Reconstruction of Metabolic Pathways by the Exploration of Gene Expression Data with Factor Analysis

David A. Henderson

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

Ina Hoeschele, Chair
David R. Notter
Eric P. Smith
Saghai Maroof
Pedro Mendes

December 14, 2001
Blacksburg, Virginia

Keywords: Microarray, Factor Analysis, Gene Network, Genetic Regulation

Reconstruction of Metabolic Pathways by the Exploration of Gene Expression Data with Factor Analysis

David A. Henderson

(ABSTRACT)

Microarray gene expression data for thousands of genes in many organisms is quickly becoming available. The information this data can provide the experimental biologist is powerful. This data may provide information clarifying the regulatory linkages between genes within a single metabolic pathway, or alternative pathway routes under different environmental conditions, or provide information leading to the identification of genes for selection in animal and plant genetic improvement programs or targets for drug therapy. Many analysis methods to unlock this information have been both proposed and utilized, but not evaluated under known conditions (*e.g.* simulations). Within this dissertation, an analysis method is proposed and evaluated for identifying independent and linked metabolic pathways and compared to a popular analysis method. Also, this same analysis method is investigated for its ability to identify regulatory linkages within a single metabolic pathway. Lastly, a variant of this same method is used to analyze time series microarray data. In Chapter 2, Factor Analysis is shown to identify and group genes according to membership within independent metabolic pathways for steady state microarray gene expression data. There were cases, however, where the allocation of all genes to a pathway was not complete. A competing analysis method, Hierarchical Clustering, was shown to perform poorly when negatively correlated genes are assumed unrelated, but performance improved when the sign of the correlation coefficient was ignored. In Chapter 3, Factor Analysis is shown to identify regulatory relationships between genes within a single metabolic pathway. These relationships can be explained using metabolic control analysis, along with external knowledge of the pathway structure and activation and inhibition of transcription regulation. In this chapter, it is also shown why factor analysis can group genes by metabolic pathway using metabolic control analysis. In Chapter 4, a Bayesian exploratory factor analysis is developed and used to analyze microarray gene expression data. This Bayesian model differs from a previous implementation in that it is purely exploratory and can be used with vague or uninformative priors. Additionally, 95% highest posterior density regions can be calculated for each factor loading to aid in interpretation of factor loadings. A correlated Bayesian exploratory factor analysis model is also developed in this chapter for application

to time series microarray gene expression data. While this method is appropriate for the analysis of correlated observation vectors, it fails to group genes by metabolic pathway for simulated time series data.

Acknowledgements

I would like to thank Dr. Ina Hoeschele for granting me the opportunity to work in her research group. I have enjoyed this collaboration and I am looking forward to collaborations with Ina in the future.

I would like to thank Drs. Peter von Rohr and Peter Sorensen for many helpful discussions. Dr. Pedro Mendes and Alberto de la Fuente for introducing me to metabolic control analysis and the biochemical simulations that made this research possible.

Lastly, I would like ot thank my wife, April, for all of the unseen support (the housework and numerous readings of papers on who knows what) she provided over these three years in Blacksburg.

Contents

List of Tables

List of Figures

Chapter 1

Literature Review

Introduction

Data from microarray, or gene expression, experiments are quickly becoming available. This data consists of indirect measurements of mRNA abundance for a prespecified set of genes. Generally, research utilizing microarray technology hopes to answer at least one of the following questions:

1. Given two or more experimental conditions, which genes are up or down regulated in one or more conditions?

2. Given a collection of genes and their expression values, either over time or across individuals, which genes are related by membership within a functional unit?

3. Given two or more experimental conditions within a specific tissue, does the tissue have a gene expression signature amenable to predicting tissue classification?

Each of these questions requires different statistical methodology to test the three different types of hypotheses.

Question 1 tries to identify genes involved in alternative responses to external (or possibly internal) stimuli. These genes could be involved in separate pathways, or alternative routes within a pathway. Initially, this type of data was analyzed utilizing a simple *t*-test of the ratios of "treatment" *vs.* "control", where "control" is used as a reference and "treatment" refers to the different conditions. Recently, more statistically rigorous methods such as mixed models and linear regression have been applied to this type of data.

Question 2 seeks to group genes into functional units, where a functional unit can be defined as either a pathway or production scheme (*e.g.* ribosome subunit production). Multivariate statistical analysis methods such as Hierarchical Clustering (Eisen *et al.*, 1998), Self Organizing Maps (Tamayo *et al.*, 1999), and Singular Value Decomposition (Alter *et al.*, 2000) are popular tools to apply to this type of data. Separation and identification of groups has mostly focused on visual tools and data reduction.

Question 3 arises mostly from research in cancer. Here, researchers hope to classify tissue types, specifically tumor types, by their gene expression patterns for a specified set of genes. Bayesian Belief Networks (Friedman *et al.*, 2000) is one method currently being explored for this use.

This dissertation focuses on answering Question 2, specifically identifying genes involved in metabolic pathways and the reconstruction of the regulatory relationships within these pathways using Factor Analysis (Johnson and Wichern, 1998). The performance of Factor Analysis for grouping genes is compared to that of Hierarchical Clustering for steady state data. Additionally, metabolic control analysis is used to explain both why factor analysis can group genes by pathway, and is used to explain the factor loadings. Lastly, a variant of factor analysis designed for time series data is used to identify independent metabolic pathways in time series microarray expression data.

Microarrays

Microarrays are essentially a tool to quantify the the abundance of mRNA expression for several (typically thousands of) genes simultaneously. There are two basic types of microarrays, oligonucleotide (Fodor *et al.*, 1993) and cDNA arrays (Schena *et al.*, 1995), although the principle behind the two is quite similar. The concept is to visually label transcripts (mRNA species), bind these transcripts to known mRNA species at fixed locations on a surface, and quantify the copy number of the transcript.

A. Oligonucleotide Arrays

Affymetrix produces this type of array in their product, "GeneChip". This is essentially a silicon wafer with thousands of genes arrayed in a grid on its surface. The dimensions of each cell in the grid are approximately 20μm by 20μm and the whole grid has the dimensions of 1.28cm by 1.28cm. Probes, or cDNA of known sequence, are "built" one nucleotide at a time using a photolithographic process similar to the process used in the manufacturing of computer chips. In Affymetrix's case, the probes are oligonucleotides of 25 nucleotides in length corresponding to unique sequences of either known genes or expressed sequence tags (EST). Affymetrix uses 16 nonoverlapping oligonucleotides per gene and includes a perfect match (all nucleotides match master sequence) and a mismatch (the $13^{\underline{th}}$ nucleotide is a mismatch to the master sequence) for each oligonucleotide. This technology is single channel, meaning that only a single flourescent dye is needed, and is highly scalable and parallel.

B. cDNA Arrays

This technology is more commonly used in the university setting due to its lower cost. Developed at Stanford University, detailed plans for the construction of equipment and laboratory protocols can still be obtained from their website (http://genome-www.stanford.edu/). cDNA arrays are usually cDNA from ESTs printed or spotted onto a glass slide in a grid array. These spots are fixed to the slide surface and labeled cDNA from whole cell RNA extractions are hybridized to each spot. This is a dual (or more) channel technology meaning that at least two color dyes are used.

C. Quantifying mRNA

Once cDNA to the whole cell RNA extractions have been hybridized onto the array, a digital image is taken of a laser excited microarray. The laser is used to excite specific flourescent dyes, one at a time. The quantity of mRNA expressed is then measured as the intensity of a single channel of colored light measured in pixels from the digital image of the microarray.

Hamadeh and Afshari (2000) provide a nice overview of microarray technology from a molecular viewpoint. While these authors mention statistical analysis, they do not provide a thorough treatment of the subject. Nevertheless, the article is a good overview of the process of obtaining expression data from cell mRNA extractions using chip technology.

Cluster Analysis

Cluster analysis is a class of multivariate statistical methods which group variables according to different distance rules so that variables with a small distance are grouped together. The use of clustering for microarray data was made popular by Eisen *et al.* (1998) who developed software

(Cluster (Eisen, 1999)) exclusively for this purpose. Cluster performs several types of data transformations including natural logarithm, mean and median centering of the data by arrays or genes, three types of hierarchical clustering (average, complete, and single linkage clustering), k-means clustering (MacQueen, 1967), and self organizing maps (SOMs) (Tamayo *et al.*, 1999). Their implementation utilizes the distance measure in equation (1.1) for all correlation based metrics, where r_{ij} is the correlation coefficient between variables i and j,

$$distance_{ij} = 1 - r_{ij} \tag{1.1}$$

although a distance metric using the absolute value of the correlation coefficient is also available.

The node averaging method used in Cluster differs from standard hierarchical clustering according to the following example. Consider grouping a node consisting of two genes with a new gene, creating a new node consisting of all three genes. In standard hierarchical clustering, the new node value is the average of the old average, from two genes, and the expression value of the new gene. In Cluster, the node value is the average expression values of all three genes.

The typical output from a cluster analysis is a tree-like figure called a dendrogram. The distance between the tips of the branches represents the distance between the variables associated with the branch tips.

SOMs is a multivariate data mining tool similar to k-means clustering. For a pre-specified grid (2-D or 3-D) of data collection nodes, the data are iteratively mapped to the nearest node, one data point at a time. The resulting map of the nodes no longer conforms to the grid specified, but has nodes separated by the average distance between data observations within the respective nodes. Some nodes may also be empty at the end of the analysis. Two disadvantages of this method are that the number of nodes must be pre-specified and test statistics cannot be compared across different node topologies.

This section does not contain a listing of all multivariate methods used to analyze microarray data, but only the two most popular. Other methods applied to microarray data are linear models (D'haeseleer, 2000), support vector machines (Brown *et al.*, 2000), and plaid models (Lazzeroni and Owen, 2000), to name a few.

Factor Analysis

Factor Analysis (FA) (Johnson and Wichern, 1998) is a method intended to describe the covariance relationships among many variables in a multivariate data set in terms of a few underlying, but unobservable, random quantities called factors. There are two types of FA models, orthogonal and oblique. This thesis will focus on the orthogonal FA model with the assumption that the factors are orthogonal to, or independent of, each other.

The general factor model is linear in the common factors **f** and is presented in equation (1.2).

$$\mathbf{y}-\mu=\mathbf{L}\mathbf{f}+\varepsilon \tag{1.2}$$

In this model, the vector **y** represents a multivariate observation, μ a vector of means, **L** a matrix of factor loadings, **f** a vector of common factors, and ε a vector of specific, or residual, factors. The vectors **f** and ε are generally not observed and assumed independent. The equations (1.3), (1.4), (1.5), (1.6), (1.7), and (1.8) help further define the orthogonal factor model.

$$E\left[\mathbf{f}\right]=\mathbf{0} \tag{1.3}$$

$$Cov\left(\mathbf{f}\right)=\mathbf{I} \tag{1.4}$$

$$E\left[\varepsilon\right]=\mathbf{0} \tag{1.5}$$

$$Cov\left(\varepsilon\right)=\boldsymbol{\Psi} \tag{1.6}$$

$$Cov\left(\mathbf{y}\right)=\mathbf{L}'\mathbf{L}+\boldsymbol{\Psi} \tag{1.7}$$

$$Cov\left(\mathbf{y},\mathbf{f}'\right)=\mathbf{L} \tag{1.8}$$

The matrix $\boldsymbol{\Psi}$ is assumed a diagonal matrix of specific, or residual, variances (Johnson and Wichern, 1998).

A property of the factor loading matrix is that it can be rotated by an orthogonal matrix without loss of information (*i.e.* the covariances or correlations between variables are maintained following rotation). The portion of the variance explained by the retained factors for a specific variable is termed the communality and the portion not explained by the retained factors is termed the specific variance. Below is a brief discussion of four common methods and one recently developed method of obtaining factor loadings.

A. Principal Components

This method is by far the simplest of all five methods presented. Principal Components FA (PCFA) proceeds by performing an eigenvalue decomposition of either the (co)variance matrix or correlation matrix of the observed data (gene expression values). This decomposes either the (co)variance matrix or correlation matrix into two new matrices, one containing the eigenvalues along the diagonal and the other the associated eigenvectors. The number of eigenvalues and eigenvectors retained determines the number of factors retained. The factor loading matrix is obtained by multiplying each of the retained eigenvectors by the square root of its associated eigenvalue.

B. Principal Factor

Principal Factor FA (PFFA) is a variant of PCFA. Since the specific variances are invariant to rotation by an orthogonal matrix, specific variance *ii* is estimated first by one of the following two methods:

Correlation matrix not singular: one over diagonal element *ii* of the inverse of the correlation matrix.

Correlation matrix singular: one over the absolute value of the largest correlation in row *i* of the correlation matrix.

A new "correlation" matrix is then formed by subtracting the specific variances from the data correlation matrix, *e.g.* $Cor^*(\mathbf{Y}) = Cor(\mathbf{Y}) - \mathbf{\Psi}$. The factor loadings are then computed as in PCFA, but using $Cor^*(\mathbf{Y})$ instead of $Cor(\mathbf{Y})$.

C. Iterated Principal Factor

This method is identical to PFFA, except the factor loadings $\mathbf{L}$ and specific variances $\mathbf{\Psi}$ are iteratively solved for. The procedure is:

1. Perform initial PFFA.

2. Perform PFFA with new estimate of $\boldsymbol{\Psi}$.

3. Calculate the convergence criteria, *e.g.* sum of squared differences between previous round and current round estimates of the factor loadings.

4. Iterate between 2 and 3 until the criterion in 3 becomes small.

This method can produce negative estimates of the specific variances called Heywood cases (Johnson and Wichern, 1998). The usual procedure for correcting the negative estimates is to set them to zero for the next round of iteration.

D. Maximum Likelihood

Maximum Likelihood FA (MLFA) estimates of $\mathbf{L}$ and $\boldsymbol{\Psi}$ are obtained from the likelihood in (1.9) given that $\Sigma = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$, where Σ is $Cov(\mathbf{y})$.

$$L(\mu, \Sigma) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(\mathbf{y}-\bar{\mathbf{y}})(\mathbf{y}-\bar{\mathbf{y}})'+n(\bar{\mathbf{y}}-\mu)(\bar{\mathbf{y}}-\mu)'\right)\right]} \tag{1.9}$$

The likelihood is degenerate for $\mathbf{L}$, so a *uniqueness condition* is imposed, namely $\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L} = \Delta$ and Δ a diagonal matrix. Rubin and Thayer (1982) developed an EM algorithm for MLFA which relies on unique starting values (such as from a PCFA), rather than the aforementioned uniqueness condition to insure identifiability of the factor loadings. The algorithm consisted of the following steps:

1. Assuming the data have been centered about the population mean, compute:

$$\mathbf{C}_{yy} = \sum_{i=1}^{n} \frac{\mathbf{y}_i\mathbf{y}_i'}{n}, \text{ a } p \times p \text{ matrix} \tag{1.10}$$

$$\mathbf{C}_{yf} = \sum_{i=1}^{n} \frac{\mathbf{y}_i\mathbf{f}_i'}{n}, \text{ a } p \times m \text{ matrix} \tag{1.11}$$

$$\mathbf{C}_{ff} = \sum_{i=1}^{n} \frac{\mathbf{f}_i\mathbf{f}_i'}{n}, \text{ a } m \times m \text{ matrix} \tag{1.12}$$

2. Compute:

$$\delta = \left(\boldsymbol{\Psi} + \mathbf{L}\mathbf{L}'\right)^{-1}\mathbf{L} \text{ and} \tag{1.13}$$

$$\Delta = \mathbf{I}_m - \mathbf{L}' \left(\Psi + \mathbf{L}\mathbf{L}'\right)^{-1} \mathbf{L} \text{ for prior starting values of } \mathbf{L} \text{ and } \Psi \tag{1.14}$$

3. Then compute the following conditional expectations:

$$E\left(\mathbf{C}_{yy}|\mathbf{Y}\right) = \mathbf{C}_{yy} \tag{1.15}$$

$$E\left(\mathbf{C}_{yf}|\mathbf{Y}\right) = \mathbf{C}_{yy}\delta \tag{1.16}$$

$$E\left(\mathbf{C}_{ff}|\mathbf{Y}\right) = \delta'\mathbf{C}_{yy}\delta + \Delta \tag{1.17}$$

4. Then compute:

$$\mathbf{L} = \left(\mathbf{C}_{yy}\delta\right)\left(\delta'\mathbf{C}_{yy}\delta + \Delta\right)^{-1} \tag{1.18}$$

$$\Psi = \operatorname{diag}\left(\mathbf{C}_{yy} - \mathbf{C}_{yy}\delta\left(\delta'\mathbf{C}_{yy}\delta + \Delta\right)^{-1}\delta'\mathbf{C}_{yy}\right) \tag{1.19}$$

5. Calculate the convergence criteria, *e.g.* sum of squared differences between previous round and current round estimates of the factor loadings.

6. Iterate between items 1 to 5 until the criterion in item 5 is small

The steps from 1 to 3 are called the E step, or Expectation step, and the steps from 4 to 5 are called the M step, or maximization step.

This method can also produce Heywood cases and the same methods to combat the negative specific variance estimates in iterated PFFA are also used here. The search for a method that would remain within the parameter space for the specific variances partly led to the next method.

E. Confirmatory Bayesian Factor Analysis - Press and Shigemasu Model

Assuming that the data are normally distributed and are mean centered, *i.e.* $E\left[\mathbf{y}_i\right] = \mathbf{0}$, the conditional density of the data $\mathbf{Y}$ given $\mathbf{L}$, $\mathbf{F}$, and Ψ is proportional to (1.20) (Press and Shigemasu, 1997),

$$p\left(\mathbf{Y}|\mathbf{L},\mathbf{F},\Psi\right) \propto |\Psi|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\left(\Psi^{-1}(\mathbf{Y}-\mathbf{F}\mathbf{L}')'(\mathbf{Y}-\mathbf{F}\mathbf{L}')\right)} \tag{1.20}$$

where $\mathbf{Y}$ is a $n \times p$ matrix of n multivariate observation vectors $\mathbf{y}_i$ and $\mathbf{F}$ is a $n \times m$ matrix of n factor score vectors $\mathbf{f}_i$. From Press and Shigemasu (1997), the joint prior distribution for $\mathbf{L}$, $\mathbf{F}$, and

$\mathbf{\Psi}$ is:

$$p(\mathbf{L},\mathbf{F},\mathbf{\Psi}) \propto p(\mathbf{L}|\mathbf{\Psi})\, p(\mathbf{\Psi})\, p(\mathbf{F}) \tag{1.21}$$

where

$$p(\mathbf{L}|\mathbf{\Psi}) \propto |\mathbf{\Psi}|^{-\frac{m}{2}}\, e^{-\frac{1}{2}tr\left(\mathbf{\Psi}^{-1}(\mathbf{L}-\mathbf{L_o})\mathbf{H}(\mathbf{L}-\mathbf{L_o})'\right)} \tag{1.22}$$

$$p(\mathbf{\Psi}) \propto |\mathbf{\Psi}|^{-\frac{\nu}{2}}\, e^{-\frac{1}{2}tr\left(\mathbf{\Psi}^{-1}\mathbf{B}\right)} \tag{1.23}$$

$$p(\mathbf{F}) \propto e^{-\frac{1}{2}tr(\mathbf{F}'\mathbf{F})} \tag{1.24}$$

The diagonal prior precision matrix $\mathbf{H}$ and (co)variance matrix $\mathbf{B}$ are both positive definite matrices as is the matrix of specific variances $\mathbf{\Psi}$. The prior distribution for the (co)variance matrix $\mathbf{\Psi}$ is Inverse Wishart with parameters ν and $\mathbf{B}$, $p(\mathbf{\Psi}) \sim \mathrm{IW}(\nu,\mathbf{B})$ and $\mathbf{\Psi}$ is assumed diagonal on average. The conditional prior distribution for the matrix of factor loadings $\mathbf{L}$ is multivariate normal with parameter vector $\mathbf{l}_o$ and $m\times m$ matrix $\mathbf{H}$, $p(\mathbf{l}|\mathbf{\Psi}) \sim \mathrm{MVN}\left(\mathbf{l}_o, \mathbf{H}^{-1}\otimes\mathbf{\Psi}\right)$ where $\mathbf{l}$ is $vec(\mathbf{L})$ and $\otimes$ denotes the Kroneckor product. The prior distribution for the matrix of factor scores $\mathbf{F}$ is equivalent to that of the product of standard normal densities for each factor score vector $\mathbf{f}_i$. This gives the joint posterior in (1.25) and the conditional sampling distributions in (1.26), (1.27), and (1.28)

$$p(\mathbf{L},\mathbf{F},\mathbf{\Psi}|\mathbf{Y}) \propto e^{-\frac{1}{2}tr\mathbf{F}'\mathbf{F}}\, |\mathbf{\Psi}|^{-\frac{n+m+\nu}{2}}\, e^{-\frac{1}{2}tr\left(\mathbf{\Psi}^{-1}\left[(\mathbf{Y}-\mathbf{F}\mathbf{L}')'(\mathbf{Y}-\mathbf{F}\mathbf{L}')+(\mathbf{L}-\mathbf{L_o})\mathbf{H}(\mathbf{L}-\mathbf{L_o})'+\mathbf{B}\right]\right)} \tag{1.25}$$

$$p(\mathbf{L}|\mathbf{F},\mathbf{\Psi},\mathbf{Y}) \propto e^{-\frac{1}{2}tr\left(\mathbf{\Psi}^{-1}\left(\mathbf{L}-\tilde{\mathbf{L}}\right)(\mathbf{H}+\mathbf{F}'\mathbf{F})\left(\mathbf{L}-\tilde{\mathbf{L}}\right)'\right)} \tag{1.26}$$

$$p(\mathbf{\Psi}|\mathbf{L},\mathbf{F},\mathbf{Y}) \propto |\mathbf{\Psi}|^{-\frac{n+m+\nu}{2}}\, e^{-\frac{1}{2}tr\left(\mathbf{\Psi}^{-1}\left[(\mathbf{Y}-\mathbf{F}\mathbf{L}')'(\mathbf{Y}-\mathbf{F}\mathbf{L}')+(\mathbf{L}-\mathbf{L_o})\mathbf{H}(\mathbf{L}-\mathbf{L_o})'+\mathbf{B}\right]\right)} \tag{1.27}$$

$$p(\mathbf{F}|\mathbf{L},\mathbf{\Psi},\mathbf{Y}) \propto e^{-\frac{1}{2}tr\left(\left(\mathbf{F}-\tilde{\mathbf{F}}\right)\left(\mathbf{I}_m+\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)\left(\mathbf{F}-\tilde{\mathbf{F}}\right)'\right)} \tag{1.28}$$

where

$$\tilde{\mathbf{L}} = \left(\mathbf{Y}'\mathbf{F}+\mathbf{L}_o\mathbf{H}\right)\left(\mathbf{H}+\mathbf{F}'\mathbf{F}\right)^{-1}$$
$$\tilde{\mathbf{F}} = \mathbf{Y}\mathbf{\Psi}^{-1}\mathbf{L}\left(\mathbf{I}_m+\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1}$$

This model was developed for confirmatory factor analysis (Rowe and Press, 1998), but it can be extended to exploratory factor analysis. This requires using vague prior distributions for the factor loadings and specific variances, rather than the sharp priors in Press and Shigemasu (1997). It can be shown that convergence of the sampler in Press and Shigemasu (1997) occurs from specifying sufficient nonzero factor loadings to make the factor loading matrix identifiable (Congdon, 2001) along with small prior variances and in the absence of these conditions, the sampler fails to converge. It should be noted that using the priors specified in Press and Shigemasu (1997), the posterior mean is completely specified through the prior, regardless of the information in the data.

F. Factor Rotation

Notice that in there are an infinite number of solutions using the likelihood in (1.9), each of which is related to the other by an orthogonal rotation matrix, say $\mathbf{T}$. This can be shown as $\boldsymbol{\Sigma} = \mathbf{L}^*\mathbf{L}^{*\prime} + \boldsymbol{\Psi} = \mathbf{L}\mathbf{T}'\mathbf{T}\mathbf{L}' + \boldsymbol{\Psi} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$ (Johnson and Wichern, 1998). Factor rotation is usually performed to clarify the interpretation of the factor loadings. The need for clarification comes from the orientation of the orthogonal factor loadings on the m-dimensional grid axes. Consider the case of two retained factors. In a two-dimensional plot of the retained factors, the factors will be orthogonal to each other but some angle of reflection away from the x and y grid axes, much like plots of principal component scores. Factor rotation is often performed subject to an objective criterion, with the Varimax procedure (Kaiser, 1958) the most popular one. This thesis will use the Minimum Entropy (Inagaki, 1994) factor rotation criterion, which follows from the assumption that the factor loadings within a given factor should be either near zero or near plus or minus one.

G. Factor Alignment

The alignment by Clarkson algorithm (Clarkson, 1979) first re-orders the factors in a sample factor loading matrix to that of the estimated factor loading matrix. The signs of the factor loading entries are aligned by multiplying the transpose of the sample factor loading matrix with the estimated factor loading matrix. The sign of the diagonal entry of the product of this multiplication is then used to adjust the signs of the factor loading entries in the sample matrix. This algorithm is often used when performing the bootstrap or jackknife. This thesis also applies this algorithm to samples from a Bayesian posterior distribution.

H. Bootstrap Confidence Intervals

Bootstrap confidence intervals for factor loadings can be obtained using the bias corrected and accelerated (BC_a) method (Efron and Tibshirani, 1993). This method differs from that presented in Ichikawa and Konishi (1995) in that it uses adjusted percentiles of the bootstrap distribution of factor loadings to find the upper and lower bounds

I. Number of Factors Retained

A ML hypothesis test of adequately approximating the sample covariance or correlation matrix with m retained factors can be performed using a likelihood ratio (LR) test of H_o: $\hat{\mathbf{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}}$ *vs.* H_1: $\hat{\mathbf{\Sigma}} = \mathbf{S}_n$ where $\mathbf{S}_n$ is the estimated (co)variance matrix of the $\mathbf{X}$'s. The LR test statistic is distributed as $\chi^2_{\frac{1}{2}\left[(p-m)^2-p-m\right]}$, where n is the number of observations, p the number of variables, and m the number of retained factors with $m \leq p \leq n$. Even with this test, the number of factors retained may still be subjective. Suppose that n is large and m is small relative to p. Even if $\hat{\mathbf{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}}$ is a close approximation to $\hat{\mathbf{\Sigma}} = \mathbf{S}_n$, the hypothesis H_o will usually be rejected.

Another, more subjective, criteria is the "Scree" plot. This is a plot with the eigenvalue on the y axis and the eigenvalue numbers (ordered from largest to smallest) on the x axis. The number of factors to retain is chosen as the eigenvalue number where a sudden drop in the line connecting the eigenvalues occurs. The term "Scree" comes from a mining term defined as a cliff with rubble at the bottom.

A more objective criterion for selecting the number of factors to retain is the "Broken Stick" criterion (Jackson, 1993). This criterion is based upon the heuristic of a stick whose length equals the number of variables p in the analysis. A series of p critical values is then computed with the $i^{\underline{\text{th}}}$ critical value equal to $\sum_{j=i}^{p} \frac{1}{j}$. The number of the last eigenvalue to exceed its critical value equals the number of factors to retain.

In the Bayesian Factor Analysis model, Rowe (1998) proposed placing a discrete prior distribution on m, the number of factors retained in the model. In Press and Shigemasu (1997), it was suggested to fit several models including different numbers of factors retained and record the Akaike's Information Criterion (AIC) for each model. The optimal number of factors retained was then inferred from the model with the minimum AIC. The method proposed in Rowe (1998) for the number of factors retained was marginally different. A prior distribution such as a uniform or Poisson was placed on the number of factors and a marginal posterior probability for the number of factors was recorded for models containing different numbers of factors. The model with the highest marginal posterior probability for the number of factors retained was used. A better method would be to use Reversible Jump Metropolis-Hastings (Green, 1995), moving the chain from a model containing n number of factors to a model with $n \pm t$ where $n > 0$ and $t \geq 0$.

Metabolic Control Analysis

In metabolic control analysis (Kacser and Burns, 1973; Heinrich and Rapoport, 1974), the relationships between the variables in a system, or network, can be described by the elasticities and control coefficients of the system (Hofmeyr *et al.*, 1993). The variables of a system consist of the reaction rates and substrate concentrations of the system constituents. The terms flux, elasticity, control coefficient, and response coefficient are defined below.

Flux The steady state reaction rate. A flux J may be defined in terms of a reaction step, *e.g.* $v_i = J_i$ where $i = 1 \ldots n$; a segment, *e.g.* $v_i = v_j = J_a$; or in terms of the entire pathway, *e.g.* $J_A + J_B + \cdots + J_I = J_Z$ where J_Z is the flux through the stem and fluxes J_A through J_I are the fluxes through each branch in the system.

Elasticity The relative change in rate of a step in a system caused by a perturbation of a single metabolite at a constant concentration of all other metabolites. Formally for a substrate s_j and local rate v_i: $\varepsilon_{s_j}^{v_i} = \frac{\partial ln\ v_i'}{\partial ln\ s_j}$.

Control Coefficient The relative change in a system variable caused by a specific modulation of an enzyme at steady state. Formally for a system variable x and a rate variable v_i: $C_{v_i}^{x} = \frac{\partial ln\ |x|}{\partial ln\ v_i}$.

Response Coefficient The response in system variable y_k to a change in the system variable x_i associated with step e_j at steady state. Formally: ${}^{e_j}R_{x_i}^{y_k} = \varepsilon_{x_i}^{e_j} C_{e_j}^{y_k}$. y_k can be either a substrate or a flux.

The relationship between the $\varepsilon_{s_j}^{v_i}$ and $C_{v_i}^{x}$ terms and their ability to describe a metabolic system can be seen in the matrix formulation of Hofmeyr and Cornish-Bowden (1996), $\mathbf{CE} = \mathbf{I}$. The expanded

$$\mathbf{CE} = \begin{bmatrix} C_1^{J_C} & C_2^{J_C} & C_3^{J_C} & C_4^{J_C} & C_5^{J_C} & C_6^{J_C} \\ -C_1^{S_1} & -C_2^{S_1} & -C_3^{S_1} & -C_4^{S_1} & -C_5^{S_1} & -C_6^{S_1} \\ -C_1^{S_3} & -C_2^{S_3} & -C_3^{S_3} & -C_4^{S_3} & -C_5^{S_3} & -C_6^{S_3} \\ -C_1^{j} & -C_2^{j} & -C_3^{j} & -C_4^{j} & -C_5^{j} & -C_6^{j} \\ -C_1^{S_4} & -C_2^{S_4} & -C_3^{S_4} & -C_4^{S_4} & -C_5^{S_4} & -C_6^{S_4} \\ -C_1^{S_5} & -C_2^{S_5} & -C_3^{S_5} & -C_4^{S_5} & -C_5^{S_5} & -C_6^{S_5} \end{bmatrix} \begin{bmatrix} 1 & -\varepsilon_1^1 & 0 & j & 0 & 0 \\ 1 & \varepsilon_1^2 & 0 & j & -\varepsilon_4^2 & 0 \\ 1 & 0 & -\varepsilon_3^3 & j-1 & 0 & 0 \\ 1 & 0 & \varepsilon_3^4 & j-1 & -\varepsilon_4^4 & 0 \\ 1 & 0 & 0 & 0 & \varepsilon_4^5 & -\varepsilon_5^5 \\ 1 & 0 & 0 & 0 & 0 & \varepsilon_5^6 \end{bmatrix}$$

$$= \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad [1.29]$$

form can be seen in equation (1.29) for the structure pictured in Figure 1.1 as parameterized by Sen (1996). Here $j = J_A/J_C$ and $J_A = J_1 = J_2$ represent the flux through the branch containing S_0 and $J_C = J_5 = J_6$ represents the flux through the stem. There are a total of three fluxes through the system in Figure 1.1: $J_A + J_B$, $J_A + J_C$, and $J_B + J_C$, where $J_B = J_3 = J_4$ is the flux through the branch containing S_2. In this structure $J_B/J_C = j - 1$. This set of equations demonstrates two important theorems in metabolic control analysis, the flux control and the connectivity flux control summation theorems.

Flux Control Summation Theorem $\sum_{i=1}^{n} C_i^{J_C} = 1$.

Flux Connectivity Summation Theorem $\sum_{i=1}^{n} \varepsilon_X^i C_i^{J_C} = \sum_{i=1}^{n} {}^iR_X^{J_C} = 0$.

These theorems state that the total flux control is complete and that the flux control is distributed by the elasticity coefficients.

C is the inverse of **E**, and **E** the inverse of **C**, thus both **C** and **E** are invertible. It should be noted, however, that there are only four internal metabolites in the system and the flux through the system is a function of the changes in the concentration of these metabolites. In the structure depicted in Figure 1.1, S_0, S_2, and S_7 are parameters of the system. S_0 and S_2 are initially determined by concentrations external to the system and the value of S_7 is buffered at a constant value. The internal metabolites can be seen in the stoichiometric matrix **N** in equation (1.30).

$$\mathbf{N} \times \mathbf{J} = \mathbf{0}$$
$$= \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} J_1 \\ J_2 \\ J_3 \\ J_4 \\ J_5 \\ J_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad [1.30]$$

This equation shows the dependencies of the fluxes of the system:

1. $J_1 = J_2$
2. $J_5 = J_2 + J_4$
3. $J_3 = J_4$
4. $J_5 = J_6$

Inspection of the stoichiometric matrix **N** shows two columns that are not pivots, columns 4 and 6. Using the parameterization of Hofmeyr and Cornish-Bowden (1996), the matrix **K** can be defined to relate the dependent fluxes J_1, J_2, J_3, and J_5 to the independent fluxes J_4 and J_6 (1.31) ($\mathbf{J}^I$ is the

$$\mathbf{C}^I\mathbf{E} = \begin{bmatrix} \mathbf{C^{J_I}} \\ \mathbf{C^{S_I}} \end{bmatrix} \begin{bmatrix} \mathbf{K} & -\varepsilon \end{bmatrix} = \begin{bmatrix} \mathbf{I_{n-r}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I_r} \end{bmatrix}$$

$$= \begin{bmatrix} C^{J_4}_4 & C^{J_4}_6 & C^{J_4}_1 & C^{J_4}_2 & C^{J_4}_3 & C^{J_4}_5 \\ C^{J_6}_4 & C^{J_6}_6 & C^{J_6}_1 & C^{J_6}_2 & C^{J_6}_3 & C^{J_6}_5 \\ -C^{S_1}_4 & -C^{S_1}_6 & -C^{S_1}_1 & -C^{S_1}_2 & -C^{S_1}_3 & -C^{S_1}_5 \\ -C^{S_3}_4 & -C^{S_3}_6 & -C^{S_3}_1 & -C^{S_3}_2 & -C^{S_3}_3 & -C^{S_3}_5 \\ -C^{S_4}_4 & -C^{S_4}_6 & -C^{S_4}_1 & -C^{S_4}_2 & -C^{S_4}_3 & -C^{S_4}_5 \\ -C^{S_5}_4 & -C^{S_5}_6 & -C^{S_5}_1 & -C^{S_5}_2 & -C^{S_5}_3 & -C^{S_5}_5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \varepsilon^4_3 & \varepsilon^4_4 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & -\varepsilon^1_1 & 0 & 0 & 0 \\ -1 & 1 & \varepsilon^2_1 & 0 & \varepsilon^2_4 & 0 \\ 1 & 0 & 0 & -\varepsilon^3_3 & 0 & -\varepsilon^5_5 \\ 0 & 1 & 0 & 0 & -\varepsilon^5_4 & \varepsilon^6_5 \end{bmatrix}$$

$$= \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad [1.32]$$

matrix of independent fluxes).

$$\mathbf{K} \times \mathbf{J}^I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} J_4 \\ J_6 \end{bmatrix}$$

$$= \begin{bmatrix} J_4 \\ J_6 \\ J_6 - J_4 \\ J_6 - J_4 \\ J_4 \\ J_6 \end{bmatrix} = \begin{bmatrix} J_4 \\ J_6 \\ J_1 \\ J_2 \\ J_3 \\ J_5 \end{bmatrix} \qquad [1.31]$$

This produces another formulation of the $\mathbf{CE} = \mathbf{I}$, namely that seen in equation (1.32). There are many other possible parameterizations of the flux through the system in Figure 1.1. For each steady state reached, the values of the fluxes through the branches and stems are different, but the ratios of these fluxes are unique.

Metabolic control analysis contains the equations and theorems providing a basis for our simulated microarray gene expression data. As the above equations describe the interlaced relationships between the enzymes and metabolites within a simple pathway, hierarchical metabolic control analysis (Hofmeyr and Westerhoff, 2001) can be used for the more complicated simulations used in this thesis involving interconnected layers of pathways. To simulate the data used in this thesis, the biochemical simulation program Gepasi was used (Mendes, 1993; Mendes, 1997; Mendes and Kell, 1998; Mendes, 2000). With this program, concentrations of mRNA, enzyme, and metabolites can be obtained for hierarchical pathways, along with many other parameters of the system including reaction rates and kinetic functions.

Additionally, in Chapter 3, metabolic control analysis is used to explain both why factor analysis can group genes by metabolic pathway and the factor loadings themselves. Use is made of both the Flux Control Summation Theorem and the Flux Connectivity Summation Theorem to propose reduced, but equivalent, response matrices that mimic the observed factor loading matrices. It is also shown that for equivalent pathway structures, differences in regulation of transcription produce different response matrices. Therefore, knowledge of both the pathway structure and the regulatory links between the genes is needed to use metabolic control analysis as an aid in interpreting factor loadings for a factor analysis of gene expression data.

References

Alter O, Brown PO & Botstein D (2000) Singular value decomposition for genome wide expression data processing and modeling. *Proc. Nat. Acad. Sci.* 97(18), 10101–10106

Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, Jr M & Haussler D (2000) Knowledge-based analysis of microarray gene expression data by using suport vector machines. *Proc. Nat. Acad. Sci.* 97, 262–267

Chen MH & Shao QM (1999) Monte -Carlo estimation of Bayesian credible and HPD intervals. *J. Comp. Graph. Stat.* 8(1), 69–92

Clarkson DB (1979) Estimating the standard errors of rotated factor loadings by jackknifing. *Psychometrika* 44(3), 297–314

Congdon P (2001) *Bayesian statistical modelling*. Wiley, West Sussex, England

D'haeseleer P (2000) *Reconstructing gene networks from large scale gene expression data*. PhD thesis, University of New Mexico

Efron B & Tibshirani RJ (1993) *An introduction to the bootstrap*. Chapman and Hall, New York, USA

Eisen MB (1999) *Cluster and TreeView Manual*. Stanford University, Palo Alto, CA, USA, 1st edition

Eisen MB, Spellman P, Brown PO & Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci.* 95, 14863–14868

Fodor SPA, Rava RP, Huang XC, Pease AC, Holmes CP & Adams CL (1993) Multiplexed biochemical assays with biological chips. *Nature* 364, 555–556

Friedman N, Linnial M, Natchman I & Pe'er D (2000) Using bayesian networks to analyze expression data. *J. Comp. Biol.* 7, 601–620

Green PJ (1995) Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–732

Hamadeh H & Afshari CA (2000) Gene chips and functional genomics. *Amer. Scientist* 88, 508–515

Heinrich R & Rapoport TA (1974) A linear steady state treatment of enzyme chains: General properties, control, and effector strength. *Eur. J. Biochem.* 42, 89–95

Hofmeyr JHS & Cornish-Bowden A (1996) Co-response analysis: A new experimental strategy for metabolic control analysis. *J. Theor. Biol.* 182, 371–380

Hofmeyr JHS & Westerhoff HV (2001) Building the cellular puzzle. control in multi-level reaction networks. *J. Theor. Biol.* 208, 261–285

Hofmeyr JS, Cornish-Bowden A & Rohwer JM (1993) Taking enzyme kinetics out of control; putting control into regulation. *Eur. J. Biochem.* 212, 833–837

Holter NS, Maritan A, Cieplak M, Fedoroff NV & Banavar JR (2001) Dynamic modeling of gene expression data. *Proc. Nat. Acad. Sci.* 98(4), 1693–1698

Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR & Fedoroff NV (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Nat. Acad. Sci.* 97(15), 8409–8414

Ichikawa M & Konishi S (1995) Application of the bootstrap methods in factor analysis. *Psychometrika* 60(1), 77–93

Inagaki A (1994) Entromin and Entromax: Criteria for factor rotation based on entropy. *Jap. J. of Behav.* 21, 10–20

Jackson DA (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74, 2204–2214

Johnson RA & Wichern DW (1998) *Applied multivariate statistical analysis*. Prentice Hall, Englewood Cliffs, NJ, USA, 4th edition

Kacser H & Burns JA (1973) The control of flux. *Symp. Soc. Exp. Biol.* 32, 65–104

Kaiser HF (1958) The Varimax criterion for analytic rotation in Factor Analysis. *Psychometrika* 23, 187–200

Lazzeroni L & Owen A (2000) Plaid models for gene expression data. Technical Report 211, Stanford University, Palo Alto, CA, USA

MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berk. Symp. Math. Stat. Prob.*, volume 1, pages 281–297, Berkely, CA, USA. Univ. CA Press

Mendes P (1993) Gepasi: a software package for modeling the dynamics, steady states, and control of biochemical and other systems. *Comp. Appl. Biosci.* 9, 563–571

Mendes P (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* 22, 361–363

Mendes P (2000) Gepasi: Numerical simulation and optimization of biochemical kinetics. In *Proc. Plant and Anim. Genome VIII*, San Diego, CA, USA

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

Peterson LE (2001) Factor analysis of cluster-specific gene expression levels from cDNA microarrays. *Comp. Meth. and Prog. in Biosci.* in press, –

Press S & Shigemasu K (1997) Bayesian inference in factor analysis - Revised. Technical Report No. 243, Department of Statistics, University of California, Riverside, California

Rowe DB (1998) *Correlated Bayesian factor analysis*. PhD thesis, Department of Statistics, University of California, Riverside

Rowe DB & Press SJ (1998) Gibbs sampling and hill climbing in Bayesian factor analysis. Technical Report No. 255, Department of Statistics, University of California, Riverside

Rubin DB & Thayer DT (1982) EM algorithms for ML factor analysis. *Psychometrika* 47(1), 69–76

Schena M, Shalon D, Davis RW & Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470

Sen AK (1996) On the sign pattern of metabolic control coefficients. *J. Theor. Biol.* 182, 269–275

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES & Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci.* 96, 2907–2912

Wade KM & Quaas RL (1993) Solutions to a system of equations involving a first-order autoregressive process. *J. Dairy Sci.* 76(10), 3026–3032

West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Marks JR & Nevins JR (2001) Predicting the clinical status of human breast cancer using gene expression profiles. *Proc. Nat. Acad. Sci.* 98, 11462–11467

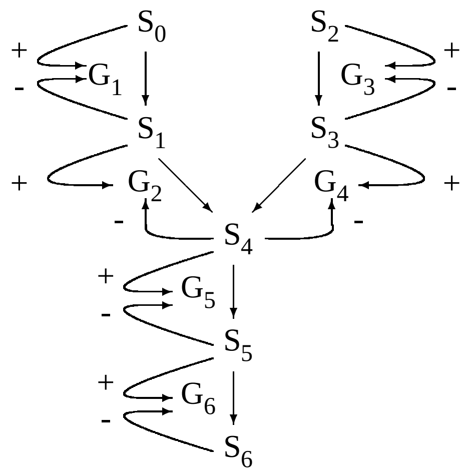

Figure 1.1. Six gened simulated pathway structure with regulatory interactions.

Chapter 2

Factor analysis for the identification of metabolic pathways from microarray expression data

Abstract

Motivation: The identification of genes belonging to different metabolic pathways could provide the input needed for causal or regulatory models and could also be used to identify candidate genes for further research within a wet laboratory setting. Results of these latter experiments could then be used to identify candidate genes for selection in animal or plant genetic improvement programs or targets for gene therapy. Previously, hierarchical clustering has been used to analyze experimental data, but its performance has only been subjectively evaluated. Factor analysis is an exploratory statistical method for multivariate data based on the assumption that the observed data are produced by a few unobserved factors. It is our assumption that these unobserved factors are the metabolic pathways of which the measured genes are constituents.

Results: Investigation of factor analysis with simulated microarray gene expression data representing sets of genes corresponding to different metabolic pathways has proven the conjecture that retained factors represent metabolic pathways and that genes can be grouped by pathway. Bootstrap confidence intervals for factor loadings can aid in interpretation of the factor loadings. Hierarchical clustering performed poorly when the signed Pearson correlation coefficient was used as a distance measure. Performance improved when the absolute value of the Pearson correlation coefficient was used as a distance measure, but additional analysis tools may be needed to infer groupings. Application of analysis tools developed for microarray expression data to simulated data leads to a better understanding of the utility of these methods as clearly demonstrated in this contribution.

Introduction

Increasingly, data on the relative expression level of thousands of genes are available for analysis from microarray gene expression experiments. The grouping, or clustering, of these genes into functional units is of particular interest, although the term functional unit is often rather loosely defined. Once identified, these groupings could then be used to determine potential targets for gene or drug therapy, or to determine genes for selection within animal or plant genetic improvement programs. To date several multivariate statistical methods have been used to group genes into functional units including hierarchical clustering (HCA) (Eisen *et al.*, 1998), self organizing maps (SOM) (Tamayo *et al.*, 1999), Bayesian belief networks (Friedman *et al.*, 2000), and singular value decomposition (Alter *et al.*, 2000; Holter *et al.*, 2000; Holter *et al.*, 2001; West *et al.*, 2001). Most of these methods have been applied to data from actual microarray experiments.

Here we use factor analysis to group genes into (independent) metabolic pathways. We also compare factor analysis with selected clustering methods in terms of their abilities to correctly group genes by metabolic pathway. The data used in this study are simulated expression data, for which the underlying metabolic pathway structure is known. Factor analysis is a statistical method often used to perform data reduction and explore multivariate data. Factor analysis views the observed multivariate data as produced by a few unobserved factors. Here we show that the first m factors represent m exclusive sets of genes corresponding to m underlying pathways.

Materials and Methods

A. Factor Analysis

Factor analysis (FA) (Johnson and Wichern, 1998) is a method to describe the covariance, or correlation, relationships among many variables in a multivariate data set in terms of a few underlying, but unobservable, random quantities called factors. Inferences about the relationships between the observed variables are made through the magnitude, and typically the sign, of the cell entries in the factor loading matrix. Given a multivariate data observation $\mathbf{y}_i$ and an unobserved factor vector $\mathbf{f}_i$, the definition of the factor loading matrix is either $Cov\left(\mathbf{y}_i, \mathbf{f}'_i\right)$ or $Cor\left(\mathbf{y}_i, \mathbf{f}'_i\right)$, depending on whether or not the data have been standardized. The number of factors m is specified based upon either subjective (*e.g.* a "Scree" plot (Johnson and Wichern, 1998)) or objective (*e.g.* "Broken Stick" (Jackson, 1993)) criteria and $m \leq p$ with p the number of variables. The portion of the total variance explained by the retained factors for a specific variable is termed the communality, and the portion not explained by the retained factors is termed the specific variance.

Equation (2.1) shows the general factor model which is linear in the common factors $\mathbf{f}_i$, or

$$\mathbf{y}_i - \mu = \mathbf{L}\mathbf{f}_\mathbf{i} + \varepsilon \tag{2.1}$$

In this model, $\mathbf{y}_i$ represents a $p \times 1$ multivariate observation vector on one experimental unit, μ a $p \times 1$ vector of means, $\mathbf{L}$ a $p \times m$ matrix of factor loadings, $\mathbf{f}_i$ a $m \times 1$ vector of the effects of the $i^{\underline{\text{th}}}$ experimental unit common factors, and ε_i a $p \times 1$ vector of specific, or residual, factors. The vectors $\mathbf{f}_i$ and ε_i are generally not observed and assumed independent. For application to gene expression data, $\mathbf{y}_i$ is the vector of expression values for p genes in sample i ($i = 1, \ldots, n$). The covariance matrix of $\mathbf{y}_i$ is then defined in equation (2.2).

$$Cov\left(\mathbf{y}_i\right) = \mathbf{L}\mathbf{L}' + \Psi \tag{2.2}$$

Here, the $p \times p$ matrix Ψ is a diagonal matrix of specific, or residual, variances (Johnson and Wichern, 1998). The covariance matrix of the vector $\mathbf{y}_i$ is invariant to rotation of the factor loadings, $\mathbf{L}$, and the factor scores, $\mathbf{f}_i$, by an orthogonal matrix. This causes a problem with the identifiability of $\mathbf{L}$ and $\mathbf{f}_i$, *i.e.* an infinite number of solutions for $\mathbf{L}$ and $\mathbf{f}$ are possible, each related to the other by an orthogonal rotation matrix.
Maximum Likelihood (ML) estimates of $\mathbf{L}$ and Ψ can be obtained from the likelihood in (2.3) given that $\Sigma = \mathbf{L}\mathbf{L}' + \Psi$, where Σ is either $Cov\left(\mathbf{y}_i\right)$ or $Cor\left(\mathbf{y}_i\right)$.

$$L\left(\mu, \Sigma\right) = (2\pi)^{-\frac{np}{2}} \left|\Sigma\right|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)\left(\mathbf{y}_i - \bar{\mathbf{x}}\right)' + n(\bar{\mathbf{y}} - \mu)(\bar{\mathbf{y}} - \mu)'\right)\right]} \tag{2.3}$$

Since there are an infinite number of solutions for $\mathbf{L}$ and $\mathbf{f}$, a *uniqueness condition*, $\mathbf{L}'\Psi^{-1}\mathbf{L} = \Delta$ with Δ being a diagonal matrix, is imposed to obtain solutions.
The likelihood in (2.3) is based on the assumption that vectors $\mathbf{f}_i$ and ε_i are normally distributed. Typically, the data from microarray experiments are not normally distributed, therefore we take the natural logarithm of the data to better approximate the normality assumption. However, the FA model is known to be rather robust to departures from normality (Johnson and Wichern, 1998).
Maximum likelihood FA estimates of factor loadings, communalities, and specific variances were obtained using the EM algorithm of Rubin and Thayer (1982). The standard errors presented in Tables 2.1 to 2.4 are calculated from the empirical variance of the eigenvalues or the absolute values of the factor loadings across 100 replicated data sets.

B. Factor Rotation

Notice that in Section A. there are an infinite number of solutions using the likelihood in (2.3), each of which is related to the other by an orthogonal rotation matrix, say $\mathbf{T}$. This can be shown as

$\Sigma = \mathbf{L}^*\mathbf{L}^{*\prime} + \Psi = \mathbf{L}\mathbf{T}'\mathbf{T}\mathbf{L}' + \Psi = \mathbf{L}\mathbf{L}' + \Psi$ (Johnson and Wichern, 1998). Factor rotation is usually performed to clarify the interpretation of the factor loadings. The need for clarification comes from the orientation of the orthogonal factor loadings on the m-dimensional grid axes. Consider the case of two retained factors. In a two-dimensional plot of the retained factors, the factors will be orthogonal to each other but some angle of reflection away from the x and y grid axes, much like plots of principal component scores. Factor rotation is often performed subject to an objective criterion, with the Varimax procedure (Kaiser, 1958) the most popular one. We will use the Minimum Entropy (Inagaki, 1994) factor rotation criterion, which follows from the assumption that the factor loadings within a given factor should be either near zero or near plus or minus one.

C. Bootstrap Confidence Intervals

Bootstrap confidence intervals for factor loadings were then obtained using the bias corrected and accelerated (BC_a) method (Efron and Tibshirani, 1993). This method differs from that presented in Ichikawa and Konishi (1995) in that it uses adjusted percentiles of the bootstrap distribution of factor loadings to find the upper and lower bounds. To prevent unequal rotation of the factor loadings, the rotated bootstrap estimates of the factor loadings were aligned with the rotated ML estimates using the alignment by Clarkson algorithm (Clarkson, 1979). The alignment algorithm first re-orders the factors in the bootstrap factor loading matrix to that of the maximum likelihood factor loading matrix. The signs of the factor loading entries are aligned by multiplying the transpose of the bootstrap factor loading matrix with the maximum likelihood factor loading matrix. The sign of the diagonal entry of the product of this multiplication is then used to adjust the signs of the factor loading entries in the bootstrap sample.

Standard errors for the factor loadings of a single data set were approximated using a complex number variant of the central difference method (Lange, 1999) with the gradient given in Jamshidian (1997). This method is less prone to underflow errors than the first-order Richardson extrapolation method presented in Jamshidian and Jennrich (2000).

D. Relationship of Factor Analysis to Singular Value Decomposition

Singular Value Decomposition (SVD) of a standardized expression matrix $\mathbf{Y}^*$ of dimension $p \times n$ is becoming a popular analysis tool for microarray data (Alter *et al.*, 2000; Holter *et al.*, 2000; Holter *et al.*, 2001; West *et al.*, 2001). The SVD of $\mathbf{Y}^*$ is: $\mathbf{Y}^* = \mathbf{U}\mathbf{S}\mathbf{V}'$, where $\mathbf{U}$ is $p \times p$, $\mathbf{S}$ is $p \times n$, $\mathbf{V}'$ is $n \times n$, and $\mathbf{U}$ and $\mathbf{V}'$ are orthogonal. The first q nonzero diagonal elements of the diagonal matrix $\mathbf{S}$ are the square roots of the first q nonzero eigenvalues in the spectral decomposition $\mathbf{Y}^{*\prime}\mathbf{Y}^{*} = \mathbf{U}\mathbf{S}^2\mathbf{U}'$. Principal Components FA can be performed using the spectral decomposition of the sample correlation matrix $\hat{P} = \mathbf{Y}^{*\prime}\mathbf{Y}^{*}$ (Johnson and Wichern, 1998).

In an appendix (Section), a method that combines hierarchical clustering and principal components factor analysis is described (Peterson, 2001). Differences between the implementation and strategy used here and that of Peterson (2001) are also presented.

E. Cluster Analysis

Cluster analysis (HCA) is a class of multivariate statistical methods which group variables according to different distance rules so that variables with a small distance are grouped together. The use of clustering for microarray data was made popular by Eisen *et al.* (1998) who developed software (Cluster (Eisen, 1999)) exclusively for this purpose. Cluster performs several types of data transformations including natural logarithm, mean and median centering of the data by arrays or genes, three types of hierarchical clustering (average, complete, and single linkage clustering), k-means clustering (MacQueen, 1967), and SOMs (Tamayo *et al.*, 1999). Their implementation utilizes the distance measure in equation (2.4) for all correlation based metrics, where r_{ij} is the correlation coefficient between variables i and j.

$$distance_{ij} = 1 - r_{ij} \tag{2.4}$$

The node averaging method used in Cluster differs from standard hierarchical clustering according to the following example. Consider grouping a node consisting of two genes with a new gene, creating a new node consisting of all three genes. In standard hierarchical clustering, the new node value is the average of the old average, from two genes, and the expression value of the new gene. In Cluster, the node value is the average expression values of all three genes.

As in the study by Eisen *et al.* (1998), we used average linkage HCA to analyze a representative data set from each transcriptional activation and inhibition scenario. The distance measure (2.4) utilized the standard centered Pearson correlation coefficient for Figures 2.6 through 2.8, and the absolute value of the standard centered Pearson correlation coefficient for Figures 2.9 through 2.11.

F. Simulation

Data were simulated for a hierarchical metabolic pathway (Hofmeyr and Westerhoff, 2001) using the biochemical simulation program Gepasi (Mendes, 1993; Mendes, 1997; Mendes and Kell, 1998; Mendes, 2000). Each step in the pathway was simulated as depicted in Figure 2.1 (Mendes, 1999). To ensure that identical steady states were not replicated throughout each data set, parameters for the steady state reaction steps (*e.g.* K_{cat} in a Michaelis-Menten equation) were drawn from normal distributions. Specific details of the simulation model can be found in Mendes (1999).

G. Independent Metabolic Pathways

Two branched metabolic pathways consisting of six genes and seven metabolites each with activation and inhibition of mRNA transcription were simulated for the activation and inhibition scenarios depicted in Figures 2.2 through 2.4. Genes were ordered as shown in Figure 2.5. As seen in Figure 2.5, the two pathways do not share any genes, enzymes, or metabolites, *i.e.* they are completely independent.

H. Semi-independent Metabolic Pathways

Two branched metabolic pathways feeding into a single branched metabolic pathway were simulated using a simple substrate activation and product inhibition transcription regulation scenario (Figure 2.12). Only the mRNA concentrations produced at steady state by Gepasi for genes G_1 to G_{12} were used in the analysis, with expression profiles on the other genes assumed unavailable.

I. Error Simulation

One hundred data sets each consisting of one hundred observation vectors were simulated and subsequently analyzed using maximum likelihood FA. The data consist of the mRNA concentrations output from Gepasi at steady state, with error added to the mRNA concentration values using the model (2.5) of Rocke and Durbin (2001).

$$y^*_{ij} = y_{ij} e^{\eta_j} + \varepsilon_{ij} \tag{2.5}$$

In this model, y_{ij} is the mRNA concentration of gene j in observation i without error, y^*_{ij} is the expression intensity value of gene j in observation i, $\eta_{ij} \sim N\left(0, 0.05\sigma^2_{\ln y}\right)$, and $\varepsilon_{ij} \sim N\left(0, 0.20\sigma^2_y\right)$.

The results of FA are reported as the average absolute value of each factor loading over the one hundred replicates along with empirical standard errors. The results of the HCA analysis are reported as dendrograms for a representative data set.

Discussion

A. Selection of the Number of Retained Factors

In the analysis of each of the three activation and inhibition scenarios, only two factors were retained. The critical value for the eigenvalue of the third factor to exceed using the "Broken Stick" criterion (Jackson, 1993) was 1.603 while the largest eigenvalue associated with the third factor was 1.445, (see Table 2.1). Most eigenvalues associated with the third factor were < 1.0. This was expected since there are only two independent pathways in our simulated data sets (see Section).

The "Broken Stick" criterion (Jackson, 1993) can produce results different from those of the "Scree Plot" method (Johnson and Wichern, 1998). In fact, it may be difficult to determine the number of factors to retain using the "Scree Plot" method for the data in Figure 2.3 (Table 2.1) since the plot would appear more as a gentle descending slope than a steep cliff with rubble at the bottom (the latter is the definition of scree, a mineral mining term). Our strict adherence to the critical value in the "Broken Stick" method appears to work well in determining the correct number of factors to retain, at least within the regulatory scenarios of our simulated branched pathway structures.

While a ML hypothesis test of the number of factors to retain can be performed (Johnson and Wichern, 1998), it is not recommended. The test is known to perform poorly on low numbers of observations due to its assymptotic nature, the current situation in most microarray data sets, and can lead to too many retained factors. Initially this may not seem problematic, but remember that the factor loading matrix requires rotation, subject to some criterion, to clarify the interpretation. In fact, all of the data sets in this study were rotated through a non-zero angle of reflection that minimized the Minimum Entropy criterion (Inagaki, 1994). The problem with too many factors retained lies in the fact that the optimal angle of reflection for two retained factors may not be the same as the optimal angle for three retained factors (Lawley and Maxwell, 1971). The interpretation of the factor loading matrix then becomes obscured.

B. Interpretation of the Factor Loadings

Average Maximum Likelihood Estimates

In data sets for scenarios where the difference between the second and third eigenvalue was large, the separation of genes into the two pathways was clear (see Tables 2.2 and 2.4). In these data sets, the average absolute factor loading for genes G_1 through G_6 is at least 10 times larger for the first factor when compared to the second factor, and *vice versa* for genes G_7 through G_{12}. The first factor represents the genes pertaining to pathway 1 (see Figure 2.5) as only genes G_1 through G_6 have high loadings on this factor. The second factor represents the genes involved in pathway 2 (see Figure 2.5).
For the scenario where the third factor explains more than 10% of the total variance (Figure 2.3, Table 2.3, fewer genes within each pathway can be correctly assigned. The factor loading pattern for the scenario using the activation and inhibition scenario for transcription is presented in Figure 2.3 is shown in Table 2.3. In Table 2.3, genes G_2 through G_6 for pathway 1 and genes G_7 through G_{11} for pathway 2 can be assigned, while it is questionable as to whether genes G_1 and G_{12} can be allocated to either pathway. Inspection of Figure 2.5 will show that the ordering of the genes within each pathway is identical. However, while G_1 and G_{12} cannot be allocated to either pathway, their function within each pathway is quite different. Two possible causes for this finding are:

1. The rotation criterion fails to find the optimal rotation.

2. The increase in the proportion of the variance explained by the third factor (Table 2.1) is due to the expression pattern of these genes.

Inspection of the communalities and specific variances for these genes in Table 2.3 indicates Reason 2 as the likely cause. This type of activation and inhibition scenario mimics the transcription behavior of prokaryotes, where multiple genes are transcribed via a single operon.

Bootstrap Confidence Intervals

A single data set simulated with the transcriptional activation and inhibition scenario of Figure 2.3 was analyzed using factor analysis. This is the same data set used in hierarchical clustering to produce Figures 2.6 and 2.10. Bootstrap 95% confidence intervals were obtained for the factor loadings using 1,000 bootstrap samples from the BC_a method. While the interpretation of the average factor loadings was difficult for this scenario (Table 2.3), bootstrap confidence intervals give an indication as to how much information is contained in each individual loading and aid in interpretation.

Inspection of the confidence intervals in Table 2.5 shows that the intervals for genes G_2 through G_4 on factor 2 are roughly symmetrical around zero, indicating that these factor loadings are simply rotating around 0. The confidence intervals for these same genes on factor 1 are limited to the negative parameter space and indicate that these genes load on this factor. While the confidence intervals for genes G_1 and G_6 are skewed for both factor 1 and 2, it is fairly clear from the size of the estimated factor loadings that these genes load on factor 1. The confidence intervals for gene G_5 are similar for both factor 1 and factor 2, and do not clarify the allocation of this gene to either factor, but a case could be made for this gene to load on factor 1.

The allocation of genes G_7 through G_{11} to factor 2 is clear as the confidence intervals for these genes strongly indicate loadings in the positive parameter space. The confidence intervals for these genes on factor 1 are also roughly symmetrical around zero. The confidence interval for gene G_{12} on factor 1 is fairly symmetrical around zero, while the interval for factor 2 is limited to the negative parameter space. Thus, the allocation of gene G_{12} to factor 2 is clarified using the confidence intervals.

Inspection of the communalities for genes G_1, G_5, G_6, and G_{12} in Table 2.5 shows that these genes do not explain much of the variance when only two factors are retained. This would indicate that these genes are actually loading on another factor that was not retained. It is interesting to note that three (G_1, G_2, and G_{12}) of these genes can be assigned to a factor through the application of bootstrap confidence intervals, although the assignment of gene G_5 to a specific factor was not clarified.

C. Hierarchical Clustering

For the activation and inhibition scenario depicted in Figure 2.2, HCA with the signed Pearson correlation coefficient correctly grouped genes G_2, G_4, G_5, and G_6 and genes G_8, G_{10}, G_{11}, and G_{12} together (Figure 2.6). In Figure 2.6 genes G_1 and G_3 and genes G_7 and G_9 are inferred more related to each other than the pathways of which they are constituents. Similar misclassifications can be seen in Figures 2.7 and 2.8.

The relatively poor performance of HCA to correctly allocate genes to pathways is evident in Figures 2.6 through 2.8. There is, however, a simple explanation. Equation (2.4) shows that the distance between genes is measured in terms of the signed Pearson correlation coefficient.

Thus negatively correlated genes are further apart than positively correlated genes of the same magnitude. An inspection of the correlation matrix of the genes in pathway 1 (data not shown) for the activation and inhibition scenario in Figure 2.2 shows that genes G_1 and G_2 are negatively correlated, as are several other genes in the data set. This finding is expected given substrate activation and product inhibition regulation of trancription. For example, Figure 2.2 shows that an increase in substrate S_1 simultaneously causes both an increase in the expression of gene G_2 and a decrease in the expression of gene G_1. These results may help to explain the presence of several of the unexpected genes in the clusters seen in Eisen *et al.* (1998).

The data sets used to produce the dendrograms in Figures 2.6, 2.7, and 2.8 were re-analyzed using the absolute value of the correlation coefficient, rather than the signed value, and average linkage HCA. The improvement of the groupings by using the absolute value of the correlation coefficient is evident in Figures 2.9 through 2.11. The two pathways are now on separate clades within each dendrogram. As an example, the resulting dendrogram in Figure 2.9, corresponding to the activation and inhibition scenario in Figure 2.2, now correctly separates the two pathways, but shows genes G_1, G_3, and G_5 form a separate outgroup distinct from genes G_2, G_4, and G_6. A similar result is seen in the dendrogram in Figures 2.10 and 2.11, where genes G_2, G_3, and G_4 are a separate outgroup from genes G_1, G_5, and G_6 in Figure 2.10 and genes G_2, G_4, and G_5 are a separate outgroup from genes G_1, G_3, and G_6 in Figure 2.11. Gene G_{12} is an outgroup from the rest of pathway 2 in both of these figures.

The genes appearing as outgroups (*e.g.* genes G_1, G_5, G_6, and G_{12} in Figure 2.10) are the same genes with low communalities in Table 2.5. These genes clearly do not contain much information about group membership, regardless of clustering method used. In factor analysis, bootstrap confidence intervals were used to clarify inference of pathway membership, and analogous approaches exist for clustering. Yeung (Yeung *et al.*, 2001; Yeung and Ruzzo, 2001) present a method to validate results from hierarchical clustering that may clarify the results seen in Figures 2.9 through 2.11. Munneke *et al.* (2001) also provide a stopping rule for cluster separation based upon a permutation test to aid clarification of the degree of cluster separation. Following 1,000 permutations, three clusters were significant at the p = 0.10 level: cluster 1 included genes G_7, G_8, G_9, G_{10}, and G_{11}; cluster 2 included genes G_1, G_2, G_3, G_4, G_5, and G_6; and cluster 3 included gene G_{12}. No clusters were significant at the $p < 0.10$ level.

It should be noted that other multivariate methods (*e.g.* SOMs and k-means clustering) that use correlation coefficients as a distance measure will improve in quality of clustering when the absolute value of the coefficient is used. Additionally, distance measures other than the correlation coefficient need to account for inverse relationships in the magnitude (*e.g.* high-low and low-high) of the signal between genes.

D. Semi-independent Metabolic Pathways

Using the "Broken Stick" criterion, the critical value for the second eigenvalue to exceed is 2.103. Inspection of Table 2.1 shows that for the semi-independent pathways of Figure 2.12 the second eigenvalue is 1.571, thus only one factor is retained. This single factor contains all twelve genes (Table 2.6) and shows that factor analysis can find all observed genes within a metabolic pathway, even when there are missing genes providing links between the observed genes. Average linkage hierarchical clustering was performed using the absolute value of the correlation coefficient as the distance measure on a single representative data set. The dendrogram in Figure 2.13 shows that parts of the pathway can be recovered (genes G_1 through G_4 and genes G_6 through G_{11}), but two genes appear to be unrelated, genes G_5 and G_{12}. Inspection of the factor loadings in Table 2.6 shows that the loading for gene G_{12} is low, but this gene is related to the others on this factor. Again, the use of additional analysis tools for hierarchical clustering can aid in interpretation by providing validation of clusters and stopping rules for the separation of clades.

Conclusions

Many analysis tools have been developed for microarray gene expression data, including clustering algorithms and methods for genetic network inference. However, these methods have typically been applied to real rather than simulated data. The utility and properties of these methods are often not well understood, but can be investigated with simulated data and used to advance our interpretational capabilities for microarray expression data. We therefore believe that it is important to evaluate analysis tools for microarray expression data with simulated data. Here, we have shown how to apply FA methodology to microarray gene expression data so that metabolic pathways can be recovered. We also show that a popular clustering method performs poorly under certain model assumptions, namely that negatively correlated genes are less related than positively correlated genes. Further, additional analysis tools may be needed to interpret results from this method when this assumption is not utilized (*e.g.* determining the degree of separation between clades on a dendrogram).

While previous studies using experimental data have found clusters of genes involved in similar functional roles (*e.g.* the ribosomal, proteosome, and histone gene clusters in Eisen *et al.* (1998)) we seek clusters based upon functional relationships between the genes, namely genes that are members of a single metabolic pathway. Under certain activation and inhibition scenarios for transcription in metabolic pathways, FA can correctly group all genes, even when the genes for a metabolic pathway are only partially observed. There are, however, some activation and inhibition scenarios for transcription where FA can only correctly group most genes. This finding indicates that the activation and inhibition scenario for transcription, and not just the structure topology, of a metabolic pathway affects the interpretation of the factor loadings. Additional metabolic pathway structures need investigation (*e.g.* looped pathways, moiety conserved cycles, and linear pathways) along with different activation and inhibition scenarios to fully evaluate the ability of FA to accurately group genes by metabolic pathway.

The results of HCA on the activation and inhibition scenarios for transcription indicate that incorrect results arise when the absolute value of the Pearson correlation coefficient is not used. Even when it is used, the dendrograms in Figures 2.9 through 2.11 and Figure 2.13 show genes within a single pathway as fairly unrelated. This is where FA provides a useful heuristic. The factor loading matrix is a matrix of correlation coefficients between the standardized $p \times 1$ data vector $\mathbf{y}_i^*$ and the $m \times 1$ vector of factor scores $\mathbf{f}_i$ for the $i^{\underline{\text{th}}}$ experimental unit, or $Cov\left(\mathbf{y}_i^*, \mathbf{f_i}'\right) = Cor\left(\mathbf{y}_i^*, \mathbf{f_i}'\right) = Cov\left(\mathbf{L}\mathbf{f}_i, \mathbf{f_i}'\right) = \mathbf{L}$. The magnitude of the factor loadings then indicate the strength of the relationship between the genes and the putative pathways. Cluster validation (Yeung *et al.*, 2001; Yeung and Ruzzo, 2001) and stopping rules for cluster divergence (Munneke *et al.*, 2001) may improve the interpretation of hierarchical clustering results.

Lastly, we comment on the bootstrap confidence intervals in Table 2.5. As a first approximation to confidence intervals, the maximum likelihood estimate ± 2 times the standard error of the estimate produces an approximate 95% confidence interval. Such intervals can be seen in Table 2.7. Notice that these intervals are smaller for many of the factor loadings than the bootstrap intervals. The approximate confidence intervals are appropriate only if the ML estimate of a factor loading is normally distributed. The discrepancy between the approximate and the bootstrap confidence intervals may be due to the estimation of factor loadings not using a 'regular' problem, *e.g.* the lack of uniqueness of the factor loadings in spite of the rotation and alignment procedure described earlier. We have recently implemented a Bayesian exploratory factor analysis (BEFA), which is described in another communication (Henderson and Hoeschele, 2002). Included results show that many BEFA highest posterior density regions are also smaller than the bootstrap confidence intervals presented here.

References

Alter O, Brown PO & Botstein D (2000) Singular value decomposition for genome wide expression data processing and modeling. *Proc. Nat. Acad. Sci.* 97(18), 10101–10106

Clarkson DB (1979) Estimating the standard errors of rotated factor loadings by jackknifing. *Psychometrika* 44(3), 297–314

Efron B & Tibshirani RJ (1993) *An introduction to the bootstrap*. Chapman and Hall, New York, USA

Eisen MB (1999) *Cluster and TreeView Manual*. Stanford University, Palo Alto, CA, USA, 1st edition

Eisen MB, Spellman P, Brown PO & Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci.* 95, 14863–14868

Friedman N, Linnial M, Natchman I & Pe'er D (2000) Using bayesian networks to analyze expression data. *J. Comp. Biol.* 7, 601–620

Giersch C (1994) Determining elasticities from multiple measurements of steady-state flux rates and metabolite concentrations: Theory. *J. Theor. Biol.* 169, 89–99

Giersch C (1995) Determining elasticities from multiple measurements of flux rates and metabolite concentrations: Application of the multiple modulation method to a reconstituted pathway. *Eur. J. Biochem.* 227, 194–201

Heinrich R & Rapoport TA (1974) A linear steady state treatment of enzyme chains: General properties, control, and effector strength. *Eur. J. Biochem.* 42, 89–95

Henderson DA & Hoeschele I (2002) Bayesian and correlated Bayesian exploratory factor analysis for the identification of metabolic pathways from microarray expression data. *Bioinformatics* (Submitted)

Hofmeyr JHS & Cornish-Bowden A (1996) Co-response analysis: A new experimental strategy for metabolic control analysis. *J. Theor. Biol.* 182, 371–380

Hofmeyr JHS & Westerhoff HV (2001) Building the cellular puzzle. control in multi-level reaction networks. *J. Theor. Biol.* 208, 261–285

Hofmeyr JS, Cornish-Bowden A & Rohwer JM (1993) Taking enzyme kinetics out of control; putting control into regulation. *Eur. J. Biochem.* 212, 833–837

Holter NS, Maritan A, Cieplak M, Fedoroff NV & Banavar JR (2001) Dynamic modeling of gene expression data. *Proc. Nat. Acad. Sci.* 98(4), 1693–1698

Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR & Fedoroff NV (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Nat. Acad. Sci.* 97(15), 8409–8414

Ichikawa M & Konishi S (1995) Application of the bootstrap methods in factor analysis. *Psychometrika* 60(1), 77–93

Inagaki A (1994) Entromin and Entromax: Criteria for factor rotation based on entropy. *Jap. J. of Behav.* 21, 10–20

Jackson DA (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74, 2204–2214

Jamshidian M (1997) An EM algorithm for factor analysis with missing data. In Berkane M, editor, *Latent variable modeling and applications to causality*, pages 247–258. Springer-Verlag, New York, USA, 1st edition

Jamshidian M & Jennrich RI (2000) Standard errors for EM estimation. *J. Roy. Stat. Soc. (series B)* 62(2), 257–270

Johnson RA & Wichern DW (1998) *Applied multivariate statistical analysis*. Prentice Hall, Englewood Cliffs, NJ, USA, 4th edition

Kacser H & Burns JA (1973) The control of flux. *Symp. Soc. Exp. Biol.* 32, 65–104

Kaiser HF (1958) The Varimax criterion for analytic rotation in Factor Analysis. *Psychometrika* 23, 187–200

Lange K (1999) *Differentiation of a finite function*. Springer-Verlag, New York, USA

Lawley DN & Maxwell AE (1971) *Factor Analysis as a statistical method*. Butterworths, London, 2$^{\underline{nd}}$ edition

MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In *Proc. 5$^{\underline{th}}$ Berk. Symp. Math. Stat. Prob.*, volume 1, pages 281–297, Berkely, CA, USA. Univ. CA Press

Mendes P (1993) Gepasi: a software package for modeling the dynamics, steady states, and control of biochemical and other systems. *Comp. Appl. Biosci.* 9, 563–571

Mendes P (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* 22, 361–363

Mendes P (1999) Metabolic simulation as an aide in understanding gene expression data. In *Proc. Workshop Comp. Biochem. Path. And Genetic Networks*, pages 27–33, Berlin, Germany

Mendes P (2000) Gepasi: Numerical simulation and optimization of biochemical kinetics. In *Proc. Plant and Anim. Genome VIII*, San Diego, CA, USA

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

Munneke B, Beavis WD, Schlauch KA & Doerge RW (2001) Gene expression cluster validation and a novel dissimilarity measure. *Bioinformatics* submitted

Peterson LE (2001) Factor analysis of cluster-specific gene expression levels from cDNA microarrays. *Comp. Meth. and Prog. in Biosci.* in press, –

Rocke DM & Durbin B (2001) A model for measurement error for gene expression arrays. *J. Comp. Biol.* In Press

Rubin DB & Thayer DT (1982) EM algorithms for ML factor analysis. *Psychometrika* 47(1), 69–76

Sen AK (1996) On the sign pattern of metabolic control coefficients. *J. Theor. Biol.* 182, 269–275

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES & Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci.* 96, 2907–2912

West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Marks JR & Nevins JR (2001) Predicting the clinical status of human breast cancer using gene expression profiles. *Proc. Nat. Acad. Sci.* 98, 11462–11467

Yeung KY, Haynor DR & Ruzzo WL (2001) Validating clustering for gene expression data. *Bioinformatics* 17(4), 309–318

Yeung KY & Ruzzo WL (2001) Principal component analysis for gene expression data. *Bioinformatics* 17(9), 763–774

Appendix

A microarray experiment produces only limited knowledge of the state of a specific sample. Knowledge is limited in that we only indirectly observe levels of the abundance of transcript (mRNA). In our simulated data, the relative levels of transcript are inextricably tied to the concentrations of the substrates and products of the metabolic reactions that make up the pathway. The concentrations of the substrates and products are tied to the reaction rates of the steps in the pathway to which they are associated. Given this description of the connected nature of our simulation, we can view the reaction rates of each step as some function of the observed mRNA abundance, or concentrations.

Metabolic control analysis (MCA) (Kacser and Burns, 1973; Heinrich and Rapoport, 1974) is a tool to investigate the control and sensitivity of members of metabolic and regulatory networks, or pathways (Giersch, 1994; Giersch, 1995; Hofmeyr *et al.*, 1993; Hofmeyr and Cornish-Bowden, 1996; Hofmeyr and Westerhoff, 2001; Sen, 1996). The theory behind our simulation model is based upon recent developments in MCA (Hofmeyr and Westerhoff, 2001). Using MCA, one can show that the control on the flux (the steady state reaction rate) of any step in pathway 2 through the concentration of any mRNA species in pathway 1 is zero, and *vice versa* through the concentration of mRNA from pathway 2 on the flux in pathway 1. Additionally, the control of the flux of any step in pathway 1 through any mRNA species in pathway 1 is not zero if any regulatory link between transcription of that mRNA species and that reaction step exists and likewise for mRNA and reaction steps in pathway 2. Given the reaction scenarios for our simulated data sets, we would expect two factors. One factor containing information about the control of the flux in pathway 1 through the mRNA concentrations observed for enzymes in pathway 1, and another for pathway 2. Since there is no cross control between the two pathways, we would expect these two factors to be independent.

Appendix

Peterson (2001) developed software (CLUSFAVOR) for joint hierarchical clustering and factor analysis of cDNA microarray data. In this software, the genes (or arrays) are grouped using average, single, or complete linkage hierarchical clustering with one of two distance measures: 1) the distance measure presented in equation (2.4), and 2) Euclidean distance. Principal components factor analysis is then performed on either the genes contained within selected clusters (branches of the dendrogram), or the entire data set. The number of factors is set at the number of eigenvalues of the correlation matrix exceeding unity (1.0) and a varimax rotation is then performed on the resulting factor loading matrix. Visual interpretation of the factor loadings is from a dual color display where within a single factor, loadings that exceed 0.45 are one color and loadings that are below -0.45 are another.

Results presented in this contribution have shown that using the signed correlation coefficient as a distance measure produces incorrect groupings. The Euclidean distance measure should perform similarly. Additionally, using the number of factors whose eigenvalues exceed unity to select the number of factors will also produce problems. The eigenvalue of the third factor for the data shown in Figure 2.3 exceeds 1.0 (Table 2.1), but its inclusion in the analysis obscures the interpretation of the factor loadings following rotation. The choice of rotation method in Peterson (2001) can also obscure the interpretation when an excess number of factors are retained as the variance explained by each factor will tend to be spread across the factors, rather than concentrated on the first few.

The factor analysis in Peterson (2001) uses the principal components method, which we show to be related to SVD. This method is based upon the approximation of the sample (co)variance or correlation matrix through a reduced number of factors, while our implementation of factor analysis utilizes the likelihood in (2.3), based upon the common factor model in (2.1).

A concern in using CLUSFAVOR is the interpretation of large factor loadings on clearly separate branches of a dendrogram produced using hierarchical clustering. Should these branches be related, although their portrayal in the dendrogram suggests otherwise? This is where our implementation of factor analysis clarifies the interpretation of the factor loadings. Each retained factor represents a single metabolic pathway, regardless of the sign on the loading entries. The magnitude of each loading entry indicates the strength of the relationship between that gene and the pathway represented by the retained factor.

Table 2.1. Average eigenvalues of correlation matrices.

Figure	Eigenvalue Number	Eigenvalue	Standard Error	Proportion of Total Variance
	1	4.8808	$\pm$ 0.0278	40.6734
2.2	2	4.1719	$\pm$ 0.0287	34.7655
	3	0.5398	$\pm$ 0.0055	4.4985
	1	3.1270	$\pm$ 0.0207	26.0580
2.3	2	2.3795	$\pm$ 0.0165	19.8293
	3	1.4451	$\pm$ 0.0104	12.0422
	1	4.8327	$\pm$ 0.0243	40.2722
2.4	2	4.1069	$\pm$ 0.0237	34.2238
	3	0.5600	$\pm$ 0.0076	4.6667
	1	8.4112	$\pm$ 0.0173	70.0934
2.12	2	1.5706	$\pm$ 0.0127	13.0887
	3	0.5325	$\pm$ 0.0062	4.4374

Table 2.2. Average of 100 samples of absolute value of Maximum Likelihood Factor loadings for pathways with the transcriptional activation and inhibition scenario in Figure 2.2.

	Rotated Factor Loadings			
Gene	Factor 1	Factor 2	Communality	Specific Variance
G_1	0.8508 ± 0.0025	0.0393 ± 0.0029	0.7268 ± 0.0042	0.2406 ± 0.0031
G_2	0.8205 ± 0.0033	0.0423 ± 0.0036	0.6774 ± 0.0053	0.2740 ± 0.0039
G_3	0.8515 ± 0.0030	0.0469 ± 0.0041	0.7298 ± 0.0051	0.2398 ± 0.0037
G_4	0.8264 ± 0.0032	0.0450 ± 0.0032	0.6870 ± 0.0053	0.2660 ± 0.0039
G_5	0.8815 ± 0.0027	0.0392 ± 0.0029	0.7802 ± 0.0047	0.2018 ± 0.0032
G_6	0.8798 ± 0.0025	0.0361 ± 0.0030	0.7769 ± 0.0043	0.2034 ± 0.0030
G_7	0.0735 ± 0.0055	0.8414 ± 0.0030	0.7173 ± 0.0049	0.2555 ± 0.0035
G_8	0.0600 ± 0.0052	0.7951 ± 0.0032	0.6394 ± 0.0049	0.2989 ± 0.0036
G_9	0.0676 ± 0.0049	0.8696 ± 0.0030	0.7639 ± 0.0049	0.2201 ± 0.0034
G_{10}	0.0676 ± 0.0050	0.8084 ± 0.0033	0.6617 ± 0.0054	0.2822 ± 0.0039
G_{11}	0.0619 ± 0.0057	0.8641 ± 0.0026	0.7545 ± 0.0044	0.2195 ± 0.0030
G_{12}	0.0707 ± 0.0053	0.8439 ± 0.0034	0.7210 ± 0.0057	0.2456 ± 0.0041

Table 2.3. Average of 100 samples of absolute value of Maximum Likelihood Factor loadings for pathways with the transcriptional activation and inhibition scenario in Figure 2.3.

	Rotated Factor Loadings			
Gene	Factor 1	Factor 2	Communality	Specific Variance
G_1	0.1825 ± 0.0139	0.0764 ± 0.0068	0.0629 ± 0.0076	0.8704 ± 0.0102
G_2	0.8248 ± 0.0062	0.0408 ± 0.0028	0.6865 ± 0.0098	0.3129 ± 0.0066
G_3	0.6645 ± 0.0068	0.0559 ± 0.0045	0.4512 ± 0.0089	0.4400 ± 0.0088
G_4	0.7618 ± 0.0070	0.0564 ± 0.0043	0.5902 ± 0.0105	0.3522 ± 0.0067
G_5	0.2135 ± 0.0123	0.0982 ± 0.0070	0.0750 ± 0.0063	0.8052 ± 0.0123
G_6	0.3143 ± 0.0088	0.0885 ± 0.0065	0.1185 ± 0.0059	0.7440 ± 0.0096
G_7	0.0871 ± 0.0068	0.5604 ± 0.0079	0.3323 ± 0.0087	0.5314 ± 0.0082
G_8	0.0790 ± 0.0057	0.7830 ± 0.0063	0.6265 ± 0.0094	0.3593 ± 0.0063
G_9	0.0649 ± 0.0053	0.7664 ± 0.0068	0.5989 ± 0.0099	0.3423 ± 0.0062
G_{10}	0.0811 ± 0.0064	0.7324 ± 0.0066	0.5513 ± 0.0095	0.3880 ± 0.0068
G_{11}	0.0837 ± 0.0062	0.6065 ± 0.0066	0.3829 ± 0.0079	0.4589 ± 0.0060
G_{12}	0.0963 ± 0.0079	0.1577 ± 0.0089	0.0482 ± 0.0042	0.8889 ± 0.0075

Table 2.4. Average of 100 samples of absolute value of Maximum Likelihood Factor loadings for pathways with the transcriptional activation and inhibition scenario in Figure 2.4.

	Rotated Factor Loadings			
Gene	Factor 1	Factor 2	Communality	Specific Variance
G_1	0.8566 ± 0.0026	0.0443 ± 0.0036	0.7376 ± 0.0043	0.2326 ± 0.0031
G_2	0.8965 ± 0.0022	0.0383 ± 0.0029	0.8065 ± 0.0038	0.1854 ± 0.0025
G_3	0.8723 ± 0.0025	0.0403 ± 0.0035	0.7644 ± 0.0043	0.2120 ± 0.0030
G_4	0.8899 ± 0.0024	0.0373 ± 0.0029	0.7947 ± 0.0042	0.1937 ± 0.0029
G_5	0.8230 ± 0.0028	0.0486 ± 0.0041	0.6821 ± 0.0045	0.2677 ± 0.0033
G_6	0.7794 ± 0.0041	0.0489 ± 0.0037	0.6129 ± 0.0063	0.3220 ± 0.0050
G_7	0.0624 ± 0.0051	0.7821 ± 0.0041	0.6199 ± 0.0061	0.3296 ± 0.0050
G_8	0.0611 ± 0.0048	0.8900 ± 0.0024	0.7988 ± 0.0043	0.1975 ± 0.0029
G_9	0.0675 ± 0.0052	0.8291 ± 0.0035	0.6959 ± 0.0058	0.2620 ± 0.0041
G_{10}	0.0599 ± 0.0047	0.8805 ± 0.0028	0.7818 ± 0.0048	0.2097 ± 0.0033
G_{11}	0.0754 ± 0.0052	0.7590 ± 0.0060	0.5881 ± 0.0084	0.3426 ± 0.0068
G_{12}	0.0627 ± 0.0050	0.7893 ± 0.0051	0.6320 ± 0.0081	0.3057 ± 0.0061

Table 2.5. Maximum Likelihood Factor loadings for pathways in Figure 2.3 with Bootstrap 95% confidence intervals, 1000 Bootstrap samples.

	Rotated Factor Loadings							
		Factor 1			Factor 2			Specific
Gene	Loading	Lower	Upper	Loading	Lower	Upper	Comm.	Variance
G_1	-0.2699	-0.6528	0.0197	-0.0876	-0.2696	0.0618	0.0805	0.8112
G_2	-0.7897	-0.9623	-0.5842	-0.0130	-0.2001	0.1406	0.6239	0.3147
G_3	-0.6418	-0.7994	-0.4592	0.0334	-0.1655	0.1739	0.4130	0.5104
G_4	-0.7785	-0.9703	-0.6167	0.0336	-0.1449	0.1700	0.6072	0.2985
G_5	0.2076	-0.1489	0.4595	0.1652	-0.0146	0.2994	0.0704	0.8040
G_6	0.2963	0.0377	0.4684	-0.0909	-0.2683	0.0581	0.0961	0.8022
G_7	-0.0472	-0.2873	0.1124	0.6196	0.4373	0.7163	0.3861	0.4737
G_8	-0.0473	-0.1934	0.0676	0.8523	0.7090	0.9193	0.7287	0.2919
G_9	0.0221	-0.1677	0.1817	0.8170	0.6406	0.8880	0.6679	0.2967
G_{10}	-0.0076	-0.1966	0.1397	0.7592	0.6136	0.8329	0.5765	0.3703
G_{11}	0.0486	-0.1557	0.2174	0.6118	0.4459	0.7084	0.3767	0.4733
G_{12}	0.1196	-0.1518	0.2900	-0.2554	-0.4578	-0.1013	0.0795	0.8523

Table 2.6. Average of 100 samples of absolute value of Maximum Likelihood rotated Factor loadings for pathways with the transcriptional activation and inhibition scenario in Figure 2.12.

Gene	Factor 1	Communality	Specific Variance
G_1	0.9108 ± 0.0015	0.8298 ± 0.0028	0.1744 ± 0.0027
G_2	0.9458 ± 0.0013	0.8948 ± 0.0024	0.1286 ± 0.0019
G_3	0.9028 ± 0.0015	0.8153 ± 0.0027	0.2020 ± 0.0028
G_4	0.9468 ± 0.0013	0.8966 ± 0.0025	0.1262 ± 0.0019
G_5	0.7631 ± 0.0039	0.5839 ± 0.0059	0.4187 ± 0.0052
G_6	0.8481 ± 0.0027	0.7201 ± 0.0045	0.2490 ± 0.0041
G_7	0.8620 ± 0.0025	0.7437 ± 0.0043	0.1857 ± 0.0029
G_8	0.7990 ± 0.0033	0.6395 ± 0.0052	0.2612 ± 0.0034
G_9	0.8621 ± 0.0022	0.7437 ± 0.0038	0.1862 ± 0.0023
G_{10}	0.8036 ± 0.0033	0.6469 ± 0.0052	0.2537 ± 0.0034
G_{11}	0.6327 ± 0.0050	0.4028 ± 0.0063	0.4964 ± 0.0056
G_{12}	0.3168 ± 0.0100	0.1102 ± 0.0059	0.9067 ± 0.0056

Table 2.7. Upper and lower confidence limits calculated from twice the standard error for pathways in Figure 2.3. Factor loading estimates are identical to those in Table 2.5.

	Rotated Factor Loadings					
	Factor 1			Factor 2		
Gene	Loading	Lower	Upper	Loading	Lower	Upper
G_1	-0.2699	-0.4388	-0.1010	-0.0876	-0.2246	0.0494
G_2	-0.7897	-0.9332	-0.6462	-0.0130	-0.1351	0.1091
G_3	-0.6418	-0.7774	-0.5062	0.0334	-0.0768	0.1436
G_4	-0.7785	-0.7810	-0.7760	0.0336	0.0171	0.0501
G_5	0.2076	0.0390	0.3762	0.1652	0.0285	0.3019
G_6	0.2963	0.1811	0.4115	-0.0909	-0.1884	0.0066
G_7	-0.0472	-0.1909	0.0965	0.6196	0.5011	0.7381
G_8	-0.0473	-0.0707	-0.0239	0.8523	0.8493	0.8553
G_9	0.0221	-0.0927	0.1369	0.8170	0.7222	0.9118
G_{10}	-0.0076	-0.1508	0.1356	0.7592	0.6374	0.8810
G_{11}	0.0486	-0.0948	0.1920	0.6118	0.4936	0.7300
G_{12}	0.1196	0.1115	0.1277	-0.2554	-0.2843	-0.2265

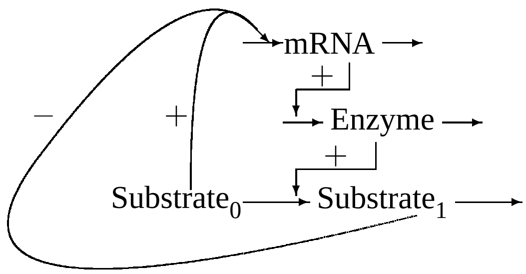

Figure 2.1. Hierarchy of simulated pathway structure.

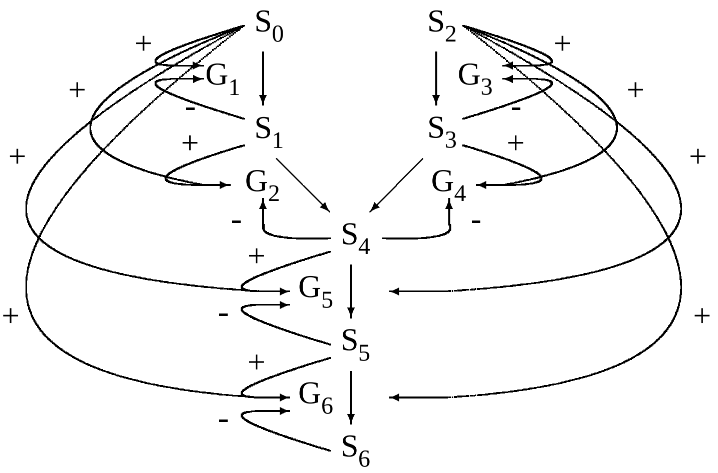

Figure 2.2. Six gened simulated pathway structure with regulatory interactions.

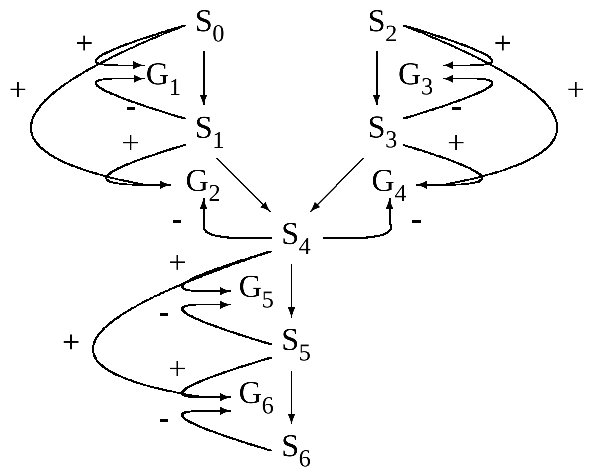

Figure 2.3. Six gened simulated pathway structure with regulatory interactions.

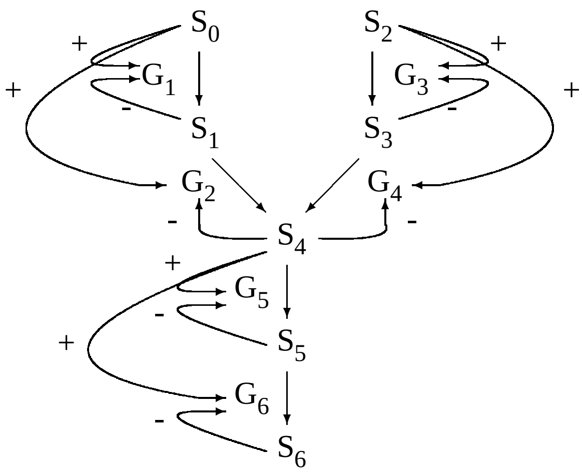

Figure 2.4. Six gened simulated pathway structure with regulatory interactions.

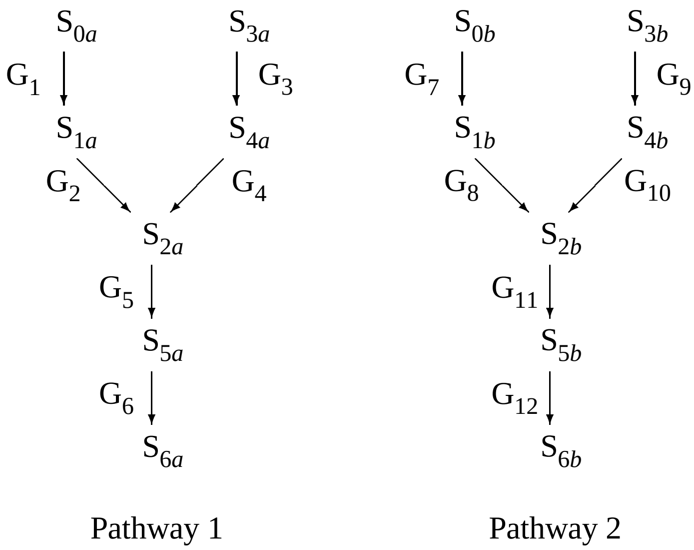

Figure 2.5. Pathway structure for two independent pathways.

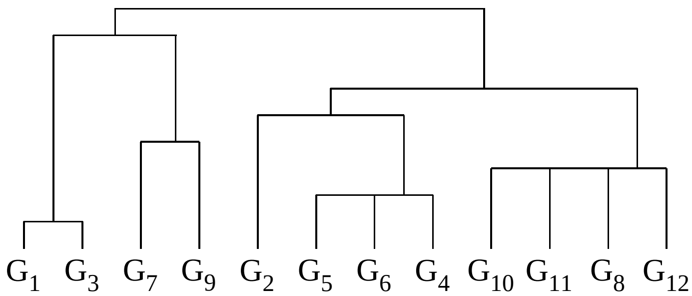

Figure 2.6. Average linkage hierarchical clustering for two independent twelve gene simulated pathways. Structure shown in Figure 2.5, activation and inhibition scheme shown in Figure 2.2.

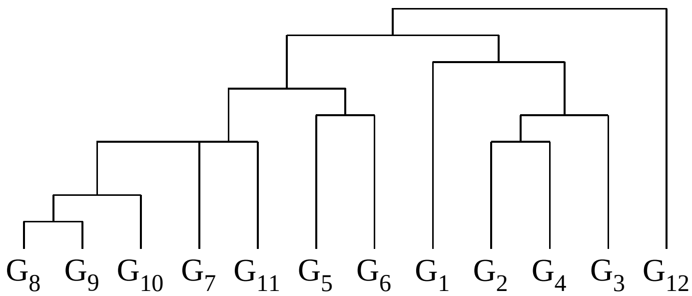

Figure 2.7. Average linkage hierarchical clustering for two independent twelve gene simulated pathways. Structure shown in Figure 2.5, activation and inhibition scheme shown in Figure 2.3.

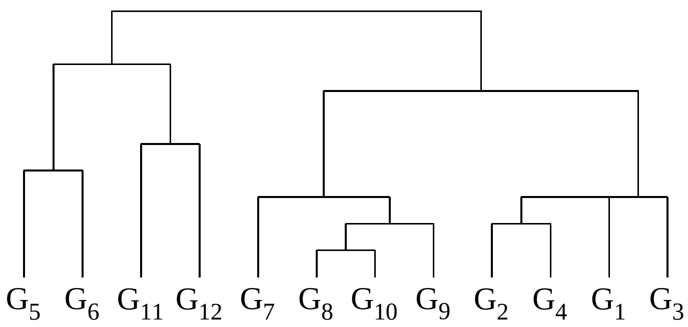

Figure 2.8. Average linkage hierarchical clustering for two independent twelve gene simulated pathways. Structure shown in Figure 2.5, activation and inhibition scheme shown in Figure 2.4.

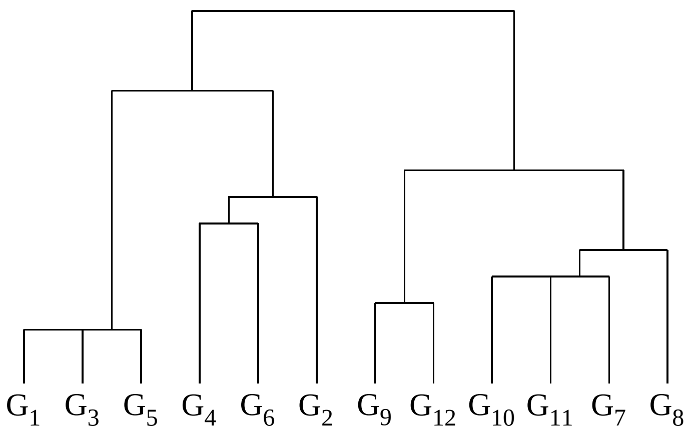

Figure 2.9. Average linkage hierarchical clustering for two independent twelve gene simulated pathways. Structure shown in Figure 2.5, activation and inhibition scheme shown in Figure 2.2. Absolute value of the Pearson correlation coefficient used as the distance measure.

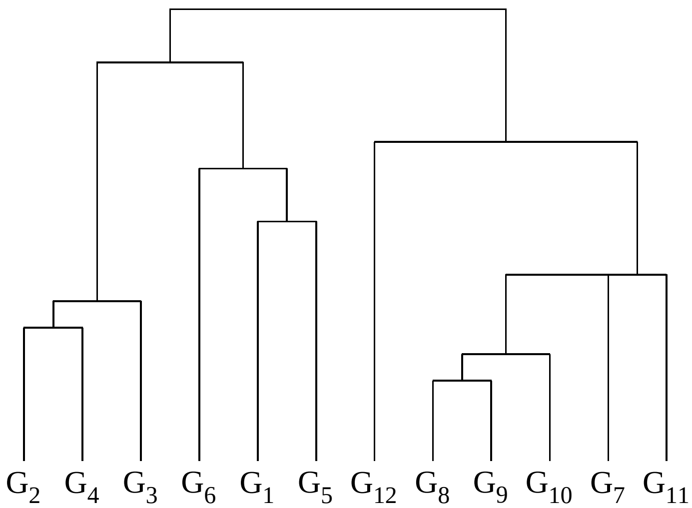

Figure 2.10. Average linkage hierarchical clustering for two independent twelve gene simulated pathways. Structure shown in Figure 2.5, activation and inhibition scheme shown in Figure 2.3. Absolute value of the Pearson correlation coefficient used as the distance measure.

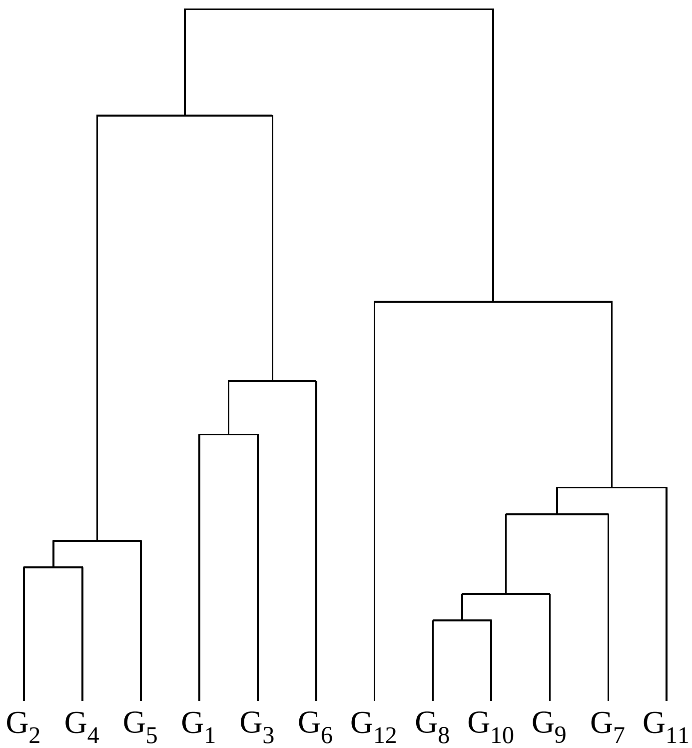

Figure 2.11. Average linkage hierarchical clustering for two independent twelve gene simulated pathways. Structure shown in Figure 2.5, activation and inhibition scheme shown in Figure 2.4. Absolute value of the Pearson correlation coefficient used as the distance measure.

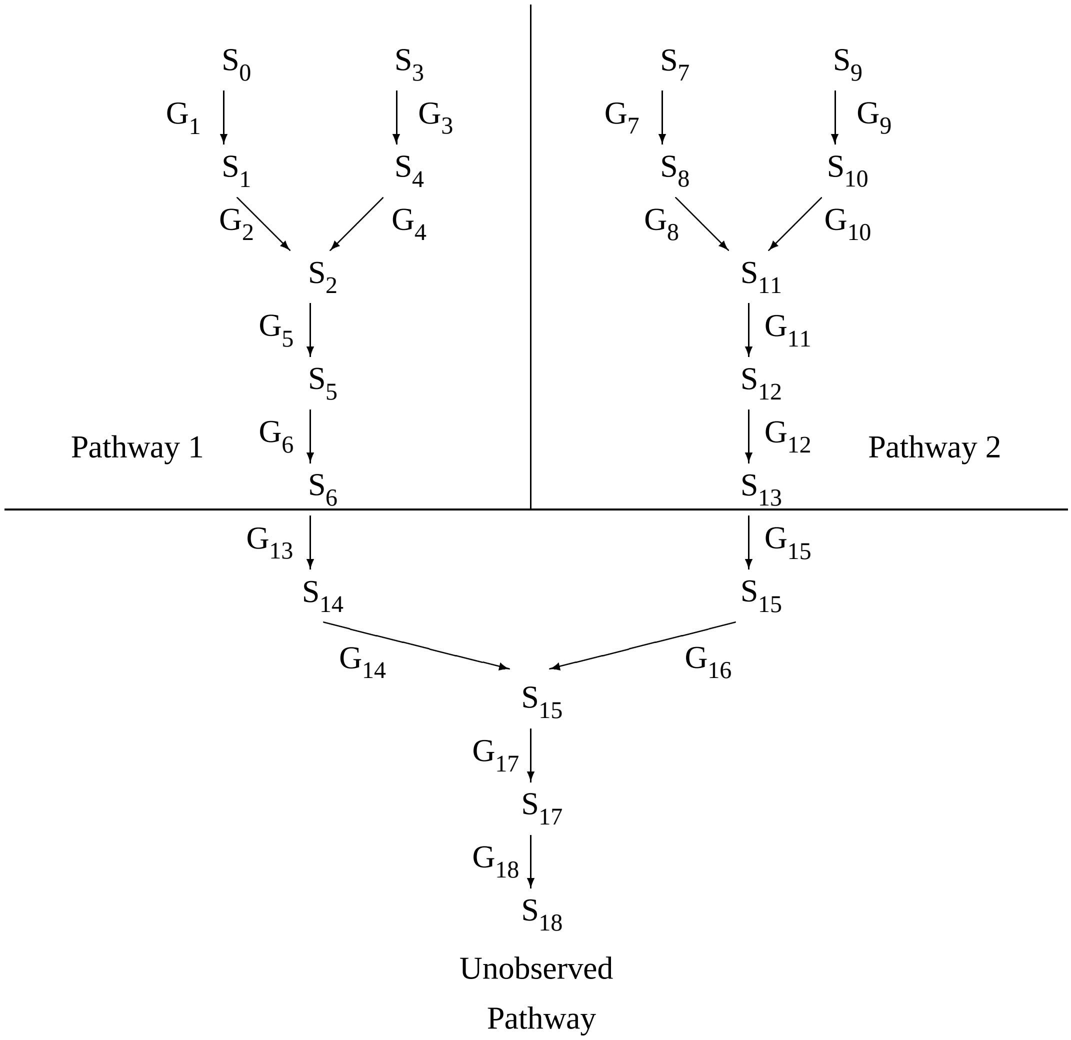

Figure 2.12. Pathway structure for two semi-independent pathways.

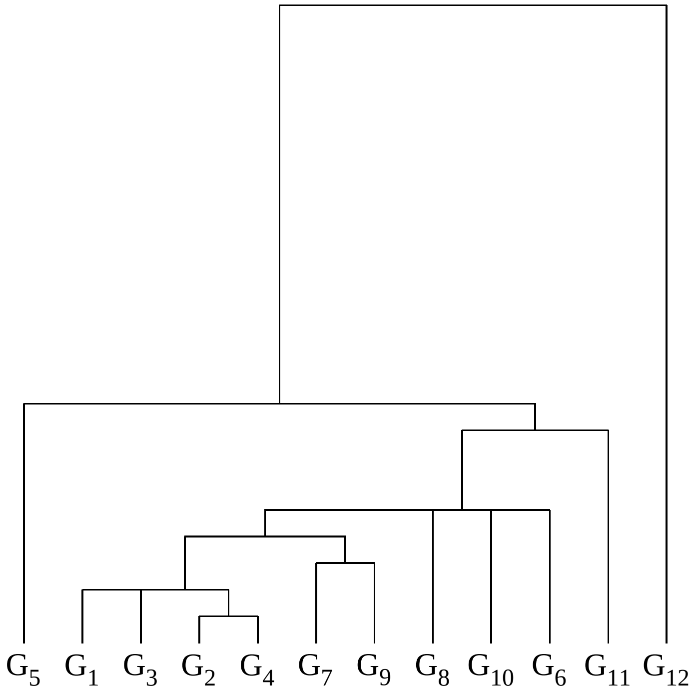

Figure 2.13. Average linkage hierarchical clustering for two independent twelve gene simulated pathways. Structure shown in Figure 2.12.

Chapter 3

Factor analysis of gene expression data for the identification and investigation of metabolic pathways and pathway features using metabolic control analysis

Abstract

Motivation: Identification of local regulatory relationships between genes involved in a single metabolic pathway could greatly improve the efficiency of molecular biology experiments to identify drug or gene therapy targets, or targets for selection within animal and plant genetic improvement programs. Ideally, biologists want to discover the causal network underlying the observed data, *e.g.* the regulatory network that maintains a metabolic pathway. Graphical or Structural Equation Models are methods which look for the most probable model given the data and include models which allow for causal reasoning, in particular Bayesian networks. However, these models may have search spaces which can be prohibitively large. Previously, factor analysis of simulated gene expression was shown to group genes according to membership within a metabolic pathway. In this communication, we use metabolic control analysis to explain why factor analysis can group genes by metabolic pathway. We further use metabolic control analysis to investigate whether factor analysis can indentify regulatory links or pathway features.

Results: Factor analysis is applied to simulated expression values of genes within a single pathway identified using factor analysis on a larger set of gene expression data. Patterns seen in the factor loadings of the first factor and one additional factor (or several additional factors with eigenvalues approximately equal to the second eigenvalue) appear to represent coefficient patterns in a response matrix derived from knowledge of the pathway structure using metabolic control analysis.

Introduction

The identification of not only the genes involved in a single metabolic pathway, but also of the regulatory relationships between them would be particularly useful. This information could be used to identify drug or gene therapy targets, or identify genes for selection in plant and animal genetic improvement programs.

In this communiction, we use metabolic control analysis to explain why factor analysis can group genes by metabolic pathway. We also investigate whether patterns in the factor loadings identify the regulatory relationships in a single metabolic pathway. Such additional inference could simplify the reconstruction of the regulatory network maintaining said pathway.

Materials and Methods

A. Factor Analysis

Factor analysis (FA) (Johnson & Wichern, 1998) is a method to describe the covariance, or correlation, relationships among many variables in a multivariate data set in terms of a few underlying, but unobservable, random quantities called factors. Inferences about the relationships between the observed variables are made through the magnitude, and typically the sign, of the cell entries in the factor loading matrix. Given a $p \times 1$ multivariate data vector $\mathbf{y}_i$ and an unobserved $m \times 1$ vector of factor scores $\mathbf{f}_i$ for experimental unit i, the factor loading matrix is either $\mathbf{L} = Cov\left(\mathbf{y}_i, \mathbf{f}_i'\right)$ or $\mathbf{L} = Cor\left(\mathbf{y}_i, \mathbf{f}_i'\right)$ for any i, depending on whether or not the data have been standardized. The number of factors m is specified based upon either subjective (*e.g.* a "Scree" plot (Johnson & Wichern, 1998)) or objective (*e.g.* "Broken Stick" (Jackson, 1993)) criteria and $m \leq p$ with p the number of variables. The portion of the total variance explained by the retained factors for a specific variable is termed the communality, and the portion not explained by the retained factors is termed the specific variance. Details of our implementation of maximum likelihood factor analysis, factor rotation and alignment, and the relationship to singular value decomposition can be found in Henderson *et al.* (2002).

B. Simulation

Data were simulated for a hierarchical metabolic pathway (Hofmeyr & Westerhoff, 2001) using the biochemical simulation program Gepasi (Mendes, 1993; Mendes, 1997; Mendes & Kell, 1998; Mendes, 2000). Each step in the pathway was simulated as depicted in Figure 3.1 (Mendes, 1999). To ensure that identical steady states were not replicated throughout each data set, parameters for the steady state reaction steps (*e.g.* K_{cat} in a Michaelis-Menten equation) were drawn from normal distributions. Specific details of the simulation model can be found in Mendes (1999).
Single branched metabolic pathways consisting of six genes and seven metabolites each were simulated for the activation and inhibition scenarios for mRNA transcription depicted in Figures 3.2 and 3.3. Genes were identified as shown in each figure.

C. Error Simulation

One hundred data sets of one hundred observation vectors each were simulated and subsequently analyzed using maximum likelihood FA. The data consisted of the mRNA concentrations output from Gepasi at steady state, with error added to the mRNA concentration values using the model (3.1) of Rocke and Durbin (2001).

$$y^*_{ij} = y_{ij}e^{\eta_j} + \varepsilon_{ij} \tag{3.1}$$

In this model, y_{ij} is the mRNA concentration of gene j in observation i without error, y^*_{ij} is the mRNA expression intensity value of gene j in observation i , $\eta_{ij} \sim N\left(0, 0.05\sigma^2_{\ln y}\right)$, $\varepsilon_{ij} \sim N\left(0, 0.20\sigma^2_y\right)$, and $\sigma^2_{\ln y}$ and σ^2_y are $\sum_{i=1}^p \frac{Var(\ln y_i)}{p}$ and $\sum_{i=1}^p \frac{Var(y_i)}{p}$, respectively.

The results of FA are reported as the average aligned value of each factor loading over the one hundred replicates along with empirical standard errors. Alignment of factors with the alignment by Clarkson algorithm (Clarkson, 1979) insures that both all retained factors are consistently in the same order and that all large factor loading entries are consistent in sign.

Discussion

A. Number of Factors Retained

In Henderson *et al.* (2002), the number of retained factors was selected using a strict adherence to the "Broken Stick" criterion (Jackson, 1993). Here, we also use the "Broken Stick" criterion, but we retain one additional factor, or several additional factors with nearly equal eigenvalues. Eigenvalues pertain to the sample correlation matrix and a single additional, or several additional, factors are chosen based on size of eigenvalue. While the first factor results from the constraint that exists for all members of a pathway (Kacser & Burns, 1973; Heinrich & Rapoport, 1974; Hofmeyr & Westerhoff, 2001), there are local constraints which are picked up by either the additional factor, or several additional factors with nearly equal eigenvalues. The additional factor(s) should provide information about dependencies within the data, which arise from the regulatory links between the genes, *i.e.* the transcriptional activation and inhibition scenarios seen in the figures. Factors associated with the remaining smaller eigenvalues contribute little to the explanation of regulatory information and the loadings on these factors are poorly estimated. Additionally, rotation including these factors often obscures interpretation.

As an example, the "broken stick" critical value to include the second factor with six variables in the model is 1.450. The eigenvalue associated with the second factor for the data set simulated using Figure 3.2 is 0.4284, so this factor is kept as the additional factor since it is the first factor not to exceed the "broken stick" critical value. The eigenvalue for the third factor is clearly smaller than the eigenvalue associated with the second, therefore the third factor is not retained.

B. Factor Loading Patterns and Metabolic Control Analysis

The first retained factor shown in Tables 3.2 and 3.3, which represents the correlation between factor (pathway) 1 and each gene, details the contribution to the control of the flux (see Appendix) of the pathway for each gene in the pathway. In Henderson *et al.* (2002), the identification of independent pathways was through this association between the contribution of each gene to the control of the flux of the pathway. The flux control summation theorem (Kacser & Burns, 1973) shows that the sum of the standardized flux control coefficients for a single metabolic pathway sums to 1.0, indicating the control is shared by all genes within the pathway (see Appendix).
Metabolic control analysis is used in the appendix to derive a matrix of response coefficients specific to the pathway in Figure 3.2. The pattern of the first two (or more) columns of the response matrix reflects the pattern of the factor loadings for the first two (or more) retained factors. Knowledge of the pathway structure (*i.e.* stem and branches connected by metabolites as seen in Figure 3.2) and the activation scheme (depicted by arrows with a plus sign in Figure 3.2) is required to derive the response matrix under the assumption of product inhibition (depicted by arrows with a minus sign in Figure 3.2). Under the additional assumption of approximate equality of the response in reaction rate at the steps associated with genes G_5 and G_6 to perturbations in substrate concentrations for metabolites S_0 and S_3, the pattern seen in factor 2 (Table 3.2) can be reproduced using metabolic control analysis (see Appendix). The assumption of approximate equality of the responses in reaction rates is met under our simulation model.
In Table 3.3, only gene G_6 has a large loading on the second factor and illustrates an important point. The activation and inhibition of transcription scenario in Figure 3.3 differs from that in Figure 3.2, even though the branched pathway structure is identical. Thus, for each new scenario, a new response matrix must and can be derived, which reflects the factor loading patterns (data not shown).

Conclusions

Patterns in the factor loadings from a FA of gene expression data from a metabolic pathway can be explained using metabolic control analysis. This supports previous findings where FA is shown to group genes by metabolic pathway, *i.e.* identify the observed constituent genes within a metabolic pathway. While the factor loading patterns can be explained using metabolic control analysis given knowledge of the pathway structure and activation/inhibition scheme, the latter cannot be inferred solely from the factor loading patterns. If some external knowledge of pathway structure and regulation is available, a FA in conjunction with metabolic control analysis may aid in identifying regulatory relationships between genes within a single pathway.

References

Clarkson, D. B. (1979) Estimating the standard errors of rotated factor loadings by jackknifing. *Psychometrika,* **44** (3), 297–314.

Heinrich, R. & Rapoport, T. A. (1974) A linear steady state treatment of enzyme chains: general properties, control, and effector strength. *Eur. J. Biochem.,* **42**, 89–95.

Henderson, D. A., Mendes, P., de la Fuente, A. & Hoeschele, I. (2002) Factor analysis for the identification of metabolic pathways from microarray expression data. *Bioinformatics,* **(Submitted)**.

Hofmeyr, J.-H. S. & Cornish-Bowden, A. (1996) Co-response analysis: a new experimental strategy for metabolic control analysis. *J. Theor. Biol.,* **182**, 371–380.

Hofmeyr, J.-H. S. & Westerhoff, H. V. (2001) Building the cellular puzzle. control in multi-level reaction networks. *J. Theor. Biol.,* **208**, 261–285.

Hofmeyr, J. S., Cornish-Bowden, A. & Rohwer, J. M. (1993) Taking enzyme kinetics out of control; putting control into regulation. *Eur. J. Biochem.,* **212**, 833–837.

Jackson, D. A. (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology,* **74**, 2204–2214.

Johnson, R. A. & Wichern, D. W. (1998) *Applied multivariate statistical analysis*. 4$^{\underline{\text{th}}}$ edition,, Prentice Hall, Englewood Cliffs, NJ, USA.

Kacser, H. & Burns, J. A. (1973) The control of flux. *Symp. Soc. Exp. Biol.,* **32**, 65–104.

Mendes, P. (1993) Gepasi: a software package for modeling the dynamics, steady states, and control of biochemical and other systems. *Comp. Appl. Biosci.,* **9**, 563–571.

Mendes, P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.,* **22**, 361–363.

Mendes, P. (1999) Metabolic simulation as an aide in understanding gene expression data. In *Proc. Workshop Comp. Biochem. Path. And Genetic Networks* pp. 27–33, Berlin, Germany.

Mendes, P. (2000) Gepasi: numerical simulation and optimization of biochemical kinetics. In *Proc. Plant and Anim. Genome VIII*, San Diego, CA, USA.

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

Rocke, D. M. & Durbin, B. (2001) A model for measurement error for gene expression arrays. *J. Comp. Biol.,* **In Press**.

Sen, A. K. (1996) On the sign pattern of metabolic control coefficients. *J. Theor. Biol.,* **182**, 269–275.

Appendix

In metabolic control analysis (Kacser & Burns, 1973; Heinrich & Rapoport, 1974), the relationships between the variables in a system, or pathway, can be described by the elasticities and control coefficients of the system (Hofmeyr *et al.*, 1993). The variables of a system consist of the reaction rates and substrate concentrations of the system constituents. The terms flux, elasticity, control coefficient, and response coefficient are defined below.

Flux The steady state reaction rate. A flux J may be defined in terms of a reaction step, *e.g.* $v_i = J_i$ where $i = 1 \ldots n$; a segment, *e.g.* $v_i = v_j = J_a$; or in terms of the entire pathway, *e.g.*

$J_A + J_B + \cdots + J_I = J_Z$ where J_Z is the flux through the stem and fluxes J_A through J_I are the fluxes through each branch in a branched metabolic pathway.

Elasticity The relative change in rate of a step in a system caused by a perturbation of a single metabolite at a constant concentration of all other metabolites. Formally for a substrate s_j and local rate v_i: $\varepsilon^{v_i}_{s_j} = \frac{\partial \ln v'_i}{\partial \ln s_j}$.

Control Coefficient The relative change in a system variable caused by a specific modulation of an enzyme at steady state. Formally for a system variable x and a rate variable v_i: $C^x_{v_i} = \frac{\partial \ln |x|}{\partial \ln v_i}$.

Response Coefficient The response in system variable y_k to a change in the system variable x_i associated with step e_j at steady state. Formally: ${}^{e_j}R^{y_k}_{x_i} = \varepsilon^{e_j}_{x_i} C^{y_k}_{e_j}$. y_k can be either a substrate or a flux.

The relationship between the $\varepsilon^{v_i}_{s_j}$ and $C^x_{v_i}$ terms and their ability to describe a metabolic system can be seen in the matrix formulation of Hofmeyr and Cornish-Bowden (1996), $\mathbf{CE} = \mathbf{I}$ where $\mathbf{C}$ is a matrix of control coefficients, $\mathbf{E}$ a matrix of elasticities, and $\mathbf{I}$ an identity matrix. The rows of $\mathbf{C}$ refer to either fluxes or metabolite concentrations while the columns refer to enzymes. In $\mathbf{E}$, the rows refer to enzymes while the columns refer to either fluxes or metabolite concentrations. While generally metabolic control analysis considers the metabolites S_0 and S_2 in Figure 3.2 as external, and therefore as parameters, we consider them as internal with the rest of the upstream pathway not observed. Additionally, since in our simulation model the product of translation is a functional protein, we can assume that the concentration of mRNA is proportional to the reaction rate. The expanded form of the matrices $\mathbf{C}$ and $\mathbf{E}$ can be seen in equation (3.2) for the structure pictured in Figure 3.2 as parameterized by Sen (1996). Here $j = J_A/J_C$ and $J_A = J_1 = J_2$ represent the flux through the branch containing S_0 and $J_C = J_5 = J_6$ represents the flux through the stem. There are a total of three fluxes through the system in Figure 3.2: $J_A + J_B$, $J_A + J_C$, and $J_B + J_C$, where $J_B = J_3 = J_4$ is the flux through the branch containing S_2. In this structure $J_B/J_C = j - 1$. This set of equations demonstrates two important theorems in metabolic control analysis, the flux control and the connectivity flux control summation theorems.

Flux Control Summation Theorem

$$\sum_{i=1}^{n} C^J_i = 1$$

$$\mathbf{CE} = \begin{bmatrix} C^{J_C}_{E_1} & C^{J_C}_{E_2} & C^{J_C}_{E_3} & C^{J_C}_{E_4} & C^{J_C}_{E_5} & C^{J_C}_{E_6} \\ -C^{S_0}_{E_1} & -C^{S_0}_{E_2} & -C^{S_0}_{E_3} & -C^{S_0}_{E_4} & -C^{S_0}_{E_5} & -C^{S_0}_{E_6} \\ -C^{S_2}_{E_1} & -C^{S_2}_{E_2} & -C^{S_2}_{E_3} & -C^{S_2}_{E_4} & -C^{S_2}_{E_5} & -C^{S_2}_{E_6} \\ -C^{j}_{E_1} & -C^{j}_{E_2} & -C^{j}_{E_3} & -C^{j}_{E_4} & -C^{j}_{E_5} & -C^{j}_{E_6} \\ -C^{S_4}_{E_1} & -C^{S_4}_{E_2} & -C^{S_4}_{E_3} & -C^{S_4}_{E_4} & -C^{S_4}_{E_5} & -C^{S_4}_{E_6} \\ -C^{S_5}_{E_1} & -C^{S_5}_{E_2} & -C^{S_5}_{E_3} & -C^{S_5}_{E_4} & -C^{S_5}_{E_5} & -C^{S_5}_{E_6} \end{bmatrix} \begin{bmatrix} 1 & -\varepsilon^{E_1}_{S_0} & 0 & j & 0 & 0 \\ 1 & \varepsilon^{E_2}_{S_0} & 0 & j & -\varepsilon^{E_2}_{S_4} & 0 \\ 1 & 0 & -\varepsilon^{E_3}_{S_2} & j-1 & 0 & 0 \\ 1 & 0 & \varepsilon^{E_4}_{S_2} & j-1 & -\varepsilon^{E_4}_{S_4} & 0 \\ 1 & \varepsilon^{E_5}_{S_0} & \varepsilon^{E_5}_{S_2} & 0 & \varepsilon^{E_5}_{S_4} & -\varepsilon^{E_5}_{S_5} \\ 1 & \varepsilon^{E_6}_{S_0} & \varepsilon^{E_6}_{S_2} & 0 & 0 & \varepsilon^{E_6}_{S_5} \end{bmatrix}$$

$$= \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad [3.2]$$

Connectivity Theorem

$$\sum_{i=1}^{n} \varepsilon^{i}_{X} C^{J}_{i} = \sum_{i=1}^{n} {}^{i}R^{J}_{X} = 0$$

These theorems state that the total flux control is complete and that the flux control is distributed by the elasticity coefficients. The Flux Control Summation Theorem is demonstrated in the multiplication of the first row of **C** with the first column of **E**. Multiplying the first row of **C** with the second column of **E** produces equation (3.3).

$$-{}^{E_1}R^{J_C}_{S_0} + {}^{E_2}R^{J_C}_{S_0} + {}^{E_5}R^{J_C}_{S_0} + {}^{E_6}R^{J_C}_{S_0} = 0 \qquad (3.3)$$

Multiplying the first row of **C** with the third column of **E** produces equation (3.4).

$$-{}^{E_3}R^{J_C}_{S_2} + {}^{E_4}R^{J_C}_{S_2} + {}^{E_5}R^{J_C}_{S_2} + {}^{E_6}R^{J_C}_{S_2} = 0 \qquad (3.4)$$

If the products $C^{J_C}_{E_5}\varepsilon^{E_5}_{S_1}$ and $C^{J_C}_{E_6}\varepsilon^{E_6}_{S_1}$ are approximately equal to the products $C^{J_C}_{E_5}\varepsilon^{E_5}_{S_3}$ and $C^{J_C}_{E_6}\varepsilon^{E_6}_{S_3}$, then equation (3.5) results.

$$-{}^{E_1}R^{J_C}_{S_0} + {}^{E_2}R^{J_C}_{S_0} \approx -{}^{E_3}R^{J_C}_{S_2} + {}^{E_4}R^{J_C}_{S_2} \qquad (3.5)$$

This approximate equality is the case for our simulated data and the above equation approximates factor 2 in Table 3.3 in both the sign and magnitude of the factor loadings. The first two columns of the resulting response matrix following the multiplications just described is seen in equation (3.6). The pattern of nonzero elements in the first two columns of the response matrix is identical to the pattern of nonzero elements in the first two or more columns of the factor loading matrix. This is illustrated in the example response matrix **R** below for the scenario in Figure 3.2 whose factor loadings are in Table 3.2.

$$\mathbf{R} = \begin{bmatrix} C^{J_C}_{E_1} & -{}^{E_1}R^{J_C}_{S_0} & \dots \\ C^{J_C}_{E_2} & {}^{E_2}R^{J_C}_{S_0} & \dots \\ C^{J_C}_{E_3} & -{}^{E_3}R^{J_C}_{S_2} & \dots \\ C^{J_C}_{E_4} & {}^{E_4}R^{J_C}_{S_2} & \dots \\ C^{J_C}_{E_5} & 0 & \dots \\ C^{J_C}_{E_6} & 0 & \dots \end{bmatrix} \tag{3.6}$$

Similar results can be obtained for the factor loading patterns of the retained factors for the other scenarios examined here and in Henderson *et al.* (2002), including looped pathway structures (data not shown).

Table 3.1. Average eigenvalues of correlation matrices.

Figure	Eigenvalue Number	Eigenvalue	Standard Error	Proportion of Total Variance
	1	4.5661	$\pm$ 0.0091	76.1018
3.2	2	0.4284	$\pm$ 0.0045	7.1404
	3	0.3248	$\pm$ 0.0030	5.4139
	1	4.5690	$\pm$ 0.0095	76.1505
3.3	2	0.4239	$\pm$ 0.0050	7.0655
	3	0.3369	$\pm$ 0.0033	5.6149

Table 3.2. Average of 100 samples of absolute value of Maximum Likelihood Factor loadings for pathways with transcriptional activation and inhibition scenario as shown in Figure 3.2.

	Rotated Factor Loadings			
Gene	Factor 1	Factor 2	Communality	Specific Variance
G_1	-0.8442 $\pm$ 0.0027	-0.1449 $\pm$ 0.0112	0.7467 $\pm$ 0.0061	0.1622 $\pm$ 0.0041
G_2	0.8221 $\pm$ 0.0030	-0.2089 $\pm$ 0.0154	0.7439 $\pm$ 0.0093	0.1600 $\pm$ 0.0055
G_3	-0.8481 $\pm$ 0.0024	-0.2003 $\pm$ 0.0151	0.7825 $\pm$ 0.0077	0.1465 $\pm$ 0.0039
G_4	0.8159 $\pm$ 0.0028	-0.2097 $\pm$ 0.0168	0.7385 $\pm$ 0.0091	0.1555 $\pm$ 0.0054
G_5	0.8668 $\pm$ 0.0025	-0.0164 $\pm$ 0.0079	0.7584 $\pm$ 0.0047	0.1895 $\pm$ 0.0030
G_6	0.8634 $\pm$ 0.0022	0.0110 $\pm$ 0.0081	0.7527 $\pm$ 0.0046	0.1918 $\pm$ 0.0028

Table 3.3. Average of 100 samples of absolute value of Maximum Likelihood Factor loadings for pathways with transcriptional activation and inhibition scenario as shown in Figure 3.3.

	Rotated Factor Loadings			
Gene	Factor 1	Factor 2	Communality	Specific Variance
G_1	-0.8459 ± 0.0028	-0.0286 ± 0.0146	0.7382 ± 0.0064	0.1851 ± 0.0044
G_2	-0.8878 ± 0.0023	-0.0383 ± 0.0080	0.7966 ± 0.0048	0.1660 ± 0.0029
G_3	-0.8547 ± 0.0023	-0.0040 ± 0.0089	0.7390 ± 0.0045	0.1938 ± 0.0032
G_4	-0.8825 ± 0.0023	-0.0424 ± 0.0083	0.7880 ± 0.0049	0.1706 ± 0.0028
G_5	0.8127 ± 0.0027	-0.0654 ± 0.0171	0.6944 ± 0.0076	0.2015 ± 0.0062
G_6	0.7722 ± 0.0042	-0.4244 ± 0.0250	0.8403 ± 0.0146	0.0901 ± 0.0069

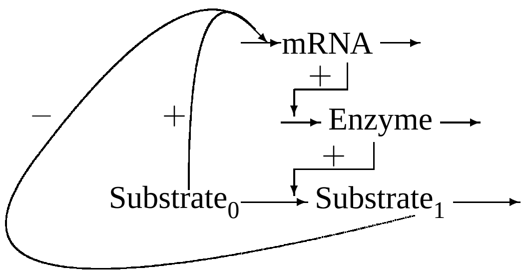

Figure 3.1. Hierarchy of simulated pathway structure.

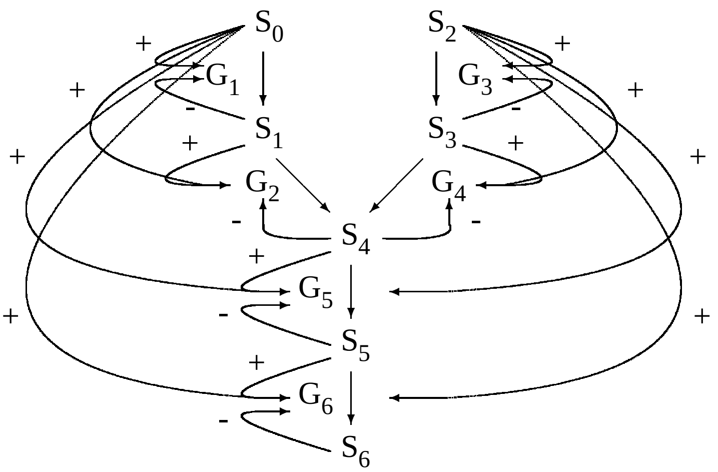

Figure 3.2. Simulated pathway with six genes. Structure with regulatory interactions.

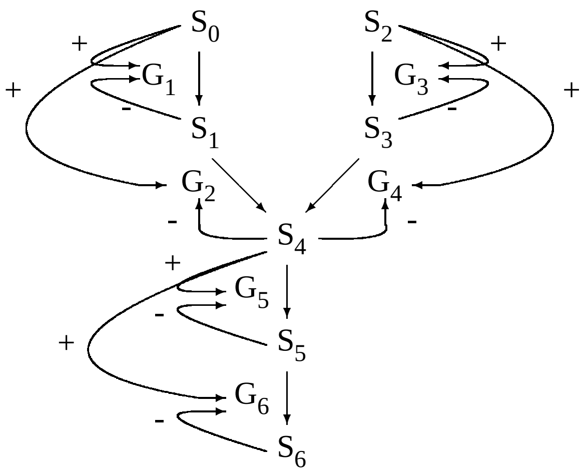

Figure 3.3. Simulated pathway with six genes. Structure with regulatory interactions.

Chapter 4

Bayesian and Correlated Exploratory Bayesian Factor Analysis

Abstract

Motivation: We have shown in a previous communication that the factors from a maximum likelihood factor analysis of gene expression data identify the metabolic pathways of which the measured genes are constituents for some activation and inhibition of transcription scenarios. Maximum likelihood estimation, however, relies on large sample approximation or requires bootstrapping to obtain confidence intervals. The development of an exploratory Bayesian factor analysis may yield highest posterior density regions with better coverage probabilities than the bootstrap confidence intervals in the maximum likelihood analysis. Also, many microarray gene expression experiments are time series and analysis methods need to account for the additional correlation that exists between observation vectors. A correlated Bayesian factor analysis using a first order autoregressive correlation structure could prove useful in the analysis of time series data. The identification of genes belonging to independent metabolic pathways could be useful in molecular biology, plant and animal genetic improvement programs, and cancer research.

Results: A Bayesian exploratory factor analysis can be applied to simulated steady state microarray gene expression data. Our Bayesian implementation of exploratory factor analysis separates genes into independent metabolic pathways, reproducing groups of genes previously reported for maximum likelihood factor analysis. The use of vague or uninformative priors differs from a previously developed confirmatory Bayesian factor analysis and allows our implementation of Bayesian exploratory factor analysis to be applied to current gene expression data where prior knowledge of the relationships between genes is not available. However, where limited knowledge of the relationships between genes within a metabolic pathway is available, this information can be incorporated into the prior information on the factor loadings. 95% highest posterior density regions can be calculated from the posterior distribution and aid in the interpretation of the factor loadings, specifically the allocation of genes to independent metabolic pathways. The present implementation of a correlated Bayesian factor analysis with first order autoregressive correlation structure, however, does not group genes by metabolic pathway for simulated time series data.

Introduction

Previously, Henderson *et al.* (2002) presented an application of maximum likelihood factor analysis (MLFA) to microarray gene expression data for the identification of metabolic pathways, where bootstrap confidence intervals for individual factor loadings were used to aid in the interpretation of the factor loadings. Due to the lack of uniqueness of the factor loading matrix, estimation of factor loadings is not a 'regular' statistical problem and bootstrap confidence intervals may be too wide and hence include zero where they should not. Here, we present a Bayesian exploratory factor analysis (BEFA) implemented with a Gibbs sampler which provides highest posterior density regions as an alternative method for obtaining confidence intervals on factor loadings.

Press and Shigemasu (1997) developed a Bayesian confirmatory factor analysis implemented with a Gibbs sampler. Confirmatory factor analysis (CFA) investigates the support by the data for a (partially) prespecified factor loading pattern. CFA therefore requires (substantial) prior knowledge, but incorporation of this prior knowledge can make the factor loading matrix unique and thus avoid convergence problems. The model of Press and Shigemasu (1997) can be extended to exploratory factor analysis, but suffers from a lack of identifiability of the factor loading matrix since restrictions on the loadings are not utilized. Here, we present an exploratory Bayesian FA which does not require any prior knowledge, but can incorporate prior knowledge, if available. Rotation and alignment of the sampled factor loadings ensures the uniqueness of the factor loading matrix, and thus convergence of the Gibbs sampler.

Rowe (1998a) extended the Press and Shigemasu model to accomodate correlated observation vectors using compound symmetry, autoregressive, and free (co)variance structures. Again, this Bayesian factor analysis model was used in a confirmatory manner, and here we present an extention of our BEFA model to correlated exploratory Bayesian factor analysis (CEBFA). The CEBFA model is applied to time series microarray gene expression data.

For both the BEFA and CEBFA models, we use Bayesian highest posterior density (HPD) regions to aid in the interpretation of the factor loadings. The HPD region is the Bayesian analogue of the confidence interval and differs in interpretation in that for a given α, a $100(1-\alpha)\%$ HPD region is a region containing the unknown value with belief probability $(1-\alpha)$ conditional on the observed data; in contrast to the classical coverage probability, where the random confidence interval contains the fixed true parameter value in $100(1-\alpha)\%$ of n repeated experiments.

Materials and Methods

A. Factor Analysis

Factor analysis (Johnson & Wichern, 1998) is a method to describe the covariance, or correlation, relationships among many variables in a multivariate data set in terms of a few underlying, but unobservable, random quantities called factors. Inferences about the relationships between the observed variables are made through the magnitude, and typically the sign, of the cell entries in the factor loading matrix. Given a multivariate data observation $\mathbf{y}_i$ and an unobserved factor vector $\mathbf{f}_i$, the definition of the factor loading matrix is either $Cov\left(\mathbf{y}_i, \mathbf{f}_i'\right)$ or $Cor\left(\mathbf{y}_i, \mathbf{f}_i'\right)$, depending on whether or not the data have been standardized. The number of factors m is specified based upon either subjective (*e.g.* a "Scree" plot (Johnson & Wichern, 1998)) or objective (*e.g.* "Broken Stick" (Jackson, 1993)) criteria and $m \leq p$ with p the number of variables. The portion of the total variance explained by the retained factors for a specific variable is termed the communality, and the portion not explained by the retained factors is termed the specific variance.
Equation (4.1) shows the general factor model which is linear in the factor scores $\mathbf{f}_i$.

$$\mathbf{y}_i - \mu = \mathbf{L}\mathbf{f_i} + \varepsilon \tag{4.1}$$

In this model, $\mathbf{y}_i$ represents a $p \times 1$ multivariate observation vector on experimental unit i, μ is a $p \times 1$ vector of means, $\mathbf{L}$ is a $p \times m$ unknown matrix of factor loadings, $\mathbf{f}_i$ a $m \times 1$ vector of the effects of the common factors for experimental unit i, and ε_i is a $p \times 1$ vector of specific factors, or residuals. The vectors $\mathbf{f}_i$ and ε_i are generally not observed and assumed independent. Furthermore, $\mathbf{f}_i$ is normally distributed with mean vector zero and an identity (co)variance matrix, or $\mathbf{f}_i \sim \text{MVN}(\mathbf{0}, \mathbf{I}_m)$. For application to gene expression data, $\mathbf{y}_i$ is the vector of expression values for p genes in sample i ($i = 1, \ldots, n$). The covariance matrix of $\mathbf{y}_i$ is then defined in equation (4.2).

$$Cov\left(\mathbf{y}_i\right) = \mathbf{L}\mathbf{L}' + \Psi \tag{4.2}$$

Here, the $p \times p$ matrix Ψ is a diagonal matrix of specific, or residual, variances (Johnson & Wichern, 1998). The covariance matrix of the vector $\mathbf{y}_i$ is invariant to rotation of the factor loadings, $\mathbf{L}$, and the factor scores, $\mathbf{f}_i$, by an orthogonal matrix. This causes a problem with the identifiability of $\mathbf{L}$ and $\mathbf{f}_i$ since an infinite number of solutions for $\mathbf{L}$ and $\mathbf{f}$ are possible, each related to the other by an orthogonal rotation matrix.

B. Maximum Likelihood

Maximum likelihood factor analysis (MLFA) estimates of $\mathbf{L}$ and Ψ are obtained from the likelihood in (4.3) given that $\Sigma = \mathbf{L}\mathbf{L}' + \Psi$, where Σ is $Cov(\mathbf{y})$.

$$L(\mu, \Sigma) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(\mathbf{y}-\bar{\mathbf{y}})(\mathbf{y}-\bar{\mathbf{y}})'+n(\bar{\mathbf{y}}-\mu)(\bar{\mathbf{y}}-\mu)'\right)\right]} \tag{4.3}$$

Maximum likelihood estimates of specific variances and factor loadings were obtained using an expectation / maximization (EM) algorithm (Rubin & Thayer, 1982) with principal components factor analysis estimates as starting values. Since restrictions on the estimation of factor loadings are not used in the EM algorithm, identifiability of factor loadings is obtained through starting values of factor loading estimates subject to the constraint $\mathbf{L}'\Psi^{-1}\mathbf{L} = \Delta$ and Δ a diagonal matrix. The MLFA method can also produce Heywood cases (*e.g.* estimates of specific variances less than 0). Traditional means to deal with Heywood cases typically include setting the estimate of that specific variance to zero for the next round of iterations. The search for a method that would remain within the parameter space for the specific variances partly led to the development of the Bayesian methods described in section D. below, in addition to the aforementioned difficulties with the construction of confidence intervals.

C. Confirmatory Bayesian Factor Analysis - Press and Shigemasu Model

Assuming that the data are normally distributed and are mean centered, *i.e.* $E\left[\mathbf{y}_i\right] = \mathbf{0}$, the conditional density of the data $\mathbf{Y}$ given $\mathbf{L}$, $\mathbf{F}$, and Ψ is proportional to (4.4) (Press & Shigemasu, 1997),

$$p(\mathbf{Y}|\mathbf{L},\mathbf{F},\Psi) \propto |\Psi|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\left(\Psi^{-1}(\mathbf{Y}-\mathbf{F}\mathbf{L}')'(\mathbf{Y}-\mathbf{F}\mathbf{L}')\right)} \tag{4.4}$$

where $\mathbf{Y}$ is a $n \times p$ matrix of n multivariate observation vectors $\mathbf{y}_i$ and $\mathbf{F}$ is a $n \times m$ matrix of n factor score vectors $\mathbf{f}_i$. From Press and Shigemasu (1997), the joint prior distribution for $\mathbf{L}$, $\mathbf{F}$, and Ψ is:

$$p(\mathbf{L},\mathbf{F},\Psi) \propto p(\mathbf{L}|\Psi)\, p(\Psi)\, p(\mathbf{F}) \tag{4.5}$$

where

$$p(\mathbf{L}|\Psi) \propto |\Psi|^{-\frac{m}{2}} e^{-\frac{1}{2}tr\left(\Psi^{-1}(\mathbf{L}-\mathbf{L_0})\mathbf{H}(\mathbf{L}-\mathbf{L_0})'\right)} \tag{4.6}$$

$$p(\Psi) \propto |\Psi|^{-\frac{\nu}{2}} e^{-\frac{1}{2}tr\left(\Psi^{-1}\mathbf{B}\right)} \tag{4.7}$$

$$p(\mathbf{F}) \propto e^{-\frac{1}{2}tr(\mathbf{F}'\mathbf{F})} \tag{4.8}$$

The diagonal prior precision matrix $\mathbf{H}$ and (co)variance matrix $\mathbf{B}$ are both positive definite matrices as is the matrix of specific variances Ψ. The prior distribution for the (co)variance matrix Ψ is Inverse Wishart with parameters ν and $\mathbf{B}$, $p(\Psi) \sim \mathrm{IW}(\nu, \mathbf{B})$ and Ψ is assumed diagonal on average. The conditional prior distribution for the matrix of factor loadings $\mathbf{L}$ is multivariate normal with parameter vector $\mathbf{l}_o$ and $m \times m$ matrix $\mathbf{H}$, $p(\mathbf{l}|\Psi) \sim \mathrm{MVN}\left(\mathbf{l}_o, \mathbf{H}^{-1} \otimes \Psi\right)$ where $\mathbf{l}$ is $vec(\mathbf{L})$ and $\otimes$ denotes the Kroneckor product. The prior distribution for the matrix of factor scores $\mathbf{F}$ is equivalent to that of the product of standard normal densitites for each factor score vector $\mathbf{f}_i$. This gives the joint posterior in (4.9) and the conditional sampling distributions in (4.10), (4.11), and (4.12)

$$p(\mathbf{L}, \mathbf{F}, \Psi|\mathbf{Y}) \propto e^{-\frac{1}{2}tr\mathbf{F}'\mathbf{F}} |\Psi|^{-\frac{n+m+\nu}{2}} e^{-\frac{1}{2}tr\left(\Psi^{-1}\left[(\mathbf{Y}-\mathbf{F}\mathbf{L}')'(\mathbf{Y}-\mathbf{F}\mathbf{L}')+(\mathbf{L}-\mathbf{L_o})\mathbf{H}(\mathbf{L}-\mathbf{L_o})'+\mathbf{B}\right]\right)} \tag{4.9}$$

$$p(\mathbf{L}|\mathbf{F}, \Psi, \mathbf{Y}) \propto e^{-\frac{1}{2}tr\left(\Psi^{-1}\left(\mathbf{L}-\tilde{\mathbf{L}}\right)(\mathbf{H}+\mathbf{F}'\mathbf{F})\left(\mathbf{L}-\tilde{\mathbf{L}}\right)'\right)} \tag{4.10}$$

$$p(\Psi|\mathbf{L}, \mathbf{F}, \mathbf{Y}) \propto |\Psi|^{-\frac{n+m+\nu}{2}} e^{-\frac{1}{2}tr\left(\Psi^{-1}\left[(\mathbf{Y}-\mathbf{F}\mathbf{L}')'(\mathbf{Y}-\mathbf{F}\mathbf{L}')+(\mathbf{L}-\mathbf{L_o})\mathbf{H}(\mathbf{L}-\mathbf{L_o})'+\mathbf{B}\right]\right)} \tag{4.11}$$

$$p(\mathbf{F}|\mathbf{L}, \Psi, \mathbf{Y}) \propto e^{-\frac{1}{2}tr\left(\left(\mathbf{F}-\tilde{\mathbf{F}}\right)\left(\mathbf{I}_m+\mathbf{L}'\Psi^{-1}\mathbf{L}\right)\left(\mathbf{F}-\tilde{\mathbf{F}}\right)'\right)} \tag{4.12}$$

where

$$\tilde{\mathbf{L}} = \left(\mathbf{Y}'\mathbf{F} + \mathbf{L}_o\mathbf{H}\right)\left(\mathbf{H} + \mathbf{F}'\mathbf{F}\right)^{-1}$$
$$\tilde{\mathbf{F}} = \mathbf{Y}\Psi^{-1}\mathbf{L}\left(\mathbf{I}_m + \mathbf{L}'\Psi^{-1}\mathbf{L}\right)^{-1}$$

While this model has previously been used in confirmatory factor analysis (Rowe & Press, 1998), it can be modified to perform exploratory factor analysis. This entails using vague prior distributions for the factor loadings and specific variances, rather than the sharp priors in Press and Shigemasu (1997). In the absence of restrictions on the factor loading matrix, the Gibbs sampler fails to converge. It can be shown that convergence of the sampler in Press and Shigemasu (1997) occurs from specifying sufficient nonzero factor loadings to make the factor loading matrix identifiable (Congdon, 2001) along with small prior variances. In fact, using the priors in Press and Shigemasu (1997), the posterior mean is completely specified through the prior, regardless of the information in the data.

D. Bayesian Exploratory Factor Analysis with Uncorrelated Residuals within Experimental Units

Consistent with the traditional implementation of factor analysis, we assume here that all residuals are uncorrelated within experimental units. In contrast, Press and Shigemasu (1997) sample $\mathbf{\Psi}$ as a full (co)variance matrix in (4.11), which has diagonal expectation ($\mathbf{B}$ is diagonal). A second difference between the confirmatory factor analysis of Press and Shigemasu and the Bayesian implementation described below is that we do not assume sharply informative priors for the factor loading matrix (see below and section F.) as generally no prior knowledge is available for exploratory factor analysis. The specific variances, or diagonal elements of $\mathbf{\Psi}$, are then sampled independently using an Inverse Gamma prior distribution, $\psi_{ii} \sim \mathrm{IG}(\alpha, \beta)$ for specific variance $i = 1 \ldots p$. The joint posterior is then given in (4.13)

$$p(\mathbf{L}, \mathbf{F}, \mathbf{\Psi} | \mathbf{Y}) \propto \prod_{i=1}^{p} \psi_{ii}^{-\frac{n+m}{2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{p} \psi_{ii}^{-1}\left[\left(\mathbf{y}_i - \mathbf{F}\mathbf{l}'_{\mathbf{i_c}}\right)'\left(\mathbf{y}_i - \mathbf{F}\mathbf{l}'_{\mathbf{i_c}}\right) + 2\beta\right] + tr(\mathbf{F}'\mathbf{F}) + tr((\mathbf{L}-\mathbf{L}_o)\mathbf{H}(\mathbf{L}-\mathbf{L}_o))\right]} \tag{4.13}$$

where $\mathbf{y}_i$ is redefined as column i of $\mathbf{Y}$ and $\mathbf{l}_{i_c}$ is column i of $\mathbf{L}$. Now, the diagonal elements of $\mathbf{\Psi}$ can be sampled independently from an Inverse Gamma distribution

$$\psi_{ii} \sim \mathrm{IG}\left(\frac{n+m}{2} + \alpha, \frac{1}{2}\left[\left(\mathbf{y}_i - \mathbf{F}\mathbf{l}_i\right)'\left(\mathbf{y}_i - \mathbf{F}\mathbf{l}_i\right) + \left(\mathbf{l}_i - \mathbf{l}_{o_i}\right)'\mathbf{H}\left(\mathbf{l}_i - \mathbf{l}_{o_i}\right)\right] + \beta\right) \tag{4.14}$$

where $\mathbf{l}_i$ is row i of $\mathbf{L}$.

Additionally, the factor loadings can be sampled independently for each variable (gene) which follows directly from equation (4.10) by replacing $\mathbf{\Psi}$ with a diagonal matrix, or

$$\mathbf{l}_i \sim \mathrm{MVN}\left(\tilde{\mathbf{l}}_i, \psi_{ii}\left(\mathbf{H} + \mathbf{F}'\mathbf{F}\right)^{-1}\right) \tag{4.15}$$

where $\tilde{\mathbf{l}}_i$ is row i of $\tilde{\mathbf{L}}$. The means $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{F}}$ of the conditional sampling densities are defined as in Press and Shigemasu (1997).

Inferences about $\mathbf{L}$ and $\mathbf{\Psi}$ are made from the marginal posterior densities which are Multivariate Normal given $\mathbf{F}$ and ψ_{ii} and Inverse Gamma given $\mathbf{F}$ and $\mathbf{L}$, respectively. Additionally, 95% highest posterior density (HPD) regions for the parameters in $\mathbf{F}$, $\mathbf{L}$, and $\mathbf{\Psi}$ can be obtained from the posterior samples (Chen & Shao, 1999).

While in MLFA the estimate of a specific variance can step outside of the parameter space, in BEFA, by simulating estimates of the specific variance from either an Inverse Wishart or Inverse

Gamma distribution, this problem cannot occur (Rowe, 1998). Some implementations of MLFA impose a (computationally convenient) uniqueness condition on $\mathbf{L}$ by forcing $\mathbf{L}'\Psi^{-1}\mathbf{L}$ to be diagonal, but this specific uniqueness condition is not necessary (Henderson *et al.*, 2002). Informative priors similar to those in confirmatory factor analysis (Press & Shigemasu, 1997; Rowe, 1998; Rowe & Press, 1998) were not used, where prior means distinctly different from zero were specified for some of the factor loadings along with small variance (see section F. below). Without restrictions and informative priors, the Gibbs sampler appears to fail to converge. A solution is to rotate and align, *e.g.* (Clarkson, 1979), each sampled factor loading matrix in each cycle following burn in. See Henderson *et al.* (2001) for a detailed description of factor alignment and rotation. We note that the rotated factor loading matrices are used only for convergence diagnostics and posterior inferences, while unrotated sampled factor loading matrices are used to sample the unknowns in the next cycle.

E. Correlated Exploratory Bayesian Factor Analysis

We derived our correlated exploratory Bayesian factor analysis (CEBFA) model using the correlated Bayesian factor analysis model of Rowe (1998a), but with the univariate sampling of the diagonal elements of Ψ. This model uses a likelihood proportional to

$$p\left(\mathbf{Y}|\mathbf{L},\mathbf{F},\Phi,\Psi\right) \propto \left|\mathbf{I}_t \otimes \Phi \otimes \Psi\right|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\left(\Psi^{-1}(\mathbf{Y}-\mathbf{F}\mathbf{L}')'(\mathbf{I}_t\otimes\Phi)^{-1}(\mathbf{Y}-\mathbf{F}\mathbf{L}')\right)} \tag{4.16}$$

where $\mathbf{I}_t$ is an identity matrix of size t, t is the number of replicates of the time series and Φ is a $q \times q$ correlation matrix for the q points in each time series structured as a first order autoregressive process. Note that $n = tq$. Prior distributions for the factor loadings and the specific variances are unaffected by the addition of the autoregressive parameter and are identical to those in the BEFA model. The prior distribution for the matrix of factor scores, $\mathbf{F}$, is now as given in (4.17), and the prior for ρ, the autoregressive parameter, is a scaled Beta distribution given in (4.18).

$$p\left(\mathbf{F}|\Phi\right) \propto \left|\mathbf{I}_t \otimes \Phi\right|^{-\frac{m}{2}} e^{-\frac{1}{2}tr\left(\mathbf{I}_t\otimes\Phi^{-1}\right)\mathbf{F}\mathbf{F}'} \tag{4.17}$$

$$p\left(\rho\right) \propto \left(\frac{1}{2}\left(1+\rho\right)\right)^{(a-1)} \left(\frac{1}{2}\left(1-\rho\right)\right)^{(b-1)} \frac{\Gamma(a)+\Gamma(b)}{\Gamma(a)\,\Gamma(b)} \tag{4.18}$$

This gives the joint posterior distribution in (4.19)

$$p\left(\mathbf{L},\mathbf{F},\mathbf{\Phi},\mathbf{\Psi}|\mathbf{Y}\right) \propto \left(\frac{1}{2}\left(1+\rho\right)\right)^{(a-1)} \left(\frac{1}{2}\left(1-\rho\right)\right)^{(b-1)} \frac{\Gamma\left(a\right)+\Gamma\left(b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} \left|\mathbf{I}_t \otimes \mathbf{\Phi}\right|^{-\frac{n}{2}} \prod_{i=1}^{p} \psi_{ii}^{-\frac{n+m}{2}}$$
$$e^{-\frac{1}{2}\left[\sum_{i=1}^{p} \psi_{ii}^{-1}\left[\left(\mathbf{y}_i-\mathbf{f_i}\mathbf{l}'_{\mathbf{i}*}\right)'\left(\mathbf{I}_t\otimes\mathbf{\Phi}\right)^{-1}\left(\mathbf{y}_i-\mathbf{f_i}\mathbf{l}'_{\mathbf{i}*}\right)+2\beta\right]+tr\left[\left(\mathbf{I}_t\otimes\mathbf{\Phi}^{-1}\right)\mathbf{F}\mathbf{F}'\right]+tr\left[\left(\mathbf{L}-\mathbf{L}_o\right)\mathbf{H}\left(\mathbf{L}-\mathbf{L}_o\right)\right]\right]} \quad [4.19]$$

The conditional posterior sampling densities are now given by equations (4.20) through (4.23).

$$\psi_{ii} \sim \mathrm{IG}\left(\frac{n+m}{2}+\alpha, \frac{1}{2}\left[\left(\mathbf{y}_i-\mathbf{F}\mathbf{l}_i\right)'\left(\mathbf{I}_t\otimes\mathbf{\Phi}^{-1}\right)\left(\mathbf{y}_i-\mathbf{F}\mathbf{l}_i\right)+\left(\mathbf{l}_i-\mathbf{l}_{o_i}\right)'\mathbf{H}\left(\mathbf{l}_i-\mathbf{l}_{o_i}\right)\right]+\beta\right) \tag{4.20}$$

$$\mathbf{l}_i \sim \mathrm{MVN}\left(\tilde{\mathbf{l}}_i, \psi_{ii}\left(\mathbf{H}+\mathbf{F}'\left(\mathbf{I}_t\otimes\mathbf{\Phi}^{-1}\right)\mathbf{F}\right)^{-1}\right) \tag{4.21}$$

$$\mathbf{F} \sim \mathrm{MVN}\left(\tilde{\mathbf{F}}, \left(\left(\mathbf{I}_m+\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)\otimes\mathbf{\Phi}^{-1}\right)^{-1}\right) \tag{4.22}$$

$$\rho \propto \left(\frac{1}{2}\left(1+\rho\right)\right)^{(a-1)} \left(\frac{1}{2}\left(1-\rho\right)\right)^{(b-1)} \frac{\Gamma\left(a\right)+\Gamma\left(b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} e^{-\frac{1}{2}\frac{k_1-k_2\rho+k_3\rho^2}{1-\rho^2}} \tag{4.23}$$

where

$$\tilde{\mathbf{L}} = \left(\mathbf{Y}'\left(\mathbf{I}_t\otimes\mathbf{\Phi}^{-1}\right)\mathbf{F}+\mathbf{L}_o\mathbf{H}\right)\left(\mathbf{H}+\mathbf{F}'\left(\mathbf{I}_t\otimes\mathbf{\Phi}^{-1}\right)\mathbf{F}\right)^{-1}$$
$$\tilde{\mathbf{F}} = \mathbf{Y}\mathbf{\Psi}^{-1}\mathbf{L}\left(\mathbf{I}_m+\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1}$$
$$k_1 = tr\left(\left(\mathbf{Y}-\mathbf{F}\mathbf{L}'\right)\mathbf{\Psi}^{-1}\left(\mathbf{Y}-\mathbf{F}\mathbf{L}'\right)'+\mathbf{F}\mathbf{F}'\right)$$
$$k_2 = tr\left(\begin{bmatrix} 0 & 1 & & & 0 \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 0 & 1 \\ 0 & & & 1 & 0 \end{bmatrix}\left(\mathbf{Y}-\mathbf{F}\mathbf{L}'\right)\mathbf{\Psi}^{-1}\left(\mathbf{Y}-\mathbf{F}\mathbf{L}'\right)'+\mathbf{F}\mathbf{F}'\right)$$
$$k_3 = tr\left(\begin{bmatrix} 0 & & & 0 \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ 0 & & & 0 \end{bmatrix}\left(\mathbf{Y}-\mathbf{F}\mathbf{L}'\right)\mathbf{\Psi}^{-1}\left(\mathbf{Y}-\mathbf{F}\mathbf{L}'\right)'+\mathbf{F}\mathbf{F}'\right)$$

The form for sampling the autoregressive parameter comes from the equation for the inverse of a first order autoregressive correlation matrix (Wade & Quaas, 1993) given in equation (4.24).

$$\Phi^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & & & 0 \\ -\rho & \left(1+\rho^2\right) & -\rho & & \\ & \ddots & \ddots & \ddots & \\ & & & \left(1+\rho^2\right) & -\rho \\ 0 & & & -\rho & 1 \end{bmatrix} \tag{4.24}$$

The conditional posterior distribution for ρ is not a standard distribution, so ρ is sampled using a univariate Metropolis-Hastings step with a uniform proposal distribution on the interval (-0.98, 0.98). Factor loading estimates are aligned and rotated as in BEFA.

F. Priors

Vague, or uninformative, prior distributions for both the factor loadings and the specific variances were specified. Specifically, the prior mean for each factor loading was zero, giving $\mathbf{L}_o = \mathbf{0}_{12\times 2}$, with a prior precision matrix $\mathbf{H} = 0.2\ \mathbf{I}_2$. This gives each factor a prior variance of 5 which is vague given that the factor loadings are should be between -1 and 1. The improper prior for the specific variances used hyperparameters α and β both equal to zero for each variable (gene) giving a prior distribution proportional to $\frac{1}{\psi_{ii}}$. In addition to the priors specified above, hyperparameters for the autocorrelation parameter ρ in CEBFA were $a = 0.0$ and $b = 0.0$ since we are drawing from a uniform proposal distribution.

G. Gibbs Sampling

The BEFA and CEBFA models presented above are implemented with a Gibbs sampler using the sampling distributions described earlier. The order of sampling is to sample row i of the matrix of factor loadings $\mathbf{L}$, followed by the specific variance ψ_{ii} for $i = 1 \dots p$, and then the matrix of factor scores $\mathbf{F}$ from their conditional distributions. The autoregressive correlation coefficient ρ in CEBFA is sampled last using a univariate Metropolis-Hastings step.

H. Data Simulation

Data were simulated for a hierarchical metabolic pathway (Hofmeyr & Westerhoff, 2001) using the biochemical simulation program Gepasi (Mendes, 1993; Mendes, 1997; Mendes & Kell, 1998; Mendes, 2000). Each step in a pathway was simulated as depicted in Figure 4.1 (Mendes, 1999). To ensure that identical steady states were not replicated throughout each data set, parameters for the steady state reaction steps (*e.g.* K_{cat} in a Michaelis-Menten equation) were drawn from normal distributions. Specific details of the simulation model can be found in Mendes (1999).

Steady State Data on Independent Metabolic Pathways

Two branched metabolic pathways consisting of six genes and seven metabolites each with activation and inhibition of mRNA transcription were simulated for the activation and inhibition scenarios depicted in Figures 4.2 through 4.5. Genes were ordered as shown in Figure 4.6. As seen in Figure 4.6, the two pathways do not share any genes, enzymes, or metabolites, *i.e.* they are completely independent. A total of 100 observation vectors were used in the analysis of each scenario.

Times Series Data

For the correlated BEFA, one branched metabolic pathway (Figure 4.2) and one looped metabolic pathway (Figure 4.5) were paired in order to obtain the maximum difference between the two pathways. These two pathways did not share any metabolites, enzymes, or genes in common and so are independent. Each pathway was initialized with high input substrates, such as plating yeast on high nutrient media, and allowed to reach an eventual oscillating steady state. Simulated expression intensity values for the genes in the branches or start of the loop started high and tapered down to a steady oscillating value, while simulated expression intensity values for genes in the stem or remaining part of the loop started low and increased to a different steady oscillating value. A total of twelve sequential samples were taken over a span of two hours, and four replicates of the time series were used in the analysis, reflecting a total of 48 observation vectors. The sequential time points were simulated using the time series function in Gepasi.

Error Simulation

The data consist of the mRNA concentrations output from Gepasi at steady state, with error added to the mRNA concentration values using the model (4.25) of Rocke and Durbin (2001).

$$y^*_{ij} = y_{ij}e^{\eta_j} + \varepsilon_{ij} \tag{4.25}$$

In this model, y_{ij} is the mRNA concentration of gene j in observation i without error, y^*_{ij} is the expression intensity value of gene j in observation i, $\eta_{ij} \sim N\left(0, 0.05\sigma^2_{\ln y}\right)$, and $\varepsilon_{ij} \sim N\left(0, 0.20\sigma^2_y\right)$. The statistics $\sigma^2_{\ln y}$ and σ^2_y are the mean variance of the $\log_e$ mRNA concentrations and mRNA concentrations, respectively.

Discussion

A. Convergence of the Gibbs Sampler

Based upon a standard trace plot of sample value versus cycle, the Gibbs sampler converged to the distribution of each factor loading and specific variance rather quickly. Convergence occured following at most 300 rounds of burn in. To be safe, 1000 rounds of burn in were used, and 6000 consecutive samples were retained for posterior inferences. This number of samples retained was sufficient to obtain effective sample sizes of at least 300 calculated from the autocorrelation structure of the samples and shown in Table 4.3. Computations took approximately 30 minutes for BEFA, including calculation of HPD regions, autocorrelations, and variances of estimates, on a dual Intel PIII 733MHz. Computational time for CEBFA was longer due to sampling $Vec(\mathbf{F})$ in a single multivariate step.

In Figure 4.9, histograms of posterior samples of two factor loadings and one specific variance are presented. The graphs closely represent normal densities, with the distribution of the specific variances slightly skewed, which is typical for variance parameters. A normal marginal posterior distribution on an unknown parameter indicates sufficient information in the data for inferences on this unknown.

B. Comparison of BEFA with Maximum Likelihood Estimates

Table 4.1 shows ML estimates of factor loadings for two independent pathways simulated under the scenario in Figure 4.2. An approximate 95% confidence interval is shown, calculated using the BC_a bootstrap method of Efron and Tibshirani (1993) with 2000 Bootstrap samples (see Henderson *et al.* (2002) for details on the bootstrap procedure). Table 4.2 shows BEFA factor loadings along with 95% HPD regions. The ML bootstrap intervals are, on average, smaller than the HPD regions in Table 4.2. The HPD regions responded to changes in the prior variance of the factor loadings. The resulting regions thus reflect both the information in the data and our degree of belief in the prior means for the factor loadings. The BEFA estimates of the factor loadings are also generally larger and can step outside of the parameter space, which also results from our large prior variance for the retained factors (see Figure 4.9). This could be solved by truncating the prior distribution. Table 4.3 shows estimates of the variances and effective sample sizes for the BEFA factor loading and specific variance estimates calculated from the posterior distribution. The low variances indicate fairly precise posterior inferences.

In a previous communication (Henderson *et al.*, 2002), we found data simulated according to the scenario in Figure 4.3 difficult to interpret using MLFA and used 95% bootstrap confidence intervals to aid in interpretation. As mentioned in the introduction, the bootstrap method for confidence intervals used in Henderson *et al.* (2002) accounts for the bias that can produce wider intervals. The degree of bias in our posterior estimates from BEFA and the degree of bias in our HPD regions is proportional to the degree of influence the prior has over the information in the data. This influence decreases with both the amount of information in the data and the level of informativeness of the prior, hence we used vague prior information.

Table 4.4 show BEFA factor loadings along with 95% HPD regions for the data set used in Table 5 of Henderson *et al.* (2002). As in our previous communication, genes G_1, G_5, and G_{12} are difficult to assign to a pathway. However, using the BEFA 95% HPD regions in Table 4.4, we can see that the loading of gene G_1 (G_{12}) is much higher on factor 1 (factor 2) than on factor 2 (factor 1) and we can then assign gene G_1 (G_{12}) to pathway 1 (pathway 2). This result was comparable only for gene G_{12} in the bootstrap intervals used in Henderson *et al.* (2002). The use of BEFA 95% HPD regions does not clarify the placement of gene G_5 on either pathway.

The factor loadings in Table 4.5 are for data simulated according to the scenario in Figure 4.4 where the use of 95% HPD regions clearly shows the separation of the genes into the two pathways, which is identical to the MLFA results.

C. Time Series Data

Table 4.6 presents results from a CEBFA on simulated time series data containing six genes simulated with the scenario in Figure 4.2 and six genes simulated with the scenario in Figure 4.5. The autoregressive parameter ρ was estimated to be -0.56 for this data set. Genes G_1 through G_4 and genes G_7 and G_{11} have high loadings on factor one, therefore we would place them on pathway 1 using the same logic used for the steady state data. However, genes G_1 through G_6 correspond to a pathway simulated with the scenario in Figure 4.2 and genes G_7 through G_{12} correspond to a pathway simulated with the scenario in Figure 4.5. The two pathways are completely independent. Inspection of the 95% HPD regions for factor 2 shows some loadings appear different from zero, however, the wide intervals suggest little information exists for this factor. The inclusion of genes from different pathways on a single factor suggest that CEBFA cannot group genes by pathway for simulated time series gene expression data.

Conclusions

We show that Bayesian factor analysis can be implemented in an exploratory setting which is in contrast to a previous implementation of confirmatory factor analysis. Consequently, Bayesian factor analysis can be used for the exploration of microarray gene expression data as an alternative to classical factor analysis. Advantages of the Bayesian implementation include no reliance on assymptotic theory for statistical inference and the ability to incorporate prior information. The latter could be used in a partial confirmatory factor analysis where nonzero prior means for genes known *a priori* to be in the same metabolic pathway could be specified on a single factor. In contrast to confirmatory factor analysis in a classical setting where prespecified factor loadings are restricted *a priori* to zero (Rubin & Thayer, 1982), vague prior variances of factor loadings can be specified for loadings on factors with both zero and nonzero factor loading means. This allows the data to indicate if sufficient evidence exists for the prior factor loading structure.

The poor performance of the correlated model for time series data could indicate that it either failed to properly account for the correlation between observation vectors, or it is not the appropriate method to capture the information from this type of data. The latter would appear more likely since the posterior mean of the autocorrelation parameter was -0.56 indicating that the method finds a correlation between the observed time points. However, at least two different correlations between time points exists, one between time points for the genes within the branches and begining of the loop and another between time points for the remainder of each pathway. Given our simulation model, the former will be negative, while the latter will be positive. The presence of heterogeneous correlations between time points violates our proposed simple correlation structure and suggests that CEBFA may perform better under less restrictive correlation structures. However, analysis with BEFA, which ignores the correlation structure, produces results similar to those presented for CEBFA. The rationale behind utilizing time series gene expression data is that the change in expression patterns over time contains information about the relationships between the genes within a pathway. Correlated exploratory Bayesian factor analysis does not extract this information from the data, at least when using a first order autoregressive correlation structure. This is in contrast with FA of steady state data, where genes are successfully grouped by pathway.

References

Chen, M.-H. & Shao, Q.-M. (1999) Monte -Carlo estimation of Bayesian credible and HPD intervals. *J. Comp. Graph. Stat.*, **8** (1), 69–92.

Clarkson, D. B. (1979) Estimating the standard errors of rotated factor loadings by jackknifing. *Psychometrika*, **44** (3), 297–314.

Congdon, P. (2001) *Bayesian statistical modelling*. Wiley, West Sussex, England.

Efron, B. & Tibshirani, R. J. (1993) *An introduction to the bootstrap*. Chapman and Hall, New York, USA.

Henderson, D. A., Mendes, P., de la Fuente, A. & Hoeschele, I. (2002) Factor analysis for the identification of metabolic pathways from microarray expression data. *Bioinformatics*, **(Submitted)**.

Hofmeyr, J.-H. S. & Westerhoff, H. V. (2001) Building the cellular puzzle. control in multi-level reaction networks. *J. Theor. Biol.*, **208**, 261–285.

Jackson, D. A. (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, **74**, 2204–2214.

Johnson, R. A. & Wichern, D. W. (1998) *Applied multivariate statistical analysis*. $4^{\underline{\text{th}}}$ edition,, Prentice Hall, Englewood Cliffs, NJ, USA.

Mendes, P. (1993) Gepasi: a software package for modeling the dynamics, steady states, and control of biochemical and other systems. *Comp. Appl. Biosci.,* **9**, 563–571.

Mendes, P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.,* **22**, 361–363.

Mendes, P. (1999) Metabolic simulation as an aide in understanding gene expression data. In *Proc. Workshop Comp. Biochem. Path. And Genetic Networks* pp. 27–33, Berlin, Germany.

Mendes, P. (2000) Gepasi: numerical simulation and optimization of biochemical kinetics. In *Proc. Plant and Anim. Genome VIII*, San Diego, CA, USA.

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

Press, S. & Shigemasu, K. (1997). Bayesian inference in factor analysis - Revised. Technical Report No. 243 Department of Statistics, University of California Riverside, California.

Rocke, D. M. & Durbin, B. (2001) A model for measurement error for gene expression arrays. *J. Comp. Biol.,* **In Press**.

Rowe, D. B. (1998). *Correlated Bayesian factor analysis*. PhD thesis, Department of Statistics, University of California Riverside.

Rowe, D. B. & Press, S. J. (1998). Gibbs sampling and hill climbing in Bayesian factor analysis. Technical Report No. 255 Department of Statistics, University of California Riverside.

Rubin, D. B. & Thayer, D. T. (1982) EM algorithms for ML factor analysis. *Psychometrika,* **47** (1), 69–76.

Wade, K. M. & Quaas, R. L. (1993) Solutions to a system of equations involving a first-order autoregressive process. *J. Dairy Sci.,* **76** (10), 3026–3032.

Table 4.1. Maximum Likelihood Factor loadings for two independent pathways simulated with scenario 4.2 (Figure 4.2). 95% bootstrap confidence intervals are estimated from 2000 bootstrap samples.

	Rotated Factor Loadings							
	Factor 1			Factor 2				Specific
Gene	Loading	Lower	Upper	Loading	Lower	Upper	Communality	Variance
G_1	0.8727	0.8227	0.9084	0.0787	-0.0659	0.1978	0.7678	0.1999
G_2	-0.7796	-0.8698	-0.7055	-0.0190	-0.1988	0.1014	0.6082	0.3056
G_3	0.8604	0.7953	0.9048	0.0287	-0.1147	0.1461	0.7411	0.2095
G_4	-0.8276	-0.9101	-0.7525	-0.0413	-0.1921	0.0747	0.6867	0.2452
G_5	-0.8974	-0.9406	-0.8557	-0.0377	-0.1822	0.0709	0.8068	0.1696
G_6	-0.8271	-0.8900	-0.7806	-0.0999	-0.2650	0.0400	0.6940	0.2505
G_7	0.0049	-0.1489	0.1148	0.8794	0.8161	0.9176	0.7734	0.1860
G_8	-0.0774	-0.2227	0.0362	-0.8122	-0.8796	-0.7580	0.6657	0.2659
G_9	0.0628	-0.0730	0.1699	0.8770	0.8200	0.9134	0.7731	0.1852
G_{10}	-0.0679	-0.1976	0.0384	-0.8858	-0.9277	-0.8480	0.7893	0.1710
G_{11}	-0.0590	-0.2246	0.0566	-0.8632	-0.9180	-0.8209	0.7486	0.2017
G_{12}	-0.0256	-0.1693	0.1010	-0.8681	-0.9166	-0.8249	0.7543	0.2011

Table 4.2. Bayesian Exploratory Factor Analysis factor loadings for two independent pathways simulated with scenario 4.2 (Figure 4.2). 95% Highest Posterior Density Regions, 2,000 samples following 5,000 rounds burn in.

	Rotated Factor Loadings							
	Factor 1			Factor 2				Specific
Gene	Loading	Lower	Upper	Loading	Lower	Upper	Communality	Variance
G_1	-0.9103	-1.0710	-0.7485	-0.0904	-0.2370	0.0496	0.8368	0.2370
G_2	0.8338	0.6673	1.0170	0.0085	-0.1488	0.1576	0.6953	0.3680
G_3	-0.9017	-1.0000	-0.7408	-0.0204	-0.1633	0.1278	0.8134	0.2603
G_4	0.8746	0.7036	1.0000	0.0442	-0.1045	0.1935	0.7669	0.3018
G_5	0.9409	0.7902	1.0000	0.0390	-0.0942	0.1753	0.8868	0.1939
G_6	0.8538	0.6768	1.0000	0.0866	-0.0570	0.2439	0.7365	0.3320
G_7	0.0127	-0.1277	0.1567	-0.9080	-1.0000	-0.7463	0.8245	0.2606
G_8	0.1104	-0.0440	0.2716	0.8281	0.6600	1.0000	0.6980	0.3724
G_9	-0.0579	-0.2070	0.0796	-0.9066	-1.0000	-0.7380	0.8253	0.2567
G_{10}	0.0570	-0.0782	0.2069	0.9070	0.7493	1.0000	0.8260	0.2566
G_{11}	0.0693	-0.0792	0.2162	0.8836	0.7132	1.0000	0.7855	0.2911
G_{12}	0.0222	-0.1114	0.1762	0.9038	0.7430	1.0000	0.8174	0.2647

Table 4.3. Bayesian Exploratory Factor Analysis standard errors, variances, and effective sample sizes for data simulated with scenario 4.2 (Figure 4.2). 2,000 samples following 5,000 rounds burn in.

	Factor 1		Factor 2		Specific Variance	
Gene	Variance[1]	Sample Size	Variance	Sample Size	Variance	Sample Size
G_1	0.0070	472	0.0053	508	0.0019	1398
G_2	0.0081	563	0.0060	608	0.0037	1709
G_3	0.0072	489	0.0053	531	0.0021	1428
G_4	0.0076	409	0.0056	509	0.0025	1538
G_5	0.0068	363	0.0046	439	0.0014	1198
G_6	0.0077	445	0.0060	590	0.0031	1601
G_7	0.0051	501	0.0074	390	0.0022	1403
G_8	0.0066	621	0.0082	576	0.0037	1595
G_9	0.0054	510	0.0073	396	0.0022	1392
G_{10}	0.0052	559	0.0076	406	0.0022	1301
G_{11}	0.0057	521	0.0077	452	0.0025	1548
G_{12}	0.0052	517	0.0076	368	0.0022	1464

[1] Variance of posterior distribution of factor loading estimated as the variance of the sample values.

Table 4.4. Bayesian Exploratory Factor Analysis factor loadings for two independent pathways simulated with scenario 4.3 (Figure 4.3). 95% Highest Posterior Density Regions, 2,000 samples following 5,000 rounds burn in.

	Rotated Factor Loadings							
	Factor 1			Factor 2				Specific
Gene	Loading	Lower	Upper	Loading	Lower	Upper	Communality	Variance
G_1	0.3123	0.0590	0.5842	0.0895	-0.1203	0.3174	0.1056	0.9114
G_2	0.8294	0.5764	1.0000	0.0152	-0.1575	0.1902	0.6881	0.3970
G_3	0.6558	0.4183	0.8818	-0.0360	-0.2193	0.1651	0.4314	0.6263
G_4	0.8424	0.5943	1.0000	-0.0252	-0.2027	0.1505	0.7102	0.3776
G_5	-0.2168	-0.4533	0.0171	-0.1729	-0.3958	0.0479	0.0769	0.9372
G_6	-0.3024	-0.5325	-0.0342	0.0931	-0.1250	0.3125	0.1001	0.9156
G_7	0.0413	-0.1468	0.2252	-0.6537	-0.8667	-0.4598	0.4291	0.6213
G_8	0.0478	-0.0767	0.1986	-0.8947	-1.0000	-0.7195	0.8028	0.2859
G_9	-0.0265	-0.1861	0.1232	-0.8520	-1.0000	-0.6767	0.7265	0.3546
G_{10}	-0.0050	-0.1652	0.1537	-0.8033	-1.0000	-0.6243	0.6453	0.4292
G_{11}	-0.0487	-0.2388	0.1400	-0.6550	-0.8512	-0.4446	0.4314	0.6216
G_{12}	-0.1272	-0.3558	0.1016	0.2637	0.0374	0.4966	0.0857	0.9274

Table 4.5. Bayesian Exploratory Factor Analysis factor loadings for two independent pathways simulated with scenario 4.4 (Figure 4.4. 95% Highest Posterior Density Regions, 2,000 samples following 5,000 rounds burn in.

	Rotated Factor Loadings							
	Factor 1			Factor 2				Specific
Gene	Loading	Lower	Upper	Loading	Lower	Upper	Communality	Variance
G_1	0.8603	0.6907	1.0000	0.0851	-0.0533	0.2433	0.7473	0.3243
G_2	0.9104	0.7493	1.0000	0.1531	0.0112	0.2946	0.8522	0.2277
G_3	0.8841	0.7253	1.0000	-0.0043	-0.1510	0.1331	0.7816	0.2944
G_4	0.9044	0.7378	1.0000	0.0969	-0.0492	0.2355	0.8273	0.2503
G_5	-0.8810	-1.0000	-0.7135	-0.1107	-0.2649	0.0245	0.7884	0.2854
G_6	-0.7746	-0.9637	-0.6090	0.0405	-0.1139	0.2098	0.6017	0.4572
G_7	0.0167	-0.1412	0.1519	0.8488	0.6777	1.0000	0.7207	0.3452
G_8	0.0545	-0.0753	0.1750	0.9732	0.8109	1.0000	0.9500	0.1354
G_9	0.1879	0.0432	0.3374	0.8584	0.6767	1.0000	0.7723	0.2961
G_{10}	0.0521	-0.0779	0.1881	0.9120	0.7421	1.0000	0.8344	0.2410
G_{11}	-0.0909	-0.2390	0.0562	-0.8116	-0.9895	-0.6305	0.6669	0.3938
G_{12}	-0.0230	-0.1894	0.1280	-0.7591	-0.9478	-0.5735	0.5767	0.4806

Table 4.6. Correlated Bayesian Exploratory Factor Analysis factor loadings for two independent pathways simulated with scenario 4.2 (Figure 4.2 and scenario 4.5 (Figure 4.5. 95% Highest Posterior Density Regions, 2,000 samples following 5,000 rounds burn in.

	Rotated Factor Loadings							
	Factor 1			Factor 2				Specific
Gene	Loading	Lower	Upper	Loading	Lower	Upper	Communality	Variance
G_1	0.6588	0.4160	1.0000	0.1179	-0.1734	0.5479	0.4479	0.1213
G_2	0.5567	0.3612	0.7532	-0.0260	-0.2695	0.2130	0.3106	0.1263
G_3	0.6173	0.3701	0.8513	-0.1058	-0.5405	0.1895	0.3923	0.1174
G_4	0.6142	0.3882	0.8519	-0.1049	-0.4412	0.2167	0.3882	0.1260
G_5	0.3252	0.0093	0.5549	-0.0758	-0.5237	0.4106	0.1115	0.1225
G_6	-0.3607	-0.6134	-0.0546	0.0970	-0.4673	0.6763	0.1395	0.1198
G_7	0.6217	0.4239	0.8436	0.0575	-0.2080	0.4068	0.3898	0.1148
G_8	0.1951	-0.0159	0.3859	0.0165	-0.5412	0.7174	0.0383	0.0813
G_9	0.0910	-0.1635	0.3626	-0.2380	-0.7713	0.5449	0.0649	0.1141
G_{10}	-0.2289	-0.5388	0.0781	-0.2383	-0.8581	0.5450	0.1092	0.1059
G_{11}	0.4506	0.1332	0.6867	-0.1298	-0.5908	0.2068	0.2199	0.1177
G_{12}	-0.1440	-0.3973	0.0897	0.0570	-0.6465	0.7122	0.0240	0.1232

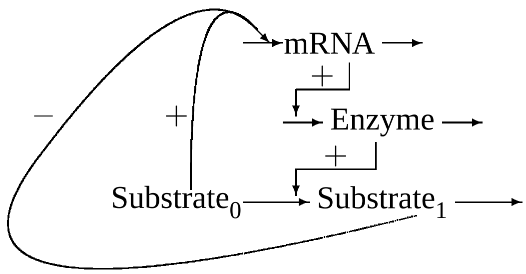

Figure 4.1. Hierarchy of simulated pathway structure.

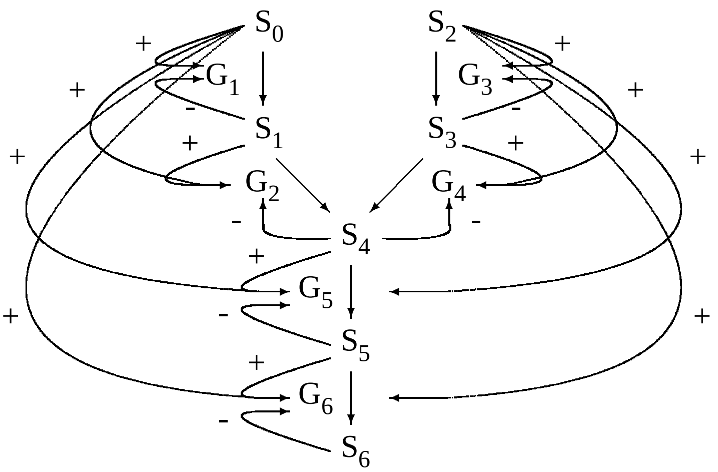

Figure 4.2. Six gened simulated pathway structure with regulatory interactions.

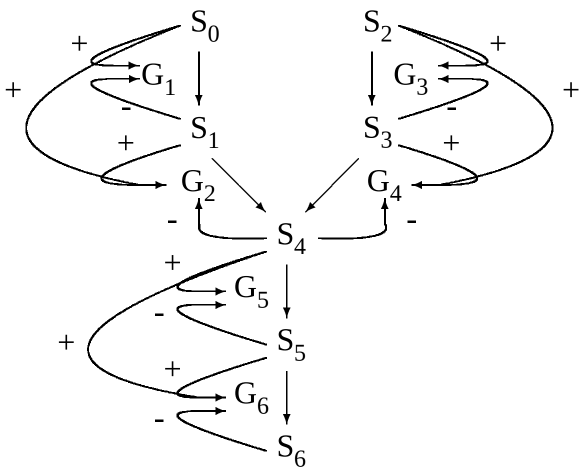

Figure 4.3. Six gened simulated pathway structure with regulatory interactions.

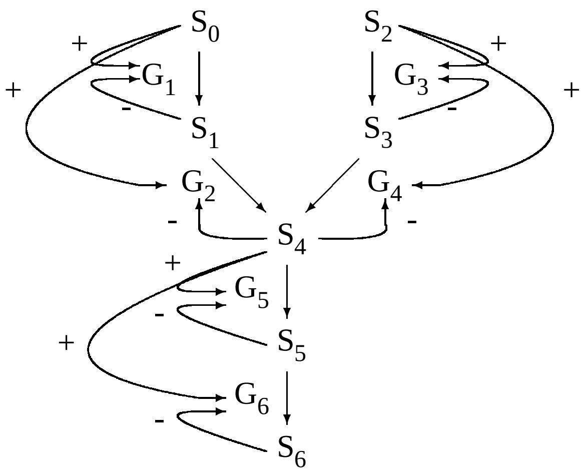

Figure 4.4. Six gened simulated pathway structure with regulatory interactions.

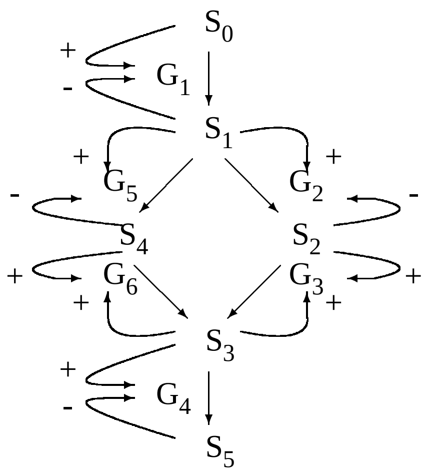

Figure 4.5. Six gene parallel looped pathway structure.

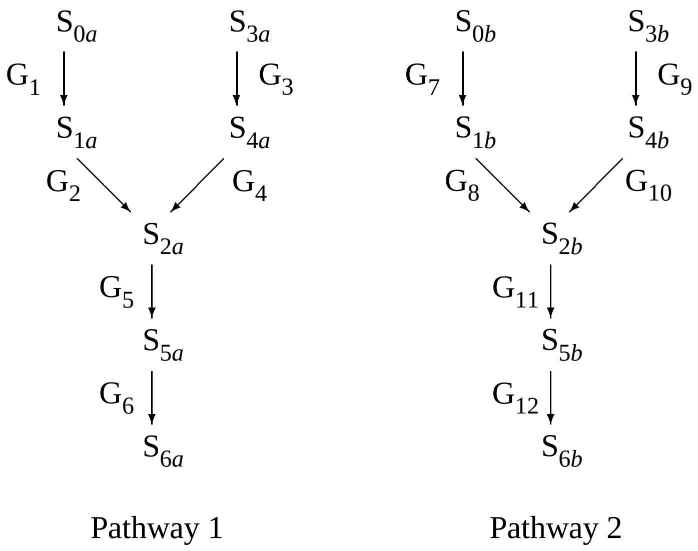

Figure 4.6. Pathway structure for two independent pathways.

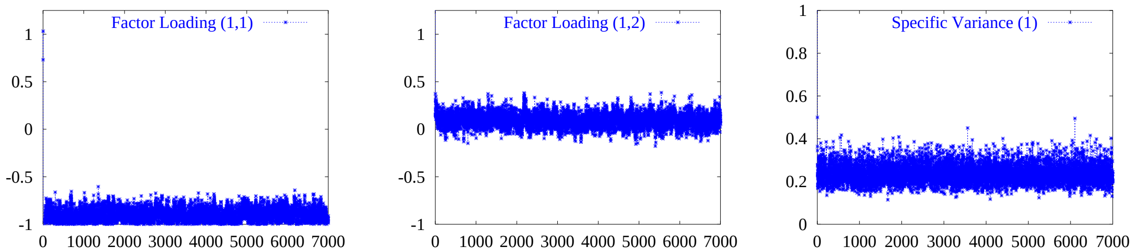

Figure 4.7. Trace plots of 7000 posterior estimates of factor loadings with no burn in.

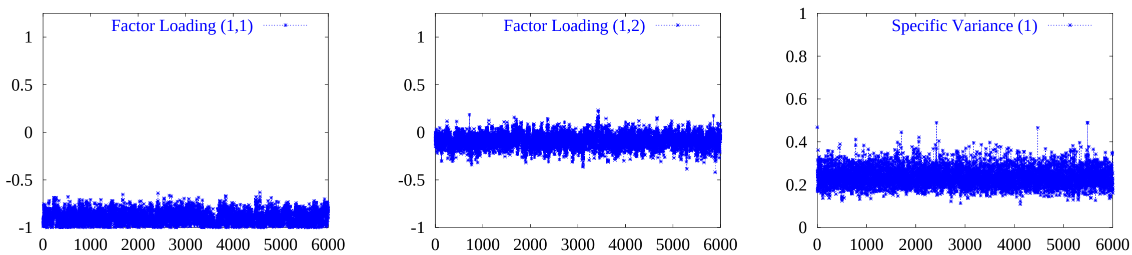

Figure 4.8. Trace plots of 6000 posterior estimates of factor loadings and specific variances following 1000 rounds of burn in.

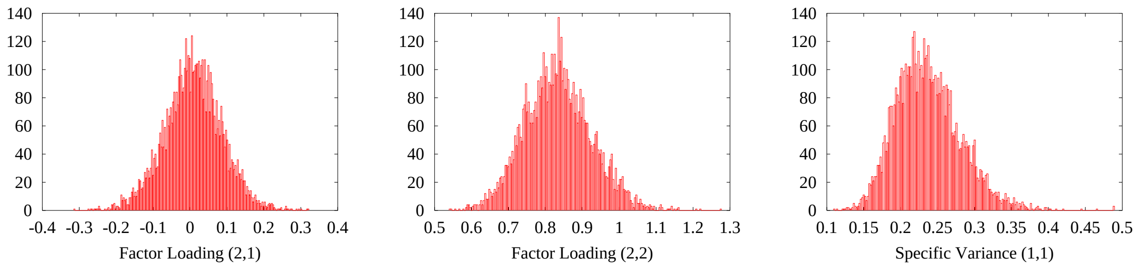

Figure 4.9. Histograms of 6000 posterior estimates of factor loadings and specific variances following 1000 rounds of burn in.

Curriculum Vitae

David Allen Henderson was born on September 13, 1971 in Urbana, IL to Lonnie and Irma Jean Henderson. At six months of age, his parents moved back to South Texas where he lived until graduation in 1989 from Pleasanton High School in Pleasanton, TX. Following graduation from high school, David attended Texas A&M University and graduated on August 13, 1993 with a Bachelor of Science degree in Animal Science. Still pursuing formal education, David moved to Brookings, SD, obtaining a Master of Science degree in Animal Science with a minor in Genetics from South Dakota State University in 1996.

Prior to the actual completion of his M.Sc. at South Dakota, David joined the Technical and Development Department of PIC USA in Franklin, KY. He worked here for three year before deciding to return to school to pursue a Ph.D. in Genetics at Virginia Tech.