Data integration and visualization for systems biology data

Hui Cheng

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics and Computational Biology

Pedro Jose Pedrosa Mendes, Chair
Ina Hoeschele
Reinhard C. Laubenbacher
Brett M. Tyler

October 27, 2010
Blacksburg, Virginia

Keywords: Systems biology, data integration, data visualization, data fusion, biplot display, Fast Fourier transform, phase spectrum

Data integration and visualization for systems biology data

Hui Cheng

Abstract

Systems biology aims to understand cellular behavior in terms of the spatiotemporal interactions among cellular components, such as genes, proteins and metabolites. Comprehensive visualization tools for exploring multivariate data are needed to gain insight into the physiological processes reflected in these molecular profiles. Data fusion methods are required to integratively study high-throughput transcriptomics, metabolomics and proteomics data combined before systems biology can live up to its potential. In this work I explored mathematical and statistical methods and visualization tools to resolve the prominent issues in the nature of systems biology data fusion and to gain insight into these comprehensive data.

In order to choose and apply multivariate methods, it is important to know the distribution of the experimental data. Chi square Q-Q plot and violin plot were applied to all *M. truncatula* data and *V. vinifera* data, and found most distributions are right-skewed (Chapter 2). The biplot display provides an effective tool for reducing the dimensionality of the systems biological data and displaying the molecules and time points jointly on the same plot. Biplot of *M. truncatula* data revealed the overall system behavior, including unidentified compounds of interest and the dynamics of the highly responsive molecules (Chapter 3). The phase spectrum computed from the Fast Fourier transform of the time course data has been found to play more important roles than amplitude in the signal reconstruction. Phase spectrum analyses on *in silico* data created with two artificial biochemical networks, the Claytor model and the AB2 model proved that phase spectrum is indeed an effective tool in system biological data fusion despite the data heterogeneity (Chapter 4). The difference between data integration and data fusion are further discussed. Biplot analysis of scaled data were applied to integrate transcriptome, metabolome and proteome data from the *V. vinifera* project. Phase spectrum combined with *k*-means clustering was used in integrative analyses of transcriptome and metabolome of the *M. truncatula* yeast elicitation data and of transcriptome, metabolome and proteome of *V. vinifera* salinity stress data. The phase spectrum analysis was compared with the biplot display as effective tools in data fusion (Chapter 5). The results suggest that phase spectrum may perform better than the biplot.

This work was sponsored by the National Science Foundation Plant Genome Program, grant DBI-0109732, and by the Virginia Bioinformatics Institute.

Acknowledgements

First and foremost I would like to sincerely thank my advisor, Dr. Pedro Mendes, for his guidance, support, understanding and patience throughout this research. His insight and stimulating discussions on the subject are invaluable in helping me move towards the goal. His constant encouragement and input always brighten my heart and led me through the difficult times during the research.

I would also like to extend my appreciation to my committee members: Dr. Ina Hoeschele, Dr. Reinhard Laubenbacher, Dr. Brett Tyler, and my former committee member, Dr. Vladimir Shulaev for their advice and suggestions through the whole process.

Special thanks to Dennie Munson for her generous assistance that made my GBCB program accomplished much smoothly. Her warm support continues to give me strength. Thank you for always being there for me. Also I would like to thank Dr. David Bevan for his support in my GBCB fulfillment.

To all my former fellow members in Mendes research group, Dr. Bharat Mehrotra, Dr. Stefan Hoops, Dr. Saroj Mohapatra, Dr. Ana Martins, Xingjing Li, Aejaaz Kamal, Kimberly Heard, Dr. Revonda Pokrzywa, Dr. Wei Sha, Dr. Diogo Camacho, and Jim Walke, I send a heartfelt thank-you. I thank you for teaching me many things and lending me lot of help both professionally and personally.

My sincere appreciation is extended to Janet Donahue for sharing her time and knowledge to enrich my graduate experience in Virginia Tech. A special thank-you to Joel Donahue for introducing me to "*The Scientist and Engineer's Guide to Digital Signal Processing"*.

My thanks also go to my motivated fellow students in GBCB program for their friendship, company and valuable conversations.

I would like to thank my parents Lingyu Cheng and Peilan Jia for their years of unending support and unconditional love. I would not be here if it weren't for you. Also, I want to thank Xiaohai's mom, Fengming Bu for her understanding and support during our difficult time.

But most of all a loving thank you to my Atticus who has brought joy to the last stage of my graduate life and my husband, Xiaohai, without whose unwavering love and devotion, this work could not be possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Systems theory and systems biology

In the 1950s biologist Ludwig von Bertalanffy and other scientists established the field of "systems theory" (Abraham 2002). This interdisciplinary study brings together theoretical principles and concepts from ontology, philosophy of science, physics, biology and engineering. Von Bertalanffy emphasized the ideas of holism, organicism, and open systems. Rather than trying to "explain observable phenomena by reducing them to an interplay of elementary units investigatable independently of each other" (reductionism), systems theory focuses on the arrangement of, and relations between, the parts which connect them into a whole (holism) (Chong and Ray 2002). Von Bertalanffy applied general systems theory not only to biology, but to geology, psychology, economics, and social science as well.

More than 50 years later, the work in understanding systems has evolved into many disparate areas of study. And the definition of system theory remains effective for systems biology as practiced today with the integration and application of mathematics, engineering, physics, and computer science to understanding a range of complex biological regulatory systems (Chong and Ray 2002; Ideker, Galitski et al. 2001).

To understand biology at the system level, we must examine the structure and dynamics of cellular and organismal function. To conduct such an analysis, a comprehensive set of quantitative data is required. Comprehensiveness in measurements requires consideration of three aspects (Kitano 2002): (i) factor comprehensiveness, which reflects the numbers of mRNA transcripts, metabolites and proteins that can be measured at once; (ii) time-line comprehensiveness, which represents the time frame within which measurements are made; and (iii) item comprehensiveness, which refers to the simultaneous measurement of multiple items, such as mRNA, metabolite and protein concentrations, localization, and so forth.

Projects already under way, such as *Medicago truncatula* and *Vitis vinifera* and many others, are making large-scale measurements on transcriptomics, metabolomics and proteomics with the ultimate goal of providing an in-depth understanding of the cellular mechanism.

1.2 *Medicago truncatula* and *Vitis vinifera* projects

1.2.1 *Medicago truncatula* project

Medicago truncatula (commonly known as "barrel medic" because of the shape of its seed pods) is a forage legume commonly grown in Australia. It is an omni-Mediterranean species

and closely related to the world's major forage legume, alfalfa (*Medicago sativa*). As a legume, and unlike the most studied genetic model plant, *Arabidopsis thaliana*, *M. truncatula* establishes symbiotic relationships with nitrogen fixing Rhizobia. The mutualistic interactions provide a plentiful supply of nitrogen to the plants that in turn result in very high protein levels in legumes. Therefore, legumes have been assimilated as a major dietary source of protein for both humans and animals. Legumes also provide nitrogen to the soil, thus reducing the need for exogenous fertilizers (Lei, Elmer et al. 2005). Legumes are also rich sources of bioactive natural products (Dixon and Sumner 2003), of which triterpene saponins (Suzuki, Reddy et al. 2005) protect the plant by nature of their anti-microbial, anti-insect and anti-palatability activity; isoflavonoids are also products of legumes (and few other plants) which have been shown to play potential roles in the prevention of cancers and cardiovascular disease.

M. truncatula has been chosen as a model legume species for genomic studies in view of its small diploid genome, fast generation time (from seed-to-seed), and high transformation efficiency (Cook 1999). *M. truncatula* is the subject of several major US genomics initiatives funded federally through NSF, USDA, and privately through the Samuel Roberts Noble Foundation (SRNF).

An integrated approach to funtional genomics and bioinformatics in a model legume (Mendes and Dixon 2001) was a project proposed by Dr. Pedro Mendes and Dr. Richard Dixon (SRNF) and funded by NSF that took place between 2001 and 2007. The main experimental approach of this project was to induce the expression of these natural products, and other areas of metabolism, by exposing cell cultures to biotic and abiotic elicitors. Use of cell suspension cultures allowed sufficient material to be collected and analyzed in parallel, while providing a laboratorial environment which was simple and easy to control. Three experimental conditions were chosen that mimic natural environmental challenges: UV radiation, a fungal elicitor (a purified gluco-mannan from yeast cell wall), and methyl jasmonate (a plant hormone involved in systemic acquired resistance). The ultimate goal of this project was to generate a truly systematic data set for control and elicited cell cultures. Such data encompasses expressed sequence-tags and the associated mRNA, protein and metabolite identities and concentrations, all collected over time after applying the elicitations.

After elicitation, samples of sufficient size were collected in triplicate and snap frozen at 21 time points: 0, 5, 15, 30, and 45 minutes; and 1, 2, 3, 4, 6, 8, 10, 12, 15, 18, 21, 24, 30, 36, 42 and 48 hours. The control experiment was carried out in parallel to each elicitation experiment (also in triplicate). Each sample was analyzed for: *i*) gene expression using DNA microarrays; *ii*) protein expression patterns using two-dimensional gel electrophoresis and mass spectrometry; *iii*) changes in a range of primary and secondary metabolites by LC/MS,

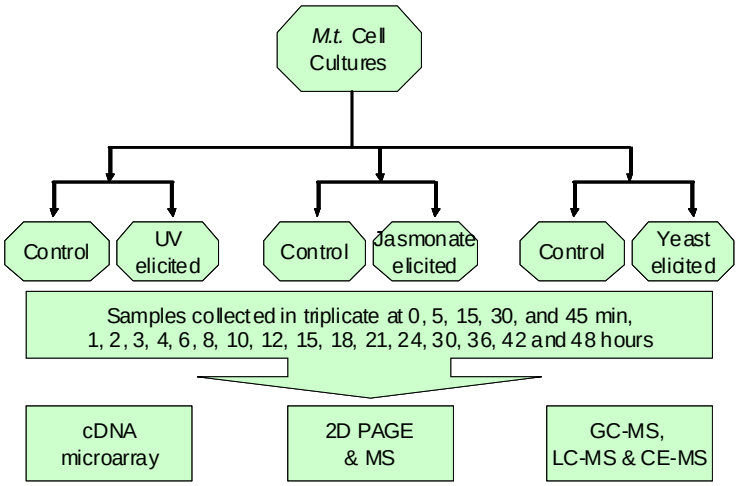


Figure 1.1 *Medicago truncatula* project experimental design.

GC/MS and CE/MS analyses (Figure 1.1). The ability to compare information from all functional levels of gene expression in a homogeneous, inducible system, will lead to a synergistic leap in our understanding of the genetic programming of cellular metabolism.

The data generated by this project provides information about the extent and nature of gene expression reprogramming in response to biotic and abiotic signals at the transcription, protein and metabolite levels. These studies are expected to result in an expansion of the scope of our understanding of induced plant defense responses to a global cellular level.

1.2.2 *Vitis vinifera* project

On a worldwide basis, the common table grape (*Vitis vinifera*) is both the most widely cultivated and economically important fruit crop**.** The United States wine, grape and grape product industries contribute more than $162 billion annually to the American economy (http://www.wineinstitute.org). Among all the grape products, wine produced from cultivars of *Vitis vinifera* has the highest economic value (Goes da Silva, Iandolino et al. 2005).

Although the consumption of alcohol is controversial, it is now well established that the consumption of wine at moderate levels is correlated with health benefits including the reduction of risk of cardiovascular disease, stroke and cancer (German and Walzem 2000; Cramer, Cushman et al. 2002). Phenolics in wines are major contributors to these health benefits. These compounds are particularly high in red wine and their amounts can be increased in the berries by exposure to light and water-deficit (Kolb, Kaser et al. 2001; Kennedy, Matthews et al. 2000).

The *Vitis* functional genomics project (Cramer, Cushman et al. 2002) was funded by the NSF Plant Genome Program (September, 2002 to August, 2006). It was part of an international effort to build resources for the International Grape Genome Community. The particular efforts were focused on the improvement of abiotic stress tolerance (cold, drought and salinity) and wine quality.

Abiotic stress in the form of drought, salinity, and cold has a major impact on grape production and quality. Several studies have shown that water-deficit-stressed grapevines produce superior quality wine (Kennedy, Matthews et al. 2000). The molecular genetic and biochemical basis for this correlation, however, remains poorly understood. An integrative and quantitative analysis of mRNA, protein, and metabolite changes following abiotic stress imposition is required to enhance production efficiency under stress conditions and to understand the plant-derived contribution to constituents of wine quality (Cramer, Ergul et al. 2007). In the *Vitis vinifera* project, the effects of cold, salinity and polyethylene glycol (PEG) generated osmotic stress on the growth of Cabernet Sauvignon grapes over the course of 24 hours were studied. PEG stress is used to mimic water-deficit condition (osmotic stress). One long-term goal of this research is to develop comprehensive genomic tools to facilitate the genetic engineering of improved abiotic stress tolerance traits in *V. vinifera* or to establish knowledge to allow a precise manipulation of environmental conditions to reliably produce wine with superior qualities.

Following multiple abiotic stresses, the shoots of *V. vinifera* cv. Cabernet Sauvignon plants were collected in triplicate and immediately frozen at 0, 1, 4, 8 and 24 hours. The control experiment was carried out in parallel. Each sample was analyzed for: *i*) gene expression using Affymetrix arrays; *ii*) protein expression patterns using two-dimensional gel electrophoresis and mass spectrometry; *iii*) changes in metabolite levels by GC/MS analyses (Figure 1.2). Ultimately, these data sets are to be integrated into a reliable prediction model for wine characteristics. The proposed research will greatly facilitate future gene discovery and enable improvements to be made in both production efficiency and wine quality under environmentally adverse growing conditions.

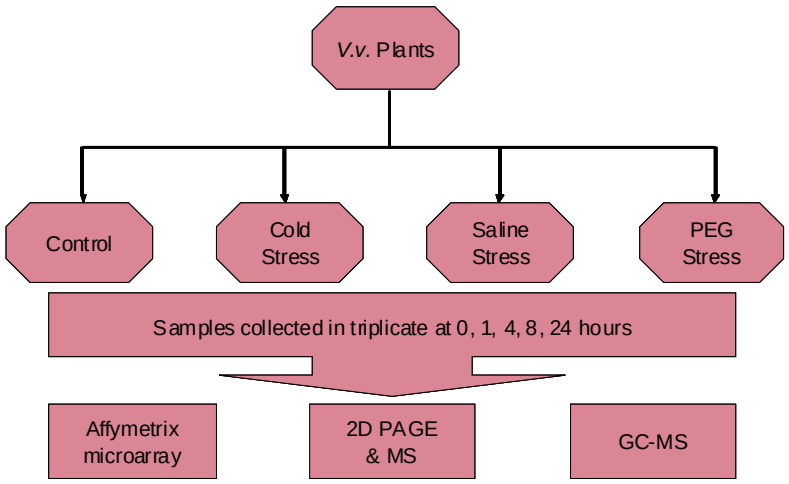


Figure 1.2 *Vitis Vinifera* project experimental design.

Systems biology study such as the two projects described above produced large-scaled comprehensive data sets enriched with transcriptomic, proteomic and metabolomic profiles. Carefully selected approaches are much needed to evaluate, analyze, and integrate these data; otherwise the conclusion drawn from them may be misleading. In the following sections, the concepts of data integration and data fusion will be reviewed (Section 1.3); the common visualization and clustering methods (Section 1.4), dimension reduction mechanisms (Section 1.5) and missing value estimation (Section 1.6) will be discussed.

1.3 Data integration and data fusion

Data integration in life sciences is a relatively recent activity, while the concept itself is hardly new. It started since 1960 when the rapid adoption of databases led to the need to share or merge existing repositories (Ziegler and Dittrich 2004). Data integration is the problem of combining data residing at different sources and providing the user with a unified view of these data (Lenzerini 2002). The availability of multiple independent, heterogeneous biological data requires an enormous effort to do data integration. After data from several sources are extracted, quality controlled, preprocessed and transported to a data warehouse and can be queried with a single schema, the next step is “data fusion”.

The concept of “data fusion” was coined in the Electrical and Electronic Engineering field, frequently called “multi-sensor data fusion”. Data fusion techniques combine data from multiple sensors, and related information from associated databases, to achieve improved accuracies and more specific inferences than could be achieved by the use of a single sensor alone (Llinas and Hall 1998). Historically, data fusion methods were developed primarily for military applications including automated target recognition, battlefield surveillance, and guidance and control of autonomous vehicles. However, in recent years these methods have been applied to civilian applications, such as monitoring of complex machinery, geographical information systems, medical diagnosis and robotics, etc.

Multi-sensor data fusion involves data association and correlation. Methods that have been used for this purpose include neural networks, template methods, and pattern recognition methods such as cluster algorithms (Hall and McMullen 2004). In order to fuse the raw data, the original sensor data must be commensurate (*i.e.*, must be observations of the same or similar physical quantities such as visual images) and must be able to be properly associated.

1.4 Distance based visualization tools for multivariate data

Current common visualization tools in biological data analysis, such as hierarchical clustering, *k*-means clustering and self-organizing map are based on geometric distance measurements. Distance measure is one way to quantify the similarity or dissimilarity of objects.

Three frequently used distance measurements are the Euclidean distance, the Manhattan distance and correlation ‘distance’ (Rencher 2002; Gibbons and Roth 2002) . Given vectors of data points $\boldsymbol{x} = (x_1,\ldots, x_n)$, $\boldsymbol{y} = (y_1,\ldots, y_n)$,

Euclidean distance: $$d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Manhattan distance: $$d_M(x,y) = \sum_{i=1}^{n}\left|x_i - y_i\right|$$

Correlation ‘distance’: $$d_C(x,y) = 1 - \frac{\sum_{i=1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}(x_i - \bar{x})^2 \sum_{i=1}(y_i - \bar{y})^2}}$$

Euclidean and Manhattan distances both measure absolute differences between vectors. Correlation 'distance' measures trends or relative differences (i.e. without scale).

Since Eisen *et al.* (Eisen, Spellman et al. 1998) implemented this technique to microarray gene expression data, hierarchical clustering has become the most popular method to cluster microarray data. In hierarchical clustering, the similarities of objects are represented in a tree structure or dendrogram. It can be carried out in two ways: agglomerative and divisive. Agglomerative clustering takes each entity as a single cluster to start off with and then builds bigger and bigger clusters by grouping similar entities together until the entire dataset is encapsulated into one final cluster. Divisive hierarchical clustering works the opposite way: the entire dataset is first considered to be one cluster and is then broken down into smaller and smaller ones until each subset consists of only one single entity.

k-means clustering (Rencher 2002; Tavazoie, Hughes et al. 1999) is one of the simplest unsupervised learning algorithms. *K*-means clustering is different from hierarchical clustering in that the number of clusters, *k,* needs to be determined at the onset. The goal is to divide the objects into *k* groups (clusters) such that the sum of squares of distances between data points in a cluster and the corresponding cluster centroid is minimal.

The self-organizing map (SOM) (Tamayo, Slonim et al. 1999) is a data visualization technique invented by Teuvo Kohonen, a Finnish academician and prominent neural network researcher. SOM reduces the dimensions of data through the use of self-organizing neural networks. The way SOMs reduce dimensions is by producing a map of usually 1 or 2 dimensions which plot the similarities of the data by grouping similar data items together. So SOM accomplishes two things, it reduces dimensions and display similarities between the groups. SOM has a number of features that make them particularly well suited to clustering and analysis of gene expression patterns. It allows one to impose partial structure on the clusters and facilitating easy visualization and interpretation. The drawback of SOM is it is very sensitive to noise, and every time it is run it may give out different results.

Clustering methods alone cannot give detailed information of data and relationships between molecules and samples. Data dimension reduction methods, such as singular value decomposition (SVD) and non-negative matrix factorization (NMF) provide insights into the complex data matrix without assuming any grouping *a priori*.

1.5 SVD and NMF

The singular value decomposition (SVD) is a generalization of the eigen-decomposition which can be used to analyze rectangular matrices (Abdi 2007). SVD can be used to approximate any rectangular matrix by a matrix of same size but of lower rank, such that the

sums of squares of the differences between the elements of the matrix and its approximation is minimized (Pittelkow and Wilson 2003). If we have an *n* by *p* matrix Y of rank *r* with $r \le p \le n$, then the SVD of Y can be written as $Y = U\Lambda V^T$, where *U* (*n* by *p*) and *V* (*p* by *p*) are matrices of singular vectors and Λ (*p* by *p*) is a diagonal matrix of singular values. *U*, named the left singular vectors of Y, is the matrix with columns corresponding to the *p* orthogonal eigenvectors of YY^T. V, named the right singular vectors of Y, is the orthogonal matrix corresponding to the eigenvectors of Y^TY. Λ =diag(λ_1, λ_2, …, λ_p) with $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_r \ge 0 = \lambda_{r+1} = \lambda_{r+2} = ... = \lambda_P = 0$. The singular values are the positive square roots of the eigenvalues of Y^TY.

The Non-negative Matrix Factorization (NMF) is a matrix factorization algorithm developed by Lee and Seung in 1999 (Lee and Seung 1999) to decompose images into recognizable features. Recently it has been increasingly used in biological data (Carmona-Saez, Pascual-Marqui et al. 2006; Du, Sajda et al. 2005; Gao and Church 2005). This technique can be applied to the analysis of multidimensional datasets in order to reduce the dimensionality, discover latent patterns and, more important, aid in the interpretation of the data (Pascual-Montano, Carmona-Saez et al. 2006).

Formally, the non-negative matrix decomposition can be described as: $V \approx WH$ where V_{nxm} is a positive data matrix with *m* variables and *n* observations, W_{nxr} are the reduced *r* basis vectors or factors, and H_{rxm} contains the coefficients of the linear combinations of the basis vectors needed to reconstruct the original data (also known as encoding vectors). The rank *r* of the factorization is generally chosen so that $(n+m)r < nm$, and the product *WH* can be regarded as a compressed form of the data in *V*. The main difference between NMF and other classical factorization models relies in the non-negativity constraints imposed on both the basis *W* and encoding vectors *H*. In this way, only additive combinations are possible. The factors produced by this method can be intuitively interpreted as parts of the data or as subsets of elements that tend to occur together in the data set (Pascual-Montano, Carmona-Saez et al. 2006).

The classic NMF uses an iterative algorithm: start from non-negative initial conditions for *W* and *H*, iteration of these update rules for *W* and *H* given as follows:

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \qquad H_{a\mu} \leftarrow H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$$

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$

Iteration of these update rules converges to a local minimum of the objective function (Lee and Seung 1997)

$$\min_{W,H} \|V - WH\|^2$$

The fidelity of the approximation enters the updates through the quotient $\frac{V_{i\mu}}{(WH)_{i\mu}}$. The update rules preserve the non-negativity of *W* and also constrain the columns of *W* to sum to unity.

1.6 Estimation of missing values

Missing values are common in gene expression data, metabolomics and proteomics data. Unfortunately many multivariate analysis methods require a complete matrix of expression values as input. A small number of rows with missing entries in the data matrix do not constitute a serious problem; one can simply discard each row that has a missing value. However, with this procedure, a small portion of missing data, if widely distributed, would lead to a substantial loss of data.

The distribution of missing values in a data set is an important consideration (Rencher 2002; Xing, Schumacher et al. 2003; Howell). Randomly missing variable values scattered throughout a data matrix are less serious than a pattern of missing values that depends to some extent on the values of the missing variables.

The row-average method that replaces a missing value by its row average reduces the variance and the absolute value of the covariance. Therefore, the resulted sample covariance matrix computed from this data matrix is biased (Rencher 2002).

Recently, for missing value estimation in microarray data analysis, weighted *k*-nearest neighbors imputation (KNNimpute) and the SVD based method (SVDimpute) (Troyanskaya, Cantor et al. 2001), and local least squares imputation method (LLSimpute) (Kim, Golub et al. 2005) have been introduced.

The KNNimpute method (Troyanskaya, Cantor et al. 2001) finds *k* other genes with expressions most similar to that of g1 and with the values in their first positions not missing. The missing value of g1 is estimated by the weighted average of values in the first positions of these *k* closest genes. For the weighted average, the contribution of each gene is weighted by the similarity of its expression to that of g1.

In the SVDimpute method (Troyanskaya, Cantor et al. 2001), the SVD of the matrix Y', which is obtained after all missing values of the Y are substituted by zero or row averages, is

computed. Then, using the *t* most significant eigengenes (Alter, Brown et al. 2000) of Y', where the specific value of *t* is either predetermined or determined based on datasets, a missing value *x* in g1 is estimated by regressing this gene against the *t* most significant eigengenes. Using the coefficients of the regression, the missing value is estimated as a linear combination of the values in the first position of *t* eigengenes. When determining these regression coefficients, the missing value g1 (1) of g1 and the first values of the *t* eigengenes are not used. The above procedure is repeated until the total change of the matrix becomes insignificant.

In the local least squares imputation method (LLSimpute) (Kim, Golub et al. 2005), a target gene that has missing values is represented as a linear combination of similar genes. The similar genes are chosen by *k* -nearest neighbors or *k* coherent genes that have large absolute values of Pearson correlation coefficients.

The missing value estimation methods developed for microarray data analysis can be used in proteomics and metabolomics data sets, where the missing value problem is more severe.

In the following chapters, Chi square Q-Q plot and violin plot will be applied to "omic" data produced from the *M. truncatula* and *V. vinifera* projects in order to characterize the distributions of the data sets (Chapter 2); Biplot display will be demonstrated with a yeast glycolysis and glycerol biosynthesis model and applied to *M. truncatula* data, the inferences drawn from Biplot will be discussed (Chapter 3); Phase spectrum computed from time course data's Fourer transform will be used to analyze the *in silico* data created from Claytor model and AB2 model (Chapter 4); Biplot and phase spectrum analysis will be implemented respectively as data fusion tools in *V. vinifera* and *M. truncatula* data, and their results are compared (Chapter 5); Chapter 6 will summarize this research and the potential benefit of systems biology in medical sciences will be briefly discussed.

This page is intentionally left blank.

Chapter 2

Characterization of systems biology data

2.1 Introduction

The massive amounts of systems biology data generated through the joint effort of numerous scientists across multiple fields have created an increased demand for quantitative analysis methods. Knowing the characteristics of the target data is the prerequisite of applying any statistical methods. Through collaboration with the Dixon and Sumner labs from Samuel Roberts Noble Foundation (SRNF) and Cramer lab from University of Nevada-Reno we have obtained comprehensive multi-omics time-couse data sets. Characterization of these data will provide a foundation for further analysis.

The multi-omics data sets share a common feature: they all use relative quantitation. Gene expression level, metabolite and protein quantitation and the accompanied low-level data processing and/or normalization largely shaped the outcome of the systems biology data.

In microarray data processing, fluorescent signals are usually base-2 logarithm transformed and log-ratios used to quantify the relative differential expression (Smyth, Yang et al. 2002). Normalization is necessary to remove systematic variations in the experiments (e.g. differences in labeling efficiency between the two fluorescent dyes). Print-tip LOESS normalization used in *Medicago truncatula* (*M. truncatula*) project provides a well-tested general purpose normalization method which has given good results on a wide range of arrays (Smyth and Speed 2003). This method is based on robust local regression and account for intensity and spatial dependence in dye biases for different types of cDNA microarray experiments (Yang, Dudoit et al. 2002).

Although some research (Hoyle, Rattray et al. 2002; Novak, Kim et al. 2006; Konishi 2004) has been done on transcriptomics data distributions since microarray data started becoming available in 1995 (Schena, Shalon et al. 1995), the equivalent studies on description of metabolomics and proteomics data is still sparse. Metabolomics is a relatively new discipline in terms of the global analysis and techniques for high-throughput metabolite profiling. Due to the chemical complexity of the metabolome there does not exist any singular technique for profiling all of the metabolites simultaneously, so a mixture of analytical technologies is used (Sumner, Mendes et al. 2003; Shulaev 2006). The common approach to metabolomics involves segregation of the metabolome into several subclasses followed by parallel analyses utilizing the selectivity of mass spectrometry (Huhman and Sumner 2002). In metabolomics, normalization is used not only to remove the systematic variation, but to allow data on different scales to be compared, by bringing them to a common scale. Normalization process is defined by the corresponding technology and samples. The standards are often set by each lab. Take *M. truncatula* study as an example, in GC-MS analysis, peak areas are normalized by dividing each peak area value by the mean peak area for that compound (Broeckling, Huhman et al. 2005); in LC-MS, values of triterpene saponins are relative molecular ion

intensities normalized on an equal cell weight basis with the level of hexose hederagenin in non-elicited cells set as 100 (Suzuki, Reddy et al. 2005); in CE-MS, relative peak areas are determined by taking the peak ratio of each component to that of the internal standard, ethionine (Williams, Cameron et al. 2007). The expansion of techniques, extensive raw data processing (Shulaev 2006) and under-developed international standards (Castle, Fiehn et al. 2006) accentuate the complexity of metabolomics data.

Since proteins are considered the functionally most important biological molecules, proteomics is instrumental in discovery of biomarkers (Verrills 2006). In a wider sense, proteomics research not only identifies and quantifies the entire protein complement in a given cell, tissue or organism, but also assesses protein activities, modifications and localization, and interactions of proteins in complexes (Winslow, Cortassa et al. 2005; Spickett, Pitt et al. 2006). Unfortunately, there is no experimental platform to systematically measure the diverse properties of proteins at high throughput. Currently, the most mature and versatile proteomic methods are based on mass spectrometry coupled with 2D-PAGE (Aebersold 2003). The protein spot volume is normalized by dividing each spot volume by the total spot volume (Lei, Chen et al. 2010). The typical problems posed in proteomic research are high variation in total spot volume detection (Lei, Elmer et al. 2005) and low protein identification success rate (Watson, Asirvatham et al. 2003).

The shibboleth that sets systems biology apart from the more traditional and more reductionist molecular biology are involvement of statistical analysis, mathematical modeling and integration of high-throughput data from different biological levels (Kell 2006; Kell 2005). Most of the multivariate inferential procedures are based on the multivariate normal distribution because of its mathematical tractability, such as the distribution can be completely described using only means, variance and covariances; if the variables are uncorrelated, they are independent; linear functions of multivariate normal variables are also normal; and even when the data are not multivariate normal, the multivariate normal may serve as a useful approximation, especially in inferences involving sample mean vector, which are approximately multivariate normal by the central limit theorem (Rencher 2002). Thus knowing the distribution of the experimental data is the prerequisite of choosing the appropriate approaches and applying these methods. In addition, most systems biological data generated today using relative quantitations based on lab specific standards. Therefore, data characterization on individual basis is recommended.

2.2 Methods

The data in this study is *M. truncatula* data from Sumner lab of Noble Foundation and *Vitis vinifera* data from Cramer lab of University of Nevada. Both projects were scrupulously

designed and carried out (as described in Chapter 1). For each experiment in each project, normality tests and distribution plots will be applied to the data set.

2.2.1 Chi-square Q-Q plot

Many tests and graphical procedures have been suggested for evaluating whether a data set likely originated from a multivariate normal population (Easton and McCulloch 1990; Holgersson 2006; Liu, Parelius et al. 1999; Rencher 2002). One possibility is to check each variable separately for univariate normality. However, marginal normality of each component doesn't imply their joint multivariate normality. Checking for multivariate normality is conceptually not as straightforward as assessing univariate normality. Because of the inherent "sparseness" of multivariate data, a goodness-of-fit test would be impractical. Unless n is very large, a multivariate sample may not provide a very complete picture of the distribution from which it was taken (Rencher 2002).

However some checks on the distribution are often desirable. The chi-square Q-Q plot (Holgersson 2006; Azzalini 1985) is an informal graphical approach to assess multivariate data normality. It's based on Q-Q plot, which is a quantile-quantile comparison of two distributions by plotting their quantiles against each other. The procedure is to compute a quantity from each multivariate observation, such that this quantity follows a known probability distribution when the data follow a multivariate normal distribution. Then a Q-Q plot of this quantity against the quantiles of the reference distribution will plot as a straight line. For each multivariate observation, compute the squared Mahalanobis distance between that observation and the sample mean vector:

$$D_i^2 = (X_i - \bar{X})' S^{-1} (X_i - \bar{X}), i = 1, \dots, n,$$

Where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad and \qquad S = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$$

D^2 is the multivariate analog of the square of the standard score (z-scores) for a single variable, z^2, which measures the distance from the mean in standard deviation units. D^2 measures the distance from the mean vector in relation to the variance covariance matrix, which takes into account the precision of the variables as well as their inter-correlations. If the observations $X_i's$ indeed follow a p-dimensional multivariate normal distribution, the distances D^2 can be approximated by a chi-square χ^2_p with p degrees of freedom (Khattree and Naik 1999). Therefore, a Q-Q plot of the ordered distance values $D^2_{(i)}$ against the corresponding quantiles of χ^2_p should approximately resemble the 45° straight line $y = x$.

By using quantiles in the chi square Q-Q plot (abbreviated as Q-Q plot following), we can visualize the behavior of each fraction (percent) of the data below the given value. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry and the presence of outliers can all be tested from this plot.

The points plotted in a Q-Q plot are always increasing when viewed from left to right. When plotted points lie on a line, but not $y = x$, it is a shifted and/or scaled normal distribution, it will agree with normal distribution after some linear transformation. If the general trend of the Q-Q plot is flatter than the line $y = x$, the distribution is more compressed than normal distribution. Conversely, if the general trend of the Q-Q plot is steeper than the line $y = x$, the distribution is more dispersed than normal distribution. Q-Q plots are often arced, or "S" shaped, indicating that the distribution is more skewed than normal distribution, or it has heavier tails than normal (Chambers, Cleveland et al. 1983). Skewness of the data is a common type of deviation from normality.

An efficient program using R─"rqqchi2" to plot the above Q-Q plot is provided by Ruey S. Tsay (Tsay 2006).

2.2.2 Violin plot

Many different statistics and graphical tools summarize the characteristics of the experimental data. Descriptive statistics provide information about location, scale, median, symmetry, and tail thickness. Other statistics and graphs investigate extreme observations (outliers) or study the distribution of data values. Commonly used diagrams in data analysis, such as stem-leaf plots, dot plots, box plots, histograms and probability plots give information about the distribution of all the observed values. Violin plots introduced by Hintze and Nelson present us an efficient tool for data analysis and exploration in systems biology study (Hintze and Nelson 1998).

The violin plot, as depicted in Figure 2.1, and implemented in R 2.11 statistical and graphical software, combines the box plot and the density plot into a single diagram. The name violin plot originated because one of the first analyses that used the envisioned procedure resulted in a graphic with the appearance of a violin (Hintze and Nelson 1998). Box plots show four main features about a data set: center, spread, asymmetry, and outliers. The density plot supplements traditional summary statistics by graphically showing the shape of the distribution, which can be important to reveal asymmetry and multimodality (Gentleman, Hahne et al. 2006). In a violin plot, the density plot is plotted symmetrically to the left and the right of the vertical box plot. There is no difference in these density plots other than the direction in which they extend. Adding two density plots gives a symmetric plot which makes

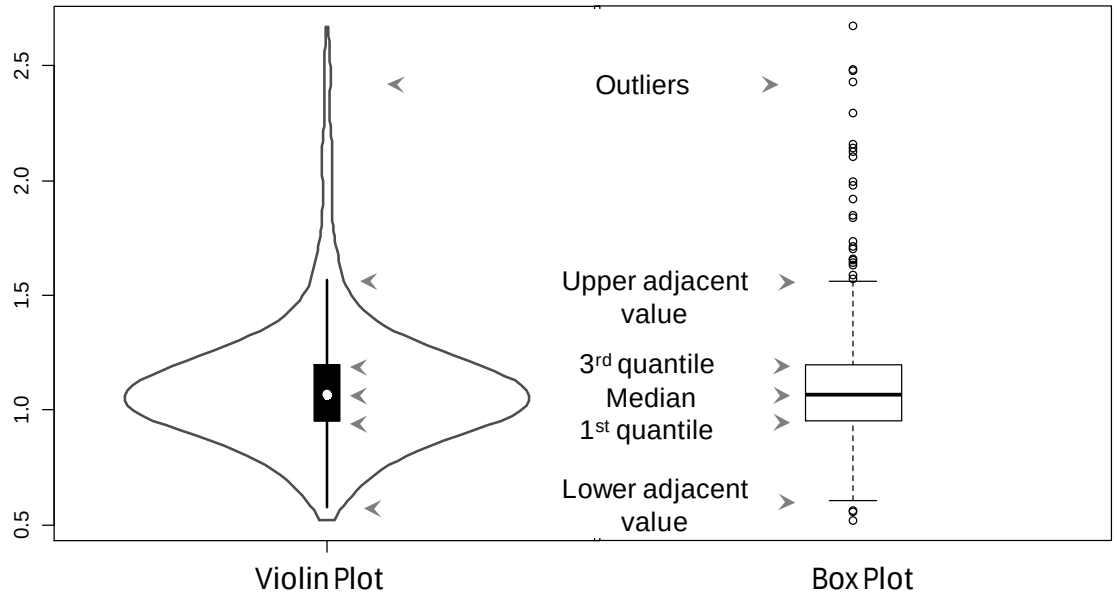


Figure 2.1: Common components of violin plot and box plot (plotted in R2.11). The data are responses of metabolites to Methyl Jasmonate (MeJa) elicitation obtained with CE-MS analyses in *Medicago truncatula* study.

it easier to see the magnitude of the density. This hybrid of the density plot and box plot allows quick and insightful comparison of several distributions.

As an example, Figure 2.1 shows the violin plot and box plot side by side for the responses of metabolites to the elicitation of Methyl Jasmonate (MeJa) in CE-MS analyses of *M. truncatula* project. The labels in the diagram identify the principal lines and points which form the main structure of the traditional box plot diagram. The lower adjacent value and the upper adjacent value at the end of the whiskers represent the lowest datum still within 1.5 IQR (Inter-quantile Range) of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile. As shown, the violin plot includes a box plot with two slight modifications. First, a circle replaces the median line which facilitates quick comparisons when viewing multiple groups. Second, outside points, which are traditionally classified as mild and severe outliers, are not identified by individual symbols (Hintze and Nelson 1998).

A built-in function in R—"vioplot" (Adler 2005) will be used for violin plotting. It starts with a box plot, and then adds a rotated kernel density plot to each side of the box plot.

2.3 Results and discussion

High throughput comprehensive 'omics' data partially symbolize systems biology study. Yet, the quality of the data is largely confined by the current available analytical technologies, and potentially bottlenecks the advancement of system biology.

M. truncatula project exposes the typical issues present in the current 'omics' data. First they are all relative values, rather than absolute quantifications; some are heavily processed such as the gene expression data, which severely hinders inter-experimental comparison and integration. Second, the missing values are especially prevalent in proteomic data. In MeJa elicitation experiment, 64% of the protein data are missing, while in yeast elicitation experiment, 37% data are missing, which presents a formidable challenge even to the most robust data analysis method. Another issue of interest is the distribution of the 'omics' data. Microarray data have been log-transformed and normalized (to isolate system error) to be forced to 'normal', but have they? The metabolomic and proteomic data are relatively new and their data distribution and characteristics are scarcely described. Here I present research that attempts to study this issue for these two comprehensive projects.

Both chi square Q-Q plot and violin plot of gene expression levels responding to MeJa-elicitation reveal the distribution to be skewed to the right (Figure 2.2). Chi square Q-Q plot shows the majority of the data fall on the line, while the right end of the pattern (upper tail) curves above the line. This suggests that the data distribution has a long tail at the right end, *i.e.* right skewed. Microarray data usually show that most genes are expressed at very low copy number and few genes are expressed at high levels. The Q-Q plot shows that log-transform has worked to some degree; presumably the distribution of the original "raw" signal intensity would be more skewed.

The violin plots of all types of 'omics' data typically display a "Hershey's kisses" shape or flat-bottomed teardrop shape (Figure 2.2). Violin plots provide a better indication of the shape of the distribution than the box plot. This includes the ability of detecting bumps or clusters in data. The bumps close to the bottom of "Hershey's kisses" indicate the mass of the distribution is on the left of the density plot; most of the 'omic' molecules have a low relative response level and they have relatively few high values.

Q-Q plots of the metabolomic and proteomic data following MeJa elicitation (Figure 2.2) both suggest that their data distributions are skewed to the right, where Q-Q plot of GC-MS analyses shows a curved pattern with slope increasing from left to right and Q-Q plot of proteins shows a general trend of a flatter line than the reference line.

Investigation of profiled metabolites with different technologies (Figure 2.3) indicates GC-MS and LC-MS analyses have distributions with positive skew (right-skewed) and data are

compressed at low level. The exception is the data from CE-MS analyses, which shows a squeezed violin shape in the violin plot. Among all the violin plots in this study, only the plots for the data from CE-MS analyses display bell-shaped distribution. The reason for this may lie in the data normalization procedure where relative peak areas are determined by taking the peak ratio of each component to that of the internal standard, ethionine (Williams, Cameron et al. 2007).

In *M. truncatula* study, ratios between elicitor and control data are chosen to represent each molecule's response to elicitation. Q-Q plots of the original elicitor data and ratio data illustrate the effect of this ratio transformation (Figure 2.4). The plot of the original elicitor data shows an obvious "S" shape; while in the plot of the ratio data, the degree of the curvature is clearly subdued and the overall trend of the data points is closer to the reference line. This result indicates that ratio transform brings the extreme high and low values to the middle ground by replacing the absolute quantities with relative values. It has justified the choice of ratio transformation in the *M. truncatula* project.

A kinetic model of the yeast glycolysis and glycerol biosynthesis pathway (Pritchard and Kell 2002; Teusink and Westerhoff 2000; Martins, Camacho et al. 2004) used with the simulation software COPASI to simulate a change in medium glucose and corresponding changes in intermediate metabolite levels provides a small simulated metabolomic profile (19×9) that can be analyzed for comparison. This data set contains no measurement error, and all the changes in metabolites can be attributed to the metabolic interactions (in the model). The violin plot of this profile displays a similar "Hershey's kisses" shape but with a less compressed lower values. Interestingly, the upper end of whisker shows a slight bump, which indicates a small degree of cluster caused by the accumulation of glycerol.

All the *M. truncatula* MeJa-elicitation data are right skewed, among them, CE-MS data is the least skewed and least spread, LC-MS-saponin is the most skewed and most spread. Other *M. truncatula* MeJa-elicitation data sets, from the least spread to most spread distribution are: mRNA, LC-MS-flavonoid, GC-MS non-polar, protein, and GC-MS polar.

All the *M. truncatula* yeast-elicitation data are also right skewed. From the least spread to most spread distribution, these data sets are: CE-MS, mRNA, GC-MS polar, LC-MS-flavonoid, and GC-MS non-polar.

V. vinifera drought stress data are less skewed compared to *M. truncatula* data. GC-MS and protein data sets display scaled normal distribution with outliers. mRNA data is slightly right skewed. The spreadness from low to high are: GC-MS, mRNA, and protein.

V. vinifera salinity stress data are less skewed as well. GC-MS data shows scaled normal distribution with few outliers. Both protein and mRNA data are right skewed, and protein data

shows a lower degree of skewness. Their spreadness from low to high are: GC-MS, protein, and mRNA.

Although chi square Q-Q plot is a quick and efficient way to detect data normality, it has its drawbacks. Interpreting Q-Q plots is more a visceral than an intellectual exercise. In the Q-Q plot, potential outliers appear as points in the upper right which are substantially above the line for the expected chi square quantiles. Unfortunately, like all classical (least squares) techniques, the Q-Q plot for multivariate normality is not resistant to the effects of outliers. A few discrepant observations not only affect the mean vector, but also inflate the variance covariance matrix. Thus, the effect of the few wild observations is spread through all the squared Mahalanobis distance.

When Q-Q plots indicate departures from normality (data skewed and heavy tailed), transformation may be considered. Logarithmic transformation has been widely used in microarray data processing. Although transformations are recommended as a remedy for outliers and breaches in normality, we must be aware of the change to the data. Studies have shown that log transform tends to make the variability more constant, resulting in loss of information associated with low channel intensities (Sharov, Kwong et al. 2004; Smyth, Yang et al. 2002).

Classical parametric multivariate methods (principal component analysis, multivariate regression, canonical correlation, discriminant analysis, etc.) are based on the parameters of populations or probability distributions, such as sample mean vector and sample covariance matrix. Mean vector and covariance matrix are optimal if the data come from a multivariate normal distribution but they are very sensitive to outlier observations and lose in efficiency in the case of heavy tailed distributions. Therefore robustified methods that are based on parametric test statistics in which the estimates of the parameters like the mean or the standard deviation are replaced respectively by robust estimates like the trimmed mean and the median absolute deviation (MAD) (Büning 2000), and non-parametric methods that require fewer assumptions about a population or probability distribution are good candicates in most cases of systems biology data.

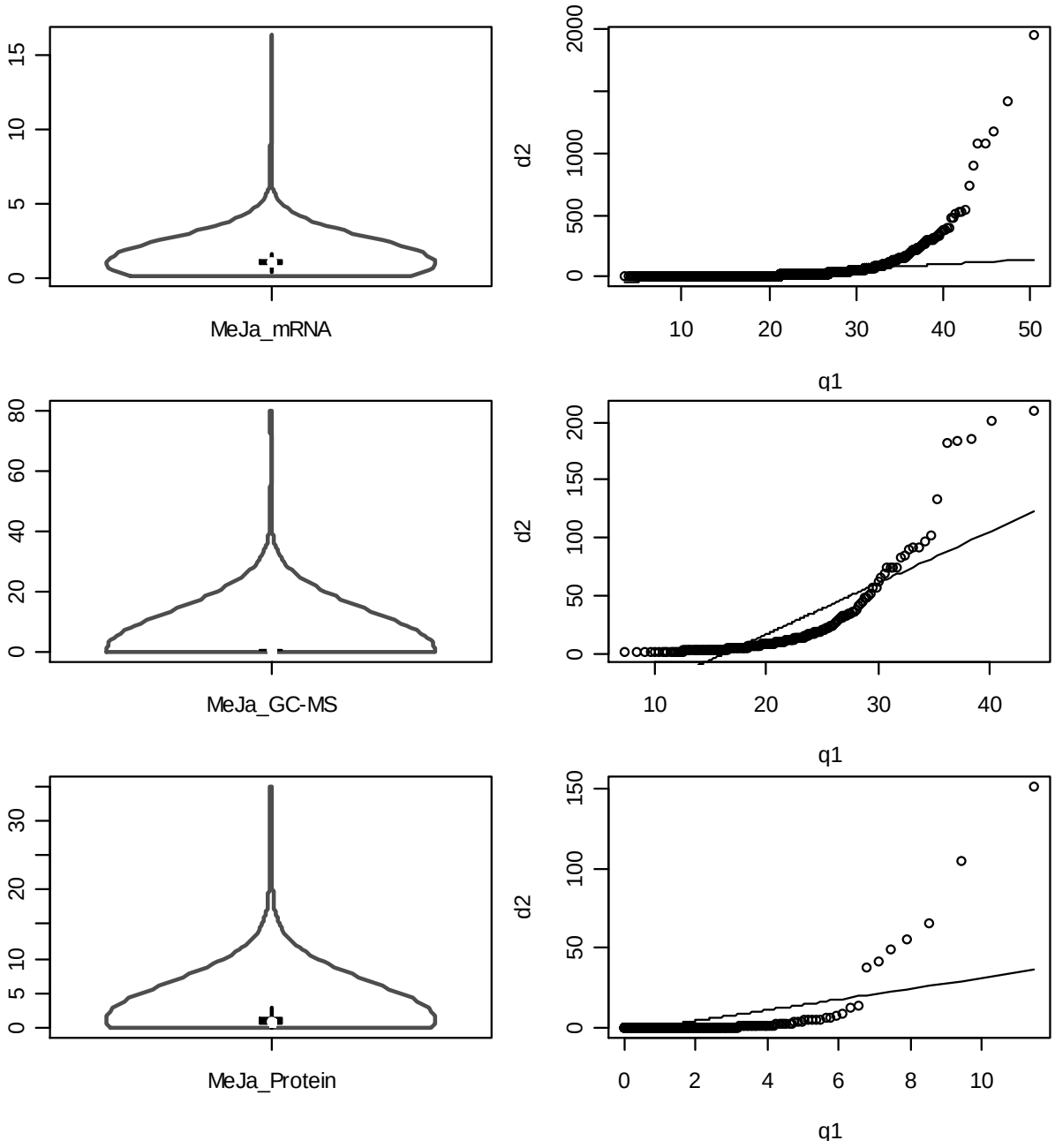


Figure 2.2: Violin plots and Chi square Q-Q plots for three types of 'omics' data in *M. truncatula* study (Plotted in R2.11). The data are produced from MeJa elicitation experiment. The first row represents the gene expression level responding to MeJa elicitation (50% genes, a total of 8000 is shown here due to the memory limit in R); the second row represents polar metabolites' responses to the elicitation in GC-MS analyses (ratio values between elicitor and control sample); the third row represents protein level at 24 hrs following elicitation.

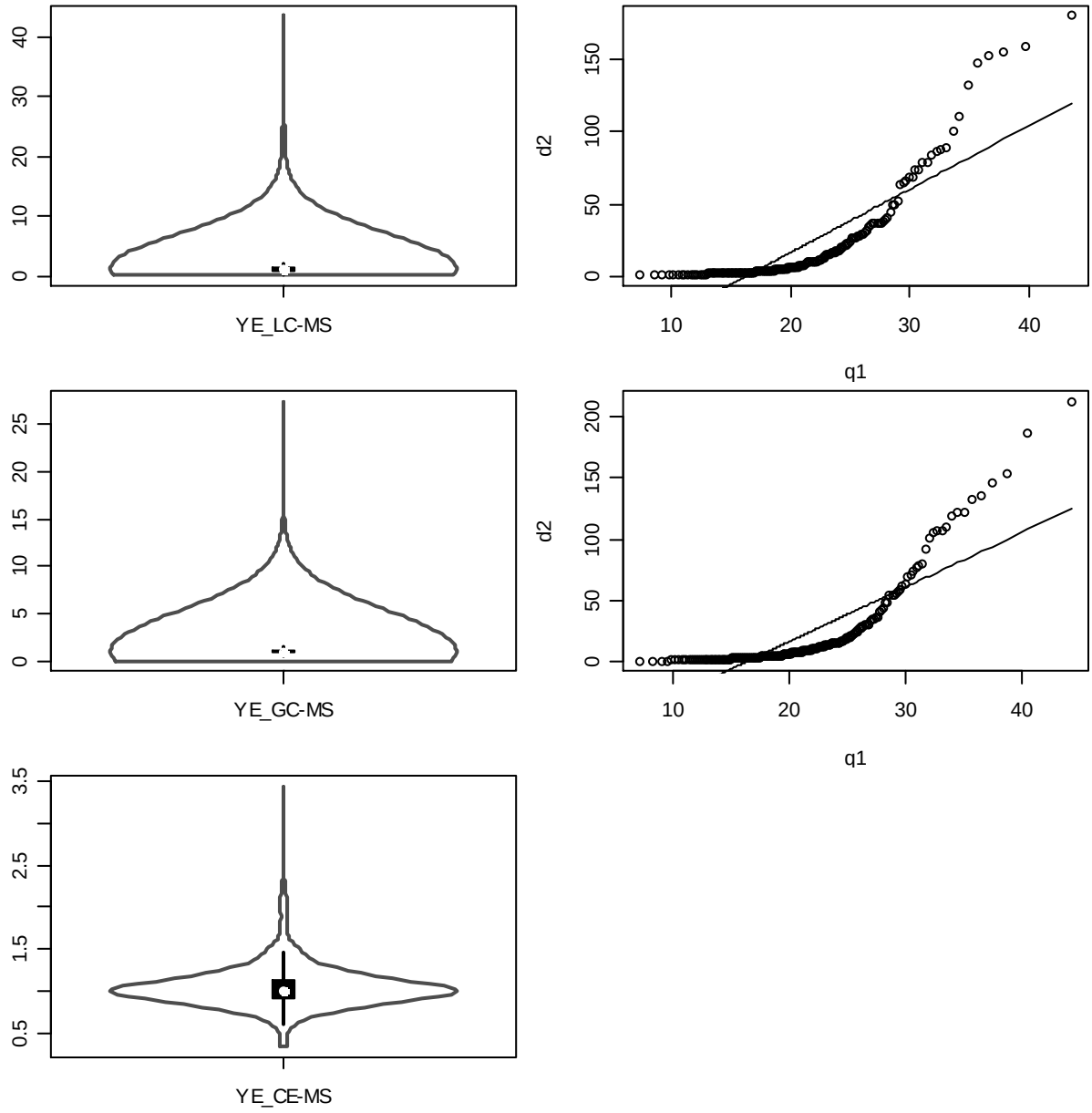


Figure 2.3: Exploring data produced with different metabolomic technologies using Violin plots and Chi square Q-Q plots. The data are metabolites' responses to yeast elicitation in *M. truncatula* project. The first row represents metabolites' relative responses in LC-MS analyses using flavonoid as internal reference; the second row represents polar metabolites' responses in GC-MS analyses; the third row represents metabolites (most are amino acids) in CE-MS analyses using ethionine as internal reference. All the values are ratios between elicitor and control sample. The CE-MS analyses data are singular.

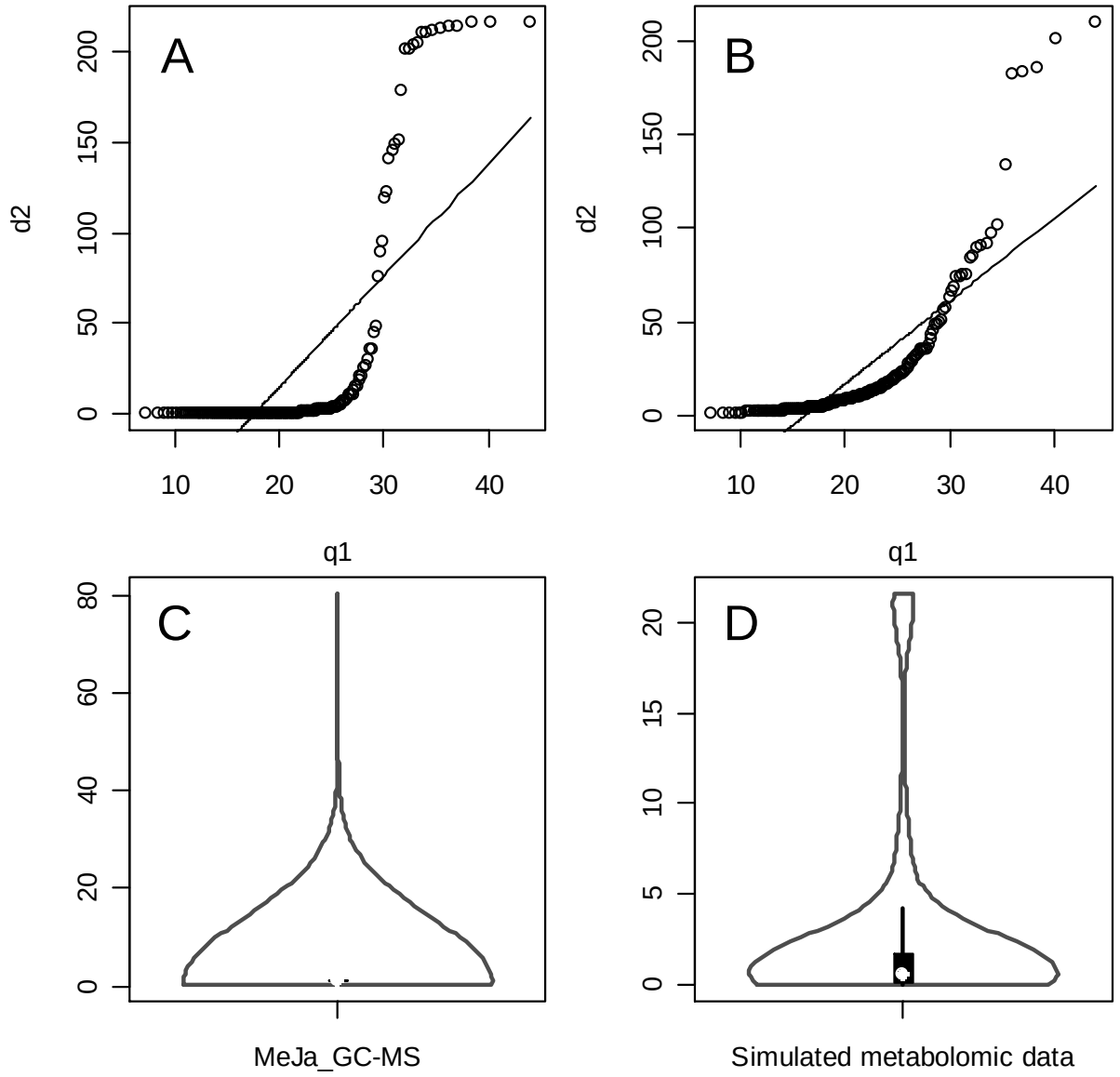


Figure 2.4: Inspection of empirical ratio data and simulated metabolomic data with Chi square Q-Q plots and Violin plots. The data of plots A, B and C are produced from the same experiment, GC-MS analyses of polar extracts elicited with MeJa in *M. truncatula* project. A: Chi square Q-Q plot for the original elicitor data. B: Chi square Q-Q plot for ratio values between elicitor and control samples. C: Violin plot for ratio values between elicitor and control samples. D: Violin plot for simulated metabolomic data of a yeast glycolysis and glycerol biosynthesis model created with Copasi.

Chapter 3

Biplot display: a visualization tool to provide insight into systems biological data

3.1. Abstract

Time-resolved systems biology data are becoming increasingly available, providing a good opportunity to uncover the molecule-molecule relationships and to characterize the dynamic properties of the underlying molecular networks of various biological processes. Comprehensive visualization methods for exploring multivariate data are needed to gain insight into the physiological processes reflected in these molecular profiles. The biplot display is particularly useful for the visualization of these data as both molecules and samples can be plotted together in a low dimension plot. With the biplot a vast amount of information about a set of experiments and their (high dimensional) results can be inferred visually from a single plot. This is particularly important in time course experiments where analysis of the biplot allows for a rapid identification of a) which parts of the time course are most related with each other, b) which molecules have similar patterns in the time line, and c) in which regions of the time course do each molecules peak.

In this work, we demonstrate the utility of biplots by applying them to singular value decomposition (SVD) analysis of metabolomics time course data. We propose that the biplot as a useful tool to reveal inter-molecule relationships and to single out the unidentified molecules that appear to have highest significance in specific experiments, and thus are good candidates for identification. Most importantly, the biplot conveys a system perspective on the dynamic observed in omics experiments.

3.2. Introduction

3.2.1 Visualization tools in systems biology

Advances in system biology have produced ever-increasing amounts of complex data that need to be analyzed. Visualization is an effective analytical technique that exploits the ability of the human brain to process large amounts of data. Visualization involves conscious decision about what message should be conveyed in a particular image and what methods are likely to achieve that goal easily and accurately (Gentleman, Hahne et al. 2006). In current high-throughput biological data analysis, visualization can be categorized in terms of their objectives as diagnostic tools or exploratory tools (Figure 3.1).

Diagnostic tools are mainly for examining raw data, comparing normalization methods, and providing information for choice of exploratory analysis methods. Some commonly used

methods are box plots, density plots and MA-plots. Box plots and density plots are used for plotting and comparing distribution of different data sets. Box plots show five main features about a data: the median, the upper and lower hinges (quartiles), and the extremes (McGill, Tukey et al. 1978). Density plots illustrate the shape of the distribution. It highlights the peaks, valleys and bumps in the distribution (Hintze and Nelson 1998). Violin plots are the synergistic combination of box plots and density plots, which have been detailed in Chapter 2. MA-plots have been developed and applied in microarray data analysis (Dudoit, Yang et al. 2002), also called Ratio-Intensity (RI) plots (Cui, Kerr et al. 2003; Sharov, Kwong et al. 2004). In microarray experiments, there are many sources of systematic variation. MA-plot is especially helpful for the detection of the intensity-dependent effects in the log-ratios and the effectiveness of normalization (Park, Yi et al. 2003; Yang, Dudoit et al. 2002). An MA-plot is a scatter plot of log-intensity ratios (M-values) versus log-intensity averages (A-values), where $M = \log_2 R - \log_2 G$ and $A = \frac{1}{2}(\log_2 R + \log_2 G)$. Data points from similar hybridizations will be centered on the $M = 0$ axis. When the scattered data points have a tendency to spread towards diagonal direction, it indicates that there is a correlation between the total intensity of a spot and its ratio.

Once raw data has been processed and diagnosed, exploratory tools will be employed to investigate data in order to discover interesting patterns, regularities or irregularities. The goal is to gain insight into the structure of the data without imposing conditions upon them. Most methods rest either on pair wise distance measure or dimension reduction. Since distance measure is one way to quantify the similarity or dissimilarity of objects, it provides a foundation for many clustering methods, such as hierarchical clustering, *k*-means clustering and self organizing map (Rencher 2002), which have been discussed in Chapter 1.

In cluster analysis we search for patterns in a data set by grouping the multivariate observations into clusters. Although we hope to find the natural groupings in the data, i.e. groupings that make sense to the researcher, it still could be strained. The biological system's concerted response to biotic or abiotic stress is far more complicated than can be described by these groupings. Dimension reduction methods, on the other hand, are concerned only with the core structure of a sample of observations on p variables. None of the variables is designated as dependent, and no grouping of observations is assumed (Rencher 2002).

Given a $n \times p$ multivariate data matrix obtained from systems biology study (Figure 3.1), n represents the number of elements, objects or molecules and p stands for the number of feature variables, experimental conditions or time points the study has taken. Normally molecules are the objects that we want to measure the distance. In a molecule distance matrix D, $D_{(i,j)}$ element of distance matrix is the distance between molecule i and j, which results in a $n \times n$ distance matrix. Likewise should distances between experimental conditions or time points be needed, a $p \times p$ distance matrix will be produced. While it's not feasible to portrait

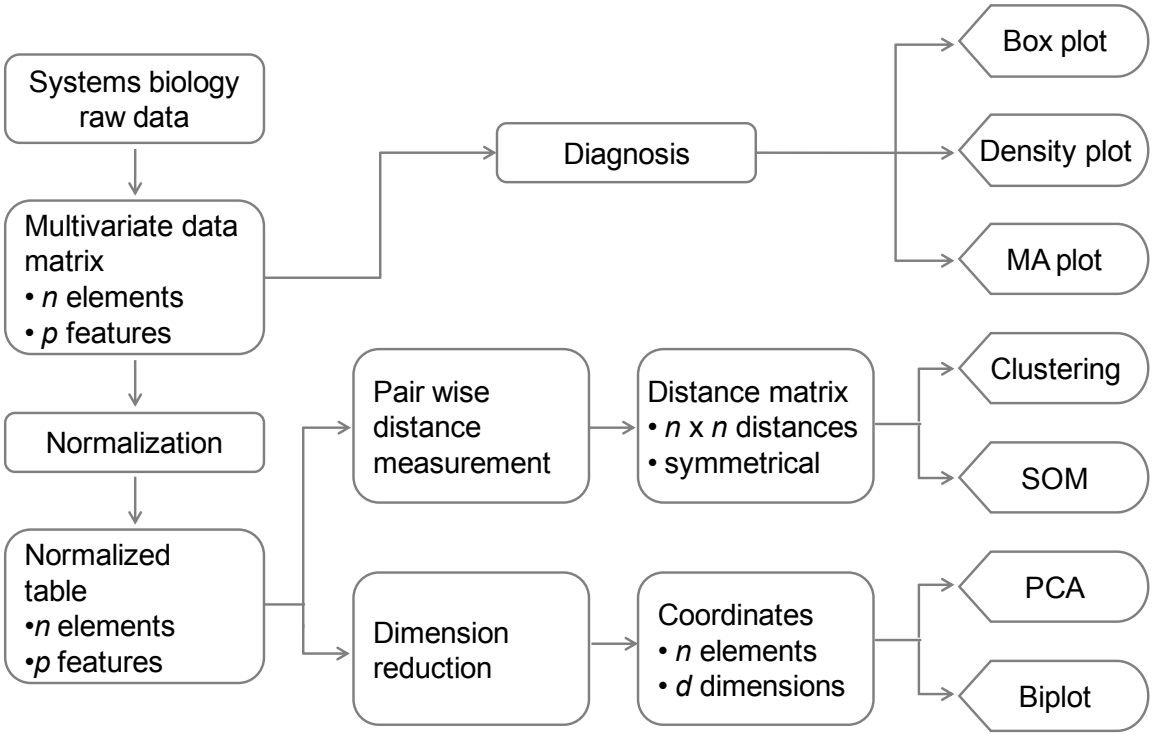


Figure 3.1: Visualization tools in systems biology multivariate data study. This flow chart shows the categories of visualization tools and the corresponding data processing steps in high-throughput biological data analysis. In this chart, visualization tools are categorized as diagnostic tools and exploratory tools. Diagnostic tools include Box plots, Density plot and MA plot. Exploratory methods rest on either distance measure or dimension reduction. The former comprises Clustering methods and Self organizing map (SOM); the latter includes Principal component analysis (PCA) and Biplot analysis.

the whole original data in Cartesian coordinate system, we can envisage each molecule as a point in a *p*-dimentional feature space. To accommodate human visual perception, dimension reduction methods have been developed to use a small number of dimensions (two or at most three) to represent the original high dimensional data.

The increasingly available time-resolved data are valuable in systems biology study. It provides a good opportunity to uncover genes-proteins-metabolites relationships and to characterize the dynamic properties of the underlying molecular network. To analyze these time resolved datasets, many analysis and visualization methods used in the transcriptomics and proteomics are directly applicable to metabolomics and are already in use, such as clustering and principal

components analysis (PCA) (Mendes 2002). While PCA has been a popular method in gene expression data and metabolomics data analysis, the essence of PCA, singular value decomposition (SVD) and SVD based biplot (Gabriel 1971, Cox and Gabriel 1980), has not yet been utilized in systems biological data study.

Biplot display is a commonly used multivariate method for graphing row (observations, i.e. metabolites in this case) and column (samples, i.e. time points in time course data) elements simultaneously using a single display. Biplot is mainly used in ecology(Barker, East et al. 1986), genetics (Yan and Hunt 2002) and biomedical (Halling, Fridh et al. 2006) research. Only recently have a few applications been found in microarray data, such as gene expression data of bone marrow samples from different type of leukemia patients (Pittelkow and Wilson 2005; Pittelkow and Wilson 2003), gene expression patterns of *Arabidopsis thaliana* leaves subjected to different treatments related to plant defense (Chapman, Schenk et al. 2001). However, no literature has shown biplot display on any time-resolved experimental data.

The biplot display provides a method for reducing the dimensionality of the systems biological data and displaying the molecules and time points jointly on the same plot. Similarities between quantities of molecules or patterns of time points may be gleaned from these types of plots. Relationships between molecules and time become obvious and can be identified instantly.

3.2.2 Biplot display

The results of systems biology experiments, can be organized in a multivariate data matrix Y, with n rows and p columns, where n is the number of the molecules, and p is the number of time points or experimental conditions. A biplot is a two-dimensional representation of the data matrix $Y_{(n\times p)}$ by means of row markers $a_1, a_2, \ldots a_n$ and column markers $b_1, b_2, \ldots b_p$. The biplot carries one marker/point for each row (observation, or molecule in this study), and one marker/point for each column (variable, or time point in this study). The principle of biplot display of matrix Y is that element $y_{i,j}$ in the i-th row and j-th column is represented by the inner product of the i-th row marker and the j-th column marker of AB^T (Cox and Gabriel 1981). The prefix "bi" refers to the two kinds of points; not to the dimensionality of the plot (Rencher 2002). The method presented here could, in fact, be generalized to a three-dimensional (or higher-order) biplot, namely bi-model because it too is a joint display of both rows and columns: the ending "model" indicates that there are three dimensions (Cox, Gabriel et al. 1981). Biplots were introduced by Gabriel (1971) and have been discussed at length by Gower and Hand (Gower and Hand 1996).

Since the biplot is planar, the row markers a_i, as well as the column markers b_i, are plotted in the plane; thus a rank 2 approximation has to be used for any matrix of rank greater than 2 (Cox,

Gabriel et al. 1981). Singular value decomposition (SVD) can be used to approximate any rectangular matrix by a matrix of same size but of lower rank, such that the sums of squares of the differences between the elements of the matrix and its approximation is minimized (Pittelkow and Wilson 2003). If we have an n by p matrix Y of rank r with $r \le p \le n$, then the SVD of Y can be written as $Y = U\Lambda V^T$, where U (n by p) and V (p by p) are matrices of singular vectors and Λ (p by p) is a diagonal matrix of singular values. U, named the left singular vectors of Y, is the matrix with columns corresponding to the p orthogonal eigenvectors of YY^T. V, named the right singular vectors of Y, is the orthogonal matrix corresponding to the eigenvectors of Y^TY. Λ =diag(λ_1, λ_2, …, λ_p) with $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_r \ge 0 = \lambda_{r+1} = \lambda_{r+2} = ... = \lambda_P = 0$. The singular values are the positive square roots of the eigenvalues of Y^TY.

Typically the matrix is approximated using the first few dominated values and vectors. The biplot display is the plot of the row (metabolites) markers G and column (time points or experimental conditions) markers H where

$$G = U_{(k)}\Lambda^{\alpha}_{(k)}$$

$$H = V_{(k)}\Lambda^{1-\alpha}_{(k)}$$

The value of k determines the dimension of the approximation (k = 2 in biplot, k = 3 in bi-model). The parameter α $(0 \le \alpha \le 1)$ can be specified to determine whether emphasis is placed on the rows or columns of Y. The matrix Y is then approximated by $\hat{Y}_{(k)} = GH^T = U_{(k)}\Lambda^{\alpha}_{(k)}\Lambda^{1-\alpha}_{(k)}V^T_{(k)} = U_{(k)}\Lambda_{(k)}V^T_{(k)}$. Plots of the coordinates associated with G superimposed over the coordinates associated with H form the biplot display (Lipkovich and Smith 2002). To enhance interpretation, lines or arrows will be drawn from the center to the coordinates associated with H.

Although any value from 0 to 1 can be used for α, three are commonly used, 1, $1/2$, and 0. When the value 1 is selected, the result is called a JK or RMP (row metric preserving) biplot. In this display the distances between pairs of rows is preserved (after any centering and scaling is performed). The Euclidean distance between two row markers is the approximation of the distance between the corresponding points (rows) in the data matrix Y. This display is useful for studying the relationship between objects (molecules in systems biological data matrix). When the value 0 is selected, the result is a GH or CMP (column metric preserving) biplot. The display preserves distances between columns and is useful for interpreting variance and relationship between variables (time points or experimental conditions in this case). When columns are not standardized, the length of the vector represents the standard deviation of the variable. The cosine of the angle between the lines (arrows) drawn to each pair of column markers approximates the correlation between the two corresponding variables (time points). Thus a

small angle between two vectors indicates that the two variables are highly correlated, two variables whose vectors form a 90° angle are uncorrelated, and an angle greater than 90° indicates that the variables are negatively correlated (Cox, Gabriel et al. 1981). The other value of α, $1/2$, gives equal scaling or weight to the rows and columns. It is useful for interpreting interaction in two factor experiments (Gower and Hand 1996). Since the ability to estimate distance between molecules and angle between time points is important for systems biological time course data, equal scaling on both rows and columns ($\alpha = 1/2$) will be adequate with no further adjustment.

Biplots are the multivariate analog of scatter plots (Gower and Hand 1996). Instead of x and y axes, biplot has p axes that correspond to p time variables (or sample variables). For all biplots, the values of the n observations in the ith variable vector (corrected for means) are related to the perpendicular projection of the point on the vectors from the origin to the points representing variables. The further from the origin a projection falls on an arrow, the larger the value of the observation on that variable. Hence the vectors will be oriented toward the observations that have larger values on the corresponding variables (Rencher 2002).

The singular value decomposition can be used as a basis for many multivariate graphical techniques, including principal component analysis (PCA) and correspondence analysis. The latter is a descriptive/exploratory technique designed to analyze two-way frequency cross tabulation tables. Before singular value decomposition is applied to PCA, some transformations usually will be made to the data. Depending on the type of the transformation we obtain different data matrices for PCA: when column centering is performed, $y_{ij}^{*} = y_{ij} - \bar{y}_{\bullet j}$, the biplot is based on the PCA of the covariance matrix; when column centering and standardization is performed, $y_{ij}^{*} = \frac{1}{\sqrt{n-1}}(y_{ij} - \bar{y}_{\bullet j})/S_j, \quad S_j = \sqrt{\frac{\sum_{i-1}^{N}(y_{ij} - \bar{y}_{\bullet j})^2}{n-1}}$, the biplot is based on the PCA of the correlation matrix; when rows and columns centering is performed, $y_{ij}^{*} = y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet}$, SVD is performed on interaction residual matrix. From the third transformation, we can construct biplot for multiplicative interaction diagnostics in two-way tables.

To improve interpretability of biplot for time course data, it is useful to draw lines joining all the points in the time sequence. We refer to these augmented biplots as phase-space biplots because they are akin to phase space diagrams. In phase-space biplots each time point represents a state of the corresponding biological system at that time. The plot shows the trajectory of the system after perturbation and the tendency to reach new equilibrium (a stationary state).

3.3. Methods

3.3.1 Kinetic model of yeast glycolysis and glycerol biosynthesis pathway

The application of biplot display was carried out using metabolic time-course data of a computer-simulated yeast glycolysis and glycerol biosynthesis pathway. The simulations were carried out with the COPASI software(Hoops, Sahle et al. 2006) version 4.0.18 on a Pentium® 4 CPU 3.40 GHz computer (Dell Corp., Round Rock, TX) running Windows 2000 (Microsoft Corp., Redmond, WA). This Copasi model (Figure 3.2) by Martins *et al*. (Martins, Camacho et al. 2004) has combined and extended two previous models of yeast metabolism: a glycolytic model created by Teusink *et al*. (Teusink and Westerhoff 2000) and modified by Pritchard *et al.* (Pritchard and Kell 2002), and a model of glycerol biosynthesis by Cronwright *et al*. (Cronwright, Rohwer et al. 2002). By merging these two models, the fixed rate step that represents glycerol synthesis in the glycolytic model was replaced by the full model of Cronwright *et al*. In addition, Martins *et al* substituted the other fixed rate steps in the glycolytic model (trehalose, succinate, and glycogen branches) by first order reactions, and explicitly added glycerol transport in the model. This model represents the exponential growth phase with high external glucose concentration (Martins, Camacho et al. 2004).

Figure 3.2 illustrates this model of glycolysis and glycerol biosynthesis pathway. The boxed area indicates the perturbation of reducing glucose in the medium. The model contains 49 state variables: 19 internal metabolites (red colored text) and 30 external metabolites (with fixed concentrations). Two internal metabolites NADH and AMP are restricted by moiety conservation (orange colored text).

In order to simulate a metabolic perturbation, glucose concentration was reduced from 50 mM to 5 mM to effect the shifting of cells to a low glucose medium. The duration time was set as 4 minutes and samples were taken at every 0.5 minutes. The resulting table of 9 time points and 19 metabolite concentrations was used for singular value decomposition and biplot display.

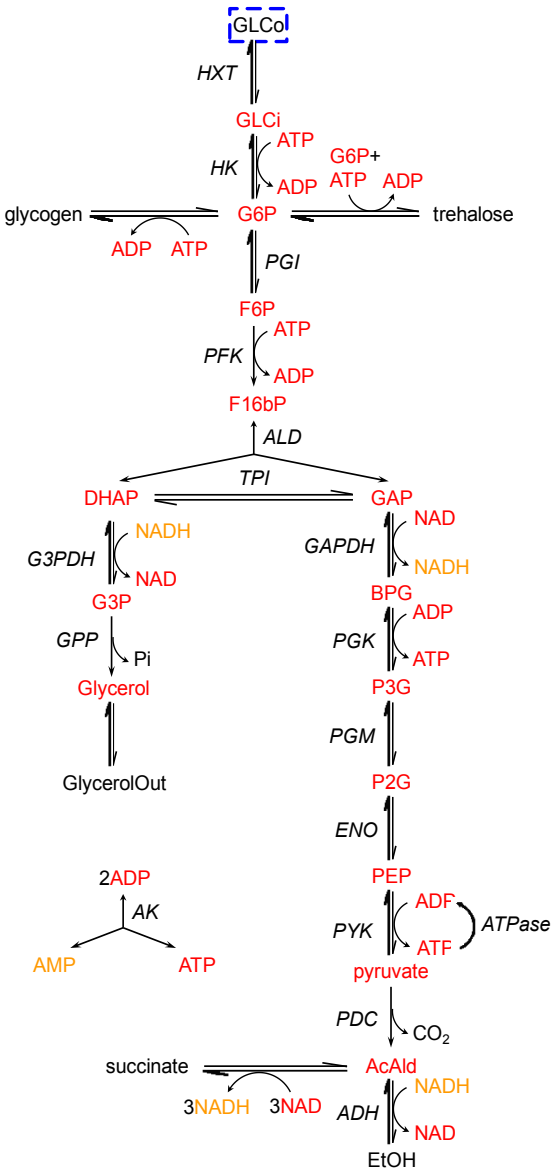


Figure 3.2: Copasi model of Yeast glycolysis and glycerol biosynthesis pathway. It's a combined and extended model by Martins *et al*. based on the work by Teusink et al., Pritchard et al. and Cronwright *et al*. (Martins, Camacho et al. 2004). The boxed area indicates the perturbation of reducing glucose in the medium. The model contains 49 state variables: 19 internal metabolites (red colored text) and 30 external metabolites. Two internal metabolites, AMP and NADH are restricted by moiety conservation (orange colored text). Abbreviations: GLCo, glucose (external); GLCi, glucose (internal); AMP, adenosine monophosphate; G6P, glucose-6-phosphate; F6P, fructose 6-phosphate; F16bP, fructose 1,6-bisphosphate; DHAP, dihydroxyacetone phosphate or glycerone phosphate; GAP, glyceraldehyde 3-phosphate; G3P, glycerol 3-phosphate; BPG, 1,3-Bisphosphoglycerate; P3G, 3-phosphoglycerate; P2G, 2-phosphoglycerate; PEP, phosphoenolpyruvate; AcAld, acetalydehyde; EtOH, ethanol; *HXT*, hexose transporter; *HK*, hexokinase; *PGI*, phosphoglucoisomerase; *PFK*; phosphofructokinase; *ALD*, aldolase; *TPI*, triosephosphate isomerase; *G3PDH*, glycerol 3-phosphate dehydrogenase; *GPP*, Glycerol-3-phosphate phosphatase; *GAPDH*, glyceraldehyde-3-phosphate dehydrogenase; *PGK*, phosphoglycerate kinase; *PGM*, phosphoglycerate mutase; *ENO*, enolase; *PYK*, pyruvate kinase; *PDC*, pyruvate decarboxylase; *ADH*, alcohol dehydrogenase; *AK*, adenylate kinase.

3.3.2 Software for generation of biplot displays

The simulated metabolite time course data after perturbation were processed by row and column centering, before applying singular value decomposition and generating the biplot display. Times and molecules were equally scaled to better identify their relationships. Two programs were used to carry out these procedures. A program written with Mathematica performed data processing, SVD, 2D and 3D biplot display. The software written by Lipkovich and Smith ─ Biplot and Singular Value Decomposition Macros for Excel© (Lipkovich and Smith 2002) also carried out the SVD and 2D biplot display. The add-in Macros for Excel allow customized editing on biplot for a better visualization.

3.3.3 Adding noise to test the robustness of biplot display

To test the robustness of biplot display and how the algorithm would withstand the presence of experimental error, noise was added to the data obtained in the COPASI simulation with a normal distribution as $N\sim(0,\mu)$, where μ is the standard deviation and ranged from 100 times to 1000 times of the smallest value in the data. The sampling of the distribution was made using R (Team 2009), ran on Windows 2000 (Microsoft Corp., Redmond, WA).

3.4. Results and discussions

3.4.1 Biplots of simulated data

After the system of yeast glycolysis and glycerol biosynthesis was perturbed, a time course data table was collected and processed. Figure 3.3A shows the biplot display for the corresponding data matrix. Time points are shown as solid circles, and metabolites are solid triangles. The center of the plot represents the data average after rows and columns have been centered. Time rays are lines drawn from the center to the time variables to help interpretation. For illustrative purposes, the distance l and angle θ are drawn and labeled. Since these 9 time arrays are akin to x and y axes in Cartesian coordinate system, we can project metabolites onto time arrays to estimate concentration of metabolites at that time. Taking F16bP as an example, a line is drawn through F16bP perpendicular to the time T0.0 array, the distance l formed in this process can be interpreted as number l on the corresponding number line. Thus, it indicates the concentration of Fructose-1,6-bisphosphate (F16bP) at 0 minute T0.0. I need to point out that l is not the true concentration of F16bP in the original time course data since the data has been processed by row and column centering. It's a proportional value for the purpose of quick comparison. Similarly

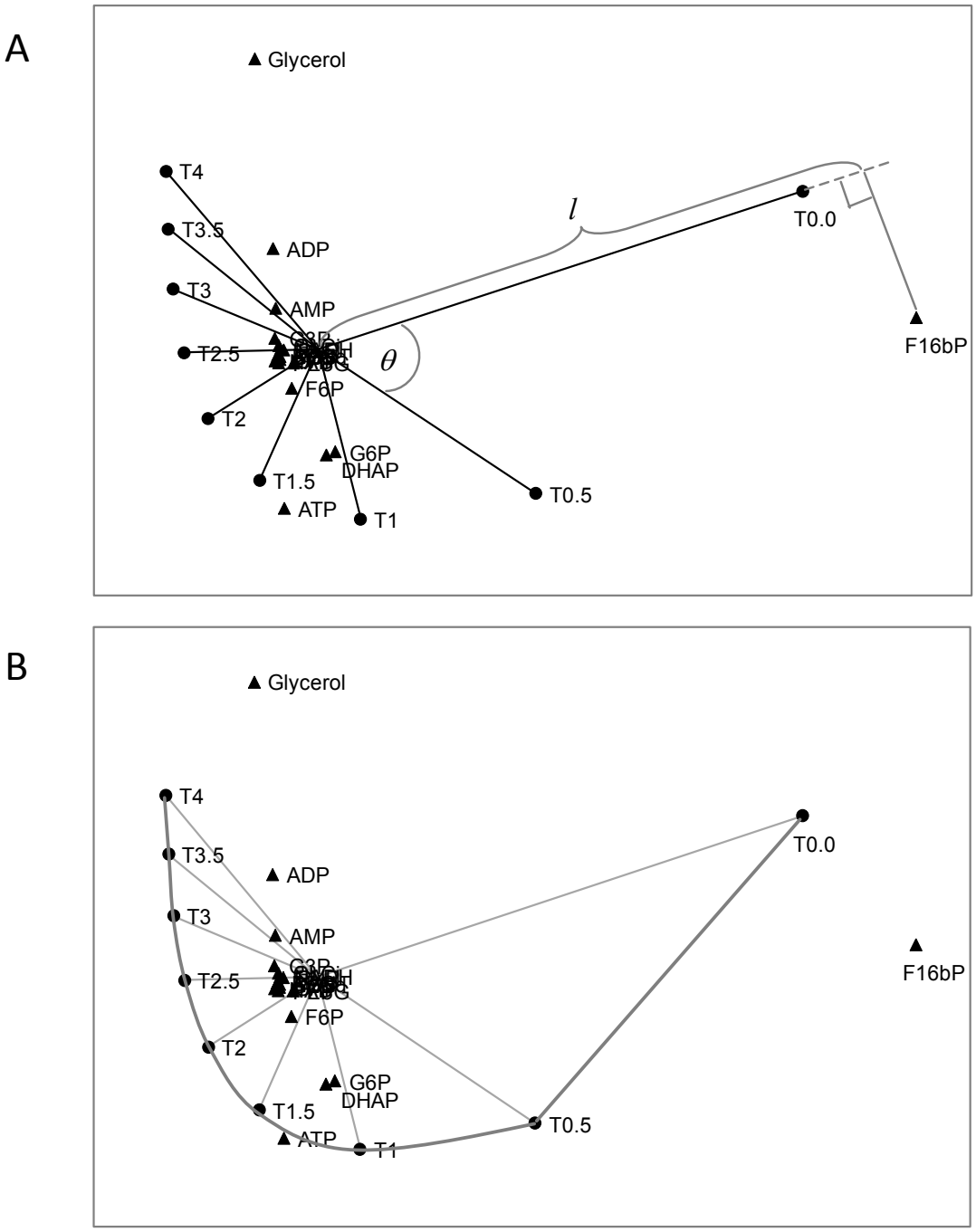


Figure 3.3: Biplot display on metabolic time-course data generated with a COPASI model (www.copasi.org) (Hoops, Sahle et al. 2006) of yeast glycolysis and glycerol biosynthesis pathway (Martins, Camacho et al. 2004). Rows and columns centering are performed before SVD. Times and metabolites are equally scaled. Solid circles represent time points from 0 minute (T0.0) to 4 minutes (T4) and solid triangles represent metabolites. A: Lines are drawn from the center (coordinates (0, 0)) to the coordinates associated with time. The number l indicates the concentration of Fructose-1,6-bisphosphate (F16bP) at 0 minute T0.0. The angle θ indicates the relatedness between 0 minute (T0.0) and 0.5 minute (T0.5). B: Phase space Biplot display of the same data in A. All the time points are joined with lines in time sequence.

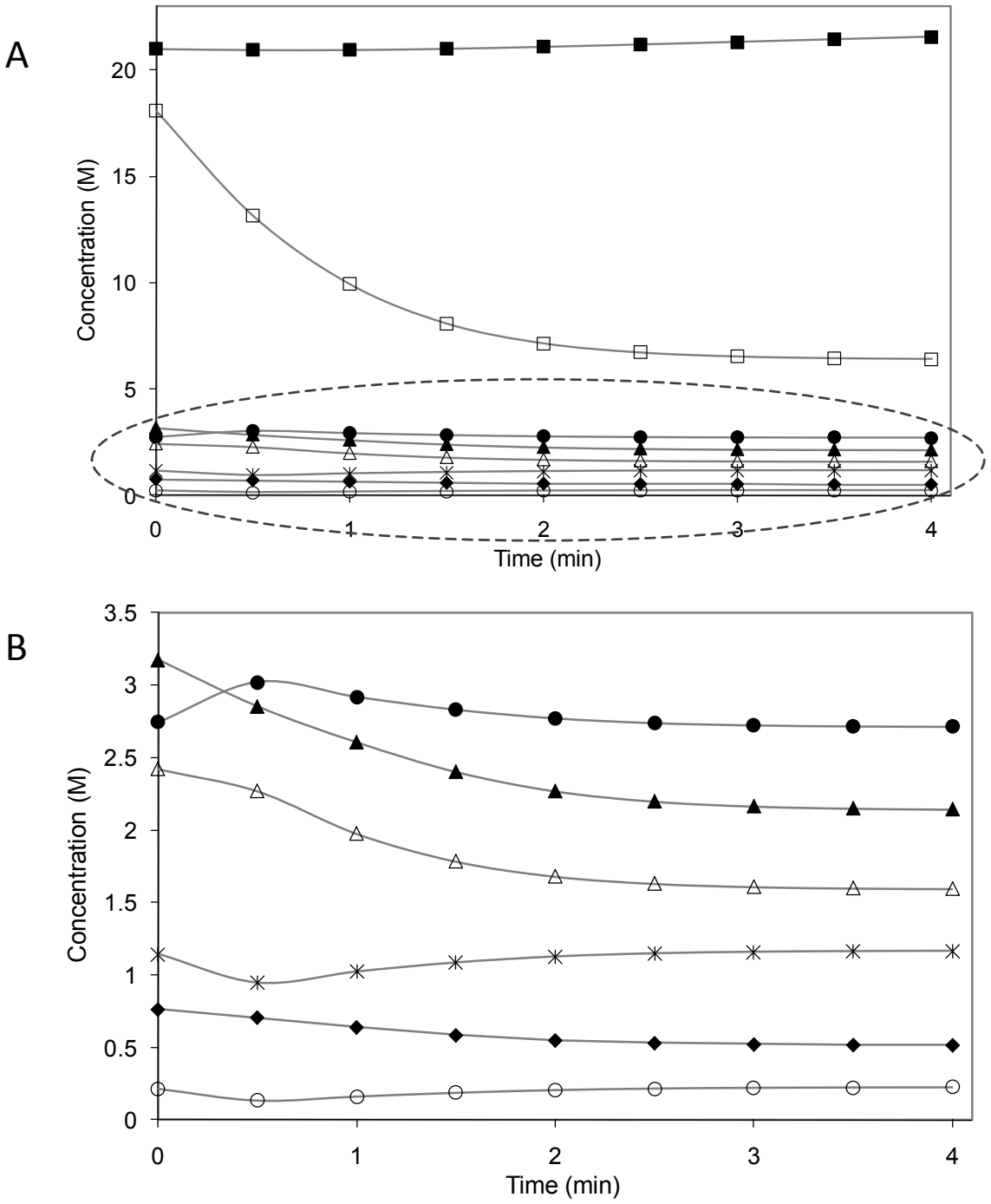


Figure 3.4: A: Time course plot of selected metabolites illustrates the dynamics of several metabolites observed in Figure 3.3A. F16bP (open squares) concentration declines over time; while Glycerol (filled squares) increases in slow motion. B: An enlarged view of the portion circled in Figure 3.4A. ATP (filled circles) decreases from the time point at 0.5 minute, meanwhile ADP (stars) and AMP (open circles) slowly increase. F6P (filled diamonds) is dropping in a slow pace. G6P (filled triangles) and DHAP (open triangles) have very similar trend—both are gradually decreasing.

the level of F16bP at each time point is approximated by drawing a line through F16bP perpendicular to each time array, and reading the points where these lines meet the axes as the numbers of these number lines. Due to the centering process, for each molecule there will be positive levels with regard to certain time points, and negative levels concerning the rest of them. Take the projections of F16bP (Figure 3.3A), it has positive projections on times T0.0, T0.5 and T1 and negative ones on times T1.5, T2, T2.5, T3, T3.5 and T4. It's found that with dominant valued molecules, the projections mostly reflect the true time course pattern accurately. For many other molecules, it's more meaningful to pay attention to their inter-molecule relationships and overall trends on biplot display.

With the projection rule, we can easily deduce that the molecules that are located around the center of the biplot have values that are close to data average (double centering); and the time arrays are oriented toward the molecules that have higher values at the corresponding times. On Figure 3.3A, F16bP is oriented adjacent to time T0.0, which suggests that F16bP has the highest value at 0 minute. The time course plot of F16bP (represented by open squares in Figure 3.4A) confirms this finding where the concentration of F16bP declines over time. Similarly glycerol is located close to time T4, which suggests it accumulates at 4 minutes; the time course plot of glycerol (marked by filled squares in Figure 3.4A) shows that it's increasing gradually and reaches its highest value at 4 min. Biplot display has reflected the glycolysis and glycerol biosynthesis network in a sense that F16bP is consumed and reducing when it's split by aldolase into two triose sugar, dihydroxyacetone phosphate (DHAP) and glyceraldehyde 3-phosphate (GAP), meanwhile glycerol is increasingly produced and transported out of the cell.

Other metabolites seen from the biplot that contribute to the formation of glycerol are G6P, F6P and DHAP. It's evident from their time series plot that all of them are reducing (Figure 3.4B). The association of these molecules with early times T1 and T1.5 (Figure 3.3A) suggests their reduction trend, which is in agreement with the time course plot. The fact that G6P and DHAP are closely adjacent to each other suggests they are highly correlated, which is consistent with Figure 3.4B, where they have similar pattern and both decrease over time; and with Table 3.1, where $r_{(G6P,DHAP)} \approx 0.99$. The relation of the locations between the metabolites, G6P, F6P and DHAP and the product glycerol clearly shows that they reside in the opposite direction

Table 3.1: Pearson's correlation matrix for metabolites in time course data generated with COPASI model of yeast glycolysis and glycerol biosynthesis pathway. This correlation matrix was computed using built-in functions in Mathematica 7.0. Correlations greater than 0.456 or less than -0.456 are significant ($P < 0.05$) (highlighted with gray color); correlations greater than 0.575 or less than -0.575 are significant ($P < 0.01$). Among 171 pair-wise correlations, 141 (82%) are significant at the 0.05 level, among which 80 are positive and 61 negative. The correlations between G6P, F6P, DHAP and Glycerol, and between ATP, ADP and AMP are highlighted; positive correlations are highlighted with orange color; negative correlations, green color.

	GLCi	ATP	G6P	ADP	F6P	F16bP	AMP	DHAP	GAP	NAD	BPG	NADH	P3G	P2G	PEP	PYR	AcAld	Glycerol	G3P
GLCi	1																		
ATP	-0.3430	1																	
G6P	0.6312	0.5121	1																
ADP	0.3435	-0.9999	-0.5116	1															
F6P	0.5731	0.5731	0.9973	-0.5728	1														
F16bP	0.7309	0.3854	0.9871	-0.3862	0.9748	1													
AMP	0.3414	-0.9993	-0.5126	0.9986	-0.5734	-0.3833	1												
DHAP	0.5683	0.5743	0.9944	-0.5751	0.9977	0.9767	-0.5721	1											
GAP	0.5683	0.5743	0.9944	-0.5751	0.9977	0.9767	-0.5721	1.0000	1										
NAD	-0.8655	-0.1698	-0.9320	0.1702	-0.9039	-0.9743	0.1687	-0.9037	-0.9037	1									
BPG	0.2355	0.8289	0.9002	-0.8297	0.9295	0.8349	-0.8260	0.9334	0.9334	-0.6900	1								
NADH	0.8656	0.1697	0.9320	-0.1701	0.9039	0.9742	-0.1686	0.9037	0.9037	-1.0000	0.6900	1							
P3G	0.7142	0.4121	0.9935	-0.4119	0.9828	0.9978	-0.4122	0.9807	0.9807	-0.9672	0.8476	0.9671	1						
P2G	0.7009	0.4292	0.9954	-0.4290	0.9861	0.9972	-0.4292	0.9841	0.9841	-0.9623	0.8575	0.9623	0.9998	1					
PEP	0.0960	0.8992	0.8307	-0.9001	0.8691	0.7488	-0.8963	0.8734	0.8734	-0.5809	0.9900	0.5809	0.7646	0.7767	1				
PYR	0.8176	0.2602	0.9624	-0.2594	0.9399	0.9859	-0.2620	0.9347	0.9347	-0.9925	0.7485	0.9925	0.9861	0.9829	0.6483	1			
AcAld	-0.8781	0.0791	-0.7404	-0.0731	-0.7022	-0.8360	-0.0932	-0.7250	-0.7250	0.9040	-0.4689	-0.9040	-0.8022	-0.7944	-0.3504	-0.8470	1		
Glycerol	-0.1558	-0.7133	-0.7280	0.7057	-0.7519	-0.6305	0.7302	-0.7221	-0.7221	0.5098	-0.7916	-0.5097	-0.6719	-0.6798	-0.7912	-0.6048	0.1306	1	
G3P	-0.3033	-0.7907	-0.9303	0.7905	-0.9545	-0.8682	0.7905	-0.9537	-0.9537	0.7365	-0.9952	-0.7364	-0.8834	-0.8920	-0.9757	-0.7960	0.4963	0.8232	1

relative to the center of the data. It indicates that their values are negatively correlated (highlighted in Table 3.1), which resonates with the fact that one is produced and the others are consumed.

One important property of biological network is the moiety conservation, a special form of mass conservation. Moieties are certain chemical groups that treated as indivisible units by many metabolic pathways. Their synthesis or degradation takes place in special pathways which need not always be included in a kinetic description. Examples of such moieties in yeast glycolysis and glycerol biosynthesis pathway are phosphate, NAD^+ and adenylic acid. Moiety conservation means that the sum of these moieties remains constant in spite of their occurrence in different metabolites and of continuous and rapid inter-conversion (Reich and Sel'kov 1981). Finding and analyzing conserved moieties may yield insights into the structure and function of a biological network. In this glycolysis and glycerol biosynthesis pathway, ATP, ADP and AMP demonstrate a strong example of moiety conservation, where

$$ATP + ADP + AMP = constant$$

And $$\frac{dATP}{dt} + \frac{dADP}{dt} + \frac{dAMP}{dt} = 0$$

The biplot display in Figure 3.3A clearly captures the relationship of the adenylate system: ATP is associated with early time, while ADP and AMP are associated with late time points. They dwell in the opposite direction relative to the center point. It indicates that ATP is negatively correlated with ADP and AMP; ADP and AMP are positively correlated. Both time course plot of these individual metabolites (Figure. 3.4B) and the correlation matrix in Table 3.1(highlighted in green and orange color) have confirmed the revelations from Figure. 3.3A.

An interesting biplot pattern, shown in Figure 3.3A has most molecules aligned roughly across the center except glycerol and F16bP. It suggests that these molecules are correlated, either positively or negatively. The Pearson correlation matrix (Table 3.1) shows that among 171 pair-wise correlations, 82% (141) of them have correlations that are significant at the 0.05 level (two-tailed) (greater than 0.456 or less than -0.456). Among these significant correlations, there are 61 negative and 80 positive. This fact is very characteristic of biological networks, in which most molecules are substrates, precursors, energy or products of the pathway. The mass conservation property of the network structure implies that the concentration of certain chemical species are linear combination of other chemical species (Mendes 2009). They subsequently have significant linear relationship.

Some of the features of the data that can be seen from this biplot (Figure 3.3) are the standard deviations of the time variables, by virtue of the time ‘rays’ drawn through the center and each time marker. The standard deviation of the time variables is represented by the length of

Table 3.2: Variances and standard deviations of time variables in time course data generated with COPASI model of yeast glycolysis and glycerol biosynthesis pathway. Where variances and standard deviations were computed with statistical functions in Excel using the following formula:

$Var = \frac{\sum(x-\bar{x})^2}{n-1}$; $stdev = \sqrt{Var}$.

Time	*T0.0*	*T0.5*	*T1*	*T1.5*	*T2*	*T2.5*	*T3*	*T3.5*	*T4*
Variance	35.63	28.60	25.37	24.12	23.75	23.76	23.92	24.13	24.37
stdev	5.969	5.348	5.037	4.911	4.873	4.874	4.890	4.913	4.937

Table 3.3: Pearson's correlation matrix for time points in time course data generated with COPASI model of yeast glycolysis and glycerol biosynthesis pathway. This correlation matrix was computed using built-in functions in Mathematica 7.0. Correlations greater than 0.798 or less than -0.798 are significant ($P<0.01$). All the correlations shown in this table are significant at the 0.01 level (two-tailed) and significant at 0.005 level (one-tailed). The vertical and horizontal gray arrows point out the decreasing patterns in the values. The columns of T2.5, T3 and T3.5 are formatted respectively to five to six decimal places in order to show the differences. The rest of the cells are formatted to four decimal places.

	T0	T0.5	T1	T1.5	T2	T2.5	T3	T3.5	T4
T0	1								
T0.5	0.9861	1							
T1	0.9570	0.9918	1						
T1.5	0.9303	0.9780	0.9967	1					
T2	0.9136	0.9681	0.9922	0.9991	1				
T2.5	0.9049	0.9626	0.9893	0.9979	0.9998	1			
T3	0.9006	0.9597	0.9878	0.9972	0.9995	0.99995	1		
T3.5	0.8983	0.9583	0.9870	0.9968	0.9993	0.99988	0.999986	1	
T4	0.8970	0.9574	0.9865	0.9965	0.9992	0.99983	0.999964	0.999995	1

the vectors when variables are not standardized. Time variable 0 minute has the longest vector and 0.5 minute comes second in length. Thus the standard deviation of 0 min is much larger than the rest of the time variables and that of 0.5 min is next in order. It is a true reflection of their standard deviations (Table 3.2). This feature of biplot instantly shows us the variability of all the variables. For the metabolomics data collected after perturbations, a time point with large standard deviation flags a period when the biological system has done some interesting work that is reflected as metabolic flux. In this simulation study, the carbon flux periods signaled by time rays 0 min, 0.5 min and 4 min are the results of early accumulation of F16bP and late production of glycerol respectively.

Another feature of biplot discussed here is nothing short of fascinating. When one looks at the biplot Figure 3.3A, one can't help noticing the pattern of all the time points that fan out in the time sequence. From this plot we can envisage the trajectory of the system (Figure 3.3B). Taking a biological system, it starts with certain initial levels of metabolites. Conversion and translocation will take place in accordance with the rate expressions. The metabolite contents change with time. If successive metabolite contents are plotted into a coordinate system (one axis for each metabolite), a continuous curve is obtained. Such a curve is called a ***trajectory*** of the system and all possible metabolic states form the points of the ***phase space***, which has the same dimension as the number of variable metabolites (Reich and Sel'kov 1981). In reality though, one can only envisage this high dimensional phase space. Thanks to the low dimension approximation of the high dimensional data matrix in Singular Value Decomposition, we can visualize the trajectory of the system in biplot display. On such a plot, each time point represents a system state at that moment. Therefore biplot display is analogous to phase space. When we join all the time point in the order, we decide to call it ***phase space biplot***.

What contributes to the formation of phase space biplot is the approximation of correlation between two time variables by the cosine of the angle between the corresponding time vectors. Angle θ on Figure 3.3A indicates correlation between variables 0 minute and 0.5 minute. A small angle between two vectors indicates that the two variables are highly correlated (e.g. Figure 3.3A, 0 minute is more correlated to 0.5 minute than to 1 minute). For *in silico* time course data (without noise), one time vector's most correlated variable is its adjacent time point(s). On a typical phase space biplot, a time variable's adjacent time is its nearest neighbor(s). Hence we see the interesting patterns in the correlation matrix for times (Table 3.3), where the correlations between two time variables are decreasing when they are farther apart in time. This pattern that the early time points have large standard deviation and that this decreases with increasing time is what should be expected in a stable dynamical system, where perturbations die out in time.

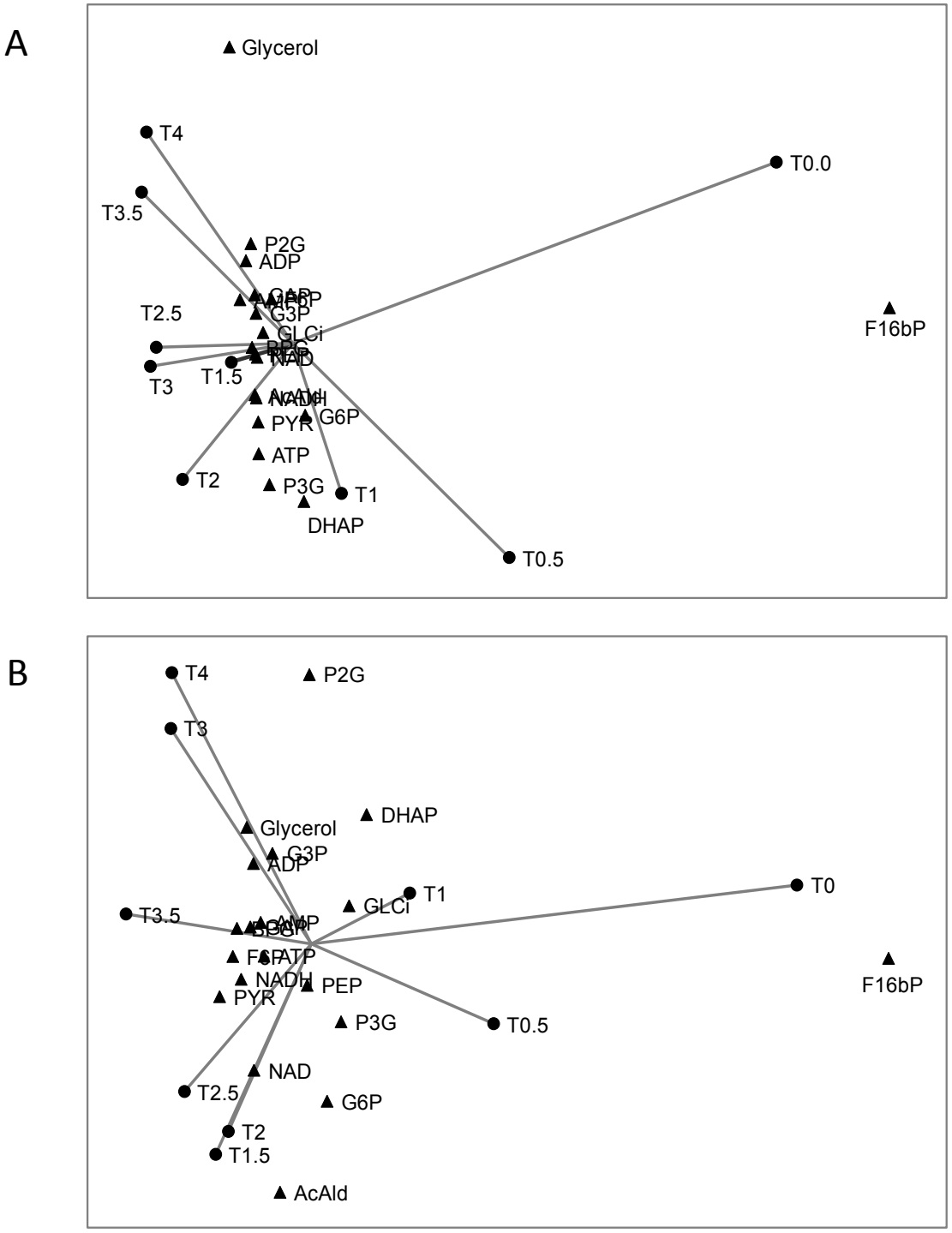


Figure 3.5: Effect of noise on biplot display. The simulated data was described in Method 3.3.1, which is the same data as in Figure 3.3, but with fixed additive noise added. A: the standard deviation of the noise is 100 times of the smallest value (0.000677) in the data, i.e. 0.0677. B: The noise is 1000 times of the smallest value, i.e. 0.677.

3.4.2 Robustness of biplot display to noise

The high-throughput datasets produced in systems biology study are prone to technical and biological noise. Technical noise arises from sensitivity of the experimental techniques (such as limited resolution of equipment or differences in hybridization strength between probes in microarrays (Klebanov and Yakovlev 2007)), or from the variation of the cell populations used. Biological noise stems from stochastic events occurring during gene expression or from differences in the local cellular environment (De Backer, De Waele et al.). The desired analytical method for systems biology study should be sensitive to changes in biological inputs, but at the same time robust to technical noises. To assess the robustness of biplot display to the level of noise in the measurements, we need to add well-defined sources of noise to the simulated data.

Considering three different types of variance that can be added, we chose to add additive noise (Mendes, Camacho et al. 2005) to simulate the noise that is introduced by the measurement process and it is incorporated into simulated data by adding appropriate random values to the data after simulation (Mendes, Camacho et al. 2005). Figure 3.5 shows how biplot display is affected after fixed additive noise was added to the original data. We found that biplot display is robust to a certain level of fixed additive noise. When noise with a standard deviation (0.0677) equivalent to 100 times of the smallest value was added to simulated data, biplot still keeps its basic shape and separation between early time points (0 – 2.5 min) and late time points (3 – 4 min). The representations of F16bP and glycerol on the plot haven't changed. The relationship between ADP, AMP and ATP is intact. But due to the noise addition, some molecules with small values become inflated and visible on the plot (e.g. P2G, P3G, PYR and GLCi). Only when we increase the measurement error to 1000 times of the smallest value (0.677), the biplot is destroyed (the patterns are no longer preserved). From Figure 3.5 A and B, we found that the time variables that have larger variance (0 min and 0.5 min) are relatively not affected to the small added noise. The disruption of the phase space biplot by the additive noise has explained why we rarely see the typical system trajectory in the biplot of the real experimental data. It's debatable whether the level of noise mentioned previously is appropriate in real experiments. A recent study has found that the observed technical noise in microarray data is quite low and does not cause any tangible bias in statistical inference (Klebanov and Yakovlev 2007). But the level of noise in metabolic and proteomic profiles is still unknown.

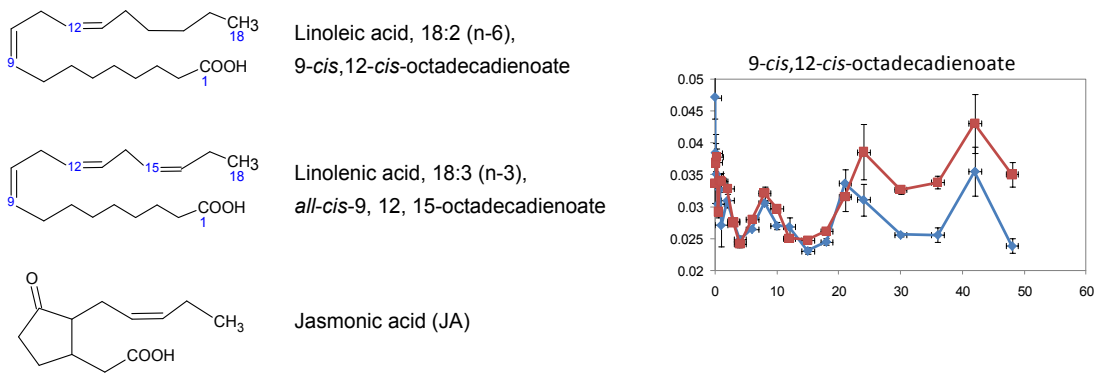


Figure 3.6: The structure of Jasmonic acid (JA) and its precursors (left) (drawn with ChemSketch 8.0), and the transient increase of 9-*cis*, 12-*cis*-octadecadienoate (linoleic acid) in MeJa-elicited sample (right). Jasmonic acid, a defense and developmental signal is derived from linoleic acid and linolenic acid, drawn in ChemSketch 8.0. The response of linoleic acid was obtained with GC-MS analyses on non-polar extracts. Y-axis values represent relative peak areas after normalization to the mean peak area for that compound. Maroon squares represent elicited sample means and blue diamonds represent control sample means. Error bars represent standard error from six replicates analyses.

3.4.3 Biplot of *Medicago truncatula* data–MeJa Elicitation

The ultimate goal of studying a methodology like biplot display is to investigate the real experimental data and to gain insight into the biological system. We applied biplot display to the experimental data from *Medicago truncatula* study (Broeckling, Huhman et al. 2005).

The time course data of metabolites response to elicitations of methyl jasmonate (MeJa), UV radiation and yeast infection were obtained with gas chromatography-mass spectrometry (GC-MS) analyses and liquid chromatography-electrospray ionization mass spectrometry (LC-MS) analyses. *M. truncatula* as an emerging biological model for forage legume study is a rich source of bioactive products (Dixon and Sumner 2003). Two classes of these secondary metabolites are of particular interest. Isoflavonoids are a class of polyphenolic compounds, which have been attributed with disease resisting and health promoting properties and are nearly exclusive to legumes (Broeckling, Huhman et al. 2005). Another important class of the secondary metabolites are the triterpene saponins, which protect plants by their allelopathic, antimicrobial, anti-insect and anti-palatability activities. *M. truncatula* has a diverse content of saponins, which is shown to be elicited by the stress and developmental signaling molecule MeJa (Suzuki, Achnine et al. 2002; Farmer 1994).

Jasmonic acid (JA), and its methyl ester (methyl jasmonate, MeJa) are linoleic and linolenic acid-derived cyclopentanone-based compounds of wide distribution in the plant kingdom. The biosynthesis of most plant JA is initiated by lipoxygenase, a non-heme iron dioxygenase that

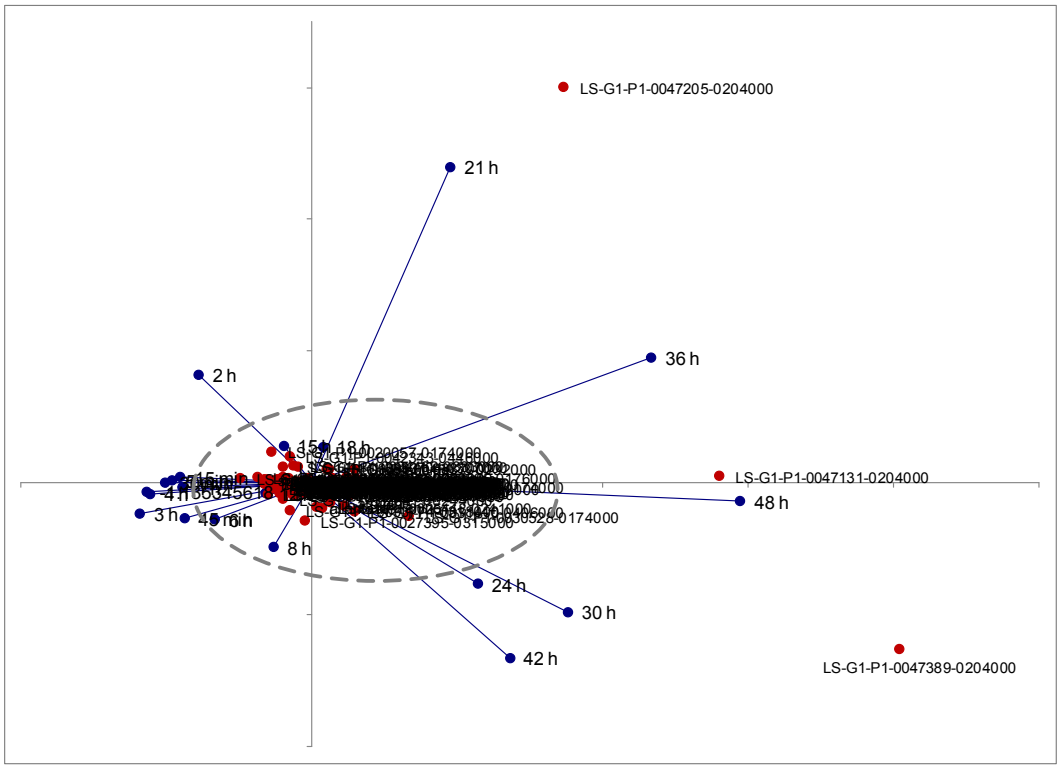
adds molecular oxygen to either the 9 or the 13 position of the C18 chain of linoleic and linolenic acids (Figure 3.6) (Creelman and Mullet 1997; Howe and Schilmiller 2002; Stumpe, Carsjens et al. 2005).

Two compounds, 9-*cis*, 12-*cis*-octadecadienoate and Beta-amyrin are the only identified non-polar metabolites that have significantly increased in MeJa-elicited sample. This result has enhanced the previous finding by Broeckling *et al.* (Broeckling, Huhman et al. 2005) where Beta-amyrin was the only identified non-polar metabolite that has responded to MeJa elicitation. 9-*cis*, 12-*cis*-octadecadienoate, commonly known as linoleic acid, is an unsaturated omega-6 fatty acid with 18 carbons (octadeca-). Together with linolenic acid (*all-cis*-9, 12, 15-octadecadienoate), they represent two essential fatty acids that humans and other animals must ingest for good health because the body requires them for various biological processes, but cannot synthesize them from other food components. Linoleic acid and linolenic acid were found at high level in *Medicago* plants (Demir and Cakmak 2007). They are the initial precursors in jasmonate biosynthesis. Mueller *et al.* (Mueller, Brodschelm et al. 1993) were able to show that linolenic acid was released in response to elicitor treatment of cells. The transient increase of linoleic acid following MeJa elicitation suggests exogenous MeJa has initiated the octadecanoic-based pathway from the C18 fatty acid linoleic and linolenic acid to jasmonic acid. This is in agreement with the previous studies that showed exogenous JA will cause transient and large changes in the concentration of JA in most plant cells before reaching internal equilibrium in tissues (Harms, Atzorn et al. 1995 ; Creelman and Mullet 1997).

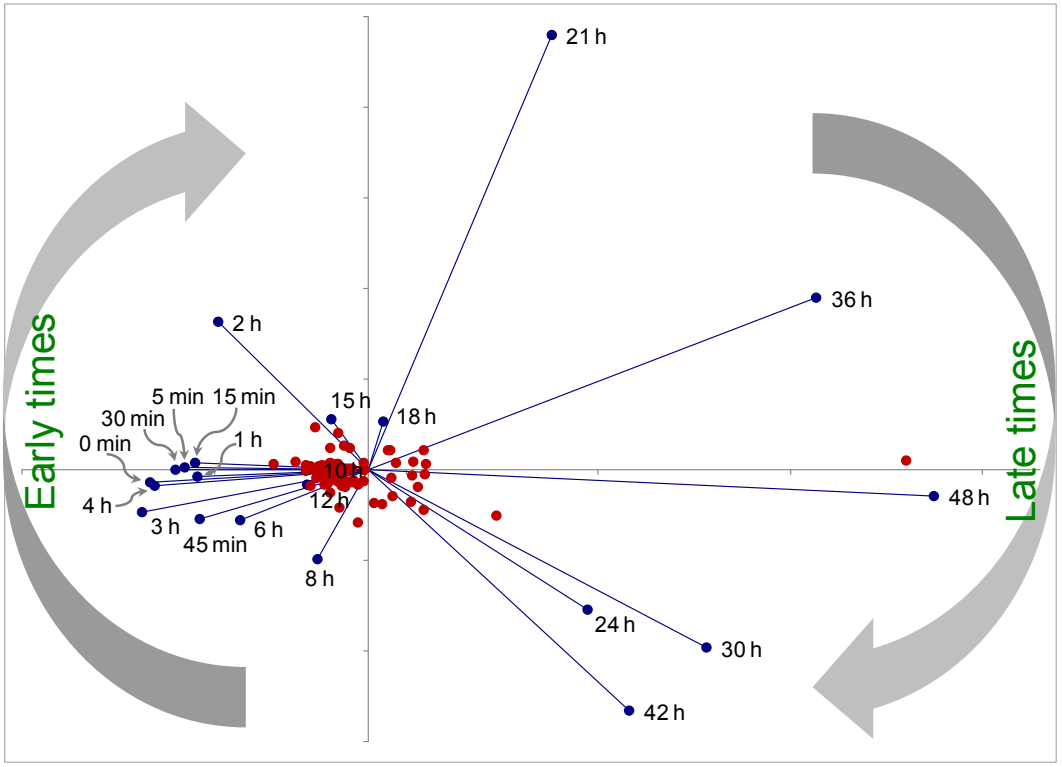
Analyses of GC-MS based metabolites reveal the effect of elicitation on the accumulation of many primary metabolites and their interrelationship. Secondary metabolites, including triterpene saponins, and isoflavanoids were analyzed using LC-MS (Broeckling, Huhman et al. 2005). Marked changes in the level of primary metabolites, including several amino acids, organic acids, and carbohydrates were observed following elicitation with MeJa. Biplot analyses on the significantly changed metabolites ($p \leq 0.01$) with MeJa elicitation indicate apparent separation between early and late times by y axis (Figure 3.7B). Early times from 0 min to 18 hrs have relatively small standard deviations, which is indicated by the short rays that mainly fall in the 2nd and 3rd quadrants of the Cartesian plane; while late times from 21 hrs to 48 hrs have large variability, whose vectors spread in the 1st and 4th quadrants. The time rays in Figures 3.7B suggest that the changes and accumulations of the primary metabolites after elicitation mostly happen in the late times.

The effect of elicitation on the primary metabolites pool is most striking following MeJa elicitation (Figure 3.7 and 3.8). A similar but dampened response was observed with yeast elicitor, whereas little response was observed with UV radiation (Data not shown). Elevated level of several amino acids, notably lysine, leucine, isoleucine, phenylalanine, arginine and

A


LS-G1-P1-0047205-0204000
21 h
36 h
2 h
LS-G1-P1-0047131-0204000
48 h
3 h
8 h
24 h
30 h
42 h
LS-G1-P1-0047389-0204000

B


Early times
Late times
21 h
36 h
2 h
5 min
15 min
30 min
0 min
1 h
15 h
18 h
10 h
12 h
4 h
3 h
45 min
6 h
8 h
48 h
24 h
30 h
42 h

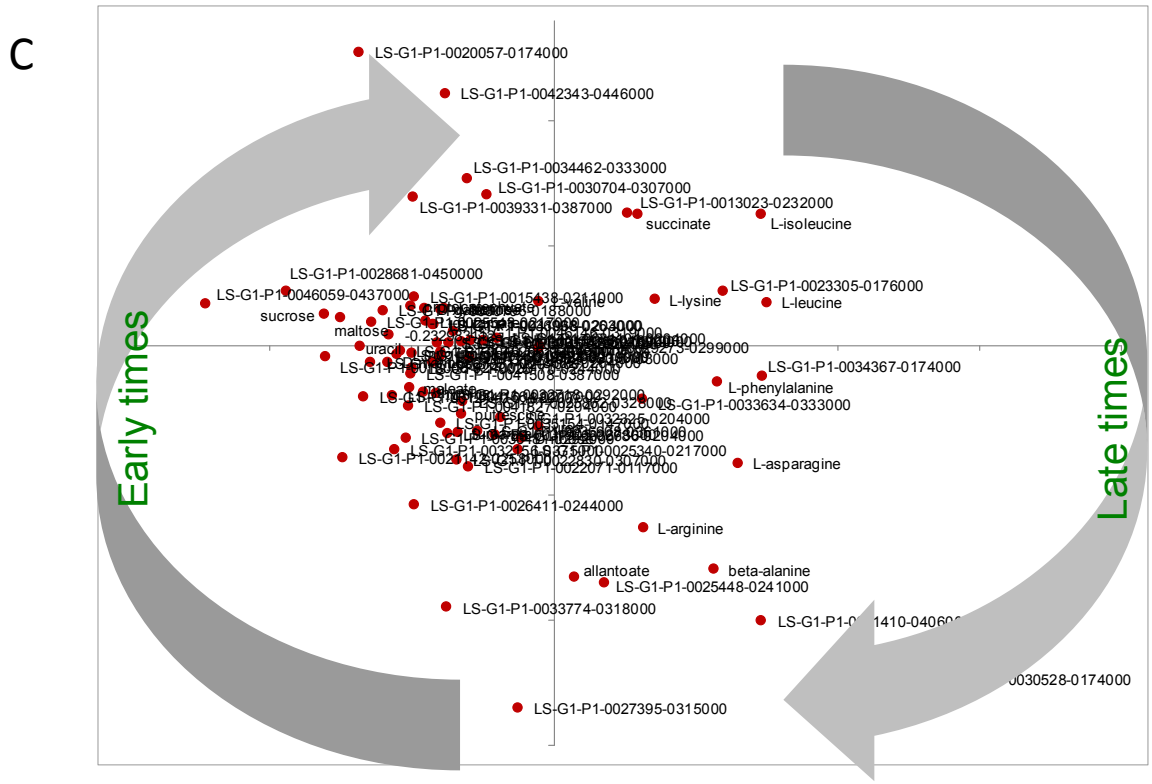


Figure 3.7: Biplot display of selected metabolites following elicitation with Methyl Jasmonate in *Medicago truncatula* study. The responses of metabolites were obtained with GC-MS analyses on polar extracts. The metabolite data were corrected by internal references and retrieved through the database for 'ome' ─ DOME (http://calvin.vbi.vt.edu/DOME/DOMEMT/index.php). The selected 88 metabolites have significant responses after the elicitation with *p* value less than 0.01. Rows and columns centering are performed before SVD. Times and metabolites are equally scaled. Blue points represent times from 0 minute to 48 hours and red points represent 88 significantly changed metabolites. A: Biplot display of the times and metabolites ─ the whole picture. B: An enlarged view of all the time points. The early times and late times are marked by curved block arrows and roughly separated by *y* axis. C: An enlarged view of the portion circled in Figure 3.6A. The early times and late times are marked by curved block arrows. All of the 88 metabolites except LS-G1-P1-0047205-0204000, LS-G1-P1-0047131-0204000 and LS-G1-P1-0047389-0204000 are represented here.

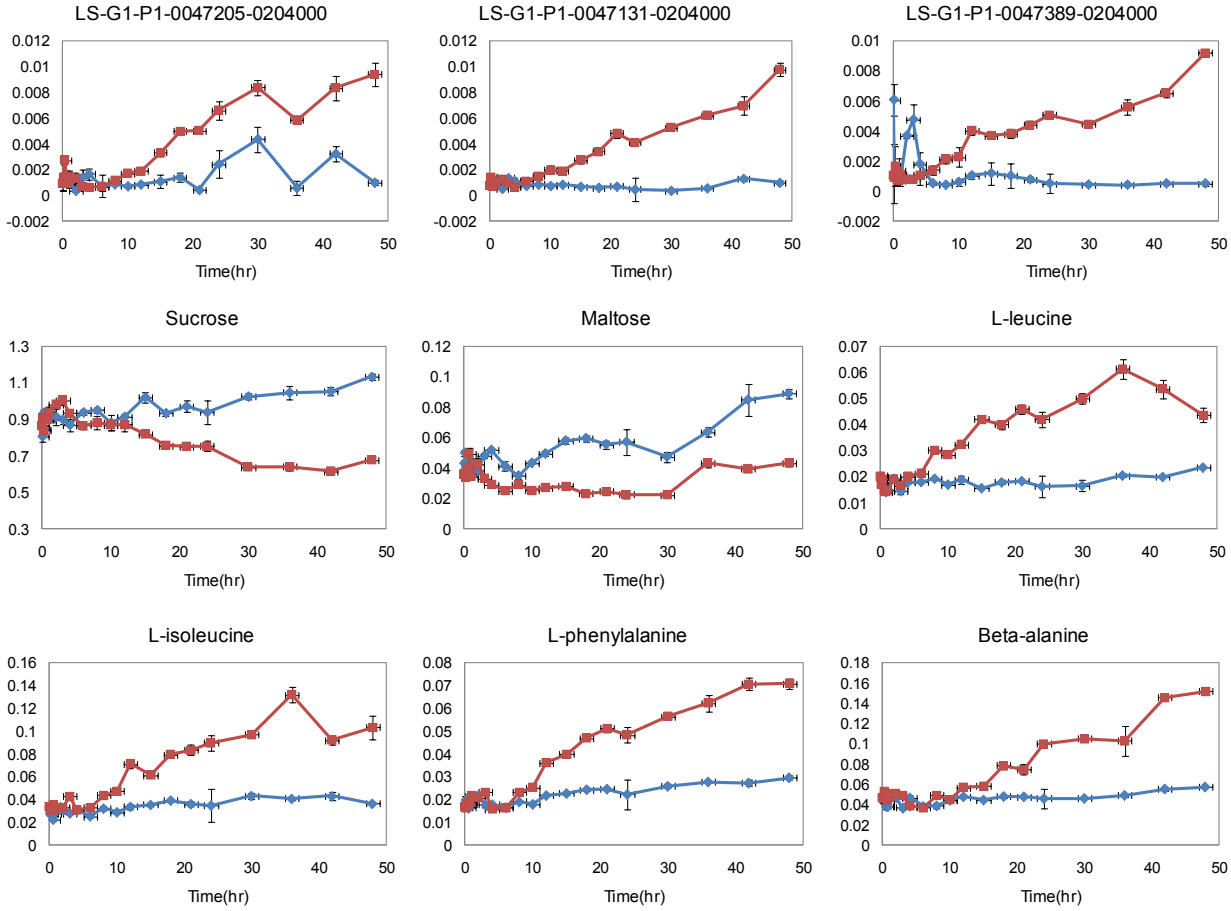


Figure 3.8: Responses of selected metabolites to elicitation with Methyl Jasmonate in *Medicago truncatula* study. The data was obtained with GC-MS analyses on polar extracts. Y-axis values represent relative peak areas after normalization to the mean peak area for that compound. Maroon squares represent elicited sample medians and blue diamonds represent control sample medians. Error bars represent standard error from six replicates analyses (three biological replicates and two injection replicates for each biological sample). The unknown compounds LS-G1-P1-0047205-0204000, LS-G1-P1-0047131-0204000 and LS-G1-P1-0047389-0204000 dominate the late times on the Biplot display of Figure 3.6A. Sucrose and Maltose are found on the second quadrant of Biplot display on Figure 3.6C and both are associated with early times. L-leucine, L-isoleucine, L-phenylalanine and Beta-alanine are located respectively on the first and fourth quadrants of Biplot display on Figure 3.6C. They are associated with late times. These amino acids increased following elicitation while sucrose and maltose decreased.

asparagine were observed over 48 hrs (Figure 3.7 C and Figure 3.8). On biplot these amino acids are associated with late times, namely from 21 hrs to 48 hrs. Their responses over time indicate an increasing pattern. Interestingly leucine and isoleucine show a spike at 36 hrs, while on biplot both of them are located in 1st quadrant of the Cartesian plane and close to 36 hrs vector. Sucrose and maltose reside in 2nd quadrant of the *xy* plane, which indicates they are decreasing over time. Their levels are lower in MeJa-elicited samples compared to the control samples (Figure 3.8). It suggests that MeJa signals not only induce specific defense related pathways including triterpene synthesis, but also trigger a wide variety of physiological responses in plants (Farmer 1994). Study has shown that MeJa regulates the growth, N uptake, N partitioning and N storage in roots of Leguminous plants (Meuriot, Noquet et al. 2004). These results indicate MeJa alters the carbon source partitioning as well. In addition, the elevation of Beta-alanine level suggests MeJa's effect on the pathways of coenzyme A, which further elucidates MeJa's role as developmental signals.

Among 32 triterpene glycosides detected in the *M. truncatula* root cell culture through LC-MS analyses, 10 of them showed significant accumulation in MeJa-elicited sample ($p \leq 0.01$). All of them are glycoside derivatives of five different triterpene aglycones: hederagenin, bayogenin, medicagenic acid, soyasapogenol B and soyasapogenol E (Figure 3.8) (Suzuki, Achnine et al. 2002; Achnine, Huhman et al. 2005). Their principal route of formation is from beta-amyrin, one of the most common triterpenes in plants. It is made of a pentacyclic carbon skeleton, derived from the precursor 2,3-squalene after a series of processes mediated by 2,3-oxidosqualene cyclase and Beta-amyrin synthase (Yendo, de Costa et al. 2010). The downstream reactions in the biosynthesis of *M. truncatula* saponins are believed to include a set of cytochrome P450-dependent hydroxylations/oxidations and several glycosyl transfer reactions catalyzed by glycosyltransferases (GTs) (Achnine, Huhman et al. 2005; Yendo, de Costa et al. 2010; Vogt and Jones 2000; Naoumkina, Modolo et al. 2010).

The triterpene beta-amyrin is another identified non-polar compound that demonstrated an elicitation response. Beta-amyrin showed a small but consistent elevation in MeJa-elicited sample since 10 hrs (Figure 3.10). LC-MS analysis revealed the spikes of most saponins after 42 hrs (Figure 3.10), which suggests that the accumulation of beta-amyrin precedes the increase of triterpene biosynthesis (Broeckling, Huhman et al. 2005). Following MeJa elicitation, soyasaponin I is the most highly induced saponin in the cell culture, with over 2000-fold increase in elicited sample. Soyasaponin I is a group B soyasaponin (Figure 3.9), it is a main form of soyasaponins in heat-treated soy products (Hu, Reddy et al. 2004). The other most strongly induced saponins are (Figure 3.10), in order, rha-hex-hex-soyasapogenol E (32-fold induction), 3-glc-ara-28-glc-hederagenin (23-fold), hex-hex-bayogenin (18-fold), hex-hederagenin (13-fold), rha-hex-hex-hederagenin and rha-hex-hex-hex-soyasapogenol E (7-fold), hex-hex-hex-bayogenin (5-fold). Nearly all of them have a striking increase at 36

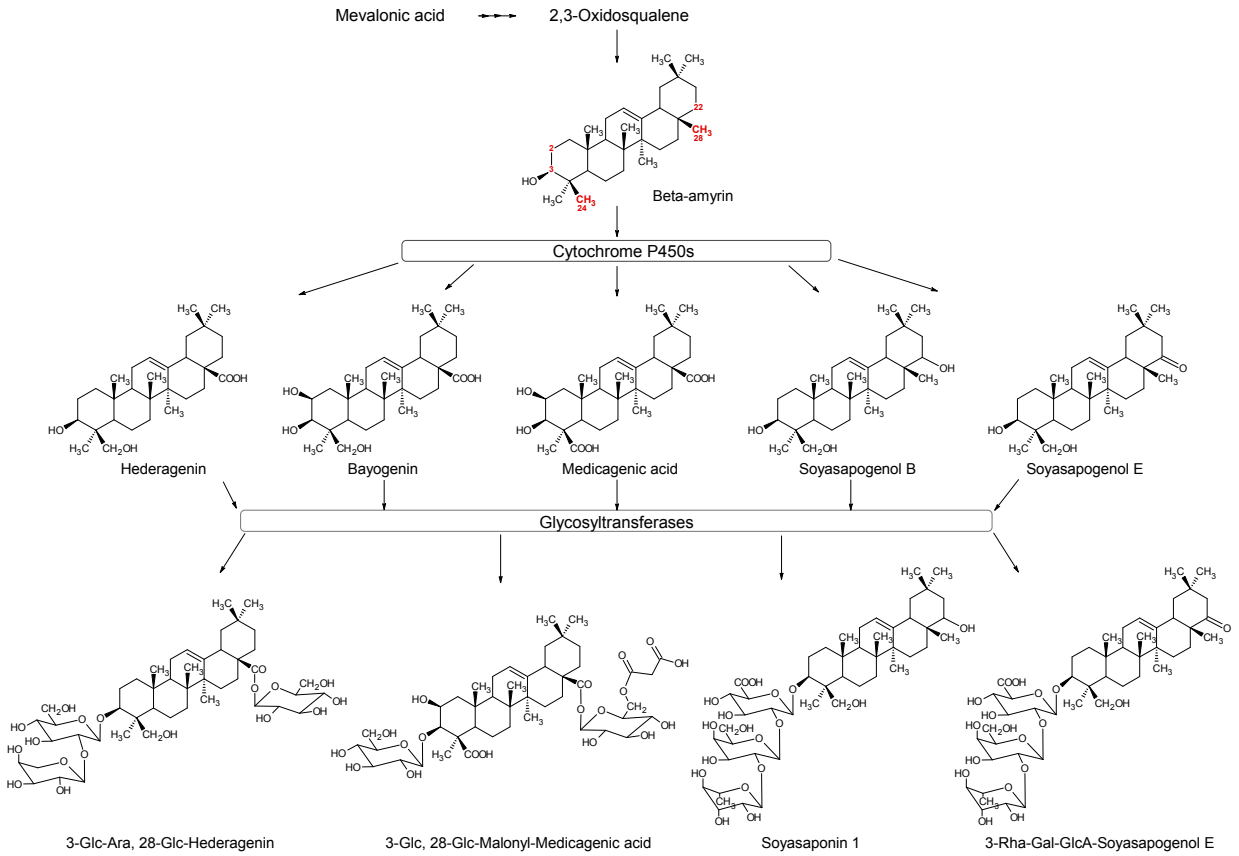


Figure 3.9: Biosynthetic pathway of the saponins in *Medicago truncatula* (drawn and created with ChemSketch 8.0). The universal precursor Beta-amyrin is converted by a series of oxidative reactions to at least five different triterpene aglycones: hederagenin, bayogenin, medicagenic acid, soyasapogenol B, and soyasapogenol E, which are converted by glycosyltransferases to over 37 different triterpene saponins (Huhman and Sumner 2002).

Figure 3.10: Levels of Beta-amyrin and selected triterpene saponins responding to MeJa elicitation in *Medicago truncatula* study. The value of Beta-amyrin was obtained with GC-MS analyses on non-polar extracts. Its *y*-axis values represent relative peak areas after normalization to the mean peak area for Beta-amyrin. Values of triterpene saponins are relative molecular ion intensities from LC/MS analysis normalized on an equal cell weight basis with the level of hexose hederagenin in non-elicited cells set as 100 (Suzuki, Reddy et al. 2005). Maroon squares represent elicited sample means and blue diamonds represent control sample means. Error bars represent standard error from six replicates analyses (three biological replicates and two injection replicates for each biological sample). Abbreviations: Hex, Hexose; Rha, rhamnose; Glc, glucose; Ara, arabinose; GlcA, glucuronic acid.

hrs. The exceptions to these are the glycosides of medicagenic acids, with a less than 2-fold consistent increase in 3-glc-28-glc-medicagenic acid and 3-glc-28-glc-malonyl-medicagenic acid in the MeJa-elicited cell culture (Figure 3.10). This result is in agreement with a previous finding, where much higher levels (15-fold) of medicagenic acids were found in the culture medium of MeJa-treated cells (Suzuki, Reddy et al. 2005). This is due to the very good water and alcohol solubility of medicagenic acid glycosides, which is attributed to its allelopathic role in plant roots, where they can readily be released to the environment (soil) as water soluble compounds and effect neighboring crop seedling growth (Oleszek and Jurzysta 1987). Unlike many other saponins that were not detected or only present in small amounts in the un-elicited cell cultures, a substantial amount of medicagenic acids and rha-hex-hex-hex-soyaspogenol E were observed in the control sample. It suggests that they are constitutive metabolites and located in roots and/or aerial parts of the plant ready to be utilized against attack (Dixon 2001).

The broad range of triterpene saponins in the study is primarily due to high degrees of conjugation in *M. truncatula* (Huhman and Sumner 2002). Saponins belonging to the same core triterpene aglycone can have a huge variation in the number and type of monosaccharide residues attached. Oligosaccharyl chains are usually attached at the C3 and/or the C28 atom(s) and they have a length of 1–4 residues in *M. trucatula*. This mechanism represents a high efficiency biosynthesis system in plant kingdom. The chemically diverse spectrum of triterpene saponin in *Medicago* species may reflect the diverse roles they play in the signaling and plant defense process.

All of the three unidentified compounds that dominate biplot display on Figure 3.7A, LS-G1-P1-0047205-0204000, LS-G1-P1-0047131-0204000 and LS-G1-P1-0047389-0204000 have same fragment ion peaks at *m/z* 204, and they elute approximately at the same time of 47 min. The plots of their responses over time show a pronounced increase in the MeJa elicited samples (Figure 3.8). Inspection of the significantly changed metabolites ($p \leq 0.01$) revealed that these compounds are abundant in the metabolite profiles following all three elicitations. There are 12 of these compounds with ions at *m/z* 204 with MeJa elicitation, 9 with yeast infection and 4 with UV radiation. The intensity of the ion *m/z* 204 (TMSiO-CH=CH-OTMSi) is closely related to the ring size. The formation of ions *m/z* 204 requires a circulation of electrons within the ring, which is favored by a six-membered cyclic structure (Starke, Holzberger et al. 2000). The fragment ion of *m/z* 204 is characteristic in the mass spectra of saccharides with a pyranose ring. Deducing the hypothetical sources of this compound discussed below, is the intriguing part of work in the metabolic profiling and systems biology study.

Studies on *Medicago* plants have provided some clues of the possible identity/identities of these abundant compounds. They could be N-Acetylglucosamine (GlcNAc) residues in Nod

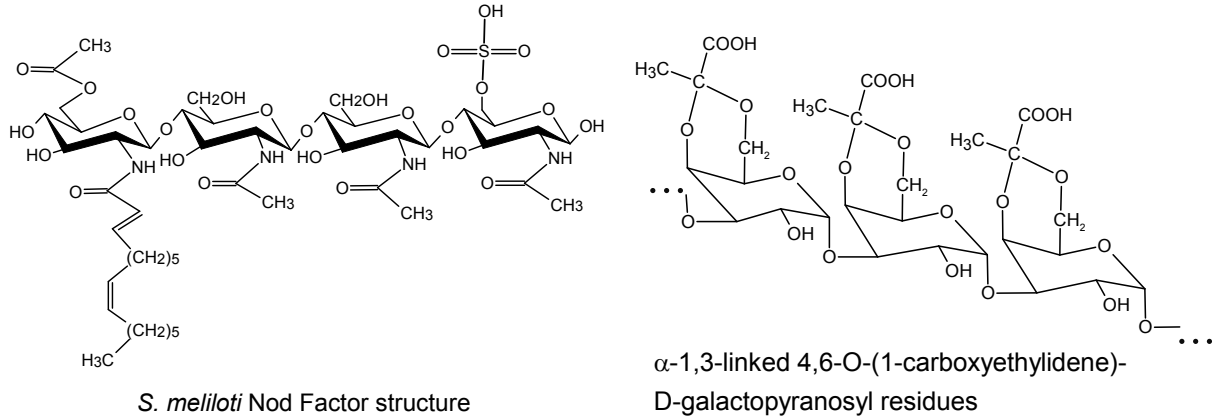


Figure 3.11: Structures of the compounds with fragment ion *m*/*z* 204 that highly responded to MeJa elicitation (drawn in ChemSketch 8.0). Left: Nod factor in *S. meliloti* (Wais, Keating et al. 2002); right: Three α-1,3-linked 4,6-O-(1-carboxyethylidene)-D-galactopyranosyl residues that constitute the structure units in the extracellular polysaccharides (EPS) released by rhizobia during the symbiotic process (D'Haeze, Glushka et al. 2004). The characteristic fragment ions of *m*/*z* 204 were attributed to the ─C_3C_4─ part of the N-Acetylglucosamine (GlcNAc) residues in Nod factor and 4,6-O-(1-carboxyethylidene)-D-galactopyranosyl residues in EPS released by rhizobia.

Factors or 4,6-O-(1-carboxyethylidene)-galactose residues (Figure 3.11) in the extracellular polysaccharides (EPS) released by rhizobia, *M. truncatula*'s symbiotic partner. Nitrogen-fixing rhizobial bacteria and leguminous plants have evolved complex signal exchange mechanism ('molecular dialogue' by De´narie´ et al.) that allows the bacteria to use plant root hair cells as a means of entry and leads to the formation of root nodules on legumes (Cooper 2007; Sun, Cardoza et al. 2006). Once rhizobia have entered the plant root, the bacteria differentiate into a new form that can convert atmospheric nitrogen into ammonia. Bacterial differentiation and nitrogen fixation are dependent on the microaerobic environment and other support factors provided by the plant. In return, the plant receives nitrogen from the bacteria, which allows it to grow in the absence of an external nitrogen source (Jones, Kobayashi et al. 2007). Chemical compounds produced by leguminous plant, such as, flavonoids, simple phenolics and jasmonates have been shown to act as *nod* gene inducers (Cooper 2007; Mabood, Souleimanov et al. 2006). Extracellular signals secreted by Rhizobia include Nod factors, surface polysaccharides and proteins, *etc*. (Farmer 1994; Cooper 2007; Sun, Cardoza et al. 2006).

The rhizobial symbiotic signal molecules─Nod factors are able to induce plant root hair deformations and nodule organogenesis. Nod factors are chitin (N–acetylglucosamine oligomers) derivatives, of which the non-reducing end is N-acylated and the reducing end is

modified by various molecules. These specific structures determine the strict specificity between Rhizobium and host legume species, and elicit both the rhizobial infection process and nodulation development (Lerouge, Roche et al. 1990; Kouchi, Imaizumi-Anraku et al. 2010). The structure of the major Nod factor from *Rhizobium meliloti*, NodRm- 1, was determined by mass spectrometry and was shown a main characteristic fragment ion at *m/z* 204 that was attributed to the C-3 and C-4 part of a hexosamine moiety (Roche, Lerouge et al. 1991). The same fragment ion peak of *m/z* 204 was also found in Nod signal molecules in *Bradyrhizobium japonicum* strain that nodulates soybeans (Carlson, Sanjuan et al. 1993) and in *Mesorhizobium loti* strain that has a symbiotic relationship with *Lotus japonicas* (Niwa, Kawaguchi et al. 2001).

The fragment ion peak at *m/z* 204 was also found in the spectrum of the trimethylsilyl derivatives of 4,6-O-(1-carboxyethylidene)-galactose (Misevic, Guerardel et al. 2004), which is the structural units of the massive amount of extracellular polysaccharides (EPS) released by rhizobia during the process of invasion and nodulation of host plant (D'Haeze, Glushka et al. 2004). EPS is one of the four types of surface polysaccharides (SPS). Accumulating data suggest that rhizobia SPSs play an important role in various stages of symbiotic development including root colonization, host recognition, infection thread formation and nodule invasion (Cooper 2007). Specifically they are important for the evasion of plant immune responses and as protecting agents against reactive oxygen species (D'Haeze, Glushka et al. 2004).

3.4.4 Biplot on *Medicago truncatula* data—Yeast Elicitation

When we apply phase space biplot to a whole metabolomic time course data set, we usually don't see nice time patterns probably because of complex occurrences of different pathways after perturbation and/or contribution of technical or biological noises during the experiment and data collection. To achieve a "perfect" phase space biplot, we need to collect metabolites data that are presumably from the same biosynthetic pathway and apply biplot display and phase space biplot to them. The *Medicago* isoflavonoids profiled by LC-MS analyses in yeast-elicited root cells provides such an opportunity. The author would like to clarify that biplot display as an exploratory data analysis tool doesn't require a prior knowledge, but for the sake of phase space demonstration, the *Medicago* flavonoids profile is chosen as input signal.

Isoflavonoids, an important group of legume natural products, function as constitutive or inducible antimicrobial or anti-insect compounds, as inducers of the nodulation genes of symbiotic Rhizobium bacteria, or as allelopathic agents (Dixon 2001; Farmer 1994; *Cook* 1999). Its subcategory isoflavones are also beneficial to human health through their estrogenic, antiangiogenic, antioxidant, and anticancer activities (Barnes 2003; Dixon 2001;

Setchell and Cassidy 1999). Isoflavones are derived from flavanones through phenylpropanoid pathway. Flavanones are ubiquitous intermediates leading to the biosynthesis of all other flavonoid subclasses (Figure 3.12 and Figure 3.13). Isoflavones are synthesized from the flavanones naringenin and liquiritigenin via migration of the B-ring from the 2- to the 3-position, followed by hydroxylation at the 2-position. This complex reaction is catalyzed by isoflavone synthase (IFS), a cytochromeP450 enzyme, and yields the immediate product 2-hydroxyisoflavanone that is subsequently dehydrated, either spontaneously or enzymatically, to the corresponding isoflavone (Deavours and Dixon 2005; Farag, Huhman et al. 2008). In this way, IFS converts liquiritigenin to daidzein and naringenin to genistein.

The biosynthetic pathway leading to the production of isoflavonoids can be elicited by the application of yeast cell wall extract (Suzuki, Reddy et al. 2005; Broeckling, Huhman et al. 2005). Out of 188 extracted and quantified metabolites in yeast-elicited cells with LC-MS analyses (flavonoid as internal reference), 116 of them have significant responses following elicitation ($p \leq 0.01$). Biplot display on these selected molecules reveals the flavonoids and other molecules that were most highly induced under the action of yeast elicitation (Figure 3.14). These metabolites and their maximum induced fold are illustrated in Figure 3.14C. The biplot interprets that the distance from a metabolite to the center represents the variability of this metabolite's data, i.e. the bigger distance, the bigger variability. Since biplot analyses use ratio of the medians between elicitation and control sample in this study, a higher variability of the ratio data indicates a stronger response to elicitation. The metabolites and their corresponding number of fold induction in Figure 3.14C confirmed this inference, where the most strongly induced flavonoid, aurone glucoside—hispidol 4'-O-glucoside (39×) is farthest away from data center; other elicited flavonoids are moving toward center in the decreasing order of fold-induction, alfalone (21×), medicarpin (20×), hispidol glucoside malonate (19×), naringenin (13×), formononetin (10×), isoliquiritigenin (9×), irisolidone isomer (7×), biochanin A (7×), hispidol (6×), afrormosin (6×), daidzin (5×).

Biplot display also reveals the dynamics of the elicited metabolites through the inter-relationship between metabolites and times. Figure 3.14 depicts the divisions among different stages following elicitation: Early times (0 min to 4 hrs) are located in the 2nd quadrant of the Cartesian plane; mid-times (6 to 15 hrs), 1st quadrant of the plane; late times (18 to 48 hrs), 4th and 3rd quadrants. Biplot shows that the major increase of certain flavonoids responding to the elicitation happened at mid to late stages from 6 or 36 hrs (Figure 3.14B). A number of important flavonoids assembled near the mid time rays, especially between 8 to 10 hours. Note that there is a bigger angle between 8 hrs and 10 hrs compared to those between 6 and 8 hrs, and to those between 10, 12 and 15 hrs. It indicates that these two time variables are less correlated, *i.e.* the dynamics of the system have largely changed during this period, which is attributed to the transient increase of those essential flavonoids, namely alfalone, medicarpin,

4', 5, 7-trihydroxyflavanone (naringenin), formononetin, isoliquiritigenin, biochanin A, irisolidone isomer and afromorsin (Figure 3.14C). Plots of these metabolites' response to yeast elicitation confirmed this finding (Figure 3.14).

Most of the flavonoids peaked at 8 to 10 hrs following elicitation (Figure 3.14). But they declined with various dynamics, which could be used to describe three groups of metabolites. First group of flavonoids show a relative quick decline close to control level and flat out. These include isoliquiritigenin, naringenin, biochanin A, formononetin and Medicarpin. Chalcone isoliquiritigenin, the earliest responding metabolite during mid times revealed by biplot (Figure 3.14C) is the immediate precursor of liquiritigenin, which together with naringenin (4', 5, 7-hydroxyflavanone) serve as entry points into the flavone and isoflavone biosynthetic pathways (Figure 3.12). Biochanin A is a naringenin-derived isoflavone, which is a phytoestrogen and anti-cancer agent (Peterson and Barnes 1993). Formononetin is liquiritigenin-derived isoflavone through two intermediates, 2-hydroxynisoflavanone and 2,7-dihydroxy-4'-methoxyisoflavanone catalyzed by 2,7,4'-trihydroxyisoflavanone 4'-*O*-methyltransferase and dehydratase respectively (Deavours and Dixon 2005). Formononetin is a precursor of the pterocarpan phytoalexin medicarpin. Labeling studies indicated formononetin is also the precursor for afrormosin and alfalone, which is discussed below.

The second group of flavonoids that peaked at 8 to 10 hrs following elicitation had a milder decline and reached a plateau before 30 hrs, and then have another smaller peak (Alfalone, Afromosin and p-hydroxybenzaldehyde) or dip (irisolidone isomer). Isoflavones, alfalone, afromosin and irisolidone all have two methyl groups at the C4', C6 or C7 positions. Alfalone is a structural isomer of afrormosin, whereas irisolidone has an additional hydroxyl group at the C5 position relative to afrormosin. In the *M. truncatula* cell cultures, afrormosin was produced constitutively and its levels were further enhanced following exposure to yeast elicitor, whereas alfalone and irisolidone were only detected following elicitation (Figure 3.14). Plots of these three isoflavones' responses to elicitation showed that alfalone and afrormosin has similar dynamics, with a second peak at 36 hrs, when irisolidone has a dip. It indicates the former two isoflavones have a higher correlation in their elicited response pattern. Pulse labeling studies using exogenous [$^{3}H_{1}$]formononetin revealed the accumulation of ^{3}H label in afrormosin and alfalone, which supports a biosynthetic link and evidence that formononetin is a precursor of afrormosin and alfalone (Farag, Huhman et al. 2008).

The only major yeast elicitation induced metabolite that doesn't belong to flavonoids family is *p*-hydroxybenzaldehyde. It's a plant cell wall-bound phenylic compound related to early phenylpropanoid pathway. Studies have shown that changes in cell wall-bound phenolics occur in plant/pathogen interactions and are proposed to be involved in an increase of cell wall resistance against enzymes of the invader as well as in strengthening of a physical barrier against pathogens (Dixon 2001). The 9-fold induction following elicitation indicates *p*-

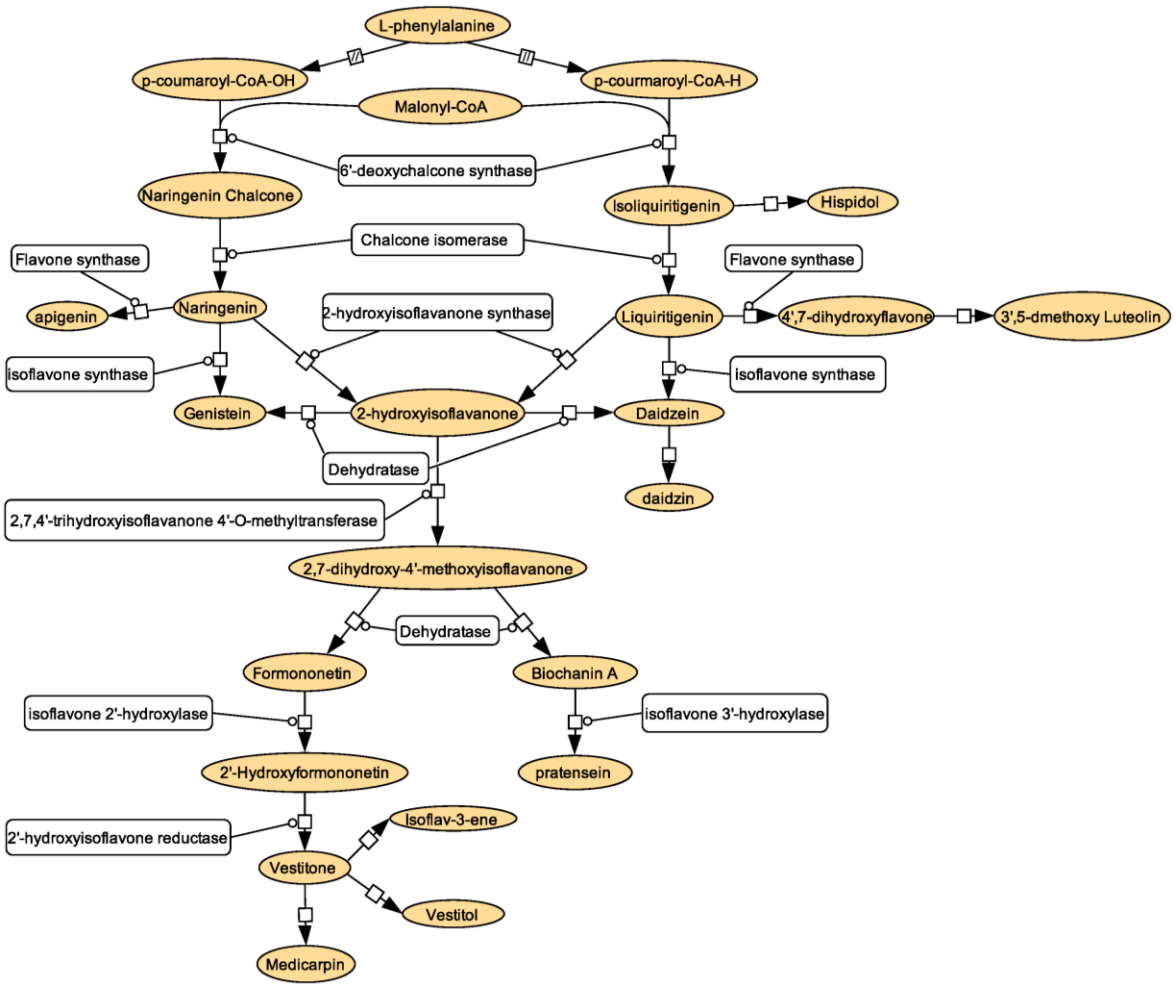


Figure 3.12: *Medicago* flavonoid and isoflavonoid biosynthesis pathway in a process diagram (created in CellDesigner 3.1(Funahashi, Matsuoka et al. 2008)). State node symbols: light amber color filled ovals represent metabolites; rounded rectangles represent proteins. Transition node symbols: arrows represent reactions; the square on an arrow represents a process node; an arrow with two slashes represents known transition omitted; a line with a circle end presents catalysis(Kitano, Funahashi et al. 2005).

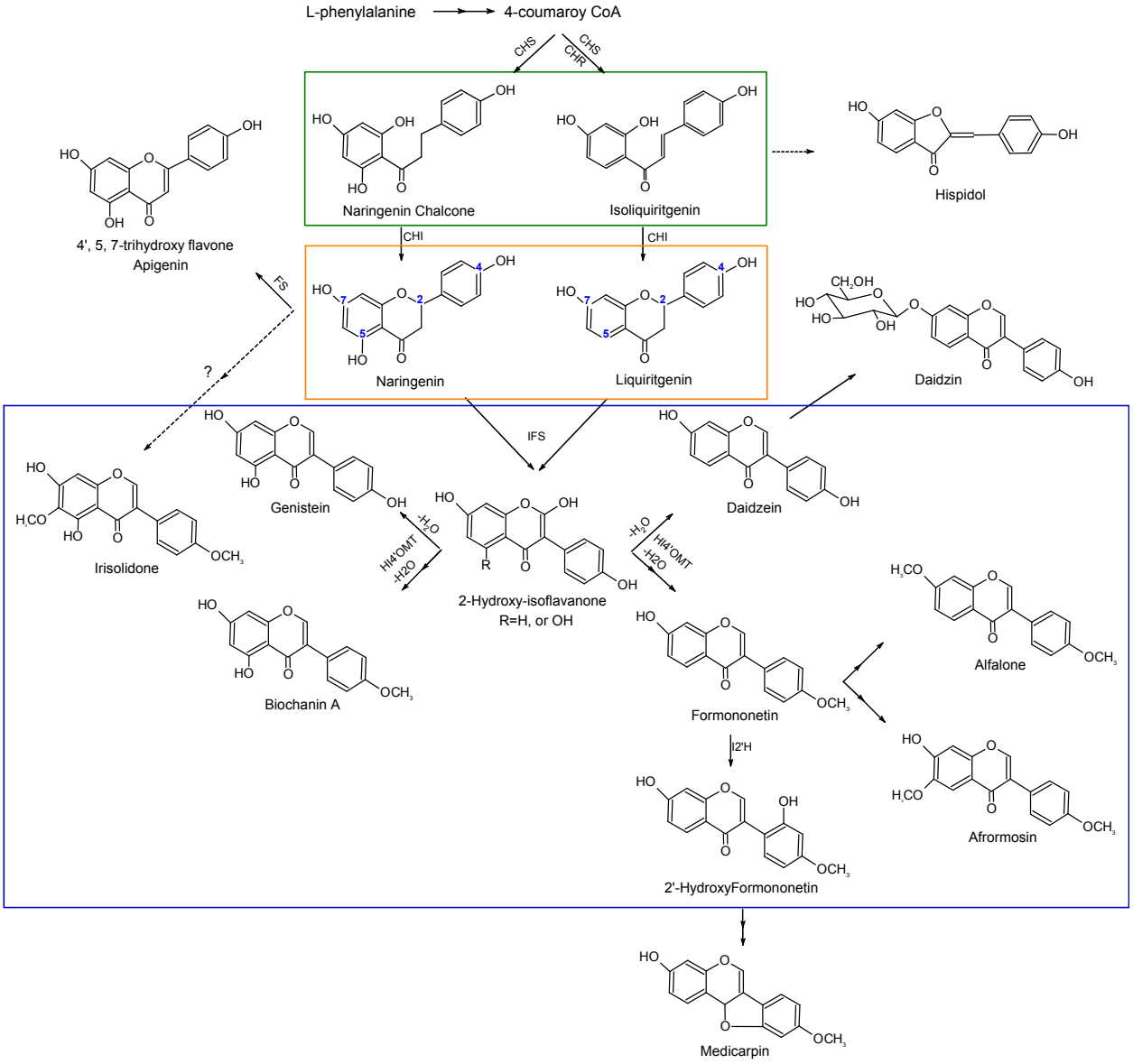
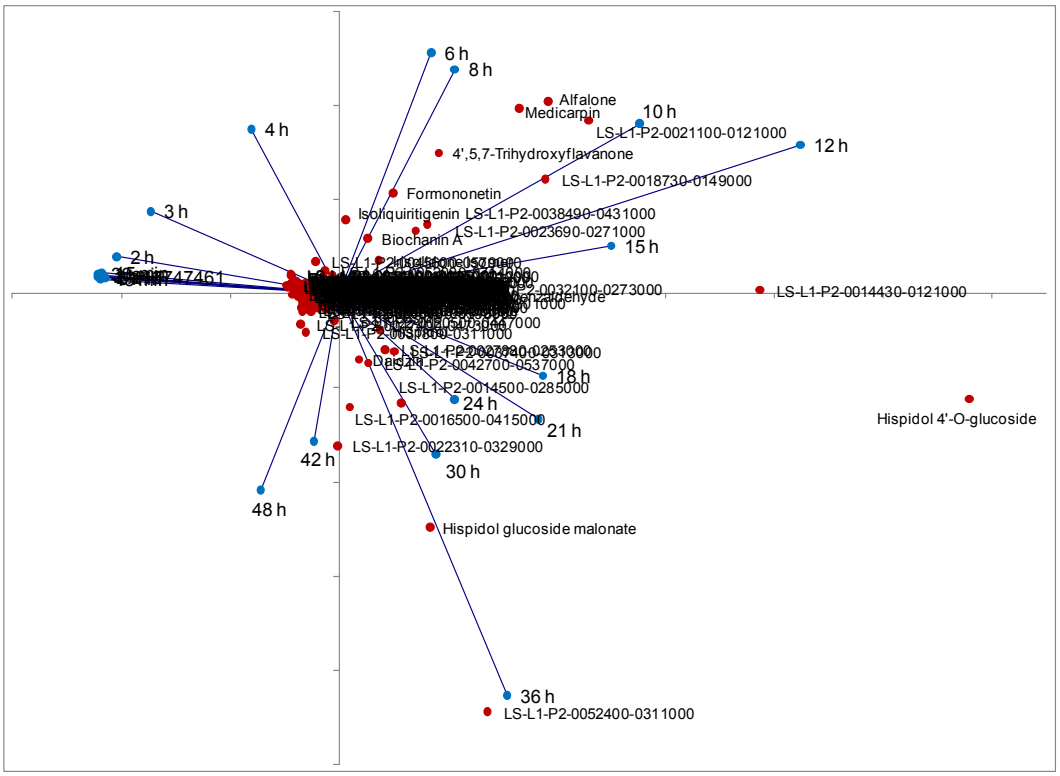
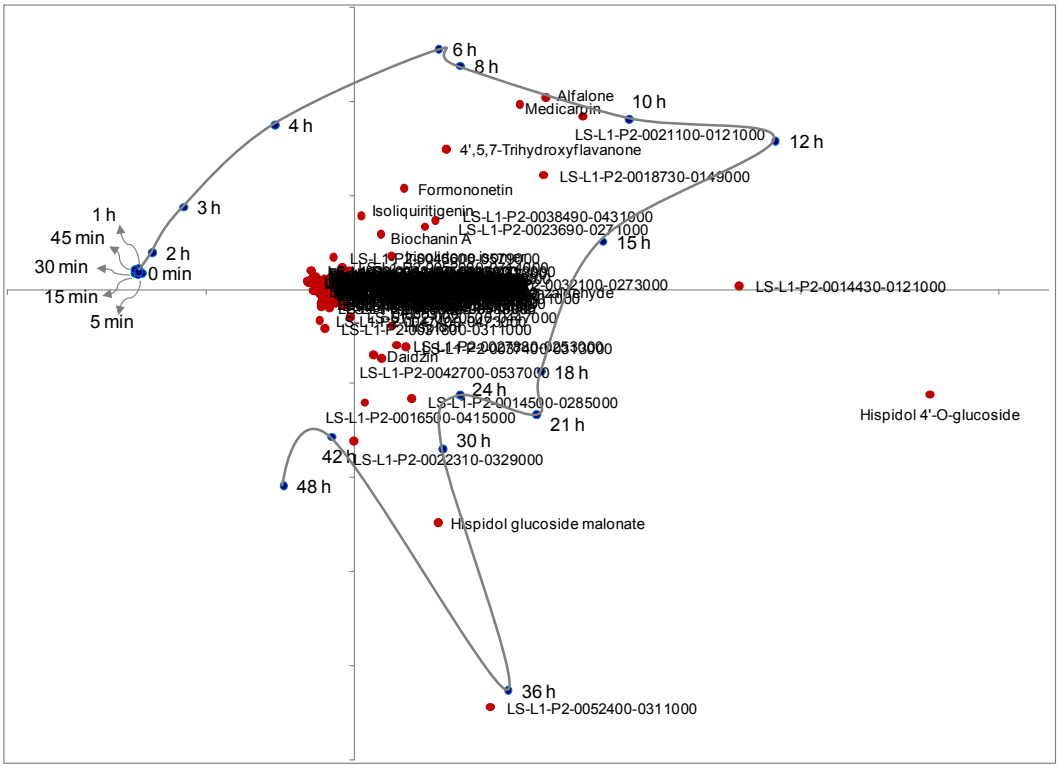


Figure 3.13: A partial diagram of the biosynthetic pathways leading to the major classes of flavonoids in *M. truncatula*: chalcones (encompassed in the green square), flavanones (orange square), isoflavones (blue square), and pterocarpans (medicarpin). Solid arrows indicate established biochemical reactions, whereas dashed arrows indicate possible steps not yet described. The double arrows indicate multiple steps in the biosynthetic pathway. The carbon numbering schema for flavanone is marked. Enzymes are as follows: CHS, chalcone synthase; CHR, chalcone reductase; CHI, chalcone isomerase; FS, flavones synthase; IFS, isoflavone synthase; HI4'OMT, 2,7,4'-trihydroxyisoflavanone 4'-*O*-methyltransferase; I2'H, isoflavone 2'-hydroxylase.

A


6 h
8 h
Alfalone
Medicarpin
10 h
LS-L1-P2-0021100-0121000
12 h
4 h
4',5,7-Trihydroxyflavanone
LS-L1-P2-0018730-0149000
Formononetin
3 h
Isoliquiritigenin
LS-L1-P2-0038490-0431000
LS-L1-P2-0023690-0271000
Biochanin A
15 h
2 h
LS-L1-P2-0032100-0273000
LS-L1-P2-0014430-0121000
LS-L1-P2-0042700-0537000
18 h
LS-L1-P2-0014500-0285000
24 h
LS-L1-P2-0016500-0415000
21 h
Hispidol 4'-O-glucoside
42 h
LS-L1-P2-0022310-0329000
30 h
48 h
Hispidol glucoside malonate
36 h
LS-L1-P2-0052400-0311000

B


6 h
8 h
Alfalone
Medicarpin
10 h
4 h
LS-L1-P2-0021100-0121000
12 h
4',5,7-Trihydroxyflavanone
LS-L1-P2-0018730-0149000
Formononetin
1 h
3 h
Isoliquiritigenin
LS-L1-P2-0038490-0431000
45 min
LS-L1-P2-0023690-0271000
Biochanin A
15 h
2 h
30 min
0 min
LS-L1-P2-0032100-0273000
LS-L1-P2-0014430-0121000
15 min
5 min
Daidzin
LS-L1-P2-0042700-0537000
18 h
24 h
LS-L1-P2-0014500-0285000
LS-L1-P2-0016500-0415000
21 h
Hispidol 4'-O-glucoside
30 h
42 h
LS-L1-P2-0022310-0329000
48 h
Hispidol glucoside malonate
36 h
LS-L1-P2-0052400-0311000

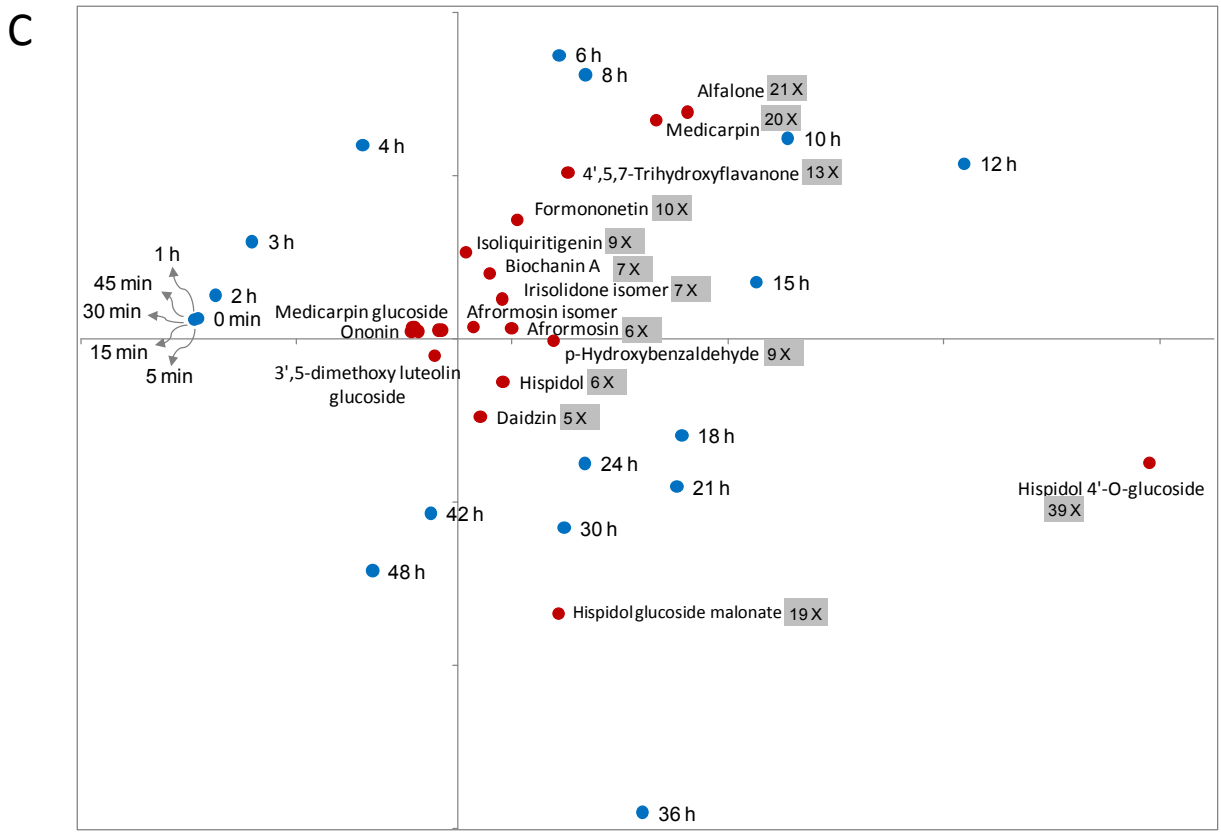


Figure 3.14: Biplot display on selected metabolites following yeast elicitation in *Medicago truncatula* study. The responses of metabolites were obtained with LC-MS analyses. The metabolite data were corrected by internal references ─ flavonoid, and retrieved through the database for 'ome's ─ DOME (http://calvin.vbi.vt.edu/DOME/DOMEMT/index.php). The selected 116 metabolites have significant responses after the elicitation with p value less than 0.01. Rows and columns centering are performed before SVD. Times and metabolites are equally scaled. Ratio of median between elicitation and control were used for SVD. Blue dots represent time points from 0 minute to 48 hours and red dots represent metabolites. A: Biplot display of the times and metabolites ─ the whole picture. B: Phase space biplot display where all the times are connected with a gray line in the sequence of times. C: Biplot display that depicts only the identified metabolites that largely contribute to the pattern of the plot. The gray boxes next to the metabolites illustrate the maximum fold induced by yeast elicitation. 4', 5, 7-trihydroxyflavanone, also known as naringenin, together with liquiritigenin serve as entry points into the flavone and isoflavone biosynthetic pathways.

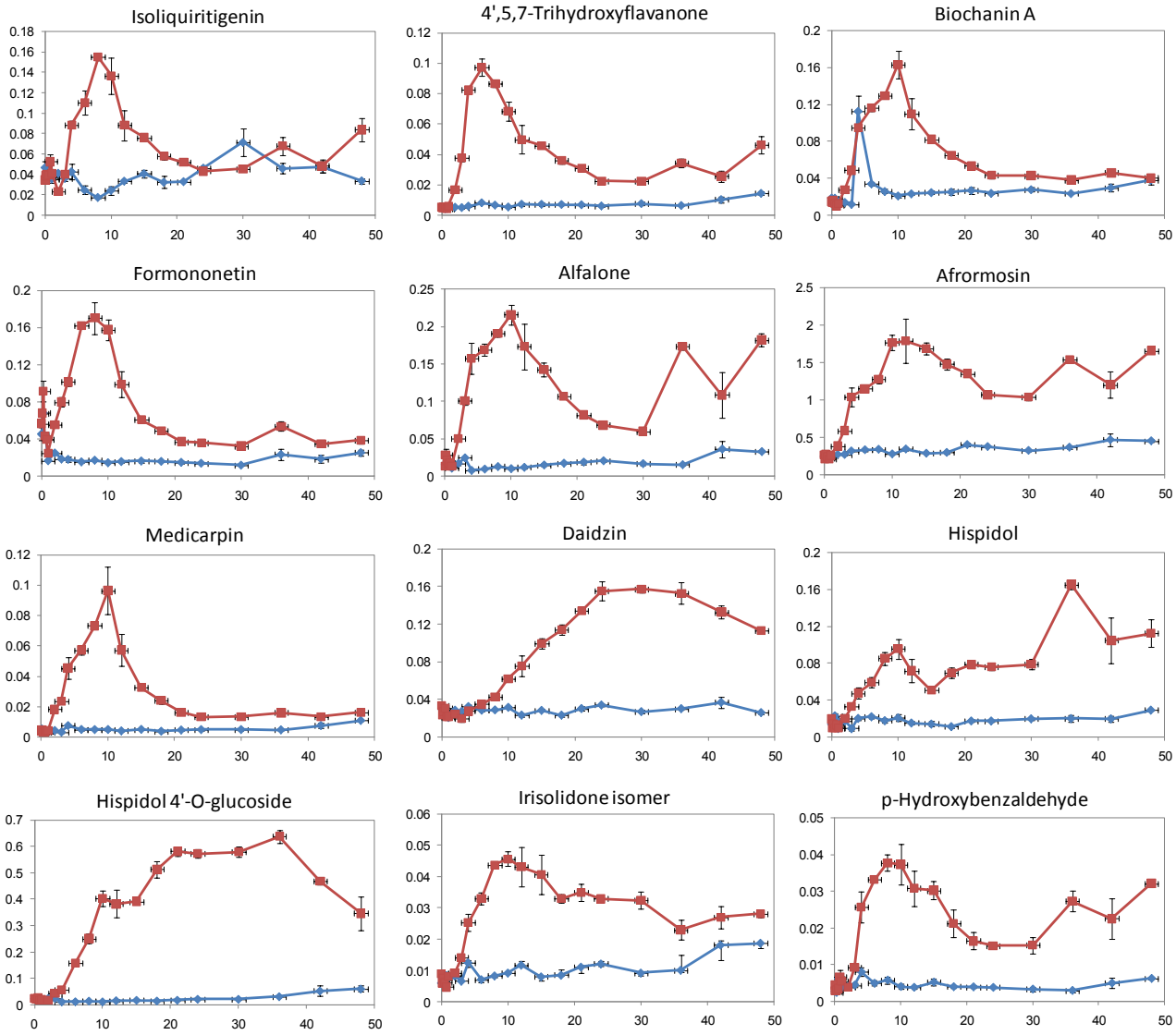


Figure 3.15: Levels of selected metabolites implicated through the Biplot display as major contributors to the plot pattern responding to Yeast elicitation in *Medicago truncatula* study. Y-axis values represent relative peak areas after normalization to the mean peak area for that compound. Maroon squares represent elicited sample means and blue diamonds represent control sample means. Error bars represent standard error from six replicates analyses (three biological replicates and two injection replicates for each biological sample). 4', 5, 7-trihydroxyflavanone is also known as naringenin.

hydroxybenzaldehyde was incorporated into the cell wall matrix following yeast elicitation and the maximum accumulation of this cell wall-bound phenolic happened at 8 to 10 hrs. This result is in agreement with a recent finding by Lei *et al.(Lei, Chen et al. 2010).*

The third group of flavonoids peaked slightly at 8-10 hrs, and declined a little, and then increased, and slowed down by a plateau until reached to a second bigger peak at 36 hrs. The members of this group include two aurones: Hispidol and its glucoside, hispidol4'-*O*-glucoside, which represents a major response to yeast elicitor in M. truncatula cells, with up to 39-fold induction. While many flavones and isoflavones are derived from naringenin and liquiritigenin, hispidol can be directly synthesized from chalcone isoliquiritigenin. Hispidol possessed significant antifungal activity relative to other *M. truncatula* phenylpropanoids tested (Farag, Deavours et al. 2009).

Interestingly, although several primary flavonoids accumulated following yeast elicitation, their corresponding glycosides were less affected with less than 1-fold increase, such as glucosides of formononetin, 2'-OH formononetin, afrormosin, biochanin A, medicarpin, and irisolidone. The exceptions to these are the previously discussed hispidol and the five-fold induced daidzin, daidzein's 7-*O*-glucoside. The response pattern of daidzin is also different from other isoflavones with a sustained increase through 48 hrs.

The phase space biplot display in this study represents a typical biplot of a biological pathway, where metabolites participating in this pathway are correlated and adjacent time variables are highly correlated. The dynamics of the system is well represented.

3.4.5 Discussion

Biplot display is a visualization tool well-suited for data inspection and exploratory data analysis in metabolomics and systems biology study. Biplot display is helpful in quickly identifying the molecules that have highly responded to the perturbation and the dynamics of their responses by the investigation of molecule-time inter-relationship; it also provides a general overview of the complete time course of the system. Biplot analyses on *M.truncatula* study show that instead of individually plotting of over 100 metabolites' responses to elicitation, biplot gives us one diagram with a vast amount of information a researcher might be interested in. It provides a solid foundation for further research.

Although biplot displays are commonly used in the analysis of data from ecological and environmental studies, they have never been utilized in metabolomics studies and only been recently introduced in gene expression data (Pittelkow and Wilson 2005; Chapman, Schenk et al. 2001). Singular value decomposition (SVD) has only been lately used in microarray data

(Wall, Rechtsteiner et al. 2003) and pathway analysis (Price, Reed et al. 2003; Wiback, Mahadevan et al. 2004).

Previous SVD studies on cell-cycle gene expression data show that cell-cycle genes appear to be relatively uniformly distributed about a ring in SVD projections (Wall, Rechtsteiner et al. 2003; Alter, Brown et al. 2000; Holter, Mitra et al. 2000). We note that phase space biplot would be appropriate for dealing with this kind of data.

As any multivariate method, biplot works best with data with less or no noise. When there is strong signal (highly accumulated metabolites) or artificial noise in data, biplot can instantly single out that signal or noise by showing it. We found that further information can be revealed after we have removed dominant signal (molecules) and cleaned the noise. To facilitate usage of biplot, Dr. Mendes has developed a open source software -- OMETER for iterative application of interactively performed biplot analysis (http://mendes.vbi.vt.edu/tiki-index.php?page=ometer).

We chose median of ratios instead of mean of ratios as measure of central tendency (statistical average) based on the advantages of the median over the mean for experimental data. The median is not influenced by extreme measurements in experimental data and median is invariant with respect to most ordinary (monotone) transformations (e.g. the median of log X is the logarithm of the median of X) (Rubin and Smith 1958). The relative advantages of the median over the mean increase as the kurtosis increases (Rubin and Smith 1958). Kurtosis measures the degree of tail heaviness and peakedness of a distribution (Ruppert 1987). And higher kurtosis means more of the variance is due to infrequent extreme deviations, i.e. heavy tail (Chissom 1970). We found that using median instead of mean of the replicates can effectively reduce the impact of variance in the replicates on the analysis.

Among three common data transformations prior to the application of SVD (variance matrix, correlation matrix and interaction residual matrix), we found that interaction residual matrix is the most effective data matrix for metabolomics time course biplot display. To calculate the interaction residual matrix, we may represent row and column “effects” and “interaction residuals” by scalar products in the following way (Bradu and Gabriel 1978):

Column effects as $y_{\bullet j} - y_{\bullet\bullet} = g_{\bullet}'(h_j - h_{\bullet})$,

row effects as $y_{i\bullet} - y_{\bullet\bullet} = (g_i - g_{\bullet})'h_{\bullet}$,

residuals as $y_{ij} - y_{i\bullet} - y_{\bullet j} + y_{\bullet\bullet} = (g_i - g_{\bullet})'(h_j - h_{\bullet})$.

The above residuals represent the multiplicative terms in an additive main effects and multiplicative interaction (AMMI) model and can be visualized through biplots (Gabriel 1971). Many studies have shown AMMI model to be effective in understanding complex genotype x environment interactions in plant breeding (Eeuwijk 1995; Ebdon and Gauch 2002). Our biplot results suggest the promising applicability of AMMI on metabolomics time course data.

When we applied a fixed amount of noise to the data, namely 100 x of the smallest value in the data set, we found that the variables with large variance didn't change much on the biplot (T0.0, T1.0, T3.5 and T4.0 in Fig. 2); while the variables that were drastically affected were with small variance. This simulation explains why we don't have perfect phase-space biplot for real experimental data when we have all kinds of noise contributed from technicians and instruments. We further speculate that when measurement noise dominates real experimental data (related to fixed noise we added), we can use those variables/rays with large variance as landmarks for phase place biplot and confidently draw information from those variables.

We argue that SVD based biplot display represents a new multivariate data analysis method for metabolomics data. Rather than separating molecules into distinct groups based on *a priori* knowledge or on statistical information extracted from the patterns as in classification methods (e.g. clustering) (Mendes 2002; Sumner, Mendes et al. 2003), SVD based biplot shows the relationship among the molecules, among the samples, and between the molecules and samples. It is more effective and informative especially in multi-dimensional space. The phase space biplot is a very compact method to display the dynamics of the large scale data sets used in functional genomics and systems biology.

Chapter 4

Data integration based on phase spectra

4.1 Introduction

4.1.1 Fourier transform

Being an interdisciplinary science, systems biology uses concepts that have come from multiple disciplines (mathematics, engineering, physics, and computer science) and applies them to biological problems (Abraham 2002; Chong and Ray 2002; Ideker, Galitski et al. 2001). The Fourier transform is just another one of these kinds of tools that come from the field of signal processing and is ready to be used in systems biology.

The Fourier transform is based on the discovery that it is possible to take any periodic function $f(x)$ and resolve it into an equivalent infinite summation of sine waves and cosine waves. The resulting infinite series is called the Fourier series (Weisstein 1999; Harris 1998):

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)]$$

For a 2π periodic function $f(x)$ that is integrable on $[-\pi, \pi]$, the numbers

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)dx$$
$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx$$
$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\,dx$$

Where

$n \geq 1$, a_n and b_n are called Fourier coefficients.

Using Euler's formula, the Fourier series can be conveniently expressed as a more compact notation, known as the complex/exponential Fourier series:

$$f(x) = \sum_{n=-\infty}^{\infty} A_n e^{inx}$$

Where:

$$A_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}dx$$

Coefficient A_n can be expressed in terms of those Fourier coefficients in the Fourier series:

$$A_n = \begin{cases} \frac{1}{2}(a_n + ib_n) & \text{for } n < 0 \\ \frac{1}{2}a_0 & \text{for } n = 0 \\ \frac{1}{2}(a_n - ib_n) & \text{for } n > 0. \end{cases}$$

The computation and study of Fourier series is a branch of Fourier analysis, and this concept has been extended to a more general field, harmonic analysis. In Fourier analysis, Fourier transform often refers to the process that decomposes a given function into the basic pieces that can be easily solved. Deriving from complex Fourier series, if we replace discrete terms with continuous terms, we will get Fourier transform (Weisstein 1999; Harrison 2003):

$$f(x) = \int_{-\infty}^{\infty} F(k)e^{2\pi ikx}dk$$

$$F(k) = \int_{-\infty}^{\infty} f(x)e^{-2\pi ikx}dx$$

Here $F(k)$ is called *forward* $(-i)$ Fourier transform, and $f(x)$, *inverse* $(+i)$ Fourier transform.

Fourier transform is widely used in signal processing and time series analysis. In these applications, Fourier transform resolves a time-domain function into a frequency spectrum; it is often called the frequency domain representation of the original function(Oppenheim and Schafer 1975; Smith 1997).

Discrete Fourier transform (DFT), occasionally called the finite Fourier transform, is a transform for Fourier analysis of discretely sampled data. Discrete Fourier transforms are extremely useful because they reveal periodicities in input data as well as the relative strengths of any periodic components (Weisstein 1999). The development of fast Fourier transform (FFT) algorithms, such as Cooley-Tukey algorithm has reduced the number of computations needed for N points from N^2 to $Nlog_2N$ (Press, Teukolsky et al. 1992).

The N- point Discrete Fourier transform of a signal x_n, $n = 0, \dots, N-1$ is defined(Press, Teukolsky et al. 1992; Weisstein 1999; Wikipedia 2009) to be a sequence of N complex numbers X_k, $k = 0, \dots, N-1$, given by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i\frac{2\pi}{N}kn} \quad k = 0, \dots, N-1$$

The original signal $x(n)$ can be recovered by the inverse transform:

$$x_n = \frac{1}{N}\sum_{n=0}^{N-1} X_k e^{i\frac{2\pi}{N}kn} \quad n = 0, \dots, N-1$$

When it's a time-domain to frequency-domain transformation, the above terms can be explained as:

x_n is input signal amplitude (real or complex) at time n (sec);

n is sampling instant (sec);

X_k is spectrum of X (complex valued), at frequency k;

k is sampling rate, frequency variable (samples/sec, or Hertz(Hz));

N is number of time samples, also is frequency samples (integer).

In signal processing, the signal is sampled at evenly spaced intervals in time. The number of times this signal is recorded in a second is called the sampling rate f_{sample} . There is a special frequency called Nyquist frequency, given by:

$$f_{Nyquist} = \frac{1}{2} f_{sample}$$

The Nyquist frequency is the highest frequency that can be coded at a given sampling rate in order to be able to fully reconstruct the signal (Press, Teukolsky et al. 1992; Landsman 1993; Weisstein 1999).

When the original signal is recorded in time, the Discrete Fourier transform has a standard interpretation as the frequency spectrum of the signal. The magnitude and phase of the complex-valued function X_k at frequency k represent the amplitude and phase of the different sinusoidal

components of the input "signal" x_n. By writing X_k in polar form, we immediately obtain the sinusoid amplitude A_k and phase φ_k, respectively(Smith 1997):

$$A_k = MagX_k = |X_k| = \sqrt{Re(X_k)^2 + Im(X_k)^2}$$

$$\varphi_k = \text{Phase}X_k = \arctan\,(X_k) = arctan\left(\frac{Im(X_k)}{Re(X_k)}\right)$$

The amplitude tells you how strong the oscillation is, the phase tells you at what stage of oscillation the system was at time 0. In DFT, instead of looking at magnitude of X_k, often times we study power spectrum of the transform(The Space Telescope Science Institute April 1994; American Society for Cell Biology. 1998).

Power spectrum $S_{(k)}$ is obtained by taking the squared magnitude of the Fourier transform. It shows the total variance of the data. One way of looking at the power spectrum is as a breakdown of this signal variance in components at frequency *k*. The function $S_{(k)}$ assigns a power to every frequency *k* and all of the powers for nonzero frequencies add up exactly to the variance. From a power spectrum, we can tell whether the variance is due primarily to low frequencies, high frequencies or some combination.

Another important property of Fourier transform is Parseval's theorem. According to this theorem (Oppenheim and Schafer 1975), the energy in the time domain is the same as the energy in the frequency domain. Thus, Parseval's theorem gives

$$\left\|\vec{x} - \vec{y}\right\|^2 \equiv \left\|\vec{X} - \vec{Y}\right\|^2$$

The above equation implies that the Euclidean distance between two signals *x* and *y* in the time domain is the same as their Euclidean distance in the frequency domain. Based on this theorem, one can detect the similarity between two time course data in the frequency domain(Agrawal, Faloutsos et al. 1993).

In the Fourier representation of signals, frequency spectrum is usually presented as amplitude and phase, both plotted versus the frequency. When the amplitude is squared, the resulting plot is referred to as a power spectrum. The phase spectrum is usually calculated by taking the arctangent of the ratio of imaginary to real parts of the Fourier transform. For a long time, the power spectrum has been the main focus in research as it answers how much a signal falls under a certain frequency. But it has been found that the phase plays more important roles than the magnitudes in the reconstruction of the original signal (Oppenheim and Lim 1981).

4.1.2 The importance of phase

In many applications of Fourier transform, such as in electronic engineering, only the amplitude information, especially square of the amplitude (power spectrum) is used and the phase information is discarded. However, despite this common practice, phase information should not be ignored. Research has long demonstrated the importance of phase in signal reconstruction (Oppenheim and Lim 1981), in applications such as 2-D and 3-D medical imaging (Vilain, Daou et al. 2001), active galactic nucleus variability in black-hole physics (Karas 1997), recovery of the temporal and spatial characteristics of a source in acoustics research (Sachse and Pierce 1990), and the retrieval and classification of audio signals from multimedia databases (Paraskevas and Chilton 2004).

In the Fourier representation of signals, spectral magnitude and phase tend to play different roles and in some situations, many of the important features of a signal are preserved if only the phase is retained (Oppenheim and Lim 1981). Oppenheim and Lim have used experiments to illustrate that if we construct synthetic images made from the amplitude information of one image and the phase information of another, it is the image corresponding to the phase data that we perceive, if somewhat degraded (Oppenheim and Lim 1981; Oppenheim, Lim et al. 1983).

Figure 4.1 gives a depiction of how important phase is in electron density maps (Read 2009). It's a remake of Oppenheim and Lim's experiment. On the top are pictures of Jerome Karle (left) and Herb Hauptman (right), who were 1985 Nobel Laureates in chemistry for their work on solving the phase problem for small molecule crystals. Read treated the pictures as density maps and calculated their Fourier transform to get amplitudes and phases. He found that if we combine the phases from the picture of Hauptman and the amplitudes from the picture of Karle, we get the picture on the bottom left. The bottom right picture is the combination of phases of Karle and amplitudes of Hauptman. Clearly phase has dominated the reconstructed image. It indicates that phase is far more important than amplitude is for determining the electron density map.

Ni and Huo have also shown the importance of phase from a statistical point of view (Ni and Huo 2007); they found that if the phases are randomly re-assigned, the reconstructed signal is likely to be severely distorted. At the same time, if the magnitudes are randomly re-assigned, the distortion is automatically controlled within a region, whose size is given by the special structure of the Discrete Fourier transform together with the distribution of the signal. Hayes *et al*. have shown that if a one-dimensional discrete-time signal is of finite length and completely specified to within a scale factor by the phase of its Fourier transform, then phase information alone is sufficient for signal reconstruction (Hayes, Jae et al. 1980).

Figure 4.1: Importance of phase in density map (Read 2009). Used with permission from the author Randy Read. The pictures of Jerome Karle and Herb Hauptman were treated as density maps and Fourier transforms were computed to get their phases and amplitudes. Top left ─ Original picture of Jerome Karle. Top right ─ Original picture of Herb Hauptman. Bottom left ─ Image synthesized from the phase of Herb Hauptman and amplitude of Jerome Karle. Bottom right ─ Image synthesized from the phase of Jerome Karle and amplitude of Herb Hauptman. The results clearly indicate that phase information dominates in reconstructed density map.

The basic idea in this study is to take the FFT of each series, compute their phase spectrum, then measure similarity between series using these spectrum parameters. The objectives of this study are to look at how phase spectrum helps us compare the behavior of different molecules and provide insight in the biochemical network.

4.2 Methods

4.2.1 *in silico* network ─ Claytor Network

Although applications of data integration algorithms to experimental data from real systems are the ultimate goal in our systems biology study, it should be noted that true assessment of these methods requires applying them to artificial data such that the outcome of their analyses can be compared with the exact mechanisms that created the data (Mendes 2009). This applies as much to network inference and parameter estimation as it does to data integration in systems biology.

Since the release of biochemical reaction simulator ─ Gepasi (Mendes 1993; Mendes 1997; Mendes and Kell 1998) and its successor COPASI (Hoops, Sahle et al. 2006; Mendes, Messiha et al. 2009), similar software packages for kinetic modeling biochemical networks have become available (Alves, Antunes et al. 2006). To test the performance of our integration algorithms, we need an *in silico* network model that includes metabolism, signal transduction and gene regulation.

Claytor network, an *in silico* network model, was created by Mendes (Mendes 2009) using the software COPASI (Hoops, Sahle et al. 2006). It includes three levels of regulation: gene expression, signal transduction and metabolism.

The Claytor network (Figure 4.2) was generated to mimic a biochemical network in order to benchmark inference algorithms (Mendes 2009). It contains a total of 59 state variables: 16 metabolites (M1─M24 except M5, M8, M14, M15, M18, M19, M21, and M23), 23 protein forms (P1─P23), and 20 genes (represented by levels of their mRNA) (G1─G20). It also contains 8 external metabolites (M5, M8, M14, M15, M18, M19, M21 and M23) that are part of the environment. The environmental perturbations are achieved by changing the concentration of the main substrate (M23) and of the toxic substance (M1).

The network contains two protein receptors and associated signaling pathways that sense two different conditions. Protein P15 is a receptor that binds to the toxic substance M1, forming a complex P22 which acts directly as a transcription factor that induces several genes (G1, G2, G3, G4, G5, G16 and G20). Protein P18 is a receptor that binds metabolite M3, and their complex is also a transcription factor (P23) that represses three genes (G7, G8 and G9) and induces another one (G10). The Claytor network also incorporates the possibility of a protein having two different states (similar to what happens with phosphorylation and dephosphorylation of proteins in signaling pathway), with P20 being “activated” by P2 in the presence of M4, with its active form being P21, that catalyzes the conversion of M9 to M10 (Camacho 2007).

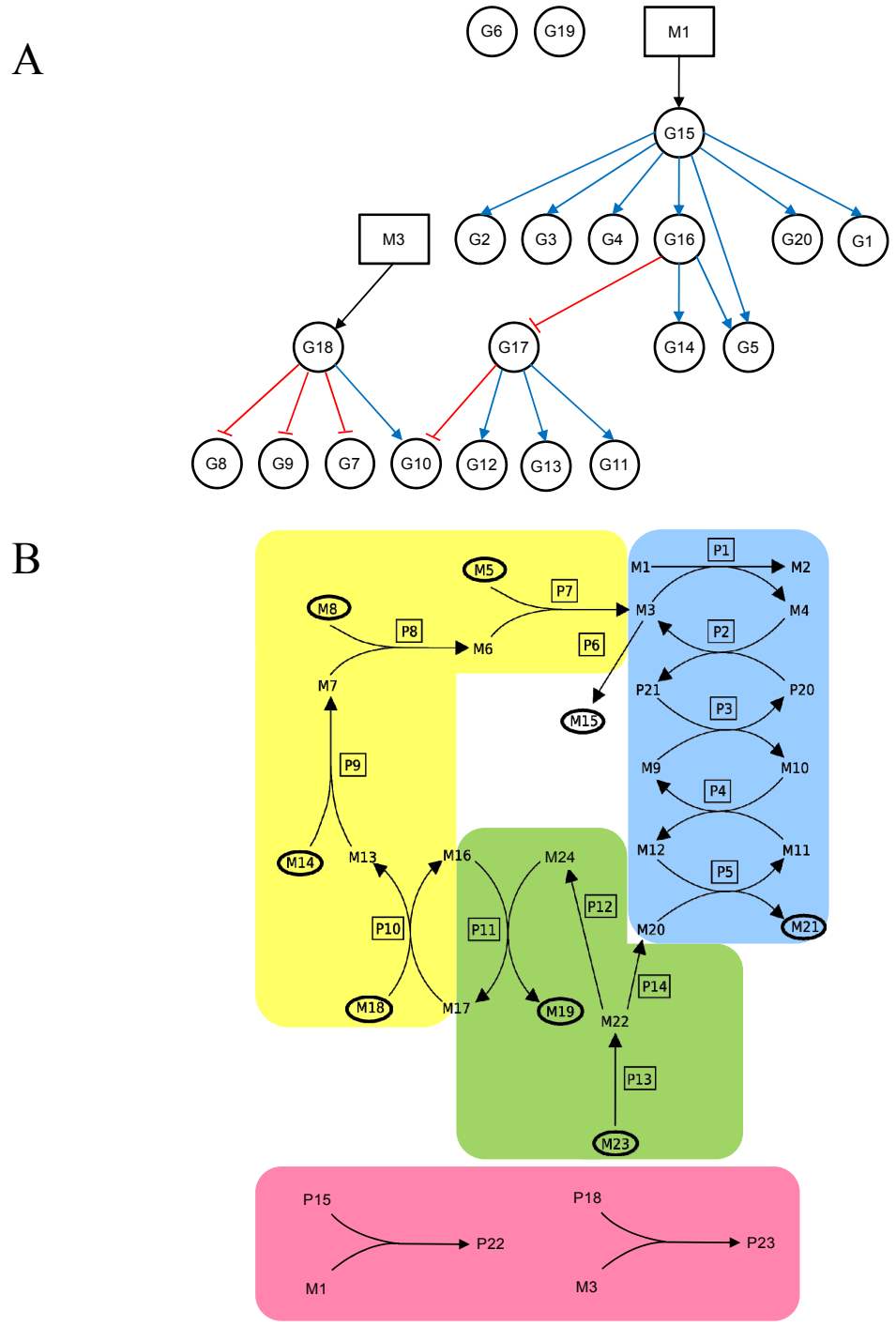


Figure 4.2: Claytor network model (Mendes 2009). A: diagram of the gene interaction; blue pointed arrows represent induction, red arrows with blunt ends represent repression, circles represent genes, squares represent metabolites. B: diagram of metabolism and signaling pathway; three pathways are indicated, green section represents catabolism which provides energy and reducing power from a substrate; yellow section represents biosynthesis, which uses energy to absorb nutrients and construct an essential molecule; blue section represents redox chain that reduces the toxic substance M1; Pink section represents two protein receptors and their formation of transcription factors. Squares represent proteins; ellipses represent metabolites that are added in the "medium"/environment.

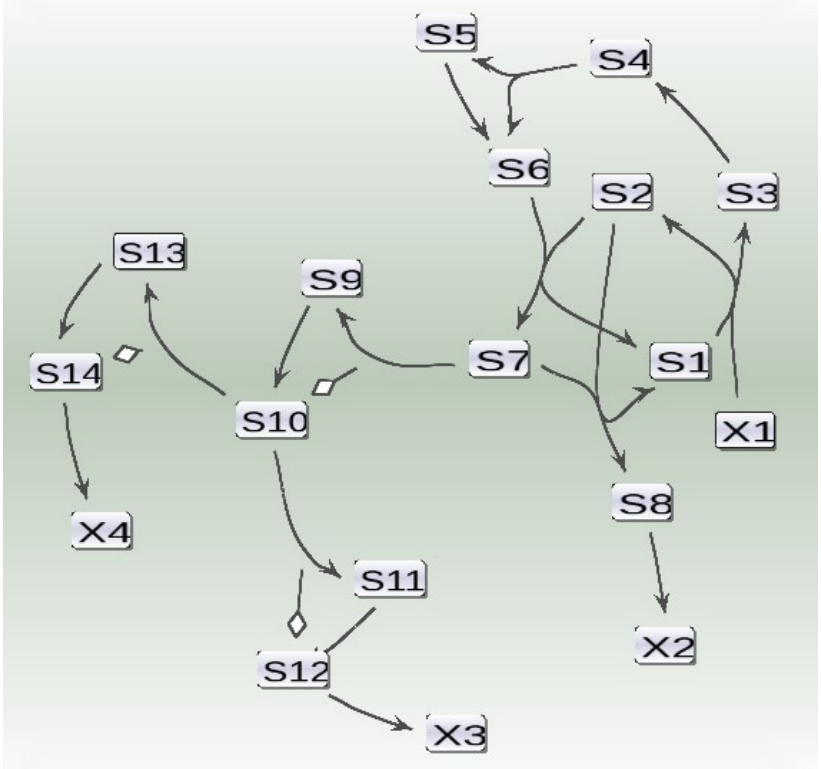


Figure 4.3: AB2 network model (courtesy of Pedro Mendes): Diagram of metabolism and signal pathway. Squares represent metabolites. Lines with diamond ends indicate inhibition, e.g. S10 inhibits the reaction from S7 to S9 and S12 inhibits the reaction from S10 to S11.

The complexity of Claytor network and its inclusion of metabolism, signal transduction and gene regulation are not normally present in benchmark models. It makes the Claytor network an ideal and only available *in silico* model to study data integration with.

4.2.2 Artificial Biological network ─ AB2 Network

To further investigate whether phase spectrum produced from Fourier transform has a broad appeal in biological networks, a second artificial biological network (Figure 4.3) was generated by Pedro Mendes. Like Claytor network, AB2 is a complex network that includes three levels of regulations: gene expression, signal transduction and metabolism. It contains a total of 57 state variables: 14 metabolites (internal) S1, S2, …, S14; 20 proteins, two of them have different states, P1, P2, …, P16, P18, P16a, P18a; and 18 genes (represented by levels of their mRNA) G1, G2, …, G18. It also contains 5 external metabolites that are part of the environment, X1, X2,…, X5. The environmental perturbations are achieved by changing the growth rate and concentration of the external metabolites (X1, X3 and X4).

4.2.3 Data generation with Claytor Network model

To generate data, the experiments are carried in the following procedures:

- the network is allowed to enter a steady state, at M1=0 and M2=0;
- the network is then perturbed in three different combinations of concentration of M1 and M23, and these experiments named as wt-1,wt-2, and wt-3 respectively (Table 4.1):

Table 4.1: Perturbations applied to the Claytor network. After the network is allowed to reach a steady state, an experiment is performed in which an environmental perturbation is applied to the system.

Perturbations	M1	M23
wt-1	0.2	0.4
wt-2	1.0	0.4
wt-3	0.2	0.01

Concentrations of each of the molecules in the network in response to the perturbations are simulated with COPASI and recorded at 50 time points, which begins at zero seconds, with duration of 120 seconds (approximately 0.417 Hz).

4.2.4 Data generation with AB2 Network model

For the AB2 network, four experiments (perturbations) are conducted. They all start from a reference steady state:

X1=1, X2=1, X3=1, X4=1, X5=1, growth rate=0.1

The following parameters were changed at time zero:

AB2-1b: growth rate = 1

AB2-2b: X1 = 0.01

AB2-3b: X3=10, X4=10, X1=0.01

AB2-4b: growth rate= 0.01, X1=10

Molecule concentrations are recorded at 256 time points, which begins at zero seconds, with a duration of 1020 seconds (approximately 0.251 Hz).

4.2.5 Fast Fourier transform with Mathematica

To conduct FFT (Fast Fourier transform) of the simulation time course data, a program called "FFT_Coeff_ COPASI" was created with Mathematica. The program takes time course data produced with COPASI as input (in which rows represent time and columns represent molecule concentration variables). It performs FFT on each molecule concentration, then extracts, computes and exports Fourier coefficients, power spectrum and phase spectrum with frequencies up to the Niquist frequency. The zero frequency term appears at position 1 in the resulting list. The value of Niquist frequency equals to half of the sampling rate. In addition, using the drop-down menu to select the desired molecules, the program produces plot displays and plot overlays of any two molecules for the time course, power spectrum and phase spectrum. The exported results are used in clustering analysis and biplot display.

4.2.6 Phase unwrapping

To remove discontinuity, a phase unwrapping algorithm is employed. "Unwrap-N" was written with Mathematica. It is intended to correct the radian phase angles in a matrix P by adding or subtracting 2π when absolute jumps between consecutive elements of P are greater than or equal to the default jump tolerance of π. The unwrapping process operates row-wise. This program will take "FFT_Coeff_ COPASI" produced phase output, in which rows are molecules and columns are frequencies.

4.2.7 *k*-Means Clustering analysis with MeV 4.3

Phase spectrum data produced with Mathematica are loaded into the software ─ MultiExperiment Viewer (MeV) *v*.4.3 (Saeed, Sharov et al. 2003). The method of figure of merit (FOM) is used to choose *k*, the number of clusters before the analysis. A figure of merit is an estimate of the predictive power of a clustering algorithm. It is computed by removing each sample in turn from the data set, clustering genes based on the remaining data, and calculating the fit of the withheld sample to the clustering pattern obtained from the other samples. The lower the adjusted FOM value is, the higher the predictive power of the algorithm. The module *k*-Means Support (KMS) is used to run the *k*-Means Clustering (KMC) algorithms multiple times using the same parameters in each run. It can generate clusters of molecules that frequently group together in the same clusters ("consensus clusters") across multiple runs. The output consists of consensus clusters in which all the member molecules clustered together in at least *x*% of the *k*-Means runs, where *x*% is set as 80%. Pearson Correlation is chosen as the distance metric to remove the effect of different magnitudes. The *k*-means clustering is set to run 100 times. During each KMC run, the maximum number of iterations is 50.

4.2.8 Biplot displays

After molecules with constant values are removed, rows and columns are centered; Singular Value Decomposition (SVD) and biplot display are conducted upon phase spectrum data. Times and molecules are equally scaled. Two programs are used to carry out these procedures. A program written with Mathematica performed the data processing, consisting of SVD, 2D and 3D biplot display. The software written by Lipkovich and Smith ─ Biplot and Singular Value Decomposition Macros for Excel© (Lipkovich and Smith 2002) carried out the SVD and 2D biplot display. The add-in Macros for Excel allowed customized editing on biplot for a better visualization.

4.3 Results and Discussion

4.3.1 Phase unwrapping

The arctangent function calculates phase angles that are constrained to an interval -π to π, although the true phase angles are not limited to this range. Consequently this computation artifact causes that any angle outside this range is wrapped around zero (Paraskevas and Chilton 2004).

We have seen these phenomena when some molecules' phase spectrum plots have abrupt jumps. Two examples are illustrated in Table 4.2 and Figure 4.4: Since arctangent function produces an inherently wrapped output, phase is returned in a form that suffers from 2π jumps. In Table 4.2, phase angle of M4 "plunged" from 2.857 rad at 0.0167 Hz to -3.133 rad at 0.025 Hz, and phase angle of P15 "jumped" from -2.965 rad at 0.0333 Hz to 3.126 rad at 0.0417 Hz (indicated with gray highlights). The corresponding phase plots show that the phase response maximum of M4 has "wrapped around" to the bottom of the plot (Figure 4.4, top left) and phase response minimum of P15 has "wrapped around" to the top of the plot (Figure 4.4 bottom left).

The discontinuity caused by wrapped output has rendered the phase data unusable. Among 48 non-constant molecules in the Claytor network experiments, 11 time series have wrapped phase responses (23%). This has created problems in the later *k*-means clustering analysis and biplot display. More than one fifth of the 'strangely-behaved' molecules inevitably distorted the clustering analysis. For instance, M4 and P15 are classified in the "unassigned" cluster before unwrapping. It is necessary to remove discontinuity by employing a phase unwrapping algorithm (Smith 2007; Paraskevas and Chilton 2004). A program I wrote in Mathematica, "Unwrap-N" has been used to detect the jumps and to remove the discontinuity. With this algorithm, consecutive phase angles are compared, when a jump over π has been detected, appropriate adjustment to the phase angle can be made to "smooth out" the phase plot. Analytically, once the jump has been identified, 2π is added to the rest of phase spectrum of the signal when the point before jump has a higher value and vice versa. M4 and P15, after unwrapping process, have achieved smooth plots, and become continuous (Figure 4.4, top right and bottom right). Previously put in "unassigned" cluster due to their discontinuity, M4 and P15 have been reclassified in Cluster 3 and 4 respectively.

The unwrapped phase data also created a meaningful biplot display and Frequency trajectory (Figure 4.10) instead of a distorted plot from the wrapped phase (unpublished results).

Table 4.2: Wrapped phase output (2π jumps) inherited from arctangent function. The listed are part of phase responses of M4 and M15 corresponding to frequencies from 0.00833 to 0.0417. The gray highlights show that phase angle of M4 "plunged" from 2.857 rad at 0.0167 Hz to -3.133 rad at 0.025 Hz, and phase angle of P15 "jumped" from -2.965 rad at 0.0333 Hz to 3.126 rad at 0.0417 Hz.

Frequency	0.00833	0.01667	0.025	0.0333	0.0417
M4	2.212	2.857	-3.133	-2.955	-2.808
P15	-2.097	-2.547	-2.831	-2.965	3.126

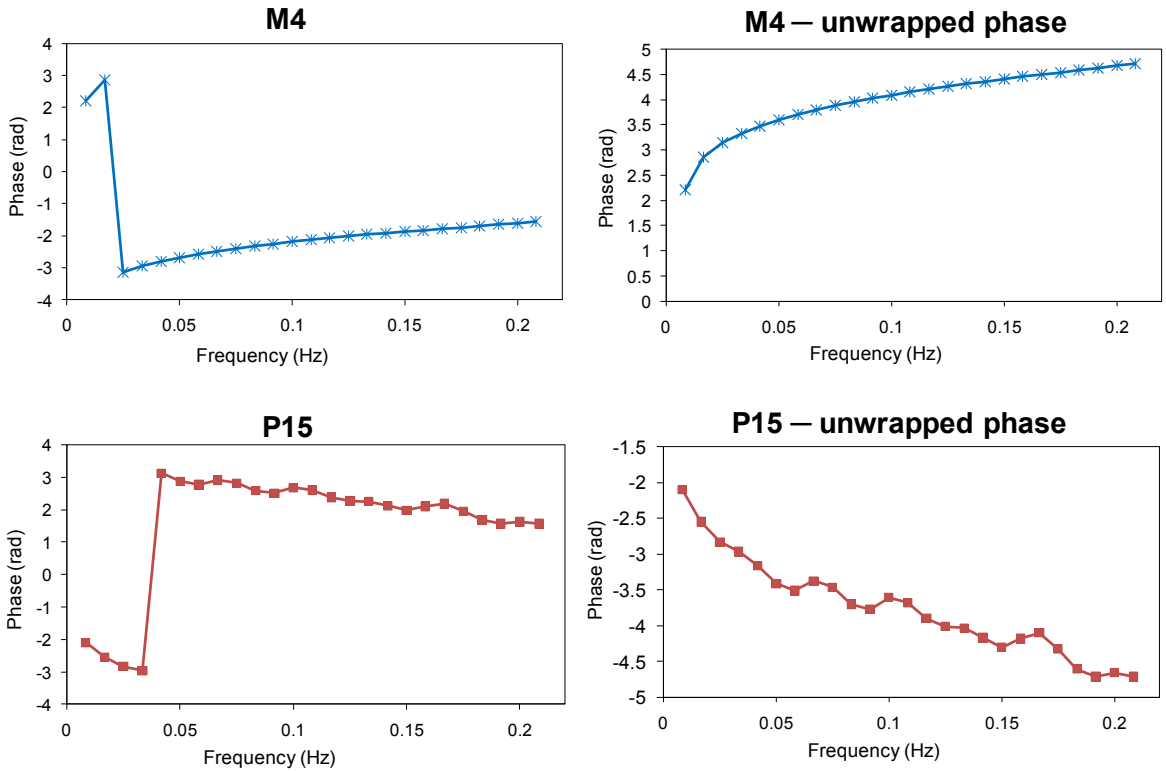


Figure 4.4: Phase unwrapping to remove phase discontinuity (on molecules M4 and P15). A program written in Mathematica, "Unwrap-N" was used to remove discontinuities. Top left ─ phase spectrum of M4 before unwrapping, phase response maximum of M4 has "wrapped around" to the bottom of the plot; Top right ─ unwrapped M4 phase plot, which is continuous; Bottom left ─ phase spectrum of P15 before unwrapping, phase response minimum has "wrapped around" to the top of the plot; Bottom right ─ unwrapped P15 phase plot, which is continuous.

4.3.2 The impact of data precision on the analysis results

Precision is one of the important indicators in data quality. Data precision in systems biology study is confined by each individual methodology that is involved in the study. For instance, the improvement of the methodologies: background adjustment (Yang, Buckley et al. 2001), normalization (Bolstad, Irizarry et al. 2003) and probe set definition (Irizarry, Bolstad et al. 2003; Sandberg and Larsson 2007) leads to a better data reproducibility in microarray data analysis. To study the impact of data precision on phase data analysis, we created data with different precisions using an artificial biological network—Ab2.

The common convention to express precision is by means of significant figures. Here we produce data with 6, 3, 2 and 1 significant figure respectively (Sig.fig.). I computed their Fourier transforms and obtained the phase data. The phase data then went through unwrapping process to remove the discontinuity. Biplot display was performed on the unwrapped phase data and results are shown on Figure 4.5. We see a nice phase space trajectory of frequency on Figure 4.5A (data Ab2-3b with six significant figures): the frequency arrays, from low to high, start from the 2nd quadrant of Cartesian plane and progress through 3rd, 4th and 1st quadrant of the plane. Molecules approximately fall into four clusters: the first group resides in 2nd quadrant, G18, P18, S8 and S9, which are clearly associated with the low frequency of the trajectory; the second group has only two molecules,

S7 and P16, which are in 3rd quadrant and associated with the mid-low frequency; the third group falls on the border between 3rd and 4th quadrants. It includes the following molecules and associates with mid-high frequency, G12, G13, S1, S2, G15, S10, S13, and S14; the last group includes the rest of the molecules and clusters in 1st quadrant. They are associated with high frequency.

When the degree of precision is reduced to three significant figures (Figure 4.5B), the low frequency part of the arrays are preserved. Although the variances of mid to high frequency vectors become erratic due to the loss of precision, the overall pattern still holds. Some molecules' projections on their associated frequencies have increased, which is illustrated as proteins P16, P18 and metabolite S8 have scattered farther away from x-axis. Many more molecules have moved towards the center of the plot, such as S1, S2, S10, S13, S14, G12, G13, G15, P2, P3, G14, S3, S4, P4, and P7.

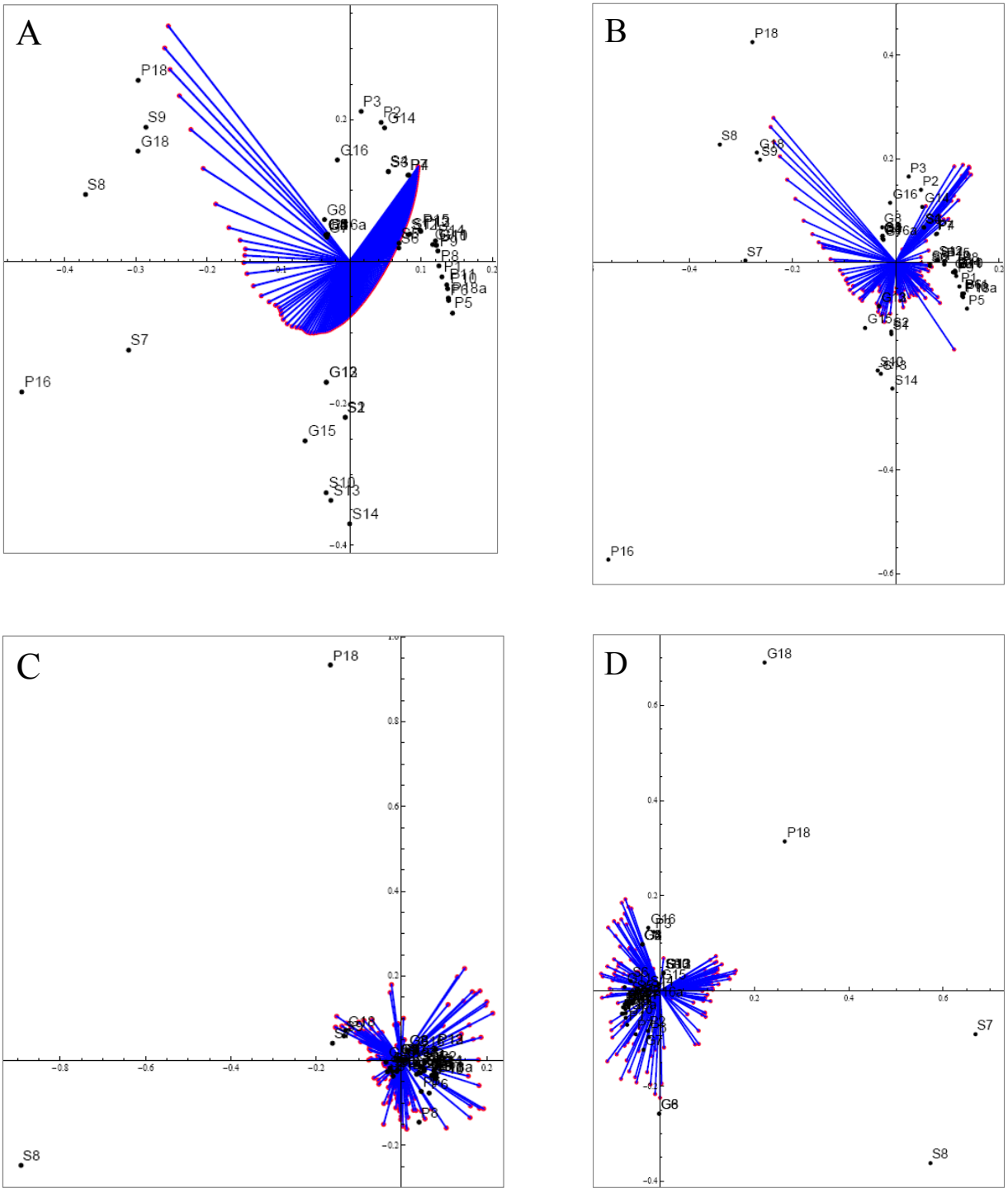


Figure 4.5: Biplot displays on unwrapped phase data of Ab2-3b with different degrees of precision. A: Ab2-3b with 6 significant figures. B: Ab2-3b with 3 significant figures. C: Ab2-3b with 2 significant figures. D: Ab2-3b with 1 significant figure. Phase data was obtained using Fourier transform of Ab2-3b and unwrapped to remove discontinuity.

While the phase space pattern of the biplot display is mostly unaltered with the data of three significant figures, further reduction of precision has categorically changed the analysis results. Please note that due to the nature of the singular value decomposition and biplot display, Figure 4.5D is better compared to others as the mirror image of itself. Figure 4.5C (data Ab2-3b with two significant figures) and Figure 4.5D (data Ab2-3b with one significant figure) indicate that the variance of the low frequencies decreases when the data precision drops to 2 and 1 significant figures, which shows as the obvious disappearance of the 'long' low-frequency arrays. Further examination of the arrays revealed that the close correlations among the adjacent frequencies were lost.

It's not difficult to understand that loss of precision has deteriorated the phase spectrum analysis and phase space biplot display. I would draw an analogy to the electron density map fitting in crystallography: Fitting individual atoms depends on the level of resolution (Evans 2005). In x-ray crystallography, we use Fourier transform to get electron density maps from diffraction patterns, then electron density maps have to be interpreted to determine the protein structure (electron density fitting). At high resolution (1.0Å), individual atoms can be fitted easily. When resolution is reduced, the fitting becomes less easy, until at 4Å, the fit is very uncertain (Figure 4.6).

Less intuitive is the observation that when precision decreases to one significant figure, more molecules seem visible from the biplot display, namely S7, G8 and G18 (Figure 4.5D). It can be explained that the variances of these molecules actually increase when one more significant figure is dropped. Table 4.3 shows the variances of some molecules at different significant figures, 6, 2 and 1 Sig.fig. Molecules like S9 and P16 have nuanced change during the experimental time, so they have very small variances to start with (E-05). Losing precisions make them *de facto* constant. Although S7, G8 and G18 have decreased variances at 2 Sig.fig., the drop of one more Sig.fig. actually increases their variability (highlighted with red color in Table 4.3). Therefore despite the fact that vast majority of the molecules have 'sunk' into the center of the plot, some molecules (S7, S8, G18 and P18) stand out and become more visible.

The artificial biological network, Claytor network and Ab2 network have allowed us to study many features that would not otherwise be possible in real experiments. For instance, from COPASI, we can choose as many significant figures as desired for output. While in real experiments, three significant figures are most feasible (sometimes less). So two questions arise: First, does phase analysis on Fourier transformed data fare well with real-experimental data, i.e. data of low numerical precision? Judging from the biplot display of phase data of Ab2-3b with different degrees of precision (Figure 4.5A and B), phase analysis is pretty robust till numerical precision is three significant figures, we can safely say that this algorithm will perform well with most real experimental data with at least three significant figures.

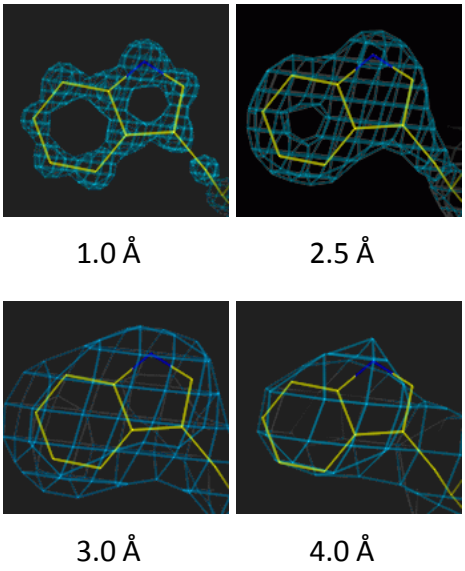


Figure 4.6: Electron density maps at different levels of resolution (Evans 2005). Used with permission from the author Phil Evans. At 1.0Å, there is no problem fitting individual atoms (and the N atom is bigger than the carbons). At 2.5Å the ring is easily fitted, at 3.0Å less easily, and at 4Å the fit is very uncertain.

Table 4.3: Losing precision has changed the variance of some molecules in Ab2-3b data. The variances of genes G8 and G18, of metabolite S7 and S9, and of protein P16 are shown respectively at 6, 2 and 1 significant figures (sig.fig.). Percents of change are compared between variances of data from the reduced precision (2 and 1 Sig.fig.) and the original 6 sig.fig. data. They are illustrated at top of each cell. Increases of %Change are highlighted with red color and decreases with green color. The variances of S9 and P16 decrease at both 2 Sig.fig. and 1 Sig.fig. While variances of S7, G8 and G18 decrease at 2 Sig.fig., but increase at 1 Sig.fig.

Variance	S7	G8	G18	S9	P16
6 Sig.fig.	3.350	0.00938	0.00625	6.58E-05	3.19E-05
%Change	-0.0982	-3.42	-0.869	-2.06	-100
2Sig.fig.	3.347	0.00906	0.00619	6.45E-05	0
%Change	1.23	23.9	14.2	-70.9	-100
1Sig.fig.	3.391	0.0116	0.00713	1.91E-05	0

The second question we might ask is what type of molecules work well with this algorithm especially under low precision? To answer this question, we examined the molecules whose variances didn't decrease under extreme situation ─ one Sig. fig. Molecules S7, S8, P18 and G18's concentrations or expression level are plotted against time in Figure 4.7. Unlike many other molecules that are monotonic increasing (e.g. S1, P11, G9, and G10), or monotonic decreasing (e.g. S2, P5, P6, and P8) during the recorded time, S7 and S8 are wavelet shaped, and P18 and G18 valley shaped. They have demonstrated a unimodal function, which refers to a function that has only one local maximum or minimum, collectively extremum under an extended definition(Pike 2001).

We know that the responses of these "virtual" molecules under a stress over time are the solutions of their corresponding coupled ordinary differential equations (ODEs). This has resulted different time course plots in response to their solutions. When the solution resembles a critically-damped or over-damped oscillation, the system returns to the equilibrium position monotonically without oscillation (Weisstein 2010; Zhu 1998; Li and Lu 2008; Kortemeyer 2009). This is the case of molecules that are monotonically increasing or decreasing. When the solution resembles a brief oscillation, like a wavelet or valley, the amplitude returns to equilibrium after one oscillation. It's conceivable that decreasing precision is less likely to change the shape of the molecule's response plot. Therefore, molecules with this feature are more tolerant with low precision over most analytical methods. Meanwhile, since Fourier analysis grew from the study of periodic functions, molecules with oscillatory behavior are better-suited for phase analysis on their Fourier transforms.

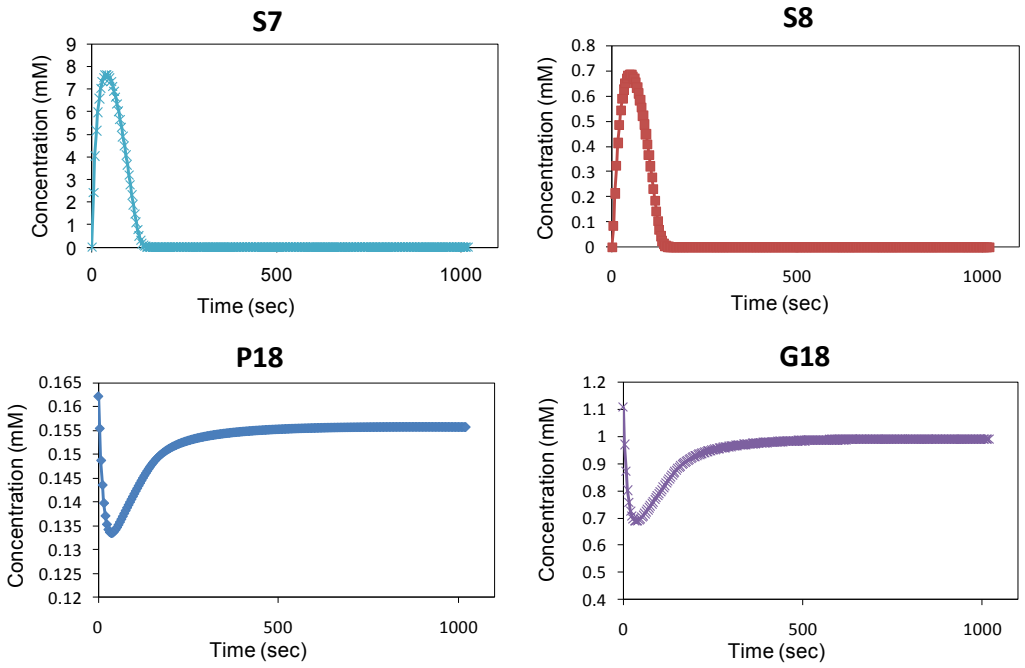


Figure 4.7: Plots of molecules in Ab2-3b data with a high tolerance for low precision. The concentrations or expression level of molecules, S7, S8, P18 and G18 are plotted against time. All the above plots demonstrate a unimodal function. S7 and S8 are "wavelet" shaped with one local maximum; P18 and G18 have a "valley" with one local minimum.

4.3.3 *k*-means clustering analysis on phase spectra

It's debatable whether *k*-means clustering (KMC) is a good tool to study systems biology data since it has some disadvantages, such as, it doesn't yield the same result with each run, its lack of assistance in choosing the number of clusters and its scaling issue (Rencher 2002; Magidson and Vermunt 2002). Since Claytor network comprises several biological pathways (Figure 4.2), it's logical to incorporate a simple and fast clustering algorithm, like KMC to study how much clustering of phase data reveals the pathway structure. To overcome the disadvantages of KMC method, I used software, MultiExperiment Viewer (MeV) *v*.4.3, in which, Figure of Merit (FOM) is employed to choose the number of clusters; the module *k*-Means Support (KMS) is used to run KMC algorithms multiple times to obtain the consensus clusters (Saeed, Sharov et al. 2003); these algorithms are detailed in Method: section 4.2.7 of this chapter. Meanwhile, application of phase data instead of the original time course data has largely resolved the heterogeneity issue that exists in the systems biology study.

After 100 KMC runs and 50 iterations per run was conducted on Claytor-wt-2's unwrapped phase data, KMS generated 4 consensus clusters and one 'unassigned' set that includes molecules that don't belong to any of the detected cluster (Figure 4.8). The molecules that belong to these clusters and their color codes are listed in Table 4.4. Each phase spectrum was represented as a gray line in Figure 4.8. Although the average phase spectrum of each cluster is illustrated (pink line) as the MeV default, Pearson correlation was used as distance metric to capture trends/patterns of dynamics irrespective of their overall magnitudes. To evaluate how the KMC output relates to the biological network, these molecules are highlighted in Claytor network model with their assigned color (Figure 4.9).

The Claytor-wt-2 data was obtained through perturbation. The Claytor system was perturbed by increasing the concentration of the main substrate M23 and of the toxic substance M1. When the system is provided with substrate M23, M23 is then modified to M22 (maybe through phosphorylation or hydrolysis) and transported into the cell by P13. The catabolism pathway breaks down M22 to smaller units (M24, M19 and M17) and releases energy. Biosynthesis pathway uses the smaller unit M17 as precursor and energy to construct other essential molecules; M17 goes through a series of chemical transformation from M17, M16, M13, M7, and M6 to finally form an essential molecule M3. This process involves enzymes P10, P9, P8, P7 and P6. When protein receptor P15 senses a lot of M1, it then binds to the toxic substance M1, forming complex P22 which acts as a transcription factor that triggers several genes, including G1, G2, G3, G4, G5, G14, G16, and G20. The redox chain transfers energy-rich electrons to reduce the toxic substrate M1 (akin to an electron transfer chain) (Mendes 2009).

The first piece of information we got in this study is from the phase analysis, even before running the KMC. The phase spectra computed through Fourier transform has detected all the molecules with constant or near constant values. These 11 molecules (Cluster 1, highlighted in pink color) includes: constitutive genes G6 and G19 and their protein products P6 and P19 (Figure 4.9A); two genes G15 and G18 which encode protein receptors P15 and P18 that bind metabolites M1 and M3 to form transcription factors (Figure 4.9A and B); three important enzymes P7, P8 and P9 that catalyze part of biosynthesis pathway leading to the formation of essential compound M3; two large metabolites M22 and M24 from degrading apparent abundant substrate M23 in catabolism pathway (Figure 4.9B).

From KMC results (Table 4.4 and Figure 4.9), we've found all the intermediate compounds in the Biosynthesis pathway, M17, M16, M13, M7 and M6 are grouped in Cluster 3 (highlighted in cyan). Cluster 3 also includes two enzymes P10 and P11 that catalyze reactions part of this process. The rest of the enzymes in this pathway, P6, P7, P8 and P9 are pre-clustered in Cluster 1 (pink). All the metabolites in the redox chain are clustered into two groups: the chain of metabolites before the protein redox pair (P20 and P21) are grouped into Cluster 3 (cyan), which includes the essential compound made through Biosynthesis pathway, M3 and its electron acceptor M4, toxic substance M1 and its less toxic reduced form M2; the metabolites after the protein redox pair are grouped into Cluster 4 (highlighted with green), which includes M9, M10, M11 and M12. One of the protein pair P21 also belongs to Cluster 4. All the enzymes (P1, P2, P3, P4 and P5) that enable this series of redox reactions are grouped into Cluster 2 (highlighted with orange color). These enzymes are most likely 'inducible', i.e. synthesized only when needed.

Metabolite M3 has dual functions: It acts as reducing agent to transfer electrons to M1 and reduce M1 to less toxic M2; M3 also binds protein receptor P18, and their complex is a transcription factor (P23) that represses three genes (G7, G8 and G9) and induces another one (G10). KMC put repressed genes G8 and G9, and P23-induced and P17-repressed gene G10 into Cluster 3 (cyan) along with M3 and P23. Some of the clustering results are not easy to interpret, for instance, the genes that induced by P22 are scattered in three clusters by KMC (Cluster 2, 4 and 5). It could be distorted partitions due to the KMC's drawbacks (inappropriate choices of initial points, local minimum instead of global minimum, etc).

4.3.4 Biplot display on phase data

The Claytor network model was created to mimic a real biochemical pathway as much as possible. It certainly exhibited a typical issue with systems biological data: heterogeneity. The levels of mRNAs, proteins and metabolites have multiple scales that span seven orders of magnitude. Thus, biplot display of the time course data couldn't clearly illustrate all the

molecules. The plot was dominated by molecules with large concentrations. Using phase spectra of time course data's Fourier transform has effectively solved this problem.

Figure 4.10 shows the biplot display of Claytor-wt-2 phase spectra. The biplot display indicates that the two smallest frequencies 0.0083 Hz and 0.0167 Hz have the largest variances among the frequencies. The frequencies spread toward the negative *x*-axis direction in the increasing order (Figure 4.10A). The placement of frequencies on the biplot display shows an interesting oscillation trend in the *y* direction. An individual plot of frequencies clearly shows a wave moves from small frequencies toward large frequencies (Figure 4.10B). As phase diagram's counterpart in the frequency domain, this succession of plotted points is analogous to the system's state over frequency. Compared to the phase space of Claytor-wt-2's time course biplot display (unpublished results), the trajectory of the system on the frequency domain indicates an oscillatory pattern. Similar to phase space, the trajectory approaches the stationary point without crossing each other, since it is known that trajectories in phase space do not cross (Reich and Sel'kov 1981). The seemingly crossing line between 0.1167 Hz to 0.125 Hz is very likely the effect of the trajectory's 2-D projection, and perhaps in 3-D this would not happen.

Instead of chopping up data into distinct clusters, the biplot display reveals the closeness of the molecules in terms of their dynamics. The P22-induced genes, G1, G2, G3, G4, G5, G14, G16, and G20 are clustered into three groups in the previous KMC study; now we see they clearly neighbor each other in Figure 4.10C. They are joined by some of their protein products: P1, P5, P2, P4, P14, P16 and P20. The metabolites in the Biosynthesis pathway closely clustered together too. These includes: M6, M7, M13, M16 and M17. The molecules that cluster in this group also include the Biosynthesis pathway's downstream product M3, M3-associated signaling pathway product, a transcription factor P23, P23-induced G10, and G10 product P10. Four redox pairs in redox chain overlap each other: M1 and M2, M3 and M4, M9 and M10, M11 and M12. But only M9, M10, M11 and M12 are close.

While majority of the molecules are distributed in the vicinity of *x*-axis, several molecules deviate from the horizontal axis and isolate themselves from the others. This includes: G7, P12, P13, P18 and M20. It's not surprising that these molecules are not assigned in the *k*-means clustering analysis (Table 4.4).

Good algorithms reveal underlying information contained in the raw data. In this purpose, both *k*-means clustering analysis and biplot display have enlightened us with information that is consistent with each other. However they are not the first level of abstraction. The extractions of phase spectra from the raw data's Fourier transforms are crucial in this process. It enables us to integrate mRNA, protein and metabolite data regardless of the difference in their magnitudes. *k*-means clustering analysis may not provide us the best groupings of the molecules, nor our molecules the best candidates for non-overlapping clustering. But it sheds some light on the

underlying pathway structure. The further study of biplot display not only complements *k*-means clustering analysis, but presents us more information in plots about the system's trajectory over the course of experiments, the closeness of molecules with each other in terms of their dynamics, and the relationships between frequencies and molecules. This level of abstraction leads us closer to the structure of the network. It likely reveals a part of the pathway. To unveil the whole picture, a carefully designed series of experiments maybe required.

Table 4.4: *k*-means clustering analysis results of Claytor-wt-2 phase spectra. Claytor-wt-2 data set was generated with COPASI and Fourier transformed using program "FFT_Coeff_COPASI". The extracted phase data was unwrapped with program "Unwrap-N" and analyzed with software MeV. An overall of 100 KMC run was conducted; and during each KMC run, the total iterations was 50; the threshold of co-occurrence was set at 80%. Pearson correlation was used as distance metric. FOM recommended 3 clusters for KMC. KMS concluded 4 consensus clusters. Please note the molecules in pink highlighted cluster 1 was excluded from KMC due to their constant or near-constant values in time course, which was detected in phase analysis. In this table, the 4 consensus clusters and 1 unassigned one are color-highlighted respectively using orange, cyan, green, blue and gray.

Cluster1	G6	G15	G18	G19	P6	P7	P8	P9	P19	M22	M24						
Cluster2	G3	G5	G13	G14	G17	P1	P2	P3	P4	P5	P14	P17					
Cluster3	G8	G9	G10	G11	G12	P10	P11	P23	M1	M2	M3	M4	M6	M7	M13	M16	M17
Cluster4	G4	G16	G20	P15	P21	P22	M9	M10	M11	M12							
Cluster5	G1	P16	P20														
Unassigned	G2	G7	P12	P13	P18	M20											

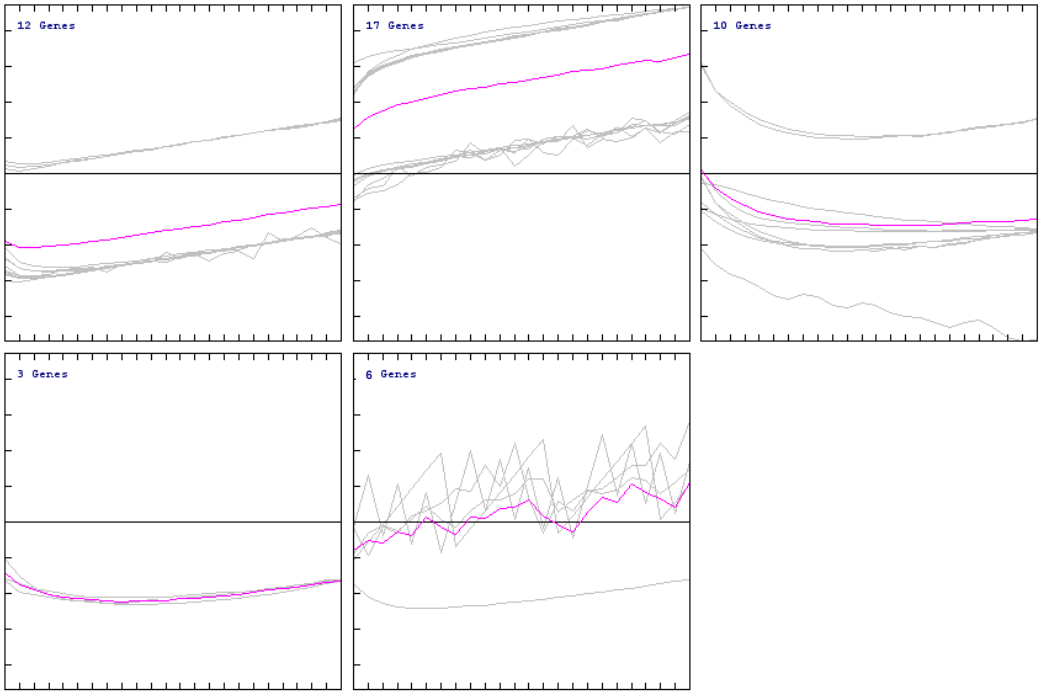
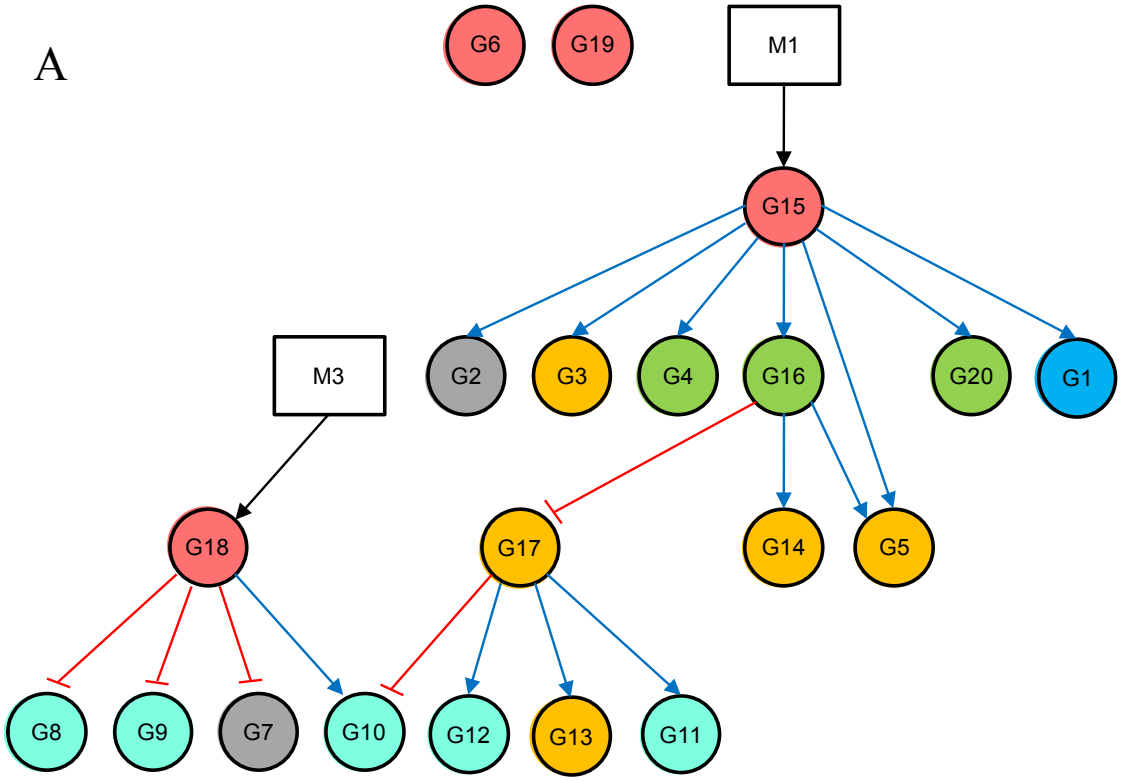


Figure 4.8: *k*-means clustering on Claytor-wt-2 data. Four consensus clusters and one unassigned set are produced. Each gray line represents the phase spectrum of an individual molecule plotted versus frequency; each pink line represents the average phase spectrum for each cluster, although Pearson correlation was used as distance metric. This analysis was performed with the software MeV. Since MeV is data analysis tool for microarray, the number of molecules in each cluster is labeled as "# Genes".


A
G6
G19
M1
G15
M3
G2
G3
G4
G16
G20
G1
G18
G17
G14
G5
G8
G9
G7
G10
G12
G13
G11

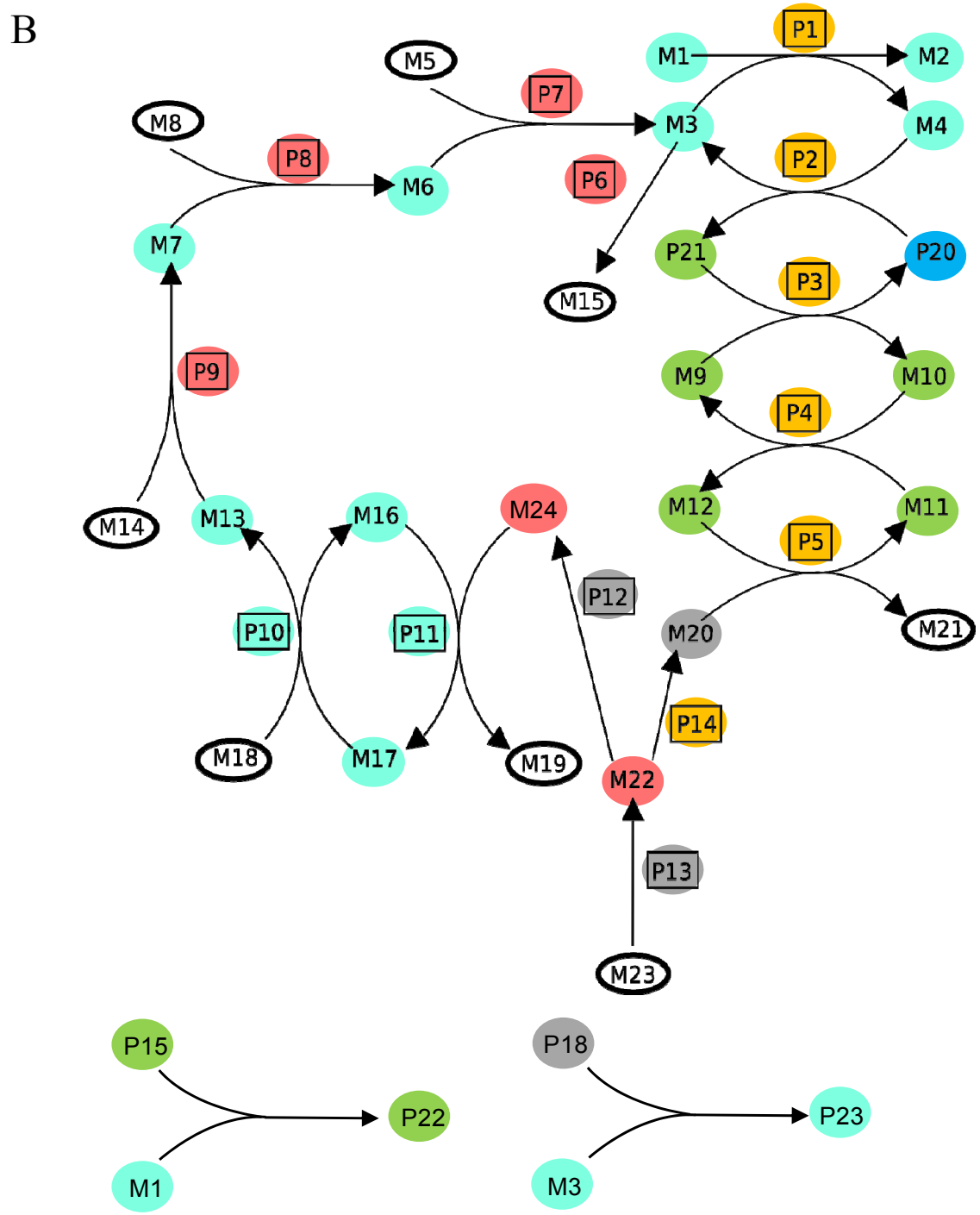
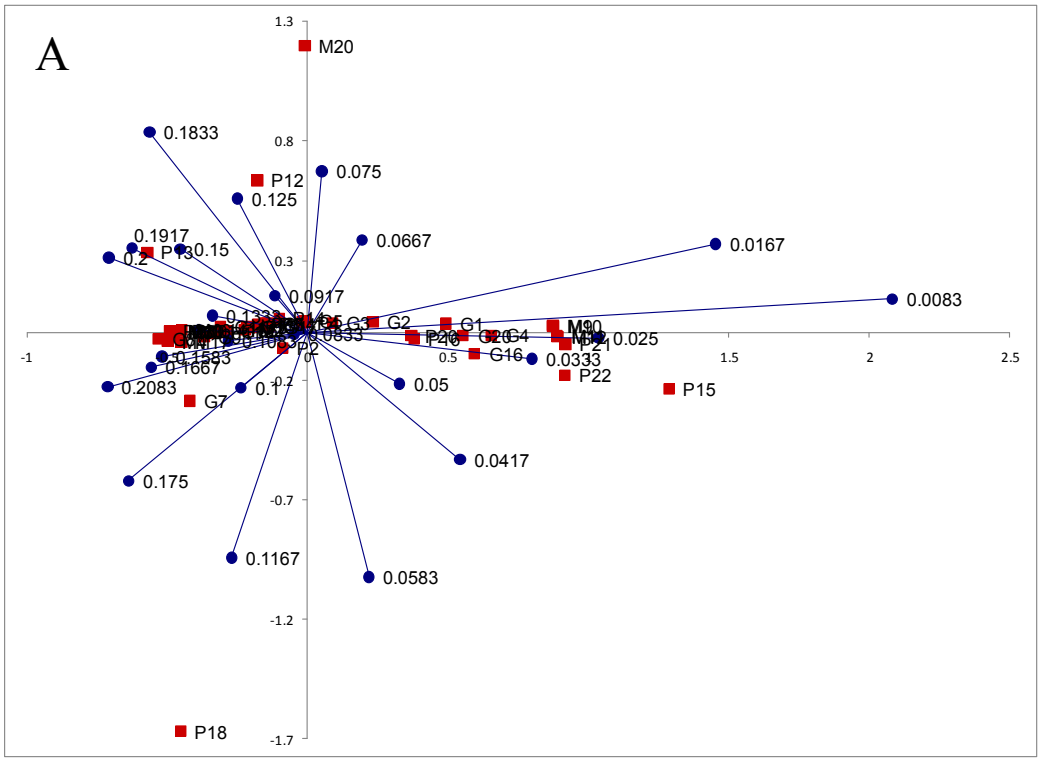
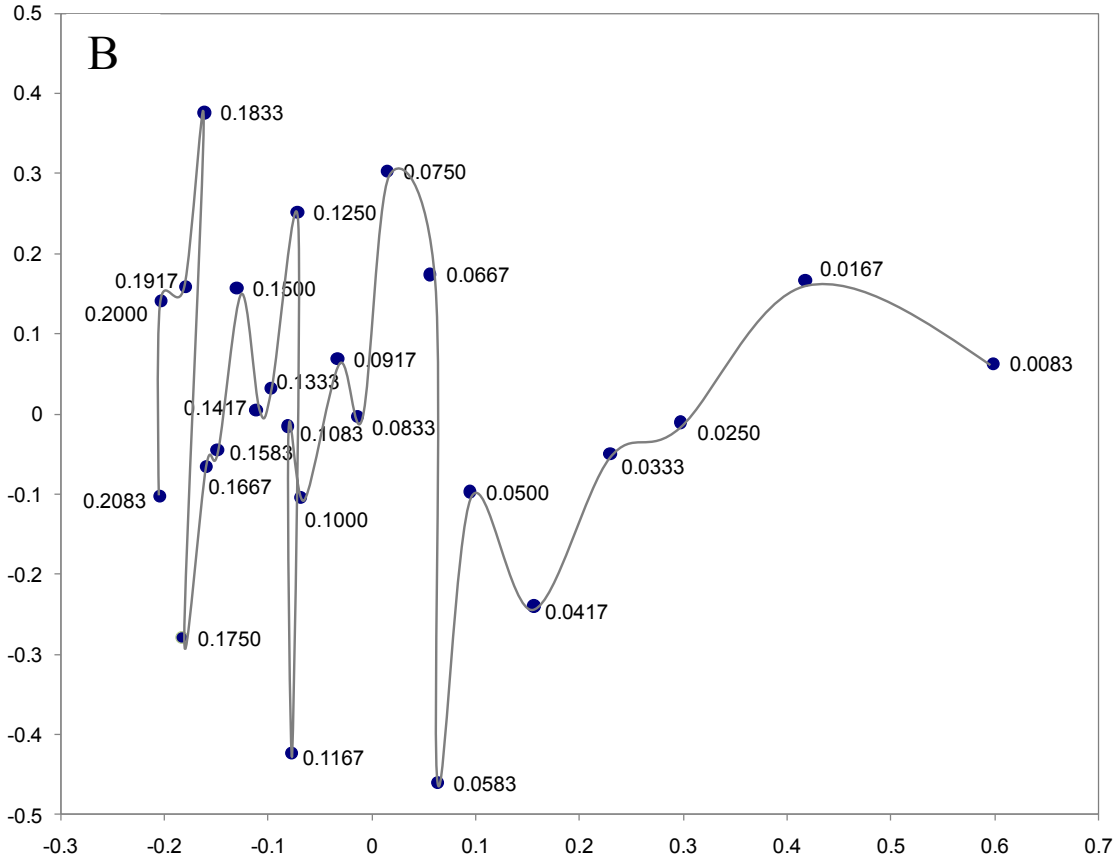


Figure 4.9: Color-highlighted *k*-means clustering on Claytor network model. A: diagram of the gene interaction. B: diagram of metabolism and signaling pathway. The pink, orange, yellow, green, cyan and blue highlighted molecules indicate respectively they belong to cluster 1 to 5 that are described in Table 4.4; the gray colored molecules represent unassigned ones in *k*-means clustering analysis.


A
M20
0.1833
0.075
P12
0.125
0.1917
0.2
P3
0.15
0.0667
0.0167
0.0917
0.0083
G2
G1
0.025
G16
0.0333
P22
P15
0.05
0.0417
0.2083
G7
0.1
0.175
0.1167
0.0583
P18


B
0.1833
0.0750
0.1250
0.0667
0.0167
0.1917
0.2000
0.1500
0.0917
0.0083
0.1333
0.1417
0.1083
0.0833
0.0250
0.1583
0.0333
0.1667
0.0500
0.2083
0.1000
0.0417
0.1750
0.1167
0.0583

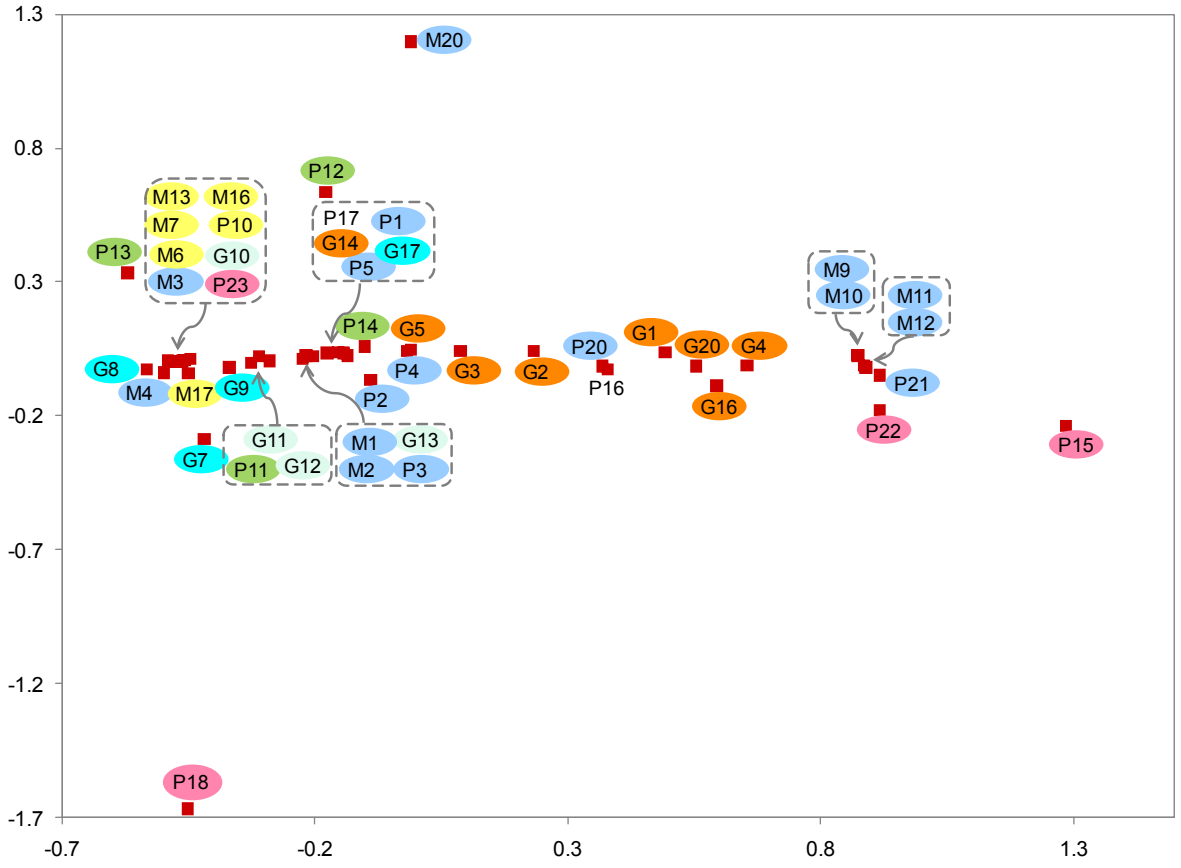


Figure 4.10: Biplot display of Claytor-wt-2 phase spectra. Phase data was extracted and unwrapped on Claytor-wt-2 data's Fourier transforms. Then the phase spectra were analyzed with Singular Value Decomposition (SVD) and displayed in Biplot, during which data was rows and columns centered and symmetrically scaled. A: Biplot display of frequencies and molecules. Frequencies are represented with dark blue filled circles and molecules bright maroon filled squares. Frequency rays are drawn. B: Trajectory of frequencies in Biplot display. Points are joined with gray line in the sequence of all frequencies occurred in the data. C: Color-highlighted Claytor network on molecules in Biplot display. The color code of three metabolism and signaling pathways is according to Figure 4.2; green, yellow, blue and pink highlighted molecules indicate respectively they belong to the following pathways: green represents catabolism which provides energy and reducing power from a substrate; yellow represents biosynthesis, which uses energy to absorb nutrients and construct an essential molecule; blue represents redox chain that reduces the toxic substance M1; pink represents two protein receptors and their formation of transcription factors. Genes that are induced (G1, G2, G3, G4, G5, G14, G16, and G20) are highlighted with orange; genes that are repressed (G7, G8, G9, and G17) are highlighted with cyan; gene G10 that undergoes both induction and repression is highlighted with bi-color, orange and cyan. Genes G11, G12 and G13 are induced by P17; however since G17 is repressed, G11, G12 and G13 are also highlighted with both color.

Chapter 5

Data integration and data fusion in systems biology

5.1 Introduction

Systems biology studies in the postgenomics era have largely been focused on multi-parallel profiling techniques for parallel monitoring of 'omes', for example, transcript, protein and metabolic profiles (Kitano 2002). These approaches have become possible mainly due to the advances in high throughput technologies for characterization of biological samples. This involves, for example, the microarray technology for transcript profiling or chromatography coupled with mass spectrometry for peptide or metabolite profiling (Weckwerth 2003). The purpose is to study organisms as integrated systems of genetic, protein, metabolic, pathway and cellular events in order to achieve a higher level of understanding of the interplay between molecular and cellular components (BylesjoÌˆ, Nilsson et al. 2008).

One classical approach of systems biology study is we treat the system as a black box, the inner structure and behavior of which can be analyzed and modeled by varying an internal or external condition (perturbation), probing it from outside and studying the effect of the variation on the external observables (Kell, Brown et al. 2005). The external observables in our case would be large amounts of transcriptomics, metabolomics and proteomics data. To understand the whole system, one must study the whole data (Goodacre, Vaidyanathan et al. 2004; Kell 2004). This is where "data integration", and more specifically "data fusion" comes into place.

In a broad sense, data integration and data fusion require combining and matching information from different sources and resolving a variety of conflicts to achieve improved accuracies and more specific inferences. These are two facets of data processing: Data integration emphasizes data management or data acquisition, which is the process of extracting, transforming, and transporting data from several source systems to a data warehouse (Lenzerini 2002); data fusion emphasizes data analysis and prediction, which involves processing digital signal, statistical analysis and making inferences (Llinas and Hall 1998).

In terms of data integration, Dr. Pedro Mendes group has developed a software system called "DOME" (http://calvin.vbi.vt.edu/DOME/DOMEMT/index.php) ─ "the database for 'ome's", to store, analyze and integrate functional genomics data. DOME allows analysis using unsupervised methods and visualization using biochemical maps. Data can be downloaded if one wants to perform analysis using other software. While we can easily query omics data of *Medicago* and *Vitis* projects from DOME, it's time to study them as a whole.

To fuse these data, one needs to consider two situations in our data set: The first is the data sets to be fused refer to the same molecules in the same sample, just having been measured by

different instruments. This situation is prominent in metabolomic data as the overlapping data problem. Among those technologies, some groups of compounds can be measured with more than one technique although they can be differentiated better on one platform than another (Smilde, van der Werf et al. 2005). For example, there are 15 amino acids detected in the methyl jasmonate experiment in *Medicago* project and were measured with both CE/MS (Capillary Electrophoresis/Mass Spectrometry) and GC/MS (Gas chromatography/Mass spectrometry). The second situation is the data to be fused refer to different molecules measured by different technologies, although they have been extracted from the same sample (which is the reason why we are seeking their relationships). In our projects, the associated mRNA, protein and metabolite identities and concentrations were obtained when cells were under biotic or abiotic stress. This problem has been increasingly recognized as a research topic of huge importance in system biology. One of the key issues here is when multiple technologies were used, the variables usually are not commensurate, that is, not similar in scale of measurement nor in their levels of measurement noise. Resolving this problem is one of the objectives in this research.

5.2 Methods

5.2.1 Using ratio scale and median as responses

One of the situations and also a major challenge we are facing in omics data fusion is the data refer to different molecules measured by different technologies, although extracted from the same biological sample. One noticeable feature of these data is that they are not commensurate. A simple but effective step to solve this problem is to use ratios of treatment versus control to remove the heterogeneity.

Ratio variables have a long history of use in multiple research areas including sociology, geology, political sciences (Lyons 1977), economics (Watson 1990), human kinetics (Liu 2003), and others. Most measurement in the physical sciences and engineering is done on ratio scales. Mass, length, time, plane angle, energy and electric charge are examples of physical measures that are ratio scales. The scale type takes its name from the fact that measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind (Michell 1997).

In cDNA microarray analysis, the red/green ratio of the two fluorescent signals at each spot is commonly used to infer the ratio of the mRNA concentrations in the two RNA samples, and further to infer the expression change. The normalized ratios are then used for various graphical and numerical study to select differentially expressed genes or to find meaningful

clusters of genes (Cui, Kerr et al. 2003; Smyth, Yang et al. 2003; Kim, Dougherty et al. 2000; Brazma and Vilo 2000; Sharov, Kwong et al. 2004; Chen, Kamat et al. 2002; Chen, Dougherty et al. 1997).

The statistical validity of using ratio variables has been investigated (Liu 2003; Bollen and Ward 1979) and debated (Lyons 1977; Uslaner 1977) by scientists. All statistical measures can be used for a variable measured at the ratio level, as all necessary mathematical operations are defined. The central tendency of a variable measured at the ratio level can be represented by, in addition to its mode, its median, or its arithmetic mean, also its geometric mean or harmonic mean. In addition to the measures of statistical dispersion defined for interval variables, such as range and standard deviation, for ratio variables one can also define measures that require a ratio, such as coefficient of variation (Stevens 1946). The ratio variables in data analysis are mainly used as measures of theoretical concepts, as a means to control an extraneous factor, or as a correction for heteroscedasticity (Bollen and Ward 1979).

In this research, if not stated otherwise, ratios of treatment versus control are used to infer the changes of molecule accumulation (or depletion). Using ratios as response has largely brought the scales of molecules produced from different profiling technologies to a similar level. Should further adjustment is needed for biplot visualization purpose, a quick scaling is performed.

Since we have up to six replicates for each measure in *M. truncatula* study, medians will be used for ratio computations. Choosing the median instead of the mean as measure of central tendency / (statistical average) is based on the advantages of the median over the mean for experimental data. The median is not influenced by extreme measurements in experimental data and the median is invariant with respect to most ordinary (monotone) transformations (e.g. the median of log X is the logarithm of the median of X) (Rubin and Smith 1958). The relative advantages of the median over the mean increase as the kurtosis increases (Rubin and Smith 1958). Kurtosis measures the degree of tail heaviness and peakedness of a distribution (Ruppert 1987). And higher kurtosis means more of the variance is due to infrequent extreme deviations, i.e. heavy tail (Chissom 1970). We found that using median instead of mean of the replicates can effectively reduce the impact of variance in the replicates on the analysis (research result from Dr. Bharat Mehrotra).

5.2.2 Phase spectra obtained with Fast Fourier transform in Mathematica

Chapter 4 has discussed the theory of Fourier transform and importance of phase information in extracting useful information from sampled signals.

A program, "FFT_Coeff_ COPASI" written in Mathematica was used to conduct FFT (Fast Fourier transform) on time course data. Phase spectrum with frequencies up to Niquist frequency was extracted and unwrapped to remove data discontinuity. Details are as described in Chapter 4 methods section 4.2.5 and 4.2.6.

5.2.3 *k*-Means Clustering analysis with MeV 4.3

Phase spectrum data produced with Mathematica are loaded into the software ─ MultiExperiment Viewer (MeV) *v*.4.3 (Saeed, Sharov et al. 2003). FOM (Figure of Merit) is used to choose *k*, the number of clusters before the analysis. The module *k*-Means Support (KMS) is used to run the *k*-Means Clustering (KMC) algorithms multiple times using the same parameters in each run. The consensus clusters were set at 80% of total *k* means runs. Cosine Correlation is chosen as the distance metric to measure similarity between different phase spectra. The *k*-means clustering is set to run 100 times. During each KMC run, the maximum number of iterations is 50.

5.2.4 Data filtering prior to biplot display analysis

For biplot display of integrated 'omes'─transcripts, metabolites and proteins in response to salinity stress in *Vitis vinifera* study, the selected 45 transcripts were MAS 5.0 normalized Affymetrix data with at least 10-fold induction in response to salinity stress. Vvi 1440 was removed due to its extremely outlying value at 24 h (610.7): 225 times higher than median. All the transcript values were scaled by dividing each value by 5 before biplot analysis. The responses of metabolites were obtained with GC-MS analyses. Total of 28 metabolites with ratios between salinity and control samples at least 1.5 were selected for biplot. M217 was removed before biplot analysis due to its outlying value at 1 h (13.25): 9 times higher than median. The metabolic data were not scaled. The total of 15 proteins with significant response to salinity was used in biplot, the cut-off value of ratio between salinity and control sample was 5. All protein values were scaled by dividing each value by 60. The scaling transformed the integrated data so that the largest value in the final data matrix is about 10.

For biplot display of integrated transcripts and metabolites following yeast elicitation in *Medicago truncatula* study, the gene expression values were cDNA microarray data processed by print tip LOWESS without background subtraction. The selected 24 gene transcripts have significant responses following yeast elicitation with *p* value less than 0.001. All the transcript values were multiplied by 5 to match up the scales of metabolites. The responses of metabolites were obtained with LC-MS analyses. The metabolite data were corrected by internal references ─ flavonoid, and retrieved from the online database DOME

(http://calvin.vbi.vt.edu/DOME/DOMEMT/index.php). The selected 39 identified metabolites have significant responses after the elicitation with p value less than 0.01. Hispidol 4'-O-glucoside didn't participate biplot analysis due to its large outlying value and known performance from the previous study in Chapter 3 (Figure 3.13).

In these data preprocessing procedures, readers may notice the heavy filtering. It is not the most appropriate approach since some molecules with subtle change may play an essential role in response to perturbation. But this approach is a reasonable candidate for data visualization.

Rows and columns centering are performed before SVD. Times and metabolites are equally scaled for biplot. These procedures were carried out using the software written by Lipkovich and Smith ─ Biplot and Singular Value Decomposition Macros for Excel© (Lipkovich and Smith 2002).

5.3 Results and discussion

5.3.1 Comparison of amino acids profiled in GC-MS and CE-MS

Data overlapping, *i.e.* same compounds are measured with different instruments is prevalent in metabolic profiling. In *M. truncatula* study, CE-MS method profiled 20 standard amino acids, of which 15 were also profiled with GC-MS analyses. How to treat these redundant measurements? To answer this question, we need to first investigate these two sets of data.

Of 20 amino acids profiled with CE-MS, five of them were not detected in GC-MS analyses; these are methionine, glutamine, tyrosin, tryptophan and histidine. Pearson's correlations were calculated on the corresponding measurements of the 15 amino acids (Table 5.1). The results indicate that 7 of them (40%) are significantly correlated ($p < 0.01$). Beta-alanine has the highest correlation between these two profiles with r of 0.971. Proline and alanine also show a high similarity between these two profiling techniques. Aspartate and arginine have the least similar measurements with negative correlations. It will be noted that 80% (12) amino acids have a higher correlation with other amino acids than with themselves. Plotting these amino acids' responses measured with two technologies during a 48 hr period at 21 time points (Figure 5.1) reveals that the CE-MS data are relatively more compressed than that of GC-MS. Recall the violin plots of metabolic profiles from different technologies: CE-MS shows a violin shape, while GC-MS has a "Hershey's kisses" shape (Figure 2.3). The heavy-tailed/skewed distribution of GC-MS data may be explained by the wide spread of data shown

Table 5.1: Correlations for amino acids profiled in CE-MS and GE-MS analyses in *M.truncatula* project. The ratios between the median response of the methyl jasmonate (MeJa)-elicited sample and control sample were used for correlation analysis. Correlations greater than 0.549 or less than -0.549 are significant ($p < 0.01$) (highlighted with gray color). Of 20 amino acids profiled in CE-MS, 5 of them were not detected with GC-MS, 7 (40%) showed significant correlations between the results from the two detection technologies.

Amino acid	$r_{GC,CE}$
Aspartate	-0.250
Alanine	0.730
Leucine	0.500
Threonine	0.586
Beta-alanine	0.971
Asparagine	0.668
Valine	0.645
Glycine	0.357
Glutamate	0.179
Arginine	-0.167
Serine	0.425
Phenylalanine	0.678
Lysine	0.417
Isoleucine	0.456
Proline	0.761
Methionine	
Glutamine	
Tyrosine	
Tryptophan	
Histidine	

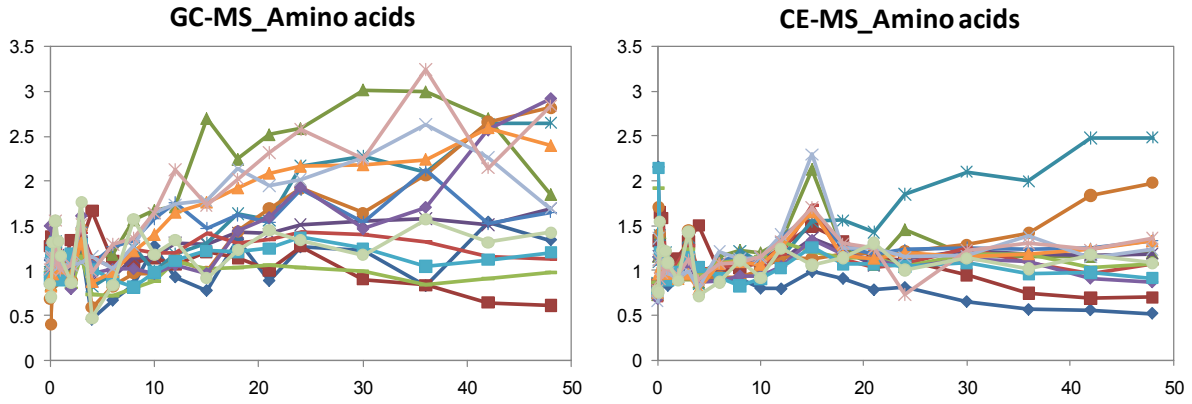


Figure 5.1: Dynamics of 15 amino acids profiled in GC-MS and CE-MS. The data are ratios between the median responses of MeJa elicited sample and control sample in *Medicago truncatula* study.

in Figure 5.1. The differences in these two technologies' data distribution are likely attributed to their different data processing and normalization procedures.

From the comparison between GC-MS and CE-MS profiled the data, I conclude that it's not a good idea to combine the data from these two sources due to the high proportion of low correlations. We need to choose one methodology that can differentiate better on the compounds of interest than the other platform, or one that will facilitate the following data analysis.

5.3.2 Biplot display of integrated 'omes' in response to salinity stress in *Vitis vinifera* study.

Grapes are grown in semi-arid environments, where drought and salinity are common problems. Gene transcript, protein and metabolite profiling under the effect of abiotic stress can be used to define genes and metabolic pathways in *Vitis vinifera* cv. Cabernet Sauvignon and to facilitate improvements to be made in both production efficiency and wine quality under environmentally adverse growing conditions (Cramer, Cushman et al. 2002).

Gene transcript, protein and metabolite's responses to salinity stress were filtered, scaled and integrated to a single data matrix. Individual omics data filtering prior to integration can resolve one issue present in systems biology, *i.e.* how much change is considerable? The answer depends on the type of data and molecules. The significance threshold of fold changes is not necessarily the same in mRNA level, protein level or metabolite level. The filtered molecules from each omics data set represent the significant change at each individual level. When we pool them together, we can use methods, such as scaling, to transform the data to same scale without losing information.

The details of the data filtering and scaling are described in Method 5.2.4. The data scaling brought the ranges of each omic data within 10, which makes it possible to visualize three types of data at the same time on biplot display. Biplot of these omics data after singular value decomposition shows clear separations between 1, 4 and 24 hrs (Figure 5.2). The vector of 8 hour lays closely to 1 hr due to the intrinsic features of the data where 8 hr is correlated to neither 4 hrs ($r_{(8h,4h)} = -0.096$) nor to 24 hrs ($r_{(8h,24h)} = 0.06$) and the contribution from protein P.1702. Three types of 'omes' are shown on this biplot. Four distinct clusters of molecules that are associated with different times are highlighted with colored ellipse (Figure 5.3). Gray ellipses enclose two groups of transcripts respectively dominate 4 and 24 hrs and largely contribute to the overall pattern of the plot; orange ellipse represents metabolite cluster and protein cluster are associated with 1 hr, while metabolite group is close to 4 hr and protein

group is between 1 and 8 hrs. The variables 4 hr and 24 hr have larger variations than that of 1 hr judging from the length of associated rays.

Salinity stress has affected the expression of a large number of genes (Cramer, Ergül et al. 2007). The most highly induced transcripts are involved in stress response, transcription, and protein synthesis and modification (Table 5.2). One of the earliest responses to salinity stress was an increase in the transcript abundance of 9-cis-epoxycarotenoid dioxygenase (NCED1), which is an enzyme in abscisic acid (ABA) biosynthesis. ABA, a plant hormone, is involved in responses to environmental stresses such as drought and high salinity, and is required for stress tolerance (Iuchi, Kobayashi et al. 2001; Qin and Zeevaart 2002). Two significant gene transcripts with increased abundance that respectively dominate 4 and 24 hrs on biplot are Vvi.7869 and Vvi.3077. Vvi.7868 is the transcript for a hypothetical small heat shock protein (Moser, Segala et al. 2005). Vvi.3077 is similar to a gene that encodes a RALF like protein, RALFL33. Rapid alkalinization factor (RALF) has been shown to be a peptide hormone that may be involved in the rapid alkalinization of the extracellular solution (Matsubayashi and Sakagami 2006; Olsen, Mundy et al. 2002). The proton gradient across the plasma membrane is important for many physiological processes in plants including ion uptake, solute transport, and cell wall growth. The transient changes in extracellular or intracellular proton concentrations and the accompanying plasma membrane depolarization or hyperpolarization are implicated in the rapid responses of cells to environmental stress (Haruta and Constabel 2003).

Other notable transcripts include Vvi.8050, which is associated with LEA (late embryogenesis abundant) protein that helps prevent the formation of damaging protein aggregates during water stress (Goyal, Walton et al. 2005), and Vvi.7397 for thaumatin-like protein serving as an defense agent against pathogenic fungi attack (Monteiro, Barakat et al. 2003; Tattersall, van Heeswijck et al. 1997).

Similar to the responses to MeJa elicitation in *M. truncatula*, the biplot shows elevated levels of amino acids, notably leucine, isoleucine, valine and beta-alanine and their accumulation at 24 hrs. Branched chain amino acids (valine, leucine, and isoleucine) may be substrates for an array of phytoanticipins and involved in plant defense (Iriti, Rossoni et al. 2005; Mayer, Cherry et al. 1990). The accumulation of beta-alanine suggests salinity stress has an effect on the pathway of coenzyme A. The biplot also shows that the accumulation of pipecolic acid at 8 hrs precedes the increase of the aforementioned amino acids. Studies indicate that pipecolic acid is involved in the conversion of D- to L-amino acids via a cyclization intermediate in plants and microorganisms (Friedman 1999). Both 1L-myo-Inositol 1-phophate and its precursor 1-alpha-D-galactosyl-myo-inositol exhibited elevated level under salinity stress. Biplot indicates they spiked at both early times (1hr and/or 8 hrs) and 24 hrs. Myo-inositol and its phosphates play an important role as the building block of phosphoinositide (PI);

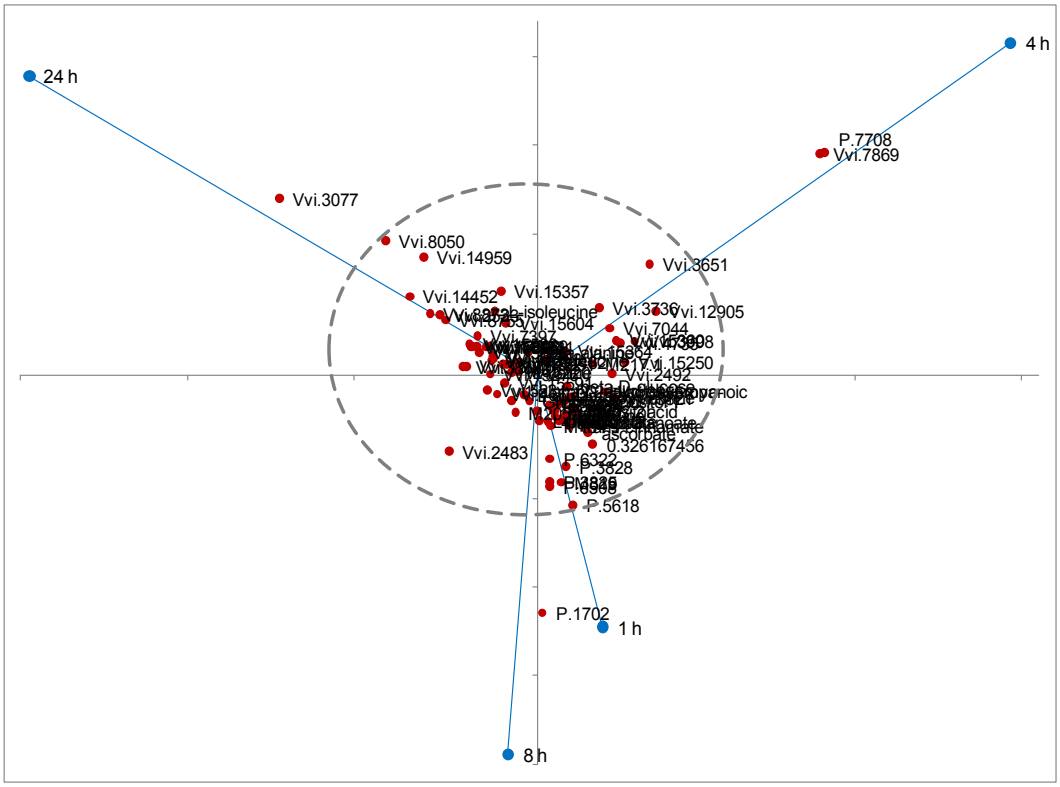


Figure 5.2: Biplot display of integrated 'omes'─transcripts, metabolites and proteins in response to salinity stress in *Vitis vinifera* study. The selected 45 transcripts were MAS 5.0 normalized Affymetrix data with at least 10-fold induction in response to salinity stress. Vvi 1440 was removed due to its extremely outlying value at 24 h (610.7): 225 times higher than median. All the transcript values were scaled by dividing each value by 5 before Biplot analysis. The responses of metabolites were obtained with GC-MS analyses. Total of 28 metabolites with ratios between salinity and control samples at least 1.5 were selected for Biplot. M217 was removed before biplot analysis due to its outlying value at 1 h (13.25): 9 times higher than median. The total of 15 proteins with significant response to salinity was used in Biplot, the cut-off value of ratio between salinity and control sample was 5. All protein values were scaled by dividing each value by 60. Rows and columns centering are performed before SVD. Times and metabolites are equally scaled. Blue points represent times 1, 4, 8 and 24 hours and red points represent 88 significantly changed transcripts, metabolites and proteins.

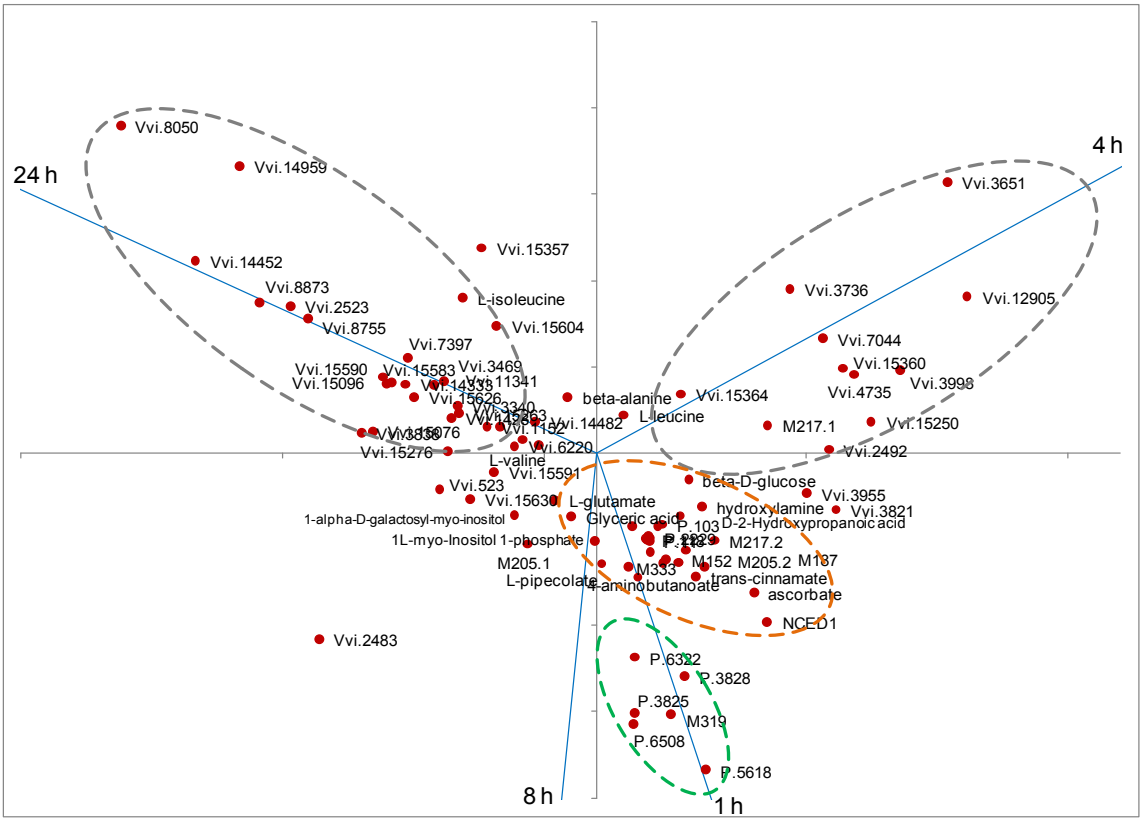


Figure 5.3: Gene transcripts, proteins and metabolites in response to salinity stress in *Vitis vinifera* study revealed by Biplot display—an enlarged view of the portion circled in Figure 5.2 A. Details are as in the legend to Figure 5.2A. Transcripts with greatest responses at 4 and 24 hr are highlighted in gray ellipse. Metabolites with greatest changes at 1 hr are highlighted in orange ellipse, proteins with greatest responses at 1 hr in green ellipse.

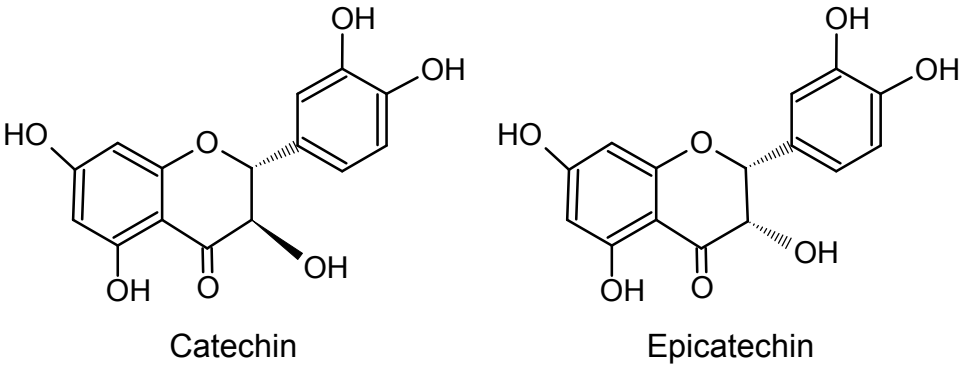


Figure 5.4: Structures of flavonoids, catechin and epicatechin that have fragment ions at *m/z* 205. Drawn in ChemSketch 8.0.

Table 5.2: Transcripts revealed by Biplot display (Figure 5.2 and 5.3), with the greatest change in transcript abundance in response to salinity stress on hours, 1, 4 and 24. *Vvi 1440 was removed before Biplot analysis due to its extremely outlying value at 24 h (610.7): 225 times higher than median.

Probe set ID	UniGene ID	Description
1 h		
1608022_at	Vvi.13945	9-cis-epoxycarotenoid dioxygenase 1 (NCED1)
4 h		
1616538_at	Vvi.7869	weakly similar to NP_175759.1 unknown protein [Arabidopsis thaliana]
1610490_at	Vvi.3651	Transcribed locus
1613510_at	Vvi.3736	Transcribed locus
1612385_at	Vvi.12905	Weakly similar to NP_200780.1 HSP18.2 (HEAT SHOCK PROTEIN 18.2) [Arabidopsis thaliana]
1609528_at	Vvi.3998	Transcribed locus
1615309_at	Vvi.7044	Moderately similar to NP_175759.1 unknown protein [Arabidopsis thaliana]
1611192_at	Vvi.4735	weakly similar to NP_196763.1 unknown protein [Arabidopsis thaliana]
1622204_at	Vvi.15250	Similar to Os03g0670700, glycine-rich RNA-binding, abscisic acid-inducible protein.
24 h		
1611272_at	Vvi.1440*	Transcribed locus, weakly similar to NP_596601.1 hypothetical protein SPBPJ758.01 [Schizosaccharomyces pombe 972h-]
1607519_at	Vvi.3077	Weakly similar to NP_567476.1 RALFL33 (RALF-LIKE 33) [Arabidopsis thaliana]
1608715_at	Vvi.8050	Weakly similar to XP_475821.1 putative LEA protein [Oryza sativa (japonica cultivar-group)]
1613616_at	Vvi.14452	Transcribed locus
1619363_at	Vvi.14959	Weakly similar to NP_176053.1 transferase, transferring glycosyl groups / transferase, transferring hexosyl groups [Arabidopsis thaliana]
1620438_at	Vvi.15357	Weakly similar to NP_193436.2 ATEXLB1 (ARABIDOPSIS THALIANA EXPANSIN-LIKE B1) [Arabidopsis thaliana]
1609355_at	Vvi.8873	Strongly similar to NP_113875.1 ubiquitin A-52 residue ribosomal protein fusion product 1 [rattus norvegicus]
1608782_at	Vvi.2523	Transcribed locus
1614289_at	Vvi.8755	Weakly similar to np_174094.1 zat10; nucleic acid binding / transcription factor/ zinc ion binding [arabidopsis thaliana]
1616695_s_at	Vvi.7397	Weakly similar to np_502360.1 thaumatin family member (thn-1) [caenorhabditis elegans]
1614207_at	Vvi.3469	DFR mRNA for dihydroflavonol reductase
1616152_at	Vvi.15583	Transcribed locus
1620715_at	Vvi.15276	Transcribed locus
1611080_at	Vvi.11341	Weakly similar to NP_179378.1 heat shock protein binding / unfolded protein binding [arabidopsis thaliana]
1621860_at	Vvi.15604	Weakly similar to xp_474035.1 osjnbb0034i13.2 [oryza sativa (japonica cultivar-group)]

The PI pathway is involved in sensing environmental stimuli and is one of the most conserved signaling pathways (Boss, Davis et al. 2006; Torabinejad, Donahue et al. 2009). Myo-inositol is also a precursor for cell-wall complex carbohydrates.

Integrated biplot display revealed the accumulation of a group of compounds in response to salinity stress that contribute to the "sensory" characteristics—aroma, color and flavor of wine and grape juice (Figure 5.3). These include: a nitrogen compound, aminobutyrate; major organic acids in grapes, ascorbate and its metabolism product, glyceric acid (Loewus 1999); and common precursor of the bioactive compounds flavonoids and stilbene, trans-cinnamate (Singleton, Timberlake et al. 1978). Two unidentified compounds with fragment ion at *m/z* 205 that eluted at different times showed elevated level in salinity treated sample. They are likely the fragment ions of flavonoids, catechin and epicatechin (structure see Figure 5.4) (Stöggl, Huck et al. 2004; Sun, Liang et al. 2007). Catechin and epicatechin are polyphenolic antioxidant compounds that are widely distributed in plant-derived foods including red wine, green tea, chocolate, and many fruits. Epidemiological and animal studies have correlated polyphenol consumption to reduced rates of heart disease as well as certain forms of cancer (German and Walzem 2000; Dixon 2001). Another abundant compound with fragment ions at *m/z* 217 may be hemiterpene glycosides (Baltenweck-Guyot, Trendel et al. 1997).

The biplot shows an interesting protein P.7869 that is highly correlated to transcript Vvi.7708 accumulating at 4 hrs. Protein P.1702 is notably accumulated at 8 hrs. Three proteins P.11, P.118 and P2229 almost overlay each other near 1 hr. Due to the prevalence of protein modifications at both transcriptional level and posttranslational level, proteomes may be two to three orders of magnitude more complex (>1000,000 molecular species of proteins) than the encoding genomes would predict (Walsh, Garneau-Tsodikova et al. 2005; Mann and Jensen 2003). It undoubtedly complicates the quantitative relationship between protein and their encoding genes. No algorithms at this level can draw more inferences than giving us clues for further investigation.

The phase space biplot of *V. vinifera* study doesn't show a typical trajectory which may be caused by the fallout of 8 hr vector. It suggests that when the sampling intervals are far apart, the time variables are less correlated or even uncorrelated. The system trajectory is more truncated to the extent that each time is like an independent event. This, together with system and biological variations can attribute to the possible failing of phase space biplot in real experiment.

5.3.3 Biplot display of integrated transcripts and metabolites in response to yeast elicitation in *Medicago truncatula* study.

Previously biplot analysis has been used on metabolites, mainly flavonoids, response to yeast elicitation in *M. truncatula*. The biplot revealed the most highly induced flavonoids and their induction level in a single diagram (Figure 3.13). By integrating transcript data with metabolite data, we can study the interplay between gene expression and metabolite accumulation responding to yeast elicitation in hope of illuminating the underlying network.

The biplot display of the integrated data has kept the overall pattern of the flavonoids' response to elicitation (Figure 5.5), where several primary flavonoids are associated with mid time. In the previous metabolomic biplot, early times from 0 min to 1 hr are so correlated that their vectors superimpose over each other. It indicates that no sizable changes observed during this period. In the integrated biplot, the rays for these early times expand across the third quadrant of the Cartesian plane. It's mainly due to the dynamics of the added gene transcripts, the majority of which exhibit an early response to the elicitation. Two noticeable trends are observed on the integrated biplot (Figure 5.5): mid times, from 4-12 hrs is still the territory of several primary flavnoids accumulated in response to yeast elicitation (highlighted in orange ellipse); Early-mid times, from 1-3 hrs are occupied by several significantly changed gene transcripts (highlighted in gray ellipse). Investigation of these transcripts revealed two chalcone synthase (CHS; EC 2.3.1.74) transcripts (Table 5.4): TC6536 and TC6559. Exposure of cells to yeast elicitor resulted in up to 7- and 2-fold induction of TC6536 and TC6559, respectively, at 2 hr and 1 hr post-elicitation. Two probe sets representing TC6536 had significant accumulation showing similar expression pattern in response to yeast elicitation. Chalcone synthase is pivotal for the biosynthesis of flavonoid antimicrobial phytoalexins and anthocyanin pigments in plants. It catalyzes the initial step of the branch of the phenylpropanoid pathway that leads to flavonoids. Chalcone synthase produces isoliquiritigenin and naringenin chalcone by condensing one *p*-coumaroyl- and three malonyl-coenzyme A thioesters into a polyketide reaction intermediate that cyclizes (Ferrer, Jez et al. 1999; Deavours and Dixon 2005). The biplot shows that the increase of Chalcone synthase precedes the accumulation of isoliquiritigenin and other important flavonoids, this is a good agreement with a hypothesis that these genes need to be induced before the accumulation of the secondary metabolites is possible.

Another gene transcript with 2-fold induction responding to yeast elicitation is pectinesterase (pectin methylesterase; EC 3.1.1.11) transcript─TC101143. Pectinesterase catalyses the de-esterification of pectin into pectate and methanol. Pectin is one of the main components of the plant cell wall. Pectinesterase produced by the plant may play an important role in cell wall metabolism during fruit ripening and during the first phase of host-pathogen interactions, leading to tissue maceration and the release of oligomers which can trigger a chain of events that activate plant's defense systems (ELAD 1997; Mendgen and Deising 1993).

Since proteins are functionally the most important biological molecules in a cell, it is essential to integrate proteome data in the study. Unfortunately the high percentage of missing values in the protein profiles hampered the integration of protein data. In the yeast elicitation experiment of *M. truncatula* project, 37% protein data are missing. Regrettably we had to leave out the proteomic data, and proceed with integration of gene transcript and metabolite data alone.

5.3.4 Using phase spectrum to integrate "omes" in response to yeast elicitation in *Medicago truncatula* study.

In Chapter 4, phase spectrum extracted from Fast Fourier transformed (FFT) *in silico* data combined with *k*-means clustering has been applied to Claytor data and obtained meaningful results. Here we will put phase spectrum algorithm to the test by implementing this method with real experimental data.

The data for phase spectrum analysis are the same as studied in previous section (Ch 5.3.3). Except that here the data were not scaled. A ratio of the median response of the elicited sample to control sample was computed at each time point and used as input signal. Phase data were computed from Fast Fourier transform of the time course data. The length of the frequency (column variable) is half of the length of the original time series. *k*-means clustering (KMC) was implemented to classify phase data to groups to facilitate our understanding of the phase spectrum. Using cosine correlation as distance metric, three consensus groups and one unassigned set were produced with the software MeV. 4.3 (Table 5.3).

KMC reveals that among the three classified clusters, Cluster 1 is mainly composed of gene transcripts, Cluster 3 is mostly constituted by metabolites, and Cluster 2 is a mix of both mRNA and metabolites. The graph representation of each cluster (Figure 5.6) in frequency domain illustrates that Cluster 1 shows a downward sloping/descending trend line, Cluster 2 exhibits an upward sloping/ascending trend line, and Cluster 3 is near horizontal trend line. Plotting of these clusters in time domain uncovered the dynamics of each cluster. Cluster 1 represents early response molecules that peaked before 4 hr, which includes most gene transcripts. Inspection of these transcripts finds that these gene transcripts are also revealed by integrated biplot and highlighted in gray ellipse in Figure 5.5. Cluster 2 mainly represents molecules with a sustained response through 48 hrs which includes daidzin, hispidol_glucoside_malonate, and irisolidone isomer that were depicted and described in Ch 3.4.4 of chapter 3. Cluster 3 represents mid time response molecules that spiked from 4 hrs to 12 hrs, which covers the primary flavonoids and chalcone including isoliquiritigenin, alfalone,

naringenin, formononetin, genistin and afrormosin. These flavonoids are also illustrated by integrated biplot and highlighted with an orange ellipse (Figure 5.5).

The consistent revelation of integrated biplot and phase spectrum analysis prompted the comparison of these two methods by showing KMC clusters on biplot (Figure 5.7). Using the SVD results produced in Figure 5.5, gene transcripts and metabolites are selectively shown on biplot based on their KMC affiliation. Therefore three biplots were created representing respectively those corresponding molecules in three KMC clusters. Cluster 1 is represented by red dots in Figure 5.7A, Cluster 2 is represented by green dots in Figure 5.7B, and Cluster 3 is represented by orange dots in Figure 5.7C.

These two results strongly contribute to deciphering molecules that accumulated at early-mid and mid times, from 45 min to 12 hours, which include the most highly induced gene transcripts and primary flavonoids. Both methods distinguished early-mid time response gene transcripts from mid time response metabolites. However they failed to agree on the interpretation of very early events (0 min to 30 min) and late events (15 hr to 48 hr). Due to the closeness of early time vectors and late time vectors, biplot couldn't differentiate molecules that responded early and late. While phase spectrum analysis classified the big cluster of molecules located in 3rd quadrant of biplot into three clusters, which suggests that phase analysis may surpass biplot in this area.

5.3.5 Using phase spectrum to integrate "omes" in response to salinity stress in *Vitis vinifera* study.

The salinity stress data from *Vitis vinifera* study was the first experimental data I tried with phase spectra integration given that all three 'omes' are available for this data set. I quickly realized that the disadvantage of few sampling points was exacerbated by the FFT, when the Niquist frequency shortened the data sampling points to half. The transformed data on the frequency domain have only two frequency variables. The follow-up analysis on phase spectra, both biplot display and *k*-means clustering based on Pearson's correlation divided the data into two groups, which doesn't meet the goal of data integration. This work was suspended and the conclusion was drawn that the data with too few sampling points wasn't suitable for phase analysis. However a later re-examination of this data revealed good clustering results when I chose Euclidean distance as the distance metric.

The data for phase spectrum analysis are the same as studied in the previous section (5.3.2) except that here the data were not scaled. Furthermore, two molecules, transcript Vvi.1440 and metabolite M217 were excluded from the biplot display due to their outlier values are now included in this analysis. A ratio of the median response of the elicited sample to control

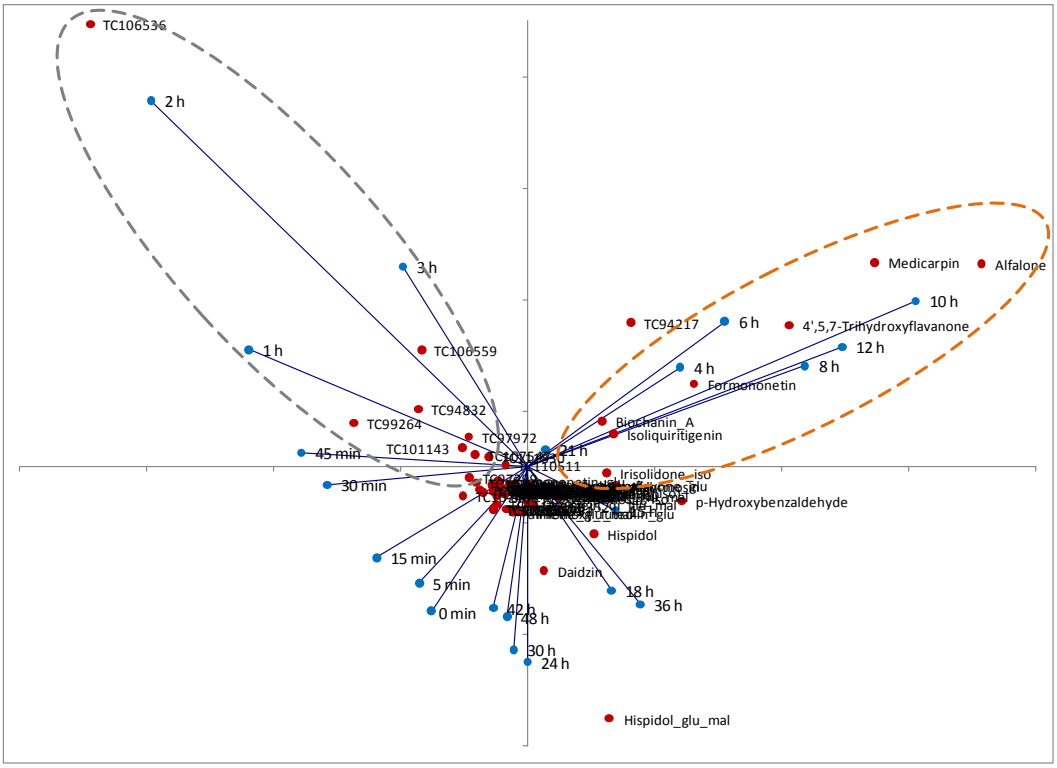


Figure 5.5: Biplot display of integrated transcripts and metabolites following yeast elicitation in *Medicago truncatula* study. The gene expression values were cDNA microarray data processed by print tip Lowess without background subtraction. The selected 24 gene transcripts have significant responses following yeast elicitation with *p* value less than 0.001. The responses of metabolites were obtained with LC-MS analyses. The metabolite data were corrected by internal references ─ flavonoid, and retrieved through the database for 'ome's ─ DOME (http://calvin.vbi.vt.edu/DOME/DOMEMT/index.php). The selected 39 identified metabolites have significant responses after the elicitation with *p* value less than 0.01. Hispidol 4'-O-glucoside didn't participate biplot analysis due to its large outlying value and known performance from the previous study in Chapter 3 (Figure 3.13). Rows and columns centering are performed before SVD. Times and metabolites are equally scaled. Ratio of median between elicitation and control were used for SVD. Blue dots represent time points from 0 minute to 48 hours and red dots represent gene transcripts and metabolites. Gene transcripts with greatest responses at early times, 1-3 hr are highlighted in gray ellipse. Major metabolites with greatest changes at middle time, 4-12 hr are highlighted in orange ellipse.

Table 5.3: Table view of *k*-means clustering on phase spectrum of the Fast Fourier transformed integrated 'omics' data described as in legend of Figure 5.4. Three consensus clusters and one unassigned set are produced. This analysis was performed with the software MeV 4.3 using cosine correlation as distance metric.

Cluster 1	Cluster 2	Cluster 3	Unassigned
TC101906	TC107649	TC95174	TC97299
TC101143	TC107547	TC105751	TC110511
TC102168	TC108735	3',5-dimethoxy_luteolin_glu	Irilone_glu_mal
TC104077	TC109352	4',5,7-Trihydroxyflavanone	Liquiritigenin_glu
TC106536	TC112363	4',7-Dihydroxyflavone_glu	Ononin
TC106559	TC94217	Isoliquiritigenin	Formononetin_glu_mal
TC107488	TC99964	Afrormosin	2'OH-formononetin glu
TC110331	3',5-dimethoxy luteolin_glu_mal	Afrormosin_glu	2'OH- formononetin _glu_mal
TC111377	Afrormosin_glu_mal	Afrormosin_glu_mal (isomer)	Medicarpin_glu
TC111830	Afrormosin isomer	Alfalone	Vestitol_glu_mal
TC94832	Biochanin_A_glu_mal	Genistein_di_glu_mal	Biochanin_A-diglu
TC97972	Daidzin	Genistin	
TC99264	Genistein_glu_mal_iso	p-Hydroxybenzaldehyde	
Biochanin_A	Hispidol_glu_mal	Formononetin	
Biochanin_A-diglu_mal	Irisolidone		
Hispidol	Irisolidone_glu_mal		
Irisolidone_glu	Irisolidone_iso		
Medicarpin_glu_mal	Isoflav-3-ene_glu_mal		
	Medicarpin		

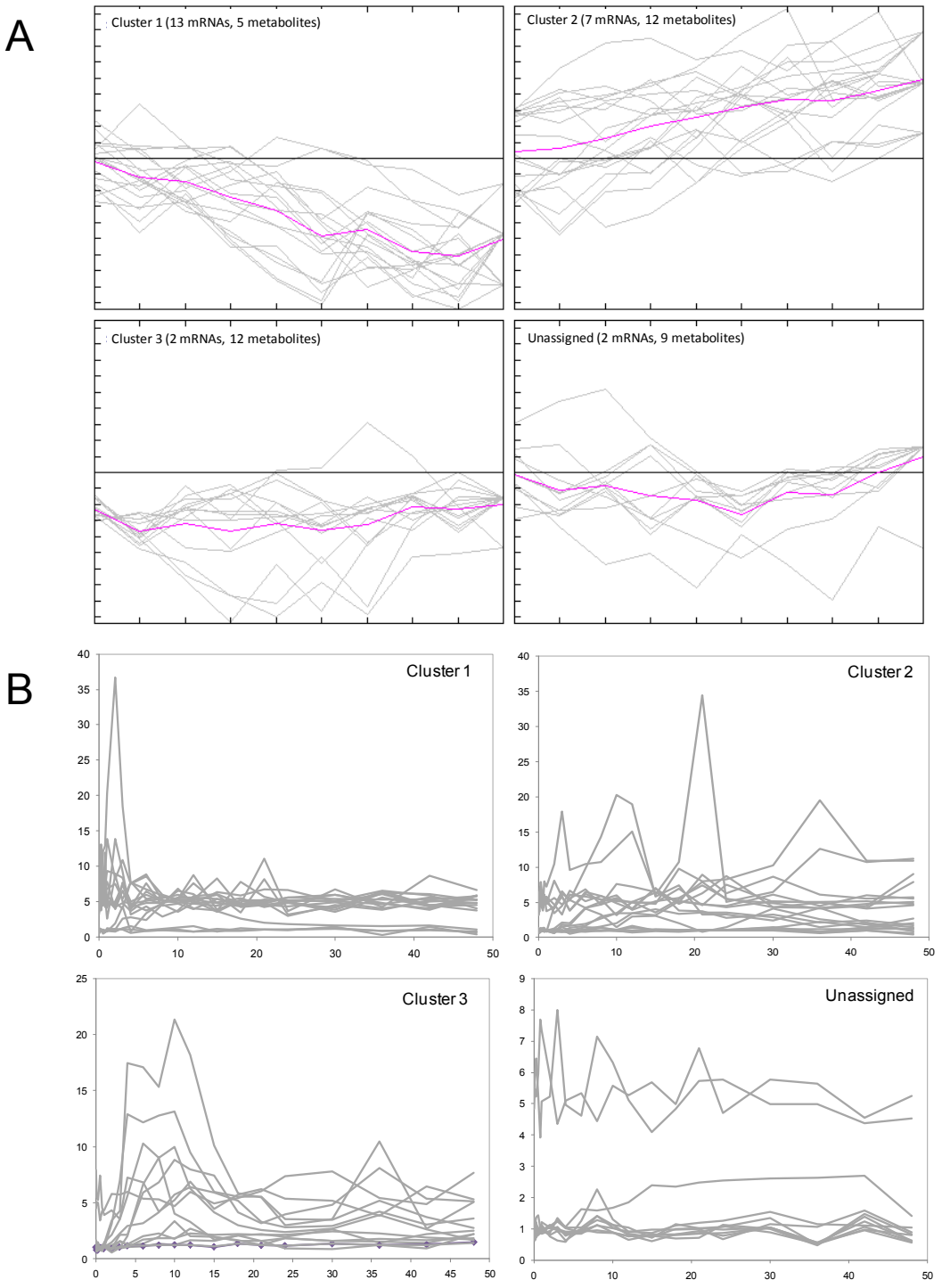
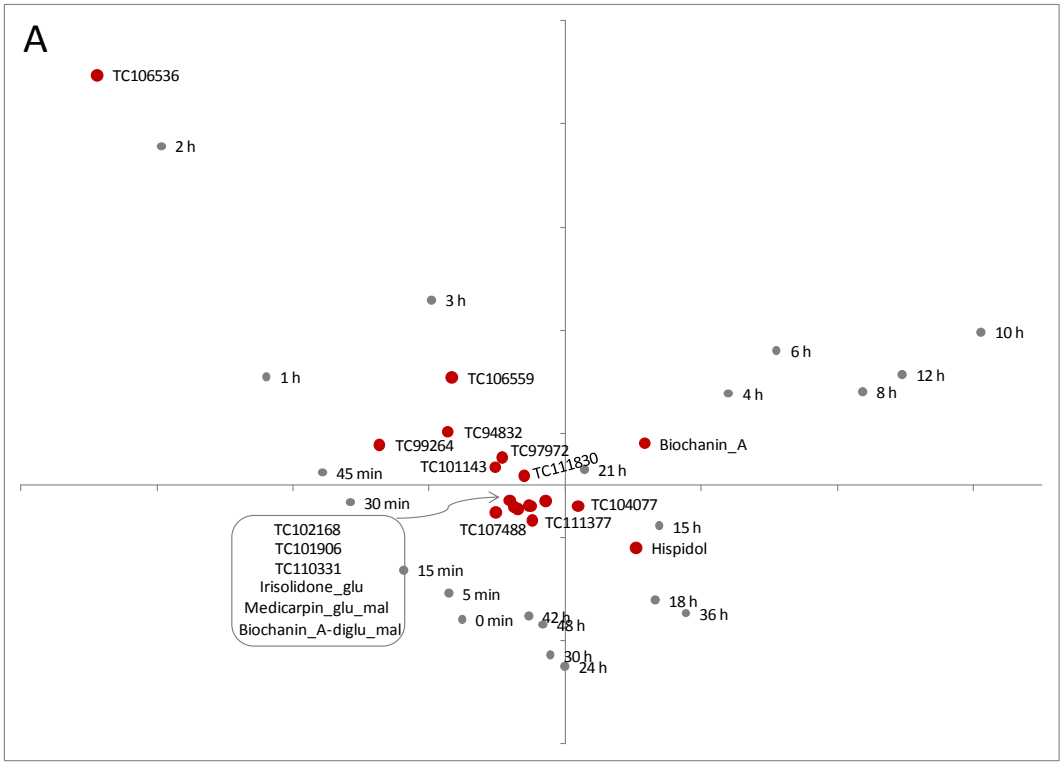
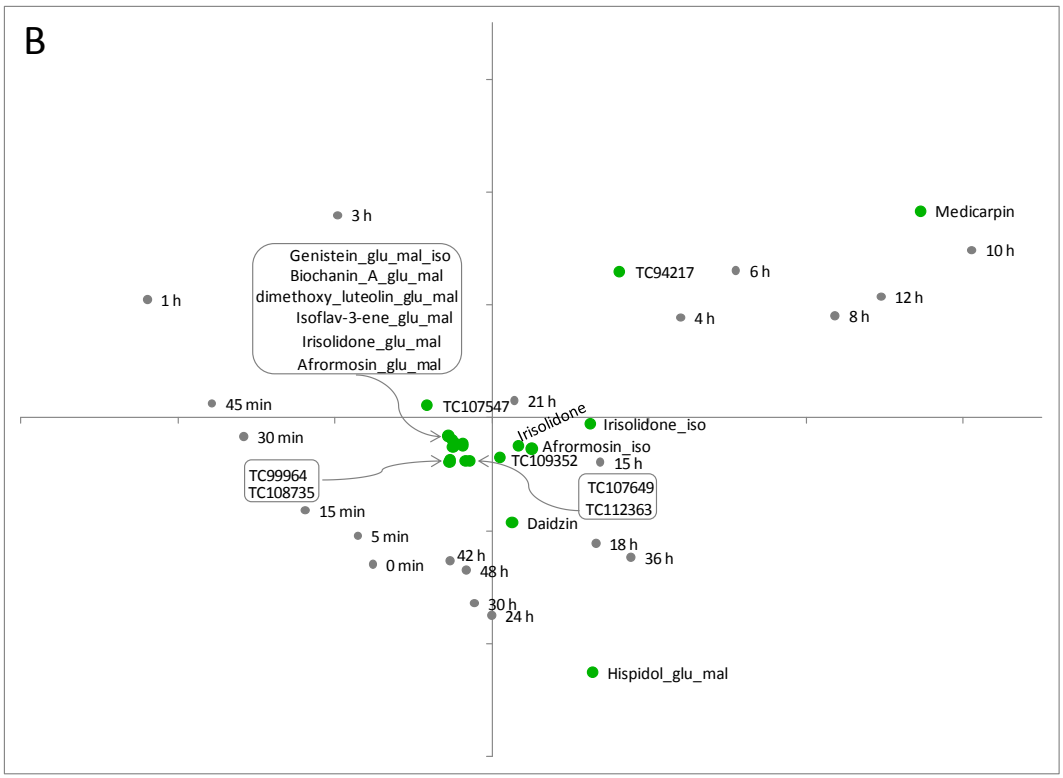


Figure 5.6: *k*-means clustering on phase spectrum of the Fast Fourier transformed integrated 'omics' data described as in legend of Figure 5.4. Three consensus clusters and one unassigned set are produced. A: Graph representations of each cluster in frequency domain. Each gray line represents the unwrapped phase spectrum of an individual molecule plotted versus frequency; each pink line represents the average phase spectrum for each cluster, although cosine correlation was used as distance metric. This analysis was performed with the software MeV 4.3. B: Responses in time domain for each cluster. The *y*-axis represents ratios of the median response of the yeast-elicited sample to control sample for the molecule. The *x*-axis represents 21 sampling points, from 0-48 hrs.

Table 5.4: *Medicago truncatula* gene transcripts with the greatest responses to yeast elicitation classified with *k* means clustering on phase spectra of the integrated 'omics' data. TCs marked with an asterisk were represented by two probe sets and showing similar expression patterns.

Tentative consensus (TC) sequence numbers	Description
Cluster 1	
TC 101906	similar to UP|Q9SJE0 (Q9SJE0) T27G7.21, partial (58%)
TC101143*	similar to UP|Q9FF77 (Q9FF77) Pectinesterase, partial (67%)
TC102168	similar to UP|Q7X6L6 (Q7X6L6) OJ000126_13.4 protein, partial (31%)
TC104077	similar to UP|Q9LMU1 (Q9LMU1) F2H15.10, partial (32%)
TC106536*	homologue to UP|CHS8_MEDSA (P30076) Chalcone synthase 8 (Naringenin-chalcone synthase 8) , complete
TC106559	homologue to PDB|1BI5_A|5542119|1BI5_A Chain A, Chalcone Synthase From Alfalfa. {Medicago sativa} , partial (35%)
TC107488*	similar to GB|AAL32012.1|16930693|AF436830 AT3g07810/F17A17_15 {Arabidopsis thaliana} , partial (39%)
TC110331	similar to UP|Q940U2 (Q940U2) AT5g05220/K18I23_2, partial (25%)
TC111377	Unknown
TC111830	similar to UP|Q9LI39 (Q9LI39) ESTs C23582(S11122), partial (16%)
TC94832	similar to UP|O64851 (O64851) Expressed protein (At2g26190/T1D16.17), partial (66%)
TC97972	Unknown
TC99264	homologue to UP|IF38_MEDTR (Q9XHM1) Eukaryotic translation initiation factor 3 subunit 8 (eIF3 p110) (eIF3c), partial (41%)
Cluster 2	
TC107649*	similar to UP|Q5Z9M1 (Q5Z9M1) Methionyl-tRNA synthetase, partial (61%)
TC107547	similar to UP|O65673 (O65673) Nonclathrin coat protein gamma-like protein, partial (51%)
TC108735	similar to UP|Q75VJ8 (Q75VJ8) Ubiquitin activating enzyme 2, partial (18%)
TC109352	weakly similar to PIR|B96764|B96764 protein integral membrane protein F25P22.12 [imported] - Arabidopsis thaliana {Arabidopsis thaliana} , partial
TC112363	similar to GB|CAA41713.1|19896|NTOEE2AG photosystem II 23 kDa polypeptide {Nicotiana tabacum} , partial (81%)
TC94217	similar to UP|DRR3_PEA (P14710) Disease resistance response protein Pi49 (PR10), complete
TC99964	similar to UP|O23243 (O23243) Beta-galactosidase like protein , partial (19%)
Cluster 3	
TC95174	similar to UP|Q6EN42 (Q6EN42) Bet v I allergen-like, partial (86%)
TC105751	weakly similar to UP|Q9LIM0 (Q9LIM0) Emb|CAB09999.1, partial (32%)
Unassigned	
TC110511	weakly similar to GB|AAR29072.1|39636757|AY426262 blight resistance protein RGA4 {Solanum bulbocastanum} , partial (9%)
TC97299	weakly similar to UP|Q6H5U7 (Q6H5U7) Receptor ser/thr protein kinase-like, partial (45%)


A
TC106536
2 h
3 h
10 h
6 h
1 h
TC106559
12 h
4 h
8 h
TC94832
TC99264
TC97972
Biochanin_A
TC101143
TC111830
45 min
21 h
30 min
TC104077
TC107488
TC111377
15 h
TC102168
TC101906
TC110331
Irisolidone_glu
Medicarpin_glu_mal
Biochanin_A-diglu_mal
Hispidol
15 min
5 min
18 h
0 min
42 h
36 h
48 h
30 h
24 h


B
3 h
Medicarpin
10 h
Genistein_glu_mal_iso
Biochanin_A_glu_mal
dimethoxy_luteolin_glu_mal
Isoflav-3-ene_glu_mal
Irisolidone_glu_mal
Afrormosin_glu_mal
TC94217
6 h
1 h
12 h
4 h
8 h
45 min
TC107547
21 h
Irisolidone
Irisolidone_iso
30 min
Afrormosin_iso
TC109352
15 h
TC99964
TC108735
TC107649
TC112363
15 min
Daidzin
5 min
18 h
0 min
42 h
36 h
48 h
30 h
24 h
Hispidol_glu_mal

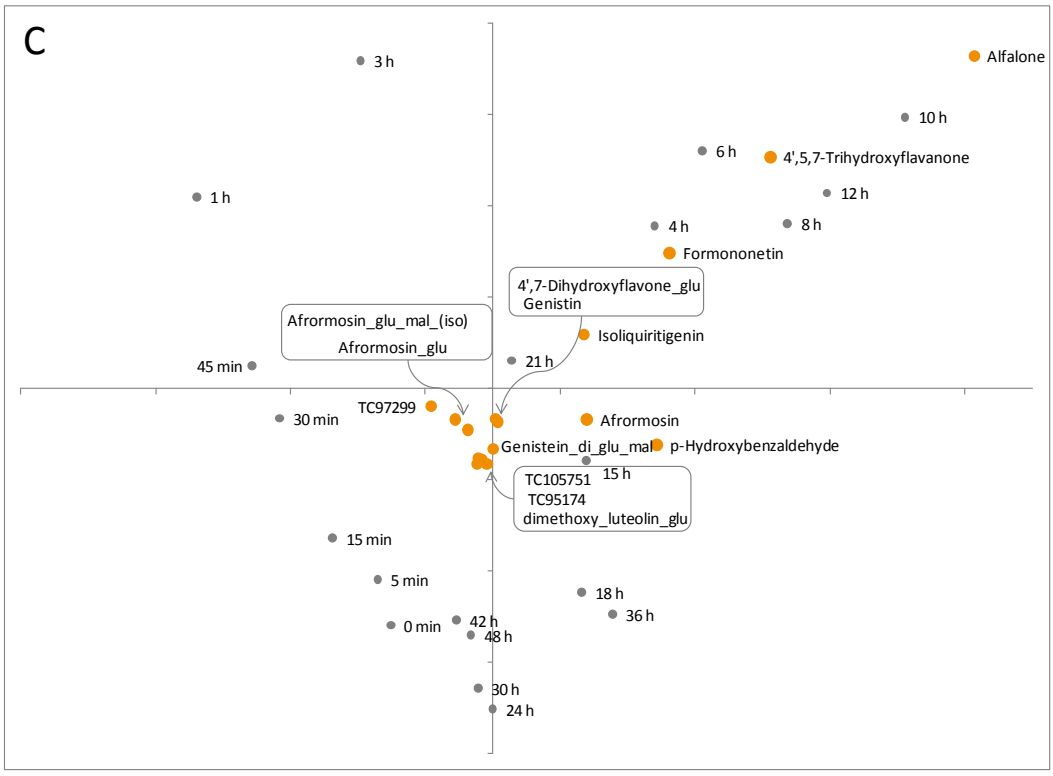


Figure 5.7: Color-highlighted *k*-means clustering on Biplot display. Three clusters produced by *k*-means clustering on phase spectrum of the integrated 'omics' data are highlighted respectively on Biplot produced in Figure 5.4. A: Cluster 1, which includes 13 gene transcripts and 5 metabolites are represented by red dots. B: Cluster 2, which includes 7 mRNA and 12 metabolites are represented by green dots. C: Cluster 3, which includes 2 mRNA and 12 metabolites are represented by orange dots.

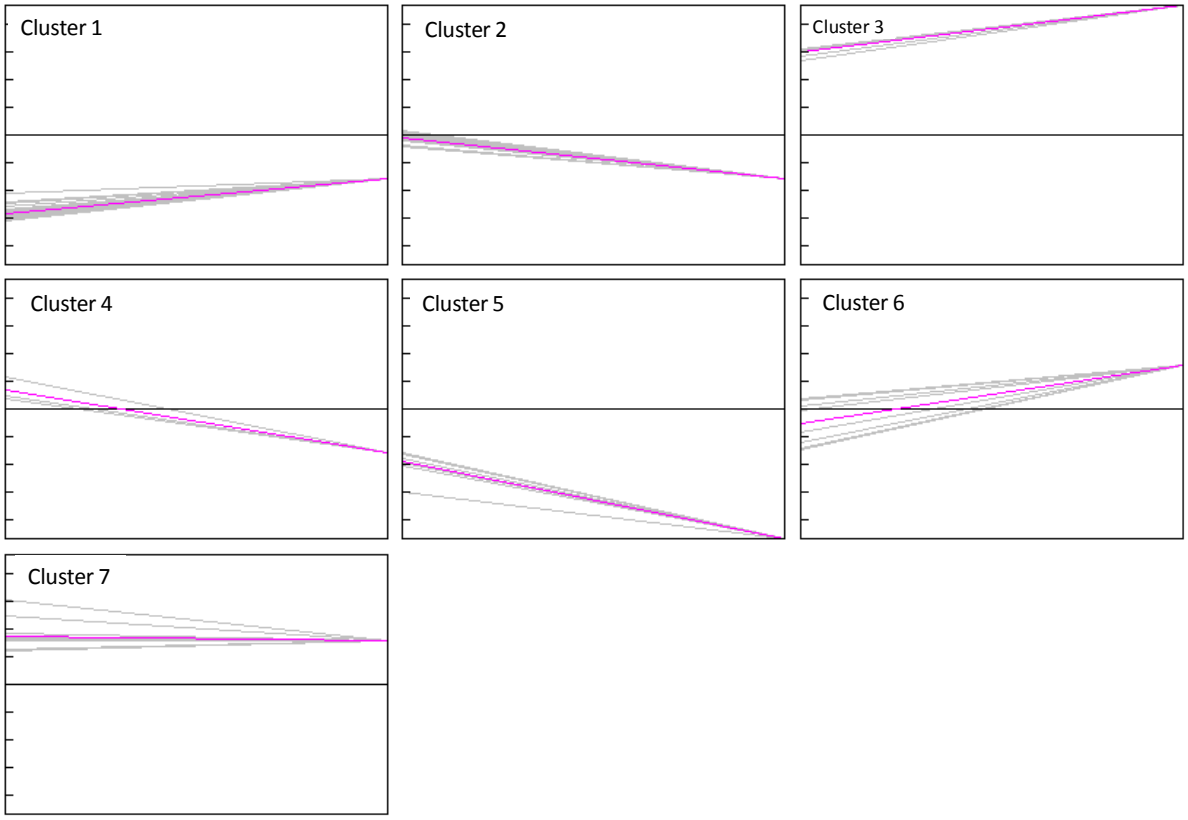


Figure 5.8: *k*-means clustering on phase spectrum of the Fast Fourier transformed integrated 'omics' data described as in legend of Figure 5.3. Seven consensus clusters are produced. Graph representations of each cluster in frequency domain. Each gray line represents the unwrapped phase spectrum of an individual molecule plotted versus frequency; each pink line represents the average phase spectrum for each cluster, although cosine correlation was used as distance metric. This analysis was performed with the software MeV 4.3.

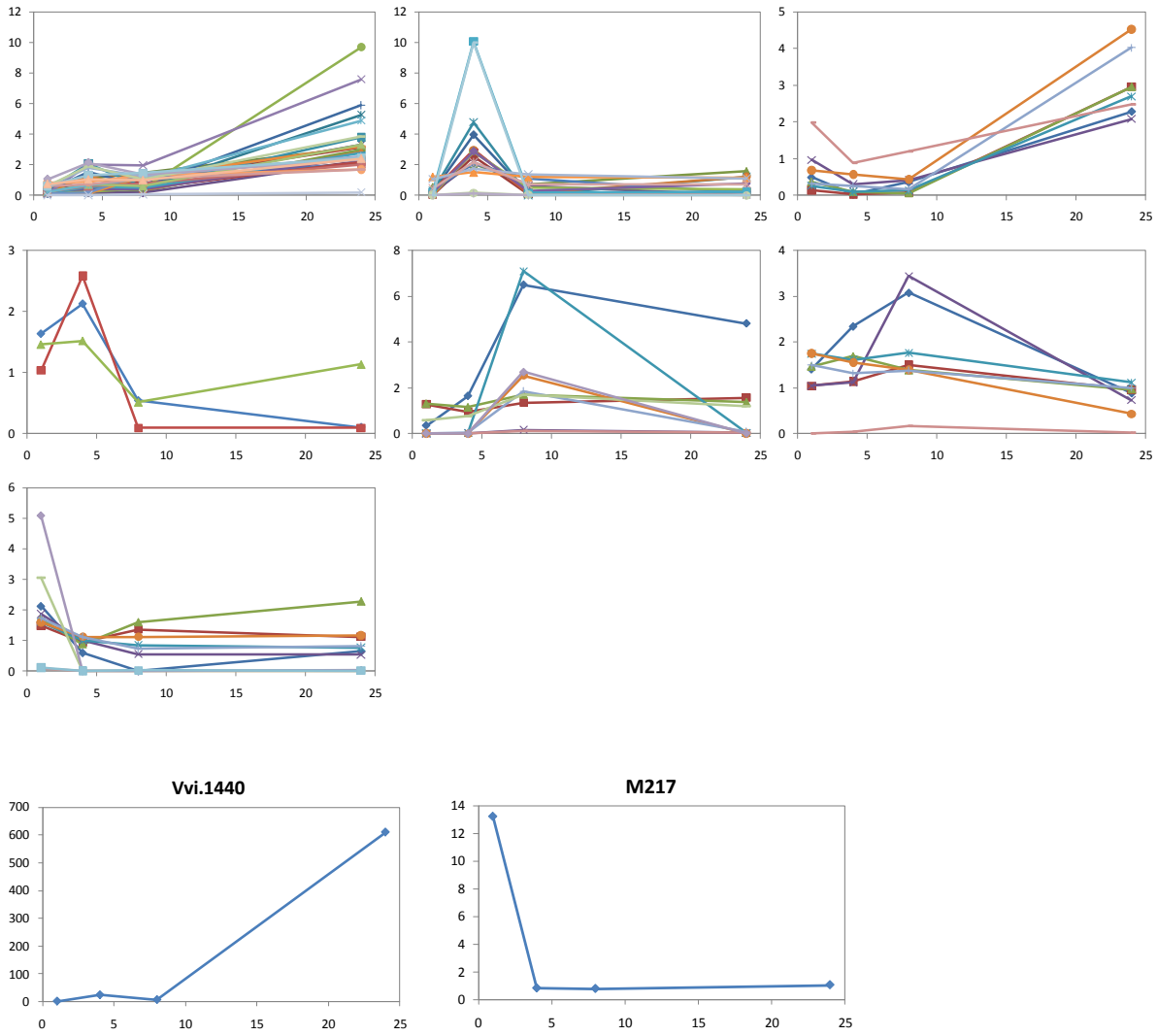
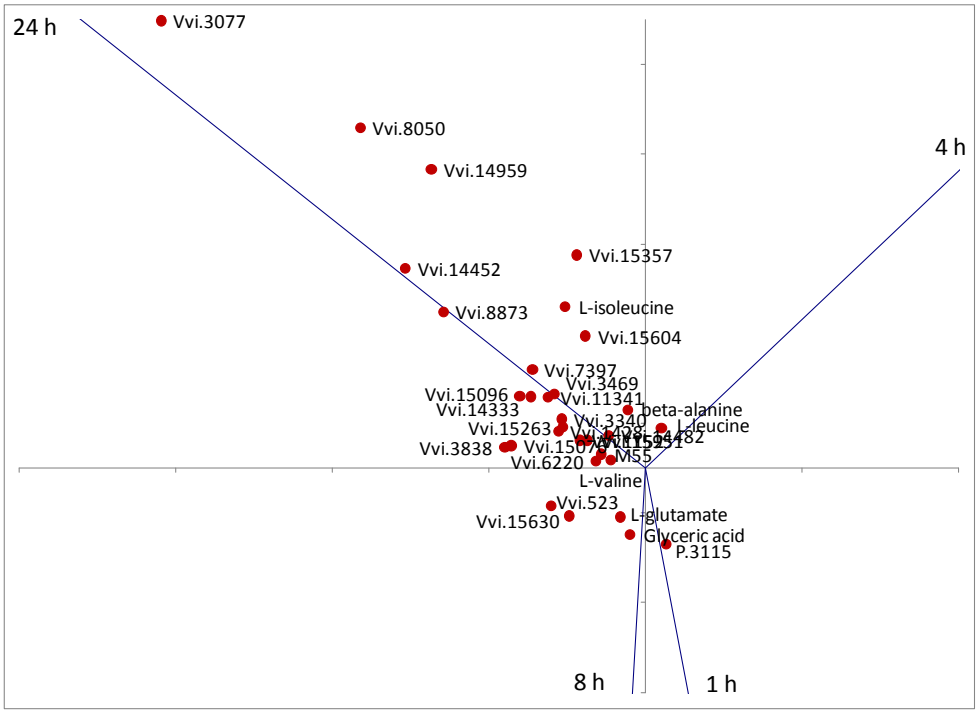
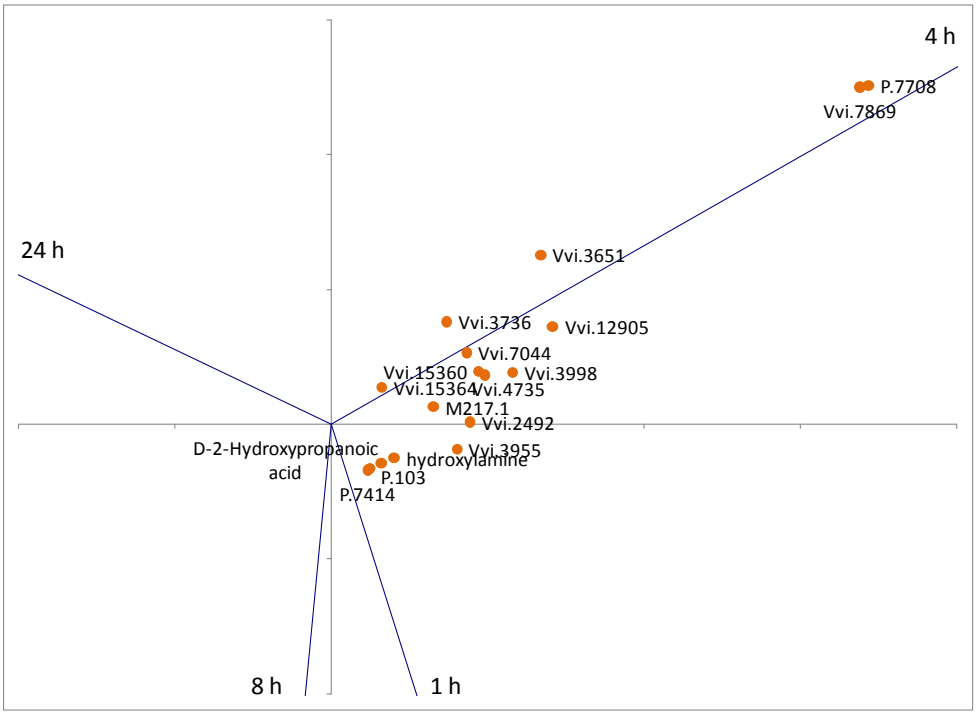


Figure 5.9: *k*-means clustering on phase spectrum of the Fast Fourier transformed integrated 'omics' data described as in legend of Figure 5.3. Seven consensus clusters are produced and the responses in time domain for each cluster were plotted. The *y*-axis represents ratios of the median response of the salinity stress sample to control sample for the molecule. The data were scaled for biplot analysis. The *x*-axis represents 4 sampling points, from 1, 4, 8, 24 hrs. The bottom two diagrams are plots of ratio responses of Vvi.1440 and M217 against time. Vvi.1440 and M217 which were removed previously due to their outlier values were added back to phase analysis, Vvi.1440 was classified in Cluster 1, and M217 was classified in Cluster 7.

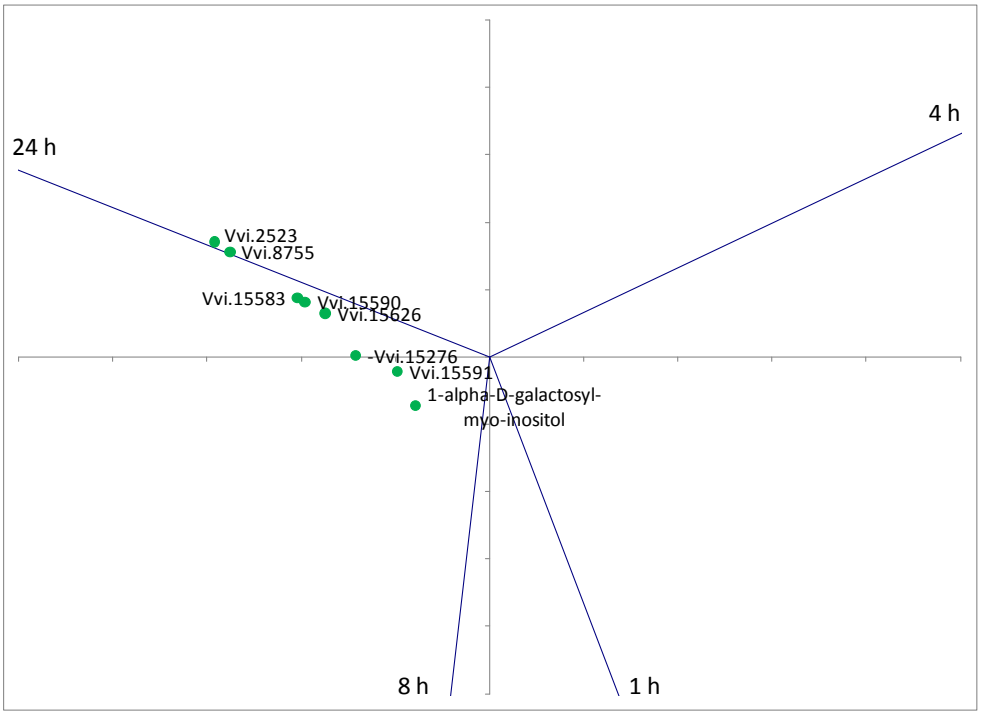
A


24 h
4 h
8 h
1 h
Vvi.3077
Vvi.8050
Vvi.14959
Vvi.14452
Vvi.8873
Vvi.15357
L-isoleucine
Vvi.15604
Vvi.7397
Vvi.3469
Vvi.15096
Vvi.11341
Vvi.14333
beta-alanine
Vvi.15263
Vvi.3340
L-leucine
Vvi.3838
Vvi.6220
M55
L-valine
Vvi.523
Vvi.15630
L-glutamate
Glyceric acid
P.3115

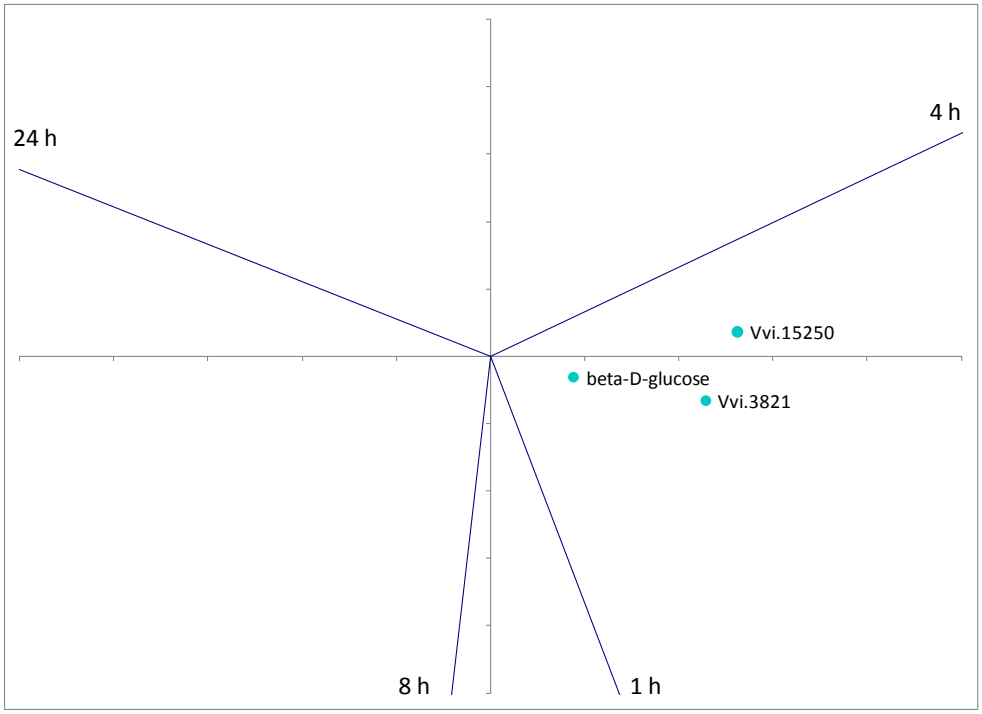
B


4 h
24 h
8 h
1 h
P.7708
Vvi.7869
Vvi.3651
Vvi.3736
Vvi.12905
Vvi.7044
Vvi.15360
Vvi.3998
Vvi.15364
Vvi.4735
M217.1
Vvi.2492
D-2-Hydroxypropanoic acid
Vvi.3955
hydroxylamine
P.103
P.7414

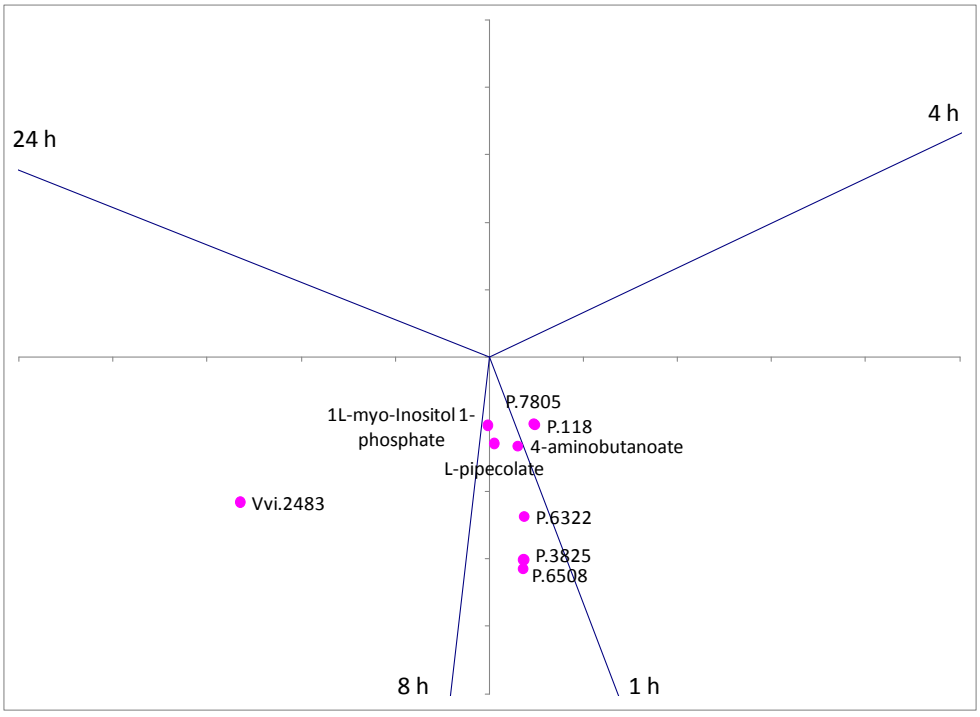
C


24 h
4 h
Vvi.2523
Vvi.8755
Vvi.15583
Vvi.15590
Vvi.15626
-Vvi.15276
Vvi.15591
1-alpha-D-galactosyl-
myo-inositol
8 h
1 h

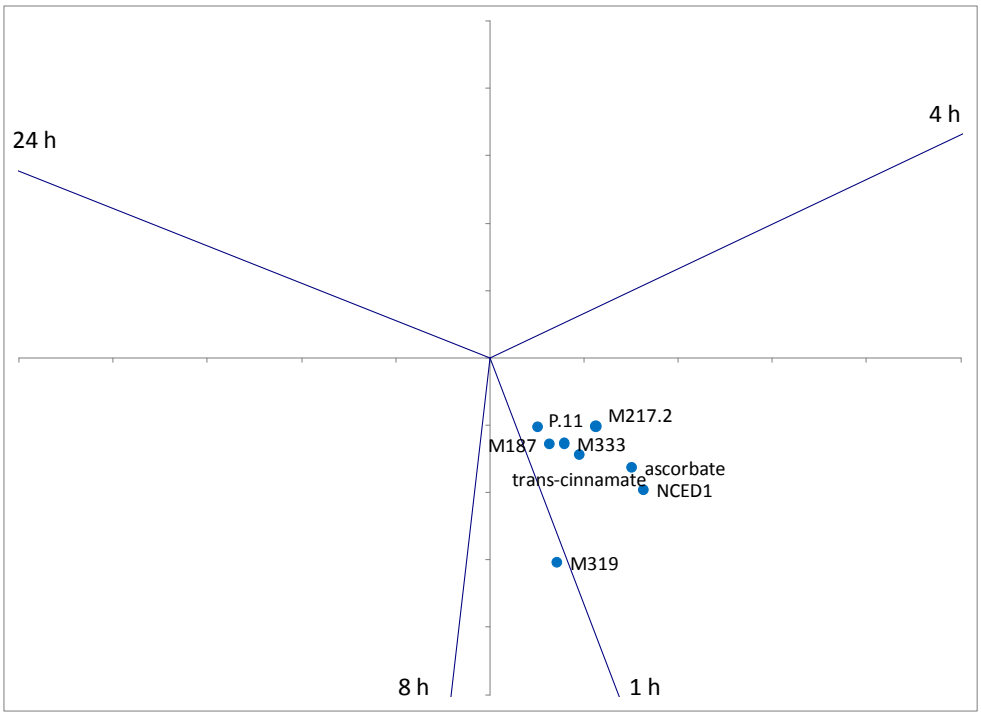
D


24 h
4 h
Vvi.15250
beta-D-glucose
Vvi.3821
8 h
1 h

E


24 h
4 h
1L-myo-Inositol 1-phosphate
P.7805
P.118
4-aminobutanoate
L-pipecolate
Vvi.2483
P.6322
P.3825
P.6508
8 h
1 h

F


24 h
4 h
P.11
M217.2
M187
M333
ascorbate
trans-cinnamate
NCED1
M319
8 h
1 h

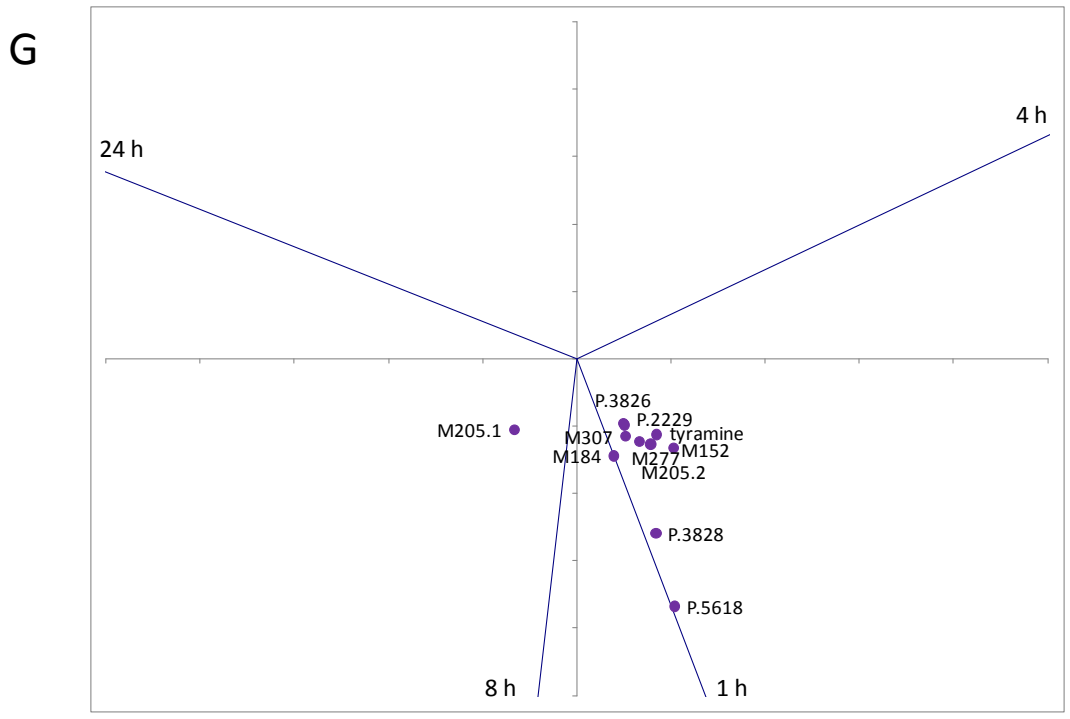


Figure 5.10: Color-highlighted *k*-means clustering on biplot display. Seven clusters produced by *k*-means clustering on phase spectrum of the integrated 'omics' data are highlighted respectively on biplot produced in Figure 5.3. A: Cluster 1, which includes 24 gene transcripts, 7 metabolites and 1 protein are represented by red dots. B: Cluster 2, which includes 11 mRNAs, 3 metabolites and 3 proteins are represented by green dots. C: Cluster 3, which includes 7 mRNAs and 1 metabolite are represented by orange dots. D: Cluster 4, which includes 2 mRNAs and 1 metabolite. E: Cluster 5, which includes 1 mRNA, 3 metabolites and 6 proteins. F: Cluster 6, which includes 1 mRNA, 6 metabolites and 1 protein. G: Cluster 7, which includes 8 metabolites and 4 proteins.

sample was computed at each time point and used as input signal. Phase data were computed from Fast Fourier transform of the time course data. The length of the frequency (column variable) is 2 (half of the length of the original time series, 4). *k*-means clustering (KMC) was implemented to classify phase data to groups to facilitate our understanding of the phase spectrum. Using Euclidean distance as distance metric, seven consensus groups were produced with the software MeV. 4.3.

KMC reveals that among the seven classified clusters, genes are mainly classified in Cluster 1, 2 and 3; metabolites are classified in Clusters 1, 6 and 7; proteins are classified in Cluster 5 and 7. The graph representation of each cluster (Figure 5.8) in the frequency domain illustrates that Clusters 1, 3 and 6 show an upward sloping/ascending trend line, Clusters 2, 4 and 5 exhibit an downward sloping/descending trend line, and Cluster 7 is near horizontal trend line. Plotting of these clusters in time domain uncovered the dynamics of each cluster (Figure 5.9). Cluster 1 and 3 represent late response molecules that peaked at 24 hrs, which include most gene transcripts. Cluster 2 and 4 represent molecules that peaked at 4 hrs, which also include most gene transcripts. Cluster 5 and 6 represent molecules that peaked at 8 hrs. Cluster 7 represents molecules that peaked at 1 hr, which includes most metabolites. Inspection of these molecules finds that Cluster 1 and 2 are also revealed by integrated biplot and highlighted in gray ellipse in Figure 5.3; Cluster 6 and 7 are highlighted in orange ellipse in Figure 5.3; Cluster 5 is highlighted in green ellipse in biplot (Figure 5.3).

The analysis results of integrated biplot and phase spectrum analysis are compared by showing the KMC clusters on the biplot (Figure 5.10). Using the SVD results produced in Figure 5.4, gene transcripts, proteins and metabolites are selectively highlighted on biplot based on their KMC affiliation. Therefore seven biplots were created representing respectively those corresponding molecules in three KMC clusters. Cluster 1 is represented by red dots in Figure 5.10A; Cluster 2 is represented by orange dots in Figure 5.10B; Cluster 3 is represented by green dots in Figure 5.10C; Cluster 4, cyan dots in Figure 5.10D; Cluster 5, pink dots in Figure 5.10E; Cluster 6, blue dots in Figure 5.10F; Cluster 7, purple dots in Figure 5.10G.

These two results strongly contribute to deciphering molecules that peaked at 4 and 24 hours, which include the most highly induced gene transcripts and metabolites. Although there are only two frequencies in phase spectra, the integration based on phase spectra can differentiate molecules peaked at different times and further classify them based on their response patterns. The results demonstrate that phase analysis is very effective in identifying the molecules' dynamics regardless of their magnitudes.

The advantage of phase spectrum analysis is that it can take data with different scales of measurement, and reveals the dynamics of the subjects without being affected by their

magnitudes. This feature is especially desired in systems biology data integration where the difference in the scales of measurement can be several orders of magnitude. The drawback of this method is the relatively high requirements on the input data, where more sampling times with a power of 2 are preferred for a less-compromised analysis.

.

Chapter 6

Summary: Looking back and looking ahead

6.1 Looking back

The concept of "systems theory" emerged more than half a century ago (Abraham 2002), defined at the time as explaining biological phenomena in terms of how the objects are related rather than what they are composed of (Mesarovic, Sreenath et al. 2004). It was not until the beginning of twenty-first century, that the systems biology's role in life sciences was recognized as to "revolutionize our understanding of complex biological regulatory systems" (Kitano 2002). The driving force for its growth is high-throughput technologies that allow us to enumerate biological components on a large scale (Palsson 2006).

The major challenges in systems biology arise from the need to integrate these large-scale data generated from multiple experimental methods to gain a system view of cells. One of the frontiers of systems biology is to capture dynamic properties of the underlying molecular networks of various biological processes with time course and spatial distribution. System-wide time course data is especially valuable, as this allows us to observe the system's response to stimuli over time. Methods to analyze these heterogeneous time course data in an integrative way are much needed. The work presented in this dissertation characterized the experimental data created in systems biology study, explored the mathematical and statistical methodologies and visualization tools to cope with the inherent problems in data fusion to gain insight from these comprehensive data obtained from different technical measurements.

Characterize experimental systems biology data: The massive amounts of data generated through the joint effort of numerous scientists across multiple fields have created an increased demand for quantitative analysis methods. Most of the multivariate inferential procedures are based on the multivariate normal distribution because of its mathematical tractability. Thus knowing the distribution of the experimental data is the prerequisite of choosing the appropriate approaches and applying these methods. Chi square Q-Q plot and violin plot have been applied to *Medicago truncatula* data (provided by Dixon and Sumner labs from Samuel Roberts Noble Foundation (SRNF) and *Vitis vinifera* data (provided by Cramer lab from University of Nevada-Reno).

Almost all 'omics' data exhibited a right-skewed distribution on their Chi square Q-Q plots, and a "Hershey's kisses" shape on their violin plots. It indicates that most of the "omic" molecules have a low relative response level and only a few display large responses. Using a response ratio between elicited and control sample for each molecule has made the distribution less skewed and close to normal. It was noted that all the 'omics' data in these two projects were relative quantities and missing values were prevalent in *M. truncatula* proteomic data (most likely as a result of misidentification of spots in gels).

Biplot display as a visualization tool: Visualization is an effective analytical technique that exploits the ability of the human brain to process large amounts of data. The biplot display provides a method for reducing the dimensionality of the systems biological data and displaying the molecules and time points jointly on the same plot. With the biplot a vast amount of information about a set of experiments and their (high dimensional) results can be inferred visually from a single plot. This is particularly important in time course experiments where analysis of the biplot allows for a rapid identification of a) which parts of the time course are most related with each other, b) which molecules have similar patterns in the time line, and c) in which regions of the time course do each molecules peak.

Using simulated data from yeast glycolysis and glycerol biosynthesis model, the features of Biplot were illustrated and discussed at length. With the aid of singular value decomposition, we can visualize the trajectory of the biological system in phase space biplot. By introducing additive noise to Biplots, we found that Biplot display is robust to a certain level of fixed additive noise. The subsequent application of Biplot to *M. truncatula*─methyl jasmonate elicitation data revealed: the response pattern in terms of the early and late times, an unidentified compound that displayed strongest response to elicitation, and the carbohydrates and amino acids that respectively accumulated at early and late times. The Biplot on *M. truncatula*─yeast elicitation data uncovered the flavonoids that were most highly induced following yeast elicitation, their relative induction levels and the dynamics of their responses. Biplot has demonstrated itself as a powerful exploratory tool for systems biological data.

Data integration based on phase spectrum: Fourier transform is widely used in signal processing and time series analysis. In these applications, Fourier transform resolves a time-domain function into a frequency spectrum. In the frequency domain representation of the original signal, frequency spectrum is usually presented as amplitude and phase, both plotted versus frequency. It has found that phase plays more important roles than magnitudes in the signal reconstruction. To find out how a phase spectrum helps us compare the behavior of different molecules, Fast Fourier transform and phase extraction were implemented to *in silico* data sets from Claytor and AB2 networks. The phase data obtained from simulations of those netwroks were unwrapped to remove data discontinuity. Studies on the impact of precision on phase analysis have found that loss of precision has deteriorated the phase spectrum analysis. Combined with either *k*-means clustering (KMC) or biplot, phase spectrum demonstrated its potential as a much needed tool in systems biology to integrate mRNA, protein and metabolite data regardless of differences in their magnitudes.

Data integration and data fusion in systems biology study: Data integration and data fusion are two facets of data processing; the former emphasizes data management or data acquisition; the latter emphasizes data analysis and prediction. In terms of data integration, Dr. Pedro Mendes group has developed a software system called “DOME”, to store, analyze

and integrate functional genomics data. This study has mainly focused on "data fusion" and oftentimes "data integration" was used to refer to "data fusion". Two situations were considered: the first is the data to be fused refers to the same molecules in the same sample, but which have been measured by different instruments; for example, data overlapping in metabolome profiling. The second situation is the data to be fused refer to different molecules measured by different technologies, although they have been extracted from the same sample, *i.e.* different 'omes' measured from the same cell culture.

To study the data overlapping problem, responses of the amino acids profiled in GC-MS and CE-MS were compared. The results revealed that although 40% same amino acids profiled by different technologies were significantly correlated, 80% amino acids had a higher correlation with other amino acids than with themselves measured by the other technique. It indicates that combining data measured from different technologies may not be appropriate. To fuse different 'omes' data, Biplot analysis was applied to scaled *V. vinifera* salinity stress data, and phase spectrum combined with *k*-means clustering was applied to *M. truncatula* yeast elicited LC-MS data and compared with Biplot analysis. Through integration, the response patterns of different highly induced molecules are revealed in terms of their dynamics. With the aid of KMC, phase spectrum analysis classified molecules into three clusters based on their dynamics. The molecules that were not separated in Biplot were differentiated by phase spectrum, which suggests phase analysis may be a better tool in systems biological data integration.

6.2 Looking ahead

The promise of systems biology is to revolutionize the practice of medicine. The current clinical medicine has been largely influenced by reductionism (Ahn, Tewari et al. 2006). It is shown especially in the practice that the focus is on a singular, dominating factor and symptoms and signs are analyzed one by one in trying to find their cause. There are limitations to the current medical science, an explanation must be sought to complement it. This explanation lies in a systems-oriented view (Laubenbacher, Hower et al. 2009).

The Traditional Chinese Medicine (TCM) has provided a "holistic" model for medical sciences. It evolved from the way body systems are viewed to affect one another and by the environment with great emphasis on balance (Marcus and Kuchera 2005). Identification of patterns indicates the process of identifying the basic disharmony that underlies all clinical manifestations. This is the essence of Chinese medical diagnosis and pathology. Identifying a pattern involves discerning the underlying pattern of disharmony by considering the picture formed by all symptoms and signs. Identifying patterns also follows the typical way of Chinese natural philosophy which looks for relationships rather than causes (Maciocia 2005).

The advancement of systems biology has a two-fold impact on the future of medicine (Weston and Hood 2004): First, systems biology will continually improve our capacity to understand and model biological systems on a more global and in-depth scale than ever before. Second, the new technologies that are driven by the needs of systems biology will enhance the efficiency, scale and precision with which cellular measurements are made. The emerging field of systems biology will have a major role in creating a predictive, preventative, and personalized approach to medicine (Hood, Heath et al. 2004). This will in turn propel medical sciences into a systems medicine era.

This page is intentionally left blank.

Bibliography

Abdi, H. (2007). Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). *Encyclopedia of Measurement and Statistics*. N. Salkind. Thousand Oaks (CA), Sage.

Abraham, R. H. (2002). The Genesis of Complexity. *Advances in Systems Theory, Complexity, and the Human Sciences*. A. Montuori.

Achnine, L., D. V. Huhman, et al. (2005). "Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume Medicago truncatula." *The Plant Journal* **41**(6): 875-887.

Adler, D. (2005). Violin plot, Comprehensive R Archive Network (CRAN)**:** A violin plot is a combination of a box plot and a kernel density plot.

Aebersold, R. (2003). "Constellations in a cellular universe." *Nature* **422**(6928): 115-116.

Agrawal, R., C. Faloutsos, et al. (1993). Efficient Similarity Search In Sequence Databases. *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*. London, UK, Springer-Verlag.

Ahn, A. C., M. Tewari, et al. (2006). "The limits of reductionism in medicine: could systems biology offer an alternative?" *PLoS Med* **3**(6): e208.

Alter, O., P. O. Brown, et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." *Proc Natl Acad Sci U S A* **97**(18): 10101-10106.

Alves, R., F. Antunes, et al. (2006). "Tools for kinetic modeling of biochemical networks." *Nat Biotechnol* **24**(6): 667-672.

American Society for Cell Biology., Ed. (1998). *Methods in cell biology*. Laser tweezers in cell biology. New York,, Academic Press.

Azzalini, A. (1985). "A Class of Distributions Which Includes the Normal Ones." *Scandinavian Journal of Statistics* **12**(2): 171-178.

Baltenweck-Guyot, R., J. M. Trendel, et al. (1997). "New Hemiterpene Glycosides in Vitis vinifera Wine." *Journal of Natural Products* **60**(12): 1326-1327.

Barker, J. S., P. D. East, et al. (1986). "Temporal and microgeographic variation in allozyme frequencies in a natural population of Drosophila buzzatii." *Genetics* **112**(3): 577-611.

Barnes, S. (2003). *Soy isoflavones—phytoestrogens and what else?* The Fifth International Symposium on the Role of Soy in Preventing and Treating Chronic Disease, Orlando, FL, supplement to The Journal of Nutrition, 134: 1225S–1228S, 2004.

Bollen, K. A. and S. Ward (1979). "Ratio Variables in Aggregate Data Analysis: Their Uses, Problems, and Alternatives." *Sociological Methods Research* **7**(4): 431-450.

Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." *Bioinformatics* **19**(2): 185–193.

Boss, W. F., A. J. Davis, et al. (2006). Phosphoinositide Metabolism: Towards an Understanding of Subcellular Signaling. *Biology of Inositols and Phosphoinositides*. J. R. Harris, B. B. Biswas and P. Quinn, Springer US. **39:** 181-205.

Bradu, D. and K. R. Gabriel (1978). "The Biplot as a Diagnostic Tool for Models of Two-Way Tables." *Technometrices* **20**(1): 47-68.

Brazma, A. and J. Vilo (2000). "Gene expression data analysis." *FEBS Lett* **480**(1): 17-24.

Broeckling, C. D., D. V. Huhman, et al. (2005). "Metabolic profiling of Medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism." *J Exp Bot* **56**(410): 323-336.

Büning, H. (2000). "Robustness and power of parametric, nonparametric, robustified and adaptive tests—The multi-sample location problem." *Statistical Papers* **41**(4): 381-407.

BylesjoÌˆ, M., R. Nilsson, et al. (2008). "Integrated Analysis of Transcript, Protein and Metabolite Data To Study Lignin Biosynthesis in Hybrid Aspen." *Journal of Proteome Research* **8**(1): 199-210.

Camacho, D. M. (2007). In silico cell biology and biochemistry a systems biology approach. [Blacksburg, Va., University Libraries, Virginia Polytechnic Institute and State University.

Carlson, R. W., J. Sanjuan, et al. (1993). "The structures and biological activities of the lipo-oligosaccharide nodulation signals produced by type I and II strains of Bradyrhizobium japonicum." *Journal of Biological Chemistry* **268**(24): 18372-18381.

Carmona-Saez, P., R. D. Pascual-Marqui, et al. (2006). "Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization." *BMC Bioinformatics* **7**: 78.

Castle, A. L., O. Fiehn, et al. (2006). "Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results." *Brief Bioinform* **7**(2): 159-165.

Chambers, J. M., W. S. Cleveland, et al. (1983). *Graphical Methods for Data Analysis*, Duxbury Press.

Chapman, S., P. Schenk, et al. (2001). "Using biplots to interpret gene expression patterns in plants." *Bioinformatics* **18**(1): 202-204.

Chen, Y., E. R. Dougherty, et al. (1997). "Ratio-based decisions and the quantitative analysis of cDNA microarray images." *Journal of Biomedical Optics* **2**(4): 364-374.

Chen, Y., V. Kamat, et al. (2002). "Ratio statistics of gene expression levels and applications to microarray data analysis." *Bioinformatics* **18**(9): 1207-1215.

Chissom, B. S. (1970). "Interpretation of the Kurtosis Statistic." *The American Statistician* **24**(4): 19-22.

Chong, L. and L. B. Ray (2002). "Whole-istic Biology." *Science* **295**(5560): 1661.

Cook, D. R. (1999). "Medicago truncatula--a model in the making!" *Curr Opin Plant Biol* **2**(4): 301-304.

Cook, D. R. (1999). "*Medicago truncatula* - a model in the making." *Curr. Opin. Plant Biol.* **2**(4): 301-304.

Cooper, J. (2007). "Early interactions between legumes and rhizobia: disclosing complexity in a molecular dialogue." *Journal of Applied Microbiology* **103**(5): 1355-1365.

Cox, C. and K. R. Gabriel (1981). *Some comparisons of biplot display and pencil-and-paper E.D.A. methods*. Rochester, N.Y., University of Rochester. Division of Biostatistics.

Cox, C., K. R. Gabriel, et al. (1981). *Some comparisons of biplot display and pencil-and-paper E.D.A. methods*. Rochester, N.Y., University of Rochester. Division of Biostatistics.

Cramer, G., A. Ergül, et al. (2007). "Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles." *Functional & Integrative Genomics* **7**(2): 111-134.

Cramer, G. R., J. C. Cushman, et al. (2002). Genomics and stress tolerance of grapes. NSF. **Plant Genome Program**.

Cramer, G. R., A. Ergul, et al. (2007). "Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles." *Funct Integr Genomics* **7**(2): 111-134.

Creelman, R. A. and J. E. Mullet (1997). "BIOSYNTHESIS AND ACTION OF JASMONATES IN PLANTS." *Annual Review of Plant Physiology and Plant Molecular Biology* **48**(1): 355-381.

Cronwright, G. R., J. M. Rohwer, et al. (2002). "Metabolic control analysis of glycerol synthesis in saccharomyces cerevisiae." *Appl. Environ. Microbiol.* **68**(9): 4448-4456.

Cui, X., M. K. Kerr, et al. (2003). "Transformations for cDNA microarray data." *Stat Appl Genet Mol Biol* **2**: Article4.

D'Haeze, W., J. Glushka, et al. (2004). "Structural characterization of extracellular polysaccharides of Azorhizobium caulinodans and importance for nodule initiation on Sesbania rostrata." *Molecular Microbiology* **52**(2): 485-500.

De Backer, P., D. De Waele, et al. "Ins and outs of systems biology vis-a-vis molecular biology: continuation or clear cut?" *Acta Biotheor* **58**(1): 15-49.

Deavours, B. E. and R. A. Dixon (2005). "Metabolic engineering of isoflavonoid biosynthesis in alfalfa." *Plant Physiol* **138**(4): 2245-2259.

Demir, R. and O. Cakmak (2007). "Investigation on fatty acids in leaves, stems and fruits of some species of Medicago." *International Journal of Agriculture and Biology* **9**(6): 934-936.

Dixon, R. A. (2001). "Natural products and plant disease resistance." *Nature* **411**(6839): 843-847.

Dixon, R. A. and L. W. Sumner (2003). "Legume natural products: understanding and manipulating complex pathways for human and animal health." *Plant Physiol* **131**(3): 878-885.

Du, S., P. Sajda, et al. (2005). "Recovery of Metabolomic Spectral Sources using Non-negative Matrix Factorization." *Conf Proc IEEE Eng Med Biol Soc* **2**: 1095-1098.

Dudoit, S., Y. H. Yang, et al. (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments." *Statistica Sinica* **12**: 111-139.

Easton, G. S. and R. E. McCulloch (1990). "A Multivariate Generalization of Quantile-Quantile Plots." *Journal of the American Statistical Association* **85**(410): 376-386.

Ebdon, J. S. and H. G. J. Gauch (2002). "Additive Main Effect and Multiplicative Interaction Analysis of National Turfgrass Performance Trials." *Crop Science* **42**: 489-496.

Eeuwijk, F. A. (1995). "Multiplicative Interaction in Generalized Linear Models." *Biometrics* **51**(3): 1017-1032.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." *Proc Natl Acad Sci U S A* **95**(25): 14863-14868.

ELAD, Y. (1997). "RESPONSES OF PLANTS TO INFECTION BY BOTRYTIS CINEREA AND NOVEL MEANS INVOLVED IN REDUCING THEIR SUSCEPTIBILITY TO INFECTION." *Biological Reviews* **72**(03): 381-422.

Evans, P. R. (2005). Fitting, Refinement & Validation. http://www-structmed.cimr.cam.ac.uk/Course/Fitting/fittingtalk.html#resolution

Farag, M. A., B. E. Deavours, et al. (2009). "Integrated Metabolite and Transcript Profiling Identify a Biosynthetic Mechanism for Hispidol in Medicago truncatula Cell Cultures." *Plant Physiol.* **151**(3): 1096-1113.

Farag, M. A., D. V. Huhman, et al. (2008). "Metabolomics Reveals Novel Pathways and Differential Mechanistic and Elicitor-Specific Responses in Phenylpropanoid and Isoflavonoid Biosynthesis in Medicago truncatula Cell Cultures." *Plant Physiol.* **146**(2): 387-402.

Farmer, E. E. (1994). "Fatty acid signalling in plants and their associated microorganisms." *Plant Molecular Biology* **26**(5): 1423-1437.

Ferrer, J. L., J. M. Jez, et al. (1999). "Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis." *Nat-Struct-Biol.* **6**(8): 775-784.

Friedman, M. (1999). "Chemistry, Nutrition, and Microbiology of d-Amino Acids." *Journal of Agricultural and Food Chemistry* **47**(9): 3457-3479.

Funahashi, A., Y. Matsuoka, et al. (2008). "CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks." *Proceedings of the IEEE* **96**(8): 1254-1265.

Gabriel, K. R. (1971). "**The Biplot Graphic Display of Matrices with Application to Principal Component Analysis**." *Biometrika* **58**(3): 453-467.

Gao, Y. and G. Church (2005). "Improving molecular cancer class discovery through sparse non-negative matrix factorization." *Bioinformatics* **21**(21): 3970-3975.

Gentleman, R., F. Hahne, et al. (2006). "Visualizing Genomic Data." *Bioconductor Project Working Papers* **Working Paper 10**.

German, J. B. and R. L. Walzem (2000). "The health benefits of wine." *Annu Rev Nutr* **20**: 561-593.

Gibbons, F. D. and F. P. Roth (2002). "Judging the quality of gene expression-based clustering methods using gene annotation." *Genome Res* **12**(10): 1574-1581.

Goes da Silva, F., A. Iandolino, et al. (2005). "Characterizing the grape transcriptome. Analysis of expressed sequence tags from multiple Vitis species and development of a compendium of gene expression during berry development." *Plant Physiol* **139**(2): 574-597.

Goodacre, R., S. Vaidyanathan, et al. (2004). "Metabolomics by numbers: acquiring and understanding global metabolite data." *Trends Biotechnol* **22**(5): 245-252.

Gower, J. C. and D. J. Hand (1996). *Biplots*. London, Chapman & Hill.

Goyal, K., L. J. Walton, et al. (2005). "LEA proteins prevent protein aggregation due to water stress." *Biochem. J.* **388**(1): 151-157.

Hall, D. L. and S. A. H. McMullen (2004). *Mathematical Techniques in Multisensor Data Fusion*, Artech House.

Halling, A., G. Fridh, et al. (2006). "Validating the Johns Hopkins ACG Case-Mix System of the elderly in Swedish primary health care." *BMC Public Health* **6**: 171.

Harms, K., R. Atzorn, et al. (1995). "Expression of a Flax Allene Oxide Synthase cDNA Leads to Increased Endogenous Jasmonic Acid (JA) Levels in Transgenic Potato Plants but Not to a Corresponding Activation of JA-Responding Genes." *Plant Cell.* **7**(10): 1645-1654.

Harris, C. M. (1998). "The Fourier analysis of biological transients." *J Neurosci Methods* **83**(1): 15-34.

Harrison, N. (2003) "Fourier Series & Fourier Transforms."

Haruta, M. and C. P. Constabel (2003). "Rapid Alkalinization Factors in Poplar Cell Cultures. Peptide Isolation, cDNA Cloning, and Differential Expression in Leaves and Methyl Jasmonate-Treated Cells." *Plant Physiol.* **131**(2): 814-823.

Hayes, M., L. Jae, et al. (1980). "Signal reconstruction from phase or magnitude." *Acoustics, Speech and Signal Processing, IEEE Transactions on* **28**(6): 672-680.

Hintze, J. L. and R. D. Nelson (1998). "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician* **52**(2): 181-184.

Holgersson, H. E. T. (2006). "A graphical method for assessing multivariate normality." *Computational Statistics* **21**(1): 141-149.

Holter, N. S., M. Mitra, et al. (2000). "Fundamental patterns underlying gene expression profiles: simplicity from complexity." *Proc Natl Acad Sci U S A* **97**(15): 8409-8414.

Hood, L., J. R. Heath, et al. (2004). "Systems biology and new technologies enable predictive and preventative medicine." *Science* **306**(5696): 640-643.

Hoops, S., S. Sahle, et al. (2006). "COPASI--a COmplex PAthway SImulator." *Bioinformatics* **22**(24): 3067-3074.

Hoops, S., S. Sahle, et al. (2006). "COPASI - a COmplex PAthway SImulator." *Bioinformatics*.

Howe, G. A. and A. L. Schilmiller (2002). "Oxylipin metabolism in response to stress." *Current Opinion in Plant Biology* **5**(3): 230-236.

Howell, D. C. (July 11, 1998). "Treatment of missing data." 2005, from http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

Hoyle, D. C., M. Rattray, et al. (2002). "Making sense of microarray data distributions." *Bioinformatics* **18** (4): 576–584.

Hu, J., M. B. Reddy, et al. (2004). "Soyasaponin I and Sapongenol B Have Limited Absorption by Caco-2 Intestinal Cells and Limited Bioavailability in Women." *J. Nutr.* **134**(8): 1867-1873.

Huhman, D. V. and L. W. Sumner (2002). "Metabolic profiling of saponins in Medicago sativa and Medicago truncatula using HPLC coupled to an electrospray ion-trap mass spectrometer." *Phytochemistry* **59**(3): 347-360.

Huhman, D. V. and L. W. Sumner (2002). "Metabolic profiling of saponins in Medicago sativa and Medicago truncatula using HPLC coupled to an electrospray ion-trap mass spectrometer." *Phytochemistry* **59**(3): 347-360.

Ideker, T., T. Galitski, et al. (2001). "A new approach to decoding life: systems biology." *Annu Rev Genomics Hum Genet* **2**: 343-372.

Iriti, M., M. Rossoni, et al. (2005). "Induction of Resistance to Gray Mold with Benzothiadiazole Modifies Amino Acid Profile and Increases Proanthocyanidins in Grape: Primary versus Secondary Metabolism." *Journal of Agricultural and Food Chemistry* **53**(23): 9133-9139.

Irizarry, R. A., B. M. Bolstad, et al. (2003). "Summaries of Affymetrix GeneChip probe level data." *Nucleic Acids Research* **31**(4): e15.

Iuchi, S., M. Kobayashi, et al. (2001). "Regulation of drought tolerance by gene manipulation of 9-cis-epoxycarotenoid dioxygenase, a key enzyme in abscisic acid biosynthesis in Arabidopsis." *The Plant Journal* **27**(4): 325-333.

Jones, K. M., H. Kobayashi, et al. (2007). "How rhizobial symbionts invade plants: the Sinorhizobium-Medicago model." *Nat Rev Micro* **5**(8): 619-633.

Karas, V. (1997). "Fourier-phase analysis of the orbiting bright-spot model for active galactic nucleus variability." *Monthly Notices of the Royal Astronomical Society* **288**(1): 12-18.

Kell, D. B. (2004). "Metabolomics and systems biology: making sense of the soup." *Curr Opin Microbiol* **7**(3): 296-307.

Kell, D. B. (2005). "Metabolomics, machine learning and modelling: towards an understanding of the language of cells." *Biochem Soc Trans* **33**(Pt 3): 520-524.

Kell, D. B. (2006). "Systems biology, metabolic modelling and metabolomics in drug discovery and development." *Drug Discov Today* **11**(23-24): 1085-1092.

Kell, D. B., M. Brown, et al. (2005). "Metabolic footprinting and systems biology: the medium is the message." *Nat Rev Microbiol* **3**(7): 557-565.

Kennedy, J. A., M. A. Matthews, et al. (2000). "Changes in grape seed polyphenols during fruit ripening." *Phytochemistry* **55**(1): 77-85.

Khattree, R. and D. Naik (1999). *Applied Multivariate Statistics with SAS Software*. Cary, NC, SAS Institute Inc.

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

Kim, S., E. R. Dougherty, et al. (2000). "Multivariate measurement of gene expression relationships." *Genomics* **67**(2): 201-209.

Kitano, H. (2002). "Computational systems biology." *Nature* **420**(6912): 206-210.

Kitano, H. (2002). "Systems biology: a brief overview." *Science* **295**(5560): 1662-1664.

Kitano, H., A. Funahashi, et al. (2005). "Using process diagrams for the graphical representation of biological networks." *Nat Biotechnol* **23**(8): 961-966.

Klebanov, L. and A. Yakovlev (2007). "How high is the level of technical noise in microarray data?" *Biology Direct* **2**(1): 9.

Kolb, C. A., M. A. Kaser, et al. (2001). "Effects of natural intensities of visible and ultraviolet radiation on epidermal ultraviolet screening and photosynthesis in grape leaves." *Plant Physiol* **127**(3): 863-875.

Konishi, T. (2004). "Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment." *BMC Bioinformatics* **5**: 5.

Kortemeyer, G. (2009). *LON-CAPA - An Open-Source Learning Content Management and Assessment System for the Sciences*. m-ICTE 2009 Conference, Lisbon.

Kouchi, H., H. Imaizumi-Anraku, et al. (2010). "How Many Peas in a Pod? Legume Genes Responsible for Mutualistic Symbioses Underground." *Plant and Cell Physiology* **51**(9): 1381-1397.

Landsman, W. B. (1993). *Astronomical Data Analysis Software and Systems II*. Astronomical Society of the Pacific Conference Series.

Laubenbacher, R., V. Hower, et al. (2009). "A systems biology view of cancer." *Biochim Biophys Acta* **1796**(2): 129-139.

Lee, D. D. and H. S. Seung (1997). Unsupervised Learning by Convex and Conic Coding. *Advances in Neural Information Processing Systems*. M. C. Mozer, M. I. Jordan and T. Petsche. Cambridge, MA, MIT Press. **9:** 515.

Lee, D. D. and H. S. Seung (1999). "Learning the parts of objects by non-negative matrix factorization." *Nature* **401**(6755): 788-791.

Lei, Z., F. Chen, et al. (2010). "Comparative Proteomics of Yeast-Elicited Medicago truncatula Cell Suspensions Reveal Induction of Isoflavonoid Biosynthesis and Cell Wall Modifications." *Journal of Proteome Research*: null-null.

Lei, Z., A. M. Elmer, et al. (2005). "A 2-DE proteomics reference map and systematic identification of 1367 proteins from a cell suspension culture of the model legume Medicago truncatula." *Mol Cell Proteomics*.

Lei, Z., A. M. Elmer, et al. (2005). "A Two-dimensional Electrophoresis Proteomic Reference Map and Systematic Identification of 1367 Proteins from a Cell Suspension Culture of the Model Legume Medicago truncatula." *Mol Cell Proteomics* **4**(11): 1812-1825.

Lenzerini, M. (2002). *Data integration: a theoretical perspective*. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Madison, Wisconsin, ACM Press New York, NY, USA.

Lerouge, P., P. Roche, et al. (1990). "Symbiotic host-specificity of Rhizobium meliloti is determined by a sulphated and acylated glucosamine oligosaccharide signal." *Nature* **344**(6268): 781-784.

Li, Y. and G. Lu (2008). University Physics Digital Teaching. D. T. Physics. Wuhan, China physics digital teaching workshop.

Lipkovich, I. and E. P. Smith. (2002, June 6). "Biplot and Singular Value Decomposition Macros for Excel." from http://www.jstatsoft.org/counter.php?id=44&url=v07/i05/BIPLOT_paper_6_6_02.pdf&ct=1.

Liu, R. Y., J. M. Parelius, et al. (1999). "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference." *The Annals of Statistics* **27**(3): 783-840.

Liu, Y. (2003, 01-SEP). "Statistical validity of using ratio variables in human kinetics research." *Research Quarterly for Exercise and Sport*, from http://goliath.ecnext.com/coms2/summary_0199-3186114_ITM&referid=2090.

Llinas, J. and D. L. Hall (1998). *An introduction to multi-sensor data fusion*. Circuits and Systems, 1998. ISCAS '98. Proceedings of the 1998 IEEE International Symposium on, Monterey, CA.

Loewus, F. A. (1999). "Biosynthesis and metabolism of ascorbic acid in plants and of analogs of ascorbic acid in fungi." *Phytochemistry* **52**(2): 193-210.

Lyons, W. (1977). "Per Capita Index Construction: A Defense." *American Journal of Political Science* **21**(1): 177-182.

Mabood, F., A. Souleimanov, et al. (2006). "Jasmonates induce Nod factor production by Bradyrhizobium japonicum." *Plant Physiology and Biochemistry* **44**(11-12): 759-765.

Maciocia, G. (2005). *The Foundations of Chinese Medicine: A Comprehensive Text for Acupuncturists and Herbalists.*, Churchill Livingstone.

Magidson, J. and J. K. Vermunt (2002). "Latent class models for clustering: A comparison with K-means." *Canadian Journal of Marketing Research* **20**: 37-44.

Mann, M. and O. N. Jensen (2003). "Proteomic analysis of post-translational modifications." *Nat Biotech* **21**(3): 255-261.

Marcus, A. and M. Kuchera (2005). *Foundations for integrative musculoskeletal medicine: an east-west approach*, North Atlantic Books.

Martins, A. M., D. Camacho, et al. (2004). " A Systems Biology Study of Two Distinct Growth Phases of Saccharomyces cerevisiae Cultures." *Current Genomics* **5**(8): 649-663.

Matsubayashi, Y. and Y. Sakagami (2006). "PEPTIDE HORMONES IN PLANTS." *Annual Review of Plant Biology* **57**(1): 649-674.

Mayer, R. R., J. H. Cherry, et al. (1990). "Effects of Heat Shock on Amino Acid Metabolism of Cowpea Cells." *Plant Physiol.* **94**(2): 796-810.

McGill, R., J. W. Tukey, et al. (1978). "Variations of Box Plots." *The American Statistician* **32**(1): 12-16.

Mendes, P. (1993). "GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems." *Comput. Appl. Biosci.* **9**(5): 563-571.

Mendes, P. (1997). "Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3." *Trends in Biochemical Sciences* **22**(9): 361-363.

Mendes, P. (2002). "Emerging bioinformatics for the metabolome." *Brief Bioinform* **3**(2): 134-145.

Mendes, P. (2009). Framework for Comparative Assessment of Parameter Estimation and Inference Methods in Systems Biology. *Learning and Inference in Computational Systems Biology*. N. D. Lawrence, M. Girolami, M. Rattray and G. Sanguinetti. Cambridge, Massachusetts, The MIT Press**:** 35-60.

Mendes, P., D. Camacho, et al. (2005). "Modelling and simulation for metabolomics data analysis." *Biochem Soc Trans* **33**(Pt 6): 1427-1429.

Mendes, P. and R. Dixon (2001). An Integrated Approach to Funtional Genomics and Bioinformatics in a Model Legume. NSF. **Grant DBI-0109732**.

Mendes, P. and D. Kell (1998). "Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation." *Bioinformatics* **14**(10): 869-883.

Mendes, P., H. Messiha, et al. (2009). "Enzyme kinetics and computational modeling for systems biology." *Methods Enzymol* **467**: 583-599.

Mendgen, K. and H. Deising (1993). "Tansley Review No. 48 Infection Structures of Fungal Plant Pathogens - a Cytological and Physiological Evaluation." *New Phytologist* **124**(2): 193-213.

Mesarovic, M. D., S. N. Sreenath, et al. (2004). "Search for organising principles: understanding in systems biology." *Syst Biol (Stevenage)* **1**(1): 19-27.

Meuriot, F., C. Noquet, et al. (2004). "Methyl jasmonate alters N partitioning, N reserves accumulation and induces gene expression of a 32-kDa vegetative storage protein that possesses chitinase activity in Medicago sativa taproots." *Physiologia Plantarum* **120**(1): 113-123.

Michell, J. (1997). "Quantitative science and the definition of measurement in psychology. ." *British Journal of Psychology* **88**: 355–383.

Misevic, G. N., Y. Guerardel, et al. (2004). "Molecular Recognition between Glyconectins as an Adhesion Self-assembly Pathway to Multicellularity." *Journal of Biological Chemistry* **279**(15): 15579-15590.

Monteiro, S., M. Barakat, et al. (2003). "Osmotin and Thaumatin from Grape: A Putative General Defense Mechanism Against Pathogenic Fungi." *Phytopathology* **93**(12): 1505-1512.

Moser, C., C. Segala, et al. (2005). "Comparative analysis of expressed sequence tags from different organs of *Vitis vinifera* L." *Functional & Integrative Genomics* **5**(4): 208-217.

Mueller, M. J., W. Brodschelm, et al. (1993). "Signaling in the elicitation process is mediated through the octadecanoid pathway leading to jasmonic acid." *Proc Natl Acad Sci U S A.* **90**(16): 7490-7494.

Naoumkina, M. A., L. V. Modolo, et al. (2010). "Genomic and Coexpression Analyses Predict Multiple Genes Involved in Triterpene Saponin Biosynthesis in Medicago truncatula." *Plant Cell* **22**(3): 850-866.

Ni, X. and X. Huo (2007). "Statistical interpretation of the importance of phase information in signal and image reconstruction." *Statistics and Probability Letters* **77**: 447-454.

Niwa, S., M. Kawaguchi, et al. (2001). "Responses of a Model Legume Lotus japonicus to Lipochitin Oligosaccharide Nodulation Factors Purified from Mesorhizobium loti JRL501." *Molecular Plant-Microbe Interactions* **14**(7): 848-856.

Novak, J. P., S. Y. Kim, et al. (2006). "Generalization of DNA microarray dispersion properties: microarray equivalent of t-distribution." *Biol Direct* **1**: 27.

Oleszek, W. and M. Jurzysta (1987). "The allelopathic potential of alfalfa root medicagenic acid glycosides and their fate in soil environments." *Plant and Soil* **98**(1): 67-80.

Olsen, A. N., J. Mundy, et al. (2002). *Peptomics, identification of novel cationic Arabidopsis peptides with conserved sequence motifs*.

Oppenheim, A. V. and J. S. Lim (1981). The Importance of Phase in Signals. *Proceedings of the IEEE*. **69:** 529-541.

Oppenheim, A. V., J. S. Lim, et al. (1983). "Signal synthesis and reconstruction from partial Fourier-domain information." *J. Opt. Soc. Am.* **73**(11): 1413-1420.

Oppenheim, A. V. and R. W. Schafer (1975). *Digital signal processing*, Prentice-Hall, Inc.

Palsson, B. Ø. (2006). *Systems Biology: Properties of Reconstructed Networks*. New York, Cambridge University Press

Paraskevas, I. and E. Chilton (2004). "Combination of magnitude and phase statistical features for audio classification." *Acoustics Research Letters Online* **5**(3): 111-117.

Park, T., S. G. Yi, et al. (2003). "Evaluation of normalization methods for microarray data." *BMC Bioinformatics* **4**: 33.

Pascual-Montano, A., P. Carmona-Saez, et al. (2006). "bioNMF: a versatile tool for non-negative matrix factorization in biology." *BMC Bioinformatics* **7**: 366.

Peterson, G. and S. Barnes (1993). "Genistein and biochanin A inhibit the growth of human prostate cancer cells but not epidermal growth factor receptor tyrosine autophosphorylation." *The Prostate* **22**(4): 335-345.

Pike, R. W. (2001). *Optimization for Engineering Systems*, Van Nostrand Reinhold Company.

Pittelkow, Y. and S. R. Wilson (2005). "Use of principal component analysis and the GE-biplot for the graphical exploration of gene expression data." *Biometrics* **61**(2): 630-632; discussion 632-634.

Pittelkow, Y. E. and S. R. Wilson (2003). *The GE-biplot for microarray data*. Proceedings of the Virtual Conference on Genomics and Bioinformatics.

Pittelkow, Y. E. and S. R. Wilson (2003). "Visualisation of gene expression data – the GE-biplot, the chip-plot and the gene-plot." *Statistical Applications in Genetics and Molecular Biology* **2**(1): Article 6.

Press, W. H., S. A. Teukolsky, et al. (1992). "Numerical Recipes in C: the art of scientific computing." *Numerical Recipes* 2nd. Retrieved 1020 pages, 2006.

Price, N. D., J. L. Reed, et al. (2003). "Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices." *Biophys J* **84**(2 Pt 1): 794-804.

Pritchard, L. and D. B. Kell (2002). "Schemes of flux control in a model of Saccharomyces cerevisiae glycolysis." *Eur J Biochem* **269**(16): 3894-3904.

Qin, X. and J. A. D. Zeevaart (2002). "Overexpression of a 9-cis-Epoxycarotenoid Dioxygenase Gene in Nicotiana plumbaginifolia Increases Abscisic Acid and Phaseic Acid Levels and Enhances Drought Tolerance." *Plant Physiol.* **128**(2): 544-551.

Read, R. J. (2009). Fourier transforms: structure factors, phases and electron density. http://www-structmed.cimr.cam.ac.uk/Course/Fourier/Fourier.html

Reich, J. G. and E. i. E. e. Sel'kov (1981). *Energy metabolism of the cell : a theoretical treatise*. London ; New York, Academic Press.

Rencher, A. C. (2002). *Methods of multivariate analysis*. New York, J. Wiley.

Rencher, A. C. (2002). *Methods of multivariate analysis*, John Wiley & Sons, Inc.

Roche, P., P. Lerouge, et al. (1991). "Structural determination of bacterial nodulation factors involved in the Rhizobium meliloti-alfalfa symbiosis." *Journal of Biological Chemistry* **266**(17): 10933-10940.

Rubin, E. and J. H. Smith (1958). "Questions and Answers." *The American Statistician* **12**(4): 24-25.

Ruppert, D. (1987). " What Is Kurtosis?: An Influence Function Approach." *The American Statistician* **41**(1): 1-5.

Sachse, W. and W. F. Pierce (1990). "Recovery of acoustic source extent by Fourier phase analysis of emitted signals." *The Journal of the Acoustical Society of America* **88**(6): 2736-2742.

Saeed, A. I., V. Sharov, et al. (2003). "TM4: a free, open-source system for microarray data management and analysis." *Biotechniques* **34**(2): 374-378.

Sandberg, R. and O. Larsson (2007). "Improved precision and accuracy for microarrays using updated probe set definitions." *BMC Bioinformatics* **8**: 48.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* **270**(5235): 467-470.

Setchell, K. D. R. and A. Cassidy (1999). "Dietary Isoflavones: Biological Effects and Relevance to Human Health." *J. Nutr.* **129**(3): 758-.

Sharov, V., K. Y. Kwong, et al. (2004). "The limits of log-ratios." *BMC Biotechnol* **4**: 3.

Shulaev, V. (2006). "Metabolomics technology and bioinformatics." *Brief Bioinform* **7**(2): 128-139.

Singleton, V. L., C. F. Timberlake, et al. (1978). "The phenolic cinnamates of white grapes and wine." *Journal of the Science of Food and Agriculture* **29**(4): 403-410.

Smilde, A. K., M. J. van der Werf, et al. (2005). "Fusion of mass spectrometry-based metabolomics data." *Anal Chem* **77**(20): 6729-6736.

Smith, J. O. (2007). Introduction to Digital Filters with Audio Applications, W3K Publishing.

Smith, S. W. (1997). *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Pub.

Smyth, G. K. and T. Speed (2003). "Normalization of cDNA microarray data." *Methods* **31**(4): 265-273.

Smyth, G. K., Y. H. Yang, et al. (2002). Statistical Issues in cDNA Microarray Data Analysis. *Functional Genomics: Methods and Protocols*. M. J. Brownstein and A. B. Khodursky. Totowa, NJ, Humana Press.

Smyth, G. K., Y. H. Yang, et al. (2003). "Statistical issues in cDNA microarray data analysis." *Methods Mol Biol* **224**: 111-136.

Spickett, C. M., A. R. Pitt, et al. (2006). "Proteomic analysis of phosphorylation, oxidation and nitrosylation in signal transduction." *Biochim Biophys Acta* **1764**(12): 1823-1841.

Starke, I., A. Holzberger, et al. (2000). "Qualitative and quantitative analysis of carbohydrates in green juices (wild mix grass and alfalfa) from a green biorefinery by gas chromatography / mass spectrometry." *Fresenius' Journal of Analytical Chemistry* **367**(1): 65-72.

Stevens, S. S. (1946). "On the Theory of Scales of Measurement." *Science* **103**(2684): 677-680.

Stöggl, W. M., C. W. Huck, et al. (2004). "Structural elucidation of catechin and epicatechin in sorrel leaf extracts using liquid-chromatography coupled to diode array-, fluorescence-, and mass spectrometric detection." *Journal of Separation Science* **27**(7-8): 524-528.

Stumpe, M., J. G. Carsjens, et al. (2005). "Lipid metabolism in arbuscular mycorrhizal roots of Medicago truncatula." *Phytochemistry* **66**(7): 781-791.

Sumner, L. W., P. Mendes, et al. (2003). "Plant metabolomics: large-scale phytochemistry in the functional genomics era." *Phytochemistry* **62**(6): 817-836.

Sun, J., V. Cardoza, et al. (2006). "Crosstalk between jasmonic acid, ethylene and Nod factor signaling allows integration of diverse inputs for regulation of nodulation." *The Plant Journal* **46**(6): 961-970.

Sun, J., F. Liang, et al. (2007). "Screening Non-colored Phenolics in Red Wines using Liquid Chromatography/Ultraviolet and Mass Spectrometry/Mass Spectrometry Libraries." *Molecules* **12**: 679-693.

Suzuki, H., L. Achnine, et al. (2002). "A genomics approach to the early stages of triterpene saponin biosynthesis in Medicago truncatula." *Plant J* **32**(6): 1033-1048.

Suzuki, H., M. S. Reddy, et al. (2005). "Methyl jasmonate and yeast elicitor induce differential transcriptional and metabolic re-programming in cell suspension cultures of the model legume Medicago truncatula." *Planta* **220**(5): 696-707.

Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." *Proc Natl Acad Sci U S A* **96**(6): 2907-2912.

Tattersall, D. B., R. van Heeswijck, et al. (1997). "Identification and Characterization of a Fruit-Specific, Thaumatin-like Protein That Accumulates at Very High Levels in Conjunction with the Onset of Sugar Accumulation and Berry Softening in Grapes." *Plant Physiology* **114**(3): 759-769.

Tavazoie, S., J. D. Hughes, et al. (1999). "Systematic determination of genetic network architecture." *Nat Genet* **22**(3): 281-285.

Team, R. D. C. (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing.

Teusink, B. and H. V. Westerhoff (2000). "'Slave' metabolites and enzymes. A rapid way of delineating metabolic control." *Eur J Biochem* **267**(7): 1889-1893.

The Space Telescope Science Institute (April 1994). STSDAS User's Guide, AURA NASA.

Torabinejad, J., J. L. Donahue, et al. (2009). "VTC4 Is a Bifunctional Enzyme That Affects Myoinositol and Ascorbate Biosynthesis in Plants." *Plant Physiol.* **150**(2): 951-961.

Troyanskaya, O., M. Cantor, et al. (2001). "Missing value estimation methods for DNA microarrays." *Bioinformatics* **17**(6): 520-525.

Tsay, R. S. (2006). "R program to compute Chi-square QQ-plot: rqqchi2.txt " *Business 41912: Applied Multivariate Analysis* Retrieved March 13, 2007, from http://faculty.chicagogsb.edu/ruey.tsay/teaching/ama/rqqchi2.txt.

Uslaner, E. M. (1977). "Straight Lines and Straight Thinking: Can all of Those Econometricians be Wrong?" *American Journal of Political Science* **21**(1): 183-191.

Verrills, N. M. (2006). "Clinical proteomics: present and future prospects." *Clin Biochem Rev* **27**(2): 99-116.

Vilain, D., D. Daou, et al. (2001). "Optimal 3-dimensional method for right and left ventricular Fourier phase analysis in electrocardiography-gated blood-pool SPECT " *Journal of Nuclear Cardiology* **8**(3): 371-378.

Vogt, T. and P. Jones (2000). "Glycosyltransferases in plant natural product synthesis: characterization of a supergene family." *Trends in Plant Science* **5**(9): 380-386.

Wais, R. J., D. H. Keating, et al. (2002). "Structure-function analysis of nod factor-induced root hair calcium spiking in Rhizobium-legume symbiosis." *Plant Physiol* **129**(1): 211-224.

Wall, M. E., A. Rechtsteiner, et al. (2003). *A Practical Approach to Microarray Data Analysis*. Norwell, MA, Kluwer.

Walsh, C. T., S. Garneau-Tsodikova, et al. (2005). "Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications." *Angewandte Chemie International Edition* **44**(45): 7342-7372.

Watson, B. S., V. S. Asirvatham, et al. (2003). "Mapping the proteome of barrel medic (Medicago truncatula)." *Plant Physiol* **131**(3): 1104-1123.

Watson, C. J. (1990). "Multivariate Distributional Properties, Outliers, and Transformation of Financial Ratios." *The Accounting Review* **65**(3): 682-695.

Weckwerth, W. (2003). "Metabolomics in systems biology." *Annual Review of Plant Biology* **54**(1): 669-689.

Weisstein, E. W. (1999) "Discrete Fourier Transform." *MathWorld--A Wolfram Web Resource*.

Weisstein, E. W. (1999) "Fourier Series." *MathWorld--A Wolfram Web Resource.*

Weisstein, E. W. (1999) "Fourier Transform." *MathWorld--A Wolfram Web Resource*.

Weisstein, E. W. (1999) "Nyquist Frequency." *MathWorld--A Wolfram Web Resource*.

Weisstein, E. W. (2010). Damped Simple Harmonic Motion--Overdamping.

Weston, A. D. and L. Hood (2004). "Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine." *Journal of Proteome Research* **3**(2): 179-196.

Wiback, S. J., R. Mahadevan, et al. (2004). "Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the Escherichia coli spectrum." *Biotechnol Bioeng* **86**(3): 317-331.

Wikipedia (2009) "Discrete Fourier transform."

Williams, B. J., C. J. Cameron, et al. (2007). "Amino acid profiling in plant cell cultures: An inter-laboratory comparison of CE-MS and GC-MS." *Electrophoresis* **28**(9): 1371-1379.

Winslow, R. L., S. Cortassa, et al. (2005). "Using models of the myocyte for functional interpretation of cardiac proteomic data." *J Physiol* **563**(Pt 1): 73-81.

Xing, C., F. R. Schumacher, et al. (2003). "Comparison of missing data approaches in linkage analysis." *BMC Genet* **4 Suppl 1**: S44.

Yan, W. and L. A. Hunt (2002). "Biplot Analysis of Diallel Data." *Crop Sci* **42**(1): 21-30.

Yang, Y. H., M. J. Buckley, et al. (2001). "Analysis of cDNA microarray images." *Brief Bioinform* **2**(4): 341-349.

Yang, Y. H., S. Dudoit, et al. (2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." *Nucleic Acids Res* **30**(4): e15.

Yendo, A., F. de Costa, et al. (2010). "Production of Plant Bioactive Triterpenoid Saponins: Elicitation Strategies and Target Genes to Improve Yields." *Molecular Biotechnology* **46**(1): 94-104.

Zhu, H. (1998) "Properties of Damped Oscillations Systems." *MIT System Dynamics in Education Project* **D-4767**.

Ziegler, P. and K. R. Dittrich (2004). *Three Decades of Data Integration - All Problems Solved?* 18th IFIP World Computer Congress (WCC 2004) Building the Information Society, Toulouse, France.