Data Mining Based on Principal Component Analysis:
Application to the Nitric Oxide Response in *Escherichia coli*

AiLing Teh3, Donovan Layton1, Daniel R. Hyduke2,
Laura R. Jarboe1, and Derrick K. Rollins, Sr.1,3

1. Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa 50011
2. Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, Los Angeles, California 90095
3. Department of Statistics, Iowa State University, Ames, Iowa 50011

This work evaluates a recently developed multivariate statistical method based on the creation of pseudo or latent variables using principal component analysis (PCA). The application is the data mining of gene expression data to find a small subset of the most important genes in a set of thousand or tens of thousands of genes from a relatively small number of experimental runs. The method was previously developed and evaluated on artificially generated data and real data sets. Its evaluations consisted of its ability to rank the genes against known truth in simulated data studies and to identify known important genes in real data studies. The purpose of the work described here is to identify a ranked set of genes in an experimental study and then for a few of the most highly ranked unverified genes, experimentally verify their importance. This method was evaluated using the transcriptional response of *Escherichia coli* to treatment with four distinct inhibitory compounds: nitric oxide, S-nitrosoglutathione, serine hydroxamate and potassium cyanide. Our analysis identified genes previously recognized in the response to these compounds and also identified new genes. Three of these new genes, *ycbR*, *yfhA* and *yahN*, were found to significantly (p-values<0.002) affect the sensitivity of *E. coli* to nitric oxide-mediated growth inhibition. Given that the three genes were not highly ranked in the selected ranked set (RS), these results support strong sensitivity in the ability of the method to successfully identify genes related to challenge by NO and GSNO. This ability to identify genes related to the response to an inhibitory compound is important for engineering tolerance to inhibitory metabolic products, such as biofuels, and utilization of cheap sugar streams, such as biomass-derived sugars or hydrolysate.

*Key words:* Data mining, principal component analysis (PCA), gene expression data analysis

**Introduction**

*E. coli* is one of the bacterial species that resides in the human digestive tract, but it is also the causative agent of diseases such as cystitis and pyelonephritis. During the course of infection, *E. coli* is affected by reactive nitrogen species (RNS) such as nitric oxide (NO) and S-nitrosoglutathione (GSNO) (Jarboe et al. 2010). Moreover, NO plays an important role in fighting infections inside the human body (MacMicking et al. 2008).
Data Mining Based on Principal Component Analysis: Application to the Nitric Oxide Response in *Escherichia coli*

Network Component Analysis (NCA), a method based on known regulatory networks, was previously used to identify the molecular targets and bacterial response networks to NO and GSNO (Hyduke et al. 2007; Jarboe et al. 2008). However, the mathematical structure of NCA means that genes which have no known regulators, many of which often have little or no functional characterization, are excluded from analysis. The purpose of this work is to evaluate the ability of a technique developed by Rollins et al. (2010) based on Principal Component Analysis (PCA), to identify treatment-specific genes without a dependence on known regulatory networks or gene function. More specifically, the identification ability of this method was evaluated in this work using the transcriptional response of *E. coli* to four distinct chemical treatments: NO and GSNO, as well as the respiration inhibitor potassium cyanide (KCN) and serine starvation inducer serine hydroxamate (SeOH). This method is broadly applicable to any transcriptome data, where transcriptome analysis is often used to identify the mechanism of inhibition by biofuels or trace contaminants in cheap, renewable sugar streams (Brynildsen and Liao 2009; Miller et al. 2009; Jarboe et al. 2011).

PCA has been widely used to analyze transcriptome data (MacMicking et al. 1997; Bogdan 2001; Voskuil et al. 2003; Raychaudhuri et al. 2000; Misra et al. 2002; Sharov et al. 2005). Two common strengths of PCA are its ability to reduce dimensionality and to create data-driven methods by exploiting its ability to find natural relationships among correlated variables (see the text book by Johnson and Wischern, 2008). The latter strength is the one we exploit in this work. Rollins et al. (2006) developed a PCA technique that uses gene contribution to a specific biological behavior of interest to obtain a ranked set (RS) of these genes. Rollins et al. (2010) extended this PCA RS technique to differential gene contribution for two groups of experimental conditions or assays. Rollins et al. (2010) evaluated this technique by comparing it to a previously identified list in a real data study and a known list in a simulated data study. In contrast, the work in this article evaluates the ability of this technique to identify new genes that affect a specific type of biological response and then to evaluate the accuracy of the identification results by experimentation. This study used the responses to four distinct chemical treatments and *E. coli* as the biological system. It applied the PCA technique to identify a ranked-order list of genes. This list included many of the genes known to be important in the response to NO. We then selected seven under-characterized genes from ranked sets to NO tolerance in *E. coli*. Experimental validation suggests that three of these tested genes may contribute to the sensitivity of *E. coli* to growth inhibition by NO.

**Methods**

**Transcriptome Data, Overexpression and Candidate Genes**

The data matrix, X, for this work has *m* = 4166 rows representing 4166 genes and *n* = 4 columns or assays (i.e., experimental conditions). The cells in this matrix are given as *x*_ij which is the expression level of the *i*th gene for the *j*th assay. Thus, *X*^mn^ = {*x*_ij}. This transcriptome data matrix is from Hyduke et al. (2007) and Jarboe et al. (2008) and it gives the 5-minute response of *E. coli* strain BW25113 to four chemical treatments: the RNS nitric oxide (8 μM diethylamine NONOate, NO, *j* = 1), and S-nitrosoglutathione (100 μM GSNO, *j* = 2), potassium cyanide (100 μM KCN, *j* = 3) and serine hydroxamate (1 mg/mL SeOH, *j* = 4). The KCN dataset has not been previously published; it is available for download from the Gene Expression Omnibus with accession number GSE23578. KCN causes respiration inhibition, one of the known effects of NO (MacMicking et al. 1997). SeOH induces amino acid starvation (Rodionov and Ishiguro 1997), which is
Data Mining Based on Principal Component Analysis: Application to the Nitric Oxide Response in Escherichia coli

an effect of both NO and GSNO treatment (MacMicking et al. 1997; Hyduke et al. 2007). KCN and SeOH are included to eliminate some of the indirect effects from NO and GSNO, such as amino acid starvation and respiration inhibition.

Candidate genes were tested by cloning the focal gene into an inducible expression vector, in order to increase its expression (i.e., “overexpress”) in a controlled manner, and analyzing the growth response to challenge with NO. Coding regions were amplified from BW25113 genomic DNA using Taq master mix (Qiagen, Valencia, CA) and cloned into pTrc-his TOPO TA Expression vectors following the manufacturer’s instructions (Invitrogen, Carlsbad, CA). Growth analysis was performed in 25 mL MOPS + 0.2% glucose, 0.01 mM IPTG with 60 mg mL⁻¹ ampicillin at 37 °C in baffled flasks 1/10th full, in an orbital shaker at 150 rpm. Cells were grown to an OD₅₅₀ of approximately 0.4 before NO challenge. Protection conferred against NO-mediated growth inhibition by the test genes was monitored relative to an empty control vector in either wild-type BW25113 with a single dose of 10μM diethylamine NONOate (DeaNO, Cayman Chemical, Ann Arbor, MI) or in BW25113 ΔΔhmpA::CmR (Hyduke et al. 2007) with 10 ΔΔM DeaNO added every five minutes. DeaNO was prepared in a Tris/saline storage buffer (pH > 10) and stored on ice.

The PCA Method

The first step in the application of the Rollins et al. methodology is to group the assays into two sets or groups, A and B. In this work, Group A is the chemical treatment of interest and Group B contains the other chemical treatments. In this application of this methodology, the goal is to rank the genes in X based on their differential expression between these two groups. This ranked set (RS) is given by the vector Ymx₁. A limitation of this methodology is that it does not provide help in choosing a reduced RS for a new investigation. Another limitation is that it ranks the genes by their absolute affect. Thus, highly ranked genes can have either a positive or negative significant impact on the biological behavior. Moreover, in the application of Y to select the most promising genes for follow up work in new discovery, one could start from the top gene in the set and move down the set in decreasing rank. This is the approach used in this study.

It is critical to note that this methodology does not make conclusions on gene significance. This method is considered effective (and useful) when this ranking is sufficiently accurate in regions where new discovery potentially exists. Note that if a methodology, such as the one here, does not make declarations (i.e., draw conclusions) for gene significance, the notion of error rate for wrong conclusions is not applicable. Thus, the determination of type 1 and 2 errors is not within the goal or application of this methodology.

The strengths of the PCA technique used in this work can be understood and appreciated by considering the following case. Suppose one is interested in finding a ranked order list of genes (i.e., Y) that express most differently for NO (Group A) response (x₁) versus the other three (Group B) responses (the average of x₂, x₃, and x₄). For a general chemical compound, this difference is given mathematically as:

\[
g_i = x_{i1} = \frac{x_{i2} + x_{i3} + x_{i4}}{3} \quad i = 1, \ldots, m
\]  

where \(g_i\) gives the differential expression of gene \(i\) according to Eq. 1. However, one problem with Eq. 1 is it was not developed based on an analysis of the data or from theoretical considerations and is, therefore, likely to result in poor gene ranking. Groups A and B will increase in separation as the sum of the \(g_i\)'s (i.e., \(\sum g_i\)) increase. What PCA does is provides a weight for each term in Eq. (1) using the inherent information
content of the data, more specifically, the correlation structure among the \(x_{ij}\)'s. The Rollins et al. method selects the set of weights provided by PCA that gives the largest value of \(\sum g_i\). Then it determines \(Y\) from the \(g_i\)'s.

There are two ways the Rollins et al. (2010) obtains weighted forms of Eq. (1). It does this by treating either the assays or genes as variables and obtains Eigenvalues (called “Eigengenes,” EG, when treating assay as variables or “Eigenassays,” EA, when treating genes as the variables). It then uses this information to establish the mathematical expression to calculate the contribution of each gene to this relationship. More specifically, for this case an EG analysis would plot the four loadings for each principal component (PC) against the assay numbers corresponding to the chemical treatments (see Fig. 1). Similarly, an EA analysis would plot the product of the vector of the loadings for each PC by each of the four genes vectors (producing four scores) against the assay numbers (see Fig. 2). The EG or EA corresponding to plots that give the largest difference between the mean levels between Groups A and B are selected to calculate \(g_i\) for all the genes. The general forms for these equations are given in Eqs. (2) and (3) below for EG and EA, respectively:

\[
g_{EGi} = \ell_1 x_{i1} - \left( \frac{\ell_2 x_{i2} + \ell_3 x_{i3} + \ell_4 x_{i4}}{3} \right) \quad i = 1, \ldots, m \tag{2}
\]

\[
g_{EAi} = \ell_i \left( x_{i1} - \left( \frac{x_{i2} + x_{i3} + x_{i4}}{3} \right) \right) = \ell_i g_i \quad i = 1, \ldots, m \tag{3}
\]

where in Eq. (2), \(\ell_j\) is the \(j\)th loading or weight for this EG vector, \(j = 1, \ldots, 4\); and \(\ell_i\) in Eq. 3 is the \(i\)th loading for this EA vector. Note that in Eq. (2), each \(x_{ij}\) for gene \(i\) has a different loading that does not change from gene to gene and that in Eq. (3), each \(x_{ij}\) for gene \(i\) has the same loading that changes from gene to gene.

This PCA methodology does not make conclusions on gene significance. The goal of the methodology in this study is to accurately rank all the genes from the most significant to the least significant. This method is considered effective (and useful) in reaching this goal when this ranking is sufficiently accurate in regions where new discovery potentially exist. Note that, if a methodology does not make declarations (i.e., draw conclusions) for gene significance, the notion of error rate for wrong conclusions is not applicable. Thus, the determination of type 1 and 2 errors is not within the goal or application of this methodology.

The number of correctly identified genes (CIG) is the number of genes in a ranked set (RS) that affects the biological behavior of interest and have been correctly identified as having an effect. The true detection sensitivity (TDS) is the percent of CIG in a RS and is given by:

\[
TDS = \frac{\text{number of CIG}}{\text{number of genes in the RS}} \times 100 \tag{4}
\]

The observed detection sensitivity (ODS) is the percent of genes in the RS known to have an effect and is given by:

\[
ODS = \frac{\text{number of CIG known to have an effect}}{\text{number of genes in the RS}} \times 100 \tag{5}
\]
Figure 1. Loading plots generated for EG analysis.

Loading 2 ($l_{i2}$) is used to identify GSNO-response genes and Loading 4 ($l_{i4}$) is used to identify NO-response genes. Blue square represents the condition of interest and the black circles represent the other conditions.
Score 1 ($S_{i1}^{EA}$) is used to identify GSNO-response gene and Score 3 ($S_{i3}^{EA}$) is used to identify NO-response genes. Blue square represents the condition of interest and the black circles represent the other conditions.
It is not possible to know for certain the TDS unless the complete truth of all the genes in a RS is known which is not likely unless the RS is very small. For a given RS, it is desirable for TDS to be high. The potential for successful discovery will be high when a RS has a large number of unknown CIG. Confidence in the results for a specific case will be low when the RS is relatively small and the ODS is low. Given that low confidence will likely result in no follow up effort for discovery, it is desirable for a technique to have high ODS for a targeted or sufficient RS size.

Eigengene Procedure

As stated above, the PCA technique has two ways of creating ranked-order lists of genes. In the context of the application of this work, we first give the steps for the EG procedure and then those for EA procedure.

The steps for obtaining a RS vector $Y^{mx1}$ using an EG procedure are as follows:

1. Standardize $X^{mxn} = \{x_{ij}\}$ to matrix $Z^{mxn} = \{z_{ij}\}$ where $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$, and $\bar{x}_j$ and $s_j$ are the sample mean and sample standard deviation, respectively, of the elements in column $j$.
2. By application of PCA, obtain the EG loading matrix, $L^{nxn}$ with column vectors $\ell_i$, $i = 1, \ldots, n$ and EG scores matrix, $S^{mn}$ with column vectors $S_i$, $i = 1, \ldots, n$.
3. Obtain a loading plot for each loading vector where each graph has a plot of the elements of $\ell_i$ against the assay number. In this work the assay numbers, 1 to 4, correspond to the four chemical treatments. (For an example see Fig. 1.)
4. For the chemical treatment of interest, assign it as Group A and the other 3 treatments as Group B. By visual inspection, select the loading plot that separates Group A from Group B with the greatest distance.
5. Use the elements of loading vector $\ell_i$ selected in Step 3 in Eq. 2 to obtain the differential contribution, $g_{EG_i}$, of each gene $i$.
6. Rank $g_{EG_i}$ from largest to smallest value to obtain the RS vector $Y^{mx1} = \{y_i\}$, $i = 1, \ldots, m$.
7. Plot the elements of $Y$ against their ranks to obtain the signature plot for the differential effects of Groups A and B. (For examples see Fig. 3.)
8. Repeat Steps 4 to 7 for each differential effect of interests. In this article, these steps are repeated for each chemical treatment of interest.

The steps for obtaining a RS vector $Y^{mx1}$ using an EA procedure are as follows:

1. By application of PCA, obtain the EA loading matrix, $L^{mxm}$ with column vectors and EA scores matrix, $S^{mn}$, with column vectors $S_i$, $i = 1, \ldots, n$.
2. Obtain a score plot for each score vector where each graph has a plot of the elements of $S_i$ against the assay number. In this work the assay numbers, 1 to 4, correspond to the four chemical treatments. (For an example of this plot see Fig. 2 for an example).
3. For the chemical treatment of interest, assign it as Group A and the other 3 treatments as Group B. By visual inspection, select the scores plot that separates Group A from Group B with the greatest distance.
4. For score plot found in Step 3, use the elements of its corresponding loading vector $\ell_i$ in Eq. 3 to obtain the differential contribution, $g_{EA_i}$, of each gene $i$. 
Figure 3. Signature plots showing gene contribution ranking for NO (left) and GSNO (right) from EG method.
Signature plot for NO response is generated using Loading 4 ($l_4$) while GSNO response is generated using Loading 2 ($l_2$) respectively from EG method.

(5) Rank $g_{\text{EG-GSNO}_i}$ from largest to smallest value to obtain the RS vector $Y^{\text{mx}}_i = \{ y_i \}, i = 1, \ldots, m$.

(6) Plot the elements of $Y$ against their ranks to obtain the signature plot for the differential effects of
Groups A and B. (For examples of this plot see Fig. 4.)

(7) Repeat Steps 3 to 6 for each differential effect of interests.

Figure 4. Signature plots showing gene contribution ranking for NO (left) and GSNO (right) from EA method. Signature plot for NO response is generated using Score 3 ($S_3$) while GSNO response is generated using Score 1 ($S_1$) respectively from EA method.
Results and Discussion

As discussed in the previous section, we generated four loading plots using the Eigengene (EG) procedure and four score plots using the Eigenassay (EA) procedure. The loading plots that gave the greatest distance from the mean values of Groups A and B were loading vectors 2 and 4; thus, they were selected for this analysis. The loading plots for Loading Vectors 2 \((l_{i2})\) and 4 \((l_{i4})\) are shown in Fig. 1; the plot for \(l_{i2}\) illustrates a clear separation of GSNO from the other assays and the one for \(l_{i4}\) illustrates a clear separation of NO from the other assays. Thus, \(l_{i2}\) and \(l_{i4}\) were selected to create the ranked-order list of genes for GSNO and NO responses, respectively. The EG contribution equations for GSNO and NO are given below by Eqs. (6) and (7), respectively.

\[
g_{EG-GSNO} = l_{2i}z_{i2} - \left( \frac{\ell_{12}z_{i1} + \ell_{32}z_{i3} + \ell_{42}z_{i4}}{3} \right) \quad i = 1, \ldots, m
\]

\[
g_{EG-NO} = l_{4i}z_{i4} - \left( \frac{\ell_{14}z_{i1} + \ell_{34}z_{i3} + \ell_{44}z_{i4}}{3} \right) \quad i = 1, \ldots, m
\]

Where \(l_{ij}\) is the \(i^{th}\) loading for the \(j^{th}\) EG vector. Similarly, under the EA procedure, the score plots that gave the greatest distance from the mean values of Groups A and B were for score vectors 1 and 3 \((S_{i1} \text{ and } S_{i3})\); thus, they were selected for this analysis. These plots are shown in Fig. 2. More specifically, we selected the first and third EA score vectors (i.e. EA1 and EA3) respectively, to create the ranked list of genes for GSNO and NO, respectively where \(S_{ij}\) is the \(i^{th}\) score for the \(j^{th}\) EA score vector. The EA contribution equations for GSNO and NO are given below by Eqs. (8) and (9), respectively.

\[
g_{EA-GSNO} = l_{1i} \left( x_{i2} - \left( \frac{x_{i1} + x_{i3} + x_{i4}}{3} \right) \right) \quad i = 1, \ldots, m
\]

\[
g_{EA-NO} = l_{3i} \left( x_{i1} - \left( \frac{x_{i2} + x_{i3} + x_{i4}}{3} \right) \right) \quad i = 1, \ldots, m
\]

Where \(l_{ij}\) is the \(i^{th}\) loading for the \(j^{th}\) EA vector.

Fig. 3 provides the EG signature plots for Eqs. (6) and (7). Similarly, Fig. 4 provides EA signature plots for Eqs. (8) and (9). From the rank ordered differential gene contributions shown in the signature plots, the distinct gaps for the top genes are clearly revealed.

The ranked set (RS) that we selected for this study needed to be large enough to have a significant number of unidentified genes and small enough to still have genes that were considered highly ranked in the full list of genes. For these reasons we selected a ranked set of 25 genes. The top 25 genes that we identified to be sensitive to direct NO target using EA and EG approaches are shown in Table 1 (ODS = 36%) and Table 2 (ODS = 56%), respectively. Similarly, direct GSNO target using the proposed approaches are shown in Tables 3 (ODS = 32%) and Table 4 (44%). From these tables, the agreement is quite strong as nearly all of the top genes identified from EG method matched with those that identified from EA method. Top genes that we have
identified have also been identified in other RNS studies, as reviewed by Jarboe et al. (2010). In addition, our list includes a number of genes that, to our knowledge, have not been previously identified as important to the NO or GSNO response.

Table 1

| Ranking | Gene   | Function                                                                 | Previously Known? |
|---------|--------|--------------------------------------------------------------------------|------------------|
| 1       | *hmp*  | Flavohemoglobin; NADPH-dependent NO dioxygenase; dihydropteridine reductase; GSNO and nitrite reductase | y                |
| 2       | *norW* | NADH:flavoruberedoxin reductase                                            | y                |
| 3       | *norV* | Anaerobic NO reductase flavorubredoxin                                     | y                |
| 4       | *ytfE* | Fe-S cluster repair protein; confers resistance to NO and hydrogen peroxide; di-iron center | y                |
| 5       | *sdhC* | Subunit of succinate dehydrogenase, carries electrons from FADH to CoQ    | n                |
| 6       | *cydA* | Contains the heme b558 component of cytochrome bd-I                       | y                |
| 7       | *carA* | Component of carbamoyl phosphate synthetase                               | n                |
| 8       | *sdhA* | Subunit of succinate dehydrogenase; contains the FAD cofactor             | n                |
| 9       | *yeeR* | Conserved protein                                                         | n                |
| 10      | *cydB* | Binds the heme b595 component and iron-chlorin component of cytochrome bd-I | y                |
| 11      | *sdhB* | Subunit of succinate dehydrogenase, delivers electrons from the FAD cofactor to ubiquinone | n                |
| 12      | *sdhD* | Subunit of succinate dehydrogenase                                        | n                |
| 13      | *iscR* | Transcriptional repressor for isc operon; contains Fe-S cluster; binds RNA in vitro | y                |
| 14      | *b0725*| Phantom gene                                                             | n                |
| 15      | *ndk*  | Nucleoside diphosphate kinase; important in the amino acid starvation stringent response | y                |
| 16      | *oppA* | Component of oligopeptide ABC transporter                                 | y                |
| 17      | *cvpA* | Required for production of colicin V                                      | n                |
| 18      | *yaiB* | Anti-adaptor protein required for stabilization of the alternative sigma factor $\sigma^S$ | n                |
| 19      | *sucA* | E1(0) component of the oxoglutarate dehydrogenase complex.               | n                |
| 20      | *pyrD* | Component of dihydroorotate oxidase                                       | n                |
| 21      | *yeeF* | Putative amino acid transporter of the APC superfamily                    | n                |
| 22      | *rnpA* | Ribonuclease P protein component                                           | n                |
| 23      | *glnL* | Sensor-histidine kinase; involved in nitrogen limitation response         | n                |
| 24      | *rplM* | 50S ribosomal subunit protein L13                                          | n                |
| 25      | *fliR* | Integral membrane component of the flagellar export apparatus            | n                |

In order to provide insight in interpretation of our results, we will briefly review the current understanding of the response to NO and GSNO, as reviewed by Jarboe et al. (2010). NO works primarily by damaging protein-associated metal complexes, including those associated with respiratory enzymes, branched-chain amino acid biosynthesis enzymes, regulation of the NO-consuming HmpA and NorVW enzymes and regulation of enzymes which repair damaged Fe-S centers. These interactions in turn activate amino acid starvation response pathways (due to the inhibition of branched-chain amino acid biosynthesis) and respiration inhibition response pathways. GSNO has two distinct chemical paths: reaction with exposed thiol groups and the release of NO. In this manner the NO response is part of the GSNO response. However, the interaction with exposed thiol groups means that GSNO exposure depletes intracellular cysteine and
Table 2
Top 25 genes for NO direct target identified from Eigenassay method. Genes marked with a bold font were employed in the validation study.

| Ranking | Gene   | Function                                                                 | Previously Known? |
|---------|--------|--------------------------------------------------------------------------|-------------------|
| 1       | norW   | NADH:flavoruberedoxin reductase                                           | y                 |
| 2       | norV   | Anaerobic NO reductase flavorubredoxin                                    | y                 |
| 3       | hmp    | Flavohemoglobin; NO dioxygenase; dihydropteridine reductase; GSNO and nitrite reductase; | y                 |
| 4       | ytfE   | Fe-S cluster repair protein; confers resistance to NO and hydrogen peroxide; di-iron center | y                 |
| 5       | yeaT   | LysR-type transcriptional regulator, Essential for growth on D-malate as sole carbon source | n                 |
| 6       | hscB   | Involved in Fe-S cluster assembly. Physically interacts with HscA and IscU | y                 |
| 7       | yfjL   | CP4-57 prophase; predicted protein of function unknown                    | n                 |
| 8       | yoaG   | Predicted protein                                                         | n                 |
| 9       | yeaR   | Conserved protein                                                         | n                 |
| 10      | argC   | N-acetylglutamylphosphate reductase; part of arginine biosynthesis        | n                 |
| 11      | hcp    | Hybrid cluster protein (HCP); possible scavenger of toxic by-products of nitrogen metabolism | y                 |
| 12      | iscU   | Scaffold protein for assembly and transfer of Fe-S clusters.              | y                 |
| 13      | nucC   | NADH:ubiquinone oxidoreductase subunit C, complex I; NADH dehydrogenase I | y                 |
| 14      | ilvC   | Ketol-acid reductoisomerase                                                | y                 |
| 15      | iscS   | Component of cysteine desulfurase, critical for repair and synthesis of Fe-S clusters | y                 |
| 16      | yahN   | Predicted neutral amino acid efflux system                                | n                 |
| 17      | fds    | Ferredoxin, an Fe-S protein; involved in assembly of other Fe-S clusters  | y                 |
| 18      | hypF   | Hydrogenase maturation protein                                            | n                 |
| 19      | ilvB   | Component of acetohydroxybutanoate synthase I                            | y                 |
| 20      | ycbR   | Putative periplasmic pilus chaperone, induced by AI-2 pheromone, function unknown | n                 |
| 21      | iscX   | Possible role in Fe-S cluster biogenesis                                  | y                 |
| 22      | pusD   | Component of y-glutamyl-y-aminobutyrate hydrolase                        | n                 |
| 23      | argG   | Component of argininosuccinate synthase                                   | n                 |
| 24      | yclE   | Conserved protein                                                         | n                 |
| 25      | ilvN   | Component of acetohydroxybutanoate synthase I                            | y                 |

The eigengene and eigenassay methods both identified the known NO-responsive genes hmp, norV, norW and ytfE as contributing to the NO response. The proteins encoded by each of these four genes are already known to contribute to defense against NO-mediated damage and growth inhibition; for example, Hmp and NorV are both known to directly consume NO (Misra et al. 2002). Additionally, the NO-important genes include Fe-S cluster repair genes iscRSUA and many amino acid biosynthesis genes. This is consistent with the previous report that NO damages the Fe-S center of branched-chain amino acid biosynthesis protein IlvD, causing amino acid starvation (Hyduke et al. 2007). For the GSNO response, the eigengene and eigenassay methods both identified a significant number of genes related to methionine and cysteine biosynthesis. This is consistent with reports that GSNO depletes the levels of cysteine and the methionine precursor homocysteine (Jarboe et al. 2008). Together, these findings validate the methodology presented here. However, this method
Data Mining Based on Principal Component Analysis: Application to the Nitric Oxide Response in Escherichia coli

also identified genes that are outside the previously-reported NO and GSNO response models. This includes several genes with minimal available functional or annotation data, such as yahN, ycbR, yciE, yeaR.

Table 3
Top 25 genes for GSNO direct target identified from Eigengene method. Genes marked with a bold font were employed in the validation study.

| Ranking | Gene     | Function                                                                 | Previously Known? |
|---------|----------|--------------------------------------------------------------------------|-------------------|
| 1       | yedN_1   | Hypothetical protein                                                     | n                 |
| 2       | metR     | Positive regulator of methionine biosynthesis genes metE and metH        | y                 |
| 3       | metA     | Homoserine O-transsuccinylase, part of methionine biosynthesis           | y                 |
| 4       | yigM     | Predicted inner membrane protein                                         | n                 |
| 5       | carA     | Component of carbamoyl phosphate synthetase                              | n                 |
| 6       | cspB     | Qin prophage; cold shock protein                                         | n                 |
| 7       | uspB     | Predicted universal stress protein B                                     | n                 |
| 8       | hmp      | Flavohemoglobin; NO dioxygenase; dihydropteridine reductase; GSNO and nitrite reductase | y                 |
| 9       | flaR     | Integral membrane components of the flagellar export apparatus          | n                 |
| 10      | ybdL     | Methionine aminotransferase, PLP-dependent                               | n                 |
| 11      | metN     | L,D-methionine transporter, ATP-binding protein; methionine sulfoximine sensitivity | y                 |
| 12      | yaiB     | Anti-adaptor protein that is required for stabilization of the alternative sigma factor | n                 |
| 13      | rnpA     | Ribonucleosyl P protein component                                        | n                 |
| 14      | dnaK     | Chaperone Hsp70; DNA biosynthesis; autoregulated heat shock proteins    | n                 |
| 15      | rpsA     | 30S ribosomal subunit protein S1                                          | n                 |
| 16      | mmuP     | S-methylmethionine permease, CP4-6 putative prophage remnant             | n                 |
| 17      | metF     | Component of 5,10-methylenetetrahydrofolate reductase                    | y                 |
| 18      | glnA     | Adenylyl-[glutamine synthetase], GlnA                                    | n                 |
| 19      | mmuM     | S-methylmethionine-(seleno)homocysteine methyltransferase; CP4-6 putative prophage remnant | n                 |
| 20      | metB     | Cystathionine gamma-synthase; homotetrameric                            | y                 |
| 21      | cysJ     | Sulfite reductase [NADPH] flavoprotein alpha-component; binds FMN and FAD | y                 |
| 22      | rpsU     | 30S ribosomal subunit protein S21                                        | n                 |
| 23      | glnL     | Sensor-histidine kinase; involved in nitrogen limitation response        | n                 |
| 24      | ytfE     | Fe-S cluster repair protein; confers resistance to NO and hydrogen peroxide; di-iron center | y                 |
| 25      | ybdH     | Predicted oxidoreductase                                                 | n                 |

We selected seven highly ranked genes that were not previously reported for contribution to the NO response for experimental assessment. They are yfhA, yahN, yeaR, ycbR, yeaT, iscS, and yoaG, with known NO-protective gene hmp included as a positive control. Note that, as seen from Tables 1 and 2, these genes are not very highly ranked in the selected ranked set (RS). The highest rank is for yahN -- 16 in Table 2 for NO. The rank for yfhA is 28 and is only ranked for the NO response in Table 2. While ycbR rank appears in three of the four lists, these ranks are 20, 34 and 45. Thus, given that these are low RS genes these results will strengthen the confidence in the sensitivity of the method if these genes are found to have an effect.

It is hypothesized that genes important to NO response might have an observable effect on the growth sensitivity of E. coli to NO challenge. The effect of overexpressing the potentially NO-protective genes was first assessed in the wild-type parent strain BW25113 relative to the same strain containing an empty
expression vector. In this experiment, strains were challenged with a single dose of $10^{-6}$ μM DEA/NO. Strains with the yeaT, yoaG, yeaR or iscS expression vectors showed no apparent difference relative to the empty expression vector and were excluded from further analysis (data not shown). The growth sensitivity of *Escherichia coli* to DEA/NO is more pronounced in the absence of the hmp gene (Hyduke et al. 2007). Therefore, we increased the expression of the candidate genes ycbR, yfhA, yahN, and hmp individually in BW25113 hmp, and tested the growth response of the resulting strains to NO challenge.

Table 4

*Top 25 genes for GSNO direct target identified from Eigenassay method. Genes marked with a bold font were employed in the validation study.*

| Ranking | Gene     | Function                                                                 | Previously Known? |
|---------|----------|--------------------------------------------------------------------------|-------------------|
| 1       | yedN_1   | Hypothetical protein                                                     | n                 |
| 2       | metR     | Positive regulator of methionine biosynthesis genes metE and metH       | y                 |
| 3       | metF     | Component of 5,10-methylenetetrahydrofolate reductase                  | y                 |
| 4       | yigM     | Predicted inner membrane protein                                        | n                 |
| 5       | metA     | Homoserine O-transsuccinylase                                           | y                 |
| 6       | ygcN     | Predicted oxidoreductase with FAD/NAD(P)-binding domain                | n                 |
| 7       | cysJ     | Sulfite reductase [NADPH] flavoprotein alpha-component; binds FMN and FAD | y                 |
| 8       | cysH     | Component of 3'-phospho-adenylsulfate reductase                           | y                 |
| 9       | cysN     | Component of sulfate adenylyltransferase                                | y                 |
| 10      | mmuP     | S-methylmethionine permease, CP4-6 putative prophage remnant            | n                 |
| 11      | ybdL     | Methionine aminotransferase, PLP-dependent                              | n                 |
| 12      | mmuM     | S-methylmethionine:(seleno)homocysteine methyltransferase; CP4-6 putative prophage remnant | n                 |
| 13      | cysD     | Component of sulfate adenylyltransferase                                | y                 |
| 14      | sbp      | Component of sulfate ABC transporter                                    | y                 |
| 15      | yeeE     | Putative transport system permease protein                              | n                 |
| 16      | metN     | L,D-methionine transporter, ATP-binding protein                          | y                 |
| 17      | metB     | Cystathionine gamma-synthase                                            | y                 |
| 18      | uspB     | Predicted universal stress (ethanol tolerance) protein B                | n                 |
| 19      | yjdK     | Hypothetical protein                                                    | n                 |
| 20      | cspB     | Qin prophase; cold shock protein                                        | n                 |
| 21      | cysI     | Sulfite reductase [NADPH] hemoprotein beta-component; has 4Fe-4S Fe-S center | y                 |
| 22      | ybdH     | Predicted oxidoreductase                                                | n                 |
| 23      | ydcD     | Hypothetical protein                                                    | n                 |
| 24      | thrA     | Component of aspartate kinase I                                         | n                 |
| 25      | dnaK     | Chaperone Hsp70; DNA biosynthesis; autoregulated heat shock proteins    | n                 |

As an NO dioxygenase, Hmp is known to be helpful in protection from NO and was determined to be highly significant in this study also. Therefore, as expected, the strain with increased expression of hmp is more tolerant of NO than the strain containing only the empty expression vector, as evidenced by significantly increased growth (the plot is not shown for space consideration). Contrastingly, yahN, ycbR, and yfhA each actually increased NO sensitivity upon their overexpression. Fig. 5 gives results for the three types of outcomes in this study. None of the tested genes showed the positive trend observed with hmp overexpression.
Figure 5. Growth ($OD_{550}$) of BW25113 $\Delta hmp::Cm^R$ during NO challenge expressing either an empty TOPO vector (control) or a gene identified in this analysis.

(A) NO-protective $hmp$, (B) potential response gene $yfhA$, and (C) potential response gene $yahN$. 
To evaluate the significance of the linear change over time, we formally tested hypotheses for no slope versus the existence of a non-zero slope using both control and test cases. Table 5 gives results on the significance of the non-zero slope based on p-values. As shown in Fig. 5c, the test run indicates a strong negative slope (p-value = $1.96 \times 10^{-6}$) while the control run (p-value = 1.0) is quite flat giving strong support to our conclusions for the treatment and control cases. Note the p-value is determined from a t-test in classical simple linear regression. The results of analysis are given in Table 5 which indicates three runs for the $ycbR$ and $yfhA$ cases and one run for the $yahN$ case. In summary, focusing on the average results in Table 5, in agreement with PCA results, the three genes indicate significance as all the slopes are indicated to be significant at the 0.002 test level with p-values of 0.00122, 0.00119 and 0.00000196 for $ycbR$, $yfhA$ and $yahN$, respectively. In contrast, none of the control cases have significant slopes for the averaged results with p-values of 0.281, 0.372, and 1.00.

### Conclusion

This work evaluated the ability of the differential PCA technique of Rollins et al. (2006) to find genes that are differentially expressed *between* assay groups. This technique is an extension of the Rollins et al. (2010) technique that focused strictly on ranking genes for a *single* subset of assays with a common biological interpretation. This work provided the first test of this PCA approach’s predictive capabilities by experimental examination of the effect of individually overexpressing candidate genes on growth sensitivity to inhibitory compounds, such as NO. Even though none of the selected high ranked genes had very high ranks in the rank set (RS) from the PCA analysis, three of these ($yahN$, $ycbR$, and $yfhA$) significantly impacted growth sensitivity to NO, showing that the technique can have strong true detection sensitivity (TDS) when a RS is relatively large.

Statistical power and TDS are both directly proportional to the number of correctly identified genes (CIG). Thus, the conclusions about power should also be true for TDS. Like statistical power, TDS can only be determined under known truth. In Rollins et al. (2010) the power of the technique was evaluated in a simulation study and the results compared quite favorably with a currently popular method. Thus, the support for high TDS in this work is consistent with the high power results in Rollins et al. (2010). While much more experimental work is required to establish its effectiveness in a broad range of applications, these results in this article are quite promising, and we, therefore, encourage the use of the proposed method.

While none of the tested genes showed a protective effect against NO challenge, we did find that three of the tested genes, $ycbR$, $yfhA$ and $yahN$, had a significant impact on *E. coli*’s NO sensitivity. In the time since...
our original analysis, additional data regarding the function of these genes have become available. *ycbR* encodes a putative chaperone protein for a fimbrial operon, meaning that it plays a role in assembling surface proteins that mediate attachment of bacteria to environmental surfaces (Korea et al 2010). *yfhA*, now known as *glmY*, regulates expression of the small RNA *glmS*, which in turn regulates expression of glucosamine-6-phosphate synthase *GlmS* (Reichenbach et al. 2009). Sequence-based analysis predicts that *yahN* encodes a neutral amino acid efflux system (The UniProt Consortium 2009). While the possible link between *ycbR*, *yfhA* and NO sensitivity is not clear, it can be proposed that *yahN*-dependent export of neutral amino acids would exacerbate the branched-chain amino acid starvation caused by NO, because all three branched-chain amino acids are neutral. However, further research will be needed to understand the relationship between these genes and the growth sensitivity of *E. coli* to NO, which is beyond the scope of this work.

**Abbreviations**

NO, nitric oxide; GSNO, S-nitrosoglutathione; PCA, Principal Component Analysis; PM, proposed method; \( T_{\text{diff}} \), absolute different in gene contribution of two assay groups; *E. coli*, *Escherichia coli*; EG, Eigengene; EA, Eigenassay; TFs, transcription factors; RNS, reactive nitrogen species; DeaNO, diethylamine nitric oxide; IPTG, Isopropyl β-D-1-thiogalactopyranoside; RS, rank set; ODS, observed detection sensitivity is missing; TDS, true detection sensitivity; CIG, correctly identified genes.

**Acknowledgements**

We wish to acknowledge experimental contributions by undergraduate researchers Jake Gillian and Kara Moeller. Funding for the experimental component of this work was provided through the Iowa State University Office of Biotechnology. Funding for the statistical component of this work was provided through a grant from the Army Research Laboratory under Cooperative Agreement W911NF-08-0036.

**References**

Bogdan C (2001) Nitric oxide and the immune response. Nature Immunology 2:907-916.

Brynildsen MP and Liao JC (2009) An integrated network approach identifies the isobutanol response network of *Escherichia coli*. *Molecular Systems Biology* 5.

Flatley J, Barrett J, Pullan ST, Hughes MN, Green J, Poole RK (2005) Transcriptional Responses of *Escherichia coli* to S-Nitrosoglutathione under Defined Chemostat Conditions Reveal Major Changes in Methionine Biosynthesis. *The Journal of Biological Chemistry* 280: 10065–10072.

Holland SM (2008) Principal Component Analysis. *UGA Stratigraphy Lab*. http://www.uga.edu/strata/software/pdf/pcaTutorial.pdf. Accessed 05 March 2010.

Huret JL (2010) Atlas of Genetics and Cytogenetics in Oncology and Haematology. INIST. http://atlasgenesisoncology.org. Accessed 05 March 2010.

Hyduke DR, Jarboe LR, Tran LM, Chou KJY, and Liao JC (2007) Integrated network analysis identifies nitric oxide response networks and dihydroxyacid dehydratase as crucial target in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 104: 8488-8489.

Jarboe LR, Hyduke DR, Tran LM, Chou KJY, and Liao JC (2008) Determination of S-nitrosoglutathione targets and response networks in *Escherichia coli* using integrated biochemical and systems analysis. *The Journal of Biological Chemistry* 283: 5148-5157.

Jarboe LR, Hyduke DR, Liao JC (2010) Nitric Oxide: Systems approaches to unraveling nitric oxide response networks in prokaryotes. In: Ignarro LJ (ed) Biology and pathobiology, 2nd edn. *Academic Press*, California, pp 103-136.

Jarboe LR, Liu P, and Royce LA (2011) Engineering inhibitor tolerance for the production of biorenewable fuels and chemicals.
Data Mining Based on Principal Component Analysis: Application to the Nitric Oxide Response in Escherichia coli

Johnson RA and Wichern DW (2008) Applied Multivariate Statistical Analysis. 6th edition. Pearson Prentice Hall (Engelwood Cliffs, NJ).

Korea CG, Badouraly R, Prevost MC, Ghigo JM, Beloin C (2010) Escherichia coli K-12 possesses multiple cryptic but functional chaperone-usher fimbriae with distinct surface specificities. Environmental Microbiology 12(7): 1957-1977.

MacMicking J, Xie QW, Nathan C (1997) Nitric oxide and macrophage function. Annu. Rev. Immunol. 15:323-350.

Miller EN et al. (2009) Furfural Inhibits Growth by Limiting Sulfur Assimilation in Ethanologenic Escherichia coli Strain LY180. Applied and Environmental Microbiology 75(19): 6132-6141.

Misra JW, Hwang D, Hsiao LL, Gullans S, Stephanopoulos G (2002) Interactive exploration of microarray gene expression patterns in a reduced dimensional space. Genome Res 12: 1112-1120.

Raychaudhuri S, Stuart JM, Altman RB (2000) Principal Component Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. Pacific Symposium on Biocomputing 5: 452-463.

Reichenbach B, Gopel Y, Gorke B (2009) Dual control by perfectly overlapping $\sigma^{54}$- and $\sigma^{70}$-promoters adjusts small RNA GlmY expression to different environmental signals. Molecular Microbiology 74(5): 1054-1070.

Rodionov DG, Ishiguro EE (1997) Ampicillin-induced bacteriolysis of Escherichia coli is not affected by reduction in levels of anionic phospholipids. FEMF Microbiology Journal 156: 85-89.

Rollins DK, Sr., Zhai D, Joe AL, Guidarelli JW, Murarka A, Gonzalez R (2006) A novel data mining method to identify assay-specific signatures in functional genomic studies. BMC Bioinformatics 7:377.

Rollins DK, Teh A (2010) An extended data mining method for identifying differentially expressed assay-specific signatures in functional genomic studies. BMC BioDataMining 3:11. doi:10.1186/1756-0381-3-11.

Sharov AA, Dudekula DB, and Ko MSH (2005) A web-based tool for principal component and significance analysis of microarray data. Bioinformatics 21: 2548-2549.

Synder JA, Haugen BJ, Buckles EL, Lockatell CV, Johnson DE, Donnenberg MS, Welch RA, Mobley HLT (2004) Transcriptome of Uropathogenic Escherichia coli during Urinary Tract Infection. Infection and Immunity 72:6373.

The UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Research 37: D169-D174.

Voskuil MI, Schnappinger D, Visconti KC, Harrell MI, Dolganov GM, Sherman DR, Schoolnik GK (2003) Inhibition of respiration by nitric oxide induces a mycobacterium tuberculosis dormancy program. The Journal of Experimental Medicine 198: 705-713.