# Integrative Analyses of Diverse Biological Data Sources

by

Wandaliz Torres García

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2011 by the
Graduate Supervisory Committee:

Deirdre R. Meldrum, Co-Chair
George C. Runger, Co-Chair
Esma S. Gel
Jing Li
Weiwen Zhang

ARIZONA STATE UNIVERSITY

May 2011

# ABSTRACT

The technology expansion seen in the last decade for genomics research has permitted the generation of large-scale data sources pertaining to molecular biological assays, genomics, proteomics, transcriptomics and other modern "omics" catalogs. New methods to analyze, integrate and visualize these data types are essential to unveil relevant disease mechanisms. Towards these objectives, this research focuses on data integration within two scenarios: (1) transcriptomic, proteomic and functional information and (2) real-time sensor-based measurements motivated by single-cell technology.

To assess relationships between protein abundance, transcriptomic and functional data, a nonlinear model was explored at static and temporal levels. The successful integration of these heterogeneous data sources through the stochastic gradient boosted tree approach and its improved predictability are some highlights of this work. Through the development of an innovative validation subroutine based on a permutation approach and the use of external information (i.e., operons), lack of a priori knowledge for undetected proteins was overcome. The integrative methodologies allowed for the identification of undetected proteins for *Desulfovibrio vulgaris* and *Shewanella oneidensis* for further biological exploration in laboratories towards finding functional relationships.

In an effort to better understand diseases such as cancer at different developmental stages, the Microscale Life Science Center headquartered at the Arizona State University is pursuing single-cell studies by developing novel technologies. This research arranged and applied a statistical framework that tackled the following challenges: random noise, heterogeneous dynamic systems with multiple states, and understanding cell behavior within and across different Barrett's esophageal epithelial cell lines using oxygen consumption curves. These curves were characterized with good empirical fit using nonlinear models with simple structures which allowed extraction of a large number of features. Application of a supervised classification model to these features and the integration of experimental factors allowed for identification of subtle patterns among different cell types visualized through multidimensional scaling. Motivated by the challenges of analyzing real-time measurements, we further explored a unique two-dimensional representation of multiple time series using a wavelet approach which showcased promising results towards less complex approximations. Also, the benefits of external information were explored to improve the image representation.

To my parents,

Efraín Torres Centeno and Wanda Ivelisse García Morales

for their unconditional love, encouragement, and support.

# ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of supportive people that in one way or another provided valuable assistance to the completion of my doctoral training at Arizona State University (ASU).

Dr. Deirdre R. Meldrum, for her wisdom and support. She is truly an inspiration to me as a woman scientist, engineer, and leader. I am grateful for the opportunities provided in the Center for Biosignatures Discoveries Automation at the Biodesign Institute at the Arizona State University and within National Institutes of Health. My integration to her laboratory gave me an incredible place to build a strong interdisciplinary career.

Dr. George C. Runger, for his extraordinary guidance throughout this research. His valuable input to this work and encouragement in all aspects of my career has made my doctoral training a very rewarding experience. His professionalism and work ethics are qualities I admire and I wish to emulate.

Dr. Weiwen Zhang, for his supervision, vision and knowledge which were important to the progress of this work and its peer-reviewed publications.

Dr. Esma S. Gel and Dr. Jing Li, for their suggestions to improve this work throughout my comprehensive exam and assessment of this dissertation.

Dr. Roger H. Johnson, Dr. Steven D. Brown, Dr. Laimonas Kelbauskas, and Dr. Shashanka Ashili, for serving as co-authors of my research papers, their insights and suggestions greatly improved this work.

My deepest gratitude to my friends and colleagues from the Industrial Engineering program in the School of Computing, Informatics, and Decisions Systems Engineering, specially Dr. Runger's lab students, and the Center for Biosignatures Discovery Automation at the Biodesign Institute at ASU.

To my parents, Efraín Torres Centeno and Wanda Ivelisse García Morales, for their unconditional love and support through this journey. To my sister, Griselle Torres García, and her sons, Sebastian and Santiago, for making my life in Arizona one full of everlasting love and smiles. To my family in Puerto Rico and my friends all over the world for their caring messages. Specially to; Laura Cabré, for always being there. Erika Murguia Blumenkranz, for her understanding and care. María Angélica Velázquez, for her encouraging words. Carelyn Torres Rivera, for making me feel as part of her family.

Lastly, I would like to thank the National Human Genome Research Institute Center of Excellence in Genomic Sciences (Grant 5 P50 HG002360), the National Science Foundation GK-12 "Down to Earth Science" project (Grant NSF/DGE 0538701), the U.S. Department of Energy Office of Science, Biological and Environmental Research at the Oak Ridge National Laboratory managed by UT-Battelle, LLC (Contract DE-AC05-00OR22725), and the More Graduate Education at Mountain States Alliance within the Alliance for Graduate Education and the Professoriate for their funding.

# TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

## LIST OF TABLES

Table Page

Table Page

## LIST OF FIGURES

Figure Page

Figure Page

# CHAPTER 1

# INTRODUCTION

## 1.1. Introduction

To enhance health in this era of genomics, the biggest challenge lies on how to extract true knowledge from the diverse libraries of biomedical data. Large scale datasets and meaningful knowledge extraction are two huge challenges for the bioinformatics community driven by technology expansion. The curse of dimensionality caused by genomic, proteomic, clinical, and lifestyle factors that interact as a system and the inherent variability of the complex system limit a simple pairwise or additive model [Robson, 2004]. These challenges present opportunities for interdisciplinary research aiming to achieve reliable biomedical decision making requiring the training of research individuals with expertise in two or more of the following areas: biology, informatics, computer science, mathematics, statistics and/or engineering [Green and Guyer, 2011]. This research focuses on integrating heterogeneous biological data sources through data mining methodologies to reach suitable biological interpretation.

There are several types of biological information in this field (see Appendix A). This research explores them in two main categories: (1) transcriptomic and proteomics in bulk cells and (2) sensor-based measurements such as single cell data sources as shown in Figure 1. The research addresses particular objectives for each of these two categories as follows:

Specific Aim 1a : Integrate static transcriptomic and proteomic data to assess how cellular protein abundance corresponds to mRNA expression levels, and to understand underlying cell mechanisms that affect their correlation through a stochastic gradient boosted tree approach. The integration model is validated through a novel subroutine developed as a permutation test with the use of external biological data (i.e., operons) (See Chapter 3).

Specific Aim 1b : Integrate temporal transcriptomic and proteomic data for missing value imputation at various levels through a multi-tiered statistical framework (See Chapter 4).

Specific Aim 2a : Investigate the data structure of sensor-based measurements of oxygen consumption curves through a statistical framework addressing unique data challenges from a novel single-cell technology developed at the Microscale Life Science Center (MLSC) headquartered at Arizona State University (See Chapter 5).

Specific Aim 2b : Address the problem of multiple time series approximation through an imaging approach using wavelets motivated by the sensor-based measurements structure (See Chapter 6).

A general diagram presenting these research objectives is depicted in Figure 1.

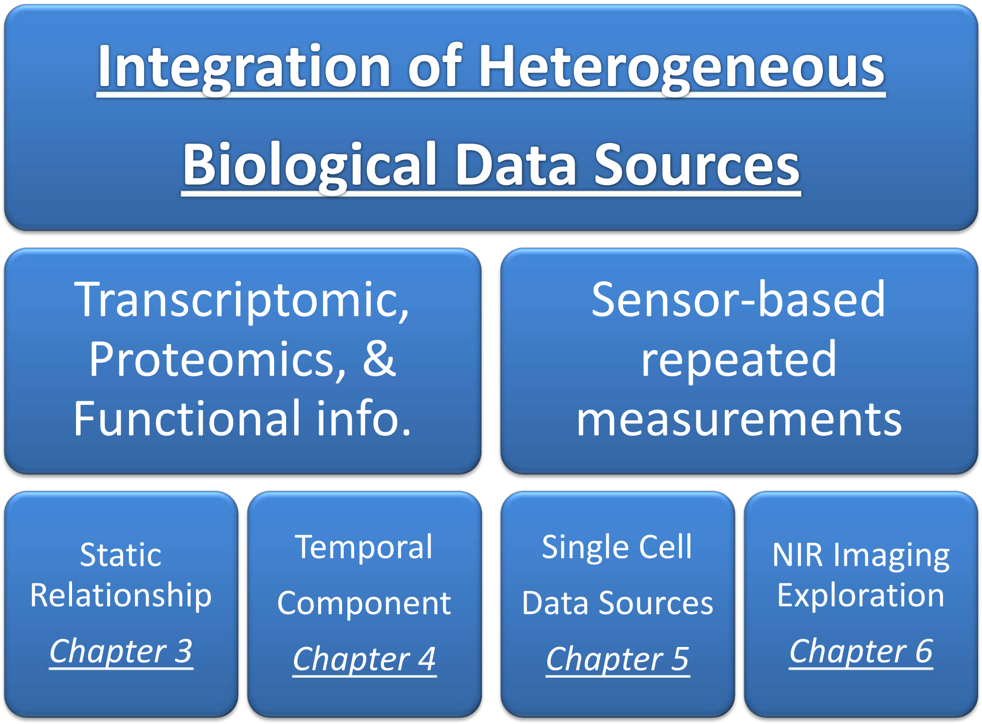

Integration of Heterogeneous Biological Data Sources
Transcriptomic, Proteomics, & Functional info.
Sensor-based repeated measurements
Static Relationship
Chapter 3
Temporal Component
Chapter 4
Single Cell Data Sources
Chapter 5
NIR Imaging Exploration
Chapter 6

Figure 1. Research proposal objectives.

## 1.2. Contributions

The need for better methodologies to analyze genomic data through cross-cutting research motivates the work presented here. The contributions outlined below address diverse challenges encountered through the integration of large and heterogeneous data sources within many scenarios at the cellular and molecular level. Based on the interdisciplinary nature of the research team, this work reports several original contributions to diverse scientific fields such as statistical data analysis, genomics, and bioinformatics.

### 1.2.1. Static transcriptomic and proteomics model

1. Successful integration of multiple heterogeneous data sources.
2. Application of non-linear models, stochastic gradient boosted trees, with improvements in predictability when compared to previous reported methods.
3. Development of a new validation procedure by using external biological information to overcome lack of a priori knowledge through a permutation approach.

4. Identification of a list of undetected proteins for *Desulfovibrio vulgaris* with similar abundances to those detected experimentally; allowing for further biological exploration in the lab towards finding functional relationships.

1.2.2. Temporal transcriptomic and proteomics model

1. Application of quality control routines to remove “noisy” genes.
2. Integration of limited temporal data with other mixed-type datasets
3. Application of non-linear methods such as smoothing cubic splines and stochastic gradient boosted trees.
4. Extension to developed validation routine in Chapter 3 that includes the use of statistical hypothesis testing to a permuted-based procedure and integration of external biological data that provides information on how similar are a group of genes and/or peptides (i.e., operons).
5. Identification of a list of undetected proteins for *Shewanella oneidensis* for further experimentation towards future functional annotations and understandings.

### 1.2.3. Single-cell statistical framework

1. Identification of unique data challenges for sensor-based measurements obtained through novel single-cell technologies at the Microscale Life Science Center (MLSC) headquartered at Arizona State University.

2. Procedure design and application of a statistical framework consisting of a set of tools within the areas of signal processing, quality control, and data mining (see Figure 14).

   - Parameters estimation study for the methods serially applied.
   - Characterization of oxygen consumption (OC) curves with good empirical fit using non-linear models with simple structures.
   - Extraction of several features from characterized curves.
   - Modeling a large number of features including the integration of experimental factors (i.e., incubation time) towards identification of biosignatures and/or patterns within and across different cell types.
   - Application of a non-linear classification models to unravel possible biosignatures across cell types.
   - Multidimensional scaling and visualization tools to capture proximity measurements from the predictors to discriminate among differ-

ent cell classes.

- Statistical comparisons of several extracted features among different cell types (Barrett's esophagus cell lines: CP-A and CP-C) and different numbers of cells within a microwell.

### 1.2.4. Imaging approach for multiple time series approximation

1. Description of a unique two-dimensional representation of multiple time series.

2. Investigated the performance of such representation through the application of wavelets decomposition approach for imaging approximations.

3. Integration of additional one-dimensional variables linked to the time series was explored. These variables were used as ordering features of the time series for the image analysis and its performance was investigated as a possible variable ranking procedure.

## 1.3. Organization of the research

Chapter 2 provides a general overview of the current literature for the problems explored in this research with more specific background information in each research chapter. In Chapter 3, the goal is to learn the inherent relationship between transcriptomic and proteomic data to predict undetected proteins. In this study, we present a nonlinear data-driven model to predict abundance for undetected proteins using two independent datasets of cognate transcriptomic and proteomic data collected from *Desulfovibrio vulgaris*. We used stochastic gradient boosted trees (GBT) which performed better than previous studies in the literature and validated its predictability performance through the development of a permuted-based subroutine that incorporates external linking relationships among genes and/or peptides (i.e., operons) .

Chapter 4 extends the work in Chapter 3 to include another level of complexity by analyzing a temporal component in the data. In this study, we focused on missing value imputation at both transcriptomic and proteomic levels for *Shewanella oneidensis*. A multi-tiered framework was used to address the problem of missing values in several spectrums (i.e., transcriptomic, proteomic). This multi-tiered framework consisted of: data preprocessing and normalization, quality control on "noisy" genes, transcriptomic temporal curve fitting, GBT fitting to transcriptomic, proteomics and other gene ontology information, and lastly model validation through external biological information such as operon and pathway data. The experience acquired by working with these complex genomic datasets highlighted elements to focus on when integrating biological resources.

Chapter 5 presents unique challenges and solution strategies to characterize sensor-based measurements specially for oxygen consumption (OC) curves at the single-cell level. The technologies developed within the Microscale Life Science Center (MLSC) headquartered at the Arizona State University can reveal complex data structures with valuable information towards understanding differences between cells and relationships with cancer. Therefore in this study, we have identified three major challenges in analyzing cellular heterogeneity from these data structures: random noise, heterogeneous dynamic systems with multiple states, and understanding cell behavior within and across different cell types. A statistical framework was arranged to address these particular challenges allowing for good empirical characterization of OC curves and extraction of features used for comparisons and classification among different Barrett's esophagus cell lines. Statistical tests were performed on the extracted features to gather differences among different cell types and experimental design. A nonlinear, supervised, and ensemble methodology (random forest) was applied to extracted features from OC curves characterization to explore further discrimination of classes. This supervised model allowed for visualization of its calculated proximity distance matrix across classes (i.e., cell types) through multidimensional scaling.

Motivated by the challenges of analyzing data structures with a temporal component such as single-cell measurements, we further explored methodologies to better approximate these measurements. Chapter 6 presents a unique representation of multiple time series through imaging and evaluates its performance in terms of model complexity and accuracy using wavelet decomposition approaches. The performance is tested on three near infrared (NIR) datasets. This two-dimensional wavelets approach was found useful to approximate certain multiple time series. Furthermore, integration of one-dimensional variables linked to each time series were investigated as a variable ranking procedure to the response of study.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1. Integration of heterogeneous data sources

A diverse array of biological databases is currently available. These sources range from experimental measurements (i.e., gene expression and proteomic assays) to annotated information derived experimentally in most cases (i.e., genome annotation and text literature). Some biological knowledge databases readily accessible are gene expression from Gene Expression Omnibus [Edgar et al., 2002], curated protein sequences from Swiss-Prot [Boeckmann et al., 2003], and the Kyoto Encyclopedia of Genes and Genomes (KEGG) which offers special information on enzymatic pathways [Ogata et al., 1999]. Joyce and Palsson (2006) briefly explained these data types in three categories: components, interactions, and functional states (See Appendix A).

### 2.1.1. Biological databases challenges

With the advancement in high-throughput technologies for biological research, the amount of data available for analysis have increased tremendously and some researchers expects it to exceed exabyte [Searls, 2005]. However, the challenge that originates from large amounts of data is not unique for biological research discipline. Current datasets in many fields has massive size such as global climate change data gathered by the National Oceanic and Atmospheric Administration and product information from large retail stores like Walmart [Joyce and Palsson, 2006].

Heterogeneity and high dimensionality encompassed the biggest challenges in the field of current genomic research. Even though other domains have encountered problems with heterogeneous and high dimensional data motivating great progress in the areas of data fusion and feature selection to address these [Esteban et al., 2005, Liu et al., 2005]; there are unique hurdles pertaining to “omics” research. Heterogeneity issues in this field are confounded within (1) techniques diversity, (2) standardization, and (3) cross-lab consistency [Joyce and Palsson, 2006, Searls, 2005].

By the vast amounts of techniques for experimental assays comes the challenge of “standardization” which affects in two different ways: normalization and annotation. Normalization is a major difficulty and has motivated scale-free methodology research. Normalization methods are vastly used to correct quantities and systematic errors between data sets [Searls, 2005]; however there is no standard approach which makes difficult further data integration [Polpitiya et al., 2008]. Standardization problem also deals with annotation. Standard annotation of information gathered is not always adopted, slowing down subsequent analysis by other researchers [Quackenbush, 2004]. Even though many data repositories are working to improve this issue; only some are close to achieve the consistency from GenBank (a database maintained by the US National Center for Biology Information) [Joyce and Palsson, 2006]. Another concern within the heterogeneity issues of "omics" data is the cross-lab consistency. Besides annotation there are many experimental parameters and information that vary greatly across different laboratories.

Finally, but not least important, is the curse of dimensionality which is a widely studied problem in operations research and data mining. Some examples within the "omics" arena are: sequencing millions of base pairs and gene information that could yield data for about $\sim 25k$ genes for humans in systems biology experiments. With the curse of dimensionality comes a bigger problem small sample size for many experimental assays. Despite the large amounts of data found in this field, there is a limitation in the amount of replicates. Some of these experimental assays are quite costly and destructive (i.e., to obtain gene expression RNA needs to be extract and hence the cells dies).

All these challenges have motivated plenty of work to integrate biological data in the hopes to better understand biological systems [Kong et al., 2006,Troyanskaya et al., 2003, Ko and Lee, 2009, Daemen et al., 2009, Li et al., 2006, Zhao et al., 2008, Ye et al., 2008]. Pairwise data integration has been the clear approach to combine the more mature technologies. Further pairwise integration such as metabolomics and fluxomics to understand metabolic pathway bottlenecks was outlined by Joyce and Palsson (2006) .

### 2.1.2. General integration methods

There is a vast list of statistical methods used in bioinformatics to combine biological sources such as gene expression and gene ontology annotations but most of them rely on univariate tests which do not account for interactions among sources [Kong et al., 2006]. To address integration of heterogenous datasets researchers have developed subspace correlations using Hotelling's $T^2$ statistic as a separation measure, Bayesian networks, and Kernel based methods [Kong et al., 2006, Li et al., 2006, Troyanskaya et al., 2003, Daemen et al., 2009, Ko and Lee, 2009]. Tree ensemble methods are recently becoming popular to address the challenges of data integration because of its functionality, accuracy and competitiveness with state-of-the-art methods for problems such as gene function prediction [Re and Valentini, 2010].

Although incorporating several data sources might seem the solution to obtain quality results, integrating more data does not always produce better predictions [Ko and Lee, 2009]. It is essential to extract relevant features from those data sources before or during the integration process. The objective of feature selection is to model a response variable $y$ with a subset of the important predictor variables. This is a general goal and quite a few more specific objectives can be identified. For example, one objective is filtering to separate irrelevant inputs. Another objective is variable ranking where the objective is to obtain the relative relevance of all input variables to the target. The ultimate target is to develop a compact and effective model to identify a small subset of independent features with the most predictive power [Tuv et al., 2009]. Furthermore, Ye et. al. 2008 proposed to integrate diverse datasets using feature selection based on kernels to predict Alzheimer's Disease [Ye et al., 2008]. Hence, it is a current challenge to extract features from several data sources.

## 2.2. Integration of transcriptomic and proteomic sources

Recent achievements in high-throughput technologies have enabled the quantitative measurements of various biological molecules abundance and their variation between different states at the genomic scale. Some of the technologies used to measure mRNA expression are DNA microarray and Serial Analysis of Gene Expression (SAGE). Discriminating between a healthy and diseased cell from gene expression data has driven researchers to implement statistical approaches for finding patterns in experimental results; for example, principal component analysis (PCA) [Jolliffe, 2002], a method to identify the most discriminating features in a dataset. Also widely implemented in this field are many well known clustering techniques that aim to partition the data into a predetermined number of categories as instances are examined, based on dissimilarity or similarity measures [Bertone and Gerstein, 2001].

Clustering is performed to group certain genes which have similarities in their gene expression profile, taking as a biological fact that genes that are similarly expressed are often involved in similar processes. Certainly, there are many clustering approaches that can provide different perspectives. Hierarchical clustering, k-means and self organizing maps are some of the most implemented clustering algorithms [Bertone and Gerstein, 2001].

Boros et. al proposed a new methodology with the objective to detect structural information from datasets. The key features of their methodology called Logical Analysis of Data (LAD) are the discovery of minimal sets of features (i.e., genes) necessary to explain trends from all observations and the detection of hidden patterns in the data capable of distinguishing observations describing positive outcome events from negative outcome events [Boros et al., 2000]. Their work is in the area of combinatorial optimization using a Boolean approach. In essence, LAD generates conjunctive patterns that characterize two different sets of outcomes using a heuristic model for a set covering formulation.

Evidence suggests that transcriptome profiling is necessary but not sufficient to characterize biological systems complexities (i.e., all regulatory processes in the cell) [Gygi et al., 1999]. Therefore, in addition to studying gene expression at the transcriptional level, large-scale proteomic analysis should be considered as well to understand living organisms' systems and pathways. Proteome-based expression analysis is generally performed by two-dimensional gel electrophoresis, in which proteins are separated according to their isoelectric point and mass [Nie et al., 2007]. This technique requires intensive labor. However, recent advances in mass spectrometry-based technology allows us to perform large-scale characterization of the proteome. More specifically, high-performance liquid chromatographic (HPLC) fractionation of protein tryptic digests, followed by automated tandem mass spectrometry (MS/MS) to identify peptide fragments, allows identification of several hundred proteins simultaneously from a single cellular extract [Nie et al., 2007].

Several researchers have integrated transcriptomic and proteomic datasets to address the relationship between cellular protein abundance and mRNA concentrations. Although the relationship between mRNA expression levels and protein abundance seems evident, it is not immediately apparent from experimental data. Most integrative approaches have not been able to find correlations [Nie et al., 2007]. Moderate correlations found in the literature might be attributed to some of the following sources of variation: (1) measurement errors in the gathering data process, (2) protein and/or transcript length, and (3) underlying biological mechanisms (e.g., translation regulation). These factors along with the challenges of data normalization, effects of measurements errors, compensation of missing values, and non-uniformity of the correlations structure are essential to capture the true correlation and have not been addressed properly.

### 2.2.1. Temporal component

There are several issues involving the study of temporal gene expression data. Some of the most important are: determination of sampling rates, variability in the timing of the biological process, synchronization, small number of replicates, missing values (lack of full repeats) and others [Bar-Joseph, 2004, Ernst et al., 2005]. These issues bring challenges for biological interpretation and computational efforts.

Determining the “right” sampling rates is an exhaustive task which is usually accomplish by biologist’s intuition. Under-sampling provides incorrect representation of temporal gene activity while oversampling is expensive and time consuming. As mentioned previously this issue is mainly dealt by experimentalist experience; however online methods have been explored for which low number samples are performed initially and its results undergo statistical inference to determine if the size is sufficient or more sampling is required. These methods should be repeated until the test statistic reaches a positive outcome which could yield to very expensive experiments. Confounded with sampling rates is the variability in the timing of the biological process such as cell cycle. The biological time is different across organisms, experimental conditions and others making very difficult comparisons using temporal information.

Another key issue for temporal analysis of gene expression is synchronization. For reliable analysis of temporal data, all cells should start at the same phase which is a very difficult task in practice but needed. Several tests involving Fourier Transform and randomization have been designed to deliberate whether or not synchronization was achieved [Shedden and Cooper, 2002]. Other methods are mentioned in Bar-Joseph 2004 review [Bar-Joseph, 2004] of temporal gene expression analysis.

Small sample size and missing values has been a limitation for all type of transcriptomic assays (static or temporal). Small sample size becomes a crucial issue because of the large number of genes capture in one sample creating a high dimensional problem. While missing values create challenges to extract continuous representation for each gene; however methods such as interpolation and splines have been studied to overcome this challenge [Aach and Church, 2001, D/haeseleer et al., 1999, Bar-Joseph, 2004, Ernst et al., 2005].

## 2.3. Sensor-based measurement analysis

Sensor-based measurements are common data structures, often collected in real time, that have offered opportunities to elucidate new insights about cellular systems mechanisms and promises further discoveries with the advancements in technologies [McNeil and Manning, 2002, Bussink et al., 2000, Otto, 2011]. As new behaviors unfold from these data types new challenges arises requiring new computational methods to extract "true" knowledge from them. Some of the problems investigated throughout this work are: modeling multiple time series, anomaly detection, change-point detection and feature extraction motivated by single-cell data curves from novel technologies.

The importance of information at the individual cell level to uncover relevant disease mechanisms and the technologies developed through the Microscale Life Science Center headquartered at the Arizona State University to achieve this objective is highlighted. The oxygen consumption curves gathered through sensor-based technologies are the foundations of this work for the applied statistical framework. Lastly, a review some of the current work related to imaging approximation methods.

### 2.3.1. General problems and methods

#### 2.3.1.1. Modeling multiple time series

Multiple time series analysis is a well studied area to find relationships among $n$ time series, $X_1(t), X_2(t), ..., X_n(t)$ for $t = [t_{initial}, t_{end}]$. An area widely explored in statistics is profile analysis. Profile analysis is one of many statistical approaches to test for significance of group differences. Profile analysis is a special case of Multivariate Analysis of Variance (MANOVA) in which all dependent variables are measured in the same scale [Tabachnick and Fidell, 2006]. This statistical methodology can study different scenarios depending on the experimental design. Moreover profile analysis answers questions such as parallelism, flatness and coincidence between population features using traditional multivariate analysis [Johnson, 2001]. These properties allow users to measure differences among subjects with certain statistical significance.

Another special case of profile analysis is repeated measures. Repeated measures models response variables resulting from measurements over time in which the between-subject and within-subject effect are of interest. Since data within a single subject is likely to be correlated, this repeated design accounts for it through the study of the response covariance structure [Weerahandi, 2004]. Two important assumptions of repeated measures are the properties of spherity and compound symmetry. Spherity states that the variance of paired differences (covariance) between all paired subjects is equal and compound symmetry states homogeneity in both variances and covariances [Stevens, 1999, Pan and Fang, 2002, Davis, 2002]. Lastly, observations across subjects are assumed to be independent and normally distributed [Vincent, 2005].

In our study we are interested on finding patterns over time with more complex variance structures. Hence, growth curves might seem more applicable based on their special covariance structure. Growth curves are a Generalized Multivariate Analysis of Variance Model (GMANOVA) and a special class of profile analysis. The growth curve model is:

$$Y_{p\times n} = X_{p\times m} B_{m\times r} Z_{r\times n} + E_{p\times n} \tag{2.1}$$

where $X$ is a time data model structure (within subject), $B$ is the parameters coefficients, $Z$ is the design matrix (across subjects), and $E$ as residual error. This method assumes that the columns of the error matrix $E$ are independent $p$-variate normal with mean vector equal to zero and unknown covariance matrix $\Sigma > 0$; $Y \sim \boldsymbol{N}_{p,n}(XBZ, \Sigma, I_n)$. Commonly, $p$ is the number of time points for each case,$(m-1)$ is the degree of polynomial, and $r$ is the number of treatments [Pan and Fang, 2002]. These types of models assumes non-independence among observations within a subject. It also assumes that change over time can be characterized with a special covariance structure [Maxwell and Delaney, 2003, Kleinbaum et al., 2008, Weerahandi, 2004].

Traditional methods such as profile analysis and growth curves could be useful to unravel information from our sensor-based measurements but must be adapted to deal with inherent assumptions (i.e normality). Other assumptions include special cases of covariance structure which might be difficult to estimate in several applications. Hence, these methods had disadvantages to model the sensor-based measurements of oxygen consumption discussed in this work because of the required a priori knowledge of covariance structure. Furthermore, only extraction of features to characterize this specific-domain time series was needed.

In the aims to model these unique data curves, we explore regression models. Regression also holds some assumptions such as: linearity among predictors, errors are i.i.d. normally distributed with mean zero and constant variance (homoscedasticity) and predictors are linearly independent [Stevens, 1999, Montgomery and Runger, 2006, Kleinbaum et al., 2008]. However, to uncover nonlinear relationships among the oxygen consumption curves, regression was used through splines models through is simple structure.

Regression splines are models to represent nonlinear, but unknown, mean functions [Berk, 2008]. The general model of splines is shown in Equation 5.7, for which $S_k(t)$ represents piecewise functions. Deciding how to partition the curves for piecewise modeling is another area well studied in the literature and explore in this dissertation (see change-point detection models below).

$$y(t) = \sum_{k=1}^{K} c_k S_k(t) \text{ where } t_{min} \leq t < t_{max} \tag{2.2}$$

In Equation 5.7, function $y(t)$ is modeled as $K$ piecewise functions partition along the time span, $[t_{min}, t_{max}]$. These partitions are determined in advanced based on domain knowledge or through the use of change-point detection methods. The piecewise functions, $S_k(t)$, combine in this model can be of linear, polynomial or more complex forms where $c_k$ are their respective coefficients for the entire model.

#### 2.3.1.2. Anomaly detection

Anomaly detection is often refer as outlier detection and for multiple time series this is growing research area. It is an important problem that has been investigated within diverse application domains, targeting general and specific areas such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities [Chandola et al., 2009]. Anomaly detection for multiple time series is a broad area and the literature has tackled different specific problems such as greedy-split algorithm [Chan and Mahoney, 2005], wavelets [Zhang et al., 2003] filtering and statistical methods [Soule et al., 2005].

Other methodologies were found in the literature to study multiple time series problems such as motif discovery [Keogh et al., 2006, Wei and Keogh, 2006]. Time series represented as symbols has been an interesting problem explored by Keogh's Lab [Patel et al., 2002, Yankov et al., 2007, Chiu et al., 2003]. Their symbolic representation has allowed the discovery of unique patterns among multiple time series and further exploration should be pursue. For our research we have focused on outlier detection through filtering and statistical quality control theory.

#### 2.3.1.3. Change-point detection models

The relevance of change-point detection methods for our sensor-based measurements comes from the preliminary studies on this unique dataset that revealed consistent patterns in its curves in the entire time span and at the end of spectrum. Two possible regions with different behaviors for most of the oxygen consumption curves was observed. Hence, investigating change-point detection models could help reveal these regions. Also, a noticeable "plateau" or tail at the end of all oxygen consumption curves was detected. This phenomenon is attributed to experimental time settings based on the constant volume capacities of the microwell. Hence, once it reaches zero no more oxygen is left and the sensor will record zero until the experiment ends. A peculiar observation among the curves reveals different time points on which the cells within a microwell reaches zero providing another change-point feature that characterizes the cell heterogeneity aspect.

The change-point detection problem has been widely studied concentrated on the single change-point detection and when $X_k$ are independent [Chen and Gupta, 2001]. Single-cell data are inherently time dependent and methods that account for this serial correlation are relevant. The effect of dependent sequences on change-point field was first studied by Johnson and Bagshaw focusing on a small number of serial values [Johnson and Bagshaw, 1974,Bagshaw and Johnson, 1975,Doukhan et al., 2003]. Various methods of time series models have focused on: mean level, shift in regression line, changing from one linear to another linear model. These problems have been addressed with approaches such as CUSUM, recursive residuals, wavelets, maximum likelihood estimates, likelihood ratio statistics and posterior distributions [Esterby and El-Shaarawi, 1981,Worsley, 1983].

Kokoszka and Leipus (1998) researched CUSUM type estimators of mean shift and Picard (1985) worked on testing and estimation of change parameter in a Gaussian stationary process . A generalization to Picard's method for non-Gaussian process was presented by Giraitis and Leipus [Giraitis and Leipus, 1990,Giraitis and Leipus, 1992,Shiohama and Taniguchi, 2003].

Bai (1995) studied fixed magnitude of shifts in linear regressions by deriving the limiting distribution of a steady change-point estimator by least squares method . Inference tests for change in parameter values at a given time were proposed in linear regression models with long-memory errors by Hidalgo and Robinson with a procedure with nonstochastic and stochastic regressors [Hidalgo and Robinson, 1996]. Bai (1997) extended the work on mean change to changes in multiple regressions through asymptotic properties of change-point estimator where the error process may include dependent observations . An upper bounded conservative method was developed for growth process by Ninomiya and Yoshimoto [Ninomiya and Yoshimoto, 2008]. In general, they start by fitting a model to the data without considering a change and evaluate the residuals. Then observe changes in the residuals and conclude those as change-points and test the change point found from residuals in the original data using a likelihood ratio test.

Some researchers have considered the test explained by null hypothesis $H_0$ where $X_1, ..., X_n$ have the same marginal distribution $F$ and alternate hypothesis $H_1$ where this distribution changes at unknown point $k^\star$, $F_1 \neq F_2$ where $P(X_k^{(i)} \leq x) = F_i(x)\ i = 1, 2$. This problem can be tested using the Kolmogorov-Smirnov test [Doukhan et al., 2003].

The literature presents a wide set of approaches to address the change-point detection problem, we have focused on likelihood methods and changes in mean based on particular behaviors from our data. Cumulative sum (CUSUM) control charts are able to detect small shifts from the mean more efficiently than Shewhart control charts [Montgomery, 2005]. To remove redundant information from a heterogeneous dynamical system, we automatically detected the time point where each curve was approximately zero (oxygen consumption stops) and removed all data following that point (zero-value tails) from further analysis using (CUSUM) control charts (See Chapter 4).

Additionally, the use of spline models to characterize the multiple time series in study permitted the accomplishments of multiple objectives. A change-point subroutine was embedded in the spline modeling of the curves in study. This change-point procedure was performed through exhaustive search every time point was tested as a possible change-point. The selection criteria is based on a likelihood measurement which evaluates minimum empirical error from the spline model fit. As mentioned previously, the special spline model, constrained piecewise linear regression, allowed for good empirical characterization of the oxygen consumption curves with a simple structure (See Chapter 5).

#### 2.3.1.4. Feature Selection

Feature selection (FS) is a needed step to summarize time series. Summaries obtained from the application of FS methods can enable comparisons and inferences across different conditions such as cell types. Through the modeling and change-detection of time series many features are extracted and ultimately studied collectively to determine their importance towards a specific group of interest. This is also refer as signatures identification where relevant features associated to particular groups or conditions are identified. For problem-based studies this step is crucial when analyzing data such as time series.

Feature selection models a response variable $y$ (or output), with a subset of the important predictor variables (inputs). FS methods can differ depending on the specific objective. For example, one objective is filtering to separate irrelevant inputs. Another objective is variable ranking where the interest is in the relative relevance of all input variables to the target. Finally, a compact and effective model where the goal is to identify a small subset of independent features with the most predictive power is usually of great interest [Tuv et al., 2009].

FS methods for integrative analysis needs to handle a number of challenges: heterogeneity of the data (categorical and numerical), missing values, interaction effects, nonlinear effects, etc. Current predictive models in the literature have embedded feature selection subroutines in their approach. Tree ensemble methods such as random forest [Breiman, 2001] and stochastic gradient boosting trees [Friedman, 2001] are some of those models. They provide importance ranking of the features in the model, given basis for elimination of redundant predictors such as the masking measure introduced by Tuv et. al (2009) . Furthermore on ensemble methods, are methodologies that uses multiple weak models to achieve better accuracy. The models can be incorporated in parallel, serially or mixture.

A popular parallel ensemble method is random forest formed by multiple tree models. Each tree model makes a classification, and the classification prediction is accounted as a voting system [Breiman, 2001]. The serial ensemble method is often more complex. Its objective lies on reducing both variance and bias and it often delivers excellent performance. A serial ensemble example is Adaboost algorithm introduced by Freund and Schapire in 1996 which concentrates on the training observations that are misclassified by the previous sequence [Tuv et al., 2009].

State-of-art visualization methods such as multidimensional scaling is used to illustrate selected features to discriminate among different conditions. The general objective of multidimensional scaling is to find a configuration of the data points explored through their similarities or dissimilarities [Chen et al., 2008]. Such statistical methods allows for visualization of N-dimensional space patterns across diverse conditions in a two or three dimensional plot.

### 2.3.2. Single-cell data analysis

Here we state the motivation on studying the unique data structure from single-cell technologies. It is becoming clear that cells are highly heterogeneous in both gene expression and phenotype [Andersen-Nissen et al., 2005]. A large amount of data proposes that in several disorders, such as cancer and inflammatory response-linked diseases, cellular heterogeneity causes transitions to disease states [Aniruddh et al., 2004, Babic et al., 2005, Barrett et al., 1999]. Moreover, heterogeneity triggers most failures of existing treatments for cancer. Genomics promises important achievements to cure major diseases, but it will be necessary to discover pathways involved in disease at the single-cell level, to both understand and control the inherent heterogeneity.

The main advantage of single cell analysis against traditional bulk cells biological experiments is that provides a more accurate representation of cell heterogeneity because variations can be masked out in the averages of bulk measurements [Yongzhong Li and Meldrum, 2010, Wang and Bodovitz, 2010, Chao and Ros, 2008, Arriaga, 2009]. The ability to identify these variations is essential to comprehend diseases and progress in future therapiesas. Adapted illustration shown in Figure 2 [Cohen, 2007].

Wang and Bodovitz (2010) studies have shown that average bulk results were not representative of any individual cell . These stochastic variations might occur through the production of RNA molecules from DNA transcription, the production of peptides from RNA translation, the degradation of RNA

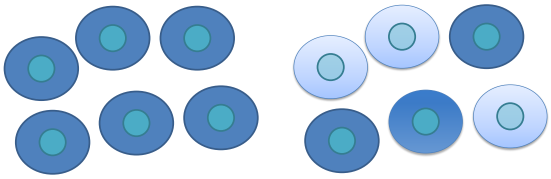


Figure 2. Cell heterogeneity.

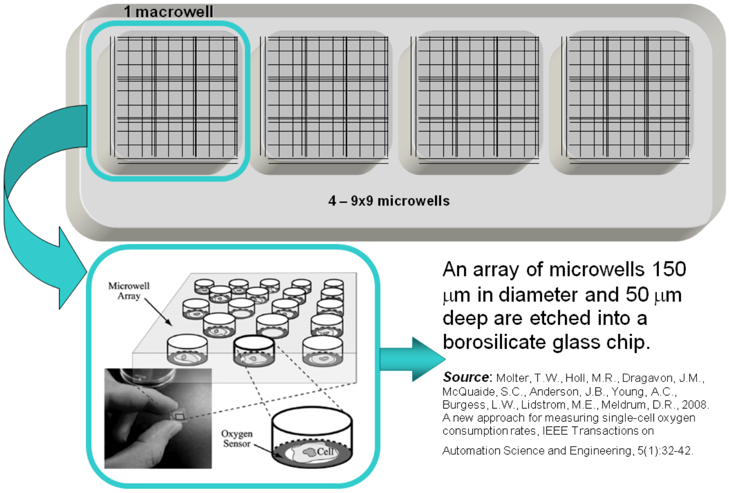
molecules, etc. [Arriaga, 2009]. Single cell analysis is a powerful methodology that: (1) reduces biological noise, (2) can target specific populations to elucidate signaling pathways and networks, (3) might reveal functional distinction between normal and cancerous cells, and many other important areas to accelerate biological research and health diagnostics [Wang and Bodovitz, 2010].

Analysis of variation at the single cell level is becoming an essential tool for understanding fundamental biological mechanisms of disease states over time. The malfunctioning of a certain cell could be determined by the over expression or suppression of a gene transcription. The suppression or over expression of a gene is believed to cause cell malfunctioning and hence be a potential indicator for health risk. This information can be obtained by microarray studies which are one of the most popular technologies used currently. Through the application of genetic and molecular biology information, microarrays studies have allowed biologists to study global gene expression in cells, over different temporal and experimental conditions, to discern key players in metabolic pathways and to assign probable function to genes [Anderle et al., 2003, Chen and Sivachenko, 2005]. Although this procedure is expensive, most molecular biology laboratories have access to high-throughput functional-genomic technology and are involved in capturing datasets consisting of tens of thousands of gene expression data points per sample measured which becomes a high dimensionality problem [Anderle et al., 2003]. Genomic profiling alone will not be sufficient to characterize the regulatory processes in the cell. Extracting relationships between temporal single-cell data at several levels (i.e., gene expression and oxygen consumption rate) could unfold important functional findings that cannot be solely discovered with only one source.

#### 2.3.2.1. Single-cell technology

The Microscale Life Sciences Center (MLSC), a National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI) Center of Excellence in Genomic Science (CEGS) headquartered at the Center of Biosignatures Discovery Automation, Biodesign Institute at Arizona State University, has developed a single cell oxygen consumption rate monitoring system with fmol/minute resolution [Molter et al., 2007, Molter et al., 2008, Molter et al., 2009, Arizona State University, 2011]. Its researchers study different types of cell models to link cell genomics to metabolic and biochemical characteristics. Traditional population-averaged physiological measurements on large numbers of cells do not adequately capture the mechanisms of disease because gene expression is highly heterogeneous and diseased cells are aberrant. Researchers in the center address cell-to-cell variations in physiological parameters by conducting studies to quantify cellular activities such as respiration and protein expression at the single-cell level.

There are many parameters of interests for the MLSC; oxygen was initially selected because of the significant correlation of this factor to cellular function [Molter et al., 2007]. Figure 3 shows the glass chip which is used to measure $O_2$ consumption rates at the single-cell level. Other sensors are being developed for multiparameters live single cell dynamic measurements [Lidstrom and Meldrum, 2003]. A chip of the size of $1 \times 1$ cm contains an array of microwells which are printed on $3$-inch borosilicate glass wafers using standard photolithography and etching techniques [Molter et al., 2008]. The research scope for this stage has focused on correlation studies of oxygen consumption measurements gathered through an experimental protocol design and developed at the MLSC headquartered at Arizona State University. Figure 4 shows the general experimental protocol and within the diagram flow the correlation studies of interest in this work are highlighted in box 4.2 of this figure. Details on the experimental procedure can be found in Chapter 5.

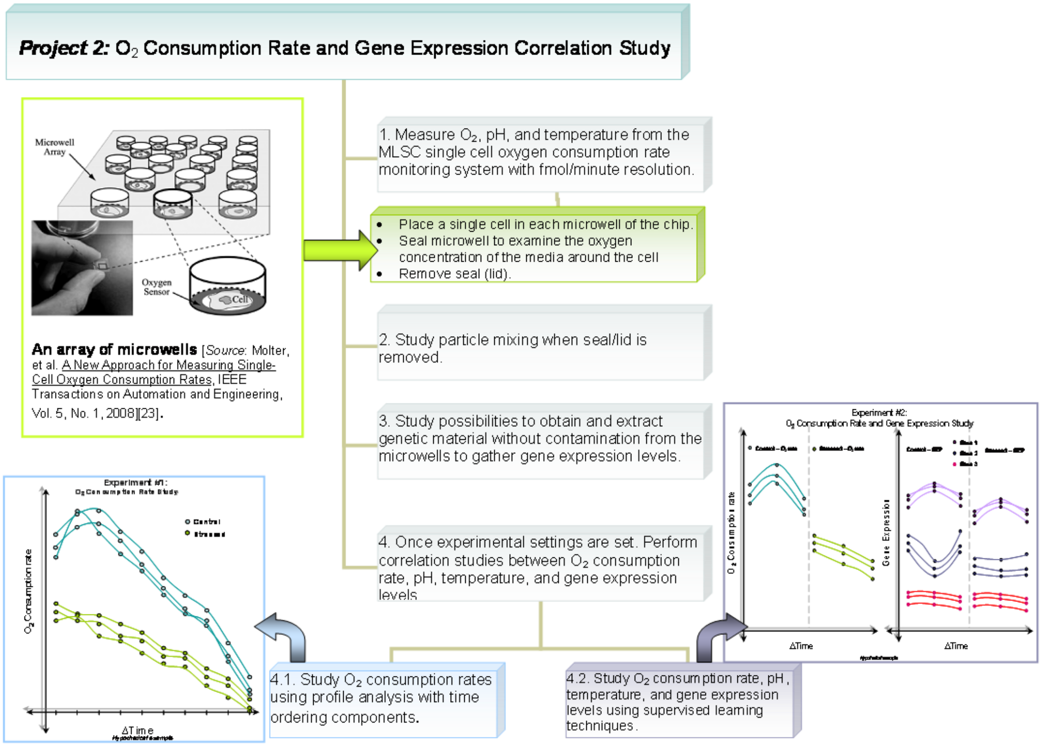


An array of microwells 150 µm in diameter and 50 µm deep are etched into a borosilicate glass chip.

***Source:*** Molter, T.W., Holl, M.R., Dragavon, J.M., McQuaide, S.C., Anderson, J.B., Young, A.C., Burgess, L.W., Lidstrom, M.E., Meldrum, D.R., 2008. A new approach for measuring single-cell oxygen consumption rates, IEEE Transactions on Automation Science and Engineering, 5(1):32-42.

Figure 3. Microwell design diagram.
An array of microwells 150 $\mu$m in diameter and 50 $\mu$m deep are etched into a borosilicate glass chip. Each microwell contains randomly seeded living cells and a platinum phosphor sensor in the shape of a ring fused to the bottom perimeter of the well (Adapted from [Molter et al., 2008]).

The single cell oxygen consumption rate measurement system consists of sealing a single cell in a microwell and examines the oxygen concentration of the media around the cell over time. Once the oxygen molecules concentration is gathered through images, the rate can be calculated based on the temporal frame



***Project 2:*** $O_2$ Consumption Rate and Gene Expression Correlation Study

1. Measure $O_2$, pH, and temperature from the MLSC single cell oxygen consumption rate monitoring system with fmol/minute resolution.

- Place a single cell in each microwell of the chip.
- Seal microwell to examine the oxygen concentration of the media around the cell
- Remove seal (lid).

2. Study particle mixing when seal/lid is removed.

3. Study possibilities to obtain and extract genetic material without contamination from the microwells to gather gene expression levels.

4. Once experimental settings are set. Perform correlation studies between $O_2$ consumption rate, pH, temperature, and gene expression levels

4.1. Study $O_2$ consumption rates using profile analysis with time ordering components.

4.2. Study $O_2$ consumption rate, pH, temperature, and gene expression levels using supervised leaming techniques.

**An array of microwells** [*Source:* Molter, et al. A New Approach for Measuring Single-Cell Oxygen Consumption Rates, IEEE Transactions on Automation and Engineering, Vol. 5, No. 1, 2008][23].

Figure 4. Single cell research plan (Array of microwells adapted from [Molter et al., 2008]).

and microwell volume [Molter et al., 2007]. Oxygen is one of the many parameters measured in the MLSC headquartered at Arizona State University, but their data structure might be similar for different sensors. These sensor-based measurements behave as a multiple time series. There are several problems associated with multiple time series that have been previously studied in the literature.

### 2.3.3. Imaging representation of multiple time series

Multiple time series occurs as a result of the single cell technology experiments. In addition to a summary of each time series, it is useful to construct features from a collection of those curves generated from common conditions (e.g., same cell type). One approach for the multiple time series problem is to represent them as images and apply two-dimensional feature methods to the collection. An interesting problem that arise from these two-dimensional representations is the ordering component of the time series. Different ordering scenarios reflects a different image and hence different features. Therefore, the use of external information for ordering could improve the image modeling of multiple time series.

In recent years, wavelet-based multiscale denoising methods have grown in popularity for images because it represents data with small number of components [Ganesan and Venkataraman, 2004]. Wavelet framework is based on the decomposition of an original signal into components. One main characteristic of the wavelet functions is that they relate to each other by simple scaling and transformation. The basic objective of wavelet research lies in finding a function that describes the signal or image of interest in a useful and informative way. Wavelet transform has been implemented in many applications such as applied mathematics and signal processing. Some of the particular advantages of wavelets decomposition are: (i) the capability to denoise frequency signals; (ii) the wavelet coefficients are decorrelated even if the input data is autocorrelated; (iii) the wavelet coefficients are normally distributed irrespective of the input data distribution; and (iv) the wavelet coefficients are stationary even if the input data is nonstationary [Ganesan and Venkataraman, 2004].

For these reasons it was suitable to implement our case study using wavelet decomposition to evaluate the two-dimensional representation performance, but in theory, this approach can be generalize towards the use of any lossless and lossy imaging compression method such as LempelZivWelch (LZW) algorithm [Welch, 1984] and Fourier-related transforms [Bovik, 2005] respectively. This general two-dimensional approach contributes to the modeling field of multiple time series providing a unique representation that results in less complex models under certain conditions.

# CHAPTER 3

# INTEGRATIVE ANALYSIS OF TRANSCRIPTOMIC AND PROTEOMIC DATA OF DESULFOVIBRIO VULGARIS: A NONLINEAR MODEL TO PREDICT ABUNDANCE OF UNDETECTED PROTEINS

## 3.1. Abstract

Gene expression profiling technologies can generally produce mRNA abundance data for all genes in a genome. A dearth of proteomic data persists because identification range and sensitivity of proteomic measurements lag behind those of transcriptomic measurements. Using partial proteomic data, it is likely that integrative transcriptomic and proteomic analysis may introduce significant bias. Developing methodologies to accurately estimate missing proteomic data will allow better integration of transcriptomic and proteomic datasets and provide deeper insight into metabolic mechanisms underlying complex biological systems. In this study, we present a nonlinear data-driven model to predict abundance for undetected proteins using two independent datasets of cognate transcriptomic and proteomic data collected from *Desulfovibrio vulgaris*. We use stochastic Gradient Boosted Trees (GBT) to uncover possible nonlinear relationships between transcriptomic and proteomic data, and to predict protein abundance for the proteins not experimentally detected based on relevant predic-

tors such as mRNA abundance, cellular role, molecular weight, sequence length, protein length, GC content and triple codon counts. Initially, we constructed a GBT model using all possible variables to assess their relative importance and characterize the behavior of the predictive model. A strong plateau effect in the regions of high mRNA values and sparse data occurred in this model. Hence, we removed genes in those areas based on thresholds estimated from the partial dependency plots where this behavior was captured. At this stage, only the strongest predictors of protein abundance were retained to reduce the complexity of the GBT model. After removing genes in the plateau region, mRNA abundance, main cellular functional categories and few triple codon counts emerged as the top ranked predictors of protein abundance. We then created a new tuned GBT model using the five most significant predictors. The construction of our nonlinear model consists of a set of serial regression trees models with implicit strength in variable selection. The model provides variable relative importance measures using as a criterion mean square error. The results showed that coefficients of determination for our nonlinear models ranged from $0.393$ to $0.582$ in both datasets, providing better results than linear regression used in the past. We evaluated the validity of this nonlinear model using biological information of operons, regulons and pathways, and the results demonstrated that the coefficients of variation of estimated protein abundance values within operons, regulons or pathways are indeed smaller than those for random groups of proteins.

## 3.2. Introduction

The last decade has seen significant growth in technologies pertaining to molecular biological assays to measure gene expression profiles. These high-throughput technologies, such as DNA microarray and Serial Analysis of Gene Expression (SAGE), have enabled the quantitative measurements of the abundance of various biological molecules and their variation between different states at the genome scale [Horak and Snyder, 2002, Smith et al., 2002, Hermeking, 2003]. However, evidence suggests that transcriptomic profiling is necessary but not sufficient to characterize biological system complexity [Gygi et al., 1999]. For example, transcript levels detected by mRNA profiling do not reflect all regulatory processes in the cell, as post-transcriptional processes, such as synthesis, processing and modification of proteins, may affect active protein concentration but are not considered. Therefore, in addition to studying gene expression at the transcriptional level, large-scale proteomic analysis should be considered as a means to understand the systems and pathways in living organisms [Nie et al., 2007]. Proteome-based expression analysis is generally performed by two-dimensional gel electrophoresis, in which proteins are separated according to their isoelectric point and mass. This technique requires intensive labor and time, and has proved effective in quantifying a cytoplasmic subset of the cellular proteome over a limited range of molecular weights and isoelectric points. In most cases, only a small set of proteins were detected [Mootha et al., 2003a,Mootha et al., 2003b,Alter and Golub, 2004]. Recent advances in gel-free proteomics technologies facilitate large-scale characterization of the proteome. High-performance liquid chromatographic (HPLC) fractionation of protein tryptic digests, followed by automated tandem mass spectrometry (MS/MS) on the peptide fragments, allows identification of several hundred or even thousands of proteins simultaneously from cellular extract [Gygi et al., 1999]. One of the major challenges in integrative analysis of large-scale transcriptomic and proteomic datasets is how to facilitate generation of new knowledge not accessible by analysis of either data type alone. In several recent studies in spite of sparse proteomic data, integrative analyses of genome-wide mRNA and protein expression patterns have enabled researchers to unravel global regulatory mechanisms and complex metabolic networks in living organisms [Hegde et al., 2003,Mootha et al., 2003a,Mootha et al., 2003b,Alter and Golub, 2004].

One of the key tasks for integrated transcriptomic and proteomic analysis is to identify relationships between protein abundance and their cognate mRNA concentrations. Although one would hypothesize that the correlation between mRNA expression levels and protein abundance will be strong based on the central dogma of molecular genetics, support from early experimental data is not immediately apparent. Most recent studies have either failed to find a correlation between protein and mRNA abundances [Gygi et al., 1999] or have observed only a weak correlation [Ideker et al., 2001, Greenbaum et al., 2002, Washburn et al., 2003]. In addition to various biological factors and limitations of current experimental protocols, it has been suggested that the poor correlation may stem from the inadequacy of available statistical tools to compensate for biases in the data collection methodologies.

While microarray analysis produces data on transcript levels for most genes in a given genome, proteomic datasets are often incomplete due to the imperfect identification of coding sequences within a genome and the limited sensitivity of current peptide detection technologies [Wilkins et al., 2006]. Current technologies allow detection of only one-third to one-half of all coded proteins [Ideker et al., 2001,Scherl et al., 2005,Zhang et al., 2006b]. In prior comparisons of transcriptomic and proteomic data, undetected proteins were often assigned a concentration value of zero, and excluded from the correlation analysis. This unrealistic simplification could adversely affect interpretation of relationships between transcriptomic and proteomic data. For instance, current technologies for proteomic analysis tend to be biased towards detection of relatively abundant proteins. Correlation patterns between transcriptomic and proteomic data for these highly-expressed genes are unlikely valid for the entire genome since correlation patterns may be different for lowly-expressed genes. Hence, improved methods of coping with missing protein abundance values are necessary for integrative analysis of transcriptomic and proteomic datasets. To address issues with the missing proteomics data, one recent tactic was to integrate Gene Ontology (GO) information into the data imputation; the approach could enhance the imputation even when the missing fraction is large [Tuikkala et al., 2006]. We also proposed a novel Zero-inflated Poisson (ZIP) regression model in which we assumed that 100p% $(0 < p < 1)$ of the genes with a proteomic abundance level of zero could be unexpressed genes or expressed genes that were undetected due to technical limitations [Nie et al., 2006a]. Thus, the proteomic abundance $(y)$ was distributed as a mixture of zeros with probability $p$ and a Poisson regression distribution with probability $(1 - p)$. Although prediction of the missing proteomic data by both GO and ZIP models has improved biological interpretation, the models' assumption that correlation patterns of transcriptomic and proteomic data are linear at the whole genome scale is not always true. For example, it has been suggested that correlations may vary in different functional categories in both prokaryotic and eukaryotic systems [Beyer et al., 2004, Nie et al., 2006b].

In this study, using two sets of cognate transcriptomic and proteomic data collected from *Desulfovibrio vulgaris*, we describe a nonlinear data-driven model to predict abundance for undetected proteins for the two datasets. We demonstrate the application of stochastic gradient boosted trees to uncover possible nonlinear relationships between transcriptomic and proteomic data. The idea is to create regression boosted trees to predict protein abundance based on several relevant predictors in both datasets: mRNA abundance, cellular role, molecular weight, sequence length, protein length, GC content, and triple codon counts in both datasets. To compare the general behavior of these factors across different experimental conditions within same species, the results are stratified into several parts: (1) variable (predictor) importance and partial dependency plots, (2) construction of the model, and (3) validation using biological information.

## 3.3. Materials and methods

### 3.3.1. Datasets

We analyzed two datasets from *Desulfovibrio vulgaris*. The experimental conditions differed between the datasets as they were obtained by independent research [Zhang et al., 2006b, Zhang et al., 2006a, Mukhopadhyay et al., 2006, Heidelberg et al., 2004]. A brief description of both datasets is provided below. We normalized the raw intensity values from both datasets with a quantile normalization using an R package (caret) available through the R project (http://www.r-project.org/). Table 1 and the following sections provide a brief description of Dataset 1 and Dataset 2 used throughout this paper.

#### 3.3.1.1. Dataset 1

The dataset consists of the whole-genome mRNA expression and LC-MS/MS proteome abundance data from *Desulfovibrio vulgaris* in two different growth stages -log and stationary- and under two distinct types of media: lactate- or formate-based. To minimize variations between microarray and proteomic measurements, identical cell samples from each growth condition were split and used to isolate both the RNA and proteins for analyses. Complete descriptions of the experimental designs and microarray and proteomic data collection methods are given elsewhere [Nie et al., 2006b, Zhang et al., 2006b, Zhang et al., 2006a]. Briefly, oligonucleotide microarrays containing 3507 ORFs of the *Desulfovibrio vulgaris* genome were designed by NimbleGen Systems, Inc. (Madison, WI) [Nuwaysir et al., 2002, Heidelberg et al., 2004]. For each experimental condition, mRNA abundances were determined from the average of four measurements for each gene: two replicates (each containing a pool of three biological replicates) that were each hybridized to duplicate microarrays [Zhang et al., 2006a]. Proteomic analysis was performed on a Finnigan model LTQ ion trap mass spectrometer (ThermoQuest Corp., San Jose, CA). The relative protein abundance was estimated based on the number of peptide hits [Qian et al., 2005]. The number of peptide hits for a given protein was the median of three LC-MS/MS measurements. The protein abundances ranged from one to several hundred [Zhang et al., 2006a].

#### 3.3.1.2. Dataset 2

The dataset consists of the whole-genome mRNA expression and LC-MS/MS proteome abundance data from *Desulfovibrio vulgaris* grown under two stress conditions which are 250 mM NaCl or KCl [Heidelberg et al.,

TABLE 1
Description of the datasets used in this study.

| | Dataset 1 | Dataset 2 |
|---|---|---|
| References | Zhang et. Al., 2006a, 2006b | Mukhopadhyay et al., 2006 |
| Conditions | Formate Log (FL)<br>Formate Stationary (FS)<br>Lactate Log (LL)<br>Lactate Stationary (LS) | Control $t = 0$ hrs(CT0)<br>Control $t = 120$ hrs(CT120)<br>Stressed NaCl $t = 120$ hrs(ST120) |
| # Variables | 70 | 70 |
| # genes analyzed | 456 (FL), 477 (FS),<br>440 (LL), and 462 (LS) | 2146 for all conditions |
| # replicates (mRNA abundance) | 4/gene | 3/gene; except ST120: 2/gene |
| # of replicates (protein abundance) | 3/gene | 2/gene |
| Number of genes removed using a threshold t | 42 (FL), 477$*$ (FS),<br>59 (LL), and 28 (LS) | 42 (CT0), 26 (CT120),<br>and 19 (ST120) |

$*$ Condition FS was eliminated from further study based on biological knowledge provided by the experts. More details on Section 3.3: Materials and Methods.

2004, Mukhopadhyay et al., 2006]. Briefly, spot signals, spot quality, and background fluorescence intensities of the microarray were quantified with ImaGene, version 5.5 (Biodiscovery Inc., Los Angeles, CA) (Raw microarray data of this dataset can also be found in NCBI, GEO accession number GSE4447). Replicate cultures from a control (time zero and $120$ min) and a stressed sample ($120$ min) were used to obtain total protein. A total of $1,356$ proteins were identified in all samples, and for $47$ of these proteins there were reproducible changes between the control and the stressed sample http://vimss.lbl.gov/SaltStress/ [Mukhopadhyay et al., 2006].

#### 3.3.1.3. Quality of datasets

The quality of both datasets was assessed by calculating Pearson correlation coefficients among multiple replicates for microarray and protein measurements. Dataset 1 shows that correlation coefficients of the microarray experiments are from $0.97$ to $0.99$ among replicate samples [Nie et al., 2006a, Nie et al., 2006b] and correlation coefficients of LC-MS/MS measurements normalized by amino acid composition are $0.86$ to $0.92$ among replicates, indicating good reproducibility. Similarly for Dataset 2, normalized microarray measurements showed correlation coefficients between $0.86$ and $0.96$ among replicates and a tight range of $0.96$ to $0.98$ for correlations between protein abundance samples for all conditions. In terms of correlation of mRNA and protein abundance using Pearson correlation, low values were found in both datasets. For Dataset 1, correlation between mRNA expression and normalized protein abundance was modest: $0.54$ to $0.63$ (P-value, $0.001$) by Pearson correlation coefficient for all conditions. Dataset 2 reflected correlation values from $0.33$ to $0.48$. These correlation levels are similar to those previously reported for yeast [Ideker et al., 2001]. The relatively poor correlation between mRNA and protein abundance suggests the fallacy of assumption of linearity in relationship between variables.

### 3.3.2. Genome information

#### 3.3.2.1. Cellular functional category

The cellular functional categories of all genes in the *Desulfovibrio vulgaris* genome were downloaded from the Comprehensive Microbial Resource (CMR) of The Institute for Genomic Research (TIGR) at http://cmr.tigr.org [Heidelberg et al., 2004]. On the basis of the original annotation, the genes and proteins are classified into 20 cellular functional categories. These categories were included in the model as possible predictors of protein abundance.

#### 3.3.2.2. Other predictor factors

Gene annotated attributes such as sequence length, protein length, molecular weight, GC content, and triple codon counts of all genes in the *Desulfovibrio vulgaris* genome were downloaded from the TIGR resource and included in our study. Continuous numerical values were gathered for the molecular weight of each gene. The GC content reflected the proportion of nucleotides G or C in the *Desulfovibrio vulgaris* genome. The triple codon information included counts for all 64 triple codon combinations in the genetic code.

#### 3.3.2.3. Operon and pathway information

The complete genome of *Desulfovibrio vulgaris* and its ORF calls and annotation were downloaded from NCBI Genbank, the TIGR resource. Genes transcribed in the same direction having intergenic regions less than 15 base pairs were defined as one operon. Although a new method has been proposed to define operons by combining intergenic distances with comparative genomic measures [Alm et al., 2005, Price et al., 2006], we opted for the distance-only approach, a relatively low threshold, to cover more of the possible operons. With this relatively low threshold, a total of 609 operons, ranging from 2 to 13 genes each, were identified in *Desulfovibrio vulgaris*. (Gene list of all operons is available upon request). The list of *Desulfovibrio vulgaris* regulons was kindly provided by Prof. Judy Wall and Dr. Chris Hemme of the Department of Biochemistry at the University of Missouri at Columbia (The regulons were identified based on their homology to the known Escherichia coli regulons) [Hemme and Wall, 2004]. Gene lists of ninety-two metabolic pathways defined for microbial genomes of interest were downloaded from the KEGG database(http://www.genome.jp/kegg/kegg2.html).

### 3.3.3. Construction of non-linear relationship model

To satisfy the need for a method amenable to mixed data types and capable of unraveling nonlinear relationships between the data previously discussed, we applied stochastic Gradient Boosted Trees (GBT) as described by [Friedman, 2002]. These models have been used in a wide range of applications such as ecological modeling and prediction, chemical concentration on rocks, and demographic survey data [De'ath, 2007, Elith et al., 2008, Friedman, 2001]. Our objective was to find an approximated function that could map a set of input variables $x$ = $x_1, ..., x_n$ to the response output y in such a way that the expected value of empirical loss was minimized as shown in Equation 4.4. Boosting fits a weighted additive expansion composed of weak classifiers (e.g., regression trees) that approximates the response y as in Equation 4.5 [Hastie et al., 2001]. Gradient boosting sequentially applies regression trees to fit residuals while minimizing squared error loss, creating new models which are encouraged to become experts in cases misclassified by previous trees.

$$\hat{y} = \arg\min_{y} E_{y,X} L(y, \hat{y}) \tag{3.1}$$

$$\hat{y} = \sum_{m=0}^{M} \beta_m T(X; \hat{\Theta}) \tag{3.2}$$

These individual trees partition the space of joint predictor variable values into disjoint regions $R_j$ with constant predictor values $j$ assigned to each region. A single tree can be formally expressed as a piecewise constant function as described in Equation 3.3. The parameter space delta is estimated by minimizing empirical risk as in Equation 3.4. To find disjoint regions and constants that minimize a particular empirical risk is a large combinatorial problem. There are several optimization methods to achieve this. The method used in this study uses a gradient approach implemented from R.

$$T(X; \Theta) = \sum_{j=1}^{J} \gamma_j I(X \epsilon R_j) \text{ where } \Theta = R_j, {\gamma_j}_1^J \tag{3.3}$$

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{x_i} \epsilon R_j L(y_i, \gamma_i) \tag{3.4}$$

The method described previously was implemented using the gbm R package available from the R project (http://www.r-project.org/). The required inputs include: loss function, number of trees, the depth of each tree, shrinkage rate, and number of folds for cross validation [Ridgeway, 2007]. Squared error loss was used as the loss function in the construction of the models for all conditions based on preliminary results where squared error and absolute loss performance were compared. The number of trees in each model was chosen to be $500$, as this is considered sufficient iteration to achieve optimality [Friedman, 2001]. To capture some degree of variable interaction a depth value of three was chosen to balance out the model complexity. For shrinkage rate we chose the recommended default of $0.005$ [Ridgeway, 2007] since we did not focus on the regularized aspect of the models. The models were specified to be built using five cross validated folds.

Cross validation is a technique for model assessment which includes randomization. Input data is partitioned into $K$ equal parts where $K - 1$ sets are used to train the model and the other unseen set is used to calculate prediction errors [Hastie et al., 2001]. This is repeated $K$ times, yielding $K$ prediction errors values, one computed at every fold. An average and standard deviation can be extracted to select the most representative model for future prediction. Once the best model has been selected based on cross validation, it is evaluated based on its coefficient of determination ($R_2$) which represents the variation explained by the model. The coefficient of determination ($R_2$) is a statistical measure representing the percentage of variance explained by the model. $R_2$ values ranges from zero to one. The closer the $R_2$ to one the better the model is explaining the variance of the data. Furthermore, as an alternative means to assess the goodness of the model, we studied the predictions of small sets of genes grouped based on pathway, operon and regulon information. In order to describe the variation within a data set, such as 'molar abundance' of proteins within one operon, we computed the coefficient of variation $(CV)$ for each set of proteins. The $CV$ is defined as the ratio of the standard deviation and the mean of the 'molar abundance' for a set of proteins [Johnson, 2005, Nie et al., 2006a] and is independent of the sample size. These coefficients of variation are computed for all pathway, operon and regulon groups and compared to a distribution of permuted CVs where permutation of genes is performed.

## 3.4. Results and discussion

### 3.4.1. Variable importance and partial variable dependence

The objective was to predict protein abundance based on the most relevant predictors. We used GBT model to uncover possible nonlinear relationships between transcriptomic and proteomic data and to incorporate categorical predictors. In a previous study using multiple regressions, [Nie et al., 2006b] found that mRNA abundance alone can explain only $20 - 28\%$ of the total variation of protein abundance, suggesting mRNA-protein correlation can not be determined solely on the basis of mRNA abundance. Other possible predictors of protein abundance include cellular role of genes, GC content and codon usage of genes, length of genes and proteins, and molecular weight of proteins [Nie et al., 2006b, Nie et al., 2006c, Nie et al., 2007].

GBT provided the implicit feature importance measures (for only the ten top ranked variables) shown in Table 2 for both Dataset 1 and Dataset 2. The relative importance measure is computed by measuring the contribution of an input variable based on its improvement on squared error loss at each tree for all trees and computing its average. This is done for all input variables. The relative influence value for a specific variable is presented as percentages of its relative contribution among all variables. Relative importance of variables measures for all seventy variables can be available upon request. Cellular role and mRNA expression level were the best predictors of protein abundance across conditions and datasets. Some triple codon sequences appear to be more relevant in modeling protein abundance than sequence length, protein length and molecular weight. These triple codon counts differ in ranking across datasets but retain similar ranking within dataset conditions.

Our findings support the known correlation of mRNA and protein abundances. Besides the variable importance measures acquired from the boosted trees, partial dependency plots were studied to gain further insight into the association of mRNA abundance with protein measurements. The partial dependency plots provide a prediction model for a given predictor variable averaged across all other predictors. Fig 1 shows prediction values for given values of mRNA for different experimental conditions for Dataset 1 (Figure 5a) and Dataset 2 (Figure 5c). Though both datasets show increasing functions, slightly different relationships are observed across datasets, with similar behavior across conditions within a dataset. Both datasets exhibit a “plateau effect” for high values of mRNA. The plateau occurs in regimens where protein abundance data is sparse with high variance where the tree models do not generate splits among the predictors. For example, in the region of high mRNA values there are a small number of genes/peptides whose protein values range from $(0, 40)$ for Dataset 1 and $(0, 500)$ for Dataset 2. This could reflect problems with the accuracy and sensitivity of current proteomic technologies.

After removing those genes/peptides with high mRNA values, the model provided a more realistic fit. The cutoff threshold was obtained as the value where the plateau starts in partial dependency plots. A different threshold is obtained for each dataset. The minimal threshold value for all conditions is $5150$

for Dataset 1 and 37975 for Dataset 2 (both in terms of absolute fluorescence intensity in the single color DNA array). Figure 5 shows partial dependency plots after eliminating genes with mRNA values higher than the corresponding cutoff threshold for Dataset 1 (Figure 5b) and Dataset 2 (Figure 5d). This provides a more realistic prediction model of protein based on mRNA. The partial dependency plots observed in Figure 5b, d) show an increasing predictive function for protein abundance as mRNA values increase. However, the curves exhibit variable slope, suggesting nonlinear modifiers to the typical linear relationship. Gradient boosted trees were rebuilt using the five most important features across conditions and after removing genes/peptides having high mRNA values. These modified models were used to predict protein abundance for genes/peptides with undetected protein values.

### 3.4.2. Construction of the non-linear correlation model

Initially, our GBT model was built using all variables to assess variable importance and to predict model behavior. Based on the plateau in regions of high mRNA values, we removed genes in those areas based on thresholds estimated from the partial dependency plots where this behavior was captured. At this stage, our aim was to reduce model complexity by selecting the most

TABLE 2

Measurements of relative importance of variables for the ten top ranked variables (after removing genes with high mRNA).

| Dataset 1 | | | | | |
|---|---|---|---|---|---|
| FL | | LL | | LS | |
| Variables | VRI (%) | Variables | VRI (%) | Variables | VRI (%) |
| mRNAmean | 24.002 | mRNAmean | 28.877 | mRNAmean | 50.473 |
| Cellular Role | 11.609 | Cellular Role | 16.694 | GCT | 14.916 |
| AAG | 9.105 | AAG | 7.785 | Cellular Role | 10.247 |
| ACC | 8.879 | GGC | 5.277 | GTT | 5.912 |
| GCT | 6.378 | GTT | 4.888 | AAG | 5.278 |
| GGT | 4.379 | GCT | 4.536 | GGT | 2.595 |
| GGC | 3.208 | ACC | 3.561 | ACC | 1.103 |
| GTT | 3.14 | ATC | 2.992 | TGC | 1.031 |
| ATG | 2.194 | ATG | 2.444 | GAA | 0.986 |
| GCG | 2.069 | GCC | 2.041 | GTG | 0.904 |
| Dataset 2 | | | | | |
| CT0 | | CT120 | | ST120 | |
| Variables | VRI (%) | Variables | VRI (%) | Variables | VRI (%) |
| mRNAmean | 38.398 | mRNAmean | 21.613 | mRNAmean | 25.149 |
| GGT | 17.351 | AAG | 13.941 | AAG | 15.046 |
| AAG | 6.532 | Cellular Role | 11.59 | GGT | 10.861 |
| Cellular Role | 6.426 | GGT | 7.291 | ACC | 6.78 |
| GTT | 5.421 | GTT | 6.294 | Cellular Role | 6.637 |
| GGG | 3.596 | GGG | 3.977 | GCT | 3.72 |
| GAA | 2.641 | CGT | 3.408 | AAC | 2.545 |
| ACC | 1.98 | GCT | 3.289 | GGG | 2.053 |
| TAC | 1.869 | GAA | 2.442 | ACG | 2.047 |
| GCT | 1.744 | ACC | 2.264 | CCC | 1.983 |

Results show two datasets where VRI specifies variable relative importance. Relative influence is computed as the average of empirical improvements in squared error of splitting the decisions trees with corresponding variable. This measure is divided by the sum of the empirical improvement of all variables in the model. The percentage of this measure is VRI. More details discussed in Section 3.4: Results and Discussion.

relevant predictors of protein abundance. As discussed previously, mRNA abundance, main cellular functional categories and few triple codon counts were top ranked after removing genes in the plateau region (See Table 2). A new, tuned GBT model was then built using the five most significant predictors. Protein abundance predictions using these newly tuned boosted trees for all conditions in both datasets are available upon request.

These predictions are depicted in Figure 6(a, c) for Dataset 1 and Dataset 2 respectively. The behavior of both datasets predictions' when plotted only against mRNA is similar, showing a large number of genes/peptides with low fitted protein values and less variability. For higher values of mRNA the magnitude and variability of the predicted protein concentration increases. Similar behavior is noted in Figure 6(b, d) were protein abundance measures where predicted for the genes/peptides with undetected values of protein abundance for both datasets.

### 3.4.3. Validation of prediction by external biological knowledge

External biological knowledge was invoked to validate the prediction of protein abundance values for the undetected proteins. The information used included gene organization information such as operon, and gene function information such as regulon and pathway. We tested the mode prediction by assuming that relationships between genes in operons, regulons and pathways are tighter than those between random gene sets. The information used for validation purposes is described in Section 3.3.2. The validation was conducted by calculating the coefficient of variation ($CV$) within conditions for every operon, regulon and pathway of *Desulfovibrio vulgaris* for both Dataset 1 and Dataset 2. These groups of genes are thought to have less dispersion than a random set of genes by virtue of their intrinsic biological relationship. Table 3 provides an example of these results for the operon groups for both datasets. The complete data for operons, regulons and pathways is available upon request. To compare CV values we also performed a permutation test in the following way. A CV was computed from the protein prediction values for a set of randomly selected genes. This step was repeated a thousand times through resampling of genes without replacement.

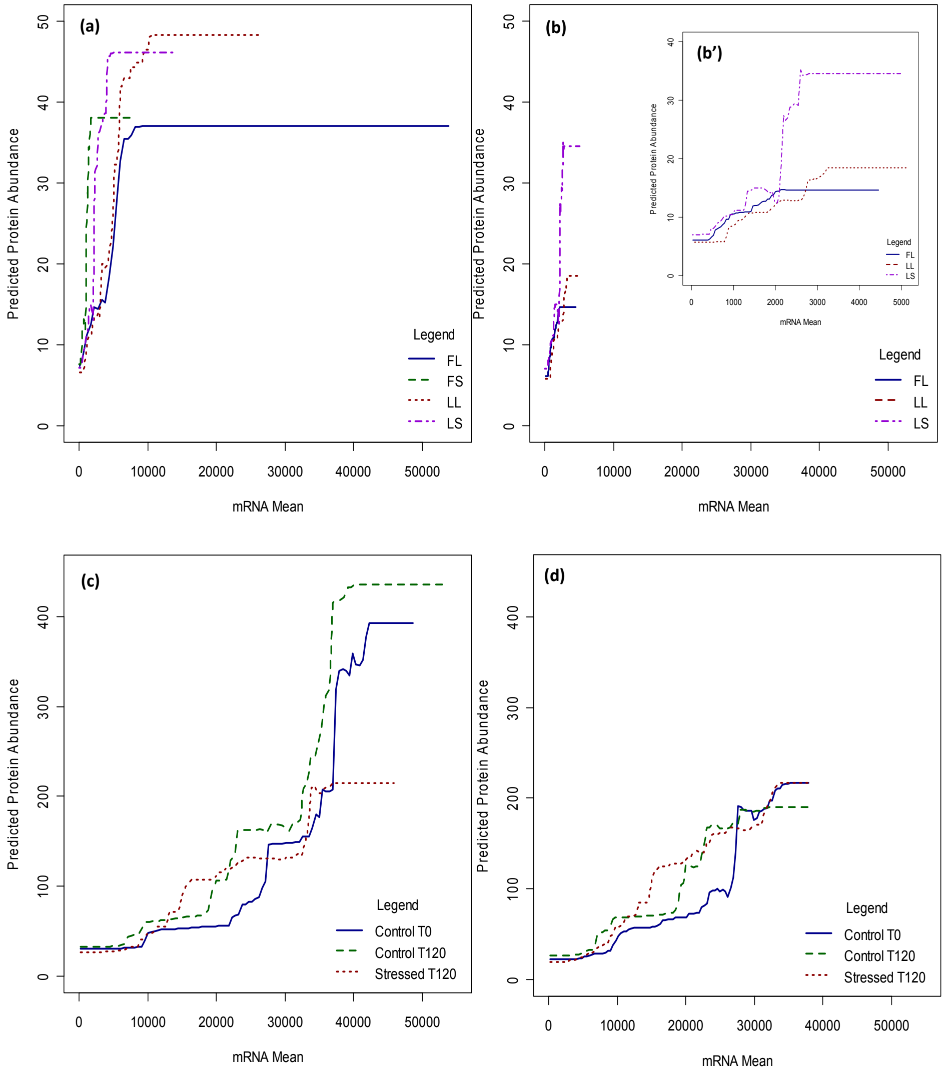


Figure 5. Partial dependency plots: mRNA versus protein prediction in *Desulfovibrio vulgaris*.

(a, c) Partial dependency prediction values for given values of mRNA for Dataset 1 and Dataset 2, respectively. (b, d) Partial dependency plots for mRNA values after eliminating genes with mRNA values higher than the corresponding cutoff threshold (for Dataset 1 and Dataset 2, respectively. (b') a zoom view to partial dependency plot for plot (b).

For example, operon $19$ contains twelve genes ($DVU0861 - 0872$). Its CV value was compared to a CV value generated through permutations where twelve genes were selected at random from the whole genome dataset (without repeating any genes) and its condition-specific prediction values were used to calculate a single CV value. Repeating this calculation a thousand times provided a CV-distribution to calculate mean, standard deviation and percentile scores for groups with random genes per condition. As a result, the CV value for this operon was $0.335$ for condition LL in Dataset $1$ and the mean of the CV values through permutations was equal to $0.769$ as shown in Table 3. Similarly, pathway path_dvu00052 (galactose metabolism) contains ten genes and its CV value was smaller than the mean of CV values through permutations $(0.431 < 0.996)$ for condition ST120 in Dataset 2. This was done in the same way for all conditions in both datasets. As shown in Table 4, for Dataset 1 $75\%$ to $79\%$ of the operon groups had smaller CV values than those computed through permutation, and $79\%$ to $88\%$ of the pathway groups had smaller CV values than those computed through permutation. However, a shift to smaller proportions for regulon groups was observed with values between $50\%$ and $67\%$. Similar results are presented Table 4 for Dataset 2. This shows that a large proportion of the biologically related groups are indeed less dispersed than unrelated groups of genes, providing some measure of validation for the predictions of our models. Furthermore, CV values from almost all operons groups were smaller than those by Zero-Inflated Poisson Regression Model [Nie et al., 2006a], suggesting the GBT model described the dataset better.

To gather more detailed information on how the CV compares with the distribution of the permuted coefficients of variation, we also calculated the percentile score. Operon 19 for LL condition in Dataset 1 showed a percentile score of $0.02$ which provides information of the position of its CV across the CV values computed through permutations. The percentile score presented is a measure of the position of the biological group CV within the thousand CV values from permutations in percentage. Because operon $19$ had a percentile score of $0.02$ this implies that $98\%$ of the thousand CV values from permutations were greater than operon $19$ CV value. Likewise, pathway path_dvu00052 showed a small percentile score of $1\%$ for ST120 condition in Dataset 2. Based on the thought that genes from pathway, operon and regulon groups should be less dispersed than permuted sets of genes, the percentile scores are expected to be very low. The calculated CV for most groups was less than the mean CV value for permuted sets of genes. For the percentile scores about half of these groups fall within a percentile less than $0.20$ as shown in Table 4. A similar trend was found when compared to the mean of permuted dispersion.

In addition, using the predicted values for each of the operon, pathway and regulon groups, we calculated the protein-mRNA correlation of these groups and compared it with the overall correlation at whole genome level. The results showed relatively strong protein-mRNA correlation for most of gene/protein pairs within operons and pathways groups for both datasets (See Figure 7 and Figure 8). Among them, pathway groups showed stronger correlation in general. About ten of these pathways groups revealed perfect correlation. However, only a small percentage of regulon groups portrayed a solid correlation. The observation that regulon groups had greater percentile values than pathway and operon groups and smaller correlation values may reflect the fact that the relationship between genes/proteins in regulons is more complicated than those in operons and pathways, and that regulon group information is also less defined and validated experimentally.

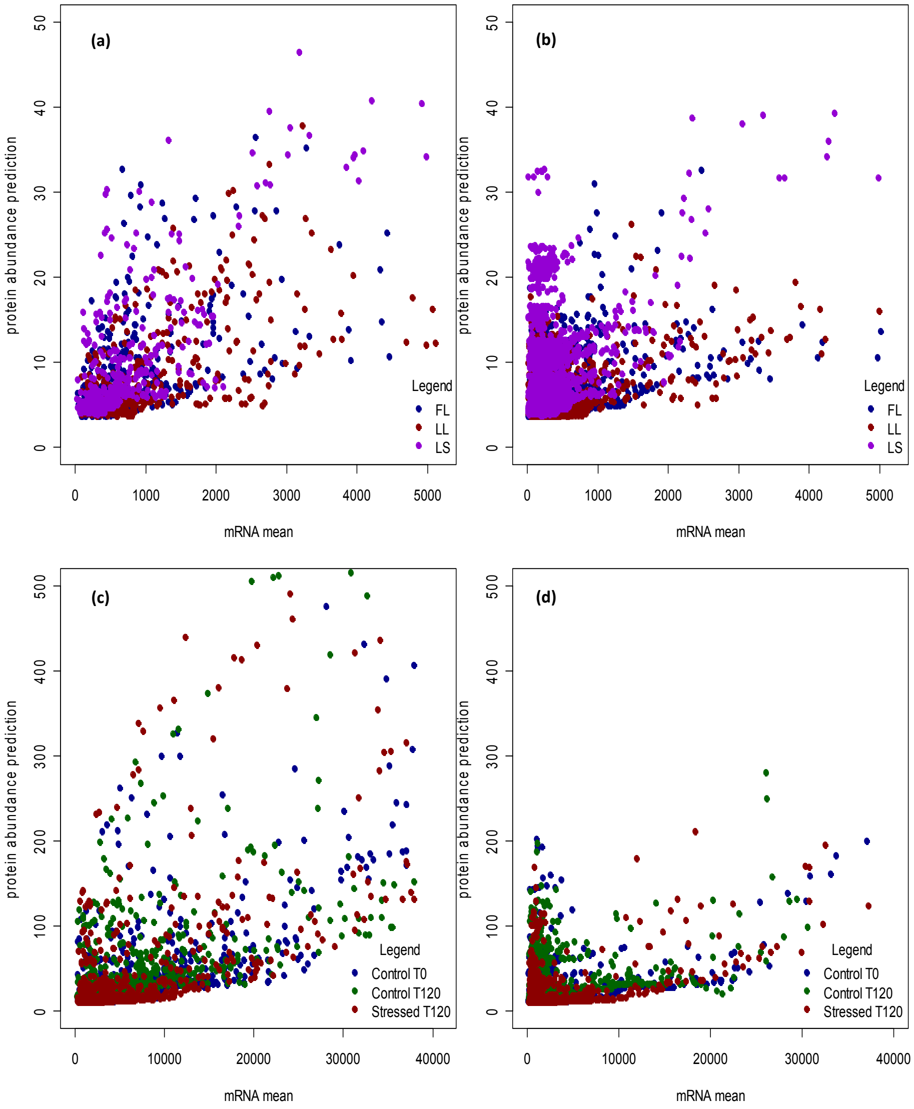


Figure 6. Prediction plot for undetected proteins.

(a, c) Protein prediction values for genes with protein values detected and used in model for Dataset 1 and Dataset 2 respectively; (b, d) protein prediction values for genes with undetected protein values for Dataset 1 and Dataset 2 respectively.

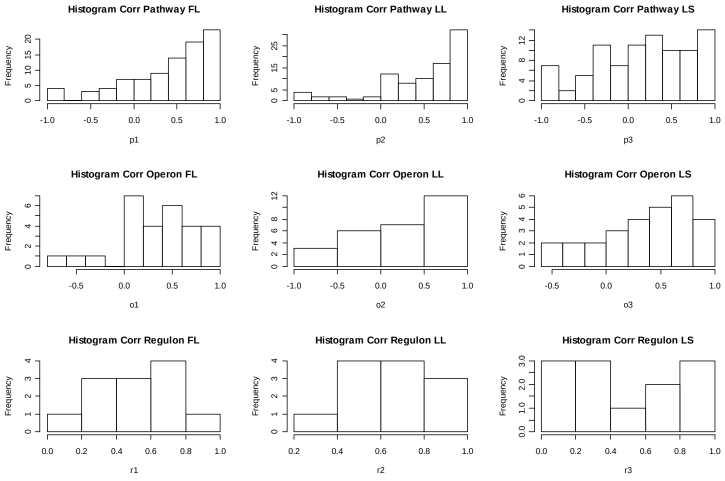


Figure 7. Histogram of correlations per biological group for Dataset 1.
Overall Pearson correlation between mRNA expression and normalized protein abundance was $0.54$ to $0.63$ for Dataset 1.

## 3.5. Conclusion

High-throughput experimentation measuring mRNA and protein expression provides rich sources of information for better understanding of the metabolic mechanisms underlying complex biological systems. The goal of this investigation, as well as our previous study [Nie et al., 2006a] is to address the problem of incomplete proteomic datasets by using statistical approaches. In the two datasets we used in this analysis, the number of undetected proteins is $3050$, $3061$, and $3057$ for FL, LL, and LS conditions, respectively for Dataset

TABLE 3

Model validation: correlated expression of proteins in some operons groups.

| | Operon | Dataset 1 | | | | | | Dataset 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FL | | LL | | LS | | CT0 | | CT120 | | ST120 | |
| | | $CV$ | $\overline{PCV}$ | $CV$ | $\overline{PCV}$ | $CV$ | $\overline{PCV}$ | $CV$ | $\overline{PCV}$ | $CV$ | $\overline{PCV}$ | $CV$ | $\overline{PCV}$ |
| 1 | DVU3025-3033 | 0.515 | 0.634 | 0.417 | 0.735 | 0.441 | 0.771 | 0.582 | 1.107 | 0.579 | 0.973 | 0.751 | 1.122 |
| 2 | DVU2399-2405 | 0.436 | 0.584 | 0.529 | 0.652 | 0.776$^\dagger$ | 0.68 | 0.932 | 1.011 | 0.256 | 0.877 | 1.915$^\dagger$ | 1.013 |
| 3 | DVU2072-2078 | 0.368 | 0.61 | 0.508 | 0.691 | 0.755$^\dagger$ | 0.73 | 1.490$^\dagger$ | 1.011 | 1.438$^\dagger$ | 0.915 | 0.618 | 1.013 |
| 4 | DVU1286-1291 | 0.338 | 0.584 | 0.394 | 0.652 | 0.54 | 0.68 | 0.376 | 0.967 | 0.381 | 0.877 | 0.373 | 0.952 |
| 5 | DVU0429-0434 | 0.47 | 0.584 | 0.417 | 0.652 | 0.601 | 0.68 | 0.504 | 0.967 | 0.501 | 0.877 | 0.481 | 0.952 |
| 6 | DVU0145-0150 | 0.482 | 0.584 | 0.401 | 0.652 | 0.343 | 0.68 | 0.472 | 0.967 | 0.417 | 0.877 | 0.518 | 0.952 |
| 7 | DVU1080-1085 | 0.329 | 0.584 | 0.313 | 0.652 | 0.497 | 0.68 | 0.307 | 0.967 | 0.423 | 0.877 | 0.382 | 0.952 |
| 8 | DVU2791-2798 | 0.29 | 0.615 | 0.29 | 0.702 | 0.448 | 0.731 | 0.507 | 1.082 | 0.346 | 0.945 | 0.267 | 1.058 |
| 9 | DVU1627-1634 | 0.588 | 0.615 | 0.441 | 0.702 | 0.732$^\dagger$ | 0.731 | 0.675 | 1.082 | 0.837 | 0.945 | 0.778 | 1.058 |
| 10 | DVU1421-1428 | 0.602 | 0.61 | 0.575 | 0.691 | 0.756$^\dagger$ | 0.73 | 0.702 | 1.011 | 0.659 | 0.915 | 1.737$^\dagger$ | 1.058 |
| 11 | DVU2978-2985 | 0.296 | 0.615 | 0.311 | 0.702 | 0.417 | 0.731 | 0.685 | 1.082 | 0.599 | 0.945 | 1.567$^\dagger$ | 1.058 |
| 12 | DVU1191-1211 | 0.649 | 0.684 | 0.75 | 0.828 | 0.67 | 0.884 | 2.305$^\dagger$ | 1.369 | 1.503$^\dagger$ | 1.091 | 1.508$^\dagger$ | 1.383 |
| 13 | DVU2558-2563 | 0.383 | 0.584 | 0.53 | 0.652 | 0.395 | 0.68 | 0.682 | 0.967 | 0.6 | 0.877 | 0.594 | 0.952 |
| 14 | DVU1242-1249 | 0.277 | 0.615 | 0.338 | 0.702 | 0.527 | 0.731 | 0.351 | 1.082 | 0.381 | 0.945 | 0.837 | 1.058 |
| 15 | DVU1552-1560 | 0.689$^\dagger$ | 0.634 | 0.828$^\dagger$ | 0.735 | 0.711 | 0.771 | 1.002 | 1.107 | 0.905 | 0.973 | 0.943 | 1.122 |
| 16 | DVU0460-0471 | 0.23 | 0.661 | 0.215 | 0.769 | 0.21 | 0.782 | 0.494 | 1.17 | 0.254 | 1.017 | 1.138 | 1.227 |
| 17 | DVU0646-0651 | 0.223 | 0.584 | 0.316 | 0.652 | 0.426 | 0.68 | 0.133 | 0.905 | 1.480$^\dagger$ | 0.877 | 1.589$^\dagger$ | 0.952 |
| 18 | DVU1908-1914 | 0.49 | 0.61 | 0.507 | 0.691 | 0.263 | 0.73 | 0.501 | 1.011 | 0.406 | 0.915 | 0.534 | 1.013 |
| 19 | DVU0861-0872 | 0.198 | 0.661 | 0.335 | 0.769 | 0.333 | 0.782 | 1.199$^\dagger$ | 1.17 | 0.771 | 1.017 | 0.552 | 1.227 |
| 20 | DVU1448-1453 | 1.081$^\dagger$ | 0.584 | 0.633 | 0.652 | 0.694$^\dagger$ | 0.68 | 0.601 | 0.967 | 0.638 | 0.877 | 0.561 | 0.952 |
| 21 | DVU1038-1044 | 0.670$^\dagger$ | 0.61 | 0.787$^\dagger$ | 0.691 | 0.712 | 0.73 | 0.661 | 1.011 | 0.776 | 0.915 | 0.998 | 1.013 |
| 22 | DVU1585-1590 | 0.394 | 0.584 | 0.338 | 0.652 | 0.464 | 0.68 | 0.433 | 0.967 | 0.314 | 0.877 | 0.467 | 0.952 |
| 23 | DVU1045-1052 | 0.675$^\dagger$ | 0.615 | 0.751$^\dagger$ | 0.702 | 0.667 | 0.731 | 1.278$^\dagger$ | 1.082 | 0.8 | 0.945 | 0.854 | 1.058 |
| 24 | DVU1278-1284 | 0.149 | 0.61 | 0.366 | 0.691 | 0.423 | 0.73 | 0.757 | 1.011 | 0.57 | 0.915 | 0.878 | 1.013 |
| 25 | DVU0807-0813 | 0.712$^\dagger$ | 0.61 | 0.739$^\dagger$ | 0.691 | 0.614 | 0.73 | 1.052$^\dagger$ | 1.011 | 1.061$^\dagger$ | 0.915 | 1.042$^\dagger$ | 1.013 |
| 26 | DVU1344-1350 | 0.739$^\dagger$ | 0.61 | 0.810$^\dagger$ | 0.691 | 0.717 | 0.73 | 0.773 | 1.011 | 0.712 | 0.915 | 0.831 | 1.013 |
| 27 | DVU1301-1330 | 0.529 | 0.695 | 0.733 | 0.855 | 0.733 | 0.948 | 0.753 | 1.499 | 0.628 | 1.123 | 0.925 | 1.526 |
| 28 | DVU2529-2537 | 1.057$^\dagger$ | 0.634 | 0.824$^\dagger$ | 0.702 | 1.856$^\dagger$ | 0.771 | 2.061$^\dagger$ | 1.107 | 1.375$^\dagger$ | 0.973 | 2.207$^\dagger$ | 1.122 |

Coefficient of Variation ($CV$) is computed by dividing standard deviation by the mean of the prediction values for protein abundance for a specific set of genes (group). The protein prediction values were normalized by molecular weight before CV calculation. CV values of selected operons based on predicted protein abundance from various experimental conditions are listed. More details in Section 3.3: Materials and Methods. $\overline{PCV}$ is the mean of CV values computed through permutation test for selected operons. $^\dagger$ CV values which value is greater than the $\overline{PCV}$.

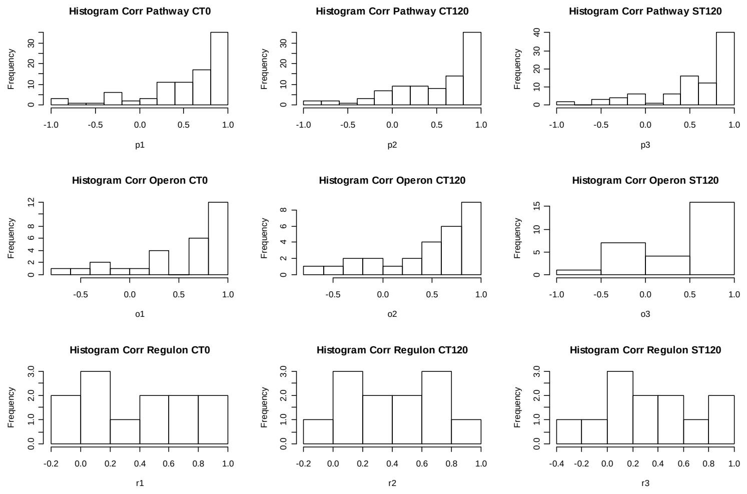


Figure 8. Histogram of correlations per biological group for Dataset 2. Overall Pearson correlation between mRNA expression and normalized protein abundance was $0.33$ to $0.48$ for Dataset 2.

1; and $2463$, $2465$, and $2463$ for CT0, CT120, and ST120 for Dataset 2 [Zhang et al., 2006b, Zhang et al., 2006a, Mukhopadhyay et al., 2006]. With only partial proteomic data, the power of integrative transcriptomic and proteomic analysis could be limited and the analyses could be biased. There exists therefore an urgent need to develop methodologies to accurately estimate missing proteomic data to provide deeper insight into metabolic mechanisms underlying complex biological systems. Estimating missing proteomic data is not a trivial task [Nie et al., 2006a, Nie et al., 2007]. One of the major difficulties is that the correlation patterns between transcriptomic and proteomic data do not follow a linear

TABLE 4

Percentages of groups with small CV value and percentile score.

| Groups | Dataset 1 | | Dataset 2 | |
|---|---|---|---|---|
| | % Groups with | % Groups with | % Groups with | % Groups with |
| | $CV < \overline{PCV}$ | Percentile Score $< 0.2$ | $CV < \overline{PCV}$ | Percentile Score $< 0.2$ |
| Operon | $75\%$ - $79\%$ | $36\%$ - $50\%$ | $75\%$- $82\%$ | $32\%$ -$54\%$ |
| Pathway | $79\%$ - $88\%$ | $45\%$ - $48\%$ | $74\%$ - $76\%$ | $37\%$ - $43\%$ |
| Regulon | $50\%$ - $67\%$ | $8\%$ - $33\%$ | $58\%$ - $83\%$ | $25\%$ - $33\%$ |

$CV$ is Coefficient of Variation. More details in Section 3.3: Materials and Methods. $\overline{PCV}$ is the mean of CV values computed through permutation test.

relationship at the whole genome scale. Recently varying correlations between different functional groups of genes/proteins [Beck-Jr. and Knecht, 2003, Beyer et al., 2004, Nie et al., 2006b], and varying strength of the correlation between different sampling times and growth conditions have been reported [Conrads et al., 2005]. However, to our knowledge, no statistical method of capturing non-linearity of correlation has been published.

In this work, we employed stochastic gradient boosting trees as a nonlinear model to understand a possible pair-wise constant relationship among transcriptomic data, proteomic data, and other factors. The boosting tree procedure is one of the favorable predictive data mining tools for many reasons. From the regression trees characteristics they inherit the positive features of robustness.

Boosted trees models are invariant under all monotone transformations of the individuals input variables which eliminates the sensitivity to long-tailed distributions and outliers [Friedman, 2001]. Moreover, implicit feature selection is intrinsic through the trees' construction and inherited by the boosting machinery. In contrast to a single tree, boosted tree models enhance stability by reducing the depth of the trees and averaging over many of them. Gradient boosting trees models may not produce exact description but they provide insights into the nature of the input-output relationship.

The GBT model constructed is a data-driven model where the input are the abundance measurements of all mRNA ($\sim 3500$) and qualified detected proteins ($<\sim 800$), and output are the predicted abundance levels for almost all proteins ($\sim 3500$) in the genome. This approach provides two major advantages over previous correlation methods. First, it allows undetected proteins (those with an assigned protein abundance value of zero) to be assigned a predicted abundance based on the mRNA levels. As output, the model provides predicted abundance levels for a large number of proteins which are undetected experimentally; and second, the model attempts to address the possible non-linearity property of the correlations between transcriptomic and proteomic data. Based on the coefficient of determination ($R_2$) which is used to assess the cross validated models, $R_2$ ranged from $0.393$ to $0.582$ in both datasets in this nonlinear model, which provided slightly better results compared to results when multiple linear regression model is applied ($R_2$) ranges from $0.27$ to $0.33$). Finally, we evaluated the validity of this model using bioinformatics approaches. For example, in a comparison of the predicted protein abundance patterns of genes belonging to the same operons (representing groups of proteins that are expected to have similar molar abundance values), the results demonstrated that the coefficients of variation of estimated protein abundance values within operons are indeed smaller than that for random groups of proteins.

## 3.6. Acknowledgements

I would like to thank the co-authors of this research work, Drs. Roger H. Johnson, Weiwen Zhang, George C. Runger, and Deirdre R. Meldrum. The authors would like to thank the Center for Biosignatures Discovery Automation at the Biodesign Institute at the Arizona State University. Funding: The work is supported by the National Human Genome Research Institute (NHGRI) Center of Excellence in Genomic Sciences (Grant No. 5 P50 HG002360) of the National Institutes of Health (NIH). Conflict of Interest: none declared.

# CHAPTER 4

# INTEGRATIVE ANALYSIS OF TRANSCRIPTOMIC AND PROTEOMIC DATA OF SHEWANELLA ONEIDENSIS: MISSING VALUE IMPUTATION USING TEMPORAL DATASETS

## 4.1. Abstract

Despite significant improvements in recent years, proteomic datasets currently available still suffer large number of missing values. Integrative analyses based upon incomplete proteomic and transcriptomic datasets could seriously bias the biological interpretation. In this study, we applied a non-linear data-driven stochastic gradient boosted trees (GBT) model to impute missing proteomic values using a temporal transcriptomic and proteomic dataset of *Shewanella oneidensis*. In this dataset, genes' expression was measured after the cells were exposed to $1$ mM potassium chromate for $5-$, $30-$, $60-$, and $90-$min, while protein abundance was measured for $45$- and $90$-min. With the ultimate objective to impute protein values for experimentally undetected samples at $45$- and $90$- min, we applied a serial set of algorithms to capture relationships between temporal gene and protein expression. This work follow four main steps: (1) quality control step for gene expression reliability, (2) mRNA imputation, (3) protein prediction, and (4) validation. Initially, $S$ control chart ap-

proach is performed on gene expression replicates to remove unwanted variability. Then, we focused on the missing measurements of gene expression through a nonlinear Smoothing Splines Curve Fitting. This method identifies temporal relationships among transcriptomic data at different time points and enables imputation of mRNA abundance at 45-min. After mRNA imputation was validated by biological constrains (i.e., operons), we used a data-driven GBT model to impute protein abundance for the proteins experimentally undetected in the 45- and 90-min samples, based on relevant predictors such as temporal mRNA gene expression data and cellular functional roles. The imputed protein values were validated using biological constraints such as operon and pathway information through a permutation test to investigate whether dispersion measures are indeed smaller for known biological groups than for any set of random genes. Finally, we demonstrated that such missing value imputation improved characterization of the temporal response of *Shewanella oneidensis* to chromate.

## 4.2. Introduction

Significant efforts to improve analytical technologies pertaining to mRNA, protein, and metabolite measurement have been made in the past decade. These efforts have led to the generation of several new "omics" research fields such as transcriptomics, proteomics, and metabolomics [Fiehn, 2001, Lin and Qian, 2007, Ishii and Tomita, 2009]. To date, a large amount of information regarding cellular metabolism has been acquired through application of these approaches [Park et al., 2005]. However, due to technical limitations of these high throughput technologies and constraints on experimental design, most of these datasets still suffer from missing values. Incomplete data availability has impeded scientists from assembling comprehensive information regarding cellular metabolism. To address this issue, attempts to apply computational tools to impute missing values in various high throughput "omics" datasets have been made [Little and Rubin, 1987, Wood et al., 2004, Polpitiya et al., 2008, Aittokallio, 2010, Albrecht et al., 2010]. The most successful examples of such efforts were for DNA microarray gene expression data [Aittokallio, 2010]. In these studies, implementation of computational tools such as k-nearest neighbors, least squares, local least squares, iterative expectation-maximization, and bayesian principal component analysis (BPCA) methods were applied to impute missing gene expression values [Troyanskaya et al., 2001, Purohit et al., 2004, Kim et al., 2005]. Alternatively, external information in the form of biological constraints [Gan et al., 2006] such as Gene Ontology (GO) annotation [Tuikkala et al., 2006] or additional microarray datasets [Hu et al., 2006] can also be used to improve the accuracy of these traditional methods [Tuikkala et al., 2008, Torres-García et al., 2009]. In general, for static gene expression datasets, imputation can be made by evaluation of specific gene patterns across samples, and can be evaluated through statistical metrics such as root mean squared error (RMSE), biological knowledge, and further experimental data [Aittokallio, 2010]. For temporal gene expression datasets, where missing values can occurred through the time span, curve fitting methods such as simple interpolation and spline interpolation have been used to estimate missing values [D′haeseleer et al., 1999, Aach and Church, 2001]. Compared with gene expression datasets, proteomic datasets suffers significantly more from missing values due to the imperfect identification of coding sequences within a genome and the limited sensitivity of current peptide detection technologies [Wilkins et al., 2006, Nie et al., 2007, Albrecht et al., 2010]. Some current proteomic technologies often allow detection of less than half of the putative proteins from a microbial genome [Nie et al., 2007], leaving a significant number of proteins experimentally undetected. Several computational methods have been adapted from the estimation of missing values in gene expression data to overcome this problem and estimate the missing values by using the available measurements from other proteins, such as the $K$-nearest neighbor method being applied to Difference Gel Electrophoresis (DGE) data [Jung et al., 2005]. Another method integrated GO information into the data imputation; this approach could enhance the imputation even when the missing fraction is large [Tuikkala et al., 2006]. In one recent study, the BPCA method was used for imputing missing values in a gel-based proteomics dataset [Pedreschi et al., 2008]. Other integrative methods have focused on improving protein identification [Ramakrishnan et al., 2009b, Ramakrishnan et al., 2009a].

Based on the assumption that there indeed exists meaningful correlation between the two types of datasets [Tartaglia and Vendruscolo, 2009, Nie et al., 2007, d. S. Abreu et al., 2009, Maier et al., 2009], we have been developing statistical tools to deal with missing values in proteomics datasets by integrating cognate transcriptomic and proteomic data. For example, Nie et al. proposed a Zero-inflated Poisson regression model to understand the correlation between transcriptomic and proteomic datasets and then predict the missing proteomic data. These models correctly predicted increased expression of Ech hydrogenase and decreased expression of Coo hydrogenase for *Desulfovibrio vulgaris* grown on formate. To further address the issue that correlations may be different in different groups of genes and that correlations may not follow a uniform pattern at the whole genome scale [Nie et al., 2007], Torres-García et al.

(2009) recently published a stochastic Gradient Boosted Trees (GBT) [Friedman, 2002] approach to uncover possible nonlinear relationships between transcriptomic and proteomic data, and to predict protein abundance for proteins not experimentally detected based on predictors such as mRNA abundance, cellular role, molecular weight, sequence length, protein length, GC content and triplet codon counts [Nie et al., 2006c, Brown et al., 2006]. The coefficient of determination $(R^2)$ was used to assess model performance and an improvement in this coefficient was observed for the nonlinear model when compared with linear regression. However, so far, missing value imputations by these models have only been performed using cognate transcriptomic and proteomic datasets.

In this study, we extended our research to temporal transcriptomic and proteomic datasets for imputation of missing proteomic values. The dataset we used was from a stress response of *Shewanella oneidensis* to potassium chromate [Brown et al., 2006]. In compiling these datasets, gene expression was measured after the cells were exposed to $1$ mM potassium chromate for $5-$, $30-$, $60-$, and $90-$min, while the protein abundance was measured after the cells were exposed to the same treatment for $45-$ and $90-$min. The goal is to uncover the relationship between temporal gene expression data and protein abundance and then use it to impute missing proteomic values for samples at $45-$

(which does not have cognate transcriptomic data) and $90-$min. Initially non-linear Smoothing Splines Curve Fitting was used to identify the temporal relationships among transcriptomic data from different time points and then impute the missing gene expression measurement for the $45-$ min sample which will be used later to predict protein abundance. After the imputation was validated by biological constraints (i.e., operons), we used a GBT model to uncover possible non-linear relationships between temporal transcriptomic and proteomic data, and to impute protein abundance for the proteins not experimentally detected in the $45-$ and $90-$min samples, based on several features (temporal mRNA gene expression data, cellular role, molecular weight, sequence length, protein length, GC content, treatment conditions, and channel dye). We validated the imputed protein values using biological constraints such as operon and pathway information through a permutation analysis which included a one-sided paired $t$-test to test whether dispersion measures are indeed smaller for known biological groups than for any set of random genes. Finally, we demonstrated that missing value imputation improved characterization of the temporal response of *Shewanella oneidensis* to chromate.

### 4.3. Materials and methods

#### 4.3.1. Dataset

##### 4.3.1.1. Description of mRNA and proteomic measurements

We analyzed a dataset from *Shewanella oneidensis* (SO) published previously [Brown et al., 2006]. A brief description of the datasets is given in Table 5. This *Shewanella oneidensis* dataset consists of whole-genome temporal transcriptomic and proteomic information gathered to understand the response to acute chromate (Cr) stress when compared to results in Luri-Bertani (LB) medium. The microarray samples were collected during the mid-exponential growth phase of cells with and without potassium chromate exposure at $5-$, $30-$, $60-$, and $90-$min. These microarray measures were validated using qRT-PCR. The data contained $9032$ probes $(2$ per gene) corresponding to 4516 genes for the study. This number does not include plasmids genes. The raw mRNA probe intensity values were normalized using quantile normalization with the R package available at http://www.r-project.org/. In parallel, differential proteomics were measured by HPLC-MS/MS at time intervals of $45-$ and $90-$min after the cells were treated with potassium chromate [Brown et al., 2006]. As shown in Table $5$, $4516$ unique SO genes were hybridized (after removing arti-

TABLE 5
Description of the dataset used in this study.

| Characteristics | Dataset |
|---|---|
| Organism | *Shewanella oneidensis* |
| Conditions | LB (Control) and Cr (Chromate Stress) |
| Number of Variables* | 12 |
| Number of replicates (mRNA abundance) | 6/gene (2 dyes ea.) |
| Number of genes analyzed | 4516 for both conditions |
| Quality of mRNA replicates | 0.85-0.98 |
| Number of replicates (protein abundance) | 2/gene |
| Number of genes analyzed | 2447 for both conditions |
| Quality of protein replicates | 0.91-0.99 |

ficial hybridized genes), only 2447 genes were detected by proteomic analysis. The number of proteins used in the models is reduced based on whether or not that protein has mRNA expression values along the time span.

#### 4.3.1.2. Quality of dataset

The quality of this dataset was assessed by calculating the Pearson correlation coefficients among multiple replicates of transcriptomic and proteomic measurements. There were three mRNA replicates per dye (Cy3 and Cy5). The experiment was repeated twice by changing the dye combination among control (LB medium) and stress response (Cr). In every chip, two probes of the same gene were measured per chip or replicate. Hence, as shown in Table 5, six mRNA replicates were quantified for each gene for which the Pearson correlation coefficients ranged from $0.85$ to $0.98$ among the $4516$ genes. The protein values showed a high Pearson correlation of $0.91 - 0.99$ between the two replicates. Based on this high correlation at linear scale and small data available per gene, we used the average of the two replicates of protein abundance for further analysis.

#### 4.3.1.3. $S$-quality control charts

Despite the high correlation coefficients among replicates for mRNA values, we further reduce the inherent variability of these mRNA measurements before using them to unravel nonlinear relationships with cognate proteomic data. Removing genes with "noisy" measurements of mRNA abundance may provide a clearer global view of the relationship under study. A quality control (QC) step was performed to accomplish this goal using $S$ - quality control charts for robustness. The $S$-chart is an approach to control process variability such as that among mRNA replicates. $S$-charts are usually used for subgroups with variable and small samples size ($n < 10$) [Montgomery, 2001] which is the case in our study. This type of control chart is commonly used in pairs: Xbar and $S$ control charts. The use of $S$ chart alone seemed more appropriate because it assumes that the variability among mRNA replicates per gene should be constant across all genes. Equation 4.1 shows how to calculate the $S$-chart control limits for each gene. In these equations $\bar{\sigma}$ was estimated as the mean of all standard deviation values which were computed for each gene. The constant $c_4$ depends on the sample size, $n$, and was calculated as shown in Equation 4.2.

$$\begin{aligned} UCL_{gene_i} &= \bar{s} + 3\frac{\bar{s}}{c_4}\sqrt{1 - c_4^2} \\ CL_{gene_i} &= \bar{s} \\ LCL_{gene_i} &= \bar{s} - 3\frac{\bar{s}}{c_4}\sqrt{1 - c_4^2} \end{aligned} \tag{4.1}$$

While the sample standard deviation ($s$) is a biased estimator of $\sigma$, $s$ is an unbiased estimator of $c_4\sigma$ when the normality assumption prevails. However, the estimation is robust under different conditions. Hence, we used this method to calculate empirical thresholds to remove "noisy" genes. Removal of these genes improved the nonlinear model performance obtaining higher $R^2$ values as discussed in detail in the Conclusion of this chapter.

$$c_4 = \sqrt{\frac{2}{n-1}}\frac{(\frac{n}{2}-1)!}{(\frac{n-1}{2}-1)!} \tag{4.2}$$

### 4.3.2. Predictors and genome information

Predictors are variables used in the model to explain the relationship between these with protein abundance measurements. The predictors in study are: mRNA(at $5-$, $30-$, $60-$, and $90-$min), cellular functional categories (Mainrole), GC content, molecular weight (MW), sequence length, protein length, treatment, and channel dye as shown in Table 5. The temporal mRNA values are shown as raw fluorescence intensity units. The cellular functional categories of all genes in the *Shewanella oneidensis* genome were downloaded from the Comprehensive Microbial Resource (CMR) of TIGR (http://cmr.tigr.org) [Heidelberg et al., 2004]. On the basis of the original annotation, the genes and proteins were classified into $20$ cellular functional categories. Gene annotated attributes such as gene and protein lengths, molecular weight, and GC content were downloaded from TIGR. These annotated attributes and cellular functional categories were included as possible predictors of protein abundance. Annotation for operon groups in *Shewanella oneidensis MR-1* was downloaded from the list published at Microbesonline [Dehal et al., 2010,Alm et al., 2005]. The list of metabolic pathway groups of interest was downloaded from the KEGG database (http://www.genome.jp/kegg/pathway.html). Operon and pathway information were used in this study for model validation.

### 4.3.3. Smoothing splines cubic curve fitting

A spline is a piecewise polynomial function as shown in Equation 5.7 where $t$ is the time parameter, $S_i(t)$ are polynomials and $c_i$ are coefficients of the piecewise function. For this particular application $y(t)$ represents mRNA abundance at time $t$. The use of piecewise low-degree polynomials results in smooth curves and avoids the problems of overfitting, numerical instability and oscillations that arise when single, high-degree polynomials are used [Bar-Joseph, 2004]. These types of curves have been widely used in several fields such as image processing [Hou and Andrews, 1978]. For our study, we invoked one of the simplest splines: the cubic smoothing spline for piecewise third-order polynomials to predict mRNA abundance at time $45$-min.

$$y(t) = \sum_{i=1}^{n} c_i S_i(t) \text{ where } t_{min} \leq t < t_{max} \tag{4.3}$$

### 4.3.4. Construction of non-linear relationship model

After handling normalization and missing values for the transcriptomic measurement values, a further study was conducted aiming to understand the global relationship between mRNA expression and protein abundance through a nonlinear method. GBT satisfies the need to handle mixed data types and is capable of finding nonlinear relationships between the data in our study. The method implemented shows several other advantages over common nonlinear methods such as robustness, invariability to monotone transformations, implicit feature selection, interpretability and its wide range of useful applications [De'ath, 2007, Elith et al., 2008, Friedman, 2001].

The main objective of this model is to map a set of input variables $x = x_1, ..., x_n$ to the response output $y$ by minimizing error (See Equation 4.4). Boosting fits a weighted additive expansion composed of weak classifiers (e.g., regression trees) that approximates the response $y$ as in Equation 4.5 [Hastie et al., 2001]. Gradient boosting sequentially applies regression trees to fit residuals while minimizing squared error loss, creating new models which are encouraged to become experts in cases misclassified by previous trees.

$$\hat{y} = \arg\min_{y} E_{y,X} L(y, \hat{y}) \tag{4.4}$$

$$\hat{y} = \sum_{m=0}^{M} \beta_m T(X; \hat{\Theta}) \tag{4.5}$$

For this particular study $y$ represents protein abundance and the input space $x$ is composed by twelve predictor variables. The method used in this study uses a gradient approach implemented from the gbm R package available at the R project (http://www.r-project.org/). The required inputs chosen were (*parameter values in parentheses*): loss function (*Squared Error Loss*), number of trees (*500*), the depth of each tree (*5*), shrinkage rate (*0.005*), and number of folds for cross validation (*5*) [Ridgeway, 2007].

### 4.3.5. Model validation

The model protein predictions are assessed through two main techniques: coefficient of determination and a permutation test. Both assessment tools allow us to gather information on how well the non-linear model predicted protein values using the data and external biological information (e.g., operons). Cross validation was used to assess model performance by reevaluating the model with randomized input data partitions [Hastie et al., 2001].A cross-validated model is assessed in terms of its coefficient of determination ($R^2$) which is the variation explained by the model. This coefficient is a statistical measure interpreted as proportion of data variability accounted ranging from zero to one. The closer the $R^2$ value is to one the better is the model in explaining the variance of the data. Furthermore, an implementation of a permutation test using external biological information (e.g., operons groups) was performed to alternatively assess the goodness-of-fit of the non-linear model. It was studied using predictions of small sets of genes/peptides grouped by pathway or operon information with the assumption that genes/peptides within a known biological group should have similar predictions. Coefficient of variation $(CV)$ was computed as the ratio of the standard deviation and the mean of the "molar abundance" for a set of proteins [Johnson, 2005, Nie et al., 2006a] and is independent of the sample size. These coefficients of variation are computed for all pathway and operon groups and compared to a distribution of permuted coefficient of variation ($PCV$). These $PCVs$ compute the $CV$ values of predicted "molar abundance" from a thousand sets of genes/peptides selected at random. These random groups are of the same size as the operon or pathway group that they will be compared with. Hence, every known biological group has a $CV$ value and an associated empirical distributions of $CV$ values computed from random sets. This list of permuted-$CVs$ is called $PCVs$. The aim is to test whether or not $CV$ values from known groups are smaller than $PCV$ values (permuted groups). This is an assumption used to validate prediction values. To test this assumption paired $t$-tests are performed as shown in Equation 4.6 and Equation 4.7. Variance for the test statistic is estimated separately for both groups and the Welch modification to the degrees of freedom is used. We have tested $CV$ against several statistics from the $PCV$ empirical distribution such as mean and lower tail percentile values $(0\%, 5\%, 10\%, 20\%, ..., 100\%)$ since we are testing whether or not $CVs$ are smaller than $PCVs$.

$$D = CV_{group[i]} - PVC_{mean||x^{th}percentile} \tag{4.6}$$

$$Hypotheses = \begin{cases} H0 : \mu = 0 \\ \\ H1 : \mu < 0 \end{cases} \tag{4.7}$$

$P$-values from these paired $t$-tests are obtained where smaller values than alpha (i.e., $0.05$) will provide information to sustain less-dispersed assumption for validation.

## 4.4. Results and discussion

### 4.4.1. Quality control

Even though high correlation was observed between mRNA replicates, a quality control (QC) step was performed aiming to remove unwanted variability that could inhibit the discovery of global nonlinear patterns between transcriptomic and proteomic data of *Shewanella oneidensis*. This step was performed primarily to reduce inherent variability of mRNA abundance within gene replicates. The criterion used to remove genes was based on their standard deviation. For example genes whose mRNA replicates' deviation exceeded a statistical threshold were removed. The genes removed at this step are excluded for further downstream modeling and validation analysis. This quality control step was performed for every condition, resulting in a range of $432 - 886$ genes removed from further studies based on this criterion. As discussed in Section 4.3.1.3, limits are created for every gene based on the number of replicates of mRNA values. Most genes had six replicates of mRNA abundance values per dye but certain genes had less if they were missing or undetected. In addition, genes with only one replicate were removed from further study. This QC step served to reduce the variability of the inputs to our model in order to better capture the nonlinear relationship between transcriptomic abundance values and proteomic peptide count.

### 4.4.2. Temporal curve fitting for mRNA prediction

Cubic smoothing splines enabled prediction of mRNA abundance at $45-$min based on temporal information of mRNA at times $5-$, $30-$, $60-$, and $90-$min. A similar method was previously reviewed by Bar-Joseph (2004) . At this stage, temporal information at the transcriptomic level is integrated through the smoothing splines to predict mRNA abundance as a continuous function of time. The method fits a nonlinear curve, a piecewise polynomial of order three, to every single gene across the temporal dimension allowing for prediction at any time point within the trained range ($5-$ to $90-$min). This allowed us to synthesize an mRNA value at $45-$min which was not experimentally obtained but was of interest to study its correlation with protein values at $45-$min. The cubic splines smoothing predictions were compared to predictions done with simple linear regression throughout the trained range ($5-$ to $90-$min), and to linear interpolation from mRNA values between $30-$ and $60-$min.

These predictions were validated using external biological information (see Section 4.3.5 for more details). The information used contained gene organization information such as operons, and gene function information such as pathways. Predictions for genes within the same operon or pathway group are assumed to be more similar than those of a random set of genes [Bild and Febbo, 2005, Charaniya et al., 2007]. Genes/peptides failing the initial QC cutoff were not used to generate the random sets. For validation, dispersion measure was calculated as the coefficient of variation ($CV$) within time and conditions for every operon and some pathway groups as shown in Table 6. This table presents corresponding $CV$ values and permuted coefficient of variation ($PCV$) per operon group. The $CV$ values calculated from the mRNA predicted values at $45-$min were, in general, smaller when compared with random set of genes in several cases for both pathway and operon data.

About $70\%$ of the $130$ operon groups had smaller dispersion measure ($CV$) than the average $CV$ for $1000$ random set of genes as presented in Table 7. We chose four main metabolic pathways for validation: Carbohydrate, Energy, Lipids, and Amino Acid Metabolism. Pathway and operon groups showed smaller dispersion values when compared to random sets of genes in most cases. For example, about $75\%$ of the lipids metabolism pathways were less dispersed than the permuted set of genes. Similarly, $50\%$ of energy, $50\%$ of carbohydrate, and $69.23\%$ of amino acid metabolism pathways had smaller $CVs$ than the $CVs$ from random sets of genes. However, pathway results were more variable than operon results; yielding two groups with no conclusive findings in terms of dispersion comparison (energy and carbohydrate metabolism). Table 7 also shows how many groups fall with a percentile score less than $20\%$. Percentile score indicates the position of the current $CV$ when compared to the distribution from $1000$ $CVs$ of permuted sets of genes [Torres-García et al., 2009]. Hence, the percentage values shown in Table 7 reflect that only a small percentage of the operon groups truly fall in the lower tail of the $PCV$ distribution. These re-

TABLE 6

Model validation: mRNA abundance predictions at $45-$min for some operons.

| Operon | | mRNA at 45min | | | |
|---|---|---|---|---|---|
| | | LB | | Cr | |
| | | $CV$ | $\overline{PCV}$ | $CV$ | $\overline{PCV}$ |
| 1 | SO0003-SO0009 | 0.924 | 0.573$^{\dagger}$ | 0.928 | 0.574$^{\dagger}$ |
| 2 | SO0023-SO0032 | 0.458 | 0.579 | 0.453 | 0.579 |
| 3 | SO0066-SO0074 | 0.47 | 0.579 | 0.464 | 0.579 |
| 4 | SO0101-SO0109 | 0.78 | 0.579$^{\dagger}$ | 0.785 | 0.579$^{\dagger}$ |
| 5 | SO0163-SO0181 | 0.536 | 0.619 | 0.526 | 0.611 |
| 6 | SO0182-SO0189 | 0.711 | 0.573$^{\dagger}$ | 0.699 | 0.574$^{\dagger}$ |
| 7 | SO0218-SO0229 | 0.605 | 0.608 | 0.604 | 0.605 |
| 8 | SO0230-SO0257 | 0.476 | 0.629 | 0.454 | 0.621 |
| 9 | SO0258-SO0269 | 0.435 | 0.579 | 0.418 | 0.579 |
| 10 | SO0272-SO0285 | 0.481 | 0.593 | 0.47 | 0.584 |
| 11 | SO0286-SO0300 | 0.535 | 0.597 | 0.532 | 0.595 |
| 12 | SO0311-SO0318 | 0.787 | 0.556$^{\dagger}$ | 0.772 | 0.544$^{\dagger}$ |
| 13 | SO0342-SO0346 | 0.388 | 0.556 | 0.39 | 0.544 |
| 14 | SO0395-SO0401 | 0.29 | 0.556 | 0.288 | 0.544 |
| 15 | SO0441-SO0456 | 0.822 | 0.579$^{\dagger}$ | 0.809 | 0.579$^{\dagger}$ |
| 16 | SO0476-SO0488 | 0.349 | 0.608 | 0.347 | 0.605 |
| 17 | SO0532-SO0536 | 0.573 | 0.556$^{\dagger}$ | 0.538 | 0.544 |
| 18 | SO0599-SO0606 | 0.742 | 0.594$^{\dagger}$ | 0.728 | 0.583$^{\dagger}$ |
| 19 | SO0608-SO0612 | 0.364 | 0.556 | 0.37 | 0.544 |
| 20 | SO0639-SO0652 | 0.53 | 0.579 | 0.52 | 0.579 |
| 21 | SO0656-SO0672 | 0.445 | 0.593 | 0.432 | 0.584 |
| 22 | SO0674-SO0690 | 0.839 | 0.597$^{\dagger}$ | 0.856 | 0.595$^{\dagger}$ |
| 23 | SO0712-SO0718 | 0.628 | 0.503$^{\dagger}$ | 0.573 | 0.512 |
| 24 | SO0842-SO0849 | 0.929 | 0.556$^{\dagger}$ | 0.92 | 0.544$^{\dagger}$ |
| 25 | SO0850-SO0854 | 0.445 | 0.556 | 0.444 | 0.544 |
| 26 | SO0877-SO0883 | 0.563 | 0.556$^{\dagger}$ | 0.559 | 0.544$^{\dagger}$ |
| 27 | SO0900-SO0909 | 0.3 | 0.573 | 0.296 | 0.574 |
| 28 | SO1008-SO1021 | 0.538 | 0.606 | 0.536 | 0.595 |
| 29 | SO1103-SO1110 | 0.703 | 0.594$^{\dagger}$ | 0.719 | 0.583$^{\dagger}$ |
| 30 | SO1155-SO1163 | 0.644 | 0.573$^{\dagger}$ | 0.623 | 0.574$^{\dagger}$ |

Coefficient of Variation ($CV$) is computed by dividing standard deviation by the mean of the prediction values for protein abundance for a specific set of genes (group). The protein prediction values were normalized by molecular weight before $CV$ calculation. $CV$ values of selected operons based on predicted protein abundance from various experimental conditions are listed. More details in Section 4.3: Materials and Methods. $\overline{PCV}$ is the mean of $CV$ values computed through permutation test for selected operons. $^{\dagger}$ CV values which value is greater than the $\overline{PCV}$.

TABLE 7

Validation mRNA predictions: Percentages of groups with small CV value and percentile score.

| Groups | mRNA at 45min | |
|---|---|---|
| | % Groups with<br>$CV < \overline{PCV}^{\dagger}$ | % Groups with<br>Percentile Score < 0.2 |
| Operon | 68.70% - 69.47% | 19.85% - 23.66% |
| Pathway | 50%-75% | 0%-75% |

Coefficient of Variation $(CV)$ is computed by dividing standard deviation by the mean of the prediction values for protein abundance for a specific set of genes (group). The protein prediction values were normalized by molecular weight before CV calculation. More details in Section 4.3: Materials and Methods.
$^{\dagger}$ $\overline{PCV}$ is the mean of CV values computed through permutation test.

sults for pathway varied more across different pathway groups such as Energy Metabolism where $75\%$ of these groups were positioned at the lower tail of the $CV$ distribution in contrast to Lipids Metabolism where all its groups had a percentile score greater than $20\%$.

### 4.4.3. Non-linear ensemble method for prediction of undetected proteins

The GBT model was built to predict protein behavior at a specific time point and to make comparisons across time periods using eleven out of the twelve variables available to assess its importance (Channel dye was removed as a predictor since it was found to be insignificant). Cellular functions, GC content and

TABLE 8
Measurements of relative importance of variables for the ten top ranked variables (after removing genes with high mRNA).

| Protein $45-$min | | Protein $90-$min | |
|---|---|---|---|
| Variables | VRI (%) | Variables | VRI (%) |
| mRNA.5min | 21.267 | mRNA.5min | 21.911 |
| Mainrole | 18.349 | Mainrole | 17.005 |
| mRNA.60min | 14.095 | mRNA.60min | 14.822 |
| mRNA.45min | 12.098 | mRNA.90min | 13.081 |
| mRNA.90min | 12.065 | mRNA.45min | 12.213 |
| GC | 7.329 | MW | 7.48 |
| MW | 6.568 | GC | 5.404 |
| mRNA.30min | 4.12 | mRNA.30min | 3.795 |
| SequenceLength | 2.863 | SequenceLength | 2.68 |
| ProteinLength | 1.225 | ProteinLength | 1.61 |
| Treatment | 0.02 | Treatment | 0 |

Results show two models for two time- protein response where VRI specifies variable relative importance. Relative influence is computed as the average of empirical improvements in squared error of splitting the decisions trees with corresponding variable. This measure is divided by the sum of the empirical improvement of all variables in the model. The percentage of this measure is VRI. More details discussed in Section 4.4.

mRNA intensities were ranked highest by the GBT, suggesting that they play important roles in predicting protein abundance for both cases ($t = 45$ min and $t = 90$ min), while factors related to protein size, such as MW, protein length or sequence length are playing less significant roles (See Table 8). Prediction values for both time points are shown in Figure 9.

Once predictions were made for mRNA abundance at $45-$min, the GBT

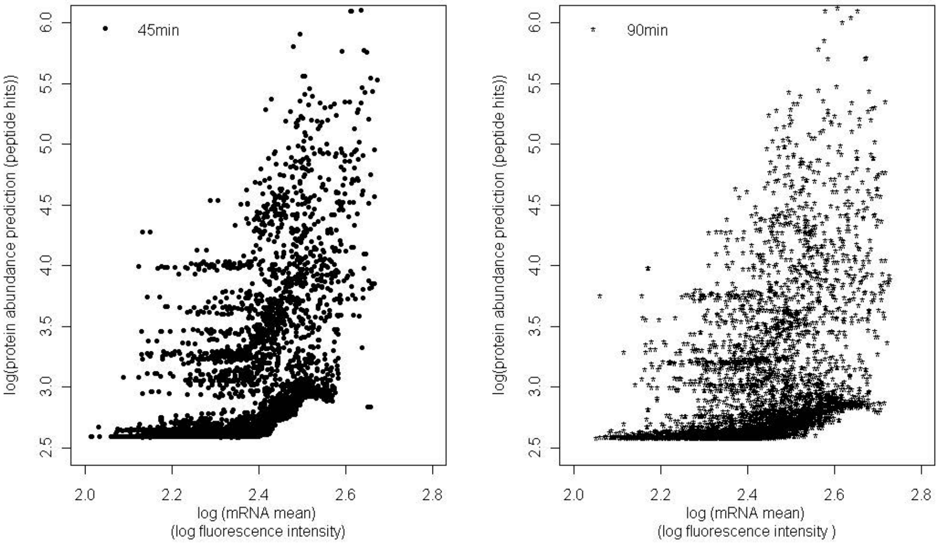


Figure 9. log-log Plot: Protein abundance predictions at $45-$min (left) and $90-$min (right).

*mRNA mean at $45-$min is based on its prediction value from Cubic Smoothing Splines.**mRNA mean at $90-$min is calculated through the average of mRNA abundance values among all replicates per gene at that time. This log-log plot shows logarithmic values of protein prediction with its corresponding mRNA measurements at logarithmic scale.

model was applied for both time periods ($45-$ and $90-$min) in an effort to elucidate a possible nonlinear relationship between transcriptome and protein abundance. Similarities and correlations between these two were studied with respect to their temporal component. The GBT model provided the implicit variable importance scores shown in Table 8 for both time points. The model revealed that the most relevant variable for prediction of protein abundance at both $45-$

and $90-$min is the amount of mRNA at $5-$min, followed by Mainrole and the mRNA values at later times. Similar results were observed in our previous study where mRNA and cellular categories were found to be most relevant [Torres-García et al., 2009]. The relative importance measure of an input variable was calculated as its ability to improve the squared error loss at each tree along the ensemble (GBT), and finally computing an average on the improvements. Our findings show high similarity of the power ranking of predictors for protein values at $45-$ and $90-$min, suggesting that these two time frames may not behave differently from a global perspective.

Figure 10 shows interesting behaviors manifested in the partial dependency plots ($PDPs$) [Hastie et al., 2001], from the GBT protein predictions at $45-$min and $90-$min. PDPs aim to show the contribution of a specific variable to the prediction in the model. In this figure we see increasing, "plateau", and decreasing patterns for mRNA at different time points. Also, we observed differences in the prediction contribution among different cellular categories. Most of these categories ranged in mid-low protein prediction values and few of these categories had higher values (i.e., Protein Synthesis and Transcription). It can also be observed that the model found high variability in protein abundance prediction for lowly expressed genes across times that could be inherent to the technology when detecting lowly expressed genes (See “plateau” effect in Figure 10). A more complex GBT model may be required to capture all this variability in the range of low protein abundance and mRNA expression levels. If we increase the depth of trees in the ensemble model, the trees will divide into smaller groups and provide more detailed predictions. On the other hand, increasing the depth of trees will increase the complexity of the model and could result in overfitting of the data.

External biological knowledge (i.e., operon groups) was employed to assess the validity of the GBT model’s predictions. By doing this, we assume that the molar concentration of gene products from the same operon will be equal or similar, while the molar concentration of gene products from different operon or unrelated genes will carry much larger variation [Bild and Febbo, 2005, Charaniya et al., 2007]. The process is similar to the validation performed for mRNA prediction values at $45-$min in the early section. The results showed that the $CV$ values computed for protein predictions were indeed smaller for operon groups than those $CVs$ calculated from random set of genes ($PCVs$) as shown in Table 9. From $130$ operon groups evaluated, about $80\%$ (ranging from $79.84\%$-$83.06\%$ for all conditions) had $CV$ values smaller than the average of $PCVs$ ($CVs$ from $1000$ random set of genes) for both $45-$ and $90-$min protein predictions. Pathway groups proved to be more dispersed when compared to operon groups. The percentage of pathway groups with smaller dispersion measure than the random set of genes ranged from $25\%$ to $100\%$. About $50\%$ of the carbohydrate metabolism pathways were less dispersed than the permuted set of genes. Similarly, $100\%$ of energy, $25\%$ of lipids, and $60\%$ of amino acid metabolism pathways had smaller $CVs$ than $PCVs$.

To further study the dispersion comparison between $CV$ values from known operon groups against $PCV$ (from permuted sets of genes), a one-sided paired $t$-test was performed on these two metrics as shown Equation 4.6 and Equation 4.7. $P$-value statistics for these paired tests showed $CV$ to be significantly smaller than the average of $PCVs$ and also smaller than $PCV$ percentile values (at $40^{th}$ percentile or greater) for all conditions. These results are shown in Figure 11 for LB condition at $45-$ and $90-$min. The dashes lines in Figure 11 represent the percentile level where $CV$ was found to be statistically smaller than $PCV$ through a one-sided paired $t$-test with corresponding $p$-values displayed as ( P: 0 ). Contrary to the solid lines where $p$-values were found insignificant for all conditions with $p$-values ranging between $0.42 - 0.68$ at the $30^{th}$ percentile. Hence, these results reveal that the smallest percentile with statistical difference is found between the $30^{th}$ and $40^{th}$ position which is in the lower side of the

TABLE 9

Model validation: Protein abundance predictions at $45$ and $90$ minutes in some operons groups.

| Operon | | Protein $45-$ min | | | | Protein $90-$ min | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LB | | Cr | | LB | | Cr | |
| | | $CV$ | $\overline{PCV}$ | $CV$ | $\overline{PCV}$ | $CV$ | $\overline{PCV}$ | $CV$ | $\overline{PCV}$ |
| 1 | SO0003-0009 | 0.76 | 0.699$^\dagger$ | 0.91 | 0.705$^\dagger$ | 0.82 | 0.675$^\dagger$ | 0.88 | 0.674$^\dagger$ |
| 2 | SO0023-0032 | 0.62 | 0.807 | 0.56 | 0.772 | 0.63 | 0.786 | 0.56 | 0.738 |
| 3 | SO0066-0074 | 0.42 | 0.951 | 0.43 | 0.907 | 0.4 | 0.889 | 0.41 | 0.883 |
| 4 | SO0163-0181 | 0.62 | 0.951 | 0.64 | 0.907 | 0.56 | 0.889 | 0.56 | 0.883 |
| 5 | SO0182-0189 | 0.42 | 0.878 | 0.42 | 0.833 | 0.43 | 0.826 | 0.43 | 0.792 |
| 6 | SO0218-0229 | 0.54 | 1.188 | 0.55 | 1.101 | 0.52 | 1.132 | 0.52 | 1.059 |
| 7 | SO0230-0257 | 0.4 | 1.37 | 0.43 | 1.308 | 0.42 | 1.315 | 0.47 | 1.205 |
| 8 | SO0258-0269 | 0.45 | 1.092 | 0.46 | 1.022 | 0.47 | 1.06 | 0.5 | 0.976 |
| 9 | SO0272-0285 | 0.37 | 1.022 | 0.38 | 0.983 | 0.4 | 0.99 | 0.4 | 0.937 |
| 10 | SO0286-0300 | 0.7 | 1.092 | 0.64 | 1.022 | 0.73 | 1.06 | 0.67 | 0.976 |
| 11 | SO0311-0318 | 0.79 | 0.807 | 0.79 | 0.772 | 0.78 | 0.786 | 0.78 | 0.738$^\dagger$ |
| 12 | SO0342-0346 | 0.28 | 0.699 | 0.27 | 0.705 | 0.22 | 0.675 | 0.21 | 0.674 |
| 13 | SO0395-0401 | 0.77 | 0.807 | 0.51 | 0.772 | 0.77 | 0.786 | 0.57 | 0.738 |
| 14 | SO0441-0456 | 0.59 | 1.092 | 0.59 | 1.022 | 0.54 | 1.06 | 0.54 | 0.976 |
| 15 | SO0476-0488 | 0.65 | 0.577$^\dagger$ | 0.65 | 0.556$^\dagger$ | 0.65 | 0.565$^\dagger$ | 0.65 | 0.551$^\dagger$ |
| 16 | SO0532-0536 | 0.86 | 0.951 | 0.83 | 0.907 | 0.8 | 0.889 | 0.76 | 0.883 |
| 17 | SO0599-0606 | 0.37 | 0.699 | 0.38 | 0.705 | 0.41 | 0.675 | 0.42 | 0.674 |
| 18 | SO0608-0612 | 0.23 | 0.878 | 0.23 | 0.833 | 0.22 | 0.826 | 0.22 | 0.792 |
| 19 | SO0639-0652 | 0.35 | 0.577 | 0.35 | 0.556 | 0.35 | 0.565 | 0.35 | 0.551 |
| 20 | SO0656-0672 | 0.86 | 0.878 | 0.86 | 0.833$^\dagger$ | 0.86 | 0.826$^\dagger$ | 0.86 | 0.792$^\dagger$ |
| 21 | SO0674-0690 | 0.62 | 0.577$^\dagger$ | 0.62 | 0.556$^\dagger$ | 0.61 | 0.565$^\dagger$ | 0.61 | 0.551$^\dagger$ |
| 22 | SO0842-0849 | 0.23 | 0.807 | 0.21 | 0.772 | 0.22 | 0.786 | 0.21 | 0.738 |
| 23 | SO0850-0854 | 0.41 | 0.577 | 0.45 | 0.556 | 0.44 | 0.565 | 0.45 | 0.551 |
| 24 | SO0877-0883 | 0.47 | 0.699 | 0.47 | 0.705 | 0.39 | 0.675 | 0.39 | 0.674 |
| 25 | SO0900-0909 | 0.56 | 1.01 | 0.58 | 0.92 | 0.56 | 0.969 | 0.57 | 0.922 |
| 26 | SO1008-1021 | 0.45 | 1.092 | 0.46 | 1.022 | 0.42 | 1.06 | 0.44 | 0.976 |
| 27 | SO1103-1110 | 1.16 | 0.807$^\dagger$ | 1.16 | 0.772$^\dagger$ | 0.87 | 0.786$^\dagger$ | 0.87 | 0.738$^\dagger$ |
| 28 | SO1155-1163 | 0.33 | 0.577 | 0.33 | 0.556 | 0.33 | 0.565 | 0.33 | 0.551 |

Coefficient of Variation ($CV$) is computed by dividing standard deviation by the mean of the prediction values for protein abundance for a specific set of genes (group). The protein prediction values were normalized by molecular weight before CV calculation. CV values of selected operons based on predicted protein abundance from various experimental conditions are listed. More details in Section 4.3: Materials and Methods. $\overline{PCV}$ is the mean of CV values computed through permutation test for selected operons.

$^\dagger$ CV values which value is greater than the $\overline{PCV}$.

$PCV$ empirical distribution. This finding validates our assumptions of smaller dispersion for predictions within known operon groups even though the test does not provide "perfect" results (statistical significance at the $0^{th}$ percentile). Figure 12 presents several histograms with corresponding percentile scores for the $CV$ values for each operon group. Percentile scores represents the position of the $CV$ among the $PCV$ distribution providing an idea of how small or large the $CV$ is compared to the $PCV$ values. These values were also found to be small with some exceptions revalidating indeed a less dispersed prediction of protein values for operon groups.

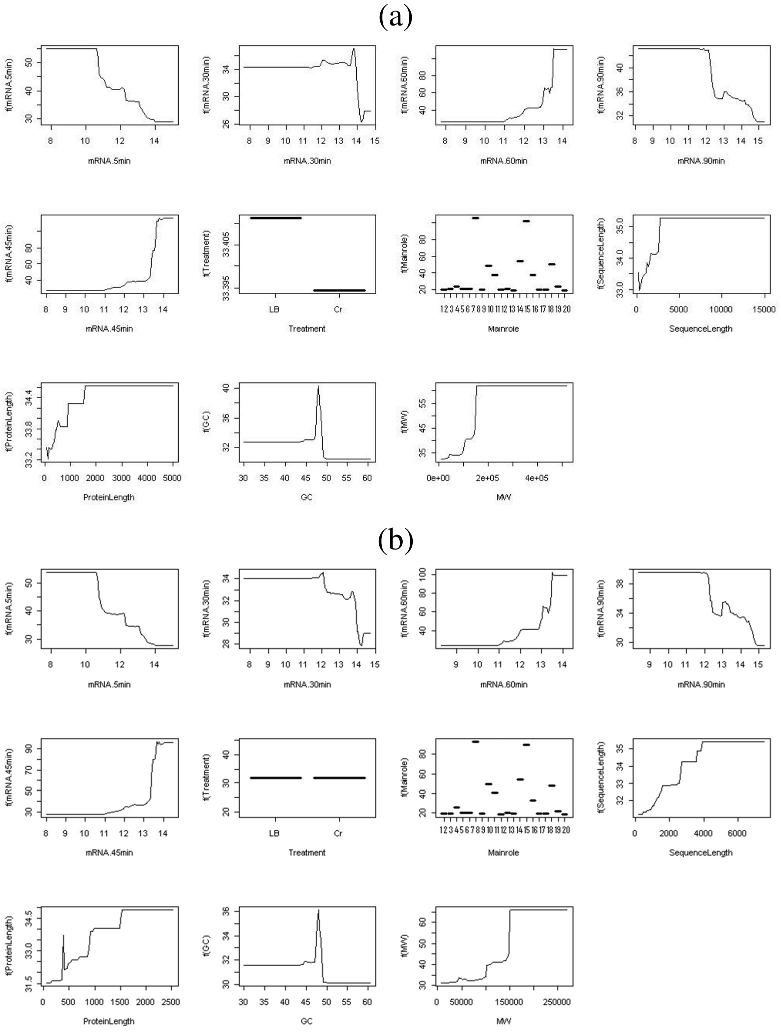


Figure 10. Partial dependency plots for protein prediction at $45$-min and $90$-min among predictors for *Shewanella oneidensis* respectively.

Every graph depicts the partial dependency prediction values for each of the $11$ predictors used in the nonlinear model when predicting protein abundance at $45-$min (a) and $90-$min (b). * $y$ axes in this figure shows protein prediction values contributed by a particular variable value using partial dependency plots in linear scale. ** $x$ axes values varies depending on the value range of each particular variable (e.g., temporal mRNA presented in linear scale, treatment=LB, Cr, mainrole=numeric index to the list of cellular functional roles as shown in Figure 13). More details about the variables presented can be found in Section 4.3.

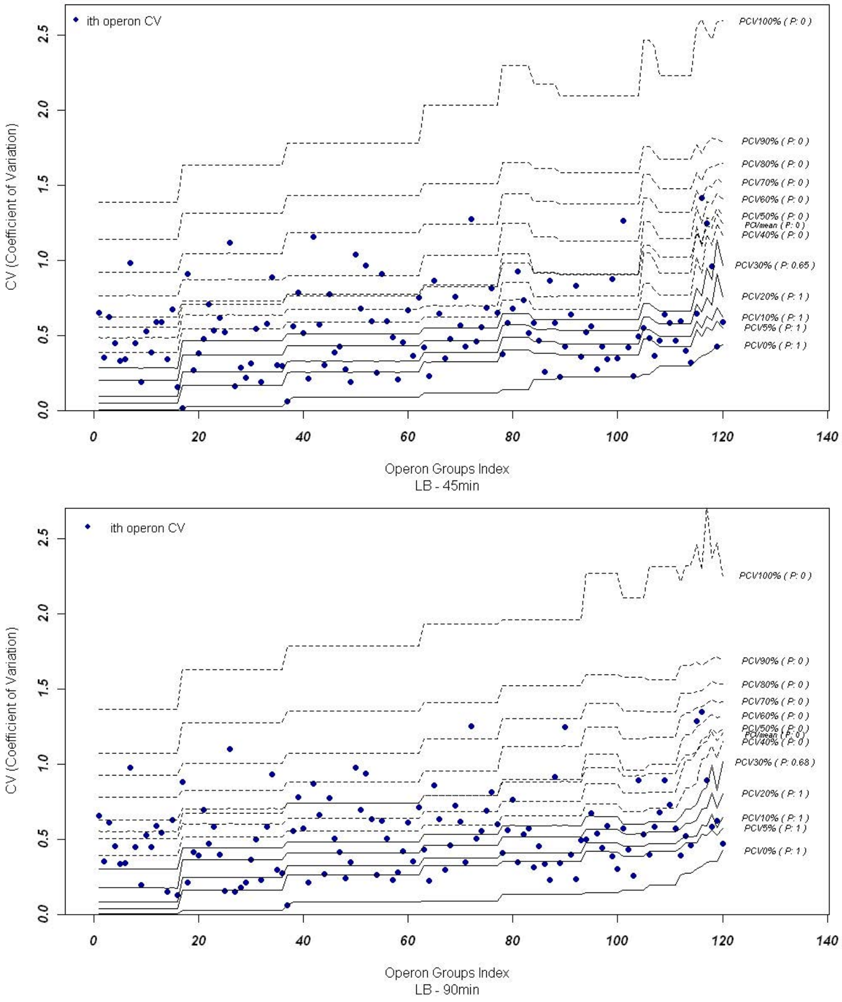


Figure 11. Permutation test for model validation.

This graph shows $CV$ values for each operon group (portrayed as dots) plotted together with $PCV$ values. These $PCV$ values are calculated from permuted set of genes at the $0^{th}$, $5^{th}$, $10^{th}$, $20^{th}$, , $100^{th}$ percentile and the mean for LB at $45-$ and $90-$min. While $CVs$ (dots) are computed for each operon group based on protein predictions of genes within the operon. Similar results were obtained for Cr at $45-$ and $90-$min. Percentiles are shown in increments of $10$ (except $5^{th}$ percentile). These are labeled as $PCVx\%$ representing the $x^{th}$ percentile value of the permuted $CV$ values. The value P next to the curves labels shows $p$-values for one-sided paired $t$-tests. The dashed lines correspond to percentiles with small $p$-values ($CVs$ are statistically smaller than $PCVs$ at the $x^{th}$ percentile). The solid lines show the percentile positions where $CVs$ was no significantly smaller than $PCVs$. See more details in Section 4.3.

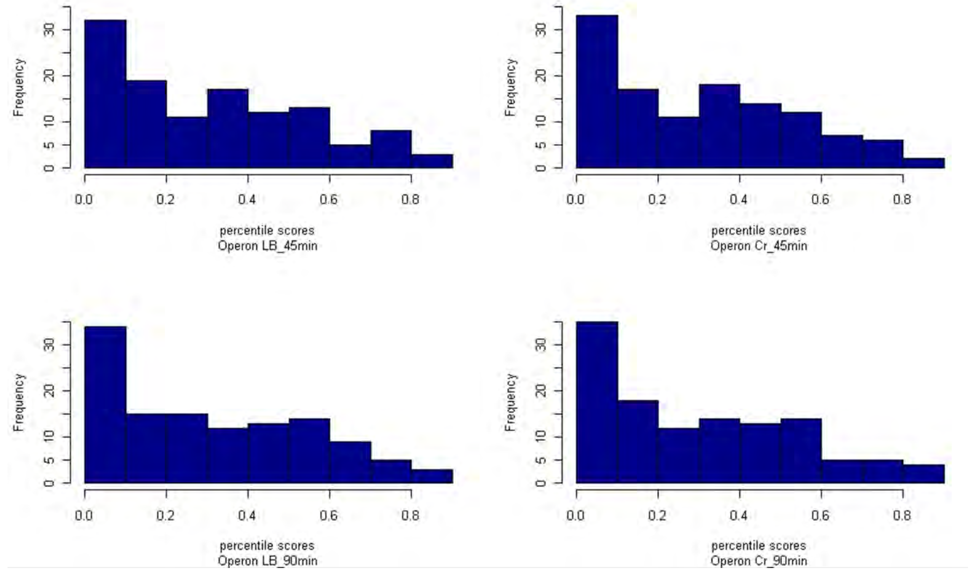


Figure 12. Model validation: Percentile scores for CV values of operon groups. These plots present frequencies for the percentile scores computed through the position of CV from known operon groups within PCV values from random set of genes for both conditions (LB and Cr) and both times ($45-$ and $90-$min).

### 4.4.4. Biological interpretation improved by using imputed data

A total of $1159$ and $1174$ proteins were experimentally identified for control and Cr(VI)-exposed samples, respectively, at $45-$min, and a total of $1054$ and $1185$ proteins for control and Cr(VI)-exposed samples, respectively, at $90-$min. This is slightly over $20\%$ of the whole proteome in the *Shewanella oneidensis* genome. Through modeling and imputation by integrating temporal transcriptomic data and proteomic data, we were able to assign abundance values to approximately $3200$ proteins. Using the imputed dataset, we first analyzed the experimentally undetected proteins with the highest predicted values. This group of proteins included ribosomal proteins, tRNA (guanine-N1)-methyltransferase, glutaminase A, FAD-dependent glycerol-3-phosphate dehydrogenase, peptidase glycyl-tRNA synthetase, prolidase, preprotein translocases, heme exporter protein CcmC and siderophore biosynthesis protein. Most of these proteins, essential to cellular metabolism, are likely excluded from detection due to technical reasons [Albrecht et al., 2010]. Protein abundance predictions for a number of undetected proteins experimentally were listed in Table 10, Table 11 and Table 12 for $45-$ and $90-$min, respectively. Full list of all protein-abundance predictions for all proteins can be provided in electronic form. The results from Tables 10, Table 10 and Table 12 showed that many experimentally undetected proteins are within important cellular functions, such as energy metabolism and regulatory or transporting roles. By imputing them through a non-linear model integrating transcriptomic and partial proteomic data and all important sequence features, we could obtain a better picture of cell metabolism.

Notable improvements were seen when comparing the functional distribution of the detected and imputed proteins for hypothetical proteins, regulatory

TABLE 10

Protein abundance prediction for selected undetected proteins for 45−min.

| Cellular Functional Categories | | | |
|---|---|---|---|
| GeneID | †Pred.-LB | †Pred.-Cr | Gene Description |
| Amino acid biosysthesis | | | |
| SO2071 | 13.32 | 13.32 | Imidazoleglycerol-phosphate dehydratase/histidinol-phosphatase (hisB) |
| Biosynthesis of cofactors, prosthetic groups, and carriers | | | |
| SO2921 | 15.951 | 15.764 | D-erythro-7,8-dihydroneopterin triphosphate epimerase (folX) |
| Cell envelope | | | |
| SO4688 | 13.974 | 13.974 | Glycosyl transferase, group 2 family protein |
| SO0067 | 46.899 | 46.002 | Penicillin-binding protein 1C, putative |
| SO2394 | 15.759 | 15.725 | Penicillin-binding protein 4 (dacB) |
| SO0853 | 15.066 | 15.066 | Pilin, putative |
| Cellular processes | | | |
| SO3279 | 14.751 | 14.751 | AcrB/AcrD/AcrF family protein |
| SO0421 | 14.707 | 14.215 | AmpD protein (ampD) |
| SO4405 | 13.988 | 13.988 | Catalase/peroxidase HPI (katG-2) |
| SO2320 | 14.196 | 14.196 | Chemotaxis protein CheA, interruption-N |
| SO3252 | 14.628 | 14.25 | Chemotaxis protein CheV (cheV-3) |
| SO0549 | 14.215 | 14.215 | Chemotaxis protein CheY/response regulator receiver domain protein |
| SO4299 | 14.059 | 14.059 | Chloramphenicol acetyltransferase (cat) |
| SO4697 | 18.522 | 17.772 | Glutathione S-transferase (gst) |
| SO4095 | 24.046 | 23.948 | Maf protein (maf) |
| SO0976 | 14.059 | 14.059 | Organic hydroperoxide resistance protein (ohr) |
| SO3432 | 33.96 | 32.147 | RNA polymerase sigma-38 factor (rpoS) |
| Central intermediary metabolism | | | |
| SO4697 | 17.092 | 16.453 | Glutathione S-transferase (gst) |
| DNA metabolism | | | |
| SO4241 | 13.328 | 13.328 | ATP-dependent DNA helicase RecQ (recQ) |
| SO0016 | 13.32 | 13.32 | DNA-3-methyladenine glycosidase I (tag) |
| SO3126 | 13.32 | 13.32 | Methylated-DNA–protein-cysteine methyltransferase (ogt) |
| SO2220 | 13.429 | 13.383 | MutT/nudix family protein |
| SO3004 | 13.32 | 13.32 | Prophage LambdaSo, DNA modification methyltransferase, putative |
| SO0952 | 13.328 | 13.328 | Single-stranded-DNA-specific exonuclease RecJ (recJ) |
| Energy metabolism | | | |
| SO2726 | 26.556 | 26.306 | Cytochrome b, putative |
| SO0978 | 80.69 | 78.805 | FAD-dependent glycerol-3-phosphate dehydrogenase, family protein |
| SO3057 | 37.776 | 37.526 | Pal/histidase family protein |
| Fatty acid and phospholipid metabolism | | | |
| SO4380 | 49.726 | 47.49 | 3-oxoacyl-(acyl-carrier-protein) synthase II, putative |

† Peptide hits predictions for control (LB) and treatment (Cr) conditions.

TABLE 11

Protein abundance prediction for selected undetected proteins for $45-$min: Cont.

| Cellular Functional Categories | | | |
|---|---|---|---|
| GeneID | †Pred.-LB | †Pred.-Cr | Gene Description |
| Protein fate and synthesis | | | |
| SO0261 | 46.132 | 44.991 | Heme exporter protein CcmC (ccmC) |
| SO3083 | 99.82 | 92.124 | Peptidase, M16 family |
| SO1334 | 50.083 | 51.294 | Prolipoprotein diacylglyceryl transferase (lgt) |
| SO4618 | 25.29 | 25.29 | Prolyl oligopeptidase family protein |
| SO2860 | 25.526 | 25.231 | Thiol:disulfide interchange protein, DsbA family |
| SO0164 | 55.36 | 54.549 | Heat shock protein 15 (hslR) |
| SO3962 | 102.26 | 104.537 | Ribosomal subunit interface protein (yfiA-3) |
| SO1620 | 55.256 | 55.256 | RNA pseudouridylate synthase family protein |
| SO1788 | 56.454 | 56.049 | tRNA-(MS[2]IO[6]A)-hydroxylase (miaE) |
| Purines, pyrimidines, nucleosides, and nucleotides | | | |
| SO1980 | 23.822 | 23.753 | Phosphoribosyl transferase domain protein |
| Regulatory functions | | | |
| SO1255 | 15.449 | 15.095 | Cyclic nucleotide phosphodiesterase, putative |
| SO2193 | 13.32 | 13.32 | DNA-binding response regulator |
| SO3901 | 14.759 | 14.833 | lacZ expression regulator (icc) |
| SO0544 | 13.328 | 13.328 | Sensory box histidine kinase |
| SO0502 | 13.32 | 13.32 | Transcriptional regulator, ArsR family |
| SO4567 | 13.32 | 13.32 | Transcriptional regulator, AsnC family |
| SO0864 | 13.32 | 13.32 | Transcriptional regulator, LuxR family |
| SO0295 | 13.32 | 13.32 | Transcriptional regulator, LysR family |
| SO1259 | 13.418 | 13.418 | Transcriptional regulator, LysR family |
| SO2202 | 13.32 | 13.32 | Transcriptional regulator, LysR family |
| SO2193 | 13.32 | 13.32 | DNA-binding response regulator |
| SO0570 | 13.722 | 13.462 | Response regulator |
| SO2366 | 13.328 | 13.328 | Response regulator |
| SO2822 | 13.328 | 13.328 | Sensor histidine kinase |
| SO3306 | 13.84 | 13.873 | Sensor histidine kinase |
| SO0352 | 13.32 | 13.32 | Sensor histidine kinase, putative |
| SO0544 | 13.328 | 13.328 | Sensory box histidine kinase |
| SO3432 | 16.178 | 15.668 | RNA polymerase sigma-38 factor (rpoS) |
| Transport and binding proteins | | | |
| SO3279 | 14.599 | 14.599 | AcrB/AcrD/AcrF family protein |
| SO1042 | 13.952 | 13.952 | Amino acid ABC transporter, ATP-binding protein |
| SO0261 | 18.065 | 18.065 | Heme exporter protein CcmC (ccmC) |
| SO1033 | 13.952 | 13.952 | Iron-compound ABC transporter, ATP-binding protein, putative |
| SO1522 | 17.891 | 18.346 | L-lactate permease, putative |
| SO2886 | 14.998 | 16.905 | Na+/H+ antiporter (nhaB) |
| SO4029 | 13.704 | 13.704 | Transporter, putative |
| SO1236 | 13.639 | 13.639 | Xanthine/uracil permease family protein |

† Peptide hits predictions for control (LB) and treatment (Cr) conditions.

TABLE 12

Protein abundance prediction for selected undetected proteins for 90−min.

| Cellular Functional Categories | | | |
|---|---|---|---|
| GeneID | [†]Pred.-LB | [†]Pred.-Cr | Gene Description |
| SO0287 | 18.886 | 17.085 | 3-dehydroquinate synthase (aroB) |
| SO2074 | 13.939 | 13.724 | ATP phosphoribosyltransferase (hisG) |
| SO1361 | 13.629 | 13.629 | Phospho-2-dehydro-3-deoxyheptonate aldolase, tyr-sensitive (aroF) |
| SO2069 | 13.472 | 13.472 | Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (hisA) |
| SO3413 | 14.174 | 14.174 | Threonine synthase (thrC) |
| Biosynthesis of cofactors, prosthetic groups, and carriers | | | |
| SO1039 | 13.388 | 13.388 | Cob(I)alamin adenosyltransferase (cobO) |
| SO1037 | 13.31 | 13.31 | Cobinamide kinase/cobinamide phosphate guanylyltransferase (cobU) |
| SO4450 | 18.032 | 13.367 | Molybdenum cofactor biosynthesis protein D (moaD) |
| Cellular processes | | | |
| SO4394 | 14.87 | 14.323 | Phage shock protein E (pspE-2) |
| SO4149 | 43.938 | 43.938 | RTX toxin, putative |
| SO3435 | 14.139 | 14.212 | Stationary-phase survival protein SurE (surE) |
| DNA metabolism | | | |
| SO2612 | 15.968 | 15.708 | DNA polymerase III, delta prime subunit (holB) |
| SO2430 | 13.407 | 13.407 | Holliday junction DNA helicase RuvA (ruvA) |
| SO2037 | 13.407 | 13.407 | Site-specific recombinase, phage integrase family |
| Energy metabolism | | | |
| SO0747 | 27.102 | 27.102 | Ferredoxin-NADP reductase (fpr) |
| SO4503 | 24.671 | 24.671 | Formate dehydrogenase accessory protein FdhD, putative |
| SO4503 | 24.671 | 24.671 | Formate dehydrogenase accessory protein FdhD, putative |
| SO1496 | 21.554 | 21.554 | Glycogen phosphorylase family protein |
| SO0293 | 54.574 | 46.119 | Phosphoglycolate phosphatase (gph) |
| SO4062 | 21.432 | 21.432 | Polysulfide reductase, subunit A (psrA) |
| Fatty acid and phospholipid metabolism | | | |
| SO0511 | 19.888 | 19.888 | Acetyl-CoA carboxylase, biotin carboxyl carrier protein (accB) |
| SO0572 | 20.784 | 20.784 | Enoyl-CoA hydratase/isomerase family protein |
| SO4372 | 20.138 | 20.138 | Thioester dehydrase family protein |
| Protein fate and synthesis | | | |
| SO0218 | 88.907 | 67.932 | Preprotein translocase, SecE subunit (secE) |
| SO1194 | 24.492 | 24.492 | Protein-export membrane protein SecF (secF-1) |
| SO1359 | 189.868 | 119.882 | tRNA (guanine-N1)-methyltransferase (trmD) |
| Purines, pyrimidines, nucleosides, and nucleotides | | | |
| SO3140 | 19.652 | 19.652 | Thymidine kinase |
| Regulatory functions | | | |
| SO0532 | 13.399 | 13.399 | Arsenical resistence operon repressor (arsR) |
| SO3389 | 14.22 | 14.161 | Sensory box protein |
| SO2244 | 13.407 | 13.407 | Transcriptional regulator, LacI family |
| SO2847 | 13.407 | 13.407 | Transcriptional regulator, LysR family |
| SO4154 | 13.31 | 13.31 | Transcriptional regulator, LysR family |
| SO4326 | 13.463 | 13.407 | Transcriptional regulator, TetR family |
| SO2540 | 14.314 | 13.798 | Response regulator |
| SO2543 | 13.725 | 13.725 | Sensor histidine kinase |
| Transcription | | | |
| SO2560 | 23.733 | 23.677 | Ribonuclease HI (rnhA) |
| Transport and binding proteins | | | |
| SO4712 | 16.164 | 15.686 | ABC transporter, ATP-binding protein, putative |
| SO1959 | 13.667 | 13.667 | ABC transporter, periplasmic substrate-binding protein, putative |
| SO1044 | 13.667 | 13.667 | Amino acid ABC transporter, periplasmic amino acid-binding protein |
| SO3134 | 14.188 | 13.821 | C4-dicarboxylate-binding periplasmic protein (dctP) |
| SO1925 | 20.666 | 20.793 | HlyD family secretion protein |

[†] Peptide hits predictions for control (LB) and treatment (Cr) conditions.

functions, transport and binding proteins and proteins with unknown function or no data associated (Figure 13), although improved coverage was observed for almost all functional categories at both $45-$ and $90-$min. Using a two-sided paired $t$-test to analyze significant changes between the number of proteins detected versus imputed we obtained $p$-values equal to $0.004$, $0.003$, $0.003$, $0.004$ for LB at $45-$min, LB at $90-$min, Cr at $45-$min and Cr at $90-$min respectively. These $p$-values showed that the difference is significant. In general, proteins in these categories are known for their relatively low abundance or short half life, which may account for their absence in the experimental proteomic datasets. In a previous experimental paper, several genes involved in DNA repair, cellular detoxification, and two-component signal transduction systems were found up-regulated in Cr(VI)-exposed cells relative to untreated cells [Brown et al., 2006]. It was proposed that genes/proteins responsible for these functions may be important for resistance to Cr(VI). Using the imputed proteomic dataset, we were able to identify several more up-regulated proteins of the same function category, such as $SO2255$ encoding a transcription-repair coupling factor (*mfd*) and $SO3961$ encoding *rpoN* ($\sigma^5 4$). These and other proteins can serve as putative candidates for further experimental validation. In general, our analysis indicates that biological interpretation may benefit from imputing missing data using computational methods by integrating temporal transcriptomic and proteomic data.

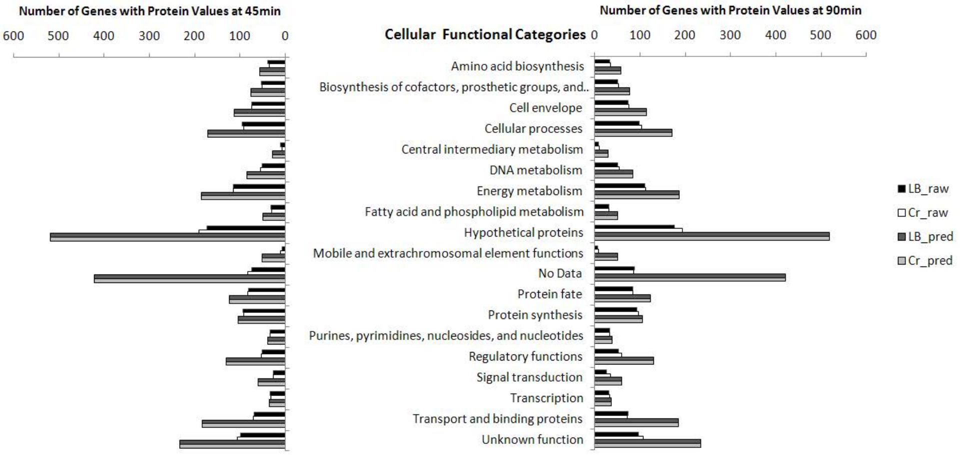


Figure 13. Number of genes for protein values predicted by model, and protein values detected (raw data), for each functional cellular category.

In this graph, changes in the amount of predictions obtained through the model are compared with the amount of data available experimentally. Some relevant differences were found for genes within cellular groups such as hypothetical proteins, regulatory functions, and transport and binding proteins.

## 4.5. Conclusion

Missing values constitute a significant void in current proteomic datasets and need to be handled properly in order to allow accurate biological interpretation. This issue has attracted some attention recently and several computational methods and relevant considerations have been proposed [Albrecht et al., 2010, Nie et al., 2007]. Temporal transcriptomic and proteomic datasets are common and powerful data types in experimental genomics to overcome issues resulting from the snapshot nature of the current technologies by revealing relevant patterns and behaviors not possible from single time point information. It is understood that biological systems are highly dynamic and that temporal information could be extremely important for exploring molecular mechanisms controlling the systems. However, since adding the temporal component to statistical analyses increases their complexity, no current method is available to properly deal with missing proteomic data in temporal datasets.

In this study, we extended the application of the GBT method to deal with missing proteomic data using temporal transcriptomic and proteomic datasets of *Shewanella oneidensis* [Torres-García et al., 2009]. The constructed GBT models assess nonlinear patterns among different types of data (transcriptomic, proteomic, external biological knowledge) providing coefficients of determination ($R^2$) which are used to evaluate the cross-validated models. This performance measure shows improvement with GBT when compared to multiple-linear regression, with $R^2$ values of $0.45$ for the GBT and $0.17$ for the multiple-linear regression model when predicting undetected proteins at $45-$min. The same improvement was seen at $90-$min. In addition, we investigated further the performance of trained ensemble method for different groups of genes combined by different combinations of cellular functional categories yielding $R^2$ values from $0.47$ to $0.68$, showing better performance for particular genes grouped by cellular function. This result validates the importance of cellular function to understand the relationship between transcriptomic and proteomic information. These $R^2$ values found for *Shewanella oneidensis* where comparable with about $0.1$ higher values than the $R^2$ values found in our previous work for *Desulfovibrio vulgaris* ($0.39 - 0.58$) [Torres-García et al., 2009]. This result might suggest a global improvement when temporal data is used but further analysis should be performed since these results are shown for different species. The validity of the trained ensemble model was also evaluated using biological information (i.e., operons), and the results demonstrated that the coefficients of variation of estimated protein abundance within operons are indeed smaller than those for random groups of proteins (e.g., Figure 11).

Although the model was well validated statistically, caution needs to be exercised when interpreting experimental data based on predicted protein expression values because the predicted abundance values are constrained by the quality of experimental proteomic data used as input, and a dearth of large-scale quantitative predictors which can be included in the model. Nevertheless, the initial success in applying the GBT method to temporal transcriptomic and proteomic datasets is encouraging, and the model could serve as a basis for developing more sophisticated models that will allow the inclusion of other relevant large-scale quantitative data types, such as RNA decay measurements, when they become available.

## 4.6. Acknowledgements

I would like to thank the co-authors of this research work, Drs. Steven D. Brown, Roger H. Johnson, Weiwen Zhang, George C. Runger, and Deirdre R. Meldrum. The authors would like to thank the Center for Biosignatures Discovery Automation at the Biodesign Institute at the Arizona State University.

Funding: This work is supported by the National Human Genome Research Institute (NHGRI) Center of Excellence in Genomic Sciences (CEGS) of the National Institutes of Health (NIH) from grant No. $5P50HG002360$. Also, supported by the National Science Foundation (NSF) GK-12 “Down to Earth Science” project from the Division of Graduate Education from grant NSF/DGE $0538701$. The data used in this research study was from a project sponsored by the U.S. Department of Energy Office of Science, Biological and Environmental Research from the Oak Ridge National Laboratory managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract $DE - AC05 - 00OR22725$.

Conflict of Interest: none declared.

# CHAPTER 5

# STATISTICAL FRAMEWORK FOR MULTIPARAMETER ANALYSIS AT SINGLE-CELL LEVEL

## 5.1. Abstract

In an effort to better understand pathogenesis and diseases such as cancer and stroke at different developmental stages, the Microscale Life Science Center is pursuing single cell studies by developing novel technologies to investigate multiparameter physiological metabolic phenotypes. Phenotype measurements with individual cells provide crucial insights into intercellular heterogeneity and allow access to information that is unavailable from bulk cell cultures. Real-time measurements of cellular metabolism such as oxygen consumption (OC) are captured with the objective to understand in situ cellular heterogeneity of cell populations. In this work, we focus on a methodology that aims to address unique challenges encountered with single parameter acquisition particularly from oxygen consumption profiles from two different Barrett's esophageal epithelial cell lines. Despite the clear knowledge that cell populations are heterogeneous, little is known on how prominent the variations of specific factors such as oxygen consumption are in the context of different cell types and/or states. Through the early exploration of OC measurements at the single-cell level, we found the

need to reduce noise or unwanted perturbations of the signals to identify "true" signals and enhance the discovery of relevant disease specific features.

Three main challenges were studied for this heterogeneous dynamic system: (1) high levels of noise, (2) lack of a priori knowledge of single cell dynamics, and (3) the role of cell behavior within and across cell types. Several strategies and solutions are presented for each of these three challenges and applied to every OC curve. To address the random noise problem, three main stages of noise reduction were performed on the signals: smoothing of negative values, low-pass filtering (i.e., Exponentially Weighted Moving Average and Savitzky-Golay), and outlier smoothing.

Once the oxygen consumption data were smoothed, the statistical framework proceeds serially to study multiple states from OC curves. This challenge led to two sub problems which we identified as: (a) removal of redundant systemic information generated by experimental settings and (b) extraction of distinctive features without a priori knowledge of the system. To remove redundant information from a heterogeneous dynamical system, we automatically detected the time point where each curve was approximately zero (oxygen consumption stops) and removed all data following that point (zero-value tails) from further analysis using cumulative sum (CUSUM) control charts. This procedure generated an empirical distribution of the time point at which curves reached an oxygen consumption value near zero, showcasing significant differences within and across different cell types.

Another strategy to model OC curves from single cells and to capture related cell behavior signatures and differences among cell types was implemented through a piecewise linear (spline) model. This was a challenging task since there was no a priori knowledge of cell behavior nor how to effectively model this type of information. In an attempt to address this challenge, we approximated the OC curves through a constrained piecewise linear regression model. This particular spline model fit two linear regressions, each with negatively-restricted slopes with a breakpoint optimally detected through a likelihood method across the entire time span. Furthermore, continuity was enforced in the model. Although such a model may not correspond to the biological function in the OC curves, it provided a good empirical fit to the experimental data with a simple structure. Furthermore, the model facilitated the extraction of relevant features that were used to characterize the OC curves. Such characterization is useful for subsequent studies of conditions and cell types.

Once features such as slopes, intercepts, breakpoint or change-point were extracted for every curve, we proceeded to compare features among different cells within the same type and between the two different types. Some interesting patterns were found in certain features extracted from the spline model. We further explored these comparisons as a classification problem with two classes (e.g., one cell type versus another) finding subtle differences between both cell types. This methodology addressed relevant challenges that arose from this novel data acquisition and experimental modalities at the single-cell level and provided a statistical framework to extract meaningful summaries from experimental data.

## 5.2. Introduction

Present day technologies enable biological experiments to be performed on bulk cells in general on the order of millions. These measurements provide results averaged over the cell population. This modality of research has shown to be fruitful in understanding many disease pathogenesis, progress and its response. Such results do not provide in-depth knowledge of the individual cell behavior or population heterogeneity, which might hold the key to understanding various diseases like cancer and stroke. These kind of diseases are not very well understood at this point and are considered heterogeneous and organotypic. One significant reason can be attributed to inter-cellular heterogeneity where from bulk cells, individual cells from are expressing at different phenotypic levels for different degrees of insults or triggers [Lidstrom and Meldrum, 2003].

A multifaceted approach to understand diseases such as cancer involves genomic, proteomic, morphometric, and phenotypic profiling at single-cell level. Holistically by acknowledging cell-to-cell and cell-to-microenvironment communications, a single cell produces its own energy in the form of ATP. In order to generate ATP, a cell consumes oxygen and glucose, resulting in successive pathway activation and finally culminating in production of adenosine triphosphate (ATP). During this process, other physiologically related parameters such as pH and other ionic contributions in and around the vicinity of the cell tend to change. A multiparametric analysis of the microenvironmental changes surrounding a single-cell has the capability to elucidate insights into inter-cellular heterogeneity by providing biosignatures related to the disease state.

Our framework is primarily focused on single-cell phenotypic correlation with disease initiation and progression. In order to elucidate such heterogeneous cellular information, identify biosignatures, and correlate with diseases, strong efforts are underway to develop novel microfluidic technologies for single-cell level. These technologies include single-cell manipulation [Anis et al., 2010], confinement devices [Zhu et al., 2009], sensitive sensors [Tian et al., 2010], and novel automated systems for data acquisition and analysis [Ashili et al., 2011, Kelbauskas et al., 2011]. The main purpose of such technologies is to enable understanding of various biophysical, biochemical and mechanistic effects at single-cell level and elucidate their correlation with disease initiation and progression.

As a first step in acquiring and analyzing multiparameter data, the Microscale Life Sciences Center (MLSC) has developed oxygen sensor and automated technologies that enable the collection of oxygen consumption profiles at single-cell level which is a direct indication of cellular metabolism. It is known that single-cell oxygen consumption rates are on the scale of fmol/min/cell. The sensor with which the dissolved oxygen concentration is sensed has intrinsic properties causing variable signal to noise ratio over the range of normal concentration to zero dissolved oxygen concentration. Apart from this, other sources of noise include readout noise from the detector, variations from the pulsed excitation source, etc. In a cell line from Barrett's esophagus, the time required for the isolated cell to consume all the oxygen within finite volume of cell media ranges between $30-90$ minutes. As noted previously, the oxygen consumption profiles encounter variable noise levels over the $90$ minute time period. This along with other constant noise sources over typical $90$ minutes puts a need to analyze the data through a more rigorous statistical framework thus by reducing noise and generating more accurate biosignatures.

As more sensors are being developed and integrated into microfludic devices to measure the extra-cellular analytes, the methodologies presented in this paper broadly serves as foundational framework for analyzing sensitive single-cell data thus generating biosignatures correlating with the disease. Three main challenges were studied for this heterogeneous dynamic system: (1) random noise, (2) understanding of multiple states, and (3) comprehension of cell behavior within and across cell types. There are a great number of different methods available to remove white noise in data structures with time component. Many with applications to a variety of fields such as chemistry, environmental issues and medicine [Orfanidis, 1996, Brocker et al., 2002]. Modeling multiple time series from dynamical systems has been explored in the literature through traditional statistical methods such a profile analysis, repeated measures and growth curves all within the same family of problems [Stevens, 1999, Pan and Fang, 2002, Maxwell and Delaney, 2003, Kleinbaum et al., 2008, Weerahandi, 2004]. A main disadvantage of this type of methods is its parametric assumptions on the nature of the time series system. Hence, there exists a need to model these real-time measurements without a priori knowledge in more effective ways to enable its curve characterization and statistical comparisons within and across conditions such as disease state.

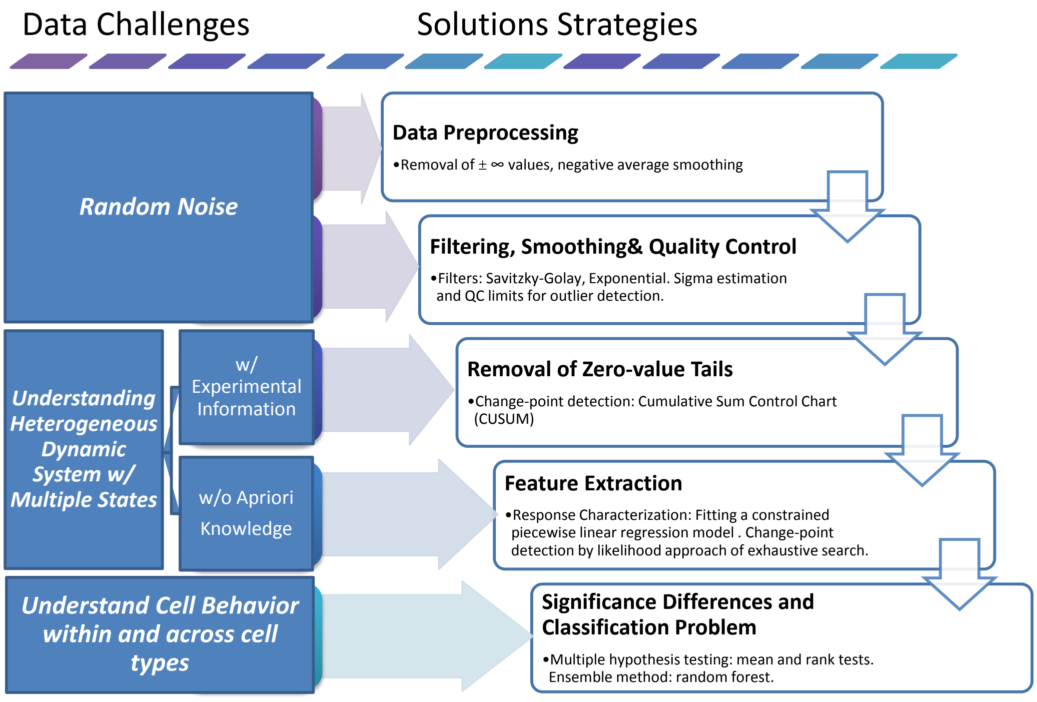

Data Challenges
Solutions Strategies
Random Noise
Understanding Heterogeneous Dynamic System w/ Multiple States
w/ Experimental Information
w/o Apriori Knowledge
Understand Cell Behavior within and across cell types
Data Preprocessing
•Removal of ± ∞ values, negative average smoothing
Filtering, Smoothing& Quality Control
•Filters: Savitzky-Golay, Exponential. Sigma estimation and QC limits for outlier detection.
Removal of Zero-value Tails
•Change-point detection: Cumulative Sum Control Chart (CUSUM)
Feature Extraction
•Response Characterization: Fitting a constrained piecewise linear regression model . Change-point detection by likelihood approach of exhaustive search.
Significance Differences and Classification Problem
•Multiple hypothesis testing: mean and rank tests. Ensemble method: random forest.

Figure 14. Statistical framework diagram.

Sequential steps to process oxygen consumption data into information through smoothing, feature extraction and classification. This figures shows the challenges encountered through the oxygen consumption curves and attempted solution strategies to address each one of them.

## 5.3. Methods

### 5.3.1. Dataset

#### 5.3.1.1. Description of oxygen consumption measurements

We analyzed several sets of data from two Barrett's esophageal epithelial cell lines (CP-A, CP-C). The quality of this dataset was assessed by expert visual inspection. The number of OC curves studied for CP-A and CP-C were $154$ and $256$ respectively. These OC measurements were obtained through a custom developed semi-automated platform [Kelbauskas et al., 2011, Ashili et al., 2011]. The cells were loaded, one per microwell, on a cassette and incubated for approximately $20$ hrs before measurements are gathered. After incubation, microwells with cells are sealed with a lid with micro-optical sensors. These oxygen sensors are excited and the emitted intensity over time is collected [Ashili et al., 2011]. It is important to mention that the effect of incubation time was evaluated on preliminary data showing no major effects. To lessen the future effect of incubation time, the curves in study in this work were kept within the range of $15$ to $30$ hrs of incubation prior data collection.

### 5.3.2. Noise reduction techniques

To reduce random effects from every signal, we smooth these out using three main stages of noise reduction: (1) Smoothing of negative values; (2) Low-pass filtering; and (3) Outlier smoothing.

#### 5.3.2.1. Smoothing of negative values

As a data preprocessing step, every curve is screened for $+\infty$ and $-\infty$ values product of sensor measurements transformation. Those scarce infinite values are removed. This step is followed by a primitive smoothing phase for negative values which consist in replacing those negative values by the average of its adjacent neighbors' values.

#### 5.3.2.2. Low-pass filtering

Two common low-pass filtering techniques were evaluated. A low-pass filter is a filter that allows passage of low frequencies and reduces the amplitude of high frequencies in signals. These two methods and its parameters estimation are described briefly here. In addition we discussed a goodness-of-fit assessment to decide which filtering technique performs better for these OC data curves.

The Savitzky-Golay filter is also called least-squares polynomial smoothing filter and is a finite impulse response (FIR) filter. In 1964 Savitzky and Golay proposed a smoothing technique involving a polynomial fit of fixed degree n to a small window of the data of size $(2m + 1)$ to estimate a midpoint as shown in Equation 5.1 and Equation 5.2. This process is repeated by moving the window of data along the total span [Savitzky and Golay, 1964, Leach et al., 1984]. This type of convoluted filter minimizes the least-squares error and is quite popular in areas such as spectroscopy and analytical chemistry because of its simplicity and speed [Persson and Strang, 2002, Alfassi et al., 2005]. If the data is evenly spaced and continuous then the smoothed value $(y_t^*)$ is the weighted summation of the points in the window frame as described in Equation 5.3. A drawback of Savitzky-Golay early methodology is that $m$ numbers of points are truncated at each side of the data window smoothing only a fraction of it. Extensions to Savitzky-Golay filter addressing initial and endpoint estimation found in the literature were implemented [Leach et al., 1984, Gorry, 1990].

$$y_t^* = \sum_{k=0}^{n} \beta_k t^k = \beta_0 + \beta_1 t + \beta_2 t^2 + ... + \beta_n t^n \qquad (5.1)$$

where $t = [-m, -(m-1), ..., 0, ..., m]$

$$\frac{\partial}{\partial \beta_k}\left[\sum_{t=-m}^{m} (y_t^* - y_t)^2\right] = 0 \qquad (5.2)$$

$$y_t^* = \frac{\sum_{t=-m}^{m} (c_t y_{j+t})}{N} \tag{5.3}$$

A second degree order polynomial fit was tested as it is commonly used in practice [Persson and Strang, 2002]. Another important parameter needed in the SG filtering is the window length ($m$). Common values for this parameter are $m = 11$ and $m = 21$, we evaluated root-mean-squared-error ($RMSE$) for a range of values in both conditions (e.g., CP-A and CP-C) as shown in Figure 15 with expected results. Results shown in this manuscript includes window size of $11$ since smoothing performance was better than $m = 21$ since we lost local pattern which were better described with a smaller $m$.

The second filter in applied is Exponentially Weighted Moving Average known as EWMA which is an infinite impulse response (IIR) filter and is a special case of moving average where its weights decay exponentially given much more importance to recent point than older data points. The smoothed value of $y_t$ is obtained through Equation 5.4 where $\lambda$ represents the decaying rate ranging from $0 \leq \lambda \leq 1$ . A small value of $\lambda$ gives more weight to older data and less to new data and vice versa [Montgomery, 2005, Walczak, 2000]. A parameter $\lambda$ equal to $0.2$ was used during the smoothing of the data curves in this study to detect small shifts. An $RMSE$ evaluation across a range of $\lambda$ values was performed as shown in Figure 15. In practice values between $0.2 - 0.3$ is used for lambda or constant smoothing rate [Hunter, 1996].

$$y_t^* = \lambda y_t + (1 - \lambda) y_{t-1}^* \tag{5.4}$$

To assess the performance of EWMA and SG filtering techniques, we evaluated average root-mean-squared-error ($RMSE$) as a goodness of fit criteria. Goodness-of-fit statistics describe how well the smoothed values are fitting the observations (i.e., coefficient of determination ($R^2$), mean squared error ($MSE$), and root-mean-squared-error ($RMSE$)). The smaller the value for the average $RMSE$ the better in terms of fit. Both techniques showed similar performances for the commonly chosen parameters as displayed in Figure 16.

5.3.2.3. Outlier smoothing

The time series data studied here contains "noisy" peaks in certain areas product of derivative transformation. We have detected these outliers using traditional control charts theory as shown in Equation 5.5.

$$L = \bar{x} \pm k\hat{\sigma} \tag{5.5}$$

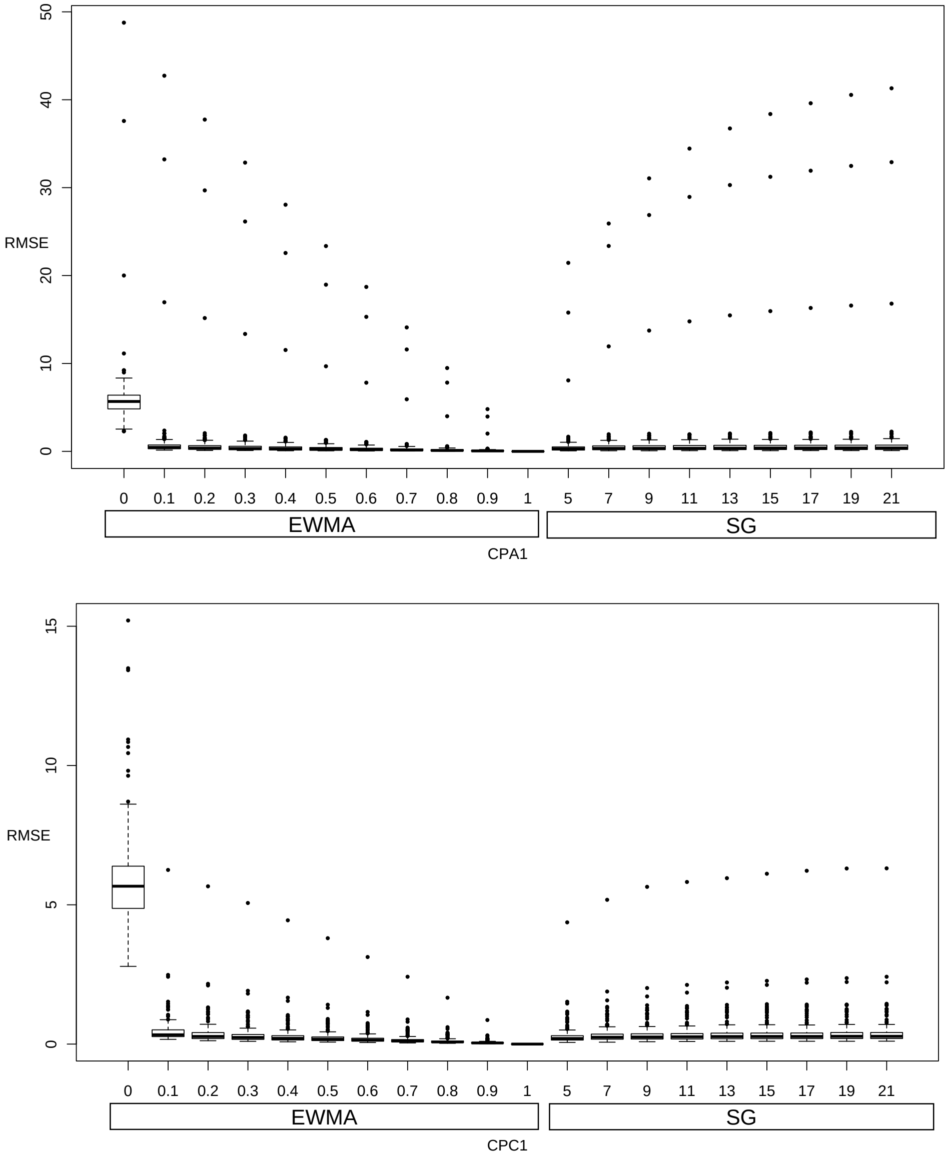


Figure 15. EWMA and SG filters parameter evaluation for CP-A and CP-C.

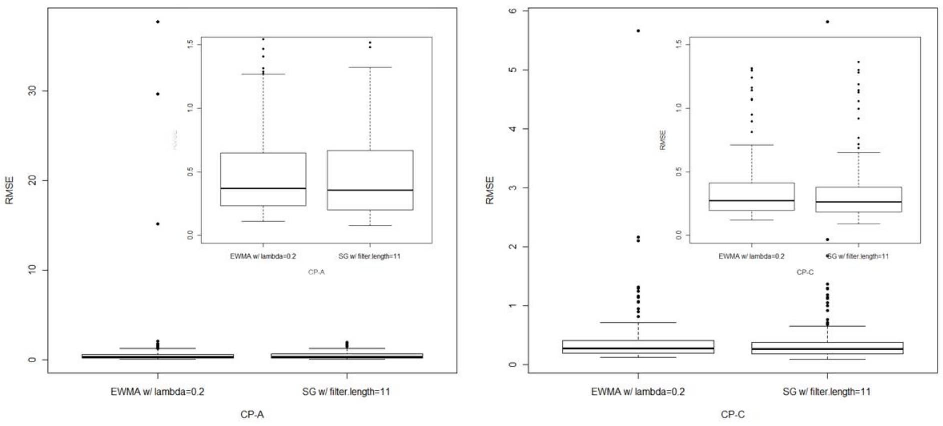


Figure 16. Smoothing filter comparison.

Outliers are detected when data observations in the time series fall away from the control limits calculated using Equation 5.5. The computation of $\hat{\sigma}$ is estimated using the $RMSE$ measure between the observations and the smoothed values from a low-pass filter as described previously. For this particular case, we are assuming $\hat{\sigma}$ to be constant across time which might not be necessarily true. However, this assumption allows for an easy estimation of $\sigma$ helping on the detection of very large outliers. Furthermore, to determine $k$ (control width constant) we studied several options. The tuned value for $k$ was chosen to be equal to 2, this value of $k$ shows an outlier detection of around $10\%$ of the points within the curve, naturally higher values of $k$ showed smaller percentages ranging from $0\%$ to $\sim 5\%$ and smaller values of $k$ such as $1$ provided a high $\sim 25\%$ outlier detection as shown in Figure 17. Hence, $k$ equal to two seemed a reasonable estimation to reduce random noise due to outliers. Once we have detected outliers from the raw signal using the previously discussed control limits, we proceed to smooth those outliers by replacing its value by averaging its close neighbor's values (symmetrical window frame of size $2$). After this update procedure to the data, we reapply the low-pass filter to proceed with further analysis.

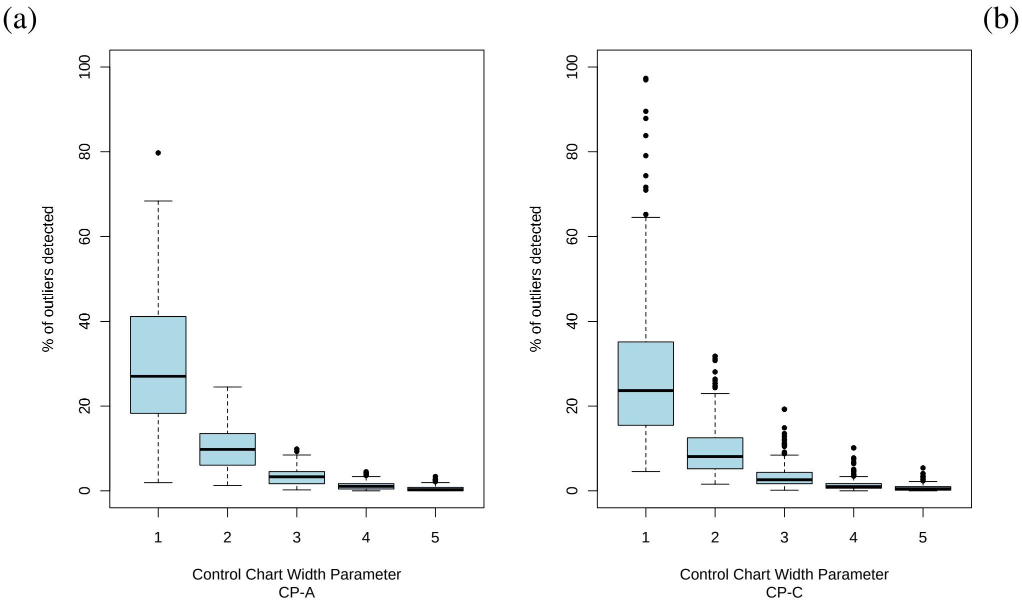


Figure 17. Control chart width paramater evaluation in terms of % of points in a curve across all curves detected as outliers.
These boxplots showcases the percentage of points within a curve detected as outliers by using a specific QC width parameters (i.e., 1, 2, 3, 4, 5) into the outlier detection methodology implemented for all curves across both cell types: CP-A (a) and CP-C (b). The smaller the QC parameter the more outliers are detected as this parameter controls the in-control region.

### 5.3.3. Feature extraction models

#### 5.3.3.1. Cumulative sum control (CUSUM) charts: Shift detection

With the use of cumulative sum control charts (CUSUM) control charts, we are able to detect small shifts from the mean more efficiently than Shewhart control charts [Montgomery, 2005]. There are two input parameters needed to calculate the CUSUM statistic ($C_k$): subgroup size ($k$) and the in-control mean (in this study $\mu_0 = 0$). The $C_k$ statistic is defined in Equation 5.6 by $k$, $\mu_0$, and the computed mean within the sample $k$ ($\bar{x}_k$). This statistic is calculated along the entire sample range.

$$C_k = \sum_{j=1}^{k} (\bar{x}_k - \mu_0) \tag{5.6}$$

Other parameters considered to detect when the process is out-of-control (in this case $\mu_0 \neq 0$) were the decision interval and the amount of shift or slack detection. Recommended values for these parameters are decision interval of size $5$ and slack value of $3$ [Pignatiello and Runger, 1990, Golosnoy et al., 2009, S.S. Prabhu and Montgomery, 1997]. To apply this CUSUM procedure, the OC curves are order-reversed to identify a deviation from zero (tail).

#### 5.3.3.2. Piecewise linear regression model

The methodology implemented in this paper for feature extraction consists on fitting piecewise linear regression model to every curve. A piecewise linear regression models a nonlinear relationship from a number of linear segments where breakpoints meet at the same point. We considered the special case of two linear regressions joint by a single change-point at time $t$ as shown in Equation5.7 with indicator variable ($I_{x \geq t}$) equal to one when $x \geq t$ [Berk, 2008] constrained by modeling negative slopes in their linear relationships.

$$\begin{gathered} y = \beta_0 + \beta_1 x + \beta_2 (x - t) I_{x \geq t} \\ \text{where } \beta_1 \leq 0 \ \text{ and } \beta_2 \leq 0 \ \forall curves \end{gathered} \tag{5.7}$$

One challenging problem is the detection of a change-point (i.e., breakpoint). A likelihood method is used to find a breakpoint which minimizes the sum squared of error ($SSE$) of fitting two linear regressions by that particular breakpoint. Exhaustive search is performed along all possible time values to fit the piecewise linear regression model to choose a time point which minimizes the fit error. Once the breakpoint is selected features from the spline model are extracted and studied across different conditions (i.e., CP-A, CP-C). Furthermore, continuity was enforced in the model. This constrained piecewise linear regression with one-breakpoint fit is statistically compared to the fit of a simple linear regression model using an $F$ test as shown in Equation 5.8.

$$\begin{aligned} F &= \frac{\dfrac{SSE_{model1} - SSE_{model2}}{p_{model2} - p_{model1}}}{\dfrac{SSE_{model2}}{n - p_{model2}}} \\ if &\rightarrow F > F_{\alpha, p_{model1}, n - p_{model2}} \\ then &\rightarrow model_2 \;\; better \end{aligned} \tag{5.8}$$

The model comparison by an $F$ test is performed multiple times. This presents a commonly known problem in multiple hypotheses testing which is the increase of false-positives. There exist several routines created to alleviate this problem such as Bonferroni correction. This widely used technique is applied when multiple statistical tests are computed simultaneously in order to reduce false-positives by reducing the value of $\alpha$. Another way in which the value of $\alpha$ can be reduced is by adjusting all the $p$-values from the individual tests as shown in Equation 5.9 where $n$ is the number of comparisons [Benjamini and Hochberg, 1995, Benjamini and Yekutieli, 2001, Shaffer, 1995].

$$p_{value.adjusted}[c] = min(p_{value}[c] \times n, 1) \;\; c \in [1, n] \tag{5.9}$$

### 5.3.4. Comparisons and classification techniques

#### 5.3.4.1. Statistical significance tests

Extracted features were studied and compare across both cell lines using traditional statistical tools such as histograms, confidence of interval plot and statistical tests on the mean and median. The mean was tested using analysis of variance (ANOVA) test which generalizes $t$-test for more than two groups but relies on several assumptions that might or might not be met for this particular data structure. It is performed with caution to get a general sense of the groups' mean from the ANOVA hypothesis shown in Equation 5.10. Based on this we also tested for differences in the median using nonparametric tests which waive those strict assumptions. The median or rank test was performed using Mann-Whitney-Wilcoxon test for a two-level group test and Kruskal-Wallis test for more than two groups. Both are nonparametric tests evaluating differences in location shift of the distribution of $x$ among each group. Equation 5.11 gives the Kruskal-Wallis test statistic where $n_i$ is the number of observations in group $i$, $r_{ij}$ is the rank of observation $j$ from group $i$, and $N$ is the total number of observations for all groups. The $p$-value related to $K$ is approximated through the $\chi^2$ distribution [Kruskal and Wallis, 1952].

$$H0 : \mu_1 = \mu_2 = ... = \mu_n \tag{5.10}$$

$$K = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2} \tag{5.11}$$

#### 5.3.4.2. Ensemble classifier: Random forests

To further explored relationships among several groups of OC curves (i.e., CP-A_1, CP-A_2, CP-C_1, and CP-C_2), we applied an ensemble classifier based on decision trees. Decision trees can be applied in almost all scenarios. Therefore, they provide a good starting point for feature selection for heterogeneous and large data sets. They apply to either a numerical or categorical response. They are nonlinear, simple and fast learners that handle also both numerical and categorical predictors well. They are scale invariant and robust to missing values. However, a single tree is produced by a greedy algorithm that generates an unstable model [Tuv et al., 2009]. Consequently, ensemble methods have been used to counteract the instability of a single tree.

Supervised ensemble methods builds a set of simple models called base learners and use a weight outcome for each base learner in a voting scheme to predict future points. In other words, ensemble methods merge outputs from multiple base learners to create a voting committee to improve performance. Many empirical studies validates that ensemble methods often outperform any single base learner [Tuv et al., 2009]. Parallel and serial are two main approaches to build ensembles. Random Forest is an example for parallel ensembles which is an improved bagging method. It grows a forest of random decision trees on bagged samples showing excelling accuracy results, comparable with the best known classifiers [Tuv et al., 2009, Breiman, 2001]. An advantageous property of random forest classifiers, they do not overfit data based on its embedded out-of-bag error estimation. Other advantages of random forest models are: simple to train and tune in many applications, runs efficiently, can handles large number of variables, provides variable importance scores, embedded method to estimate missing data, generation of proximity matrix among cases, detects variable interactions, adapted to balance error due to unbalance datasets, and capable of extending to unlabeled data for unsupervised clustering, data views and outlier detection [Breiman, 2001].

Algorithm : A simple pseudocode for random forest construction is shown below: [Hastie et al., 2009, Breiman, 2001].

1. Draw a sample number of cases with replacement from the original dataset to build the training data.

2. Use training data to grow a tree.
    (a) Select $m$ variables at random from the total number of input variables ($M$) where $m << M$.
    (b) Best variable among the $m$ predictors is chosen.
    (c) Split the chosen node into two daughter nodes.
3. Repeat Step 2 until all trees are built.
4. Output ensemble of trees.

Important features of random forests are out-of-bag sampling (OOB), variable importance and proximity plots. OOB sampling is identical to cross-validation and since random forest is performed in parallel trees, cross-validation can be done along the way. Variable importance is a key feature of random forests and gradient boosted trees. The variables are ranked based on their improvement on the empirical loss function among all trees, meaning that variables that are chosen often in the trees provide better predictive power or they minimize the loss function. Lastly, proximity plots provide a visual of which observations are closer together [Hastie et al., 2009]. These proximity distances are measured by putting all the data, training and out-of-bag, through the grown trees. If instances $i$ and $j$ are in the same terminal node their proximity increases by one and so on through all the trees [Breiman, 2001]. Then proximities are normalized by the number of trees in the model.

State-of-art visualization methods such as multidimensional scaling is used to illustrate selected features to discriminate among different conditions using scaled version of these proximity measures [Cox and Cox, 1994]. Multidimensional scaling finds a configuration of the data points explored through their similarities or dissimilarities [Chen et al., 2008].

## 5.4. Results and discussion

### 5.4.1. Data preprocessing

Through the early exploration of oxygen consumption (OC) measurements at single-cell level we found a need to reduce noise or unwanted perturbations from the signals. This is the first challenge we encountered with this unique data structure as portrayed in Figure 14. Reducing noise from the signals will enhance the discovery of relevant features related to the “true” signal’s behavior. Three main stages of noise reduction are performed to the signals: (1) Smoothing of negative values; (2) Low-pass filtering; and (3) Outlier smoothing.

Smoothing of negative values is a simple approach which is used to attenuate negative values in the signal (see Methods section). Then common filtering techniques were applied to these unique oxygen consumption signals in two different ways. First, a filter is applied to each OC curves to estimate a metric of variation ($\sigma$) and use this as a parameter for outlier detection on the unsmoothed data. This outlier smoothing process uses traditional control charts in which any data point from each OC curves that falls away from curve-specific control limits is considered an outlier and hence its value is smoothed by neighborhood averaging (See details in the Methods Section). This step aims to smooth outliers or peaks which might be results of artifacts (i.e., sensitivity of the substrate sensor). Once the outliers are smoothed out the signal is processed through a low-pass filter.

These low-pass smoothing filters are Savitzky-Golay (SG) and Exponential Weighted Moving Average (EWMA), their methodology and parameter estimation is briefly explained in the Methods section. These low-pass filters provide a useful procedure to smooth data while preserving certain signal's features and both perform better than other filters in most cases [Press et al., 1992, Orfanidis, 1996, Tranter, 2000]. We assessed both filters through their goodness-of-fit statistics such as root-mean-squared-error ($RMSE$) showing similar performances (See Figure 16). Their similar smoothing performance and usage in practice have led us to choose the exponential filter based on its local patterns of fit and the number of parameters to be estimated is less than the Savitzky-Golay smoothing filter. Figure 18 displays an example of the smoothing outcome from this exponential filter for a particular OC time series.

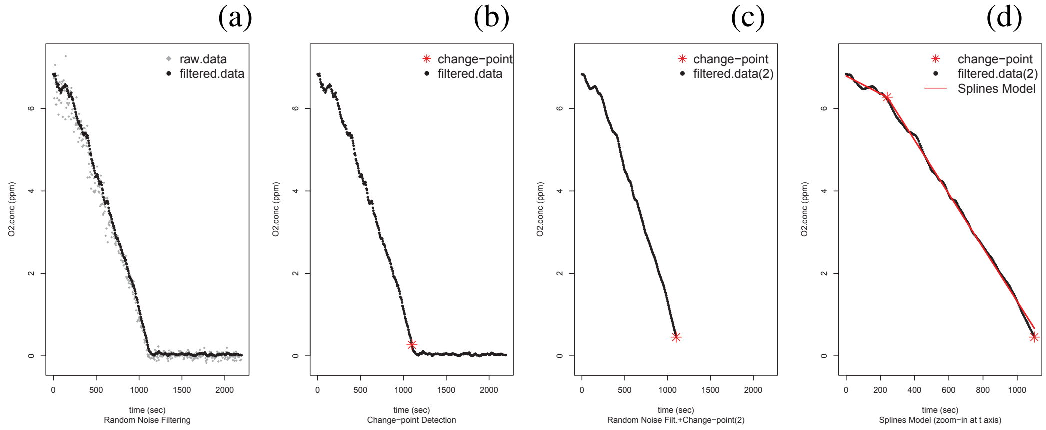


Figure 18. Step-by-step statistical framework example.
This figure shows main steps to characterize OC curves: (a) data filtering, (b) change-point detection (TimetoZero) using CUSUM, (c) removal of zero-valued tails, (d) identification of multiple states features using spline model (zoom in $x$ axis).

The importance of this noise reduction platform lies in the ability to automatically reduce random artifacts from multiple heterogeneous curves that will inhibit further data analysis towards knowledge discovery through feature extraction.

### 5.4.2. Feature Extraction

Once the oxygen consumption measurements were smoothed out, the data analysis procedure continues with the aims to understand the OC curves which are signals gathered from a heterogeneous dynamical system with multiple states such as molecular cells. The previous reflects challenge number two in Figure 14 which intersects into two sub problems: (1) removal of redundant information defined by experimental parameters comprehension and (2) extraction of distinctive features without systems' apriori knowledge.

#### 5.4.2.1. Time to reach zero detection

To address removal of redundant information from a heterogeneous dynamical system like Barrett's esophagus cells, we proceed to automatically detect the time point where each curve reaches zero and remove data from that point on (called zero-value tails) from further analysis. Hence, redundant information is defined in the context of experimental understanding of the limiting capacities of oxygen within the single-cell technology over time. This is explained by the experimental design of the data where cells are placed in finite-volume wells and enclose in it to study OC response. Further experimental details related to the data used in this study can be found at the Methods section. Removing this zero-value tails from each OC curves facilitates the modeling of these curves in regions of interests allowing for feature extraction within and across signals. This is a trivial problem if we analyze one or two curves and/or if the time to reach zero was the same for all curves, however this is not the case. The heterogeneity of the cells gets reflected on how fast or slow they can reach this point making it a challenge to identify this characteristic automatically for each OC curve.

Cumulative sum (CUSUM) control charts is commonly used statistical tool to detect small shifts. Its application allowed us to automatically detect the time point at which each OC curve reaches a value near zero (See Methods section). Figure 19 shows two plots portraying the high variation within and across cell types for this particular feature: time to reach zero. Figure 19a shows a histogram plot where frequencies of this features are accounted across the entire time spectrum with a higher concentration in $[0, 2000]$ seconds. We used this feature as a reference point to remove data that fluctuates near zero and not providing meaningful information for further data modeling. However, through this removal of redundant information we have captured a unique characteristic through change-point detection methodology that is later studied to understand cell heterogeneity.

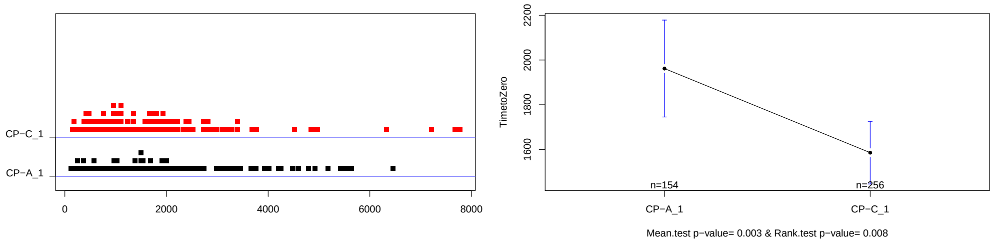


Figure 19. Features histogram and significance tests between CP-A and CP-C for TimetoZero.
The left graphic shows the frequency of values detected for times where OC curves where near zero val-ues and the right one displays $95\%$ confidence interval on the means for both cell types for this particular feature.

#### 5.4.2.2. OC characterization and other features

Modeling OC curves from single-cell technology is quite a challenging task since there is neither a priori knowledge of cell behavior nor how to effectively model this type of information. Despite the clear notion that cells are heterogeneous, little is known on how prominent the variations of specific factors such as oxygen consumption are in the context of different cell types and/or states [Molter et al., 2007, Lidstrom and Meldrum, 2003]. We address this challenge by approximating OC curves behavior through a constrained piecewise linear regression model. This particular spline model fits two linear regressions with positively-restricted slopes with one breakpoint optimally detected through a likelihood method across the entire time span. This model allows us to capture features in different segments of the data. It seemed appropriate to study OC curves by means of fitting two linear models from preliminary studies analysis which reveal this pattern (See Figure 18b).

Our efforts with this model are compared in terms of the goodness-of-fit against the simple linear regression model. We have performed the comparison of these two models by using an $F$ test on each curve. These multiple comparisons raise a commonly known problem in multiple hypothesis testing: increase of false-positives. To address this problem, we have corrected all computed $p$-values using a Bonferronni correction method. Through the evaluation of these tests, we found that $99.3\%$ and $97.7\%$ of the curves for CP-A and CP-C respectively had a better fit with the constrained piecewise linear regression model for $\alpha = 0.001$. Figure 20 shows the percentage of curves that showed better fit with the studied spline model across different levels of Type-I error for the $F$ test ($\alpha$). Both cell types showed a high percentage ($> 90\%$) of their curves have a statistically significant fit with the constrained piecewise model across different values of $\alpha$ with slightly smaller percent for CP-C than for CP-A.

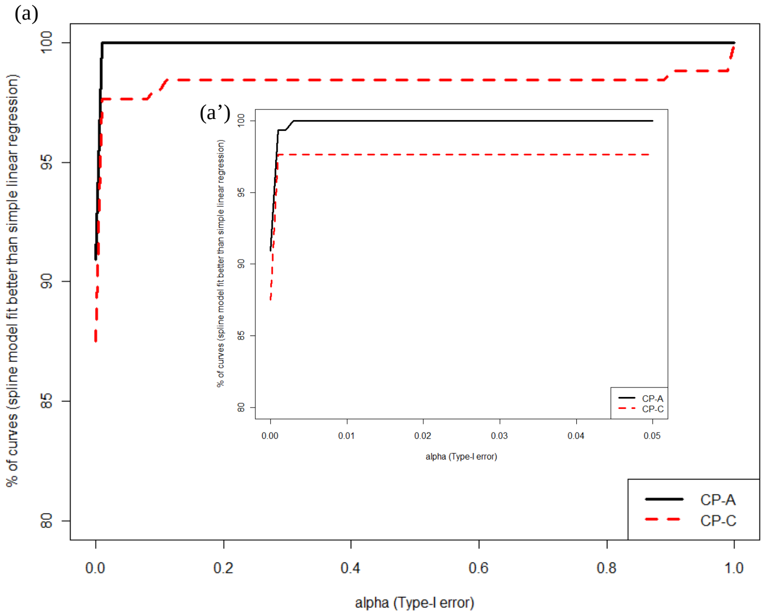


Figure 20. Multiple hypothesis testing for comparisons tests between spline model vs linear regression fit.

(a) Percentage of OC curves per cell type that revealed a better fit with the spline model discussed in this work are shown in this figure as a function of $\alpha$ (Type-I error). Bonferroni correction was applied to the individual test $p$-values to alleviate the multiple hypothesis testing problems with false-positives when multiple comparisons are performed. (a') Inner box showing alpha range of $[0, 0.05]$.

The model facilitated the extraction of relevant features that were used to characterize the OC curves. Besides the regular features from fitting linear regressions (i.e., intercept and slopes), we were able to detect other features as the ones shown in Table 13 with their respective descriptions. These useful features are extracted for each curve from both cell types (CP-A and CP-C) and

TABLE 13
Features extracted.

| **Features** | **Description** |
|---|---|
| Change-point.Time | Time value at which the best fit of two linear regressions meet |
| Change-point.Oxygen | Oxygen consumption value at which the best fit of two linear regressions meet |
| Intercept Coefficient ($B0$) | Intercept of left linear regression |
| Left Slope Coefficient ($B1$) | Slope of left linear regression[†] |
| Right Slope Coefficient ($B0$) | Slope of right linear regression[†] |
| Kurtosis | Measure of peakedness. Higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations. |
| Skewness | Measure of the asymmetry. |
| Minimum MSE | The Mean Squared Error value for the best two-linear regression fit hence minimum. |
| TimetoZero | Time at which the oxygen consumption has reached a value of zero |

Brief description of features extracted from curves after application of smoothing and filtering techniques.

[†] Slope magnitudes extracted from the spline model are divided by two for curves which conditions contained two cells per well.

analyzed further within and across cell type through their empirical distribution. Although such a model may not correspond to the biological function in the OC curves, it provided a good empirical fit to the experimental data with a simple structure permitting subsequent studies of conditions and cell types.

### 5.4.3. Biological inferences and interpretation

Once features such as slopes, intercepts, breakpoint or change-point were extracted for every curve, we proceeded to compare features among different cells within the same type and between the two different types. Some interesting patterns were found in certain features extracted from the spline model for the following features: TimetoZero, Change-point.Oxygen, Left.Slope and Right.Slope (details on Table 13). To detect differences between CP-A and CP-C features we computed two sets of significance tests. We statistically tested for the means and the medians of the features within the two cell lines.

The captured empirical distribution on the time point that each curve takes to reach oxygen concentration value near zero (TimetoZero) showcased a high range of values within and across cell types as mentioned previously (Figure 19a). Also, significant differences between cell types were found through mean and median statistical tests with $p$-values equal to $0.003$ and $0.008$ respectively as shown in Figure 19b. The different values for TimetoZero detected through CUSUM charts as a shift detection tool reaffirms the known notion of cell heterogeneity and its importance to study them at their single-cell level.

Another feature of interest in this study is the value of oxygen at the point where the two linear regressions from the studied spline model meet (Change-point.Oxygen). At the breakpoint of the spline model two features can be captured: oxygen and time. The Change-point.Oxygen feature presented a very interesting behavior in its empirical distribution. The resemblance of two subpopulations can be observed in Figure 21a with more definition on our larger sample cell type (CP-C). This thought-provoking pattern presents another dimension of the heterogeneity of cells even within same cell type. Across cell types there seems to be some subtle differences between both cell types. However, the mean and median test reflected $p$-values of $0.076$ and $0.085$ respectively for each test resulting in non-statically significant differences if the typical value of $\alpha$ is chosen to be equal to $0.05$.

Two other features of interests are the slopes or rates of the OC curves in study. Understanding how fast or how slow the cells consume oxygen is of great interest to scientists and researchers. The negatively-valued slopes showed a more tailed shape in its empirical distribution (see Figure 21(b,c)) concentrating values in the small ranges $[-0.02, 0]$ . For both types of slopes (left and right slope), no differences were found in terms of its mean when tested between both cell types. However, statistical differences were found for the right.slope in terms of its median values with $p$-value equal to $0.002$.

We further explored these comparisons as a classification problem with two classes (e.g., one cell type versus another) finding subtle differences between both cell types using an ensemble method: random forest. The classification problem raised an out-of-bag error rate of $30\%$ when classifying single-cell CP-A and CP-C from the extracted features discussed here and presented in Table 13. A multidimensional plot from the tested random forest is shown in Figure 22. The characterization of the OC curves through the constrained piecewise linear model facilitated the extraction of these features and its evaluations within and across cell types.

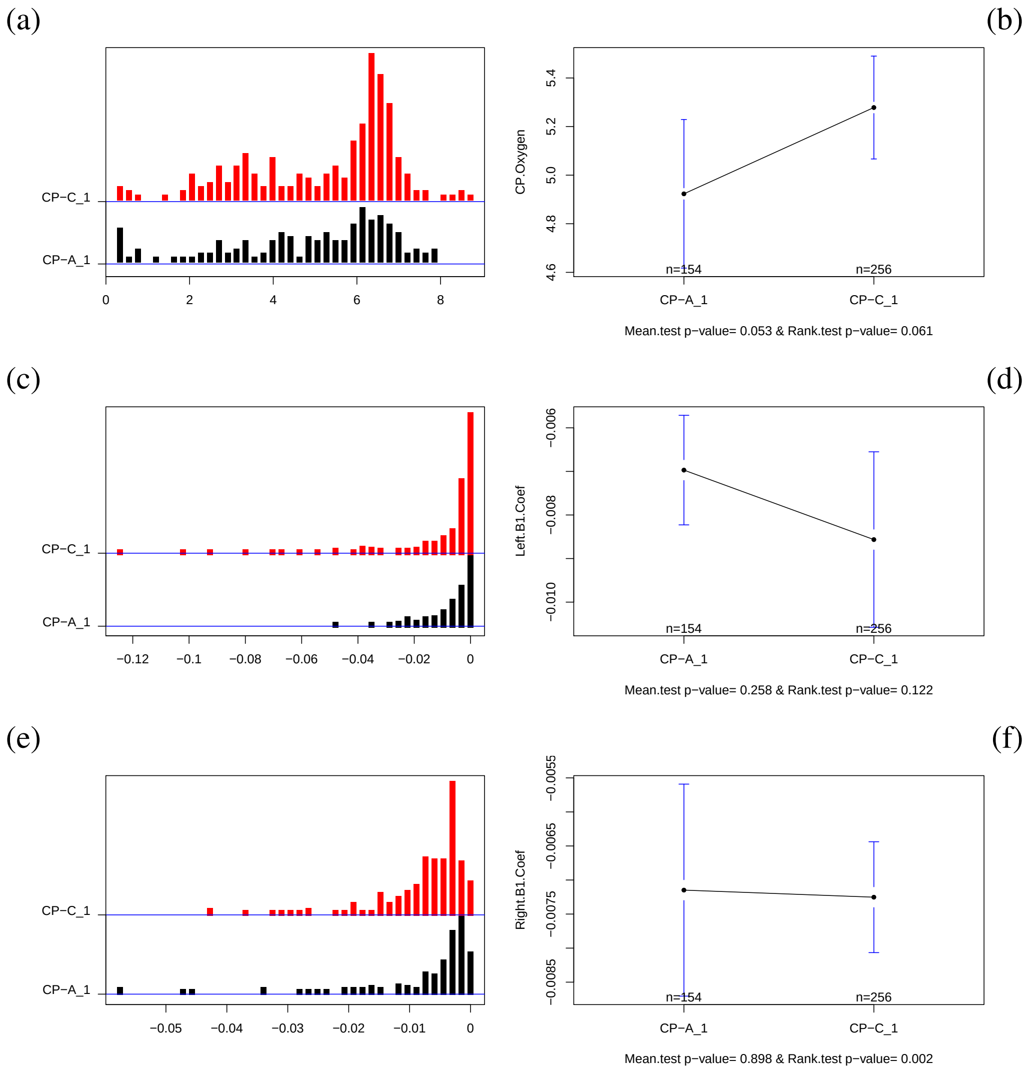


Figure 21. Features comparisons between CP-A and CP-C by means of a spline model.

Three main features extracted through the constrained piecewise linear model: (a,b) oxygen at the change-point, (c,d) left and (e,f) right slopes. Figures on the left show feature frequency values and those on the right show $95\%$ confidence interval of the features means.

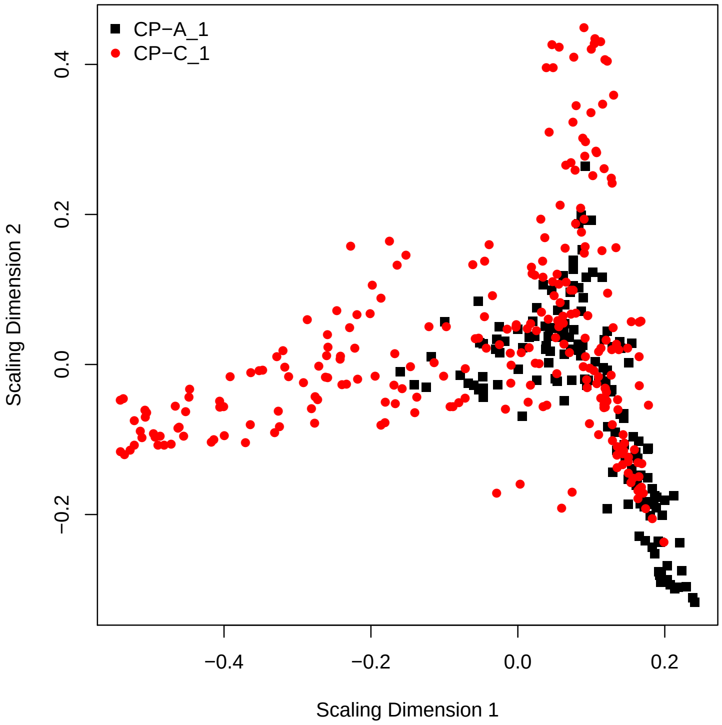


Figure 22. Multidimensional scaling plot: A random forest model for CP-A vs CP-C single-cell.

This plot provides a visual on the scaling coordinates of the proximity matrix obtained from a random forest to classify CP-A versus CP-C at the single-cell level.

### 5.4.4. Exploring comparisons beyond single-cell

The data analysis presented up to this point considered the single-cell scenario in which we represented the two cell lines as CP-A and CP-C. To further explore cell heterogeneity, OC curves were obtained through experiments beyond the single-cell level and characterized through the methodology presented in this paper. We compared features extracted from single-cells (i.e., CP-A_1 and CP-C_1) with double cells within a well (i.e., CP-A_2 and CP-C_2). The same statistical framework was applied to CP-A_2 and CP-C_2 OC curves with a small modification to certain features.

This modification is to adjust the slopes features from the spline model where magnitudes are divided by two to account for multiple cells. An assumption is made through this adjustment that respiration rates are equal for both cells within a well (which might not be necessarily true). However, this adjustment is needed to make somewhat comparable comparisons on the rates, hence lessen the effect of having multiple cells in a microwell.

The first aspect investigated among OC curves is the goodness-of-fit of the spline model presented in this paper. We statistically compared the spline model fit against the simple linear regression through multiple hypothesis testing with Bonferroni correction as described previously in Section 6.3. Similar to results presented for our two cell lines at the single-cell level, the spline model fit was statistically better than the simple linear regression model for all double cells measurements as shown in Figure 23.

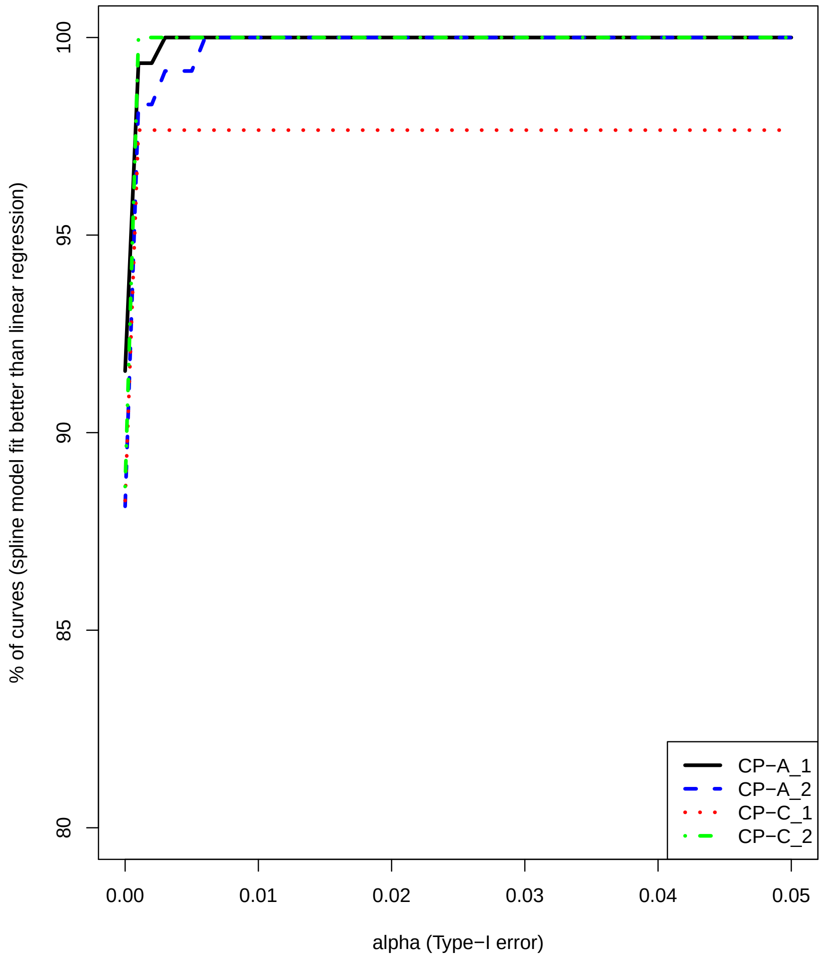


Figure 23. Multiple hypothesis testing for comparisons between spline models vs a simple linear regression model.

Percentage of OC curves per cell types that revealed a significantly better fit with the spline model discussed in this work are shown in this figure as a function of $\alpha$ (Type-I error). Bonferroni correction was applied to the individual test $p$-values to alleviate the multiple hypothesis testing problems with false-positives when multiple comparisons are performed.

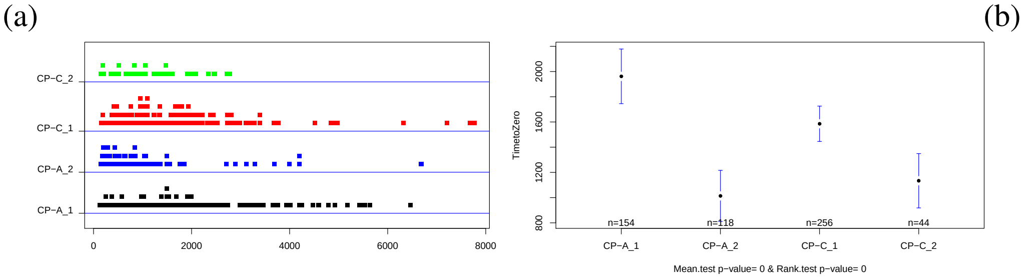


Figure 24. Time to zero extracted from single- and double-cells for CP-A and CP-C oxygen consumption curves.

Time to zero is a time feature extracted after the removal of zero-valued tails using CUSUM method. (a) This plot shows frequency values for this feature across single-cell curves (CP-A_1 and CP-C_1) and double-cells curves (CP-A_2 and CP-C_2). (b) $95\%$ Confidence interval plot of the means of TimetoZero per cell group. Testing for statistical differences for the mean and location shift (e.g., median) showed $p$-values equal to $0$ for both tests.

Throughout the implementation of the constrained piecewise linear regression model, a set of features from CP-A_1, CP-A_2, CP-C_1, and CP-C_2 data curves were extracted. Similar patterns as the ones discussed for the case of single-cell OC curves were found for OC curves with double-cells on features such as TimetoZero, Change-point.Oxygen, Left.Slope, and Right.Slope (See Table 13). Statistical differences in both mean and median were found for the distinct four groups of OC curves for the feature TimetoZero as shown in Figure 24. This difference was found previously among CP-A_1 and CP-C_1, however the mean difference is less marked for CP-A_2 and CP-C_2. Other features presented in Figure 25 shows subtle differences across the four groups studied in this section. These are expected differences as the more cells in the well, the faster the oxygen in the well will be consumed.

The features extracted through the applied statistical framework allowed for multiple comparisons for different phenotypes. As seen before, the empirical distributions for each of the features permitted comparisons and showcased subtle differences. This is an important contribution from this work. To further analyze the OC curves through the extracted features, an ensemble classifier is applied to these features with the objective of classifying the four groups of interest (CP-A_1, CP-A_2, CP-C_1, and CP-C_2). A random forest was applied to the extracted features to unravel nonlinear relationships among relevant features.

Preliminarily, we built random forest models for pair of classes (i.e., CP-A_1 vs. CP-A_2, CP-C_1 vs. CP-C_2, etc.) obtaining error rates $\sim 20\% - 30\%$ for all pairs (detailed results available upon request). These models included all features including TimetoZero. Including all four classes in one random forest model, the classification error rates were around $40\%$ and $50\%$ for all features and without the TimetoZero feature respectively (See Figure 26(a,b)). TimetoZero was removed since curves with higher number of cells will result in smaller times to reach a value of zero in oxygen. Hence, this is a natural feature

TABLE 14
Confusion matrices from random forest classification models.

All features included.

| | CP-A 1 | CP-A 2 | CP-C 1 | CP-C 2 | class.error |
|---|---|---|---|---|---|
| CP-A 1 | 75 | 24 | 51 | 4 | 0.512987 |
| CP-A 2 | 4 | 81 | 1 | 32 | 0.3135593 |
| CP-C 1 | 61 | 22 | 165 | 8 | 0.3554688 |
| CP-C 2 | 5 | 20 | 2 | 17 | 0.6136364 |

Without TimetoZero feature.

| | CP-A 1 | CP-A 2 | CP-C 1 | CP-C 2 | class.error |
|---|---|---|---|---|---|
| CP-A 1 | 74 | 29 | 45 | 6 | 0.5194805 |
| CP-A 2 | 20 | 61 | 13 | 24 | 0.4830508 |
| CP-C 1 | 60 | 28 | 142 | 26 | 0.4453125 |
| CP-C 2 | 7 | 17 | 8 | 12 | 0.7272727 |

Individual error rates per cell type and different number of cell within a microwell is shown for random forest models with and without TimetoZero. The out-of-bag error rates are $40.9\%$ and $49.5\%$ respectively.

to be important to the response but mainly due to the experimental settings.

Table 14 shows detailed error types from the model confusion matrices for both models with all features and without TimetoZero feature. As shown in Table 14, the number of curves among the four different classes is unbalanced. To address this problem, downsampling was perform on all random forest models presented here. In addition, Table 15 presents variable importance scores for both random forest models. It can be observed that TimetoZero received the highest score to predict these classes and when this feature is removed there is no clear set of features that ranked better than others.

TABLE 15
Variable importance scores from random forest classification models.

|  | MeanDecreaseGini |  |
|---|---|---|
| Features | All features model. | Without TimetoZero model. |
| Change-point.Time | 11.91773 | 17.06686 |
| Change-point.Oxygen | 14.5082 | 17.46699 |
| Left.B0.Coef | 13.42228 | 17.01607 |
| Left.B1.Coef | 13.69469 | 16.533 |
| Right.B1.Coef | 18.22898 | 16.48466 |
| TimetoZero | 22.60232 | - |
| Kurtosis | 12.06001 | 15.68385 |
| Skewness | 11.72529 | 14.98327 |
| MSE.min | 13.8405 | 16.76531 |

Variable importance scores are shown as the mean decrease in Gini index. This value measures the node impurity from splitting on the variable for which the importance score is been computed.

The results showed from the random forests models, although their predictability measures are not high, shows semi-defined clusters within same condition or cell type. Figure 26 shows how the data points from same cell type tend to agglomerate in regions with certain overlap among other cell types. This random forest model highlights a way to extract nonlinear patterns among the features to discriminate among different classes if these patterns exist. Furthermore, as these cell lines belong to premalignant stages of a particular cancer, its relationship is closed and the need for further features from these OC curves or any other data parameters might be required to elucidate clearer differences among them.

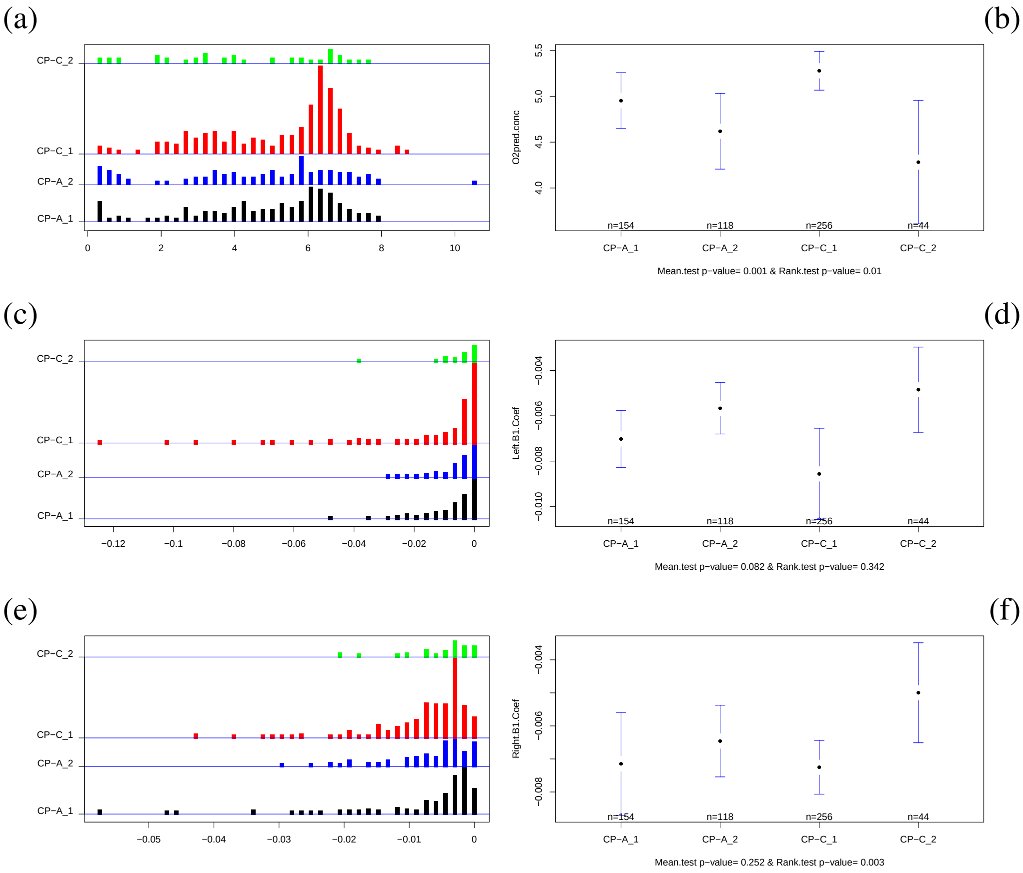


Figure 25. Other features of interest extracted from single- and double- cells for CP-A and CP-C oxygen consumption curves.

Figures on the left show feature frequency values and those on the right show 95% confidence interval of the features means. Figures (a,b) are oxygen values at the spline model breakpoint. Figures (c,d) are related to the slope of the first linear regression on the spline model (Left.Slope). Figures (e,f) are the slopes from the second linear regression (Right.Slope).

## 5.5. Conclusion

The analysis and interpretation of cellular heterogeneity has been a fundamental challenge of biology since much of the information available relies on collections of data. A great deal of interest is found in the scientific community for its relationship to cellular process such as carcinogenesis [Altschuler and Wu, 2010, Kelbauskas et al., 2011]. Therefore innovative technologies are been developed to perform single cell studies; allowing for data acquisition of complex structures of information never analyzed before such as oxygen consumption curves. The effective analysis of this type of information faces a list of challenges that no current methodology has addressed before. In this study, we have identified three major challenges in analyzing cellular heterogeneity from real-time measurements at the single-cell level. These major data challenges are: random noise, heterogeneous dynamic systems with multiple states, and understanding cell behavior within and across different cell types (Figure 14).

The problem of high-levels of noise was handled through a three main step procedure: average smoothing of negative values, low-pass filtering, and outlier smoothing. Many recognized techniques in the area of statistical signal processing were evaluated as well are their required parameters. Through this extensive evaluation of methods, we were able to establish techniques and best suited parameters to reduce random noise due to random effects such as sensitivity of the sensor.

Identification of relevant states from OC curve measurements in cellular systems is highly challenging based on the heterogeneity and unknown cell dynamics. Every curve possesses unique features that should be extracted by taking into account its individual identity. To understand system dynamics of the cell, we characterized OC curves as a set of features (i.e., time to reach near zero values, oxygen values in change-points, rates, etc.) through change-point detection techniques and a piecewise linear model that approximates these type of measurements. These techniques provided a good empirical fit to the data and provided a framework to analyze it effectively. Although such a model may not correspond to the biological function in the OC curves, it provided a good empirical fit to the experimental data with a simple structure.

The extraction of features per curve allowed for inferences within and across cell types revealing some interesting patterns and subtle differences between CP-A and CP-C with single- and double-cells per well. We extended the application of several set of tools used in signal processing and statistics for approximation and feature extraction. This methodology addresses relevant challenges that arise from this novel data acquisition and experimental modalities at single-cell level and provides a statistical framework to extract meaningful summaries from the data. Caution should be employed when making inferences in its biological impact. This study reveals the need for more complex model to characterize these curves better.

## 5.6. Acknowledgments

I would like to thank the co-authors of this research work, Drs. Shashanka Ashili, Laimonas Kelbauskas, Weiwen Zhang, George C. Runger, and Deirdre R. Meldrum. The authors would like to thank the Center for Biosignatures Discovery Automation at the Biodesign Institute at the Arizona State University. Funding: The work is supported by the National Human Genome Research Institute (NHGRI) Center of Excellence in Genomic Sciences (Grant No. 5 P50 HG002360) of the National Institutes of Health (NIH).

Conflict of Interest: none declared.

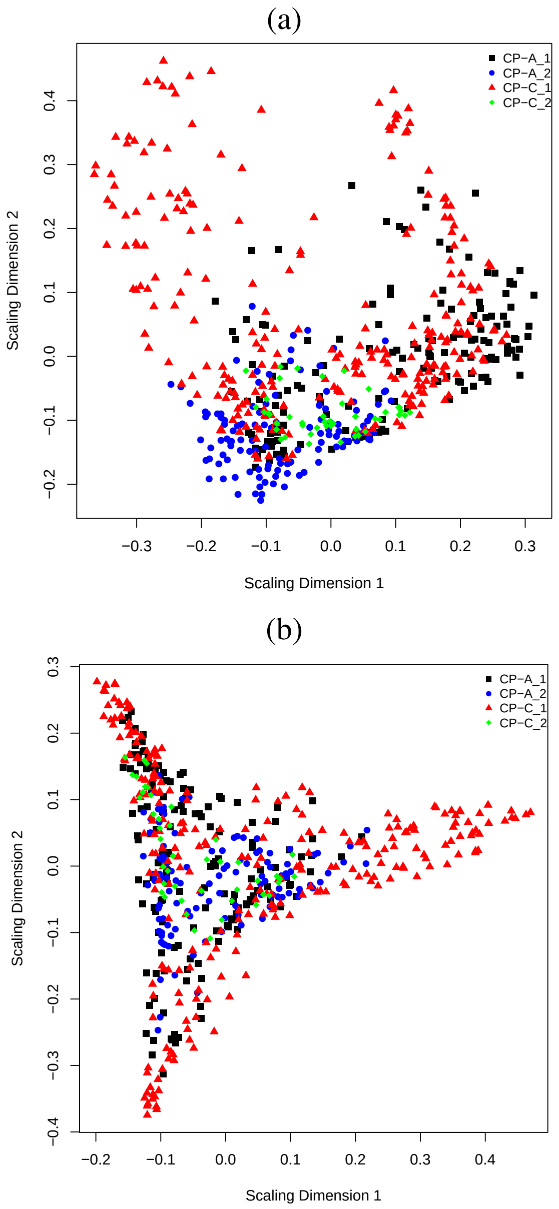


Figure 26. Multidimensional scaling plots: A random forest model for CP-A vs CP-C at single- and double-cell level.

This plot provides a visual on the scaling coordinates of the proximity matrix from random forest performed to classify CP-A versus CP-C at the single- and double-cells level. (a) Results using all features as described in Table 13. (b) Results using all features except TimetoZero.

# CHAPTER 6

# NOVEL TWO-DIMENSIONAL REPRESENTATION FOR MULTIPLE TIME SERIES

## 6.1. Abstract

Motivated by the challenges of analyzing real-time measurements, we further explored a unique two-dimensional representation of multiple time series using a wavelet approach which showcased promising results towards less complex approximations. A case study was performed using near infrared (NIR) sensor-based measurements to evaluate the performance of the two-dimensional representation. Also, ordering effect of the two-dimensional image was studied and explored to improve the image approximation through the benefits of external variables.

## 6.2. Introduction

There is a great magnitude of functional data available in several applications such as manufacturing, health care, financial businesses, web traffic, and others [Ramsay and Silverman, 2002]. As its name reveals functional data is data gathered as a function of a particular parameter which in most cases is time but could include others such as wavelength, probability, location, etc. Focusing on time component, the problem becomes a multiple time series. Extensive work has been done in this area to approximate, characterize, extract features such as change-point and model building for the data curves [Hannan, 1970, Lutkepohl, 2005, Picard, 1985, Shiohama and Taniguchi, 2003, Wei and Keogh, 2006]. Most of the literature focuses on traditional statistical methodologies like profile analysis, repeated measures, and growth curves, this last one addressing more the element of time dependencies.

Motivated by the continuous generation of multiple time series with unique features such as the oxygen consumption data curves discussed in Chapter 5 and the advancements achieved by current methodologies to analyze data in two dimensional (2D) structure (i.e., images), this research presents a unique 2D representation of multiple time series. This unique representation facilitates the application of state-of-art methodologies that does not rely on strong assumptions made by traditional methods as growth curves where special covariance structures among the time series is required. Additionally, we explored the integration of one-dimensional (1D) data information cognate to the multiple time series that serves as ordering parameters and finally could achieve variable ranking measurements.

To assess the performance of the novel 2D representation, this work investigate multiple time series approximation problem using a wavelet approach and tested through three publicly available datasets gathered through near infrared technology (NIR) from Eigenvector Research Inc. [Eigenvector Research Inc., 2009]. The 2D representation approximation performance is analyzed in terms of the trade-off measures: accuracy (i.e., sum squared error (SSE)) and model complexity (i.e., number of wavelet coefficients) and compared to a multiple one-dimensional (M1D) wavelet approach.

An extensive computational study to assess the 2D representation performance is discussed here. The study revealed less complex approximations using the novel 2D form of multiple time series than the one-dimensional structure. However, for the more noisy datasets, regions where M1D perform better than 2D in terms of accuracy but same level of complexity was found. Lastly, we examined the ordering effect of the time series on its approximation using related 1D data variables. Different ordering of the multiple time series can results in a different 2D image representation and consequently a different approximation. Ordering the time series by external 1D variables improved the complexity of the approximations when certain variables were used to order the curves. This allows for measurement of variable importance and it was tested for the NIR dataset with additional 1D variables available.

## 6.3. Methodology

### 6.3.1. Data description

This study studies near infrared (NIR) data with spectral intensities over several wavelengths. The three datasets evaluated here were freely available from Eigenvector Research Inc. [Eigenvector Research Inc., 2009]. These spectral intensities were recorded from about 800 nm to 2500 nm which reflects the near infrared wavelength range. The three datasets are labeled throughout this paper as follows: Corn, Diesel, and Shootout and described further a continuation.

Near Infrared Corn Dataset: The original dataset includes 80 samples of corn samples measured on three different NIR spectrometers over a wavelength range of 1100-2498nm with 2 nm intervals (with about 700 data points per sample). The first row of plots shown in Figure 27 presents three different data representations of this corn dataset. The first is the image used for 2D wavelet analysis. The middle plot in this first row shows a plot of the raw spectral intensities across a range of different wavelengths. The last plot shows the 256-scaled matrix of intensity values used on the multisignal-one-dimensional analysis. For these study we focused on spectra data from one instrument, instrument m5. In addition to the spectra data, we incorporated information of external variables such as moisture, oil, protein, and starch in this study as ranking criteria which are presented on Subsection 6.4.2.

Near Infrared Diesel Dataset: This dataset includes 784 samples with raw spectra intensities in the wavelength range of 750nm - 1550nm. An image, plot of the raw data points, and a plot of the scaled-raw points is presented in the second row of Figure 27. Some partial information on the properties of the diesel samples were available but not incorporated in this study based on their large amount of missing values.

Near Infrared Shootout Dataset: The International Diffuse Reflectance Conference (IDRC) reported a set of spectra information for 654 pharmaceutical tablets from two spectrometers. These were stratified in three different sets (calibration, validation, and test). We focused on the calibration data which contained 155 tablets samples. The data is includes the following wavelength range 600nm - 1900nm. Similarly to the Corn and Diesel dataset, we showed the image and plots which describes this dataset on Figure 27 at the last row of plots.

### 6.3.2. Two dimensional representation of multiple time series

#### 6.3.2.1. Data preprocessing

The raw intensity values for spectral numbers of the three datasets were scaled to 0-$255$ to make equivalent comparison between both representations ($2D$ and $M1D$) through wavelets methods. This scaling step is automatically implemented through the 2D wavelet Matlab interface where the images are scaled to a $256$ color scale. Scaling the data accordingly in both data formats (M1D: value-matrix and 2D: image) ensured comparable SSE comparisons.

#### 6.3.2.2. Imaging representation details

Representing multiple time series in a two-dimensional manner is possible by displaying every value as a pixel on an image where its color intensity depends on the value. Every time series in the dataset takes on one horizontal line of pixels in an image with intensities showcasing its scaled numerical value. For example, the $256$ color image portrays values from zero to $256$ as pure black to white with a gradient of grays in between.

This novel representation permits the usage of methods previously developed to analyzed image patterns more effectively in practice. Figures 27 a shows the constructed $2D$ representations from its matrix structure. The time series raw and scaled values are shown respectively in Figures 27 (b,c).

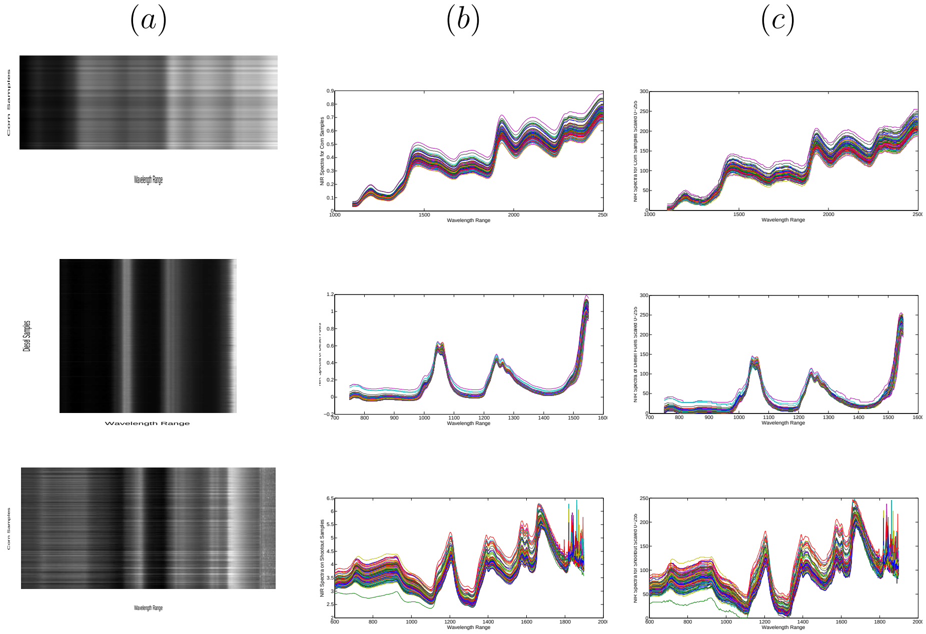


Figure 27. NIR spectra data plots for corn, diesel, and pharmaceutical tablets (shootout) samples.

The figure shows: (a) image representations, (b) raw data plots, (c) $256$-color scaled plots.

### 6.3.3. Wavelet decomposition implementation

#### 6.3.3.1. Background

Wavelet analysis consists of decomposing a signal(s) or an image(s) in a stratified manner into approximations and details. Wavelet theory is based on this decomposition approach over localized time-frequency domain [Mallat, 1999]. Any function can be expressed as a linear combination of basis functions and in wavelet transformation these basis represents "wavelet" which have time-widths adapted to their frequency [Daubechies, 1992]. A wavelet $\psi$ is usually called the Mother Wavelet which has zero bias and zero unit norm as shown in Equation 6.1.

$$\int_{\infty}^{-\infty} \psi(t)dt = 0 \;\; and \;\; \int_{\infty}^{-\infty} \psi(t)^2 dt = 1 \tag{6.1}$$

A general wavelet function can be represented as shown in Equation 6.2 were $s$ is a scale parameter and $u$ is a translation parameter. The scale $s$ focused on the frequency changes and the translation $u$ moves the wavelet and concentrates on local features.

$$\psi_{s,u}(t) = \frac{1}{\sqrt{s}}\, \psi\left(\frac{t-u}{s}\right) \tag{6.2}$$

This study was performed for the most common wavelet family: Haar. Haar is a simple piecewise constant function $\psi(t)$ and its scaling function $\phi(t)$ both described in Equation 6.3 [Haar, 1910].

$$\psi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1/2; \\ -1 & \text{if } 1/2 \leq t < 1; \\ 0 & \text{“otherwise”.} \end{cases} \quad \phi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1; \\ 0 & \text{“otherwise”.} \end{cases} \tag{6.3}$$

Wavelet decomposition is executed by levels. The level of decomposition will depend on the nature of the signal and the expertise of the analyst. Approximation and detail coefficients are computed at every level of decomposition $j$ such that $A_j$ is the $j$-level approximation and $D_j$ is the $j$-level detail or deviation of the signal from the $j$-level approximation. The wavelet level of decomposition is an important parameter that was fixed to a value recommended in Ganesan et al. [Ganesan and Venkataraman, 2004]. Let $\tau$ be the level of decomposition equal to $x/2$ where $N = 2^x$ and N is the number of instances in the dataset. For this study in particular we used $\tau = 5$ more details in Section 6.3.3.2.

This section summarized some background information about discrete wavelet transform (DWT). Since we aim to compare wavelets performance using 2D wavelets against the traditional M1D wavelet which means repeated

1D wavelets for each data profile we finish this section with some details on these two. As described above 1D analysis is based on one scaling function $\phi$ and one wavelet $\psi$ and one scaling function $\phi(x_1, x_2)$ and three wavelets $\psi_1(x, y) = \phi(x)\psi(y)$, $\psi_2(x, y) = \psi(x)\phi(y)$, and $\psi_3(x, y) = \psi(x)\psi(y)$ and the 2D analysis uses one scaling function $\phi(x_1, x_2)$ and three wavelets [The Mathworks Inc., 2009b].

#### 6.3.3.2. Wavelet implementation

The methodology presented was implemented using the Wavelet Toolbox in Matlab [The Mathworks Inc., 2009a]. This toolbox offers both 2D and M1D options for wavelets which facilitated the implementation of this study. After scaling the data in both in matrix and image form (see Section 6.3.1), we continue to apply wavelets in the Matlab platform in an Intel$^{®}$ Core$^{\text{TM}}$ 2 CPU of 2.00 GHz. The signals(or images) were decomposed first using discrete wavelet decomposition and later compressed using different compression thresholds. Details to accomplish wavelet decomposition at the 2D and M1D level are discussed below together with the estimation and selection of appropriate parameters for the decomposition procedure.

Parameters : Several parameters were used for decomposition and compression.

The essential parameters for both 2D and M1D wavelets decomposition are level of decomposition ($\tau$) and wavelet family ($w$). These two parameters were kept the same for all analysis and all datasets with values of $\tau = 5$ and $w = haar$. Based on the recommended method by Ganesan et al. [Ganesan and Venkataraman, 2004], recommended level of decomposition for the three NIR datasets are: $\tau_{corn} = 3.16$, $\tau_{diesel} = 4.81$, and $\tau_{shootout} = 3.64$. We chose a close $\tau$ value between all three datasets, hence, we set $\tau_{all} = 5$. In terms of the wavelet family we decided to use an orthogonal wavelet family and the most commonly used in the literature. Once signals (or images) are decomposed we proceed to compress them. The compression method is an relevant parameter available in Matlab some of the options are energy ratio, zero coefficients ratio, global threshold, and manual method. Depending of the compression method chosen, a threshold value needs to be assigned. For example, if zero coefficients ratio method is picked a respective threshold value, in this case, a percentage value should be specified such as $20\%$. This will reflect $20\%$ of the coefficients are set to zero. For our study we used global thresholding method with threshold values ranging from $0$ to maximum threshold value which for all three datasets was $1000$ in interval of $0.5$. Thresholding type is another parameter used in compression and there exists two types: soft and hard thresholding. We evaluated both types and gathered equivalent results. Furthermore, Matlab provides the option of keeping the approximation coefficients away from thresholding. For this study we applied compression to all coefficients including the approximation details, however both options were tested with no significant differences. This section compiles all parameters set during our Matlab implementation. These parameters were kept fixed for comparison purposes of both wavelets methods in study which we briefly explain below.

MultiSignal-One-Dimensional Wavelets with Matlab : The Wavelet Toolbox in Matlab [The Mathworks Inc., 2009a] includes several functions for wavelet analysis for multisignals (i.e., a set of one dimensional signals ordered by row or column in a matrix form). This add-in basically performs the wavelet analysis for every signal (either row or column) in a computational efficient manner. The $mdwtdec$ command corresponds to the decomposition stage for multisignal structure. An important input of this decomposition function is the data which is imported as a matrix. Our datasets had signals organized by rows and wavelength by columns (more details at 6.3.1). The object obtained from $mdwtdec$ command is used for compression in Matlab function $mswcmp$. This compression function was used with the parameters values specified in Subsection 6.3.3.2.

Two Dimensional Wavelets with Matlab : Similar to the Multisignal-One-Dimensional Wavelets, 2D wavelets commands includes a first step of decomposition performed by function $wavedec2$ following parameters values previously stated. Decomposition is followed by a compression stage which uses the resultant decomposition object. This compression step is executed using Matlab function $wdencmp$ with inputs described at Subsection 6.3.3.2. In contrast to the multisignal, the data is imported as an image. We have converted the data matrix into a gray image scaled to $256$ colors by default through the image toolbox in Matlab. More details on the data preprocessing and scaling issues of image conversion for comparison purposes in discussed in Section 6.3.1.

## 6.4. Results and discussion

In this section we describe the results obtained in terms of 2D and M1D wavelets performance for the NIR datasets. We have evaluated two important components of performance in signal processing: complexity and approximation accuracy. Complexity was measured through the number of nonzero coefficients after compression and a residual measurement known as Sum Squared of Error ($SSE$) was used for approximation accuracy. Interesting outcomes were obtained and they are presented below. One summarizes the performance findings between both wavelet methods when the data was used with no ordering (see Subsection 6.4.1). Ordering plays an important role for the 2D representation of the time series implemented through wavelet image analysis and no effect to the multisignal method. Hence, we explored how does ordering the signal response by available external variables could affect the performance of the 2D wavelets analysis. We present these findings in Subsection 6.4.2.

### 6.4.1. Wavelet's performance with no response ordering

As stated before, we focus our interest in evaluating performance in terms of number of coefficients and SSE after compression. The 2D wavelet analysis showed to performed much better than the M1D wavelet method for Diesel dataset as shown in Figure 28. The 2D method clearly shows smaller values of SSE and number of nonzero coefficients than the M1D wavelet method. However, this behavior was not similarly found in the Corn and Shootout datasets. These two sets of data reflected some tradeoffs between both measurements of performance. As shown in Figure 28, the 2D wavelets analysis overperforms the M1D in terms of complexity in certain regions but M1D shows a better accuracy with equal complexity in certain regions for Corn and Shootout dataset. Even though the multisignal wavelet method performs generally well in accuracy cannot reach the small values of complexity from the 2D wavelets analysis. As seen in Table 16, 2D wavelets can reach reductions in the hundreds for both datasets in comparison of thousands for the M1D method with similar accuracy measurements. This is similarly seen for the Diesel dataset but with complexity reductions of thousands for 2D and ten thousands for M1D (see Table 16). However, this is only observed in the range of very small number of nonzero coefficients because SSE for 2D wavelets worsens when it moves away from this range.

Through this study we thought that the 2D wavelet analysis will outperforms the traditional M1D method since the image wavelet incorporates information across and between signals which is missing in the M1D analysis. Hence, we evaluated a measure of randomness for each dataset postulating that datasets with more clear patterns across and between signals will be better analyzed by the 2D instead of M1D. For this we calculated a measure of entropy for each set of data showing the following entropy values: $e_{corn} = 7.5740,\ e_{diesel} = 6.5045$, and $e_{shootout} = 7.3090$. It seems that randomness of the signal could determine

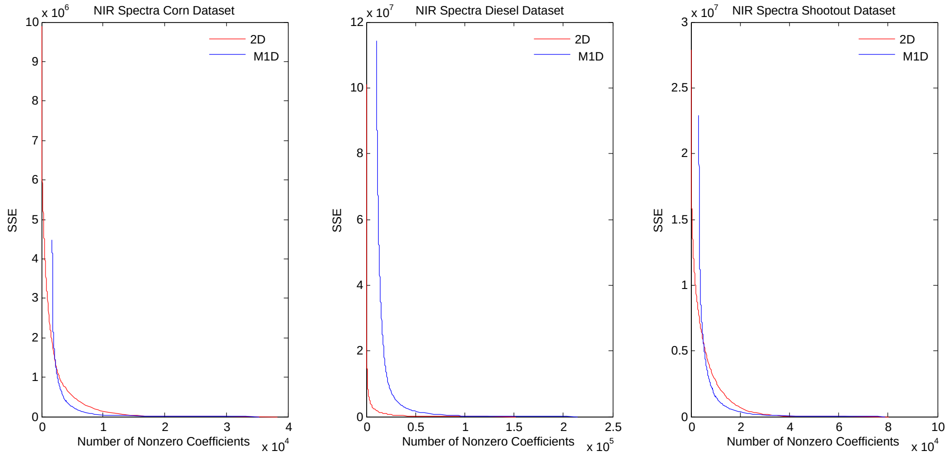


Figure 28. Results SSE and number of nonzero coefficients for all datasets.

which method could perform better in the 2D platform than multisignal since entropy for the diesel dataset showed a smaller value than the other two NIR datasets which in comparison are quite similar as their results in performance between wavelets methods. This shows some interesting results that could be explored further.

### 6.4.2. Comparison of wavelet performance with response ordering by external variables

To explore the ordering effect on the set of signals, we sorted the curves by external variables and measure its 2D performance using wavelets analysis.

TABLE 16

Coefficients comparison between 2D and M1D wavelets after compression with no ordering.

| GCT[1] | NIR Spectra - Corn Dataset | | | | NIR Spectra - Diesel Dataset | | | | NIR Spectra - Shootout Dataset | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2D Wavelets | | M1D Wavelets | | 2D Wavelets | | M1D Wavelets | | 2D Wavelets | | M1D Wavelets | |
| | Total Coefs=56064 | | Total Coefs=56080 | | Total Coefs=315880 | | Total Coefs=317520 | | Total Coefs=101691 | | Total Coefs=101215 | |
| | #Coefs | SSE | #Coefs | SSE | #Coefs | SSE | #Coefs | SSE | #Coefs | SSE | #Coefs | SSE |
| 0 | 38296 | 0 | 35313 | 0 | 151795 | 0 | 214544 | 0 | 80113 | 0 | 78607 | 0 |
| 0.5 | 29884 | 1145 | 33063 | 541 | 123127 | 3868 | 203540 | 2722 | 62712 | 2331 | 75764 | 665 |
| 10 | 6205 | 383047 | 4927 | 256816 | 14424 | 1264479 | 45336 | 1995870 | 19867 | 707946 | 14509 | 738556 |
| 20 | 2850 | 1024204 | 3253 | 624711 | 6931 | 2712127 | 30834 | 4853014 | 9727 | 2834522 | 9193 | 1833865 |
| 30 | 2112 | 1502870 | 2544 | 1041801 | 4262 | 4242716 | 23818 | 8832095 | 5205 | 5544257 | 6939 | 3213333 |
| 40 | 1268 | 2485229 | 2274 | 1361620 | 3033 | 5665914 | 20613 | 12960998 | 2664 | 8567553 | 5498 | 4940322 |
| 50 | 776 | 3431116 | 2013 | 1858869 | 2071 | 7500830 | 19224 | 15732749 | 1614 | 10597806 | 4625 | 6642446 |
| 60 | 546 | 4119400 | 1919 | 2134466 | 1734 | 8492880 | 17866 | 20121164 | 1024 | 12327079 | 4207 | 7882650 |
| 62.5 | _*459_ | _4449593_ | 1906 | 2183083 | 1649 | 8798832 | 17218 | 22550502 | 961 | 12559835 | 4144 | 8117673 |
| 63 | _*445_ | _4504531_ | 1904 | 2190937 | 1645 | 8814535 | 17101 | 23011138 | 945 | 12621716 | 4136 | 8149081 |
| 70 | 277 | 5250176 | 1876 | 2314041 | 1479 | 9557815 | 16244 | 26755380 | 662 | 13855431 | 4068 | 8446107 |
| 80 | 176 | 5801836 | 1846 | 2473965 | 1389 | 10077647 | 14824 | 34492988 | 518 | 14664039 | 3979 | 8957475 |
| 92 | 129 | 6114151 | 1842 | 2502271 | _*1240_ | _11189509_ | 14436 | 37092159 | 425 | 15331684 | 3769 | 10524318 |
| 92.5 | 127 | 6131102 | 1842 | 2502271 | _*1207_ | _11469489_ | 14420 | 37208612 | 408 | 15476090 | 3757 | 10626338 |
| 100 | 122 | 6167628 | 1840 | 2520531 | 1135 | 12135564 | 14059 | 40447002 | 369 | 15835984 | 3641 | 11691743 |
| 367 | 72 | 7549203 | 1760 | 4492689 | 524 | 32981932 | 10192 | 114463578 | _*129_ | _22868148_ | 3255 | 22932731 |
| 367.5 | 72 | 7549203 | 1760 | 4492689 | 524 | 32981932 | 10192 | 114463578 | _*128_ | ***23003113*** | 3255 | 22932731 |
| 1000 | 66 | 9895484 | _*1760_ | _4492689_ | 341 | 100979836 | _*10192_ | _114463578_ | 110 | 27857538 | _*3255_ | _22932731_ |
| 1500 | 66 | 9895484 | 1760 | 4492689 | 341 | 100979836 | 10192 | 114463578 | 110 | 27857538 | 3255 | 22932731 |

[1] Global Compression Threshold

To perform this task, we study the Corn dataset. This dataset included information on four external variables discussed in Section 6.3.1, hence we sorted the spectral profiles by these variables and investigated their 2D wavelets performance by variable as shown in Figure 29. In this figure, we can see how the variable moisture seem to improve 2D wavelet performance indicating a stronger relationship between this factor and the spectral intensities. In general, sorting this dataset by oil, starch, and protein not seem to improve the performance when compared when random order. When compared with M1D wavelet decomposition, we observed similar results as the ones shown in Subsection 6.4.1. Table 17 shows how the novel 2D image representation can attained number of nonzero coefficients in the hundreds compared with thousands for M1D when tested through wavelet decomposition approach. It is also observable how the ordering by moisture and protein reduced even more the complexity at similar SSE measures when compared with no ordering of the spectral response.

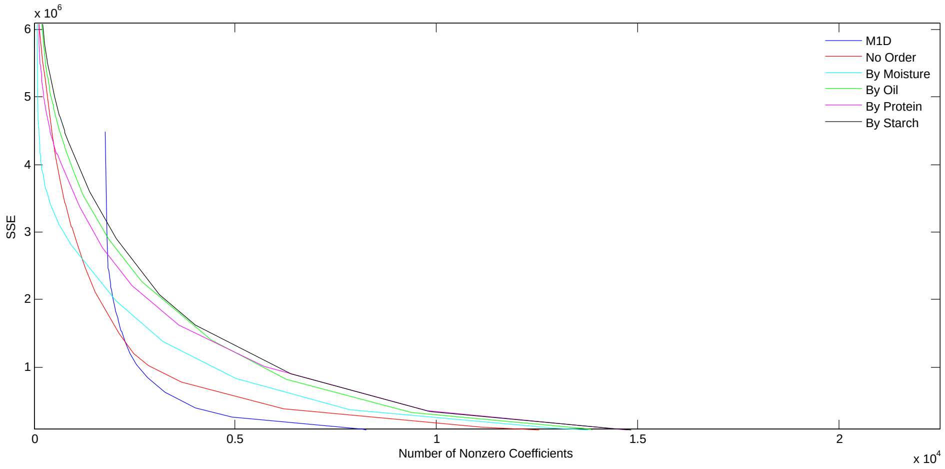


Figure 29. Results SSE and number of nonzero coefficients by ordering by external variables for corn dataset.
Trade-off metrics, accuracy and model complexity, are shown here for different ordering schemes performed on the image by one-dimensional variables: moisture, oil, protein and starch.

## 6.5. Conclusions

The analysis of multiple time series is an area widely explored in the literature of statistics where most methods rely in special cases and assumptions

# TABLE 17
## Number of coefficients for 2D and M1D wavelets after compression with ordered response (Corn: moisture, oil, protein, starch).

| GCT[1] | Wavelets analysis results ordering response. | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M1D | | 2D | | | | | | | | | |
| | | | Not Ordered | | Ordered by Moisture | | Ordered by Oil | | Ordered by Protein | | Ordered by Starch | |
| | #Coefs | SSE | #Coefs | SSE | #Coefs | SSE | #Coefs | SSE | #Coefs | SSE | #Coefs | SSE |
| 0 | 35313 | 0 | 38296 | 0 | 39960 | 0 | 40030 | 0 | 39733 | 0 | 40089 | 0 |
| 0.5 | 33063 | 541 | 29884 | 1145 | 31185 | 1252 | 31160 | 1254 | 31274 | 1155 | 31411 | 1221 |
| 1 | 20880 | 7542 | 22585 | 5026 | 23670 | 4925 | 23463 | 5061 | 24085 | 4741 | 24190 | 4691 |
| 2 | 13649 | 25562 | 18007 | 14812 | 19061 | 14771 | 19126 | 14208 | 19957 | 13367 | 19999 | 13552 |
| 3 | 10478 | 42386 | 15536 | 30987 | 16660 | 30359 | 16760 | 29652 | 17693 | 28003 | 17847 | 27506 |
| 4 | 8647 | 67849 | 13220 | 60866 | 14640 | 55909 | 14974 | 51841 | 15860 | 51268 | 15959 | 51042 |
| 5 | 8135 | 77488 | 11128 | 105805 | 12924 | 92839 | 13806 | 76599 | 14461 | 81118 | 14517 | 82071 |
| 10 | 4927 | 256816 | 6205 | 383047 | 7854 | 373048 | 9375 | 325876 | 9764 | 360812 | 9836 | 346467 |
| 15 | 4007 | 397249 | 3666 | 780783 | 5024 | 830172 | 6281 | 817035 | 5711 | 1003775 | 6386 | 900706 |
| 20 | 3253 | 624711 | 2850 | 1024204 | 3212 | 1373435 | 4398 | 1402019 | 3594 | 1620877 | 4015 | 1628087 |
| 25 | 2817 | 839162 | 2483 | 1204893 | 2035 | 1979321 | 2688 | 2259533 | 2441 | 2209346 | 3123 | 2075882 |
| 30 | 2544 | 1041801 | 2112 | 1502870 | 916 | 2814119 | 1855 | 2891309 | 1707 | 2771792 | 2046 | 2901687 |
| 35 | 2385 | 1202733 | 1529 | 2117433 | 628 | 3103576 | 1225 | 3542706 | 1129 | 3383336 | 1383 | 3597170 |
| 40 | 2274 | 1361620 | 1268 | 2485229 | 408 | 3420857 | 992 | 3881455 | 641 | 4044277 | 851 | 4332869 |
| 42.4999 | 2173 | 1533086 | 1077 | 2805327 | 372 | 3481195 | 775 | 4249172 | 574 | 4157970 | *769 | 4474441 |
| 42.50 | 2169 | 1540311 | 1074 | 2810745 | 372 | 3481195 | 766 | 4265428 | 571 | 4163389 | *755 | 4499729 |
| 44.9999 | 2081 | 1707945 | 941 | 3061483 | 344 | 3534122 | *655 | 4476196 | 498 | 4301677 | 644 | 4709955 |
| 45 | 2079 | 1711995 | 934 | 3075658 | 343 | 3536147 | *639 | 4508596 | 494 | 4309777 | 633 | 4732230 |
| 49.3749 | 2019 | 1844023 | 786 | 3406185 | 301 | 3628695 | 494 | 4824234 | *411 | 4491641 | 534 | 4947152 |
| 49.3750 | 2019 | 1844023 | 786 | 3406185 | 299 | 3633571 | 494 | 4824234 | *410 | 4494079 | 533 | 4949590 |
| 50 | 2013 | 1858869 | 776 | 3431116 | 293 | 3647097 | 484 | 4848936 | 401 | 4516171 | 509 | 5007634 |
| 55 | 1947 | 2043904 | 654 | 3767594 | 266 | 3716947 | 406 | 5057897 | 329 | 4710353 | 389 | 5336279 |
| 60 | 1919 | 2134466 | 546 | 4119400 | 232 | 3828822 | 366 | 5186838 | 286 | 4847096 | 352 | 5450843 |
| 62.7500 | 1905 | 2186999 | *448 | 4492687 | 204 | 3932067 | 345 | 5267180 | 261 | 4942340 | 334 | 5519462 |
| 62.7501 | 1905 | 2186999 | *446 | 4500562 | 204 | 3932067 | 344 | 5271118 | 261 | 4942340 | 331 | 5531275 |
| 65 | 1897 | 2219719 | 395 | 4710907 | 187 | 4000575 | 331 | 5323835 | 245 | 5007195 | 311 | 5612719 |
| 70 | 1876 | 2314041 | 277 | 5250176 | 181 | 4027914 | 293 | 5496285 | 235 | 5051600 | 269 | 5800709 |
| 75 | 1854 | 2426400 | 225 | 5521626 | 163 | 4120602 | 258 | 5678833 | 199 | 5239943 | 238 | 5964536 |
| 80 | 1846 | 2473965 | 176 | 5801836 | 153 | 4173302 | 244 | 5762671 | 183 | 5327635 | 209 | 6133124 |
| 85 | 1844 | 2487630 | 136 | 6062083 | 149 | 4200521 | 240 | 5790436 | 169 | 5418793 | 192 | 6239991 |
| 90 | 1842 | 2502271 | 129 | 6114151 | 146 | 4222691 | 219 | 5948339 | 164 | 5453298 | 180 | 6330174 |
| 95 | 1841 | 2511249 | 125 | 6148703 | 142 | 4249541 | 188 | 6207691 | 159 | 5491138 | 175 | 6371485 |
| 100 | 1840 | 2520531 | 122 | 6167628 | 138 | 4282891 | 157 | 6497434 | 154 | 5534380 | 170 | 6409478 |
| 125 | 1840 | 2520531 | 113 | 6273526 | 123 | 4477482 | 115 | 6960194 | 133 | 5791005 | 115 | 7107503 |
| 133.5625 | 1840 | 2520531 | 113 | 6273526 | *122 | 4485358 | 112 | 7008074 | 129 | 5850527 | 112 | 7156057 |
| 133.5626 | 1840 | 2520531 | 113 | 6273526 | *121 | 4494277 | 112 | 7008074 | 129 | 5850527 | 112 | 7156057 |
| 150 | 1821 | 2922896 | 104 | 6424664 | 113 | 4645729 | 105 | 7127865 | 107 | 6282099 | 107 | 7235767 |
| 300 | 1760 | 4492689 | 75 | 7309436 | 75 | 5782973 | 75 | 8025779 | 75 | 7218504 | 75 | 8173287 |
| 600 | 1760 | 4492689 | 69 | 7928940 | 69 | 6403648 | 69 | 8645041 | 69 | 7838170 | 69 | 8792646 |
| 1000 | *1760 | 4492689 | 66 | 9895484 | 66 | 8371629 | 66 | 10611171 | 66 | 9804570 | 66 | 10758552 |

[1] Global Compression Threshold Model complexity is measured by the number of coefficients needed to approximate the image at a specific GCT. The accuracy metric used is SSE and it is calculated from the 256-color scaled data. Improvements in model complexity are observed for certain threshold of accuracy.

that could limit the application to more complex data structures. Multiple time series information are widely generated in many applications such as biology, medicine, finance, and others [Bar-Joseph, 2004, Ramsay and Silverman, 2002] posing a need to new methods and analysis procedures. Interested in analyzing more complex curves structures, we investigated the effectiveness of approximating time series through a novel two-dimensional representation. This imaging representation changes with different ordering scenarios. Hence, we also explored the integration of external one-dimensional variables used as sorting criteria to the curves. This sorting mechanism can serve as a variable ranking procedure when external information about the time series is available.

The unique 2D representation presented here showcased an improved performance in terms of model complexity when compared to its primitive M1D decomposition. However, certain regions were found where M1D performed better in terms accuracy but with equal complexity for datasets portraying higher entropy. Another important result from this work is the variable ranking procedure. When external information on the time series is accessible, different ordering scenarios on the imaging representation can be evaluated and its performance can lead us to detecting relevant variables to describe the curves response. For the Corn dataset, ranking its time series by the one-dimensional variable, moisure, resulted in higher reduction in the approximation complexity tested through wavelet decomposition. This study shows interesting results and raised some interesting research questions for follow up such as measurement of randomness and ordering as an effect to extract and approximate better original NIR signals.

## 6.6. Acknowledgments

I would like to thank the co-author of this research work, Dr. George C. Runger. The authors would like to thank the School of Computing, Informatics, and Decisions Systems Engineering at Arizona State University.

# CHAPTER 7

# CONCLUSION

The last decade has seen significant growth in technologies pertaining to molecular biological assays, genomics, proteomics, microarrays to measure gene expression profile, and computational and statistical tools of data analysis. The knowledge gained in these areas need to be integrated effectively to derive efficient translational bioinformatics methodologies, which will form the foundation for building specific bedside tools for treatment. As evident from recent literature, this moves towards understanding underlying biological mechanisms that affects disease progression and therapy success. There still unclear cell pathways for many diseases and understanding of these mechanisms is the key to unravel information needed to close the gaps in current methods for early diagnosis and early intervention. However this huge task will need significant research attention and investment before useable tools can be made available for medical decision making.

Once the whole human genome was sequenced and published in 2003 [Lande et al., 2001,International Human Genome Sequencing Consortium, 2004], unprecedented advances in genomics have made genome research an important area in biomedical research and with this many more undiscovered challenges encountered to foster more research opportunities towards understanding of diseases. Also, the National Human Genome Research Institute (NHGRI) has stated three cross-cutting areas that are fundamental in the progress of genomic research: bioinformatics and computational biology, education and training, and genomics and society [Green and Guyer, 2011]. Some of the challenges stated for the bioinformatics and computational biology research community includes the need of novel methodologies for data analysis and integration of heterogeneous data types. Methodologies closely coupled with data production and designed for robust applications for diverse set of researchers and scientists. To meet the computational challenges of mature and newly acquired genomics data, a new breed of scientists should be trained with expertise and proficiency in two or more of the following fields: biology, informatics, computer science, mathematics, statistics and/or engineering [Green and Guyer, 2011]. My goal in this research is to contribute to this endeavor.

This motivates the research presented here which proposes several statistical and data mining tools to address diverse challenges encountered through the integration of large and heterogeneous data sources within many scenarios at the cellular and molecular level. Based on the interdisciplinary nature of the research team, this work reports several original contributions to diverse scientific fields such as statistical data analysis, genomics, and bioinformatics. Some general contributions of this research are:

- Successful integration of multiple heterogeneous data sources
- Application of several non-linear methods such as ensemble methods and spline models.
- Development of a validation procedure to evaluate predictability performance of a model using external knowledge sources.
- Interesting biological relationships were found.
- Establishing a statistical framework containing a series of tools towards the characterization of unique data source (oxygen consumption curves from single-cell technology).
- Extraction and modeling of a large number of features towards biological inferences.
- Pursue a unique way to represent multiple time series through imaging and evaluate its performance in terms of model complexity and accuracy using wavelets.

The diversity component of the research team in this dissertation allowed investigation of problems in need of cross-cutting research focus. The scope of expertise among the research team advanced growth in different fields and contributions in both the areas of data analysis and biology/genomics, by targeting problem-based questions.

# CHAPTER 8

# FUTURE RESEARCH

Despite the increasing molecular knowledge and the technological advances to gather data from biological processes, there remains opportunities for new computational methods in data analysis to achieve suitable biological interpretation. Throughout my participation in the Planning the Future of Genomics conference by the National Human Genome Research Institute (NHGRI) on July $6^{th}$-$8^{th}$, $2010$, it was clear the need for more initiatives in cross-cutting research. As a result of this meeting, NGHRI published their vision for the future of genomic research and elaborated on a path towards genomic medicine [Green and Guyer, 2011]. There is a special mention on some interesting areas and opportunities for researchers with interests in bioinformatics and computational biology such as data integration which has been our primary contribution with this work. Robust methods and techniques are necessary to meet challenges within: data analysis, data integration, visualization, computational tools and infrastructure, and training in genomics [Green and Guyer, 2011].

Some comprehensive databases publicly available [Green and Guyer, 2011] that are venues for cross-cutting research within main interests with data integration and data analysis (additional information in Appendix A).

## 8.1. Systems biology perspective on transcriptomics and proteomics

Chapters 3 and 4 reveals strong global relationships between transcriptomic and proteomic data through the implementation of an nonlinear model and the integration of annotated functional information. However, specific relationships were not captured. It will be very interesting to study these data types with additional experimental information on physiological states of the cell. Other possible extensions are: (1) complexity exploration on gene-protein relationship model, (2) behavior studies on "noisy" genes, and (3) time-lag models between gene expression and protein abundance.

1. Exploring models that includes more complex interactions among gene expression and peptide abundance besides the one-to-one dependency will be an interesting extension, especially for data from human cell lines. This task certainly raises other challenges such as a lack of external knowledge for model validation. Nevertheless, it is unquestionably an appealing problem to investigate and Appendix C offers brief ideas on how to implement this.

2. In this work, we focused on the global relationship of gene and peptide expression for two different bacterial species, for this objective quality con-

trol steps were performed where "noisy" genes were removed from the study. This allowed for a more robust characterization of the global correlation of gene expression and protein abundance. Nonetheless, analyzing the behavior of "noisy" genes could capture relevant signatures to understand particular phenotypes. Focusing in a smaller set of genes (through feature selection) will help reveal discoveries for particular mechanisms in the cell which has been studied in the literature. However, making robust statistical inferences from very noisy information will be an difficult task.

3. Through the analysis of temporal data, the study reflected some insights into time lagging between gene expression and protein abundance which is a known characteristic of the cell regulation process. Developing a more detailed model to understand time lags between these data sources could provide insights into post-transcriptomic changes.

## 8.2. Sensor-based measurements

The work presented here establishes a statistical framework as an initial benchmark on how to extract knowledge from multiple oxygen consumption curves (i.e., sensor-based measurements) obtained from single-cell technology within Barrett's esophageal epithelial cell lines. Many extensions to the attempted solution strategies addressing unique data challenges can be pursued. As previously mentioned in Figure 14, the challenges addressed in this work are: (1) random noise, (2) understanding multiple states, and (3) understanding cell behavior.

1. The smoothing step in this work has without a doubt alleviated the random noise problem. However, improvements can be pursue to better model the non-constant variation of the signal-to-noise ratio.

2. The core of our single-cell work focused on characterizing OC curves by models with simple structures and extracting features related to those characterizations over multiple series of data. Other models were explored for stochastic characterization of the OC states such as Markov Processes. Challenges like non-stationarity assumptions make troublesome the implementation of some methods to the studied OC curves. This requires a more complex Markovian model that voids stationarity and parametric estimators.

3. An extensive set of features were extracted from the spline model allowing for comparisons among different cell types and experimental conditions. Feature extraction is certainly an area of further exploration and improve-

ment besides the substantial set of features identified. Other appropriate variables related to hypoxia and the Barrett's premalignant cell lines are in need of integration for better understanding of this cellular system.

- Dynamic time warping models [Oates et al., 1999, Oates et al., 2001, Berndt and Clifford, 1994, Aach and Church, 2001] have been used to define similarities between time series and these methods can be applied to cluster oxygen consumption curves.
- Integrating other sources of information such as gene expression at the single-cell level could reveal interesting patterns to help elucidate cell heterogeneity mechanisms. The experimental protocol for this type of data has been optimized and is currently explored within the MLSC headquartered at Arizona State University [Yongzhong Li and Meldrum, 2010, Gao et al., 2011]. Gene expression data under same experimental conditions as to oxygen consumption will be coming available for a subset of genes chosen to have known functional relationships with hypoxia activities in the cell. It's relationships with oxygen consumption is of great interest and its integration is thought to provide clearer understanding on cell behavior and help with future disease diagnostics for Barrett's esophagus.

- As the MLSC headquartered at Arizona State University continues their novel efforts to develop multiparameter technologies, other features such as temperature, pH and $K^+$ will become available for data analysis and integration. The amounts of heterogeneous data sources for the particular system studied here allows for natural extensions to this work in the area of feature extraction and integration.

Many venues of exciting projects towards data analysis are available for future exploration in the area of genomics and bioinformatics. This work serves as a step forward for more collaborative research among researchers from diverse disciplines (i.e., engineering, statistics, computer science, genomics, biology, material sciences, physics, etc.).

# REFERENCES

[Aach and Church, 2001] Aach, J. and Church, G. M. (2001). Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508.

[Aebersold and Mann, 2003] Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422:198–207.

[Aittokallio, 2010] Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief Bioinform*, 11(2):253–264.

[Albrecht et al., 2010] Albrecht, D., Kniemeyer, O., Brakhage, A. A., and Guthke, R. (2010). Missing values in gel-based proteomics. *Proteomics*, 10(6):1202–11.

[Alfassi et al., 2005] Alfassi, Z. B., Boger, Z., and Ronen, Y. (2005). *Statistical Treatment of Analytical Data*. CRC Press: Blackwell Science, Boca Raton, FL.

[Alm et al., 2005] Alm, E. J., Huang, K. H., Price, M. N., Koche, R. P., Keller, K., Dubchak, I. L., and Arkin, A. P. (2005). The microbesonline web site for comparative genomics. *Genome research*, 15(7):1015–1022.

[Alter and Golub, 2004] Alter, O. and Golub, G. (2004). Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between dna replication and rna transcription. *Proc. Natl. Acad. Sci. USA*, 101:16577–16582.

[Altschuler and Wu, 2010] Altschuler, S. J. and Wu, L. F. (2010). Cellular heterogeneity: Do differences make a difference? *Cell*, 141(4):559 – 563.

[Anderle et al., 2003] Anderle, P., Duval, M., Draghici, S., Kuklin, A., Littlejohn, T. G., Medrano, J. F., Vilanova, D., and Roberts, M. A. (2003). Gene expression databases and data mining. *Biotechniques*, pages 36–44.

[Andersen-Nissen et al., 2005] Andersen-Nissen, E., Smith, K. D., Strobe, K. L., Barrett, S. L. R., Cookson, B. T., Logan, S. M., and Aderem, A. (2005). Evasion of Toll-like receptor 5 by flagellated bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 102(26):9247–9252.

[Aniruddh et al., 2004] Aniruddh, S., Shastry, A., and Lal, R. (2004). A Selective Single Cell Electroporator with a Microfabricated Sense-Porate Aperture. In *IEEE Conference on Micro Electro Mechanical Systems (MEMS)*, Maastricht, Holland.

[Anis et al., 2010] Anis, Y., Holl, M., and Meldrum, D. (2010). Automated selection and placement of single cells using vision-based feedback control. *Automation Science and Engineering, IEEE Transactions on*, 7(3):598 –606.

[Arizona State University, 2011] Arizona State University (2011). Center for biosignatures discovery automation overview. `http://www.biodesign.asu.edu/`.

[Arriaga, 2009] Arriaga, E. A. (2009). Determining biological noise via single cell analysis. *Analytical and Bioanalytical Chemistry*, 393(1):73–80.

[Ashili et al., 2011] Ashili, S. P., Kelbauskas, L., Houkal, J., Smith, D., Tain, Y., Youngbull, C., Zhu, H., Anis, Y. H., Hupp, M., Lee, K. B., Kumar, A. V., Vela, J., Shabilla, A., Johnson, R. H., Holl, M. R., and Meldrum, D. R. (2011). Automated platform for multiparameter stimulus response studies of metabolic activity at the single-cell level. In Becker, H. and Gray, B. L., editors, *SPIE Proceedings: Microfluidics, BioMEMS, and Medical Microsystems IX*, volume 7929.

[Babic et al., 2005] Babic, N., Beeson, C., and Dovichi, N. J. (2005). Effect of High Density Lipoproteins on Protein Expression in Myoblast Cell Lines. *Journal of Proteome Research*, 4(2):344–348.

[Bagshaw and Johnson, 1975] Bagshaw, M. and Johnson, R. A. (1975). The effect of serial correlation on the performance of cusum tests ii. *Technometrics*, 17(1):73–80.

[Bai, 1995] Bai, J. (1995). Least absolute deviation estimation of a shift. *Econometric Theory*, 11(3):403–436.

[Bai, 1997] Bai, J. (1997). Estimation of a change point in multiple regression models. *The Review of Economics and Statistics*, 79(4):551–563.

[Bar-Joseph, 2004] Bar-Joseph, Z. (2004). Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–2503.

[Barrett et al., 1999] Barrett, M. T., Sanchez, C. A., Prevo, L. J., Wong, D. J., Galipeau, P. C., Paulson, T. G., Rabinovitch, P. S., and Reid, B. J. (1999). Evolution of neoplastic cell lineages in barrett oesophagus. *Nature Genetics*, 22:106–109.

[Beck-Jr. and Knecht, 2003] Beck-Jr., G. R. and Knecht, N. (2003). Osteopontin regulation by inorganic phosphate is erk1/2-, protein kinase c-, and proteasome-dependent. *The Journal of biological chemistry*, 278(43):41921–41929.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 289–300.

[Benjamini and Yekutieli, 2001] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):pp. 1165–1188.

[Berk, 2008] Berk, R. A. (2008). *Statistical Learning From a Regression Perspective*. New York: Springer.

[Berndt and Clifford, 1994] Berndt, D. J. and Clifford, J. (1994). Using Dynamic Time Warping to Find Patterns in Time Series. In *Proceedings of KDD-94: AAAI Workshop on Knowledge Discovery in Databases*, pages 359–370, Seattle, Washington.

[Bertone and Gerstein, 2001] Bertone, P. and Gerstein, M. (2001). Integrative data mining: the new direction in bioinformatics. *Engineering in Medicine and Biology Magazine, IEEE*, 20(4):33 –40.

[Beyer et al., 2004] Beyer, A., Hollunder, J., Nasheuer, H., and Wilhelm, T. (2004). Posttranscriptional expression regulation in the yeast saccharomyces cerevisiae on a genomic scale. *Mol Cell Proteomics*, 3(11):1083–1092.

[Bilban et al., 2002] Bilban, M., Buehler, L., Head, S., Desoye, G., and Quaranta, V. (2002). Normalizing dna microarray data. *Curr Issues Mol Biol*, 4(2):57–64.

[Bild and Febbo, 2005] Bild, A. and Febbo, P. G. (2005). Application of a priori established gene sets to discover biologically important differential expression in microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15278–15279.

[Boeckmann et al., 2003] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Research*, 31(1):365–370.

[Boros et al., 2000] Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E., and Muchnik, I. (2000). An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering*, 12:292–306.

[Bovik, 2005] Bovik, A. (2005). *Handbook of image and video processing*. Communications, Networking and Multimedia. Elsevier Academic Press.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

[Brocker et al., 2002] Brocker, J., Parlitz, U., and Ogorzalek, M. (2002). Nonlinear noise reduction. *Proceedings of the IEEE*, 90(5):898 –918.

[Brown et al., 2006] Brown, S. D., Thompson, M. R., VerBerkmoes, N. C., Chourey, K., Shah, M., Zhou, J., Hettich, R. L., and Thompson, D. K. (2006). Molecular Dynamics of the Shewanella oneidensis Response to Chromate Stress. *Molecular and Cellular Proteomics*, 5(6):1054–1071.

[Bussink et al., 2000] Bussink, J., Kaanders, J. H. A. M., Strik, A. M., Vojnovic, B., and Kogel, A. J. v. d. (2000). Optical sensor-based oxygen tension measurements correspond with hypoxia marker binding in three human tumor xenograft lines. *Radiation Research*, 154(5):pp. 547–555.

[Chan and Mahoney, 2005] Chan, P. K. and Mahoney, M. V. (2005). Modeling multiple time series for anomaly detection. In *Fifth IEEE International Conference on Data Mining*, pages 90–97. IEEE Computer Society.

[Chandola et al., 2009] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41:15:1–15:58.

[Chao and Ros, 2008] Chao, T.-C. and Ros, A. (2008). Microfluidic single-cell analysis of intracellular compounds. *Journal of The Royal Society Interface*, 5(Suppl 2):S139–S150.

[Charaniya et al., 2007] Charaniya, S., Mehra, S., Lian, W., Jayapal, K. P., Karypis, G., and Hu, W.-S. (2007). Transcriptome dynamics-based operon prediction and verification in streptomyces coelicolor. *Nucleic Acids Research*, 35(21):7222–7236.

[Chen et al., 2008] Chen, C.-h., Hardle, W., Unwin, A., Cox, M., and Cox, T. F. (2008). Handbook of data visualization. In *Springer Handbooks Comp. Statistics*, chapter Multidimensional Scaling, pages 315–347. Springer Berlin Heidelberg.

[Chen and Gupta, 2001] Chen, J. and Gupta, A. (2001). On change point detection and estimation. *Comm. Statistics-Simulation and Computation*, pages 665–697.

[Chen and Sivachenko, 2005] Chen, J. and Sivachenko, A. (2005). Data mining in protein interactomics. *Engineering in Medicine and Biology Magazine, IEEE*, 24(3):95 –102.

[Chiu et al., 2003] Chiu, B., Keogh, E., and Lonardi, S. (2003). Probabilistic discovery of time series motifs. In *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining*, KDD '03, pages 493–498, New York, NY, USA. ACM.

[Cohen, 2007] Cohen, J. (2007). Tr10: Single-cell analysis. *Technical Review Published by MIT*.

[Conrads et al., 2005] Conrads, K. A., Yi, M., Simpson, K. A., Lucas, D. A., Camalier, C. E., Yu, L. R., Veenstra, T. D., Stephens, R. M., Conrads, T. P., and Jr, G. R. B. (2005). A combined proteome and microarray investigation of inorganic phosphate-induced pre-osteoblast cells. *Molecular and cellular proteomics : MCP*, 4(9):1284–1296.

[Cox and Cox, 1994] Cox, T. F. and Cox, M. A. (1994). *Multidimensional scaling*. London: Chapman and Hall.

[d. S. Abreu et al., 2009] d. S. Abreu, R., Penalva, L. O., Marcotte, E. M., and Vogel, C. (2009). Global signatures of protein and mrna expression levels. *Mol. BioSyst.*, 5:1512–1526.

[Daemen et al., 2009] Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J. A., Sempoux, C., pascal Machiels, J., Haustermans, K., and Moor, B. D. (2009). A kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine*, (4):39.

[Daubechies, 1992] Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[Davis, 2002] Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. Springer Verlag New York, Inc.

[De'ath, 2007] De'ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology*, 88(1):243–251.

[Dehal et al., 2010] Dehal, P. S., Joachimiak, M. P., Price, M. N., Bates, J. T., Baumohl, J. K., Chivian, D., Friedland, G. D., Huang, K. H., Keller, K., Novichkov, P. S., Dubchak, I. L., Alm, E. J., and Arkin, A. P. (2010). MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucl. Acids Res.*, 38:D396–400.

[Do and Choi, 2006] Do, J. and Choi, D. (2006). Normalization of microarray data: Single-labeled and dual-labeled arrays. *Mol Cells*, 22(3):254–61.

[Doukhan et al., 2003] Doukhan, P., Oppenheim, G., and Taqqu, M. S. (2003). *Theory and applications of long-range dependence*. Birkhauser.

[D′Agostino, 2004] D′Agostino, R. B. (2004). *Tutorials in Biostatistics: Statistical modelling of complex medical data*. Wiley Inc.

[D′haeseleer et al., 1999] D′haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. (1999). Linear modeling of mrna expression levels during cns development and injury. *Pac Symp Biocomput*, pages 41–52.

[Edgar et al., 2002] Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

[Eigenvector Research Inc., 2009] Eigenvector Research Inc. (2009). Data sets available to download. http://software.eigenvector.com/Data/index.html.

[Elith et al., 2008] Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *The Journal of animal ecology*, 77(4):802–813.

[Ernst et al., 2005] Ernst, J., Nau, G., and Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, 21 Suppl 1.

[Esteban et al., 2005] Esteban, J., Starr, A., Willetts, R., Hannah, P., and Bryanston-Cross, P. (2005). A review of data fusion models and architectures: towards engineering guidelines. *Neural Comput. Appl.*, 14(4):273–281.

[Esterby and El-Shaarawi, 1981] Esterby, S. R. and El-Shaarawi, A. H. (1981). Inference about the point of change in a regression model. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 30(3):277–285.

[Fiehn, 2001] Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp Funct Genomics*, 2(3):155168.

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. In *International Conference on Machine Learning*, pages 148–156.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

[Friedman, 2002] Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378.

[Gan et al., 2006] Gan, X., Liew, A. W.-C., and Yan, H. (2006). Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucl. Acids Res.*, 34(5):1608–1619.

[Ganesan and Venkataraman, 2004] Ganesan, R., D.-T. K. and Venkataraman, V. (2004). Wavelet-based multiscale statistical process monitoring: A literature review. *IIE Transactions*, 36(9):787–806.

[Gao et al., 2011] Gao, W., Zhang, W., and Meldrum, D. R. (2011). Rt-qpcr based quantitative analysis of gene expression in single bacterial cells. *Journal of Microbiological Methods*, In Press, Uncorrected Proof.

[Gerstein et al., 2010] Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dos, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D. M., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Rtsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., modENCODE Consortium, Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., and Waterston, R. H. (2010). Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, 330(6012):1775–1787.

[Giraitis and Leipus, 1990] Giraitis, L. and Leipus, R. (1990). Functional clt for nonparametric estimates of the spectrum and change-point problem for a spectral function. *Lithuanian Mathematical Journal*, (4):302–322.

[Giraitis and Leipus, 1992] Giraitis, L. and Leipus, R. (1992). Testing and estimating in the change-point problem of the spectral function. *Lithuanian Mathematical Journal*, (1):15–29.

[Golosnoy et al., 2009] Golosnoy, V., Ragulin, S., and Schmid, W. (2009). Multivariate cusum chart: properties and enhancements. *AStA Advances in Statistical Analysis*, 93(3):263–279.

[Gorry, 1990] Gorry, P. A. (1990). General least-squares smoothing and differentiation by the convolution (savitzky-golay) method. *Analytical Chemistry*, 62(6):570–573.

[Green and Guyer, 2011] Green, E. D. and Guyer, M. S. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333):204–213.

[Greenbaum et al., 2002] Greenbaum, D., Jansen, R., and Gerstein, M. (2002). Analysis of mrna expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics (Oxford, England)*, 18(4):585–596.

[Gygi et al., 1999] Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mrna abundance in yeast. *Molecular and cellular biology*, 19(3):1720–1730.

[Haar, 1910] Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69:331–371.

[Hannan, 1970] Hannan, E. J. (1970). *Multiple Time Series*. John Wiley and Sons, Inc.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning - Data Mining, Inference, Prediction.* Springer Verlag.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning - Data Mining, Inference, Prediction.* Springer Verlag, 2nd. edition.

[Hegde et al., 2003] Hegde, P., White, I., and Debouck, C. (2003). Interplay of transcriptomics and proteomics. *Current opinion in biotechnology*, 14:647–651.

[Heidelberg et al., 2004] Heidelberg, J. F., Seshadri, R., Haveman, S. A., Hemme, C. L., Paulsen, I. T., Kolonay, J. F., Eisen, J. A., Ward, N., Methe, B., Brinkac, L. M., Daugherty, S. C., Deboy, R. T., Dodson, R. J., Durkin, A. S., Madupu, R., Nelson, W. C., Sullivan, S. A., Fouts, D., Haft, D. H., Selengut, J., Peterson, J. D., Davidsen, T. M., Zafar, N., Zhou, L., Radune, D., Dimitrov, G., Hance, M., Tran, K., Khouri, H., Gill, J., Utterback, T. R., Feldblyum, T. V., Wall, J. D., Voordouw, G., and Fraser, C. M. (2004). The genome sequence of the anaerobic, sulfate-reducing bacterium desulfovibrio vulgaris hildenborough. *Nature biotechnology*, 22(5):554–559.

[Hemme and Wall, 2004] Hemme, C. L. and Wall, J. D. (2004). Genomic insights into gene regulation of desulfovibrio vulgaris hildenborough. *Omics : a journal of integrative biology*, 8(1):43–55.

[Hermeking, 2003] Hermeking, H. (2003). Serial analysis of gene expression and cancer. *Current opinion in oncology*, 15(1):44–49.

[Hidalgo and Robinson, 1996] Hidalgo, J. and Robinson, P. (1996). Testing for structural change in a long-memory environment. *Journal of Econometrics*, (1):159–174.

[Hoffman et al., 2004] Hoffman, E. P., Awad, T., Palma, J., Webster, T., Hubbell, E., A.Warrington, J., Spira, A., Wright, G., Buckley, J., Tricheare, T., Davis, R., Tibshirani, R., Xiao, W., Jones, W., Tompkins, R., and West, M. (2004). Expression profilingbest practices for data generation and interpretation in clinical trials. *Nature Reviews Genetics*, (5):229–237.

[Horak and Snyder, 2002] Horak, C. and Snyder, M. (2002). Global analysis of gene expression in yeast. *Funct Integr Genomics*, 2(4-5):171–180.

[Hou and Andrews, 1978] Hou, H. H. and Andrews, H. C. (1978). Cubic splines for image interpolation and digital filtering. *IEEE Trans. Acoust., Speech, Signal Process*, (26):508–517.

[Hu et al., 2006] Hu, J., Li, H., Waterman, M., and Zhou, X. (2006). Integrative missing value estimation for microarray data. *BMC Bioinformatics*, 7(1):449.

[Hunter, 1996] Hunter, J. (1996). The exponentially weighted moving average. *J. Quality Technol.*, 18(4):203–210.

[Ideker et al., 2001] Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (New York, N.Y.)*, 292(5518):929–934.

[International Human Genome Sequencing Consortium, 2004] International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945.

[Irizarry et al., 2006] Irizarry, R., Wu, Z., and Jaffee, H. (2006). Comparison of affymetrix genechip expression measures. *Bioinformatics*.

[Ishii and Tomita, 2009] Ishii, N. and Tomita, M. (2009). *Systems Biology and Biotechnology of Escherichia coli*. Springer Netherlands, 1 edition.

[Johnson, 2001] Johnson, R. A. (2001). *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey.

[Johnson, 2005] Johnson, R. A. (2005). *Miller and Freund′s Probability and Statistics for Engineers*. Pearson Prentice Hall, New Jersey.

[Johnson and Bagshaw, 1974] Johnson, R. A. and Bagshaw, M. (1974). The effect of serial correlation on the performance of cusum tests. *Technometrics*, 16(1):103–112.

[Jolliffe, 2002] Jolliffe, I. (2002). *Principal component analysis*. Springer series in statistics. Springer.

[Joyce and Palsson, 2006] Joyce, A. R. and Palsson, B. O. (2006). The model organism as a system: integrating ’omics’ data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210.

[Jung et al., 2005] Jung, K., Gannoun, A., Sitek, B., Meyer, H., Stuhler, K., and Urfer, W. (2005). Analysis of dynamic protein expression data. *REVSTAT Statistical Journal*, (2):99–111.

[Kelbauskas et al., 2011] Kelbauskas, L., Ashili, S., Houkal, J., Smith, D., Mohammadreza, A., Lee, K., Kumar, A., Anis, Y., Paulson, T., Youngbull, C., Tian, Y., Johnson, R., Holl, M., and Meldrum, D. (2011). A novel method for multiparameter physiological phenotype characterization at the single-cell level. In Farkas, D. L., Nicolau, D. V., and Leif, R. C., editors, *SPIE Proceedings: Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues IX*, volume 7902.

[Keogh et al., 2006] Keogh, E. J., Lin, J., Fu, A. W.-C., and Herle, H. V. (2006). Finding unusual medical time-series subsequences: Algorithms and applications. *IEEE Transactions on Information Technology in Biomedicine*, pages 429–439.

[Kim et al., 2005] Kim, H., Golub, G. H., and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198.

[Kleinbaum et al., 2008] Kleinbaum, D. G., Kupper, L. L., and Muller, K. E., editors (2008). *Applied regression analysis and other multivariable methods*. PWS Publishing Co., Boston, MA, USA, 4th edition.

[Ko and Lee, 2009] Ko, S. and Lee, H. (2009). Integrative approaches to the prediction of protein functions based on the feature selection. *BMC Bioinformatics*, 10(1):455.

[Kong et al., 2006] Kong, S. W., Pu, W. T., and Park, P. J. (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, 22(19):2373–2380.

[Kruskal and Wallis, 1952] Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):pp. 583–621.

[Lande et al., 2001] Lande, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., and et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

[Laurell et al., 2001] Laurell, T., Nilsson, J., and Marko-Varga, G. (2001). Proteomics-protein profiling technology: the trend towards a microfabricated toolbox concept. *TrAC Trends in Analytical Chemistry*, 20(5):225 – 231.

[Leach et al., 1984] Leach, R. A., Carter, C. A., and Harris, J. M. (1984). Least-squares polynomial filters for initial point and slope estimation. *Analytical Chemistry*, 56(13):2304–2307.

[Li et al., 2006] Li, J., Li, X., Su, H., Chen, H., and Galbraith, D. W. (2006). A framework of integrating gene relations from heterogeneous data sources: an experiment on Arabidopsis thaliana. *Bioinformatics*, 22(16):2037–2043.

[Lidstrom and Meldrum, 2003] Lidstrom, M. and Meldrum, D. (2003). Life-on-a-chip. *Nature Reviews Microbiology*, pages 158–164.

[Lin and Qian, 2007] Lin, J. and Qian, J. (2007). Systems biology approach to integrative comparative genomics. *Expert Rev. Proteomics*, (1):107–119.

[Little and Rubin, 1987] Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data.* John Wiley and Sons, Inc., New York.

[Liu et al., 2005] Liu, H., Dougherty, E. R., Dy, J. G., Torkkola, K., Tuv, E., Peng, H., Ding, C., Long, F., Berens, M., Parsons, L., Zhao, Z., Yu, L., and Forman, G. (2005). Evolving feature selection. *IEEE Intelligent Systems*, 20(6):64–76.

[Lutkepohl, 2005] Lutkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, 1 edition.

[Maier et al., 2009] Maier, T., Gell, M., and Serrano, L. (2009). *Correlation of mRNA and protein in complex biological samples*, volume 583.

[Mallat, 1999] Mallat, S. (1999). *A Wavelet Tour of Signal Processing, Second Edition (Wavelet Analysis & Its Applications)*. Academic Press.

[Maxwell and Delaney, 2003] Maxwell, S. E. and Delaney, H. D. (2003). *Designing Experiments and Analyzing Data: A Model Comparison Perspective, Second Edition*. Lawrence Erlbaum.

[McNeil and Manning, 2002] McNeil, C. J. and Manning, P. (2002). Sensor-based measurements of the role and interactions of free radicals in cellular systems. *Reviews in Molecular Biotechnology*, 82(4):443 – 455.

[modENCODE Consortium et al., 2010] modENCODE Consortium, T., Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., Washietl, S., Arshinoff, B. I., Ay, F., Meyer, P. E., Robine, N., Washington, N. L., Di Stefano, L., Berezikov, E., Brown, C. D., Candeias, R., Carlson, J. W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M. Y., Will, S., Alekseyenko, A. A., Artieri, C., Booth, B. W., Brooks, A. N., Dai, Q., Davis, C. A., Duff, M. O., Feng, X., Gorchakov, A. A., Gu, T., Henikoff, J. G., Kapranov, P., Li, R., MacAlpine, H. K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S. K., Riddle, N. C., Sakai, A., Samsonova, A., Sandler, J. E., Schwartz, Y. B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K. H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S. E., Brent, M. R., Cherbas, L., Elgin, S. C. R., Gingeras, T. R., Grossman, R., Hoskins, R. A., Kaufman, T. C., Kent, W., Kuroda, M. I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J. W., Ren, B., Russell, S., Cherbas, P., Graveley, B. R., Lewis, S., Micklem, G., Oliver, B., Park, P. J., Celniker, S. E., Henikoff, S., Karpen, G. H., Lai, E. C., MacAlpine, D. M., Stein, L. D., White, K. P., and Kellis, M. (2010). Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797.

[Molter et al., 2008] Molter, T., Holl, M., Dragavon, J., McQuaide, S., Anderson, J., Young, A., Burgess, L., Lidstrom, M., and Meldrum, D. (2008). A new approach for measuring single-cell oxygen consumption rates. *Automation Science and Engineering, IEEE Transactions on*, 5(1):32–42.

[Molter et al., 2007] Molter, T., McQuaide, S. C., Zhang, M., Holl, M. R., Burgess, L. W., Lidstrom, M. E., and Meldrum, D. R. (2007). Algorithm advancements for the measurement of single cell oxygen consumption rates. In *IEEE International Conference CASE 2007*, pages 386–391. Automation Science and Engineering.

[Molter et al., 2009] Molter, T. W., McQuaide, S. C., Suchorolski, M. T., Strovas, T. J., Burgess, L. W., Meldrum, D. R., and Lidstrom, M. E. (2009). A microwell array device capable of measuring single-cell oxygen consumption rates. *Sensors and Actuators B: Chemical*, 135(2):678 – 686.

[Montgomery, 2001] Montgomery, D. (2001). *Introduction to Statistical Quality Control.* John Wiley and Sons, Inc., 4 edition.

[Montgomery, 2005] Montgomery, D. (2005). *Introduction to Statistical Quality Control*. John Wiley and Sons, Inc., 5 edition.

[Montgomery and Runger, 2006] Montgomery, D. and Runger, G. C. (2006). *Applied Statistics and Probability for Engineers*. John Wiley and Sons, Inc., 4 edition.

[Mootha et al., 2003a] 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., Patterson, N., Lander, E. S., and Mann, M. (2003a). Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell*, 115(5):629–640.

[Mootha et al., 2003b] Mootha, V. K., Lepage, P., Miller, K., Bunkenborg, J., Reich, M., Hjerrild, M., Delmonte, T., Villeneuve, A., Sladek, R., Xu, F., Mitchell, G. A., Morin, C., Mann, M., Hudson, T. J., Robinson, B., Rioux, J. D., and Lander, E. S. (2003b). Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proceedings of the National Academy of Sciences of the United States of America*, 100(2):605–610.

[Mukhopadhyay et al., 2006] Mukhopadhyay, A., He, Z., Alm, E. J., Arkin, A. P., Baidoo, E. E., Borglin, S. C., Chen, W., Hazen, T. C., He, Q., Holman, H. Y., Huang, K., Huang, R., Joyner, D. C., Katz, N., Keller, M., Oeller, P., Redding, A., Sun, J., Wall, J., Wei, J., Yang, Z., Yen, H. C., Zhou, J., and Keasling, J. D. (2006). Salt stress in desulfovibrio vulgaris hildenborough: an integrated genomics approach. *Journal of Bacteriology*, 188(11):4068–4078.

[NCBI, 2010] NCBI (2010). Microarrays: chipping away at the mysteries of science and medicine. `http://www.ncbi.nlm.nih.gov/`.

[Nie et al., 2006a] Nie, L., Wu, G., Brockman, F. J., and Zhang, W. (2006a). Integrated analysis of transcriptomic and proteomic data of desulfovibrio vulgaris: zero-inflated poisson regression models to predict abundance of undetected proteins. *Bioinformatics (Oxford, England)*, 22(13):1641–1647.

[Nie et al., 2007] Nie, L., Wu, G., Culley, D. E., Scholten, J. C., and Zhang, W. (2007). Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Critical reviews in biotechnology*, 27(2):63–75.

[Nie et al., 2006b] Nie, L., Wu, G., and Zhang, W. (2006b). 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.

[Nie et al., 2006c] Nie, L., Wu, G., and Zhang, W. (2006c). Correlation of mrna expression and protein abundance affected by multiple sequence features related to translational efficiency in desulfovibrio vulgaris: a quantitative analysis. *Genetics*, 174(4):2229–2243.

[Ninomiya and Yoshimoto, 2008] Ninomiya, Y. and Yoshimoto, A. (2008). Statistical method for detecting structural change in the growth process. *Biometrics*, 64(1):46–53.

[Nuwaysir et al., 2002] Nuwaysir, E. F., Huang, W., Albert, T. J., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorski, T., Berg, J. P., Ballin, J., McCormick, M., Norton, J., Pollock, T., Sumwalt, T., Butcher, L., Porter, D., Molla, M., Hall, C., Blattner, F., Sussman, M. R., Wallace, R. L., Cerrina, F., and Green, R. D. (2002). Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome research*, 12(11):1749–1755.

[Oates et al., 2001] Oates, T., Firoiu, L., and Cohen, P. (2001). Using dynamic time warping to bootstrap hmm-based clustering of time series. In Sun, R. and Giles, C., editors, *Sequence Learning*, volume 1828 of *Lecture Notes in Computer Science*, pages 35–52. Springer Berlin and Heidelberg.

[Oates et al., 1999] Oates, T., Firoiu, L., and Cohen, P. R. (1999). Clustering time series with hidden markov models and dynamic time warping. In *In Proceedings of the IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*, pages 17–21.

[Ogata et al., 1999] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27(1):29–34.

[Orfanidis, 1996] Orfanidis, S. (1996). *Introduction to Signal Processing*. Prentice-Hall.

[Otto, 2011] Otto, A. M. (2011). Cell cultivation and sensor-based assays for dynamic measurements of cell vitality. In Choi, S., Boob-Bavnbek, B., Klosgen, B., Larsen, J., Pociot, F., and Renstrom, E., editors, *BetaSys*, volume 2 of *Systems Biology*, pages 221–240. Springer New York.

[Pan and Fang, 2002] Pan, J. X. and Fang, K. T. (2002). *Growth Curve Models and Statistical Diagnostics*. Springer Series in Statistics.

[Park et al., 2005] Park, S., Lee, S., Cho, J., Kim, T., Lee, J., Park, J., and Han, M. (2005). Global physiological understanding and metabolic engineering of microorganisms based on omics studies. *Appl. Microbiol. Biotechnol.*, (5):567–579.

[Patel et al., 2002] Patel, P., Keogh, E., Lin, J., and Lonardi, S. (2002). Mining motifs in massive time series databases. In *In Proceedings of IEEE International Conference on Data Mining (ICDM 02)*, pages 370–377.

[Pedreschi et al., 2008] Pedreschi, R., Hertog, M. L. A. T. M., Carpentier, S. C., Lammertyn, J., Robben, J., Noben, J., Panis, B., Swennen, R., and Nicolai, B. M. (2008). Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. *Proteomics*, (7):1371–1383.

[Persson and Strang, 2002] Persson, P. and Strang, G. (2002). *Mathematical systems theory in biology, communications, computation, and finance*. Springer.

[Picard, 1985] Picard, D. (1985). Testing and estimating change-points in time series. *Advances in Applied Probability*, 17(4):841–867.

[Pignatiello and Runger, 1990] Pignatiello, J. and Runger, G. (1990). Selection of the subgroup size and sampling interval for a cusum control chart. *J. Qual. Technol.*, 22:173–186.

[Polpitiya et al., 2008] Polpitiya, A. D., Qian, W.-J., Jaitly, N., Petyuk, V. A., Adkins, J. N., Camp, David G., I., Anderson, G. A., and Smith, R. D. (2008). DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics*, 24(13):1556–1558.

[Press et al., 1992] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press, 2 edition.

[Price et al., 2006] Price, M. N., Arkin, A. P., and Alm, E. J. (2006). Opwise: operons aid the identification of differentially expressed genes in bacterial microarray experiments. *BMC bioinformatics*, 7:19.

[Purohit et al., 2004] Purohit, P. V., Rocke, D. M., Viant, M. R., and Woodruff, D. L. (2004). Discrimination models using variance-stabilizing transformation of metabolomic nmr data. *OMICS: A Journal of Integrative Biology*, 8(2):118–130.

[Qian et al., 2005] Qian, W. J., Liu, T., Monroe, M. E., Strittmatter, E. F., Jacobs, J. M., Kangas, L. J., Petritis, K., 2nd, D. G. C., and Smith, R. D. (2005). Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and sequest analysis: the human proteome. *Journal of proteome research*, 4(1):53–62.

[Quackenbush, 2004] Quackenbush, J. (2004). Data standards for 'omic' science. *Nature Biotechnology*, pages 613–614.

[Ramakrishnan et al., 2009a] Ramakrishnan, S. R., Vogel, C., Kwon, T., Penalva, L. O., Marcotte, E. M., and Miranker, D. P. (2009a). Mining gene functional networks to improve mass-spectrometry-based protein identification. *Bioinformatics*, 25(22):2955–2961.

[Ramakrishnan et al., 2009b] Ramakrishnan, S. R., Vogel, C., Prince, J. T., Wang, R., Li, Z., Penalva, L. O., Myers, M., Marcotte, E. M., and Miranker, D. P. (2009b). Integrating shotgun proteomics and mrna expression data to improve protein identification. *Bioinformatics*, 25(11):1397–1403.

[Ramsay and Silverman, 2002] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer-Verlag, New York.

[Re and Valentini, 2010] Re, M. and Valentini, G. (2010). Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction. *Systems Biology*, 8:98–111.

[Ridgeway, 2007] Ridgeway, G. (2007). Generalized boosted models: A guide to the gbm package.

[Robson, 2004] Robson, B. (2004). The dragon on the gold: Myths and realities for data mining in biomedicine and biotechnology using digital and molecular libraries. *Journal of Proteome Research*, (6):1113–1119.

[Savitzky and Golay, 1964] Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639.

[Scherl et al., 2005] Scherl, A., Francois, P., Bento, M., Deshusses, J. M., Charbonnier, Y., Converset, V., Huyghe, A., Walter, N., Hoogland, C., Appel, R. D., Sanchez, J. C., Zimmermann-Ivol, C. G., Corthals, G. L., Hochstrasser, D. F., and Schrenzel, J. (2005). Correlation of proteomic and transcriptomic profiles of staphylococcus aureus during the post-exponential phase of growth. *Journal of microbiological methods*, 60(2):247–257.

[Schulze and Usadel, 2010] Schulze, W. X. and Usadel, B. (2010). Quantitation in mass-spectrometry-based proteomics. *Annual Review of Plant Biology*, 61(1):491–516.

[Searls, 2005] Searls, D. (2005). Data integration: challenges for drug discovery. *Nat Rev Drug Discov*, 4(1):45–58.

[Shaffer, 1995] Shaffer, J. (1995). Multiple hypothesis testing. *Annu. Rev. Psychol.*, 46(1):561–584. An expanded version appeared as Multiple hypothesis testing: A review. National Institute of Statistical Sciences Technical Report No. 23, September, 1994.

[Shedden and Cooper, 2002] Shedden, K. and Cooper, S. (2002). Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc Natl Acad Sci U S A*, 99(7):4379–84.

[Shiohama and Taniguchi, 2003] Shiohama, T. and Taniguchi, M. (2003). Asymptotic estimation theory of change-point problems for time series regression models and its applications. *Lecture Notes-Monograph Series*, 41:257–284.

[Smith et al., 2002] Smith, R. D., Anderson, G. A., Lipton, M. S., Masselon, C., Pasa-Tolic, L., Shen, Y., and Udseth, H. R. (2002). The use of accurate mass tags for high-throughput microbial proteomics. *Omics : a journal of integrative biology*, 6(1):61–90.

[Soule et al., 2005] Soule, A., Salamatian, K., and Taft, N. (2005). Combining filtering and statistical methods for anomaly detection. In *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, IMC ’05, pages 31–31, Berkeley, CA, USA. USENIX Association.

[S.S. Prabhu and Montgomery, 1997] S.S. Prabhu, G. R. and Montgomery, D. (1997). Selection of the subgroup size and sampling interval for a cusum control chart. *IIE Transactions*, 29(6):451–457.

[Stevens, 1999] Stevens, J. P. (1999). *Intermediate Statistics A Modern Approach Second Edition*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

[Tabachnick and Fidell, 2006] Tabachnick, B. G. and Fidell, L. S. (2006). *Using Multivariate Statistics*. Allyn and Bacon, Inc.

[Tartaglia and Vendruscolo, 2009] Tartaglia, G. and Vendruscolo, M. (2009). Correlation between mrna expression levels and protein aggregation propensities in subcellular localisations. *Mol. BioSyst.*, 5:1873–1876.

[Templin et al., 2002] Templin, M., Stoll, D., Schrenk, M., Traub, P., Vohringer, C., and T.O., J. (2002). Protein microarray technology. *Trends Biotechnol.*, (4):160–166.

[The Mathworks Inc., 2009a] The Mathworks Inc. (2009a). Matlab version 7.5.0.342 (r2007b).

[The Mathworks Inc., 2009b] The Mathworks Inc. (2009b). Wavelet toolbox users guide.

[Tian et al., 2010] Tian, Y., Shumway, B. R., Youngbull, C., Li, Y., Jen, A. K.-Y., Johnson, R. H., and Meldrum, D. R. (2010). Dually fluorescent sensing of ph and dissolved oxygen using a membrane made from polymerizable sensing monomers. *Sensors and Actuators B: Chemical*, 47(2):714 – 722.

[Torres-García et al., 2009] Torres-García, W., Zhang, W., Runger, G. C., Johnson, R. H., and Meldrum, D. R. (2009). Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics*, 25(15):1905–1914.

[Tranter, 2000] Tranter, R. (2000). *Design and analysis in chemical research*. CRC Press.

[Troyanskaya et al., 2001] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.

[Troyanskaya et al., 2003] Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8348–8353.

[Tuikkala et al., 2008] Tuikkala, J., Elo, L., Nevalainen, O., and Aittokallio, T. (2008). Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics*, 9(1):202.

[Tuikkala et al., 2006] Tuikkala, J., Elo, L., Nevalainen, O. S., and Aittokallio, T. (2006). Improving missing value estimation in microarray data with gene ontology. *Bioinformatics (Oxford, England)*, 22(5):566–572.

[Tuv et al., 2009] Tuv, E., Borisov, A., Runger, G. C., and Torkkola, K. (2009). Best subset feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 1.

[Venancio and Aravind, 2009] Venancio, T. and Aravind, L. (2009). Reconstructing prokaryotic transcriptional regulatory networks: lessons from actinobacteria. *Journal of Biology*, 8(3):29.

[Vincent, 2005] Vincent, W. (2005). *Statistics in kinesiology*. Number v. 10 in Statistics in kinesiology. Human Kinetics.

[Walczak, 2000] Walczak, B. (2000). *Wavelets in chemistry*, volume 22. Elsevier Science.

[Wang and Bodovitz, 2010] Wang, D. and Bodovitz, S. (2010). Single cell analysis: the new frontier in 'omics'. *Trends in biotechnology*, 28(6):281–290.

[Washburn et al., 2003] Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, R. R., Plouffe, D., Deciu, C., Winzeler, E., and 3rd, J. R. Y. (2003). Protein pathway and complex clustering of correlated mrna and protein expression analyses in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6):3107–3112.

[Weerahandi, 2004] Weerahandi, S. (2004). *Generalized inference in repeated measures: Exact methods in MANOVA and mixed models*. Wiley-Interscience.

[Wei and Keogh, 2006] Wei, L. and Keogh, E. J. (2006). Semi-supervised time series classification. In *KDD'06*, pages 748–753.

[Welch, 1984] Welch, T. A. (1984). A technique for high-performance data compression. *Computer*, 17:8–19.

[Wilkins et al., 2006] Wilkins, M. R., Appel, R. D., Eyk, J. E. V., Chung, M. C., Gorg, A., Hecker, M., Huber, L. A., Langen, H., Link, A. J., Paik, Y. K., Patterson, S. D., Pennington, S. R., Rabilloud, T., Simpson, R. J., Weiss, W., and Dunn, M. J. (2006). Guidelines for the next 10 years of proteomics. *Proteomics*, 6(1):4–8.

[Wood et al., 2004] Wood, J., White, I. R., and Cutler, P. (2004). A likelihood-based approach to defining statistical significance in proteomic analysis where missing data cannot be disregarded. *Signal Processing*, 84(10):1777 – 1788.

[Worsley, 1983] Worsley, K. J. (1983). Testing for a two-phase multiple regression. *Technometrics*, 25(1):35–42.

[Yankov et al., 2007] Yankov, D., Keogh, E., Medina, J., Chiu, B., and Zordan, V. (2007). Detecting time series motifs under uniform scaling. In *SIGKDD*.

[Ye et al., 2008] Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., Bae, M., Janardan, R., Liu, H., Alexander, G., and Reiman, E. (2008). Heterogeneous data fusion for alzheimers disease study. In *KDD*.

[Yongzhong Li and Meldrum, 2010] Yongzhong Li, Hansa Thompson, C. H. F. H. J. F. R. H. J. W. Z. and Meldrum, D. R. (2010). An improved one-tube RT-PCR protocol for analyzing single-cell gene expression in individual mammalian cells. *Analytical and Bioanalytical Chemistry*, 397(5):1853–1859.

[Zhang et al., 2003] Zhang, J., Tsui, F.-C., Wagner, M. M., and Hogan, W. R. (2003). Detection of Outbreaks from Time Series Data Using Wavelet Transform. In *AMIA Symposium*.

[Zhang et al., 2006a] Zhang, W., Culley, D. E., Scholten, J. C., Hogan, M., Vitiritti, L., and Brockman, F. J. (2006a). Global transcriptomic analysis of desulfovibrio vulgaris on different electron donors. *Antonie van Leeuwenhoek*, 89(2):221–237.

[Zhang et al., 2006b] Zhang, W., Gritsenko, M. A., Moore, R. J., Culley, D. E., Nie, L., Petritis, K., Strittmatter, E. F., 2nd, D. G. C., Smith, R. D., and Brockman, F. J. (2006b). A proteomic view of desulfovibrio vulgaris metabolism as determined by liquid chromatography coupled with tandem mass spectrometry. *Proteomics*, 6(15):4286–4299.

[Zhao et al., 2008] Zhao, Z., Wang, J., Liu, H., Ye, J., and Chang, Y. (2008). Identifying biologically relevant genes via multiple heterogeneous data sources.

[Zhu et al., 2009] Zhu, H., Holl, M., Ray, T., Bhushan, S., and Meldrum, D. R. (2009). Characterization of deep wet etching of fused silica glass for single cell and optical sensor deposition. *Journal of Micromechanics and Microengineering*, 19(6).

# APPENDIX A

# BIOLOGICAL DATABASES DESCRIPTION

## A.1. General biological datasets

In this era of modern genomics, novel technologies are facilitating innovative discoveries increasing new genomic areas generating diverse types of data. Some are discussed below in three categories: components, interactions, and functional states [Joyce and Palsson, 2006].

### Components

Genomics : This is the most mature field in the "omics" research arena. Involves the study of the whole genome sequence and the information contained within. The sequence data from different species is allowing comparative analysis to identify gene-regulatory elements, to understand speciation, and to refine evolutionary tree of life. Besides the clear advantage of genome sequences, there are broad sources of information regarding gene annotation.

Transcriptomics : This type of data provides information about RNA transcript abundance present in the cell for particular genes.

Proteomics : It identifies and measures cellular levels for proteins by focusing at unique characteristic peptides for each protein.

Metabolomics : This area studies the metabolome which represents the output resulting from integration of the transcriptome, proteome, and interactome. These metabolites information provides functional information about particular cellular state. Data is gathered through mass spectrometry, NMR spectrometry and vibrational spectrometry.

Localizomics : This area aims to identify proteins subcellular location in the cell. Data from this particular field is quite expensive, requiring extraordinary efforts.

Lipidomics : Inventory of lipids identification and their associated factors within the cell. This is relative new field and very few data sets have been generated.

Glycomics : Inventory of carbohydrates and glycans in a cell. This is area is on its infancy stages and not much data available similarly to lipidomics.

Fluxomics : Quantify all the metabolites in regulatory networks.

Interactions

Protein-DNA interactome : It is widely known of the importance of interaction occurring at the cellular level. This type of data focuses on the interactions between proteins and DNA, particularly between transcription factors and their target promoters. This is helpful to understand genetic regulatory networks of the cell and its responses to environmental, extracellular, intracellular and intercellular changes. The most commonly used technology for high throughput explanation of gene-regulatory interactions is ChIP-chip or genome-wide location analysis.

Protein-Protein interactome : Learning protein-protein interactions will be important for understanding the structure and function of the integrated cellular network.

Functional

Phenomics : Studies the physical and biochemical properties of cellular organisms by genetic and environmental perturbations.

### A.1.1. Transcriptomic data

The advancements in last past decades in technologies to gather transcriptomic and proteomic data have evolved greatly with improved outcomes. In this section, some major types of these particular data kinds. Microarrays are very often used to measure mRNA transcript levels on a genome-wide scope. The general platform flow of current technologies to gather transcriptomic data can be found at [NCBI, 2010].

The most common types of microarray technology platforms for transcriptomic experimental are: (1) Spotted cDNA arrays, (2) Spotted oligonucleotide arrays, and (3) Affymetrix GeneChips. Every spot in the spotted cDNA array has DNA analog or complementary to the mRNA that we want to measure. These arrays typically use sets of plasmids of specific cDNAs in gridded liquid aliquots which are amplified by PCR [Hoffman et al., 2004] and two mRNA samples (interest and reference sample) with distinct fluorescent dyes are hybridized together in a single array [D′Agostino, 2004]. This particular type microarray platform is known for its flexibility and low cost. Similar to spotted cDNA arrays are the spotted oligonucleotide arrays, these arrays are also done by liquid handling on glass slides combining two color-tagged samples. Their distinction falls on the kind of input solution: synthetic oligonucleotide (often 60-70 mers) [Hoffman et al., 2004]. These synthesized oligonucleotides are custom-made for specific fragments. [Templin et al., 2002] depicts the platform flow for spotted arrays which functions by competitive binding of Control (tagged w/ Cy5 dye) and Sample (tagged w/ Cy3 dye) to bind to specific targets on the glass substrate. Once the control and sample have been incubated and hybridized on the glass, two different wavelength lasers are emitted and images are gathered. The two fluorescent dyes have different excitation and emission wavelengths visualizing in different colors. Hence, two pictures are obtained; one as a result of excitation of Cy5 showing red intensities for every spot on the glass and another one with green intensities from excitation of Cy3. These two pictures are then merged to obtain an image with relative intensities with colors ranging from red, orange, yellow to green providing information on which particular sample was under or over expressed.

Another very common technology used in transcriptomic experiments are Affymetrix GeneChips. These GeneChips are designed and synthesized in a factory using software to choose a series of 11 short oligonucleotides (25-mers) probes from the 3' end of each transcript. In contrast to spotted arrays, Affy chips are performed in a one-color platform meaning the intensity values are not relative to a reference sample. This allows for multiple comparisons using a gene chip for each sample of interest. Glass spotted arrays and Affymetrix GeneChip are two types of experimental platforms for transcriptomic. Templin et al. (2002) and Staal et al. (2003) provides extensive literature on these data types together with the advantages and limitations of each experimental platform for microarrays.

### A.1.2. Proteomic data

These matured experimental platforms for transcriptomic analysis have been very popular in the research community and well studied. Moving to protein experimental platforms and its current technologies; it is significant to understand that advances in mass spectrometric techniques have made quantitative proteomic profiling possible. Strategies for differential proteomics where quantitative measures for protein expression of stressed sample compared to undisturbed are obtained are commonly classified as: gel-based and mass spectrometry-based approaches. There are also label-free techniques. Mass spectrometry (MS) has been boosted by the development of soft protein ionization methods (i.e., electrospray ionization (ESI)). The general steps for MS approaches are as follow:

- Protein isolation from cell lysate or tissues
- Degradation of proteins into peptides by enzymatic digestion (commonly by trypsin digestion)
- Peptides are separated by liquid chromatography into capillaries. Separate molecules into ions by an ion source (i.e., MALDI and ESI).
- Mass is analyzed (i.e., Time-of-Flight (TOF)) .
- Computer identification and measurement of abundance.

More detail at [Aebersold and Mann, 2003], [Schulze and Usadel, 2010], and [Laurell et al., 2001].

A.2. Publicly available biological datasets

Some comprehensive databases publicly available [Green and Guyer, 2011] that are venues for cross-cutting research within main interests with data integration and data analysis are outline as follow (additional information in Appendix A).

OMIM : Online Mendelian Inheritance in Man contains a compendium of human genes and genetic phenotypes. [http://www.ncbi.nlm.nih.gov/omim]

GWAS : A Catalog of published Genome-Wide Association Studies. (http://www.genome.gov/gwastudies/)

SNP : Consortium and the HapMap Project: A detailed catalogue to identify variations in the human genome through Single Nucleotide Polymorphisms (currently under development:http://hapmap.ncbi.nlm.nih.gov/)

ENCODE and modENCODE The ENCyclopedia Of DNA Elements to catalog functional elements in the genomes of *homo sapiens*, *C. elegans* [Gerstein et al., 2010], and *D. melanogaster* [modENCODE Consortium et al., 2010] (currently under development: http://www.genome.gov/10005107).

GETx : Genotype-Tissue Expression project maps sites in the human genome to gene expression (http://commonfund.nih.gov/GTEx/).

LINCS : The Library of Integrated Network-based Cellular Signatures which is developing a database of molecular signatures describing different types of cells response to a variety of reagents (http://commonfund.nih.gov/LINCS/).

Cancer Genomes : Encompassing the characterization of cancer genomes to store information about somatic mutations are currently developing such as http://www.icgc.org/ and http://www.sanger.ac.uk/genetics/CGP/.

Others : A serial of complete databases in genomics such as nucleotide, sequencing, gene expression, SNP, and many others can be found at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/sites/gquery). Also, functional information such as pathway can be found in databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [Ogata et al., 1999] and the J. Craig Venter Institute with comprehensive microbial library (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi).

# APPENDIX B

# TRANSCRIPTOMIC NORMALIZATION TECHNIQUES

## B.1. Microarray normalization techniques

There are several methods aiming to reduce noise from microarray data such as quantile normalization one of the techniques used within this work being one of the most popular in software packages such as R-bioconductor.org. It is important to stress the different ways for noise reduction (within array and across arrays noise). Within array normalization is usually performed for two-color array and for the 11-25mers probes within Affymetrix GeneChips using perfect (PM) and mistmatch (MM) probes. We have focused in our work on reducing noise across replicates (across different arrays).
The most primitive normalization method includes a constant global shift of the measurement values based on the overall mean/median. This global normalization has certain assumptions such as no effect from channel dyes hence some local normalization are commonly used [Bilban et al., 2002]. Some methods that are worth trying are multi-step methods that include: background correction (remove local errors), normalization (remove array effect), and summarization (combine probe intensity into one value of gene expression) such as MAS 5.0, dChip, RMA, ZAM, and GL [Do and Choi, 2006].

MAS5 and Robust Multiple-Array Average (RMA) are some of the most common within Affymetrix data. MAS5 is an Affymetrix algorithm that uses PM and MM probes intensities by simple linear scaling; this approach is not effective on datasets with very large chip to chip differences [Do and Choi, 2006]. RMA uses global correction and quantile normalization. Irizarry et al. (2006) reviewed several normalization methods for microarray data finding a trade-off of variance and bias between the techniques . There is no clear “best” methodology that addresses optimal normalization parameters; the users should decide based on the particular applications [Irizarry et al., 2006].

Smoothing and filtering techniques are worth trying methods to reduce noise; their challenges are computation of optimal parameters to remove/smooth out random noise and not important signals (information). LOWESS or LOESS (locally weighted scatterplot smoothing) is a method of local regression. This method has been used to normalize microarray data mainly for two-color arrays since it needs a reference value. Their idea of the ’MA-plot’, where M is the difference in log expression values and A is the average of the log expression values is basically a nonlinear fit to M with respect to A (See Equation). From the fit an estimated value of M (M’) is found and through backwards equations of Equation B.1 new values of probes intensities are calculated. In this equation $k = 1, 2, ..., p$ represents the probe, and $x_{k1}, x_{k2}$ are probe intensities in array $1$ and $2$ respectively [Do and Choi, 2006]. Other very effective filtering techniques should be considered such as moving average filters (i.e., Savitzky-Golay filter) and by distribution fitting (i.e., Gaussian smoothing).

$$M_k = log_2\left(\frac{x_{k1}}{x_{k2}}\right); A_k = \frac{1}{2}log_2(x_{k1}x_{k2}) \tag{B.1}$$

# APPENDIX C

# EXTERNAL BIOLOGICAL DATA DESCRIPTION AND IDEAS

## C.1. Operon description

The information used for validation comes from curated databases which are available to all researchers in the area. These databases are built through bioinformatics efforts from known molecular information generated by genome projects [Ogata et al., 1999]. Researchers all over the world have worked on several experiments and have placed their findings in data repositories most of them stored through the National Institutes of Health. An example of this type of bioinformatics tool available is Kyoto Encyclopedia of Genes and Genomes (KEGG). KEGG is a knowledge base for systematic analysis of gene functions in terms of the networks of genes and molecules [Ogata et al., 1999]. Hence, the work presented in the proposal uses data from these types of knowledge based tools.

No biological experiments were run from our part to validate protein predictions. The type of data that is used for validation is somewhat different to the non-observable values of protein expression. These validation sets of information (i.e., operons, regulons, and pathways) provide network relationships of the genome for several species (i.e., *Desulfovibrio vulgaris* and *Shewanella oneidensis*) that could be used for inference. For example, operon information provides clusters of genes that are known to have same promoter, meaning, that these consecutive genes in the genome are known to activate together to get expressed. Figure shows a picture of a well studied operon in yeast. Figure 30 illustrates the definition of an operon, illustration adapted from [Venancio and Aravind, 2009].

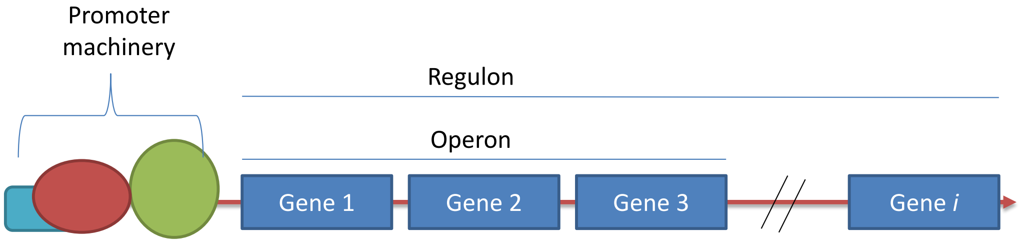


Figure 30. General diagram of operon and regulon.

Since these genes will be expressed together, it is natural to think that their corresponding protein expression will be similar, since they will be expressed at the same time. Hence, we used this type of information to check validate protein predictions because of their theoretical relationship which is available in resources such as KEGG. We showed in our studies how genes within operons had less dispersed protein predictions than any random set of genes; justifying theoretical results.

### C.2. Pathway description

Furthermore, pathway information includes networks of relationships between the genes such as metabolism and protein synthesis pathway. When we look into pathway information the genes relationship is not as linear as in the operon. Pathway information includes relationship of sets of genes within a specific pathway; however if a gene is active in a pathway all genes in the pathway will be expressed (even though might not be 1:1 relationship). Still using this type of information as validation provides some degree of information that otherwise is not available. Our studies showed that the validation wasn't as good for pathways as it was for operon but still fair which it is understandable. Figure 31 below shows a hypothetical example of a pathway network.

Lastly, we have used theoretical knowledge such as protein expression of genes clusters together by operon, regulon or pathway will be more similar than if they were not together in these groups. This information is available for most sequenced genome species and could certainly help with validation of protein expression which is much more difficult data structure to obtain. There is certain degree of "inaccuracy" inherent in those data resources but it will be minimal or at least less than expression data because of the nature on how this relationships have been discovered (such as sequencing).

### C.3. Integration of external information for modeling: another perspective

Biological knowledge such as operons, regulons and pathways information could be used as additional variable(s) in the model for prediction of protein abundance as an alternative to uncover more patterns. In the work presented in this dissertation used this type of information for needed validation. However, this linking information can be used in the model itself as a predictor(s). We present here two ways of integrating this information as predictors.

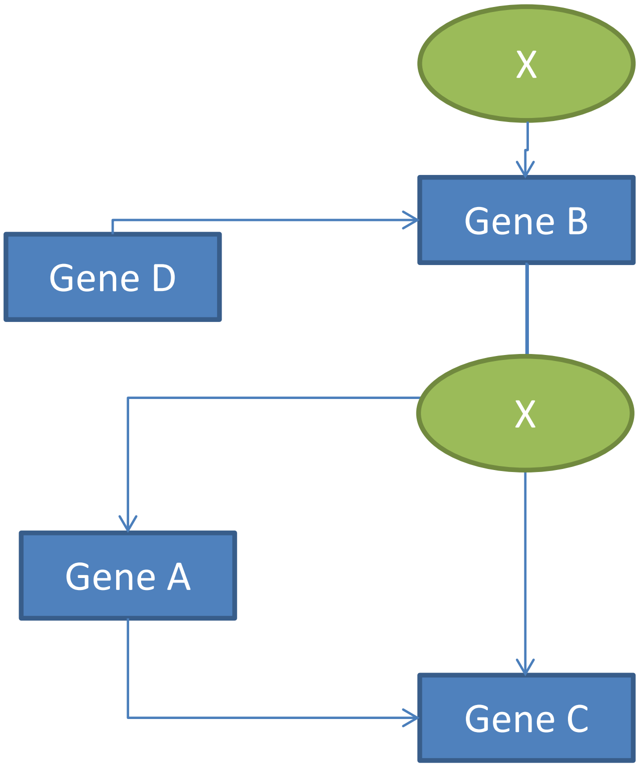


Figure 31. Hypothetical example of a pathway.

### C.3.1. Research idea 1: One-to-one relationship

If a one-to-one relationship can be established between the genes and operon assignment then the following criteria could be used for integrating external information such as operon in the modeling procedure. Note that this one-to-one relationship was the particular case of study for research questions explored in Chapters 3 and 4.

External biological data can provide you with particular connectivity relationships for the genes in study. This connections and special grouping could be relevant factors towards understanding a particular response. Hence, based

TABLE 18
Simple description to incorporate biological knowledge idea 1.

| GeneID | Bio-Knowledge Group (i.e., operon) |
|---|---|
| Gene 1 | Operongroup2 |
| Gene 2 | Operongroup1 |
| | |
| Gene 3000 | Operongroup2 |

on the one-to-one relationship, every gene can be identified to be part of a particular group (i.e., operon group i). This provides a column-like predictor as the one shown in Table 18. If genes are strictly part of only one "bio" (i.e., operon, regulon or pathway) group; a categorical variable(s) can be added providing the particular group for each gene. For example, suppose Gene 1 and Gene 3000 are both part of operon group 2. This portrays categorical information that can be incorporated as a variable in predictive techniques such as ensembles methods.

### C.3.2. Research idea 2: Network relationship

The previous scenario will be very unlikely for several organisms since a gene could be part of several network groups from the biological knowledge sources. Hence indicator variables could be added in the model to portrayed this linking relationships. For example, Gene 3 could be part of PathwayGroup1 and also PathwayGroup2, the indicator variables $(0-1)$ will be able to represent this relationship in the model, 1 if related to the group or 0 if no relationship exists. This is an useful technique that generalizes how to integrate grouping information to the model. Also, pathway information is readily available in databases such as KEGG. An example shown in Table 19.

### C.3.3. Remarks

The only problem that arises from including this information as part of the model is that no knowledge will be available to assess model performance for undetected proteins.

TABLE 19
Simple description to incorporate biological knowledge idea 2.

| GeneID | PathwayGroup1 | PathwayGroup2 |
|---|---|---|
| Gene 1 | 0 | 1 |
| Gene 2 | 1 | 0 |
| Gene 3 | 1 | 1 |
| | | |
| Gene 3000 | 0 | 1 |