the SVM classifier. Figure 5.6 shows the classification accuracies of stomach cancer with respect to the number of genes (features) used to generate the classifier. Gene expressions were binarized using B4 binarization technique. Both

Table 5.3: Microarray measurements classification and clustering accuracies for different cancer types. Bootstrap sampling with 200 trails were used with 200 best genes in each experiment. Cont. stands for continuous data and B1, B2, B3, B4 stand for global mean, gene by gene mean, global GMM and gene by gene GMM binarization techniques respectively.

Cancer	Data Type	KNN AUC	SVM AUC	K-Means AUC	Spectral AUC
Lung	Cont. Data	0.96 ± 0.04	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01
	B1 Data	0.93 ± 0.06	0.98 ± 0.02	0.98 ± 0.02	0.99 ± 0.01
(+186/-17)	B2 Data	0.94 ± 0.05	0.99 ± 0.01	0.99 ± 0.01	0.96 ± 0.03
	B3 Data	0.93 ± 0.06	0.99 ± 0.01	0.98 ± 0.01	0.96 ± 0.03
	B4 Data	0.92 ± 0.07	0.99 ± 0.01	$0.99 \pm 2.88 \times 10^{-4}$	0.97 ± 0.03
Ovary	Cont. Data	1 ± 0	1 ± 0	1 ± 0	1 ± 0
	B1 Data	1 ± 0	1 ± 0	1 ± 0	0.96 ± 0.04
(+186/-10)	B2 Data	0.99 ± 0.01	1 ± 0	1 ± 0	1 ± 0
	B3 Data	1 ± 0	1 ± 0	0.96 ± 0.03	0.92 ± 0.07
	B4 Data	1 ± 0	1 ± 0	1 ± 0	0.96 ± 0.02
Soft Tissue	Cont. Data	0.99 ± 0.01	1 ± 0	1 ± 0	0.96 ± 0.03
	B1 Data	0.99 ± 0.01	1 ± 0	1 ± 0	0.92 ± 0.06
(+125/-37)	B2 Data	0.99 ± 0.01	1 ± 0	1 ± 0	0.96 ± 0.03
	B3 Data	0.99 ± 0.01	1 ± 0	0.94 ± 0.06	0.94 ± 0.03
	B4 Data	0.99 ± 0.01	1 ± 0	0.93 ± 0.07	0.95 ± 0.04
Head & Neck	Cont. Data	0.95 ± 0.02	0.99 ± 0.01	0.99 ± 0.01	$0.99 \pm 2.08 \times 10^{-15}$
	B1 Data	0.96 ± 0.02	0.99 ± 0.01	0.99 ± 0.01	$0.99 \pm 2.06 \times 10^{-15}$
(+31/-28)	B2 Data	0.96 ± 0.02	0.99 ± 0.01	0.99 ± 0.01	0.98 ± 0.01
	B3 Data	0.94 ± 0.03	0.99 ± 0.01	0.91 ± 0.07	0.98 ± 0.01
	B4 Data	0.96 ± 0.02	0.99 ± 0.01	$0.99 \pm 1.33 \times 10^{-15}$	0.98 ± 0.02
Colon	Cont. Data	0.82 ± 0.07	0.91 ± 0.04	0.91 ± 0.03	0.89 ± 0.02
	B1 Data	0.76 ± 0.07	0.87 ± 0.06	0.87 ± 0.07	0.88 ± 0.03
(+40/-22)	B2 Data	0.78 ± 0.07	0.84 ± 0.07	0.84 ± 0.08	0.83 ± 0.07
	B3 Data	0.78 ± 0.07	0.87 ± 0.06	0.80 ± 0.02	0.80 ± 0.01
	B4 Data	0.81 ± 0.07	0.86 ± 0.06	0.82 ± 0.11	0.89 ± 0.12
Bladder	Cont. Data	0.92 ± 0.07	0.99 ± 0.01	0.99 ± 0.01	0.97 ± 0.02
	B1 Data	0.94 ± 0.03	0.90 ± 0.02	0.96 ± 0.04	0.96 ± 0.02
(+51/-09)	B2 Data	0.95 ± 0.03	0.98 ± 0.02	0.98 ± 0.02	0.89 ± 0.01
	B3 Data	0.90 ± 0.10	0.96 ± 0.04	0.96 ± 0.01	0.96 ± 0.02
	B4 Data	0.90 ± 0.09	0.99 ± 0.01	0.91 ± 0.05	0.91 ± 0.06
Stomach	Cont. Data	0.99 ± 0.01	0.94 ± 0.04	0.98 ± 0.01	0.96 ± 0.02
	B1 Data	0.90 ± 0.07	0.96 ± 0.01	0.89 ± 0.05	0.90 ± 0.06
(+38/-28)	B2 Data	0.99 ± 0.01	0.95 ± 0.03	$0.99 \pm 4.25 \times 10^{-04}$	0.88 ± 0.12
	B3 Data	0.98 ± 0.01	0.97 ± 0.01	$0.0.94 \pm 0.04$	0.94 ± 0.01
	B4 Data	0.99 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	0.90 ± 0.05

RNA-Seq and microarray measurements produce similar results and the binarization did not affect the classification accuracies with reasonable number of features to perform the classification task. The top two genes from RNA-Seq and microarray data sets are **CENPO** and **TP53INP1** respectively. **CENPO** was ranked as 14th from the RNA-Seq top genes and **TP53INP1** as 12th from microarray

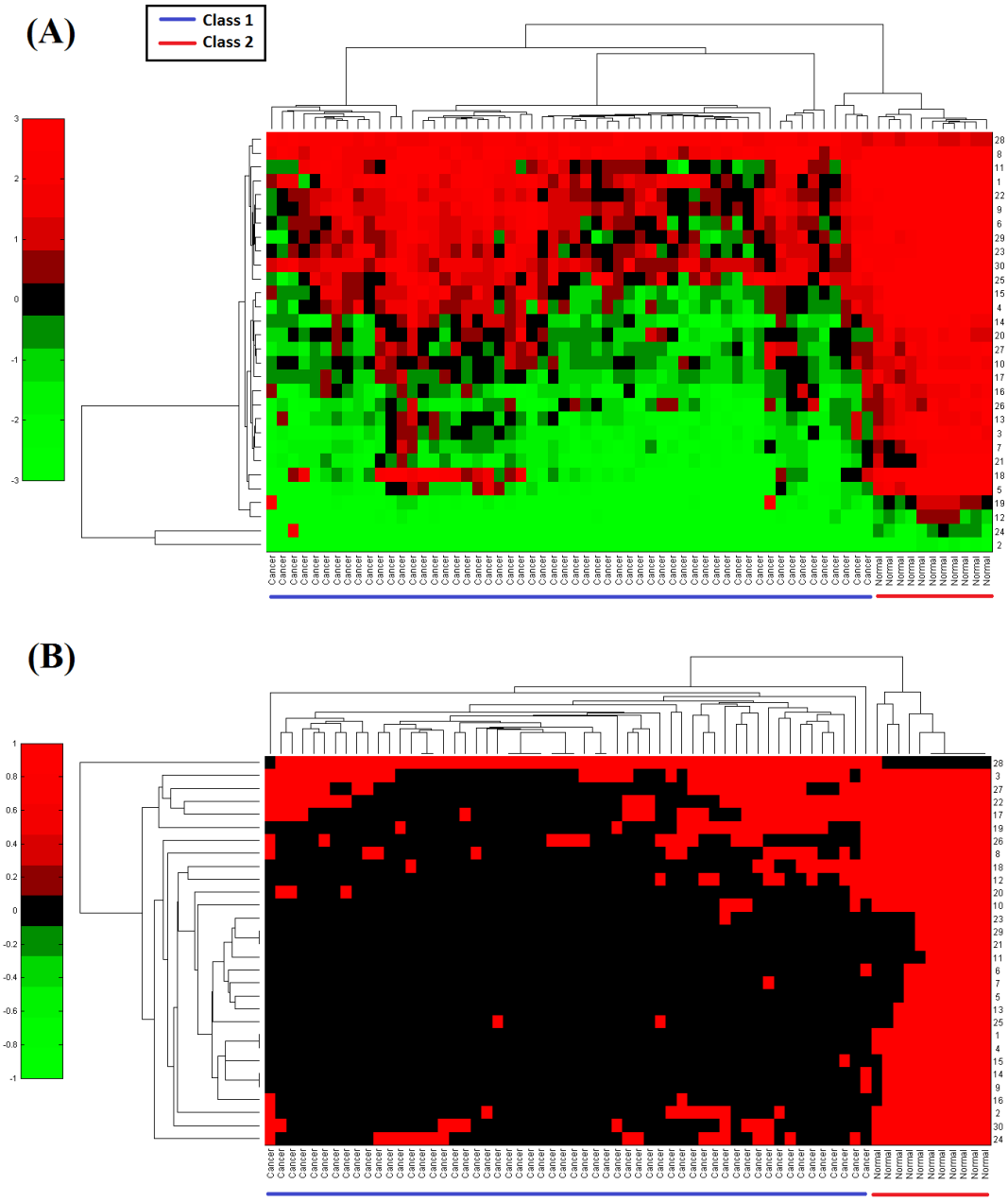


Figure 5.4: Hierarchical clustering of RNA-Seq bladder cancer data (using 30 top genes in each experiment). (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

gene list. Further, previous studies have shown that **CENPO** and **TP53INP1** are directly associates with stomach cancer (Jiang et al., 2006; Thiru et al., 2014).

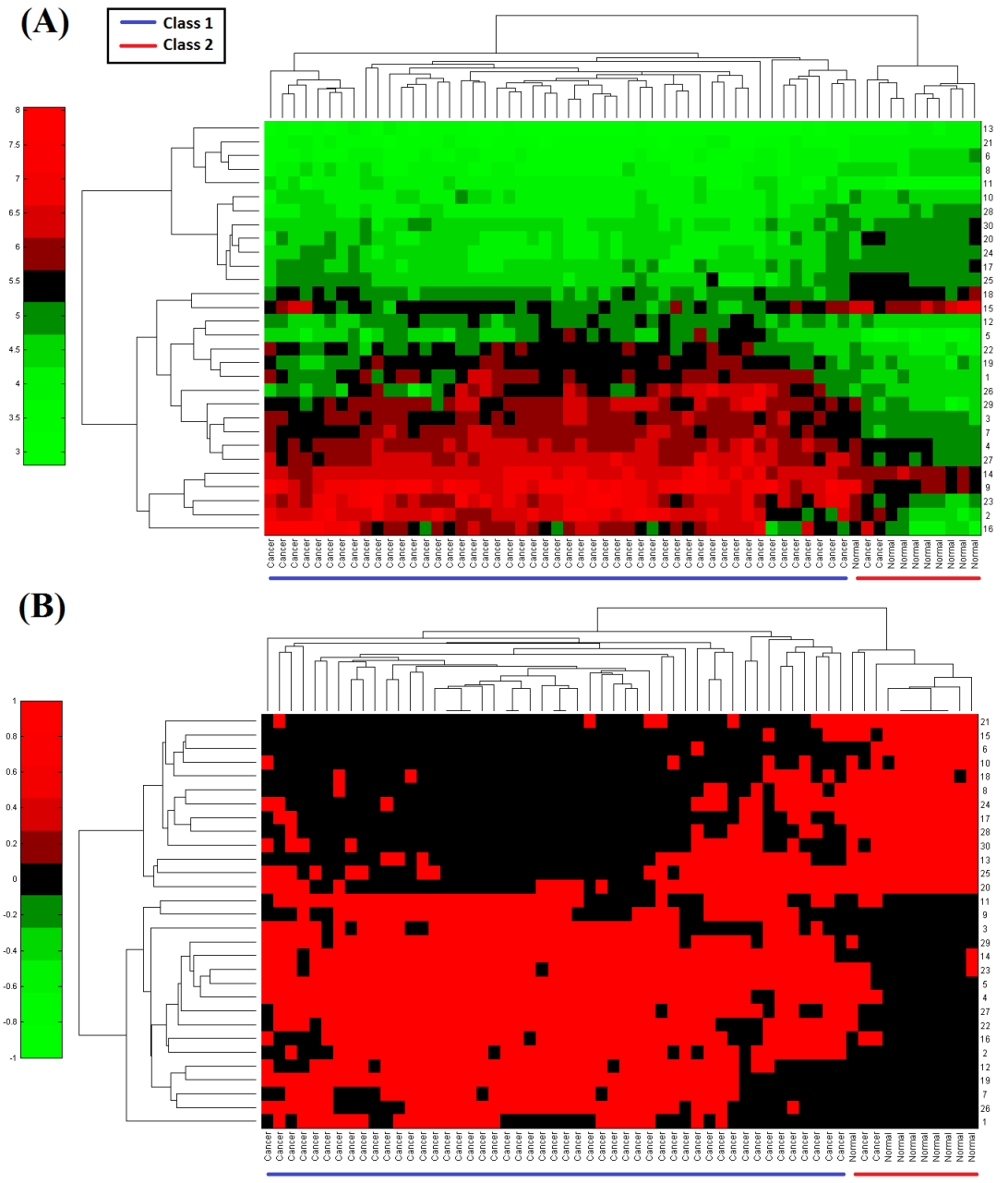


Figure 5.5: Hierarchical clustering of microarray bladder cancer data (using 30 top genes in each experiment). (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data generate same two cluster classes. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

Thus, these two genes are strong candidates to classify cancer and normal patients using gene expression data. Figure 5.7 represents that in both techniques (RNA-Seq and microarray) cancer and normal samples represent a gap/difference between gene expression levels. Thus, in this cancer type single gene is sufficient

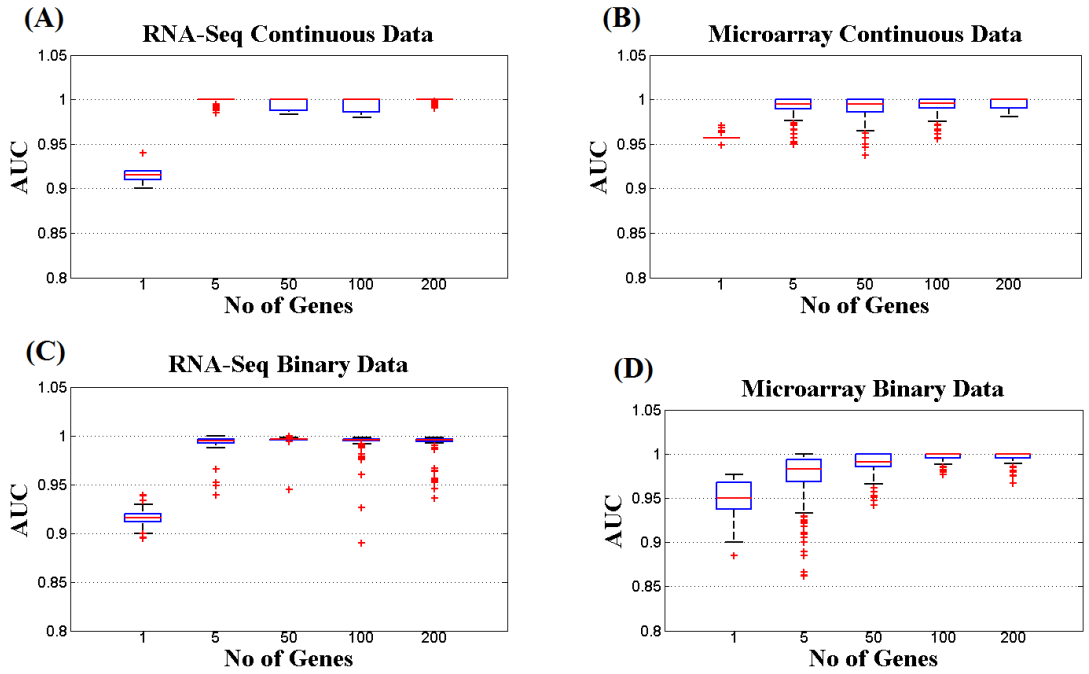


Figure 5.6: Variation of classification performance of stomach cancer with the number of best genes/features selected by the Fisher Score technique. (A) and (C) related to the RNA-Seq continuous and binarized data respectively and (B) and (D) for the microarray data. B4 (gene by gene GMM threshold based) binarization method was employed to convert continuous data into binary (using 200 top genes in each experiment).

to provide good classification accuracy. However, when you increase the number of genes the accuracy improves with both transcriptome data sets. Wang and Gotoh, 2009 also showed single gene can provide good classification accuracies in cancer studies.

5.2.4 Time Series Data Analysis

Tuna and Niranjan, 2010 used *Drosophila melanogaster*'s development cycle microarray measurements by Hooper et al., 2007 to show that both continuous and binary time series data have similar number of up-regulated and down-regulated genes along the time-course. This experiment expands the transcriptome inference by not just using cancer patient data but also by observing the changes along the development time course. However, Tuna and Niranjan, 2010 only compared the number of up/down regulated genes using microarray continuous and binary measurements. Here we perform a quantitative analysis by counting number of up/down regulated genes using RNA-Seq data of *Drosophila melanogaster*'s development cycle by Graveley et al., 2011 and also perform a qualitative analysis

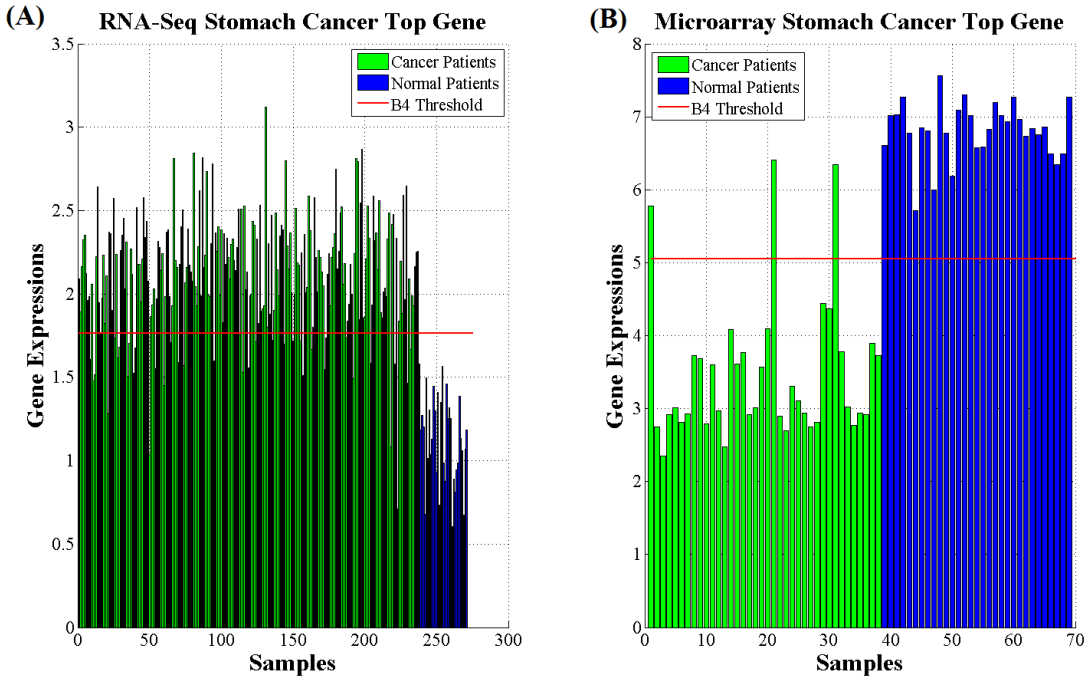


Figure 5.7: Gene expression levels of cancer and normal pateints of the best gene selected by the Fisher Score feature selection method. (A) represents the RNA-Seq measurements of gene **CENPO** and (B) shows the microarray measurements of **TP53INP1** gene obtained for the stomach cancer. Both of these genes represents a expression level change between cancer and normal patients

using gene ontology key words to observe the functional information lose in both microarray and RNA-Seq measurements with low precision binarized data along the development time course.

Following Hooper et al., 2007, we employed local convolution two steps function with RNA-Seq embryonic development data to obtain up and down regulated genes with respect to the time parameter. Following patterns were used as two step convolution functions;

- $-1-1-1-1+1+1+1+1$, to detect up regulations; and
- $+1+1+1+1-1-1-1-1$, to detect down regulations.

We selected the genes which exceeded 0.9 correlation coefficient (similar to Hooper et al., 2007) with the above patterns and compared the number of up and down regulated genes of continuous and binary measurements along the development time course.

Further, GO enrichment analysis was carried out using GOEAST web tool Zheng and Wang (2008) for the time points with highest number of up/down regulated

proteins to observe the biological enrichment between continuous and binary measurements. Here we ranked the GO terms based on the p-value and obtained the top 50 GO terms which gave highest confidence levels and extracted best terms related to the development life cycle to compare the confidence levels between continuous and binary data. Next, we used all the GO terms identified by the GOEAST web tool at the highest number of up/down regulated time points were used to generate gene ontology scatter plots. REVIGO web tool (Supek et al., 2011) was employed to generate these gene ontology scatter plots and compared the GO term clusters of continuous and binary data.

5.2.4.1 Results

Figure 5.8 and Figure 5.9 compare the number of up-regulated and down-regulated genes of continuous and binary data obtained by B2 (gene based mean) and B4 (gene based GMM) techniques respectively. In all cases, RNA-Seq transcriptome continuous and binary data have a similar number of up-regulated (or down-regulated) genes along the development time course. Thus, RNA-Seq binary data does not remove useful information throughout a development time course. In fact the lowest precision binary measurements carry same amount of information as high precision continuous measurements.

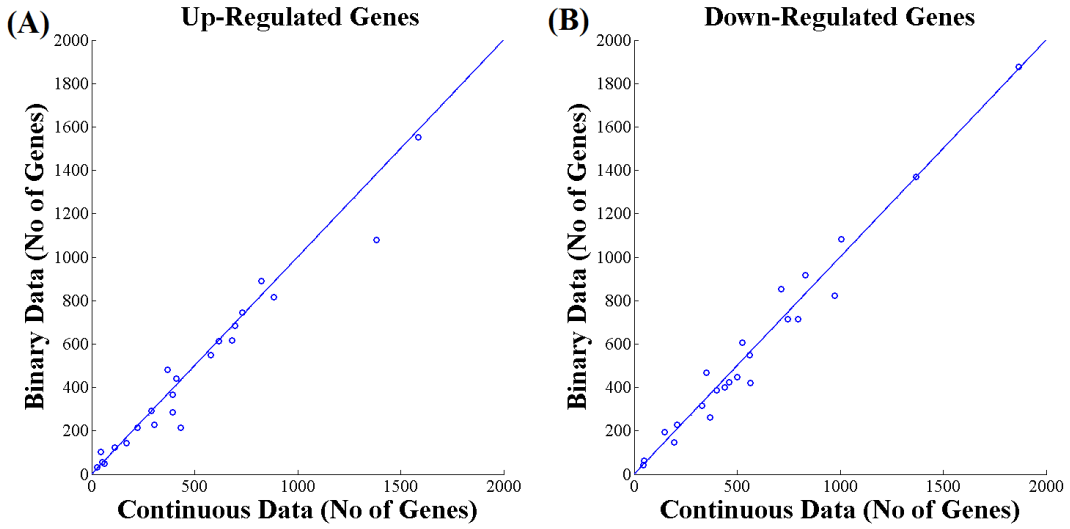


Figure 5.8: Comparison of significantly (A) up and (B) down regulated genes using continuous and B2 binarized RNA-Seq expression data over a developmental time series *i.e.* *Drosophila melanogaster*'s embryonic stage data (Graveley et al., 2011)

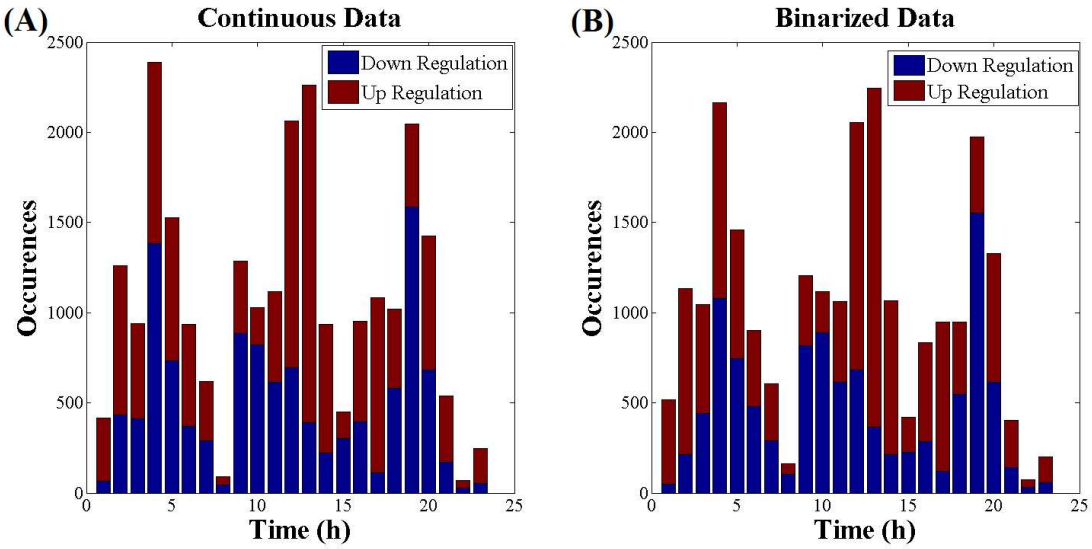


Figure 5.9: RNA-Seq up/down regulated genes of *Drosophila melanogaster*'s development time series data. (A) Continuous and (B) B4 Binarized data using B4 techniques show a same number of up/down regulated genes and a similar pattern along the development time course.

Table 5.4 represents GO annotations found related to the development biological processes from the top 50 (lowest p-value) GO terms in each experiment. Comparison of top set of GO terms and their p-values between continuous and binary data showed that binarization did not remove any important biological interpretations (found same set of GO terms) and provided similar high level of confidence for the development processes with respect to the continuous data. In fact, the binary data is sufficient to provide important biological interpretations of development life cycle using both microarray and RNA-Seq measurements. Further, Figure 5.10 shows the GO term scatter plot of continuous and binary data of RNA-Seq development time point where the highest number of down regulated genes were detected. See Appendix F for RNA-Seq up regulation and microarray up and down regulation GO term scatter plots. These results also show that both continuous and binary data have similar gene ontology terms and binary data is sufficient provide gene enrichment analysis along the development time course. We also compared the continuous and binary (using B4) data GO terms by simply taking random time points in up/down regulated RNA-Seq and microarray time series data. Appendix F REVIGO gene ontology scatter plots show that both RNA-Seq and microarray data provide same GO terms with continuous and binary measurements.

Table 5.4: Comparison of GO annotations and their statistical confidence levels related to time points which gave higest up/down regulated genes during the development process of *Drosophila melanogaster* (Cont. and Bin stand for continuous and binary data respectively). Here we only list the significant GO terms related to development life cycle found within the top 50 (lowest p-values) GO terms. B4 gene by gene GMM threshold binarization technique was employed to convert to binary measurements.

Regulation Type	GO ID	GO Annotation	Cont. p-val	Bin. p-val
RNA-Seq	$GO:0048731$	system development	6.83×10^{-77}	8.37×10^{-69}
	$GO:0007399$	nervous system development	3.46×10^{-75}	1.81×10^{-58}
	$GO:0030154$	cell differentiation	2.06×10^{-65}	3.48×10^{-50}
	$GO:0007275$	multicellular-organism development	2.91×10^{-65}	1.01×10^{-53}
	$GO:0048869$	cellular developmental process	1.12×10^{-63}	7.76×10^{-49}
Up Reg.	$GO:0048856$	anatomical structure development	3.63×10^{-57}	6.91×10^{-50}
	$GO:0048513$	organ development	6.75×10^{-54}	1.50×10^{-65}
	$GO:0032502$	developmental process	7.34×10^{-53}	1.50×10^{-65}
	$GO:0007444$	imaginal disc development	8.97×10^{-51}	9.45×10^{-51}
RNA-Seq	$GO:0048731$	system development	7.70×10^{-49}	6.74×10^{-45}
	$GO:0030154$	cell differentiation	7.05×10^{-48}	6.50×10^{-45}
	$GO:0048513$	organ development	3.01×10^{-47}	2.72×10^{-45}
	$GO:0048869$	cellular developmental process	4.56×10^{-47}	7.01×10^{-37}
	$GO:0044767$	single-organism deve. process	4.73×10^{-44}	3.66×10^{-37}
Down Reg.	$GO:0048468$	cell development	1.08×10^{-43}	7.01×10^{-37}
	$GO:0048856$	anatomical structure development	5.01×10^{-43}	1.75×10^{-35}
	$GO:0007399$	nervous system development	7.97×10^{-43}	7.62×10^{-35}
	$GO:0007275$	multicellular-organism development	4.02×10^{-41}	1.04×10^{-38}
	$GO:0032502$	developmental process	7.49×10^{-41}	2.17×10^{-36}
Microarray	$GO:0042335$	cuticle development	1.89×10^{-04}	1.87×10^{-05}
	$GO:0051146$	striated muscle differentiation	0.04	0.05
	$GO:0055001$	muscle cell development	0.01	0.05
Up Reg.	$GO:0055002$	striated muscle cell development	0.01	0.05
	$GO:0040003$	chitin-based cuticle development	0.01	2.54×10^{-04}
	$GO:0030427$	site of polarized growth	0.05	0.01
Microarray	$GO:0030154$	cell differentiation	3.15×10^{-39}	7.31×10^{-52}
	$GO:0048869$	cellular developmental process	1.10×10^{-38}	1.42×10^{-50}
	$GO:0007399$	nervous system development	8.96e-28	8.62×10^{-45}
Down Reg.	$GO:0048731$	system development	8.58×10^{-20}	6.49×10^{-33}
	$GO:0048856$	anatomical structure development	3.79×10^{-18}	3.01×10^{-26}
	$GO:0032502$	developmental process	1.77×10^{-17}	2.28×10^{-24}
	$GO:0007275$	multicellularorganism development	2.45×10^{-16}	8.68×10^{-25}
	$GO:0048468$	cell development	1.06×10^{-08}	7.24×10^{-15}
	$GO:0003006$	developmental in reproduction	2.49×10^{-07}	6.20×10^{-09}
	$GO:0045595$	regulation of cell differentiation	1.70×10^{-06}	3.88×10^{-07}

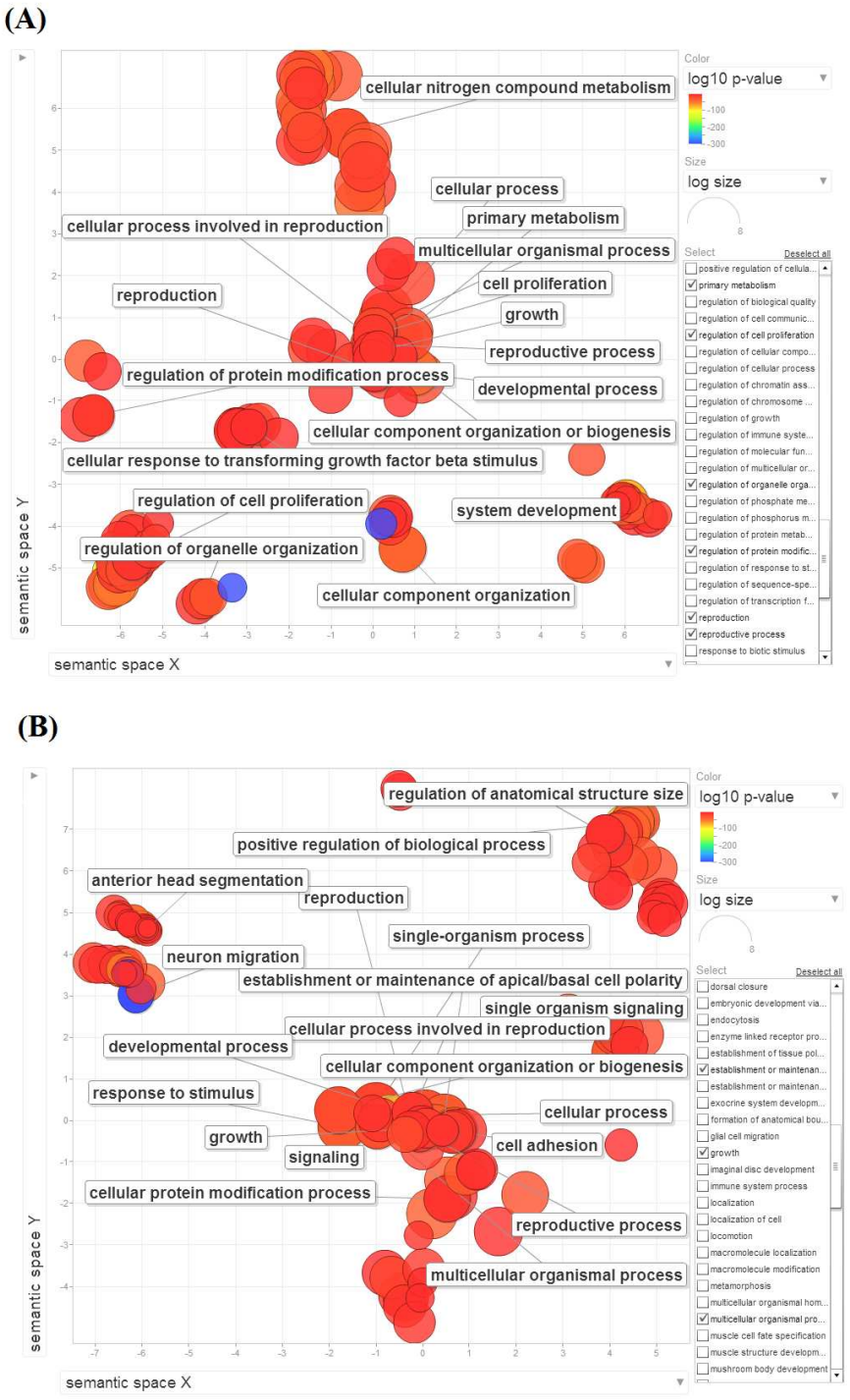


Figure 5.10: GO terms scatter plots of (A) continuous and (B) binary RNA-Seq measurements obtained at the highest number of up regulated genes detected time point of the *Drosophila melanogaster*'s developmental time course Graveley et al. (2011)

5.2.5 Cross Platform Analysis

In order to compare the two major transcriptome measuring techniques in a common environment, a cross platform analysis was carried out. We selected four cancer types, where we obtained both microarray and RNA-seq measurements to perform classification tasks. In this experiment, well studied microarray data was used to train the SVM classifier (as the training environment) and the recently developed RNA-Seq data was employed as the testing environment. Feature selection was carried out at the training environment (with microarrauy data) using Fisher Score technique. Same features (genes) were selected as the inputs for the testing environment with RNA-Seq data. Different combinations of continuous and binary data were incorporated with the SVM classifiers to obtain the cross platform analysis, those are;

- **C + C :** Training using continuous (microarray) data and testing also using continuous (RNA-Seq) data
- **C + B :** Training using continuous (microarray) data and testing using binary (RNA-Seq) data
- **B + C :** Training using binary (microarray) data and testing using continuous (RNA-Seq) data
- **B + B :** Training using binary (microarray) data and testing also using binary (RNA-Seq) data

Here we used B4 gene by gene GMM threshold binarization techniques, which was identified as the best method to binarize data with classification and clustering experiments. This method will be able deal with data distribution of cancer and normal patient data as a two component mixture model. SVM was used as the main classifier because it performs well with data measured under different scales and dimensions. In all cases, transcriptome data was converted in the log scale before using it for experimental analysis.

5.2.5.1 Results

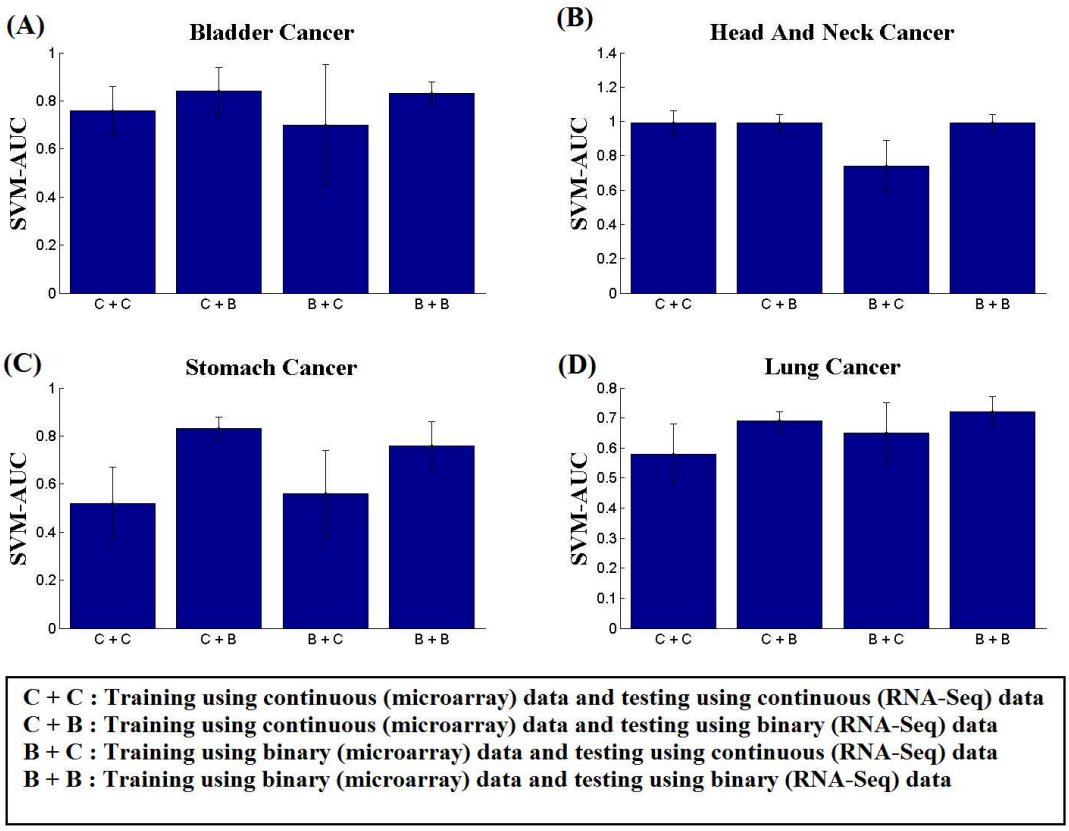


Figure 5.11: Cross Platform Analysis: SVM was trained using Microarray data and tested on RNA-Seq data. Feature selection was performed on microarray environment. B4 gene by gene GMM threshold binarization techniques was employed to convert continuous data into binary.

Figure 5.11 illustrates the cross platform analysis between microarray and RNA-Seq measurements. We observed that, in all four cancer types, training on microarray continuous data and testing on RNA-Seq binary data ($\mathbf{C + B}$) and both training and testing using binary data ($\mathbf{B + B}$) gave similar or better accuracies compared to traditional continuous training and testing experiments ($\mathbf{C + C}$). Therefore, even under different platforms binarized transcriptome data perfom similar or better with respect to continuous data. In fact, binarization removes noise from the testing data and improves the accuracy in $\mathbf{C + B}$ approach and in binary representation itself is sufficient to produce good classification accuracies. We also observed that training on binary and testing on continuous data ($\mathbf{B + C}$) reduced the accuracy of continuous train and testing experiment ($\mathbf{C + C}$) due to the noise addition to the testing data by using continuous data.

5.3 Transcriptome-Proteome Inferences

In previous section (Section 5.2) we observed that microarray and RNA-Seq high precision expression levels do not provide additional information whereas binary representation is sufficient to perform machine learning inferences. However, in our study we are more interested in modelling transcriptome-proteome interface. Therefore, here we investigate the correlation, prediction capability of protein abundance and PTR detection as outliers with high and low numerical precision of microarray and RNA-Seq transcriptome measurements. B4 gene by gene GMM threshold binarization technique was employed to convert continuous data into binary in following experiments.

5.3.1 Correlation

We used yeast (*Saccharomyces cerevisiae* - strain S2883) organism RNA-Seq and microarray transcriptome measurements with proteome data to compare the correlation of these two measurements with respect to the proteome measurements (in log scale). Microarray (Greenbaum et al., 2003) and RNA-Seq (Dang et al., 2014) data were obtained under exponentially growing conditions. Figure 5.12 shows the correlations between transcriptome and proteome levels for RNA-Seq ($R^2 = 0.65$) and microarray ($R^2 = 0.70$) measurements of yeast organism. Thus, both techniques showed a similar correlation with respect to the protein abundance. Fu et al., 2009 carried out a similar experiment using human brain tissue and they found RNA-Seq ($R^2 = 0.62$) correlates better than microarray ($R^2 = 0.75$) measurements. However, the relationship between transcriptome and proteome levels in human cells are much complicated than yeast due to the alternative splicing process (Lundberg et al., 2010; Nilsen and Graveley, 2010). Therefore, we believe yeast organism data is more appropriate to observe direct relationship between transcriptomic and proteomic measurements.

5.3.2 Protein Abundance Predictor

We also developed protein abundance predictors using microarray and RNA-Seq data to compare the transcriptome-proteome modelling capabilities of these two transcriptome measurements. A combination of mRNA levels, translation efficiencies and sequence derived codon bias information were incorporated to develop the protein abundance predictor (*i.e.* input features are - mRNA, tRNA adaptation index (tAI), codon bias, ribosome density and occupancy) according to Chapter 3.

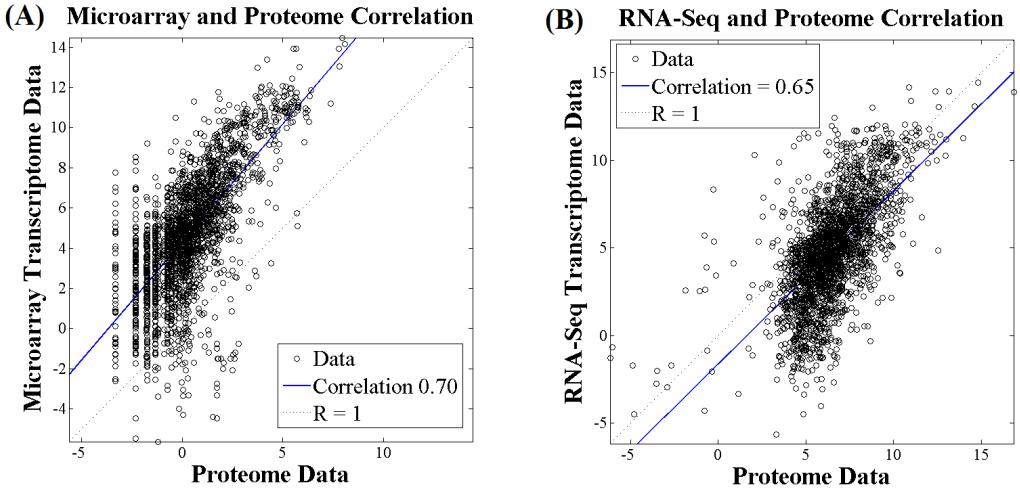


Figure 5.12: Correlation comparison of yeast (*Saccharomyces cerevisiae* - strain S2883) transcriptome and proteome data. (A) Microarray (Greenbaum et al., 2003) and (B) RNA-Seq (Dang et al., 2014) techniques were used to measured transcriptome measurements. Protein data was downloaded from PaxDB (Wang et al., 2012a).

In fact, with two predictors, only mRNA input was changed to microarray and RNA-Seq measurements respectively and the rest of the inputs remained same. Next, we used binarized mRNA levels (using B4 binarization techniue) to observe the prediction changes with the precision reduction. We also used neural net technique to develop a non-linear predictor. Table 5.5 shows the continuous and binary data prediction accuracies of two transcriptome measurements. Figure 5.13 represents linear regression outputs for different transcriptome-proteome combinations. We observed that both microarray and RNA-Seq continuous data provide good prediction levels with linear and non-linear regression models. Interestingly, with the binary data, both microarray and RNA-Seq techniques gave very close prediction levels as similar to the continuous data. Therefore, with both measurements, binarization did not remove any important information to develop an accurate protein abundance predictor.

Table 5.5: Linear and non-linear protein abundance predictor regressions. B4 gene by gene GMM threshold binarization technique was employed to binarize data.

Regression Type	Data Type	RNA-Seq	Microarray
Linear Regression	Continuous	0.85	0.86
	Binary	0.85	0.86
Non-linear Regression	Continuous	0.81	0.80
	Binary	0.82	0.82

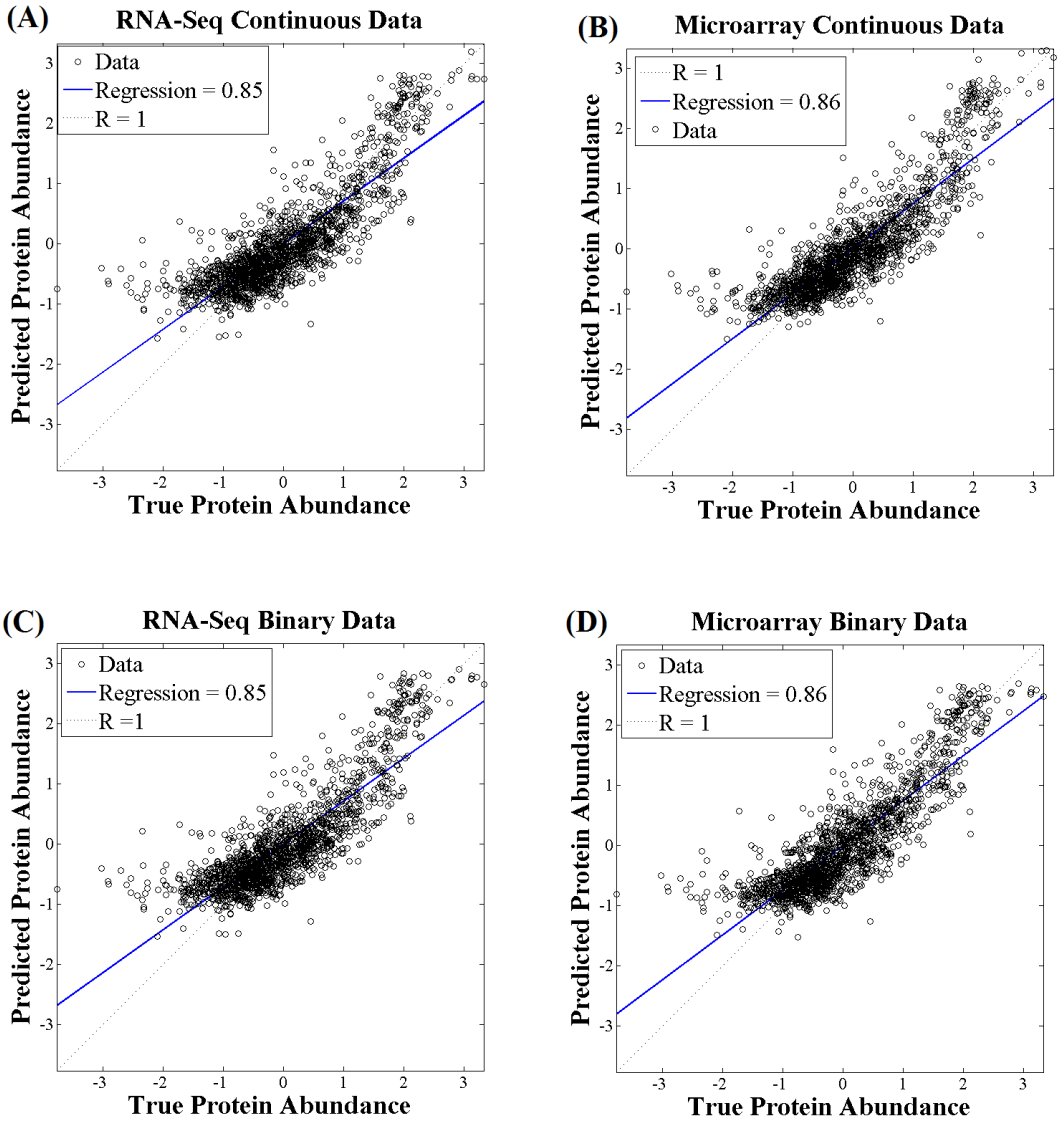


Figure 5.13: Comparison of linear protein abundance prediction accuracies of microarray and RNA-Seq measurements. Five input features were used in all the experiments *i.e* mRNA abundance, tRNA adaptation index, codon bias, ribosome density and occupancy (similar to Chapter 3). Only mRNA abundance was changed as (A) RNA-Seq continuous (B) microarray continuous (C) RNA-Seq binary and (D) microarray binary data. B4 gene by gene GMM threshold binarization techniques was employed to binarize data.

5.3.3 PTR Detection

Though the quantitative regression models performed similarly with both continuous and binary data, here we would like to investigate the effects of binarization with respect to a qualitative attribute which is detecting post-translationally regulated proteins as outliers from the transcriptome-proteome interface. We used four

transcriptomic data types (*i.e* microarray continuous, microarray binary, RNA-Seq continuous and RNA-Seq binary) with our other four input features (tAI, codon bias, ribosome density and ribosome occupancy) to predict protein abundance using quantile regression model. According to Chapter 4, quantile regression gave the highest confidence level in detecting post-translationally regulated proteins as outliers. Fifty outliers were selected from each regression model and coarse and finer level functional annotation checks were performed as described in Section 3.4.

Figure 5.14 shows the Venn diagram of outlier distribution among the four transcriptomic measurements and 70% of the proteins were common to all four measurements. Microarray continuous and binary inputs have seven and six unique proteins respectively and 11 proteins were different in both outlier sets. However, RNA-Seq continuous and binary outliers highly coincide and only one protein differ between two sets. Therefore, RNA-Seq binarization did not remove any outlier information from the regression approach. Table 5.6 shows the coarse and finer level functional annotation confidence levels of four outliers sets. We observed that all four outlier sets are highly confidence with post-translationally regulated proteins. However, there is a slight information loss with microarray binarization and RNA-Seq performed similar in both continuous and binary outlier detections. In fact, RNA-Seq performed marginally better in detecting outliers with respect to microarray regression models. Therefore, RNA-Seq transcriptomic data provide more information with respect to qualitative properties such as detecting post-translationally regulated proteins as outliers at the transcriptome-proteome interface.

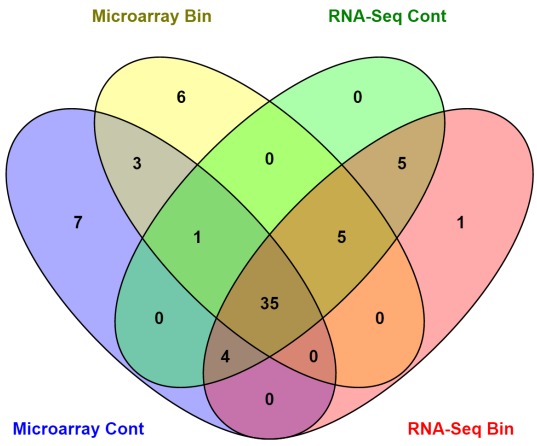


Figure 5.14: Outliers obtained by quantile regression using four types of transcriptomic measurements. *i.e* other four input properties (tAI, codon bias, ribosome density and ribosome occupancy) and proteins abundances are similar in all 4 regression models. *Cont* and *Bin* stand for continuous and binary transcriptomic measurements respectively.

Table 5.6: Coarse and finer level PTM annotation check for four outlier sets. 1000 random trials were used in each case. B4 gene by gene GMM threshold binarization technique was employed to binarize data.

mRNA Input	Coarse Level		Finer Level	
	No of genes	Confidence Level	No of genes	Confidence Level
RNA-Seq Continuous	44	$p \leq 0.02$	40	$p \leq 4.38 \times 10^{-13}$
RNA-Seq Binary	44	$p \leq 0.02$	40	$p \leq 4.38 \times 10^{-13}$
Microarray Continuous	45	$p \leq 9.89 \times 10^{-04}$	38	$p \leq 2.94 \times 10^{-11}$
Microarray Binary	41	$p \leq 0.030$	35	$p \leq 8.31 \times 10^{-09}$

5.4 Summary

In this chapter, we explored the numerical precision inference of the two main transcriptomic measuring techniques which are microarray and RNA-Seq. Tuna and Niranjan, 2009 showed that microarray binarized low numerical precision expression data is sufficient to provide good machine learning inference. Here, we used most recently developed RNA-Seq expression values and compared inference accuracies with the microarray measurements under high and low precision. Firstly, by only considering transcriptomic measurements we observed that classification, clustering, and time series data produced high accuracies with both continuous and binary measurements. Thus, binary measurement is sufficient to provide good machine learning inferences. Secondly, we investigated the influence of high numerical precision in transcriptome-proteome modelling approach. Here we combined the five feature model explained in Chapter 3 and outlier detection in Chapter 4 to perform our analysis. Both transcriptomic measurements (microarray and RNA-Seq) gave similar regression values under high and low numerical input values. Further, binary mRNA inputs in the regression model did not reduce the capability of detecting post-translationally regulated proteins as outliers of the regression approach. In fact, RNA-Seq measurements performs marginally better than microarray measurements in detecting outliers with both continuous and binary data. Thus, this study shows that in machine learning quantitative inferences, high numerical precision obtained by mRNA amplification process in RNA measuring techniques do not provide more information with respect to gene switch on/off status given by binary measurements.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This dissertation is on data-driven modelling applied to the integrated analysis of high-throughput *omic* measurements. Specifically we focus on the interface between transcriptome and proteome levels where the mRNA abundances are often used as proxies for protein concentrations. While several authors have looked for correlation between these two levels of measurements and noted only week relationship on a genomic scale (Gygi et al., 1999; Futcher et al., 1999; Beyer et al., 2004; Wu et al., 2008), we follow a notably different approach, similar to Tuller et al., 2007, in constructing a predictor of protein levels from mRNA levels and other transcriptomic variables may influence the corresponding protein concentrations on yeast (*Saccharomyces cerevisiae*).

In Chapter 3, we show that, such a predictor, using LASSO regularized linear regression, shows increased ability to predict protein levels than simply looking for correlation between mRNA and protein concentrations. Further, LASSO regularization suppress irrelevant features and selects mRNA, codon bias, tAI, ribosome density and occupancy as best features to predict protein abundance more accurately. Additionally, we also show that non-linear models did not help in improving prediction ability at the transcriptome-proteome interface. We then use this protein abundance predictor in a novel manner to identify post-translationally regulated protein by taking model failures which give large errors between the predicted and the actual measurements due to the protein stability disruption by post-translation regulation. However, detecting PTR proteins using mass spectrometry on a laboratory experimental setting is challenging, time consuming and costly. In fact, lack of prior knowledge on the modifications to be detected and

technical limitations to measure low abundances of minor sites induce difficulties to detect PTR proteins effectively (Chandramouli and Qian, 2009; Arnott et al., 2003). Hence, our data-driven model selects sub set of proteins (outliers) as post-translationally regulated proteins and cut-down the number of samples to be tested using mass spectrometry experimental setting. Addtionally, finer level annotation check provides information on PTM types to be detected by these outlier proteins. Thus, experimentalist can use this prior knowledge to detect variations among these modifications and also few number of samples will allow to amplify the low abundance peptide ratios to detect minor sites with less cost and time.

In Chapter 4, we introduce two formulations, (i) Outlier Rejecting Regression (ORR) model and use (ii) Quantile Regression (QR) model, to detect robust outliers at the transcriptome-proteome interface in a systematic manner (or mathematically) to confirm our hypothesis on post-translational regulation. We compare the model failures or outliers of the three models (including simple linear regression model in Chapter 3) using over-representation of functional annotations related to post-translational regulation. All three outlier sets showed good statistical confidence levels providing evidence that outlier proteins are likely candidates for post-translational regulation. Quantile regression with the asymmetric loss model gave the highest confidence level suggesting that among the methods we considered, QR is the best technique to detect post-translationally regulated proteins at the transcriptome-proteome interface. Additionally, all these outlier samples (from three regression models) showed high enrichment of the $p53$ pathway during pathway analysis. Further, Shin et al., 2013 showed that the protein degradation process by post-translation regulation enables $p53$ regulation. Therefore, pathway analysis reconfirmed our hypothesis providing further biological evidence.

In Chapter 5, we focus our study on the numerical precision of high-throughput transcriptome measurements. Tuna and Niranjan, 2009 and Tuna and Niranjan, 2010 studies have shown that high precision microarray measurements provide no additional information with respect to quantized binary data. In fact, binary data is sufficient to obtain high accuracy for machine learning inferences as classification, clustering etc. Hence, in our study we use a novel approach with RNA-Seq measurements, which is considered as a more sensitive and accurate technique to measure mRNA concentrations with respect to microarray technique (Wang et al., 2009; Malone and Oliver, 2011; Fu et al., 2009; Marioni et al., 2008), to perform similar inferences using continuous and binarized data. Further, we compare microarray and RNA-Seq measurements by developing a protein abundance predictor and using model failures to identify post-translationally regulated proteins. We

also combine the five feature predictor in Chapter 3 and quantile regression outlier detection in Chapter 4 to compare microarray and RNA-Seq measurements for high (continuous) and low (binary) precision data. Transcriptomic inferences show that RNA-Seq also behaves similarly to microarray data, where there is not much information loss with binarization. Classification, clustering and time series data analysis have shown that both continuous and binary data give similar accuracies. Similarly, protein abundance prediction accuracies also show that binarized data is sufficient to obtain good regression accuracies for both microarray and RNA-Seq measurements. However, RNA-Seq performs better in detecting PTR as outliers using binarized data compared to microarray binarized measurements which can be considered as a qualitative attribute.

6.2 Future Work

In this thesis, we bridge the gap between the transcriptome and proteome interface and identify post-translationally regulated proteins as outliers using yeast (*Saccharomyces cerevisiae*) data. As the next step, we would like to use human transcriptome and proteome measurements with our data-driven framework. However, modelling human data using a regression approach is difficult due to the alternative splicing process (Lundberg et al., 2010; Nilsen and Graveley, 2010). Therefore, as the first approach, we can expand our input feature space by including microRNA (miRNA) information to deal with alternative splicing at the isoform level. miRNAs are short RNA molecules which regulate mRNAs by directly binding to the 3'UTR region (Bartel, 2004). These molecules will inhibit the translation or mRNA degradation process. Similarly we would like to use ribo-seq measurements to obtain translation efficiency rates (Barry and Hartigan, 1993). Secondly, we can use a probabilistic non linear approach to model human *omic* measurements. As we discussed in the literature review, Kannan et al., 2007 used a Bayesian model which links microarray mRNA measurements with mass spectrometry protein measurements for the entire genome of laboratory mouse, *Mus musculus*. They learn the model and score the genes as a measurement of the strength of the relationship between mRNA and protein data using probabilistic inferences. However, their Bayesian network lack of knowledge in translation process to obtain the correct relationship of mRNA and protein data; *e.g* mRNA is translated in to protein but their Bayesian network represents this relationship in the opposite direction. Therefore, we can incorporate new transcriptome properties such as miRNA and ribo-seq information with isoform data and develop a

probabilistic Bayesian model along the human genome. Thus, least probable proteins can be extracted as candidates for post-translationally regulated proteins, which can be used as biomarkers for cancer and other diseases.

We can then use this data-driven framework with cancer and normal patient data. In Chapter 4, during the pathway analysis, we observed $p53$ pathway as a dominant pathway with all of three outlier sets. However, $p53$ is a tumour suppression protein, where the mutation would be directly involved with cancer studies (Butz et al., 1995; Bodner et al., 1992; Hollstein et al., 1991). Comparing the outliers of our data-driven model using cancer and normal patient data we would be able to provide information about cancer causing proteins due to $p53$ mutations. Further, mutations caused by cancer can increase protein concentrations with respect to cancer free (normal) protein. Olsen et al., 2007 have shown that HER-2 protein concentration is higher than a normal reference protein. Similar behaviour have shown with lung cancer gene apolipoprotein E by in a study by Trost et al., 2008. Thus, we can change our hypothesis to observe outlier proteins in the lower region of the regression plot where the measured protein concentration is higher than the predicted ($P > \hat{P}$) concentration in order to detect proteins that increase their concentration with cancer mutations. By comparing the cancer and the normal outliers, we will be able to uncover new proteins with potential cancer mutations. We can perform a statistical test to obtain the significance of these proteins using currently available literature and biologists can then perform laboratory experiments on these potential cancer causing proteins. Therefore, we can use our data-driven model to discover new biomarkers and therapeutic interventions for human cancer.

Appendix A

Linear Predictor

As mentioned in the Literature Review chapter, Tuller et al., 2007 have used the linear regressor to develop their protein abundance predictor. In this model we assume that the relationship between input and target data is linear and the main objective of this method to minimize the squared error or loss (Rogers and Girolami, 2012b).

$$\min \left\{ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \right\} \tag{A.1}$$

Assume we have p samples and n features for the input matrix $\mathbf{X}$ and p targets in target vector $\mathbf{y}$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & . \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$$

We can generate output vector $\mathbf{t}$ by the dot product of $\mathbf{X}^\top$ and $\mathbf{w}$ (assuming relationship between input and target is linear)

$$\begin{matrix} \mathbf{t} \\ \\ \end{matrix} \quad = \quad \begin{matrix} \mathbf{X}^\top \\ \\ \end{matrix} \quad \begin{matrix} \mathbf{w} \\ \\ \end{matrix}$$

$$\begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{pmatrix} \quad \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ x_{13} & x_{23} & \cdots & x_{n3} \\ \vdots & \vdots & \cdots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \quad \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{pmatrix}$$

Thus, we can define the error vector $\epsilon = \mathbf{X}^\top \mathbf{w} - \mathbf{y}$

Least Squared Error

The squared error is given as below,

$$\begin{aligned}
\mathbf{E}(\mathbf{w}|\mathbf{D}) &= \|\epsilon\|^2 = \|\mathbf{X}^\top \mathbf{w} - \mathbf{y}\|^2 \\
&= (\mathbf{X}^\top \mathbf{w} - \mathbf{y})^\top (\mathbf{X}^\top \mathbf{w} - \mathbf{y}) \\
&= (\mathbf{w}^\top \mathbf{X} - \mathbf{y}^\top)(\mathbf{X}^\top \mathbf{w} - \mathbf{y}) \\
&= \mathbf{w}\mathbf{X}\mathbf{X}^\top \mathbf{w} - \mathbf{y}^\top \mathbf{X}^\top \mathbf{w} - \mathbf{w}^\top \mathbf{X}\mathbf{y} + \mathbf{y}^\top \mathbf{y} \\
&= \mathbf{w}\mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y} + \mathbf{y}^\top \mathbf{y}
\end{aligned}$$

In order to find minimum values of the squared error or loss function, we take the partial derivative of the $\mathbf{E}(\mathbf{w}|\mathbf{D})$ with respect to the weight vector $\mathbf{w}$ ($\frac{\partial \mathbf{E}(\mathbf{w}|\mathbf{D}}{\partial \mathbf{w}} = \nabla \mathbf{E}(\mathbf{w}|\mathbf{D}) = 0$).

$$\begin{aligned}
\nabla \mathbf{E}(\mathbf{w}|\mathbf{D}) &= \nabla(\mathbf{w}\mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y} + \mathbf{y}^\top \mathbf{y}) \\
&= 2(\mathbf{X}\mathbf{X}^\top \mathbf{w} - \mathbf{X}\mathbf{y}) \\
&= 2(\mathbf{X}\mathbf{X}^\top \mathbf{w} - \mathbf{X}\mathbf{y}) = 0 \\
\mathbf{w} &= (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}
\end{aligned}$$

$(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$ is also known as pseudo inverse. This method models linear relationships among data, where we can predict outputs for unseen data in a precise manner.

Appendix B

Detecting Outlier Using Gaussian Mixture Model (GMM)

We also employed GMM appraoch to model our transcriptome-proteome measurements using a probabilistic model assuming that all data points are generated from a mixture of Gaussian distributions. A review by Pimentel et al., 2014 describes the use of GMM algorithm in detecting outliers or one-class classification with cancer patients data with image information. Several authors used GMM with image information to diagnose breast cancer (Li et al., 2012; Lederman et al., 2011). Therefore, here we use *omic* measurements (five input features and protein abundance) with GMM approach to detect post-translationally regulated proteins.

As mentioned in Chapter 2 (Literature Review), each Gaussian density (component) $N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$ has its own mean ($\boldsymbol{\mu}_k$) and standard deviation($\boldsymbol{\sigma}_k$),

$$p(\mathbf{x}) = \sum_{k=1}^{M} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) \tag{B.1}$$

where π_k represents the mixing coefficient satisfying $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{M} \pi_k = 1$.

Expectation Maximization (EM) algorithm (Dempster et al., 1977a) was used to estimate the parameters of these Gaussian densities. Afterwards, we obtained the negative log likelihood of the samples based on the estimated parameters and extract the least probable (with lowest likelihood) samples as outliers from the Gaussian mixture model.

Figure B.1 shows an example of randomly generated data using a two dimensional Gaussian data distribution and fitting a single component GMM. Least probable 10 samples are circled in red. Here, we observe that these samples with lowest likelihood (circles in red) are lying as outliers from the total data set.

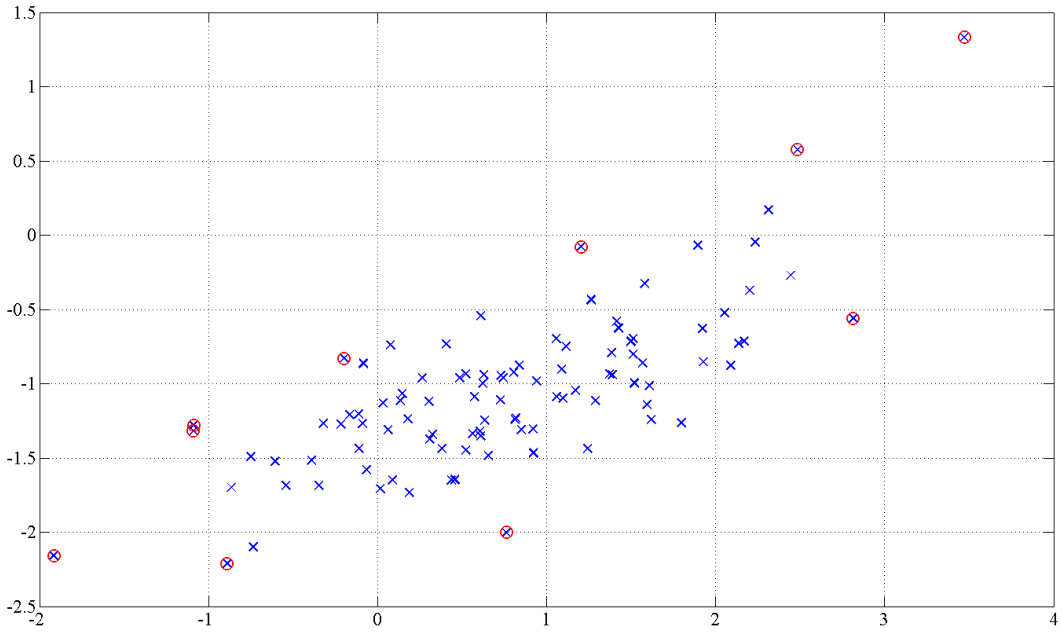


Figure B.1: Randomly generated data using two dimensional Gaussian distribution and fitting a single component GMM. The mean vector is are $\mu = [1, -1]$ and covariance matrix is $\sigma = [0.9, 0.4; 0.4, 0.3]$. Red circles represent the samples with least likelihood probability as outliers.

Similarly, we fitted a 6 component GMM to our transcriptomic and proteomic measurements (five input features and protein abundance) using `Netlab` package in `MATLAB` environment. Least probable 50 samples as outliers assuming 50 as the benchmark number of outliers with respect to the linear regression model. Since we cannot represent the 6 dimensions GMM plot with the outliers, we used our linear regression plot and circled the new set of outliers obtained from GMM in pink as shown in Figure B.2. However, GMM model only selected 13 samples similar to linear regression model (Figure 3.6) and the functional annotation checks gave low statistical confidence levels for coarse ($p \leq 0.93$) and finer ($p \leq 0.13$) levels, showing that these outliers are not significant with respect to post-translational regulation. We also observed that same set of outliers were detected under different initializations for parameters. Further, we changed the number of mixture components and observed the detected outliers and their significance in post-translational regulation. Table B.1 shows the PTR detection significance under different number of mixture components. In all cases, GMM was not able to detect

post-translationally regulated proteins with a high statistical confidence. GMM has the limitation of selecting a functional form for the input data distribution which may not be suitable to generate a good model with *omic* measurements where the underlying functional form is unknown (Pimentel et al., 2014). We believe that assuming all input features have normal (Gaussian) distribution and parameter estimation errors (random starting points), reduced the potential of capturing protein stability disruption property as outliers. Therefore, linear regression model outperforms in detecting post-translationally regulated proteins as outliers with respect to GMM approach.

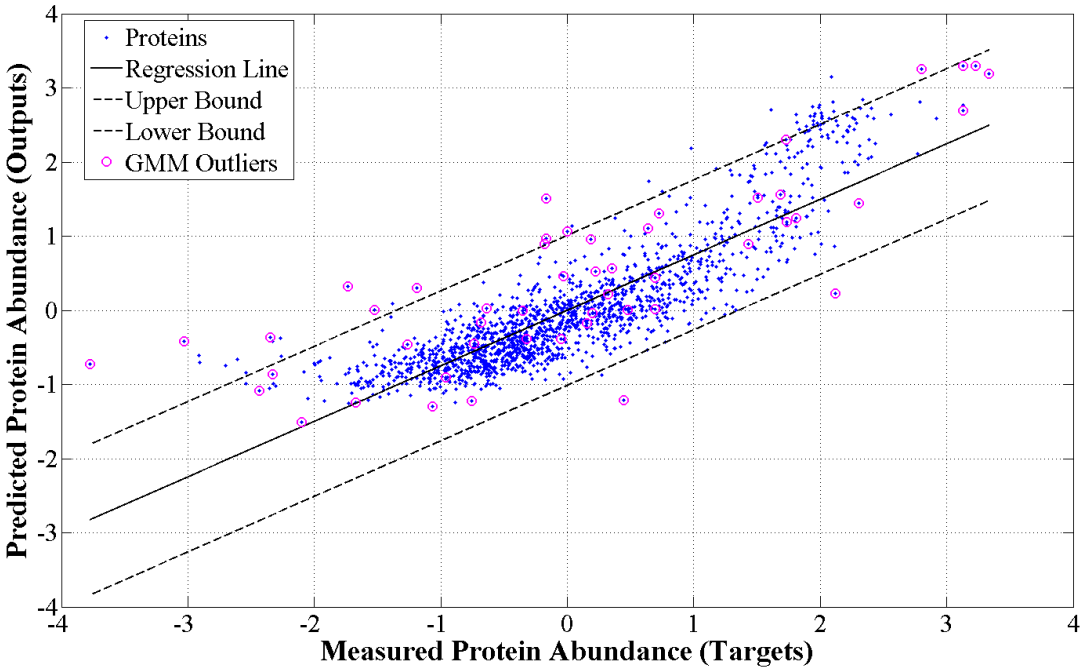


Figure B.2: Fifty outlier proteins detected by the GMM model are circled in pink colour. Only 13 proteins are similar with linear regression outliers. Majority of the GMM outliers are inline with the regression plot. Therefore, this outliers do not show protein stability disruption property by post-translational regulation.

Table B.1: PTR detection under different number of mixture components. Lowest probable 50 outliers were selected in each experiemnt and 1000 random trials were used to obtain the p-values

No of Components	Coarse Level		Finer Level	
	PTR Proteins	Confidence Level	PTR Proteins	Confidence Level
1	30	$p \leq 0.88$	20	$p \leq 0.13$
2	27	$p \leq 0.98$	18	$p \leq 0.29$
3	28	$p \leq 0.96$	19	$p \leq 0.20$
4	29	$p \leq 0.93$	20	$p \leq 0.13$
5	29	$p \leq 0.93$	20	$p \leq 0.13$
6	29	$p \leq 0.93$	20	$p \leq 0.13$
10	28	$p \leq 0.96$	19	$p \leq 0.20$
100	22	$p \leq 0.99$	17	$p \leq 0.41$

Appendix C

Difference of Convex functions Algorithm (DCA) in ORR Model

In this section we describe the steps of solving the DCA in Outlier Rejecting Regression (ORR) model. Suppose we have a set of m samples $\{(x_i, y_i)\}_{i=1,\dots,m}$ where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ and the main objective is to predict y_i as as $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ with smallest error.

Objective function for the regression model can be written as below,

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\eta}} \quad & \frac{1}{(1-\mu)m} \sum_i \eta_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) + \lambda \|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & \textstyle\sum_i (1-\eta_i) \le \mu m, \quad 0 \le \eta_i \le 1, \ \forall i, \end{aligned} \tag{C.1}$$

where $\mu \in [0, 1)$ and $\lambda \in (0, \infty)$ are hyper parameters.

We can re-write the objective function (C.1) as a difference of convex functions as below:

$$\begin{aligned} &\min_{\boldsymbol{w},b} \quad \frac{1}{(1-\mu)m} \left\{ \sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) - \mu m \phi_{1-\mu}(\boldsymbol{w}, b) \right\} + \lambda \|\boldsymbol{w}\|^2, \\ &\min_{\boldsymbol{w},b} \quad \underbrace{\frac{1}{(1-\mu)m} \sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) + \lambda \|\boldsymbol{w}\|^2}_{\text{convex}} - \underbrace{\frac{\mu}{1-\mu} \phi_{1-\mu}(\boldsymbol{w}, b)}_{\text{convex}}, \end{aligned} \tag{C.2}$$

where $\phi_{1-\mu}(\boldsymbol{w}, b)$ is $(1-\mu)$-Conditional Value-at-Risk (CVaR) which is defined as the mean of μ-tail distribution (the white area in Figure C.1).

Rockafellar and Uryasev (2002) proposed Conditional Value-at-Risk (CVar) as a financial risk measure, which can be written as

$$\phi_{1-\mu}(\boldsymbol{w}, b) = \max_{\boldsymbol{\eta}} \left\{ \tfrac{1}{\mu m} \textstyle\sum_i (1-\eta_i)\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) : \sum_i (1-\eta_i) = \mu m, \quad 0 \le \eta_i \le 1, \ \forall i \right\}.$$

Takeda and Sugiyama (2008)'s study shows that ν-SVM is able to minimizes the financial risk CVaR. With respect to the *clipped* loss function we minimizes the mean loss of the gray area in Figure C.1.

Next, we use optimal solution of η^* in Equation (C.1) to define the set of outliers as Θ,

$$\Theta := \{i \in \{1, \ldots, m\} : \eta_i^* < 1\}$$

Thus, the clipped loss function of regression model in Equation (C.2) can be written as

$$\frac{1}{(1-\mu)m} \left\{ \sum_{i=1}^{m} \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) - \sum_{i \in \Theta} \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) \right\}.$$

DC algorithm sequentially linearizes the concave part of Equation (C.2) and solves the convex subproblem. Let (w_k, b_k) be the solution obtained in the $(k-1)$th iteration. In the kth iteration, we solve the following subproblem:

$$h(\boldsymbol{w}_{k+1}, b_{k+1}) := \min_{\boldsymbol{w}, b} \quad \frac{1}{(1-\mu)m} \left\{ \sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) - \mu m (\boldsymbol{g}_w^{k\top} \boldsymbol{w} + g_b^k b) \right\} + \lambda \|\boldsymbol{w}\|^2 \tag{C.3}$$

where $\boldsymbol{g}_w^k \in \partial_w \phi_{1-\mu}(\boldsymbol{w}_k, b_k)$ and $g_b^k \in \partial_b \phi_{1-\mu}(\boldsymbol{w}_k, b_k)$ are a subgradient of $\phi_{1-\mu}(\boldsymbol{w}, b)$ at $(\boldsymbol{w}_k, b_k)$ which can be calculated by sorting the loss $\ell_i(x_i, y_i; w_k, b_k)$.

Sequence $\{(\boldsymbol{w}_k, b_k)\}$ obtained by Algorithm 2 has good convergence properties as below:

- Objective value is decreasing in each step (*i.e.* $h(\boldsymbol{w}_{k+1}, b_{k+1}) \le h(\boldsymbol{w}_k, b_k)$)
- Every limit point of the sequence is a critical point of Equation (C.2) (*i.e.* critical points are defined in Pham Dinh and Le Thi (1997))

The critical point is also known as *generalized KKT point* which is a necessary condition to obtain a local solution.

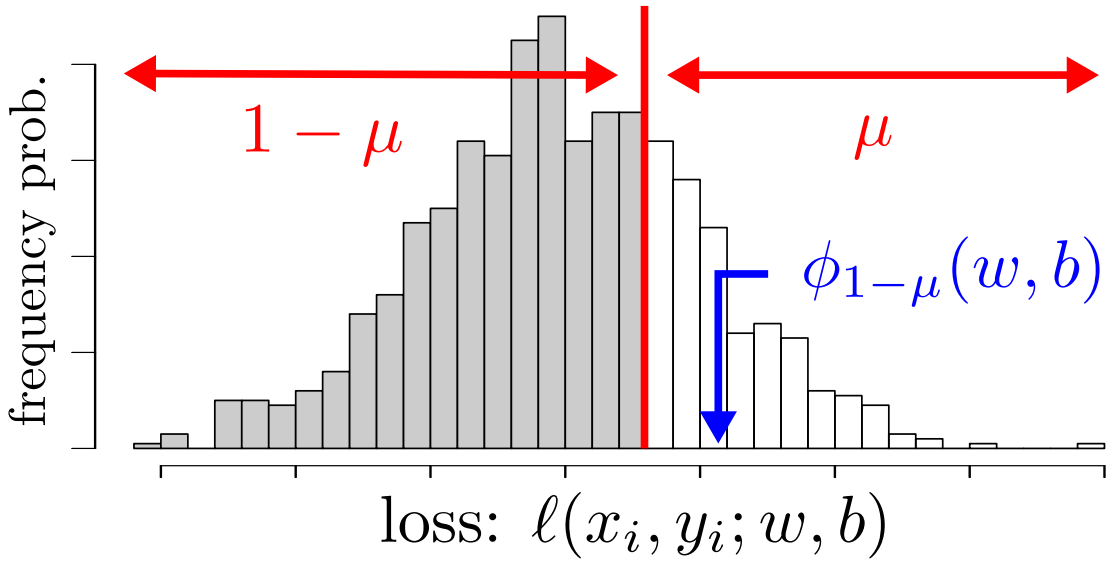


Figure C.1: ORR model minimizes the mean loss in gray area. $(1-\mu)$-CVar denotes mean loss of the white area. Difference between total mean loss and $(1-\mu)$-CVar will give the mean loss (gray area) of the ORR model.

Definition C.1. $(\boldsymbol{w}^*, b^*)$ is said to be a critical point of $u(\boldsymbol{w}, b) - v(\boldsymbol{w}, b)$ if $\partial u(\boldsymbol{w}^*, b^*) \cap \partial v(\boldsymbol{w}^*, b^*) \neq \emptyset$.

According to Definition C.1, a critical point $(\boldsymbol{w}^*, b^*)$ has $\boldsymbol{g}_u \in \partial u(\boldsymbol{w}^*, b^*)$ and $\boldsymbol{g}_v \in \partial v(\boldsymbol{w}^*, b^*)$ such that $\boldsymbol{g}_u - \boldsymbol{g}_v = 0$ which is a necessary condition for local minimum of Equation (C.3).

Matlab Codes for ORR Model

Following are the Matlab codes for ORR Model using Algorithm 1 and 2. All functions require NormalizeData() function to normalize data.

Data Normalization

```
function [Y, f] = NormalizeData(Y,f)
% Y = Input Data, f = targets

[N, p] = size(Y);

for j=1:p
    Y(:,j)=Y(:,j)-mean(Y(:,j));
    Y(:,j)=Y(:,j)/std(Y(:,j));
end
```

```
f = f - mean(f);
f = f/std(f);
```

ORR - Algorithm 1

```
function [Outliers] = ORR1(Y,f,Mu,lambda,eps)
% function inputs are Y = Input data,f = targets,
% Mu = fraction needed as outliers (between 0-1)
% lambda , eps user define variables

% Normalize the data, zero mean, unit standard deviation
[Y, f] =  NormalizeData(Y,f);

% Initialize
s = size(Y);
n = s(1);
m = s(2);
% Define required number of outliers
noOfOutliers = Mu*m;
f_dif = 1;

% Least squares regression to obtain initial weights
YY = [Y ones(n,1)]
w = inv(YY'*YY)*YY'*f;
fh = YY*w;
w0 = w(1:n,1);
b0 = w(n+1,1);

%DCA
while f_dif >= eps
   % Get the outlier list
   fh = Y*w0 + b0;
   Error = (f - fh).^2;
   [E iE] = sort(Error,'descend');
   iOut = iE(1:noOfOutliers);
   Yout = Y(iOut,:);
   fout = f(iOut,1);

   % Get Gradient gw
   A = zeros(noOfOutliers,n);
   for i=1:noOfOutliers
       t = fout(i) - ((Yout(i,:)*w0) + b0);
```

```
        A(i,:) = t*Yout(i,:);
    end
    gw = ((-2)*(sum(A)))/(noOfOutliers);

    % Get Gradient gb
    B = zeros(noOfOutliers,1);
    for i=1:noOfOutliers
        B(i) = fout(i) - ((Yout(i,:)*w0) + b0);
    end
    gb = ((-2)*(sum(B)))/(noOfOutliers);

    % Calculate Constants
    k = 1/(m-noOfOutliers);
    K= noOfOutliers*k;

    % Do the convex programming
    cvx_begin
        variables w(n) b;
        minimize ((square_pos(norm((Y*w +b) -f))*(k)) - ((gw*w + gb*b)*(K)));
    cvx_end

    f_old = ((square_pos(norm((Y*w0 +b0) -f))*(k)) - ((gw*w0 + gb*b0)*(K)));
    f_new = ((square_pos(norm((Y*w +b) -f))*(k)) - ((gw*w + gb*b)*(K)));

    % Assign new values to the next iteration
    w0 = w;
    b0 = b;

    f_dif = abs((f_old - f_new));
end

% Get the Best Outliers
fh = Y*w +b;
Error = (f - fh).^2;
[E iE] = sort(Error,'descend');
Outliers = iE(1:noOfOutliers);
```

ORR - Algorithm 2

```
function [Outliers] = ORR2(Y,f,Mu,lambda,eps)
% function inputs are Y = Input data,f = targets,
% Mu = fraction needed as outliers
% lambda , eps user define variables
```

```
% Normalize the data, zero mean, unit standard deviation
[Y, f] =  NormalizeData(Y,f);

% Initialize
s = size(Y);
n = s(1);
m = s(2);
%Get Mum number of outliers
noOfOutliers = Mu*m;
f_dif = 1;

% Least squares regression to obatin initial weights
w = inv(Y'*Y)*Y'*f;
fh = Y*w;
z = zeros(m,1);
S = [eye(m) z; z' 0];
a0 = w;
neta = ones(n,1);
D = diag(neta);
Dh = sqrt(D);

% DCA
while f_dif >= eps
   % Step 1
   cvx_begin quiet
      variable a(m+1)
      minimize(square_pos(norm(Dh*Y*a - Dh*f)) + lambda*square_pos(norm(S*a)))
   cvx_end

   % Step 2
   f_old = (square_pos(norm(Dh*Y*a0 - Dh*f)) + lambda*square_pos(norm(S*a0)));
   Error = (Y*a-f) .^2;
   [E iE] = sort(Error);
   k= 1-Mu;
   Outliers = iE(ceil(k*n):n);
   neta = ones(n,1);
   neta(Outliers)=0;
   D = diag(neta);
   Dh = sqrt(D);
   f_new = (square_pos(norm(Dh*Y*a - Dh*f)) + lambda*square_pos(norm(S*a)));
   f_dif = abs((f_old - f_new));
 a0 =a;
end

% Get the Best Outliers
```

```
Error = (Y*a-f) .^2;
[E iE] = sort(Error);
k= 1-Mu;
Outliers = iE(ceil(k*n):n);
```

Appendix D

Outlier Proteins Detected by Three Regression Models

Table D.1 shows the union set of 92 proteins and their corresponding outlier detection technique.

ORF Name	Gene Name	Model 0	Model 1	Model 2
YAL007C	YAL007C	x	✓	✓
YAL015C	NTG1	x	✓	✓
YAR075W	YAR075W	✓	✓	✓
YBL0613	RPS8A	✓	x	✓
YBR010W	HHT1	✓	x	x
YBR038W	CHS2	✓	✓	✓
YBR106W	PHO88	✓	✓	x
YBR1317	RPS9B	✓	x	✓
YBR150C	TBS1	x	✓	✓
YBR246W	ERE1	x	x	✓
YBR248C	HIS7	x	✓	x
YCR010C	ADY2	✓	✓	✓
YCR012W	PGK1	✓	x	x
YCR031C	RPS14A	✓	x	x
YDL014W	NOP1	✓	x	x
YDL048C	STP4	✓	✓	✓
YDL080C	THI3	✓	✓	x
YDL081C	RPP1A	✓	x	x
YDL083C	RPS16B	✓	x	x
YDL093W	PMT5	✓	✓	✓

YDL130W	RPP1B	✓	x	x
YDL195W	SEC31	x	✓	x
YDL217C	TIM22	✓	✓	✓
YDL222C	FMP45	✓	✓	✓
YDR006C	SOK1	✓	✓	✓
YDR064W	RPS13	✓	x	x
YDR067C	OCA6	x	✓	✓
YDR098C	GRX3	x	✓	✓
YDR127W	ARO1	x	✓	x
YDR144C	MKC7	x	x	✓
YDR225W	HTA1	✓	✓	✓
YDR243C	PRP28	x	✓	✓
YDR342C	HXT7	✓	✓	✓
YDR382W	RPP2B	✓	x	✓
YDR407C	TRS120	x	✓	✓
YDR447C	RPS17B	✓	x	x
YDR464W	SPP41	x	x	✓
YDR490C	PKH1	x	x	✓
YEL027W	VMA3	✓	x	x
YEL034W	HYP2	✓	x	x
YEL039C	CYC7	x	x	✓
YER023W	PRO3	x	✓	x
YER026C	CHO1	x	✓	✓
YER030W	CHZ1	x	x	✓
YER069W	ARG5,6	x	✓	x
YER070W	RNR1	✓	✓	x
YER093C	TSC11	x	✓	✓
YER102W	RPS8B	✓	x	x
YER117W	RPL23B	✓	x	x
YER183C	FAU1	x	✓	✓
YFL029C	CAK1	✓	✓	✓
YFL036W	RPO41	x	✓	x
YFR039C	YFR039C	x	x	✓
YGL009C	LEU1	x	✓	x
YGL012W	ERG4	x	x	✓
YGL030W	YGL030W	✓	x	x
YGL123W	RPS2	✓	x	x
YGL159W	YGL159W	✓	✓	✓
YGL191W	COX13	x	x	✓

YGL234W	ADE5,7	x	✓	x
YGR009C	SEC9	x	✓	✓
YGR034W	RPL26B	✓	x	x
YGR078C	PAC10	x	✓	x
YGR148C	RPL24B	✓	x	x
YGR172C	YIP1	x	✓	✓
YGR253C	PUP2	x	✓	x
YGR279C	SCW4	x	x	✓
YGR282C	BGL2	✓	x	✓
YHL032C	GUT1	x	✓	x
YHR021C	RPS27B	✓	x	x
YHR031C	RRM3	✓	✓	✓
YHR094C	HXT1	✓	✓	✓
YHR138C	YHR138C	x	x	✓
YHR141C	RPL42B	✓	x	x
YIL148W	RPL40A	✓	x	x
YIL169C	YIL169C	✓	✓	✓
YIR023W	DAL81	x	✓	✓
YJL129C	TRK1	✓	✓	✓
YJL177W	RPL17B	✓	x	✓
YJL200C	ACO2	x	✓	x
YJR006W	POL31	x	✓	✓
YJR095W	SFC1	x	✓	x
YJR124C	YJR124C	✓	✓	✓
YJR140C	HIR3	x	✓	✓
YKL056C	TMA19	✓	x	x
YKL082C	RRP14	x	x	✓
YKL152C	GPM1	✓	x	x
YKL180W	RPL17A	✓	x	x
YKL210W	UBA1	x	✓	x
YKL216W	URA1	x	✓	x
YKR059W	TIF1	✓	x	✓
YLR044C	PDC1	✓	x	x
Total No of Genes		50	50	50

Appendix E

Hierarchical Clustering Results: Continuous and Binary Data

Here we show hierarchical clustering results obtained by continuous and binarized transcriptome measurements (using B4 - gene by gene GMM threshold method) of RNA-Seq and microarray techniques.

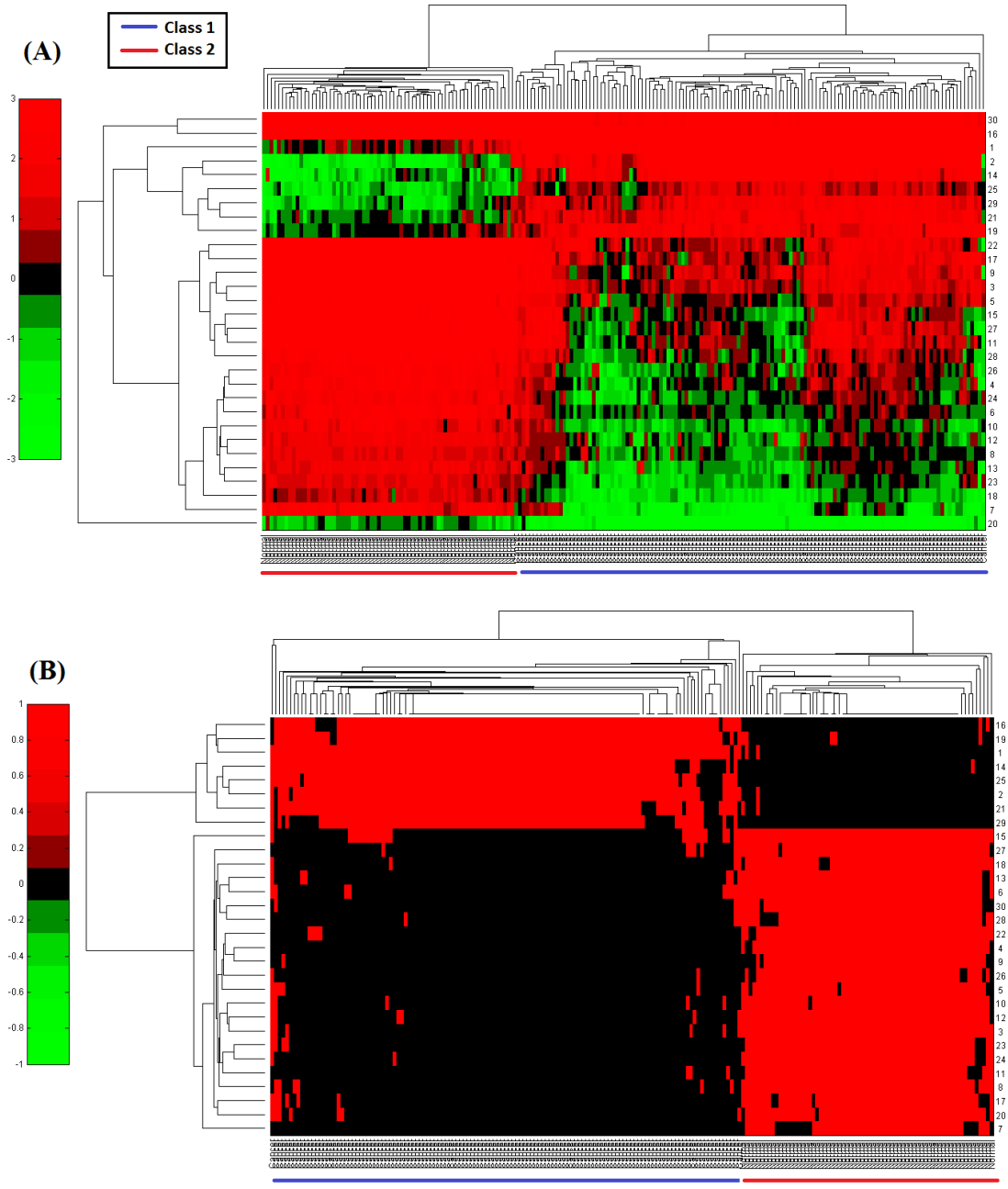


Figure E.1: RNA-Seq Breast Cancer - Hierarchical clustering using top 30 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

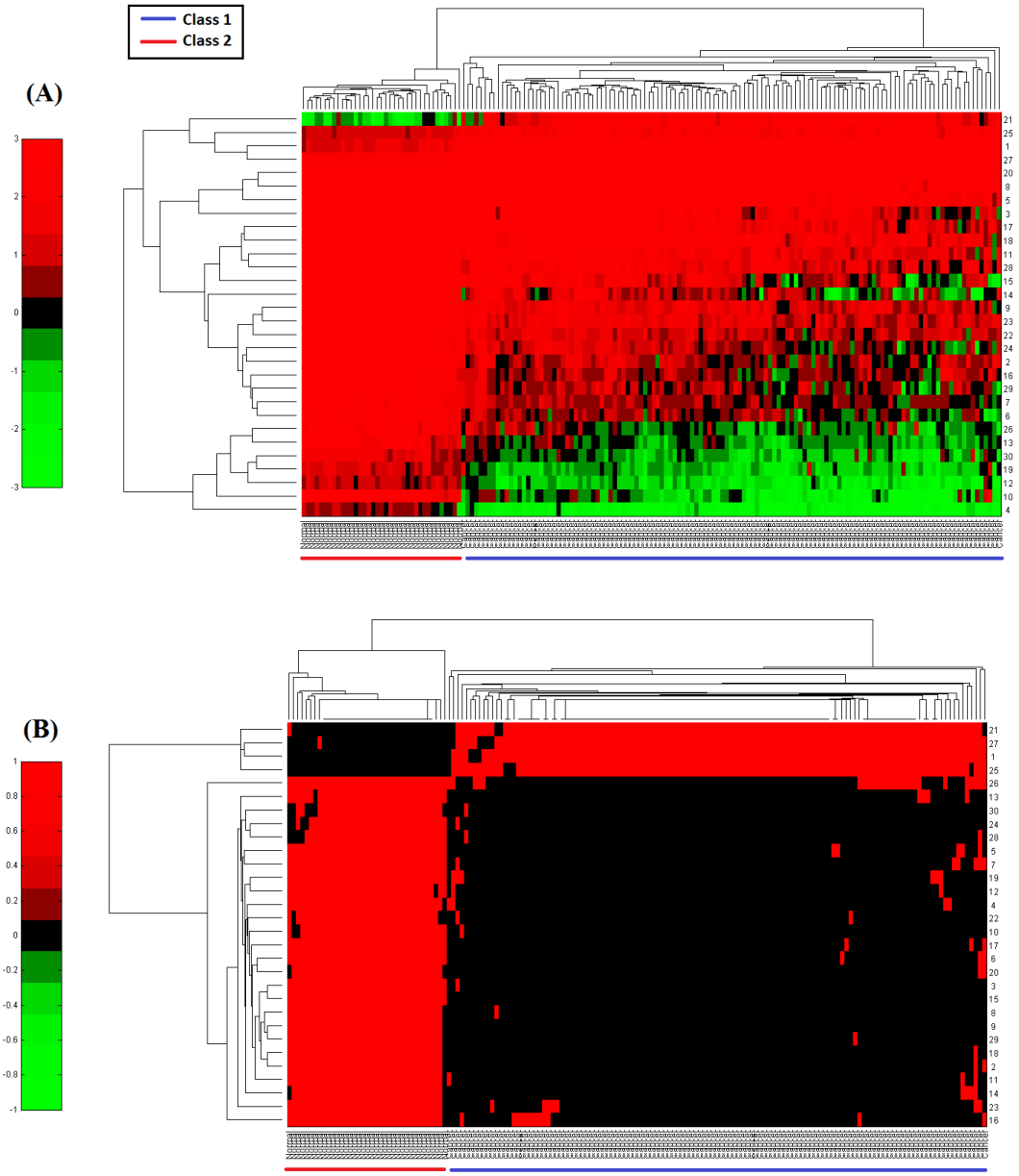


Figure E.2: RNA-Seq Lung Cancer - Hierarchical clustering using top 30 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

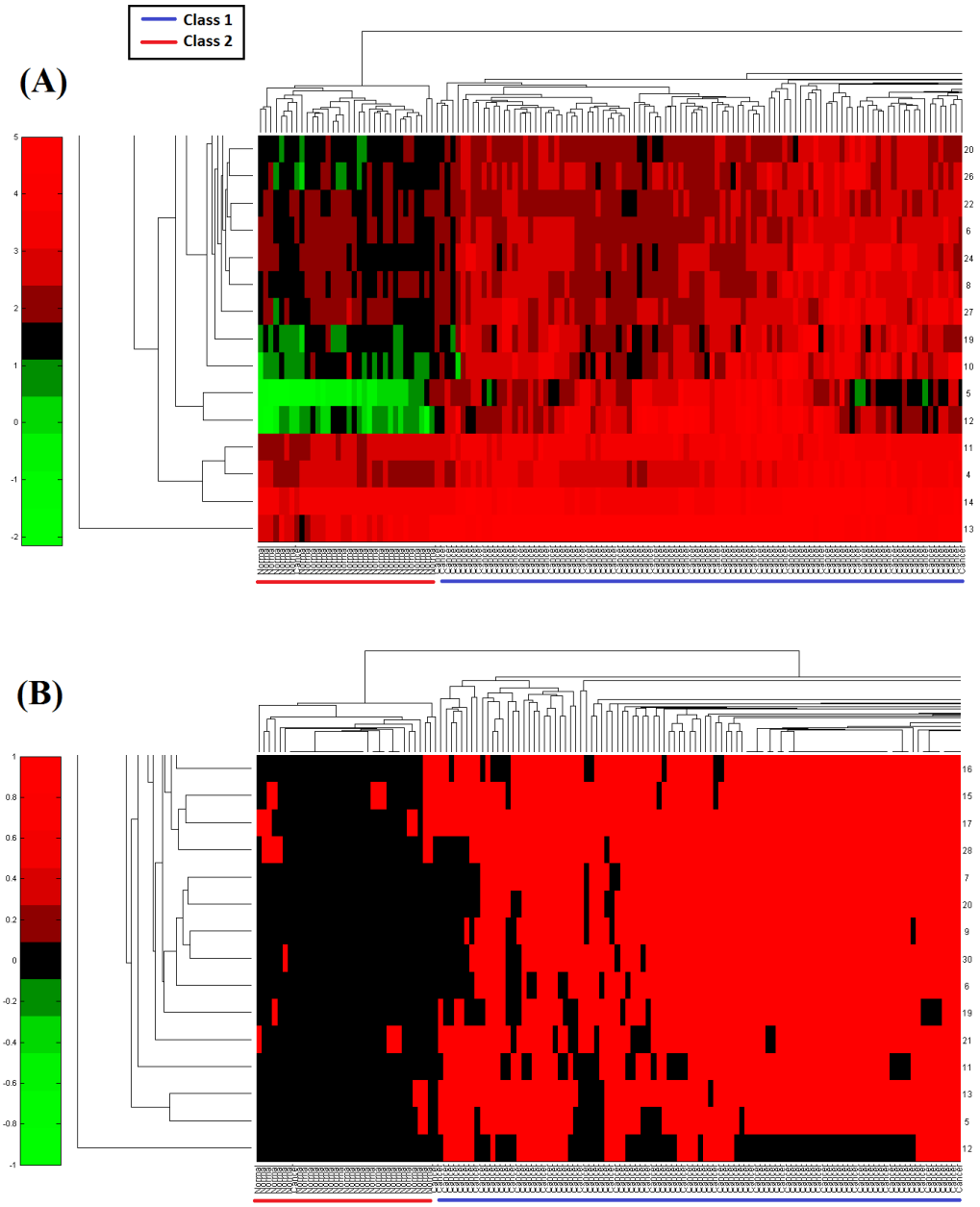


Figure E.3: RNA-Seq Stomach Cancer - Hierarchical clustering using top 30 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively..

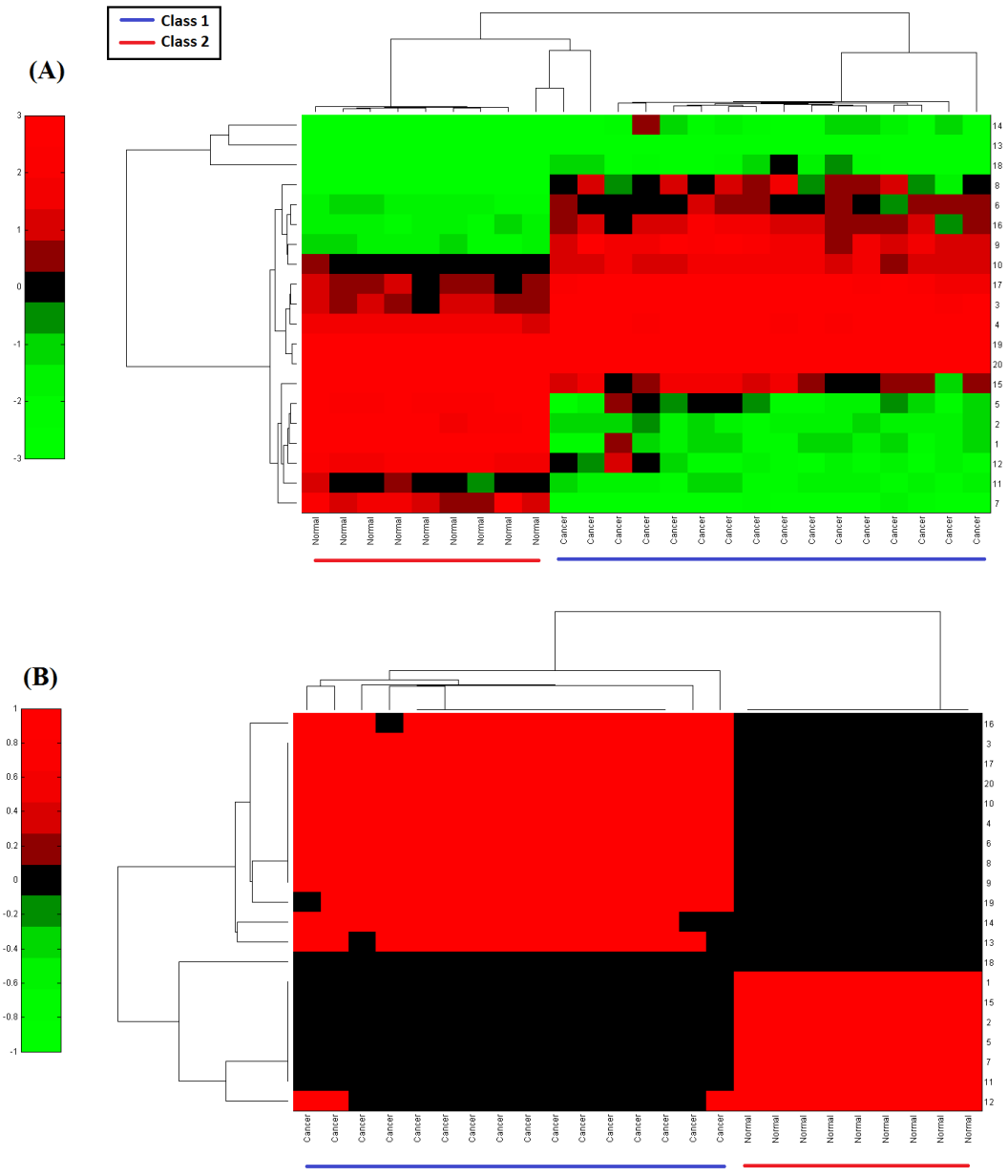


Figure E.4: RNA-Seq Liver Cancer - Hierarchical clustering using top 20 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

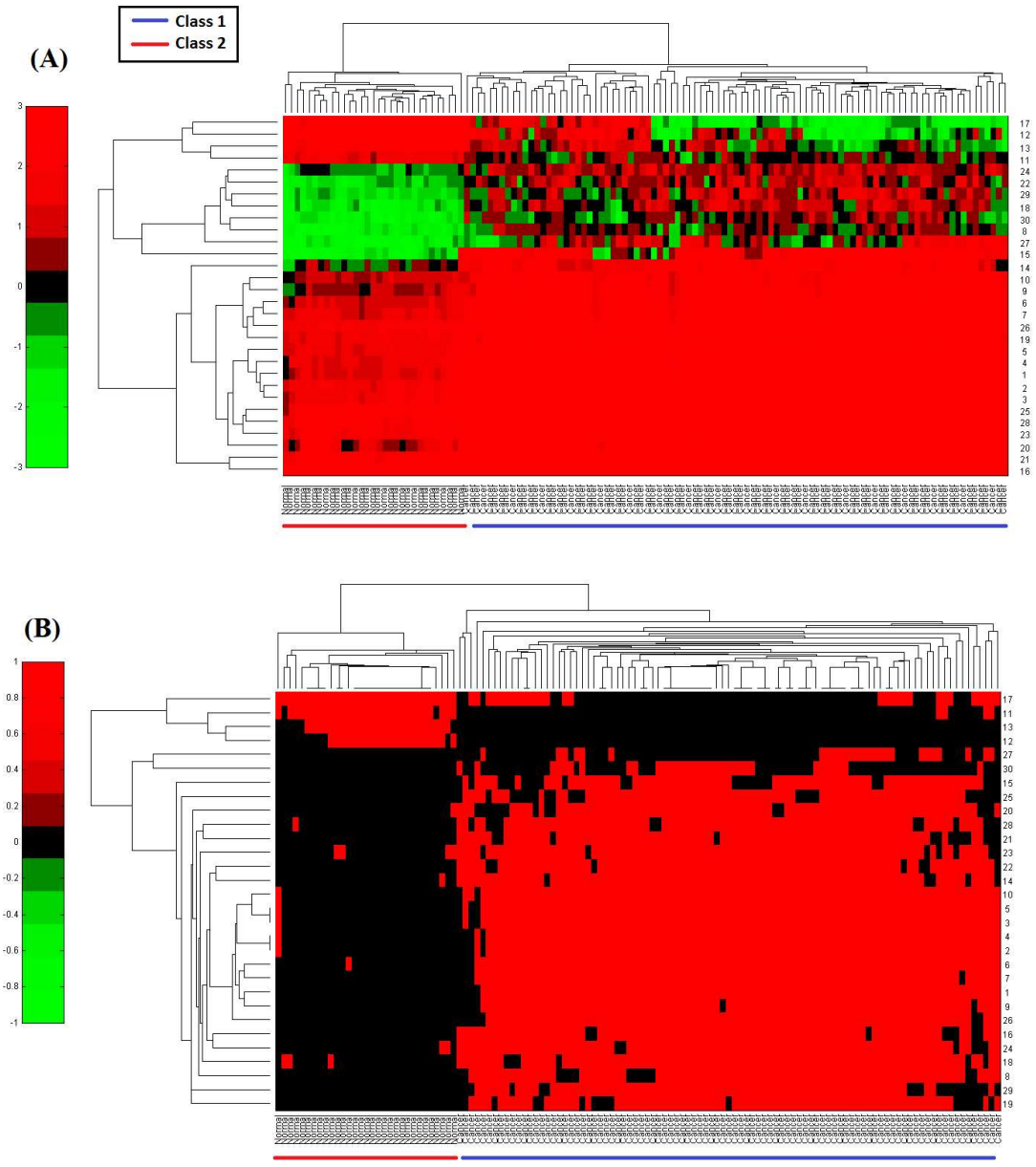


Figure E.5: RNA-Seq Head and Neck Cancer - Hierarchical clustering using top 30 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

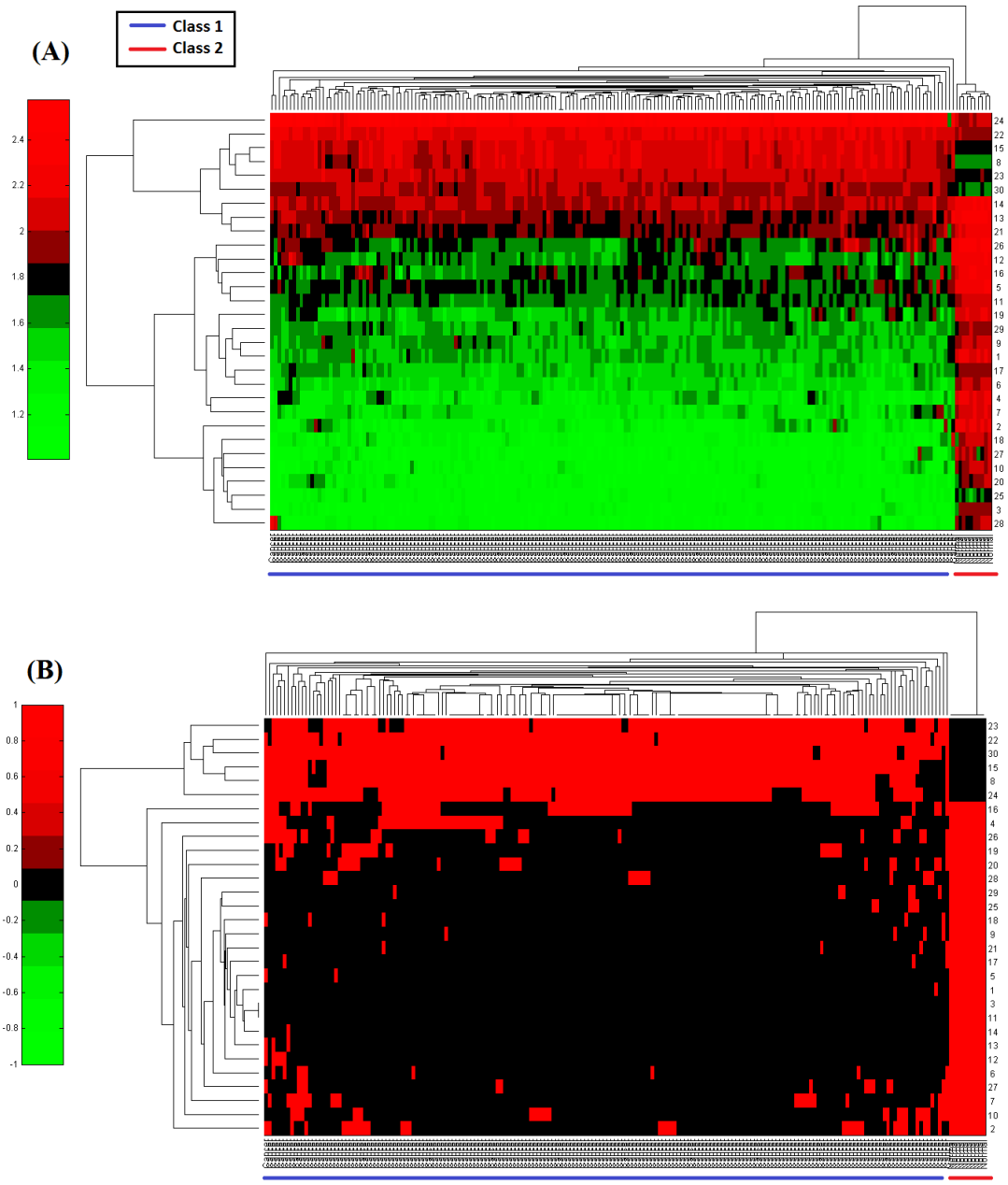


Figure E.6: Microarray Ovarian Cancer - Hierarchical clustering using top 30 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

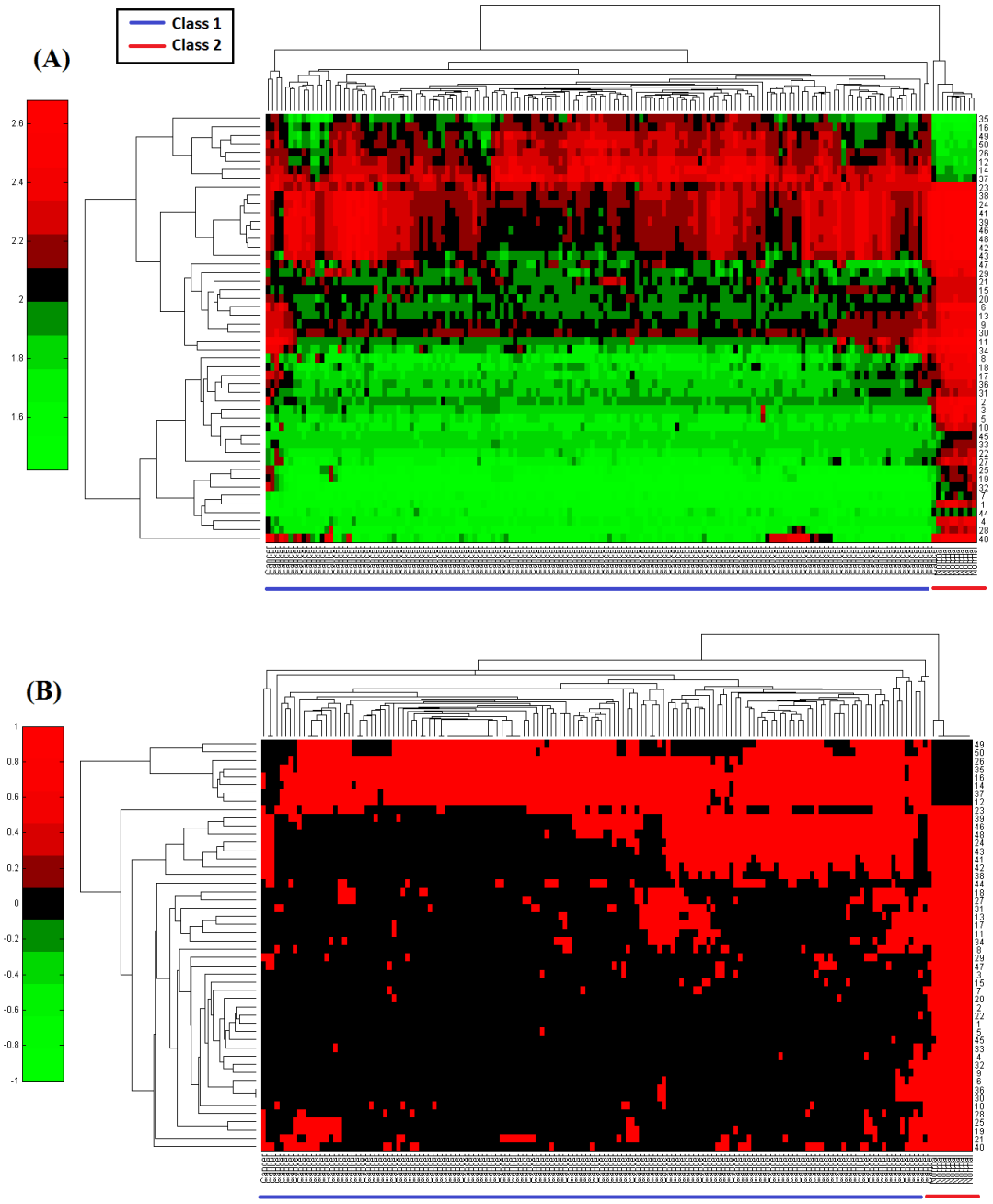


Figure E.7: Microarray Soft Tissue Cancer - Hierarchical clustering using top 50 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups.

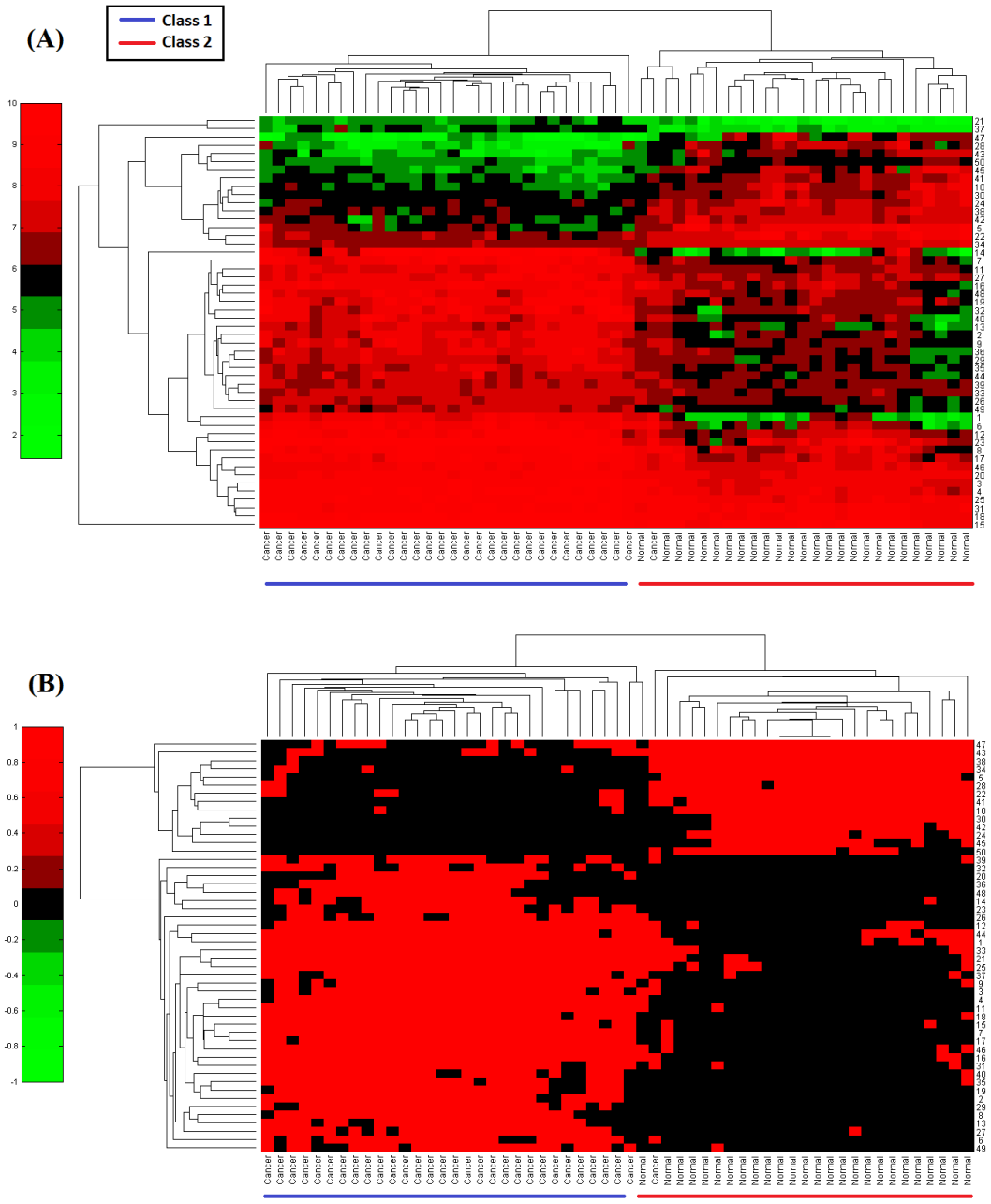


Figure E.8: Microarray Head and Neck Cancer - Hierarchical clustering using top 50 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

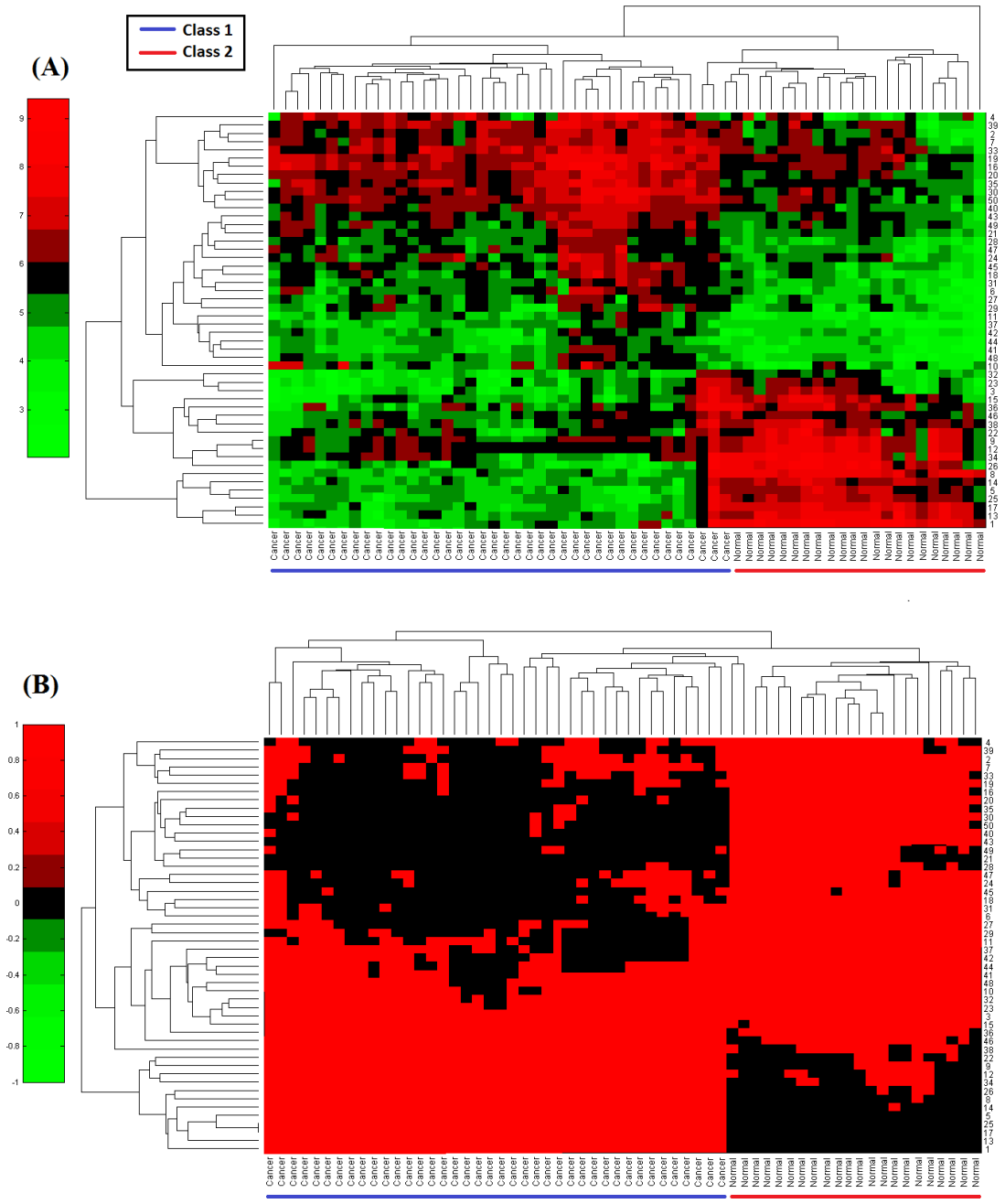


Figure E.9: Microarray Colon Cancer - Hierarchical clustering using top 50 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

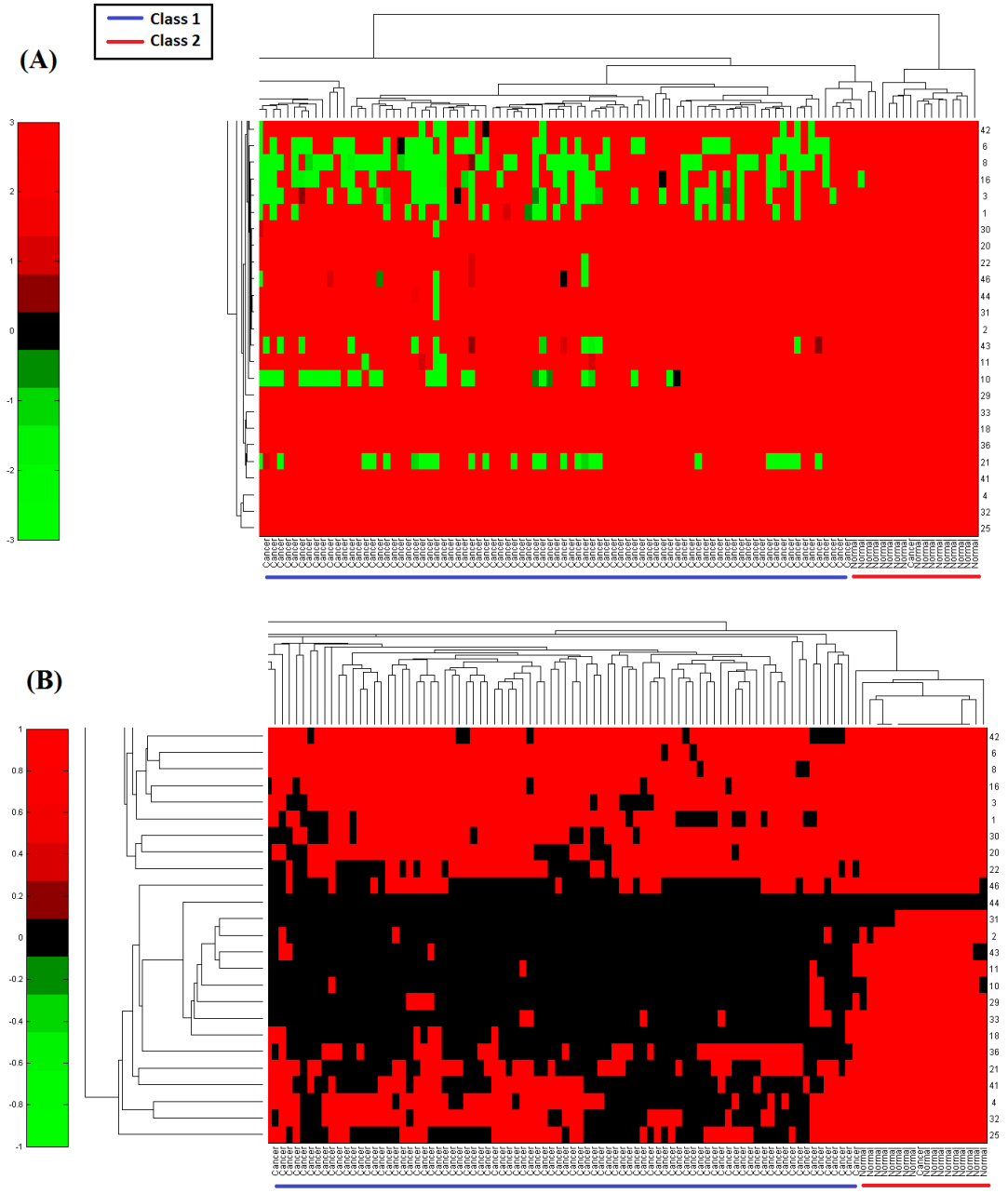


Figure E.10: Microarray Lung Cancer - Hierarchical clustering using top 50 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

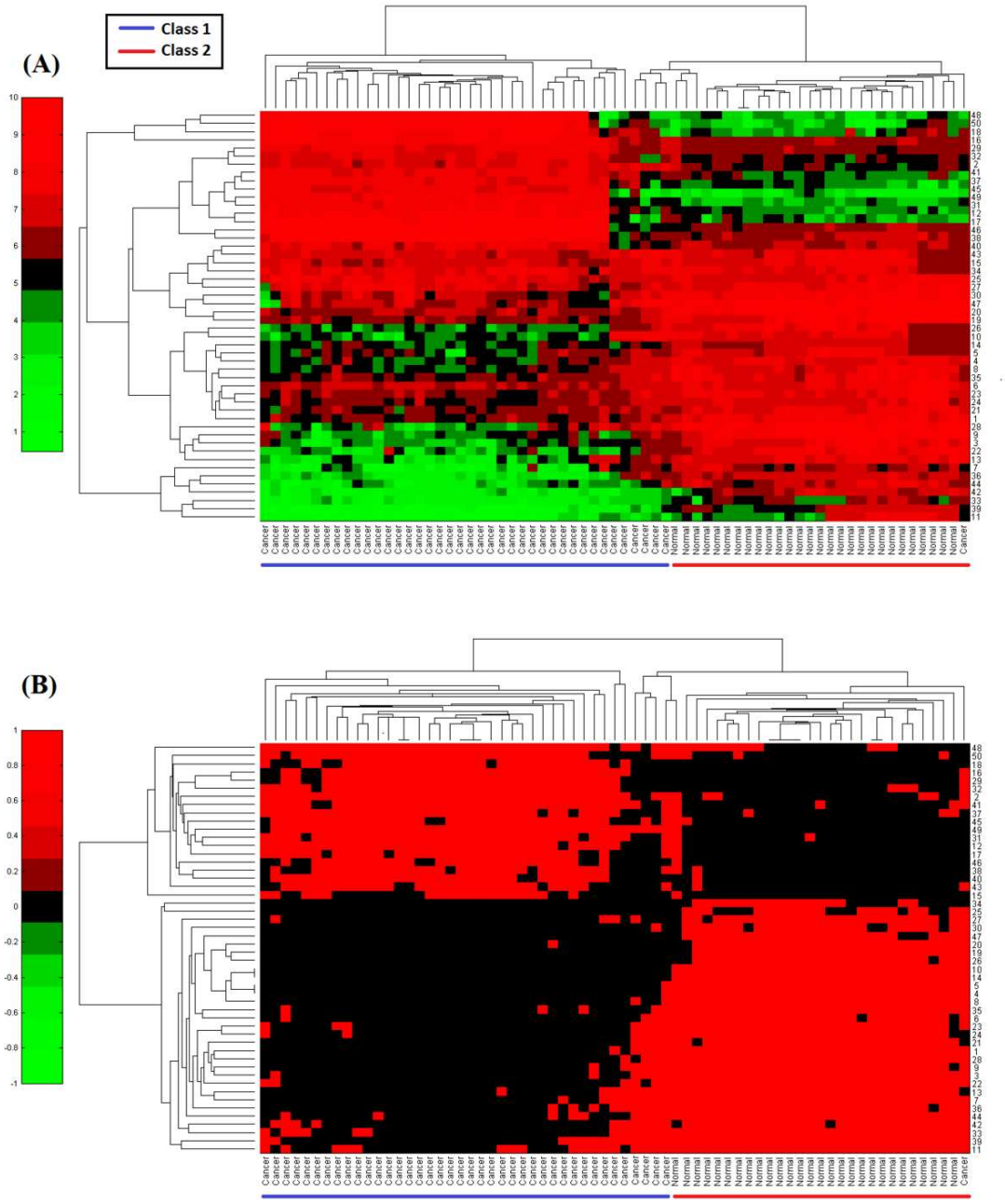


Figure E.11: Microarray Stomach Cancer - Hierarchical clustering using top 50 genes. (A) shows the continuous data clustering (B) shows binarized data using B4 (gene by gene GMM threshold binarization) clustering. Both continuous and binary data clustered cancer and normal patients into two groups. Class 1 and class 2 have majority patients with cancer and majority patients with no cancer (normal) respectively.

Appendix F

Gene Ontology Scatter Plots: Continuous and Binary Data

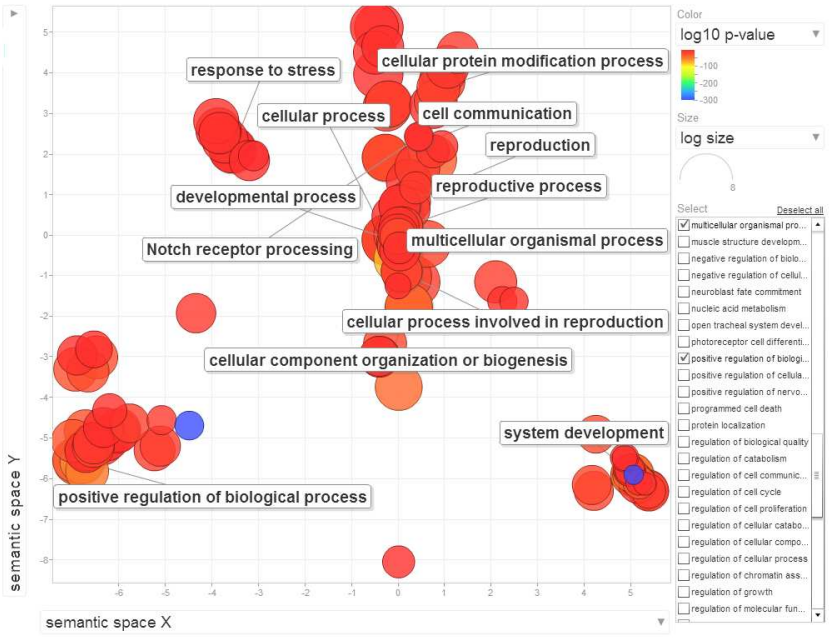
F.1 GO Scatter Plots for Times Points with Highest Number of Up/Down Regulated Proteins

REVIGO (Supek et al., 2011) gene ontology visualization plots for all the GO terms identified by the GO analysis for the time points with the highest number of up or down regulated proteins using both RNA-Seq and microarray measurements under *Drosophila melanogaster*'s developmental time course are shown here. B4 gene by gene GMM threshold binarization technique was employed to convert continuous data into binary in following experiments.

F.2 GO Scatter Plots for Random Times Points of Up/-Down Regulated Proteins

Following figures show the gene ontology scatter plots generated by REVIGO (Supek et al., 2011) web tool for continuous and binary (using B4) data by simply considering random time points of RNA-Seq and microarray *Drosophila melanogaster* development time course.

(A)



(B)

Figure F.1: GO terms scatter plots of (A) continuous and (B) binary RNA-Seq measurements obtained at the highest number of down regulated genes detected time point of the *Drosophila melanogaster*'s developmental time course Graveley et al. (2011)

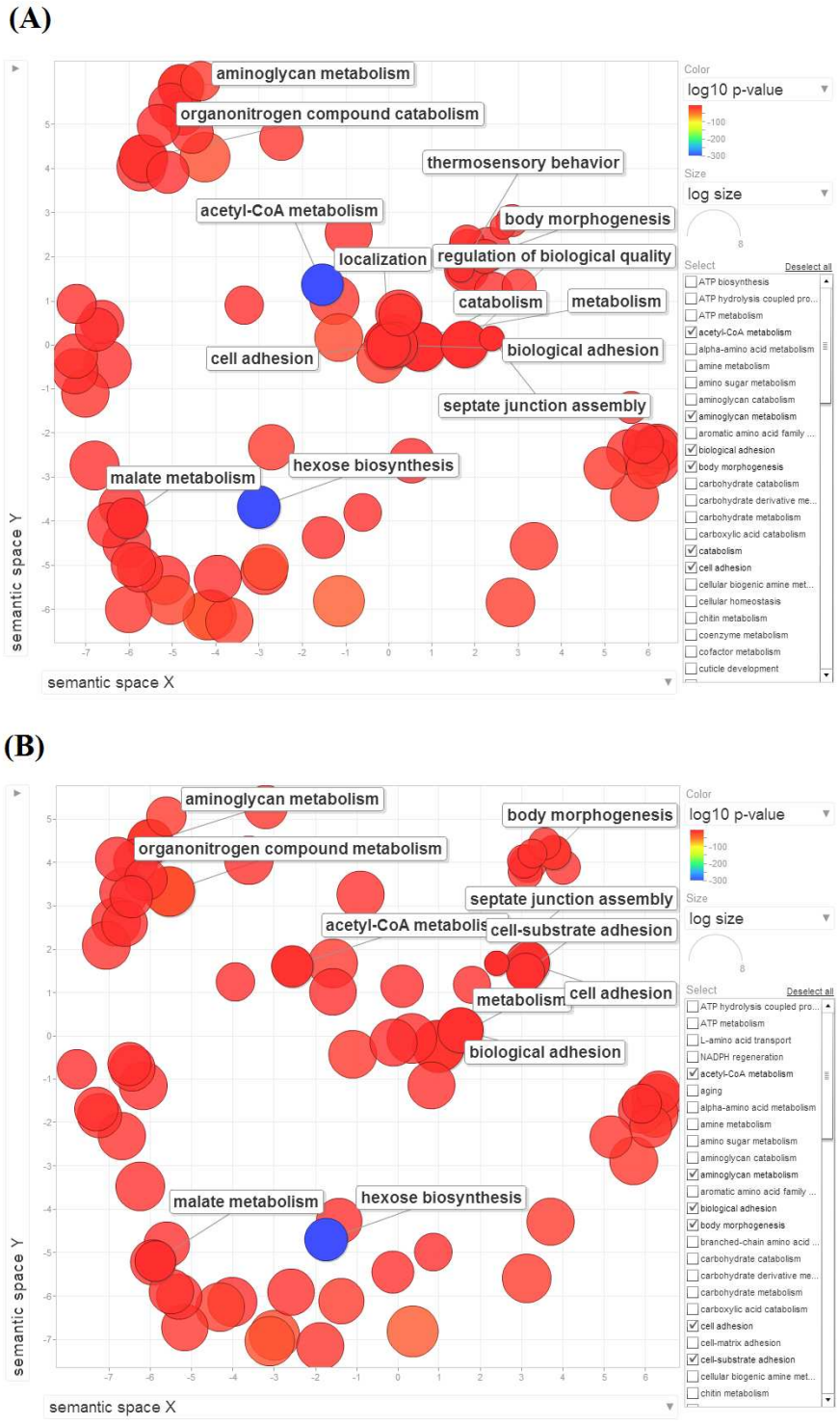


Figure F.2: GO terms scatter plots of (A) continuous and (B) binary microarray measurements obtained at the highest number of up regulated genes detected time point of the *Drosophila melanogaster*'s developmental time course Hooper et al. (2007)

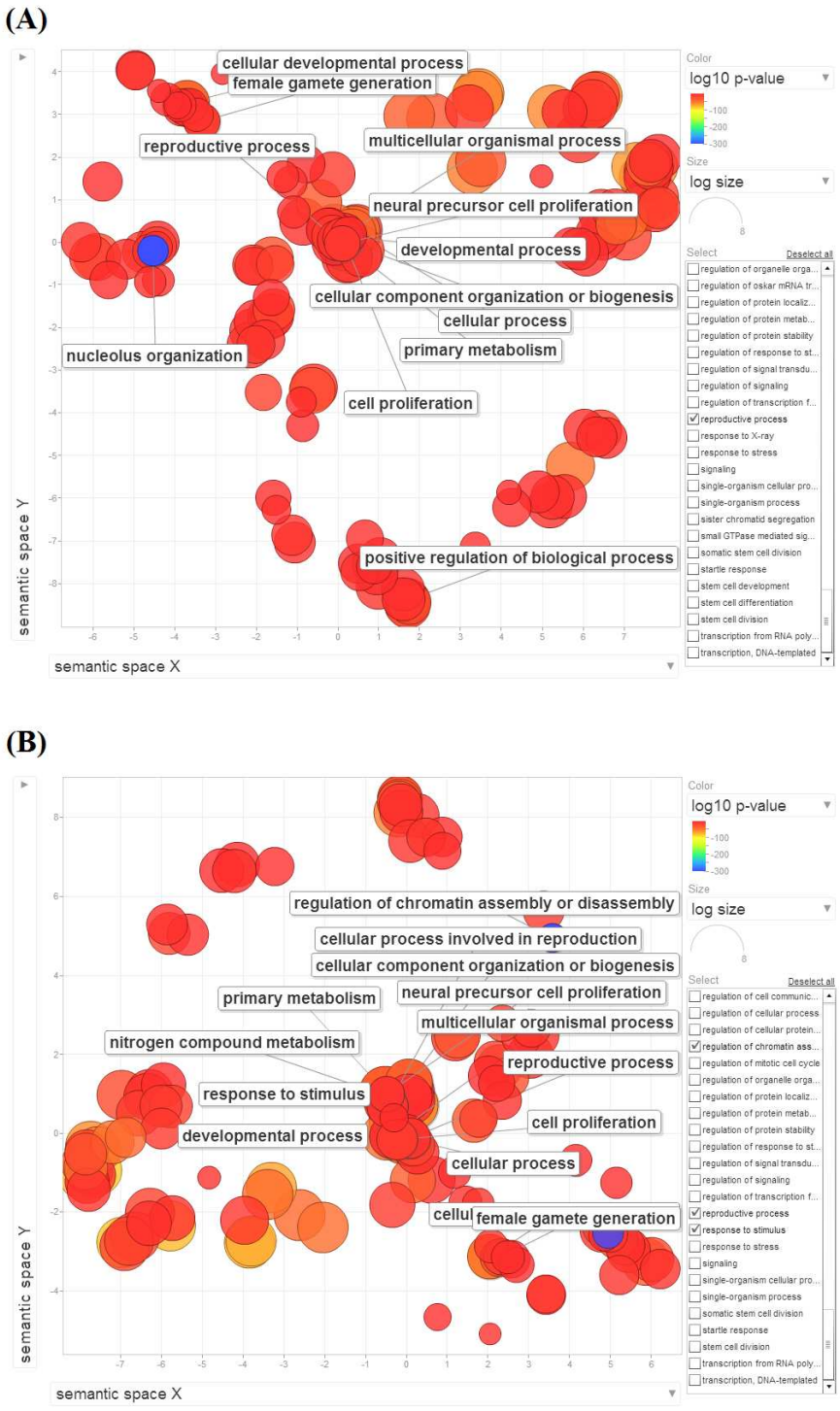


Figure F.3: GO terms scatter plots of (A) continuous and (B) binary microarray measurements obtained at the highest number of down regulated genes detected time point of the *Drosophila melanogaster*'s developmental time course Hooper et al. (2007)

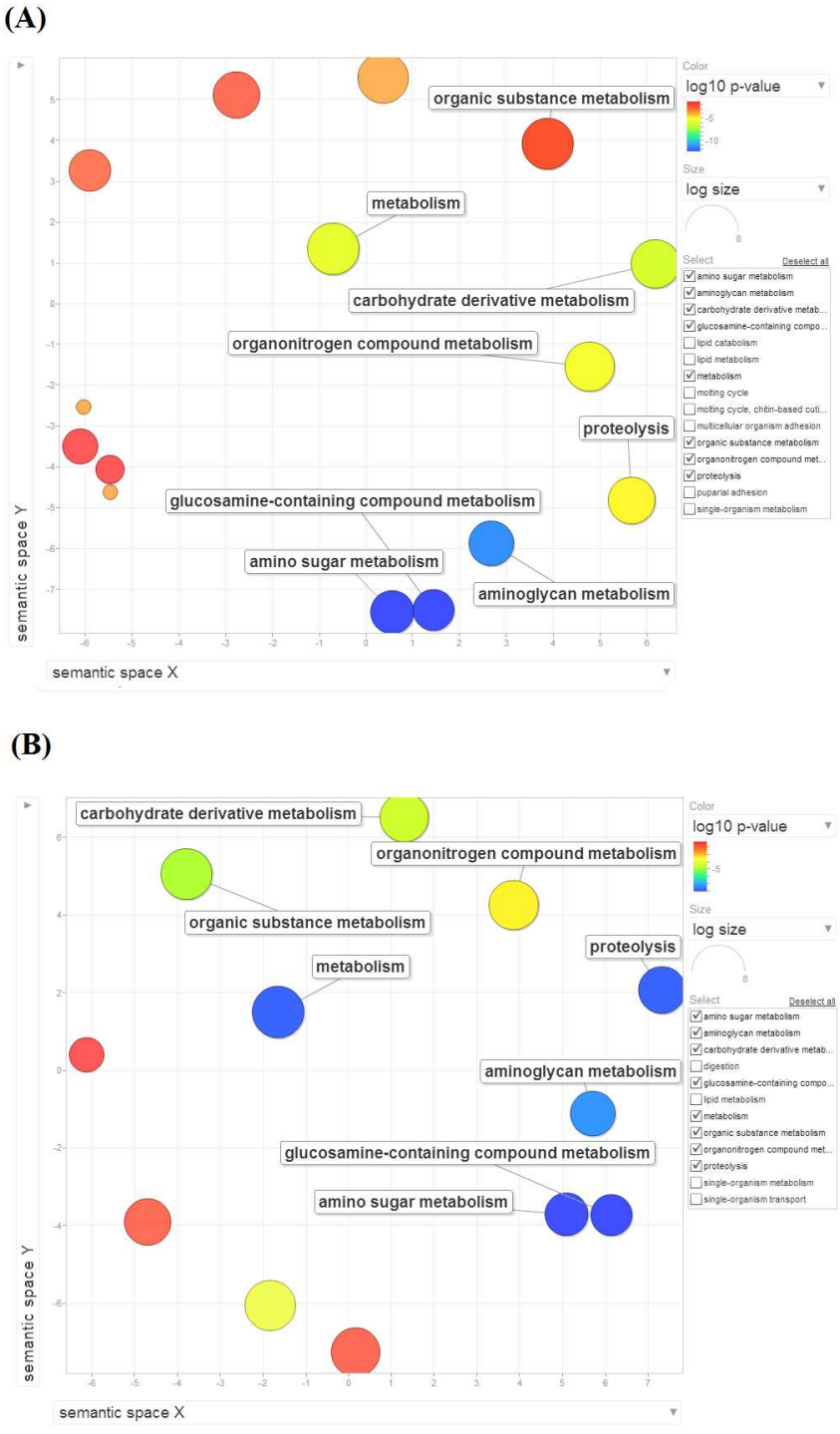


Figure F.4: GO terms scatter plots of (A) continuous and (B) binary RNA-Seq measurements obtained at a random time point with up regulated genes of *Drosophila melanogaster*'s developmental time course Graveley et al. (2011)

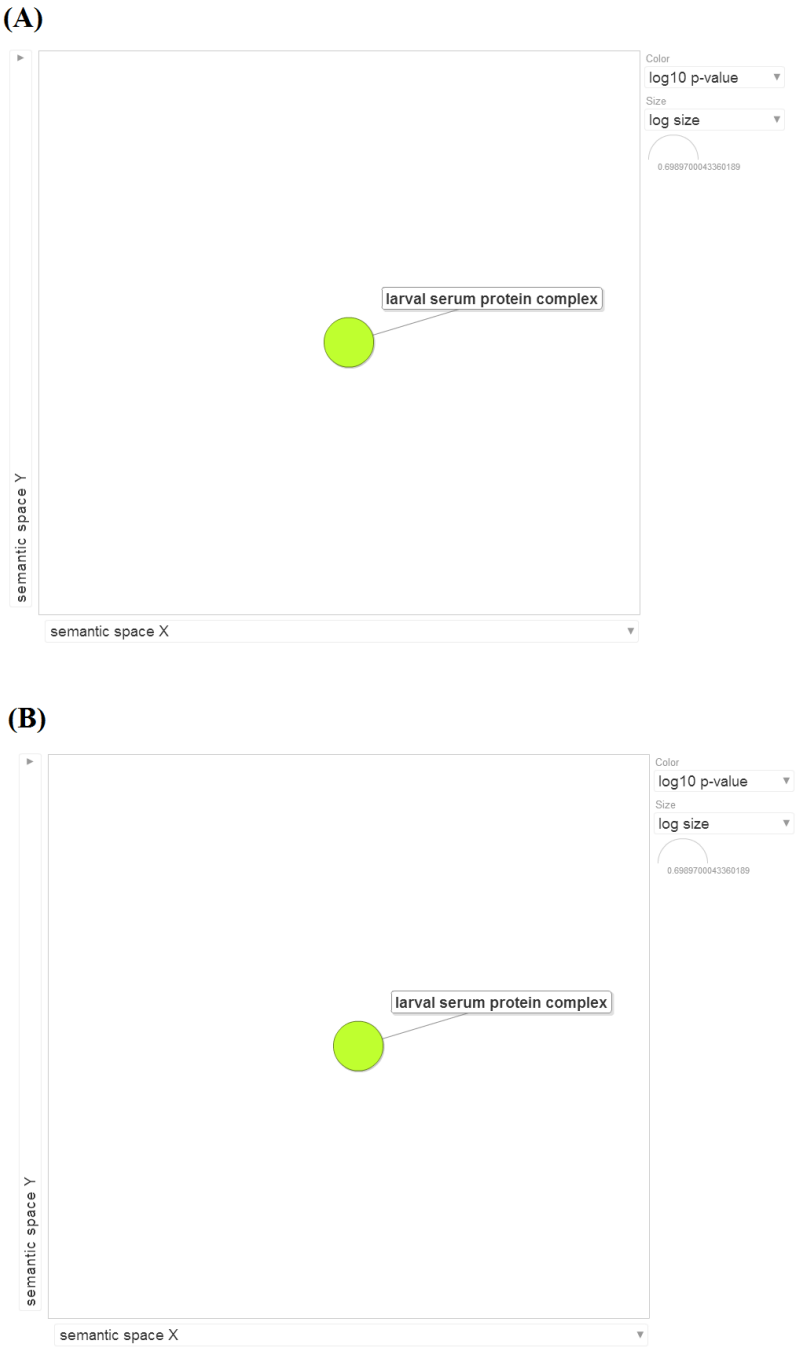


Figure F.5: GO terms scatter plots of (A) continuous and (B) binary RNA-Seq measurements obtained at a random time point with down regulated genes of *Drosophila melanogaster*'s developmental time course Graveley et al. (2011)

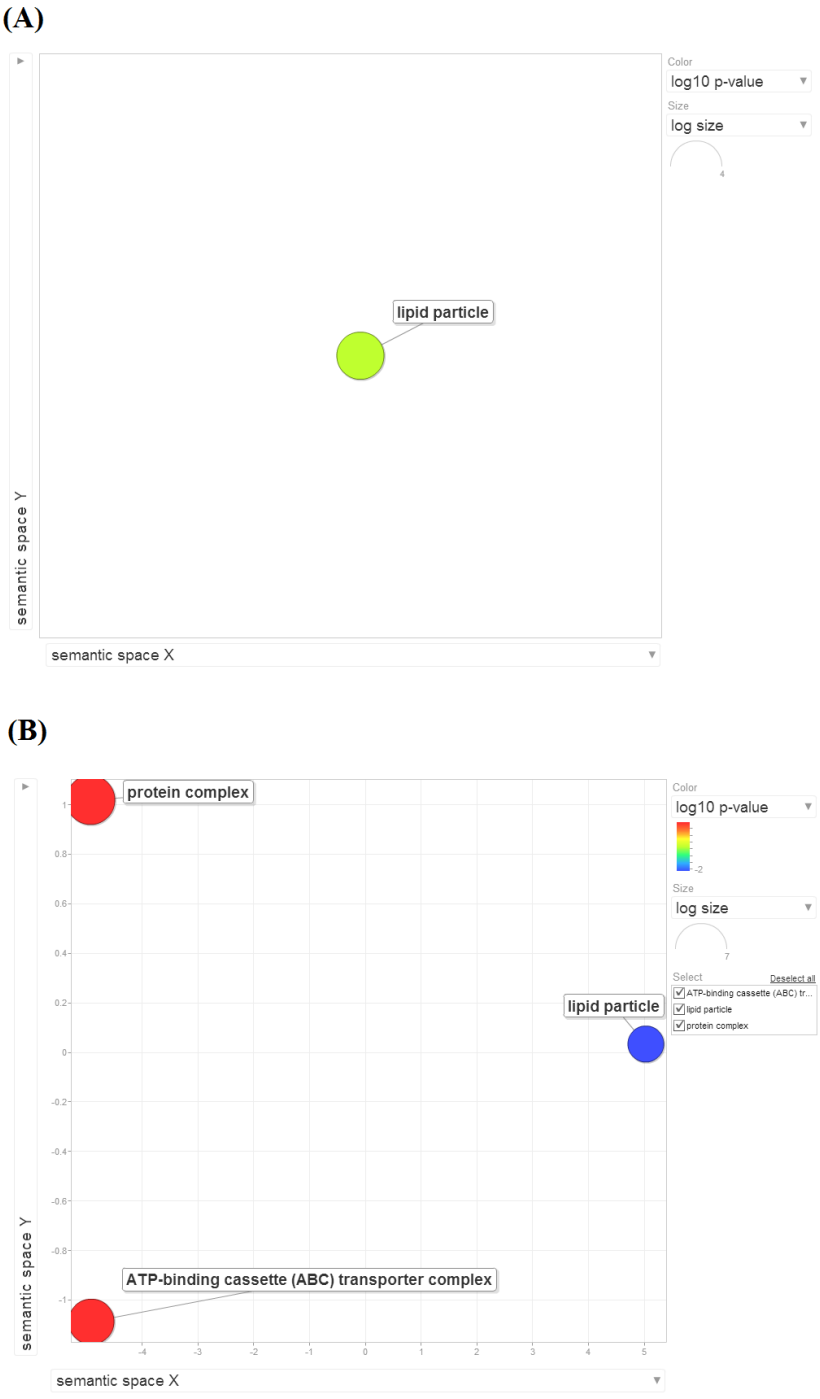


Figure F.6: GO terms scatter plots of (A) continuous and (B) binary microarray measurements obtained at a random time point with up regulated genes of *Drosophila melanogaster*'s developmental time course Hooper et al. (2007)

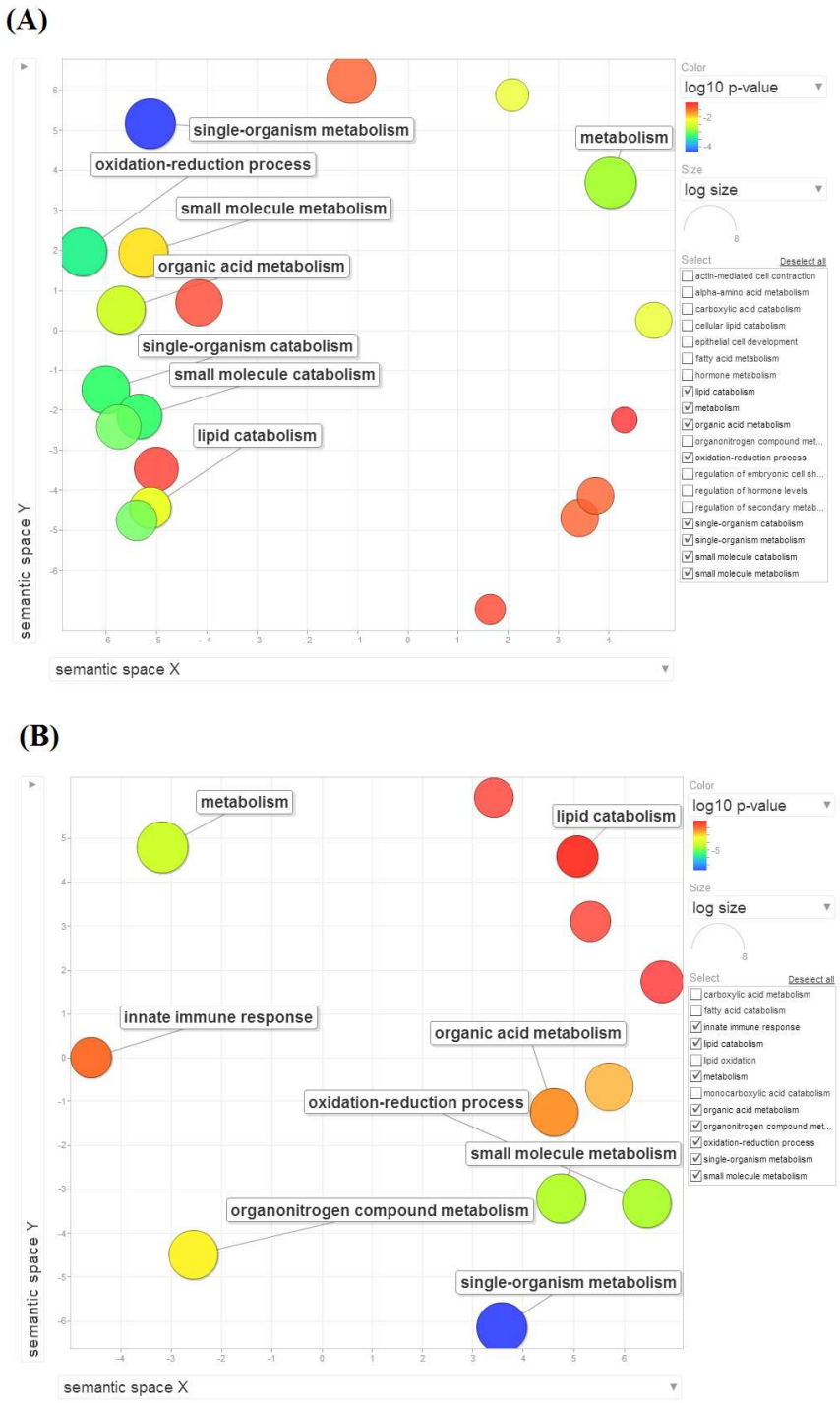


Figure F.7: GO terms scatter plots of (A) continuous and (B) binary microarray measurements obtained at a random time point with down regulated genes of *Drosophila melanogaster*'s developmental time course Hooper et al. (2007)

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750.

Anderson, J. R., Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (1986). *Machine learning: An artificial intelligence approach*, volume 2.

Arava, Y., Wang, Y., Storey, J., Liu, C., Brown, P., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 100(7):3889–3894.

Arbeitman, M. N., Furlong, E. E. M., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., and White, K. P. (2002). Gene expression during the life cycle of drosophila melanogaster. *Science*, 297(5590):2270–2275.

Arnott, D., Gawinowicz, M. A., Grant, R. A., Neubert, T. A., Packman, L. C., Speicher, K. D., Stone, K., and Turck, C. W. (2003). ABRF-PRG03: phosphorylation site determination. *Journal of biomolecular techniques: JBT*, 14(3):205.

Atkinson, A. and Riani, M. (2000). *Robust diagnostic regression analysis*. Springer.

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

Barrera, L. O., Ren, B., et al. (2006). The transcriptional regulatory code of eukaryotic cells? insights from genome-wide analysis of chromatin organization and transcription factor binding. *Current opinion in cell biology*, 18(3):291–298.

Barretina, J., Taylor, B. S., Banerji, S., Ramos, A. H., Lagos-Quintana, M., DeCarolis, P. L., Shah, K., Socci, N. D., Weir, B. A., Ho, A., et al. (2010). Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nature genetics*, 42(8):715–721.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297.

Belle, A., Tanay, A., Bitincka, L., Shamir, R., and OShea, E. K. (2006). Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences*, 103(35):13004–13009.

Benson, L. J., Gu, Y., Yakovleva, T., Tong, K., Barrows, C., Strack, C. L., Cook, R. G., Mizzen, C. A., and Annunziato, A. T. (2006). Modifications of H3 and H4 during chromatin replication, nucleosome assembly, and histone exchange. *Journal of Biological Chemistry*, 281(14):9287–9296.

Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). Biochemistry: International version (hardcover).

Beyer, A., Hollunder, J., Nasheuer, H., and Wilhelm, T. (2004). Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Molecular and Cellular Proteomics*, 3(11):1083–1092.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24):13790–13795.

Bishop, C. M. (1995). *The Multi-layer Perceptron*. Oxford University Press, UK.

Block, T. M., Comunale, M. A., Lowman, M., Steel, L. F., Romano, P. R., Fimmel, C., Tennant, B. C., London, W. T., Evans, A. A., Blumberg, B. S., et al. (2005). Use of targeted glycoproteomics to identify serum glycoproteins that correlate with liver cancer in woodchucks and humans. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):779–784.

Bode, A. and Dong, Z. (2004). Post-translational modification of p53 in tumorigenesis. *Nature Reviews Cancer*, 4(10):793–805.

Bodner, S. M., Minna, J. D., Jensen, S. M., D'amico, D., Carbone, D., Mitsudomi, T., Fedorko, J., Buchhagen, D. L., Nau, M. M., and Gazdar, A. F. (1992). Expression of mutant p53 proteins in lung cancer correlates with the class of p53 gene mutation. *Oncogene*, 7(4):743–749.

Bonome, T., Levine, D. A., Shih, J., Randonovich, M., Pise-Masison, C. A., Bogomolniy, F., Ozbun, L., Brady, J., Barrett, J. C., Boyd, J., et al. (2008). A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer research*, 68(13):5478–5486.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorg, W., Ball, C., Causton, H., Gaasterland, T., Glenisson, P., Holstege, F., Kim, I., Markowitz, V., Matese, J., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (miame)toward standards for microarray data. *Nature Genetics*, 29:365–371.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brockmann, R., Beyer, A., Heinisch, J., and Wilhelm, T. (2007). Posttranscriptional expression regulation: What determines translation rates? *PLoS computational biology*, 3(3):e57.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000a). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000b). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267.

Brunet, J., Tamayo, P., Golub, T. R., and Mesirov, J. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169.

Burnett, G. and Kennedy, E. P. (1954). The enzymatic phosphorylation of proteins. *Journal of Biological Chemistry*, 211(2):969–980.

Burton, J. L. and Solomon, M. J. (2001). D box and KEN box motifs in budding yeast Hsl1p are required for APC-mediated degradation and direct binding to Cdc20p and Cdh1p. *Genes & development*, 15(18):2381–2395.

Butler, J. M. (2005). *Forensic DNA typing: biology, technology, and genetics of STR markers*.

Butz, K., Shahabeddin, L., Geisen, C., Spitkovsky, D., Ullmann, A., and Hoppe-Seyler, F. (1995). Functional p53 protein in human papillomavirus-positive cancer cells. *Oncogene*, 10(5):927–936.

Callis, J. (1995). Regulation of protein degradation. *The Plant Cell*, 7(7):845.

Carroll, A., Heazlewood, J., Ito, J., and Millar, A. (2008). Analysis of the arabidopsis cytosolic ribosome proteome provides detailed insights into its components and their post-translational modification. *Molecular & cellular proteomics*, 7(2):347–369.

Causton, H., Quackenbush, J., and Brazma, A. (2009). *Microarray gene expression data analysis: a beginner's guide*.

Chan, E. T., Quon, G. T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R. A., Aubin, J., Ratcliffe, M. J., Wilde, A., Brudno, M., et al. (2009). Conservation of core gene expression in vertebrate tissues. *J Biol*, 8(3):33.

Chandramouli, K. and Qian, P. (2009). Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Human Genomics and Proteomics*, 1(1):239204.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 853–867.

Chen, K., Calzone, L., Csikasz-Nagy, A., Cross, F., Novak, B., and Tyson, J. (2004). Integrative analysis of cell cycle control in budding yeast. *Molecular Biology of the Cell*, 15(8):3841–3862.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S., and Wong, E. D. (2012). Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):D700–D705.

Chung, F. R. K. (1997). *Spectral graph theory*, volume 92.

Cipolla, R., Battiato, S., and Farinella, G. M. (2012). *Machine Learning for Computer Vision*.

Clark, J. (2015). The mass spectrometer. Available online at http://www.chemguide.co.uk/analysis/masspec/howitworks.html.

Cohen, P. (2000). The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends in biochemical sciences*, 25(12):596–601.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine*, 11(10):1305–1319.

Colin, C. (2002). Robust regression and outlier detection with the robustreg procedure. *SUGI Paper 265-27*.

Collobert, R., Sinz, F., Weston, J., and Bottou, L. (2006). Trading convexity for scalability. In *International Conference on Machine Learning*, pages 129–136.

Core, L. J. and Lis, J. T. (2008). Transcription regulation through promoter-proximal pausing of rna polymerase ii. *Science*, 319(5871):1791–1792.

Cortes, C. and Vapnik, V. (1995a). Support-vector networks. *Machine learning*, 20(3):273–297.

Cortes, C. and Vapnik, V. (1995b). Support-vector networks. *Machine learning*, 20(3):273–297.

Crick, F. et al. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*.

Da W. H., B. T. S., Lempicki, R. A., et al. (2008). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57.

Dang, W., Sutphin, G. L., Dorsey, J. A., Otte, G. L., Cao, K., Perry, R. M., Wanat, J. J., Saviolaki, D., Murakami, C. J., Tsuchiyama, S., et al. (2014). Inactivation of yeast isw2 chromatin remodeling enzyme mimics longevity effect of calorie restriction via induction of genotoxic stress response. *Cell metabolism*, 19(6):952–966.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977a). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977b). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

D'Errico, M., de Rinaldis, E., Blasi, M. F., Viti, V., Falchetti, M., Calcagnile, A., Sera, F., Saieva, C., Ottini, L., Palli, D., et al. (2009). Genome-wide expression profile of sporadic gastric cancers with microsatellite instability. *European Journal of Cancer*, 45(3):461–469.

Dubey, H. and Grover, A. (2001). Current initiatives in proteomics research: the plant perspective. *Current Science*, 80(2):262–270.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *Systems, Man and Cybernetics, IEEE Transactions on*, (4):325–327.

Dyrskjøt, L., Kruhøffer, M., Thykjaer, T., Marcussen, N., Jensen, J. L., Møller, K., and Ørntoft, T. F. (2004). Gene expression in the urinary bladder a common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer Research*, 64(11):4040–4048.

Ehrlich, M. (2002). Dna methylation in cancer: too much, but also too little. *Oncogene*, 21(35):5400–5413.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

Esposito, F. and Malerba, D. (2001). Machine learning in computer vision. *Applied Artificial Intelligence*, 15(8):693–705.

Estilo, C. L., Pornchai, O., Talbot, S., Socci, N. D., Carlson, D. L., Ghossein, R., Williams, T., Yonekawa, Y., Ramanathan, Y., Boyle, J. O., et al. (2009). Oral tongue cancer gene expression profiling: Identification of novel potential prognosticators by oligonucleotide microarray analysis. *BMC cancer*, 9(1):11.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., et al. (2009). Estimating accuracy of rna-seq and microarrays with proteomics. *BMC genomics*, 10(1):161.

Futcher, B., Latter, G. I., Monardo1, P., McLaughlin, C. S., and Garrels, J. I. (1999). A sampling of the yeast proteome. *Molecular and Cellular Biology*, 19(11):7357–7368.

García-Alai, M. M., Gallo, M., Salame, M., Wetzler, D. E., McBride, A. A., Paci, M., Cicero, D. O., and de Prat-Gay, G. (2006). Molecular basis for phosphorylation-dependent, pest-mediated protein turnover. *Structure*, 14(2):309–319.

Gariglio, P., Bellard, M., and Chambon, P. (1981). Clustering of rna polymerase b molecules in the 5 moiety of the adult β-globin gene of hen erythrocytes. *Nucleic acids research*, 9(11):2589–2598.

Glozak, M. A., Sengupta, N., Zhang, X., and Seto, E. (2005). Acetylation and deacetylation of non-histone proteins. *Gene*, 363:15–23.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999a). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999b). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537.

Gong, C., Liu, F., Grundke-Iqbal, I., and Iqbal, K. (2005). Post-translational modifications of tau protein in Alzheimers disease. *Journal of neural transmission*, 112(6):813–838.

Gonze, D., Halloy, J., and Goldbeter, A. (2002). Robustness of circadian rhythms with respect to molecular noise. *Proceedings of the National Academy of Sciences*, 99(2):673–678.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., et al. (2011). The developmental transcriptome of drosophila melanogaster. *Nature*, 471(7339):473–479.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, 4(9):117.

Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., and Aebersold, R. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Molecular and Cellular Proteomics*, 1(4):323–333.

Grolleau, A., Bowman, J., Pradet-Balade, B., Puravs, E., Hanash, S.and Garcia-Sanz, J. A., and Beretta, L. (2002). Global and specific translational control by rapamycin in t cells uncovered by microarrays and proteomics. *Journal of Biological Chemistry*, 277(25):22175–22184.

Gulmann, C., Espina, V., Petricoin, E., Longo, D. L., Santi, M., Knutsen, T., Raffeld, M., Jaffe, E. S., Liotta, L. A., and Feldman, A. L. (2005). Proteomic analysis of apoptotic pathways reveals prognostic factors in follicular lymphoma. *Clinical Cancer Research*, 11(16):5847–5855.

Gygi, S., Rochon, Y., Franza, B., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, 19(3):1720–1730.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 761–771.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato.

Harbison, C., Gordon, D., Lee, T., Rinaldi, N., Macisaac, K., Danford, T., Hannett, N., Tagne, J., Reynolds, D., J., Y., Jennings, E., Zeitlinger, J., Pokholok, D., Kellis, M., Rolfe, P., Takusagawa, K., Lander, E., Gifford, D., Fraenkel, E., and Young, R. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102.

Hartmann-Petersen, R. and Gordon, C. (2004). Proteins interacting with the 26S proteasome. *Cellular and molecular life sciences: CMLS*, 61(13):1589–1595.

Hawkins, D. M., Bradu, D., and Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, 26(3):197–208.

Heagerty, P. J. and Pepe, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):533–551.

Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J., and Dimopoulos, G. (2005). Bayesian coclustering of anopheles gene expression time series: study of immune defense response to multiple experimental challenges. *Proceedings of the National Academy of Sciences of the United States of America*, 102(47):16939–16944.

Hendricks, W. and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, 87(417):58–68.

Higham, D. J., Kalna, G., and Kibble, M. (2007). Spectral clustering and its use in bioinformatics. *Journal of computational and applied mathematics*, 204(1):25–37.

Hodges, P. E., Payne, W. E., and Garrels, J. I. (1998). The Yeast Protein Database (YPD): a curated proteome database for saccharomyces cerevisiae. *Nucleic acids research*, 26(1):68–72.

Hofmann, K. and Bucher, P. (1996). The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway. *Trends in biochemical sciences*, 21(5):172–173.

Hofmann, K. and Falquet, L. (2001). A ubiquitin-interacting motif conserved in components of the proteasomal and lysosomal protein degradation systems. *Trends in biochemical sciences*, 26(6):347–350.

Hofmann, T. G., Moller, A., Sirmat, H., Zentgraf, H., Tayas, Y., Droge, W., Will, H., and Schimtz, M. L. (2001). Regulation of p53 activity by its interaction with homeodomain-interacting protein kinase-2. *Nature cell biology*, 4(1):1–10.

Hollstein, M., Sidransky, D., Vogelstein, B., and Harris, C. C. (1991). p53 mutations in human cancers. *Science*, 253(5015):49–53.

Holzer, H. and Heinrich, P. C. (1980). Control of proteolysis. *Annual review of Biochemistry*, 49(1):63–91.

Hood, W., De La Morena, E., Grisolia, S., et al. (1977). Increased susceptibility of carbamylated glutamate dehydrogenase to proteolysis. *Acta biologica et medica Germanica*, 36(11-12):1667.

Hooper, S. D., Boué, S., Krause, R., Jensen, L. J., Mason, C. E., Ghanim, M., White, K. P., Furlong, E. E., and Bork, P. (2007). Identification of tightly regulated groups of genes during drosophila melanogaster embryogenesis. *Molecular systems biology*, 3(1).

Houchmandzadeh, B., Wieschaus, E., and Leibler, S. (2002). Establishment of developmental precision and proportions in the early Drosophila embryo. *Nature*, 415(6873):798–802.

Huber, P. J. (2011). *Robust statistics*. Springer.

Hwang, C., Shemorry, A., and Varshavsky, A. (2010). N-terminal acetylation of cellular proteins creates specific degradation signals. *Science*, 327(5968):973–977.

Jensen, O. (2004). Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current opinion in chemical biology*, 8(1):33.

Jiang, P., Motoo, Y., Garcia, S., Iovanna, J., Pébusque, M., and Sawabu, N. (2006). Down-expression of tumor protein p53-induced nuclear protein 1 in human gastric cancer. *World journal of gastroenterology: WJG*, 12(5):691–696.

Kai, T., Williams, D., and Spradling, A. C. (2005). The expression profile of purified¡ i¿ drosophila¡/i¿ germline stem cells. *Developmental biology*, 283(2):486–502.

Kannan, A., Emili, A., and Frey, B. J. A. (2007). A Bayesian model that links microarray mRNA measurements to mass spectrometry protein measurements. In *Research in Computational Molecular Biology*, pages 325–338. Springer.

Kearney, P. and Thibault, P. (2003). Bioinformatics meets proteomicsbridging the gap between mass spectrometry data analysis and cell biology. *Journal of bioinformatics and computational biology*, 1(01):183–200.

Keene, J. (2007). RNA regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, 8(7):533–543.

Kent, S. B. H. (2009). Total chemical synthesis of proteins. *Chemical Society Reviews*, 38(2):338–351.

Kiemer, L., Bendtsen, J. D., and Blom, N. (2005). NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics*, 21(7):1269–1270.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Koenker, R. and Geling, O. (2001). Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454):458–468.

Krueger, K. E. and Srivastava, S. (2006). Post-translational protein modifications current implications for cancer detection, prevention, and therapeutics. *Molecular & Cellular Proteomics*, 5(10):1799–1810.

Lasdon, L. S. (2013). *Optimization theory for large systems*.

Lechner, J. and Wieland, F. (1989). Structure and biosynthesis of prokaryotic glycoproteins. *Annual review of biochemistry*, 58(1):173–194.

Lederman, D., Zheng, B., Wang, X., Sumkin, J. H., and Gur, D. (2011). A GMM-based breast cancer risk stratification using a resonance-frequency electrical impedance spectroscopy. *Medical physics*, 38(3):1649–1659.

Levine, M.and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945):147–151.

Levine, R. L. (1983). Oxidative modification of glutamine synthetase. i. inactivation is due to loss of one histidine residue. *Journal of Biological Chemistry*, 258(19):11823–11827.

Li, Z., Shin, S., Jeon, S. I., Son, S. H., and Pack, J. K. (2012). A new histogram-based breast cancer image classifier using gaussian mixture model. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, pages 143–147.

Liao, J. C., Boscolo, R., Yang, Y.-L., Tran, L. M., Sabatti, C., and Roychowdhury, V. P. (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences*, 100(26):15522–15527.

Liu, W. and Niranjan, M. (2011). The role of regulated mRNA stability in establishing bicoid morphogen gradient in Drosophila embryonic development. *PLoS ONE*, 6(9):e24896.

Liu, X. and Niranjan, M. (2012). State and parameter estimation of the heat shock response system using Kalman and particle filters. *Bioinformatics*, 28(11):1501–1507.

Liu, Z., Yuan, F., Ren, J., Cao, J., Zhou, Y., and Yang, Q.and Xue, Y. (2012). GPS-ARM: Computational Analysis of the APC/C Recognition Motif by Predicting D-Boxes and KEN-Boxes. *PloS one*, 7(3):e34370.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). Molecular cell biology. *New York*.

Lovell, D., Dance, C., Niranjan, M., Prager, R., Dalton, K., and Derom, R. (1998). Feature selection using expected attainable discrimination. *Pattern Recognition Letters*, 19(5):393–402.

Lu, Y., Zhou, Y., Qu, W., Deng, M., and Zhang, C. (2011). A Lasso regression model for the construction of microrna-target regulatory networks. *Bioinformatics*, 27(17):2406–2413.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Älgenäs, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. *Molecular Systems Biology*, 6(1).

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004a). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004b). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312.

Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., Harris, P. L., Haserlat, S. M., Supko, J. G., Haluska, F. G., et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449.

Magrane, M. and Consortium, U. (2011). Uniprot knowledgebase: a hub of integrated protein data. *Database*, 2011.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS letters*, 583(24):3966–3973.

Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9(1):34.

Man, O. and Pilpel, Y. (2007). Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nature genetics*, 39(3):415–421.

Man, O., Sussman, J. L., and Pilpel, Y. (2006). Examination of the trna adaptation index as a predictor of protein expression levels. In *Systems Biology and Regulatory Genomics*, pages 107–118.

Mann, M. and Jensen, O. N. (2003a). Proteomic analysis of post-translational modifications. *Nature biotechnology*, 21(3):255–261.

Mann, M. and Jensen, O. N. (2003b). Proteomic analysis of post-translational modifications. *Nature biotechnology*, 21(3):255–261.

Marchal, C., Haguenauer-Tsapis, R., and Urban-Grimal, D. (1998). A PEST-like sequence mediates phosphorylation and efficient ubiquitination of yeast uracil permease. *Molecular and cellular biology*, 18(1):314–321.

Marioni, J., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517.

Marks, P. A., Richon, V. M., Miller, T., and Kelly, W. K. (2004). Histone deacetylase inhibitors. *Advances in cancer research*, 91:137–168.

Markstein, M., Markstein, P., Markstein, V., and Levine, M. S. (2002). Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proceedings of the National Academy of Sciences*, 99(2):763–768.

Martinez, L. O., Agerholm-Larsen, B., Wang, N., Chen, W., and Tall, A. R. (2003). Phosphorylation of a pest sequence in abca1 promotes calpain degradation and is reversed by apoa-i. *Journal of Biological Chemistry*, 278(39):37368–37374.

Mattick, J. S. (2001). Non-coding rnas: the architects of eukaryotic complexity. *EMBO reports*, 2(11):986–991.

Mattick, J. S. and Gagen, M. J. (2001). The evolution of controlled multitasked gene networks: the role of introns and other noncoding rnas in the development of complex organisms. *Molecular Biology and Evolution*, 18(9):1611–1630.

McDonald, J. H. (2009). *Basic concepts of hypothesis testing*, volume 2. Sparky House Publishing Baltimore, MD.

Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., Mayer, A., J. M., Dölken, L., et al. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular systems biology*, 7(1):458–470.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature genetics*, 30(1):13–19.

Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M., Wisniewski, J. R., Stahl, E., Bolouri, M. S., Ray, H. N., Sihag, S., Kamal, M., et al. (2003). Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell*, 115(5):629–640.

Moreira, D., Kervestin, S., Jean-Jean, O., and Philippe, H. (2002). Evolution of eukaryotic translation elongation and termination factors: variations of evolutionary rate and genetic code deviations. *Molecular biology and evolution*, 19(2):189–200.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628.

Muppirala, U. K., Honavar, V. G., and Dobbs, D. (2011). Predicting rna-protein interactions using only sequence information. *BMC Bioinformatics*, 12(1):489–501.

Nalivaeva, N. N. and Turner, A. J. (2001). Post-translational modifications of proteins: acetylcholinesterase as a model system. *Proteomics*, 1(6):735–747.

Nie, L., Wu, G., and Zhang, W. (2006). Correlation between mRNA and protein abundance in desulfovibrio vulgaris: A multiple regression to identify sources of variations. *Biochemical and biophysical research communications*, 339(2):603–610.

Nilsen, T. W. and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463.

Ning, K., Fermin, D., and Nesvizhskii, A. I. (2012). Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-seq gene expression data. *Journal of proteome research*, 11(4):2261–2271.

Nygaard, V., Løland, A., Holden, M., Langaas, M., Rue, H., Liu, F., Myklebost, O., Fodstad, Ø., Hovig, E., and Smith-Sørensen, B. (2003). Effects of mRNA amplification on gene expression ratios in cDNA experiments estimated by analysis of variance. *BMC Genomics*, 4(1):11.

Olsen, D. A., Østergaard, B., Bokmand, S., Wamberg, P. A., Jakobsen, E. H., and Brandslund, I. (2007). Her-2 protein concentrations in breast cancer cells increase before immunohistochemical and fluorescence in situ hybridization analysis turn positive. *Clinical Chemical Laboratory Medicine*, 45(2):177–182.

Oueslati, A., Fournier, M., and Lashuel, H. (2010). Role of post-translational modifications in modulating the structure, function and toxicity of α-synuclein: Implications for Parkinsons disease pathogenesis and therapies. *Progress in brain research*, 183:115–145.

Ozsolak, F. and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98.

P., M. (1997). Bacterial glycoproteins. *Glycoconj J*, 14:3–11.

Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., Lindeman, N., Boggon, T. J., et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497–1500.

Pancaldi, V. and Bähler, J. (2011). In silico characterization and prediction of global protein–mRNA interactions in yeast. *Nucleic Acids Research*, 39(14):5826–5836.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pfleger, C. M. and Kirschner, M. W. (2000). The ken box: an apc recognition signal distinct from the d box targeted by cdh1. *Genes & development*, 14(6):655–665.

Pham Dinh, T. and Le Thi, H. A. (1997). Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355.

Pimentel, M. A. F., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

Preiss, T. and WHentze, M. (2003). Starting the protein synthesis machine: eukaryotic translation initiation. *Bioessays*, 25(12):1201–1211.

Products, T. S. P. P. B. (2013). Overview of post-translational modifications (ptms). Available online at http://www.piercenet.com/browse.cfm?fldID=7CE3FCF5-0DA0-4378-A513-2E35E5E3B49B.

Qin, X., Ahn, S., Speed, T. P., Rubin, G. M., et al. (2007). Global analyses of mRNA translational control during early Drosophila embryogenesis. *Genome Biol*, 8(4):R63.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics*, 3(1):30.

Reimers, M. (2010). Making informed choices about microarray data analysis. *PLoS computational biology*, 6(5):e1000786.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends in genetics*, 16(6):276–277.

Richard, H., Schulz, M. H., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., et al. (2010). Prediction of alternative isoforms from exon expression levels in rna-seq experiments. *Nucleic acids research*, 38(10):e112–e112.

Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1472.

Rogers, S. (2011). Statistical methods and models for bridging omics data levels. In *Bioinformatics for Omics Data*, pages 133–151.

Rogers, S. and Girolami, M. (2012a). *Classification*. CRC Press, US.

Rogers, S. and Girolami, M. (2012b). *Linear Modelling: A Least Squares Approach*. CRC Press, US.

Rogers, S., Girolami, M., Kolch, W., Waters, K. M., and Liu, T. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*, 24(24):2894–2900.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*, volume 94. John Wiley & Sons.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Saenger, W. (1984). *Principles of nucleic acid structure*, volume 7.

Sanguinetti, G., Lawrence, N. D., and Rattray, M. (2006). Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22(22):2775–2781.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

Schwanhäusser, B., Wolf, J., Selbach, M., and Busse, D. (2013). Synthesis and degradation jointly determine the responsiveness of the cellular proteome. *BioEssays*.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905.

Shi, L., Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., et al. (2006). The

microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9):1151–1161.

Shiio, Y. and Aebersold, R. (2006). Quantitative proteome analysis using isotope-coded affinity tags and mass spectrometry. *NATURE PROTOCOLS-ELECTRONIC EDITION*, 1(1):139.

Shin, Y., Chen, K., Sayed, A. H., Hencey, B., and Shen, X. (2013). Post-translational regulation enables robust p53 regulation. *BMC systems biology*, 7(1).

Shmulevich, I. and Zhang, W. (2002). Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*, 18(4):555–565.

Silencing, T. U. T. F. G. (2011). Small RNA - Introduction. Available online at http://2011.igem.org/Team:DTU-Denmark/Background_sRNA#References.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209.

Šmardová, J., Šmarda, J., and Koptíková, J. (2005). Functional analysis of p53 tumor suppressor in yeast. *Differentiation*, 73(6):261–277.

Solomon, V.and Lecker, S. H. and Goldberg, A. (1998). The n-end rule pathway catalyzes a major fraction of the protein degradation in skeletal muscle. *Journal of Biological Chemistry*, 273(39):25216–25222.

Stadtman, E. (1990). Covalent modification reactions are marking steps in protein turnover. *Biochemistry*, 29(27):6323–6331.

Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl01):D535–D539.

Statnikov, A., Wang, L., and Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1):319–329.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800.

Takeda, A. and Sugiyama, M. (2008). ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th international conference on Machine learning*, pages 1056–1063. ACM.

Thiru, P., Kern, D. M., McKinley, K. L., Monda, J. K., Rago, F., Su, K., Tsinman, T., Yarar, D., Bell, G. W., and Cheeseman, I. M. (2014). Kinetochore genes are coordinately up-regulated in human tumors as part of a foxm1-related cell division program. *Molecular biology of the cell*, 25(13):1983–1994.

Thomas, J. D. and Johannes, G. (2007). Identification of mRNAs that continue to associate with polysomes during hypoxia. *Rna*, 13(7):1116–1131.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). Panther: a library of protein families and subfamilies indexed by function. *Genome research*, 13(9):2129–2141.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515.

Tritchler, D., Fallah, S., and Beyene, J. (2005). A spectral clustering method for microarray data. *Computational statistics & data analysis*, 49(1):63–76.

Trost, Z., Marc, J., Sok, M., and Cerne, D. (2008). Increased apolipoprotein e gene expression and protein concentration in lung cancer tissue do not contribute to the clinical assessment of non-small cell lung cancer patients. *Archives of medical research*, 39(7):663–667.

Tuller, T., Kupiec, M., and Ruppin, E. (2007). Determinants of protein abundance and translation efficiency in S.Cerevisiae. *PLoS Computational Biology*, 3(12):e248.

Tuller, T., Waldman, Y. Y., Kupiec, M., and Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences*, 107(8):3645–3650.

Tuna, S. and Niranjan, M. (2009). Classification with binary gene expressions. *Journal of Biomedical Science and Engineering*, 2(06):390.

Tuna, S. and Niranjan, M. (2010). Inference from low precision transcriptome data representation. *Journal of Signal Processing Systems*, 58(3):267–279.

Vapnik, V. N.and Vapnik, V. (1998). *Statistical learning theory*, volume 2.

Vapnyarskii, I. (2012). *Lagrange multipliers*.

Varshavsky, A. (1997). The n-end rule pathway of protein degradation. *Genes to Cells*, 2(1):13–28.

Vogel, J. and Wagner, E. G. H. (2007). Target identification of small noncoding RNAs in bacteria. *Current opinion in microbiology*, 10(3):262–270.

Waldman, Y. Y., Tuller, T., Shlomi, T., Sharan, R., and Ruppin, E. (2010). Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Research*, 38(9):2964–2974.

Wall, D., Hirsh, A., Fraser, H., Kumm, J., Giaever, G., Eisen, M., and Feldman, M. (2005). Functional genomic analysis of the rates of protein evolution. Number 15, pages 5483–5488. National Acad Sciences.

Walsh, C. (2006). *Posttranslational modification of proteins: expanding nature's inventory*. Roberts & Company.

Walsh, C. et al. (2006). Posttranslational modification of proteins.

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schimpf, P., Hengartner, M., and Mering, C. (2012a). Paxdb, a database of protein abundance averages across all three domains of life. *Molecular and cellular proteomics : MCP*, 11(8):492–500.

Wang, N., Chen, W., Linsel-Nitschke, P., Martinez, L. O., Agerholm-Larsen, B., Silver, D. L., and Tall, A. R. (2003). A PEST sequence in abca1 regulates degradation by calpain protease and stabilization of abca1 by apoa-i. *Journal of Clinical Investigation*, 111(1):99.

Wang, T., Li, H., Hu, X., and Cao, X. (2012b). Predicting rna-protein interactions using a novel method. In *Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference on*, pages 981–985. IEEE.

Wang, X. and Gotoh, O. (2009). Cancer classification using single genes. *Genom Informatics*, 23(1):179–188.

Wang, Y., Liu, C., Storey, J. D., Tibshirani, R., Herschlag, D., and Brown, P. O. (2002). Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences*, 99(9):5860–5865.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl 2):W214–W220.

Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in biochemical sciences*, 24(11):437–440.

Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, R. R., Plouffe, D., Deciu, C., Winzeler, E., and Yates, J. R. (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.

Wei, X. and Li, L. (2009). Mass spectrometry-based proteomics and peptidomics for biomarker discovery in neurodegenerative diseases. *International journal of clinical and experimental pathology*, 2(2):132.

Wilson, C. L., Candela, G. T., and Watson, C. I. (1994). Neural network fingerprint classification. *Journal of Artificial Neural Networks*, 1(2):203–228.

Wong, N. K., Easton, R. L., Panico, M., Sutton-Smith, M., Morrison, J. C., Lattanzio, F. A., Morris, H. R., Clark, G. F., Dell, A., and Patankar, M. S. (2003). Characterization of the oligosaccharides associated with the human ovarian tumor marker CA125. *Journal of Biological Chemistry*, 278(31):28619–28634.

Wu, G., Nie, L., and Zhang, W. (2008). Integrative analyses of posttranscriptional regulation in the yeast Saccharomyces cerevisiae using transcriptomic and proteomic data. *Current Microbiology*, 57:18–22.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Xing, E. P. and Karp, R. M. (2001). Cliff: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*, 17(suppl 1):S306–S315.

Xu, L., Crammer, K., and Schuurmans, D. (2006). Robust support vector machine training via convex outlier ablation. In *AAAI*, pages 536–542. AAAI Press.

Yang, X. and Seto, E. (2008). Lysine acetylation: codified crosstalk with other posttranslational modifications. *Molecular cell*, 31(4):449–461.

Yates, J. R., Ruse, C. I., and Nakorchevsky, A. (2009). Proteomics by mass spectrometry: approaches, advances, and applications. *Annual review of biomedical engineering*, 11:49–79.

Yeakley, J. M .and Fan, J., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M. S., and Fu, X. (2002). Profiling alternative splicing on fiber-optic arrays. *Nature biotechnology*, 20(4):353–358.

Yeh, I. C. (1998). Modeling concrete strength with augment-neuron networks. *Journal of Materials in Civil Engineering*, 10(4):263–268.

Yoon, S. and Seger, R. (2006). The extracellular signal-regulated kinase: multiple substrates regulate diverse cellular functions. *Growth factors*, 24(1):21–44.

Yu, Y., Yang, M., Xu, L., White, M., and Schuurmans, D. (2010). Relaxed clipping: A global training method for robust regression and classification. In *Neural Information Processing Systems*, pages 2532–2540.

Zhang, B., Kirov, S., and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research*, 33(suppl 2):W741–W748.

Zhang, H., Holden, A. V., Kodama, I., Honjo, H., Lei, M., Varghese, T., and Boyett, M. R. (2000). Mathematical models of action potentials in the periphery and center of the rabbit sinoatrial node. *American Journal of Physiology - Heart and Circulatory Physiology*, 279(1):H397–H421.

Zhang, Y., Sturgill, D., Parisi, M., Kumar, S., and Oliver, B. (2007). Constraint and turnover in sex-biased gene expression in the genus drosophila. *Nature*, 450(7167):233–237.

Zheng, Q. and Wang, X.-J. (2008). Goeast: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic acids research*, 36(suppl 2):W358–W363.

Zheng-Bradley, X., Rung, J., Parkinson, H., and Brazma, A. (2010). Large scale comparison of global gene expression patterns in human and mouse. *Genome Biology*, 11(12):R124.

Zhou, X., Wang, X., and Dougherty, E. R. (2003). Binarization of microarray data on the basis of a mixture model1. *Molecular cancer therapeutics*, 2(7):679–684.