Global Transcriptional Analysis of an *Escherichia coli* Recombinant Protein Process during Hypoxia and Hyperoxia

by

William Bryon Perry

B.S. Chemical Engineering
B.A. Biochemistry
University of Colorado, Boulder, 1998

M.S. Chemical Engineering Practice
Massachusetts Institute of Technology, 2000

SUBMITTED TO THE DEPARTMENT OF CHEMICAL ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY OF CHEMICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[September 2004]
JUNE 2004

Signature of Author: _____
Department of Chemical Engineering
June 24, 2004

Certified by: _____
Charles L. Cooney
Professor of Chemical and Biochemical Engineering
Thesis Supervisor

Accepted by: _____
Daniel Blankschtein
Professor of Chemical Engineering
Chairman, Committee for Graduate Students

Global Transcriptional Analysis of an *Escherichia coli* Recombinant Protein Process during Hypoxia and Hyperoxia

by

William Bryon Perry

Submitted to the Department of Chemical Engineering on June 24, 2004
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy of Chemical Engineering

ABSTRACT

Both exposure to oxygen and recombinant protein production are known to have adverse effects on microbial fermentation, including increased proteolytic and oxidative damage to the product. In an effort to characterize the effects of these stresses on the cell, DNA microarrays were used to monitor global gene expression of *E. coli* producing recombinant human α1-antitrypsin (α_1AT) during exposure to defined aeration conditions. Recombinant α_1AT has been shown to undergo oxygen-dependent degradation during production in *E. coli*, due in part to activation of the heat-shock response. The goal of this work is to better understand the effects of oxygen in order to improve this recombinant protein production process.

In order to study the effects of oxygen extremes, global expression analysis was performed on α_1AT-producing cultures exposed to pure nitrogen, air, and pure oxygen. The most notable effects of oxygen exposure were those of superoxide. This reactive oxygen species is generated upon oxygen exposure and is known to oxidize iron-sulfur clusters. In response to hyperoxic conditions, the SoxRS stress response was activated, as were genes involved in iron uptake and the Isc and Suf repair systems for Fe-S clusters. Supplementation of iron in the growth medium resulted in expression changes consistent with improved formation of Fe-S clusters. Iron supplementation also decreased superoxide stress at the expense of a short-term increase in the peroxide (OxyR) stress response. In addition, iron supplementation dramatically reduced the oxygen dependence of recombinant α_1AT degradation. Regeneration of Fe-S clusters is proposed to improve protein folding and limit activation of the heat-shock response.

The effects of recombinant protein production were observed through expression analysis of induced, uninduced, and Empty-Vector cultures. As expected, recombinant α_1AT production led to increased expression of heat-shock genes, including proteases and chaperones that are known to be involved in α_1AT degradation. Based on expression analysis data, production of recombinant α_1AT also resulted in catabolite repression and decreased amino acid biosynthesis.

This work demonstrates the utility of DNA microarrays in analyzing and improving microbial fermentations. Global expression studies have suggested several strategies for increasing the resistance of bioprocesses to the damaging effects of oxygen and recombinant protein production.

Thesis Supervisor: Charles L. Cooney
Title: Professor of Chemical and Biochemical Engineering

Acknowledgements

The funny thing about graduate school is that you enter suspecting that you don't know very much, and you leave knowing exactly how little you know. Completing this document required help from a lot of people who knew more than I did on various topics.

I am very grateful to my advisor Charlie Cooney, whose balance of patience and drive provided an environment in which I was able to fail, learn, and ultimately succeed. I am particularly thankful for incredible opportunities to travel and speak about my research, as well as the freedom to explore professional development opportunities outside of the lab. I was very fortunate to have a wonderful thesis committee, consisting of Professors Alan Grossman, Greg Stephanopoulos, and Dane Wittrup. My committee not only took the time to gain a deep understanding of my work, but also provided invaluable feedback and insight.

I would also like to thank the undergraduate researchers that contributed to the overall understanding of *E. coli* gene expression. Jina Sinskey contributed to the initial full-genome DNA-microarray analyses and helped to define the final protocol. She also applied this protocol to some of the samples presented in Chapter 5. Swapna Panuganti explored improved normalization techniques as well as the effects of cysteine addition on *E. coli* cultures. John Liu contributed to the analysis of protein extracts during long growth experiments.

I would also like to acknowledge several labs that contributed material to this work. Dr. Susan Lovett and Vincent Sutera at Brandeis University performed 4,000+ PCR's and generously donated the products, which were eventually spotted on our DNA microarrays. Dr. Susan Gottesman at the National Cancer Institute donated the *E. coli* BL21 ClpA^{-} strain and Dr. Myeong-Hee Yu at the Korea Research Institute of Bioscience and Biotechnology for donating the original pEAT8 α1-antitrypsin plasmid. I am also grateful to the NIH Biotechnology Training Program for financial support.

I am very appreciative of Professor Jonathan King, Cammie Haase-Pettingel, Claire Ting, Jacqueline Piret, Peter Weigele, Welkin Pope, and Shannon Flaugh who generously welcomed me into their lab and answered my biology questions. I also appreciate help from the laboratories of Greg Stephanopoulos, Dane Wittrup, and Daniel Wang in generously sharing materials and equipment. Saliya Silva and Kohei Miyaoku deserve thanks as well for their guidance in performing amino acid analysis. I am very thankful for the help of Brett Roth, Patsy

Sampson, Janet Fischer, Elaine Aufiero, Suzanne Easterly, Jennifer Shedd, Mary Keith, and Darlene Ray for thinking of every detail and allowing me to focus on my research.

Several groups were particularly helpful in passing along their knowledge of DNA microarrays and applications for prokaryotic systems: from Alan Grossman's lab, Rob Britton, Elke Kuester-Schoeck, Natalia Comella, and Jennifer Black; from Greg Stephanopoulos' lab, Ryan Gill, Bill Schmitt, Jatin Misra, and Ilias Alevizos; from Anthony Sinskey's lab, Andrea Loos and Philip Lessard; and from the MIT BioMicro Center, Sean Milton. Their patience and guidance allowed me to find my way in this new field of research.

As a new graduate student, I was particularly grateful to Mike Laska for his help with fermentations, his guidance with my project proposal, and afternoons at the Muddy. At the other end of my time at MIT, I appreciated editing help from Filipe Mergulhao and Brian Mickus on the final draft of this thesis. In addition to these, other members of the Cooney Lab, past and present, have provided technical guidance and friendship: Steve Griffiths, Junfen Ma, Reuben Domike, Brian Baynes, Asti Goyal, Samuel Ngai, Yu Pu, Ben Waters, Jared Johnson, Lakshman Pernenkil, Jean-François Hamel, G.K. Raju, C.K. Lai, May Sun, Maria-José Ibanez, Alex Sotiriadis, Rasmus Bjerre-Nielsen, Kilian Aviles, Arno Biwer, and Aleks Engel.

Finally, I would like to acknowledge my parents and my family for providing the foundation that allowed me to meet the challenges of this project and for their love and support throughout this process. I am also extremely grateful to my wife Anne for standing by my side and for believing in me during all the times when I couldn't do the same.

Table of Contents

List of Figures

List of Tables

1 Introduction

> "All you have to do is pick up a baseball. It begs to you: throw me. If you took a year to design an object to hurl, you'd end up with that little spheroid: small enough to nestle in your fingers but big enough to have some heft, lighter than a rock but heavier than a hunk of wood. Its even, neat stitching, laced into the leather's slippery white surface, gives your fingers a purchase. A baseball was made to throw. It's almost irresistible."
>
> ***—Dave Dravecky***

1.1 DNA Microarrays

DNA microarray technology and the promise it held was initially met with a great deal of excitement by the scientific community. While this initial excitement has dwindled, as realism has set in, the technology has made a significant contribution to areas such as cancer research (Golub *et al.* 1999), microbiology (DeRisi *et al.* 1997), and biotechnology (DeLisa *et al.* 2001). This tool provides relative quantification of thousands of sequences within a nucleic acid pool all at the same time. The high-throughput nature of this technology is extremely powerful. Most commonly, microarrays are applied to analysis of gene expression, in which the pool of mRNA transcripts within the cell is characterized.

Developed in the mid-1990's by both the Patrick Brown lab at Stanford (Schena *et al.* 1995) and Affymetrix® (Chee *et al.* 1996), DNA microarrays consist of a substrate—typically glass, quartz, or nylon membrane—spotted with DNA. Each spot contains single-stranded DNA molecules, or probes, covalently attached to the substrate. Each probe corresponds to a different gene or region of interest. When a labeled nucleic acid solution is introduced to the array, the probes bind to complementary sequences in this labeled sample. By detecting the labels on the substrate, it is possible to quantify the amount of nucleic acid bound to each spot. Thus, DNA microarrays give information about the sequences present in the nucleic acid sample as well as the amount of each sequence.

DNA microarray experiments are similar to Northern Blot hybridizations in that binding of probe to target allows quantification of the target nucleic acid. As with any binding assay, both of these techniques must have excess probe in order to accurately quantify levels of the target molecule. These two techniques differ in that, for microarray experiments, the *target* is

labeled and is in solution, while for Northern Blots, the *probe* is labeled and is in solution. For this reason, some confusion exists in the probe/target terminology. Throughout this document, the term *probe* will refer to the substrate-bound DNA, while *target* will refer to the labeled nucleic acid sample. As described below, another difference is that DNA microarrays typically use a cDNA target, rather than an RNA target. Of course, the major difference between Northern Blots and DNA microarrays is that Northern Blots allow quantification of only a handful of transcripts per experiment, while DNA microarrays can quantify thousands of transcripts in a single experiment.

1.1.1 Manufacture of DNA Microarrays

There are several options in the manufacture of DNA microarrays. In general, the two types are oligonucleotide (oligo) arrays and cDNA arrays, each using a different type of probe. In addition, arrays can be produced either by spotting DNA onto slides or by photolithographic printing. These different manufacturing options are briefly described below.

1.1.1.1 Chips and Slides

The different manufacturing methods for DNA microarrays have created two distinct types of microarrays. The term "chips" refers to Affymetrix GeneChips®, which are produced by photolithographic printing. This method involves synthesizing DNA molecules of 25 nucleotides directly on the substrate (Figure 1.1). Photolithographic masks are applied to the substrate to activate particular regions of the chip. Free nucleotides react only in the activated regions to extend the DNA chain one link at a time. This technique is exclusively applied to making oligo arrays. For this type of printing, the gene sequence must be known in order to design not only oligos that will specifically bind, but also the masks.

Both cDNA arrays and oligo arrays can also be produced by directly spotting DNA samples onto glass slides. Robotic arrayers can print DNA samples from 96- or 384-well plates at a density of roughly 15,000 spots per 75 mm × 25 mm slide. Typically, these arrayers can print at least 100 slides per run. The pins used for this operation are finely machined to release volumes consistently below 1 nL. Robotic arrayers can print not only oligo arrays, but cDNA arrays as well.

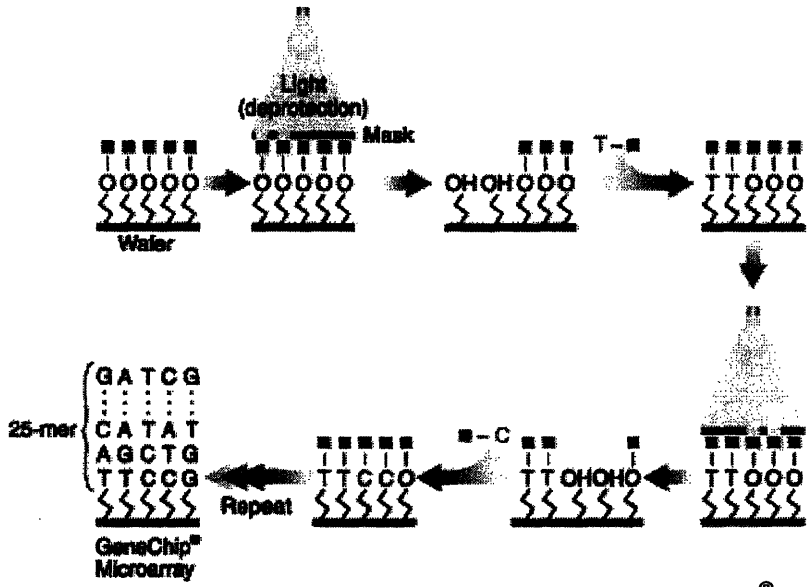


Figure 1.1: Photolithographic Printing of Affymetrix GeneChips®
Affymetrix GeneChips® are manufactured by a series of steps involving light exposure through a mask to deprotect a particular region on the substrate, and reaction with protected nucleotides to extend the DNA chain one link at a time (Affymetrix® 2004).

1.1.1.2 Probe Preparation

DNA microarrays on glass slides can be produced by using a variety of different probes. Probes for spotting can be designed based on knowledge of the organism's genomic sequence, but an annotated genomic sequence is not required. Expressed sequence tags (EST's) or selected cloned sequences can be used to generate cDNA probes (DeRisi *et al.* 1996). This option does not require any knowledge of the genomic sequence; however, the ability to connect each spot to an annotated gene is extremely valuable. Another drawback of EST's is the low specificity that can result if a particular clone sequence spans multiple transcripts.

Probes can also be generated by performing Polymerase Chain Reaction (PCR) using a genomic DNA template. The primers for these reactions are generated using phosphoramidite chemistry and are designed based on the genomic sequence of the organism, such that they bind to either end of each open reading frame (ORF) in the genome. An ORF is the part of the gene that contains the transcript information, essentially beginning at the start codon and ending at the stop codon. An ORF may or may not have an expressed protein product—those with no identified product are said to produce a "hypothetical protein." As probes, PCR products have several inherent disadvantages. As with EST's, there are specificity issues. Several ORF's

within a given genome might have a high degree of similarity. Therefore, using the full ORF as a probe could potentially result in a large amount of cross-hybridization (Richmond *et al.* 1999). In addition, PCR products consist of two complementary strands, both of which will act as probes. While the technical issue of separating these strands prior to printing has been overcome (*e.g.* by printing in a reducing buffer such as 50% DMSO), these two strands can potentially result in decreased specificity. It is possible for genes transcribed in opposite directions to overlap. Therefore, a spot corresponding to one of these genes might also detect the other.

Oligonucleotide probes overcome some of these specificity issues. Oligos for spotting can be synthesized at a relatively inexpensive cost with the commonly used phosphoramidite chemistry. The sequences of these 50-70mer probes can be carefully designed to correspond to regions that are unique to a particular gene. Therefore, the specificity can be significantly better from oligo probes than from PCR products (Kane *et al.* 2000). In terms of sensitivity, oligo arrays have been found to be just as good as cDNA arrays.

1.1.2 Differential Gene Expression Experimental Methods

The most common DNA microarray experiment is the analysis of Differential Gene Expression (DGE), which allows investigators to probe expression of the same genome under a variety of conditions. A typical DGE experiment is performed using two-color hybridization with two different fluorescent labels (Figure 1.2). First, samples are collected from cells grown under different conditions. For instance, these samples could be *E. coli* grown at different temperatures or cells from human muscle, liver, and skin tissue. Next, total RNA is purified from each of these samples. In some cases, mRNA is enriched from total RNA; however, this separation may not always be easy to achieve, as discussed later. Once RNA is purified, the most common labeling scenario is that RNA from each sample is reverse transcribed to cDNA using labeled nucleotides.

The final product of this reaction is a labeled nucleic acid sample that is ready to hybridize to the DNA microarray. If a complementary sequence of DNA appears in one of the spots on the array, the labeled nucleic acid will bind the probes in that spot. When the array is scanned, the amount of label in each spot is quantified. If the same label is used for both samples, which is typically the case with radioactive labels, the two samples are hybridized to separate arrays. The signals at each spot are subtracted from one another to determine which

genes are highly expressed under each of the test conditions. It is more common, however, to use two different fluorescent labels, which allows the two samples to be simultaneously hybridized to a single array. In this case, the array is scanned at two different wavelengths (or channels) to determine the expression under each condition, and the images are merged to determine differential expression. Figure 1.2 illustrates this two-color hybridization scenario.

In an effort to validate these microarray techniques, initial studies compared the results from microarrays to those from more established techniques such as Northern blots (Pomposiello *et al.* 2001; Taniguchi *et al.* 2001). These studies found that transcripts expressed at low levels were more likely to be detected by Northern Blots than by microarrays. However, the benefit of the high-throughput analysis outweighs this drawback.

1.1.3 Application of Experimental Methods to Prokaryotic Systems

Application of array technology to prokaryotic organisms was initially hampered by the question of mRNA enrichment from total RNA. Although mRNA comprises only a small fraction of the total cell RNA (Voet and Voet 1995), it is unclear whether removal of ribosomal RNA gives a significantly higher signal on the arrays. This step is much easier with eukaryotic organisms because their mRNA has a distinguishing feature that can be used to separate it from total RNA. Most eukaryotic mRNAs contain a 3'-poly(A) tail of 20-50 nucleotides (Voet and Voet 1995). To separate these mRNA molecules, total RNA is purified, or selectively reverse transcribed using oligo(dT) primers. Prokaryotes, on the other hand, do not have a distinctive marker for mRNA molecules. As a result, investigators have been forced to either use methods to copy only the mRNA sequences or use total RNA.

Some of the approaches that have been taken to purify mRNA from total RNA are the mRNA-specific primers method (Fislage *et al.* 1997), differential expression by customized amplification library (Alland *et al.* 1998), and subtractive hybridization (Plum and Clark-Curtiss 1994). Another approach is to selectively reverse transcribe rRNA species to generate cDNA:rRNA complexes, which can be selectively degraded by RNase H and DNase I (Rosenow *et al.* 2001). Yet another approach takes advantage of the ability of poly(A) polymerase I to selectively modify the 3' termini of mRNA over rRNA (Wendisch *et al.* 2001). The mRNA species, modified to have poly(A) tails, can then be treated as eukaryotic RNA samples, *i.e.* purified or selectively labeled using oligo(dT) primers.

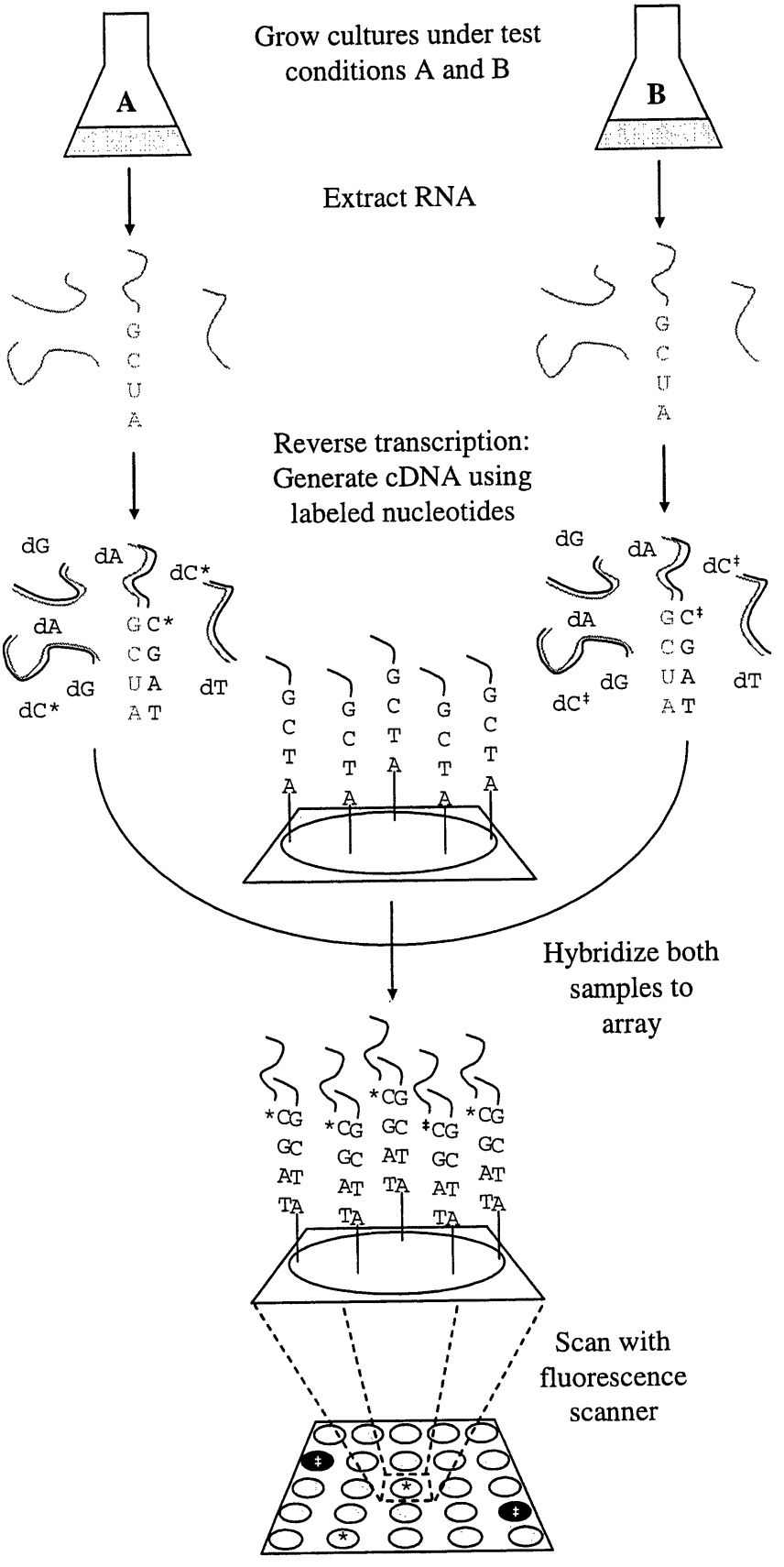

Grow cultures under test conditions A and B
A
B
Extract RNA
Reverse transcription: Generate cDNA using labeled nucleotides
Hybridize both samples to array
Scan with fluorescence scanner

Figure 1.2: A Typical Differential Gene Expression Experiment
(opposite) Cultures are grown under conditions A and B, and RNA is extracted from both. These RNA samples are used to generate cDNA in a reverse transcription reaction, which incorporates labeled nucleotides. Two different fluorescent labels are used in this scenario (illustrated as * for the culture-A sample and ‡ for the culture-B sample). Following degradation of the RNA template, the labeled cDNA samples are simultaneously hybridized to the array. If the complementary sequence exists in one of the spots on the array, the labeled cDNA will bind to that spot. By scanning with a fluorescence scanner, each label is detected and the relative amount of each sequence in the original RNA sample is thereby quantified. Although only four nucleotides are shown here, the complementary binding regions are actually much longer.

To use total RNA for microarray analysis, random primers, instead of poly(dT) primers, are used in the reverse-transcription labeling. This method has several drawbacks. First and foremost, this method cannot distinguish mRNA from rRNA; therefore rRNA molecules, which have no complementary sequences on the array, will be labeled and will non-specifically bind to both spots and background. Another disadvantage of labeling with random primers is that it tends to favor the 5' end of the transcript. Reverse transcriptase starts its reaction at the primer and moves toward the 5' end of the transcript until the end of the strand is reached. Primers that bind at the 3' end of the transcript will presumably result in a complete copy. However, most primers will bind somewhere in the middle and may not copy the 3' end of the transcript. As a result, genes from the same transcription unit may not show equivalent signal. Despite these difficulties, several studies have shown that total RNA gives reasonably good results with DNA microarrays. For example, several experiments using Affymetrix® arrays were performed using total RNA and high hybridization backgrounds were not observed (de Saizieu *et al.* 1998). Other experiments used total RNA on custom-made arrays with no mention of difficulty in interpreting the signals (Richmond *et al.* 1999; Tao *et al.* 1999). These experiments understandably raise questions as to the necessity of mRNA purification; nonetheless, such techniques are available if needed.

1.1.4 Analysis of DNA Microarray Data

The goals of microarray data analysis are to convert large numbers of signal measurements to meaningful measurements of transcript levels and to discover patterns in those transcript levels. There are a variety of methods for performing microarray data analysis. However, to this point, no single method has been universally adopted. In fact, each method

comes with its own requirements and assumptions and may not be universally applicable. This section reviews methods for achieving these goals and ends with methods used to analyze the reproducibility of microarray experiments.

1.1.4.1 Evaluating Spot Quality

An unavoidable fact of cDNA microarray experiments is that not every spot will produce high quality data. Therefore, before subsequent analysis can be performed, low quality data should be identified and treated accordingly.

One standard approach to identifying low-quality spots is to examine signal-to-noise ratios for each spot (Chen *et al.* 2002). These ratios are typically calculated as follows:

$$\text{Signal-to-Noise Ratio} = \frac{(I_F)_{Med} - (I_B)_{Med}}{\sigma_B} \tag{1.1}$$

In this equation, σ_B is the standard deviation of the background pixel intensities, while $(I_F)_{Med}$ and $(I_B)_{Med}$ are the median intensities of all pixels in the feature and background regions, respectively. Typically, a signal-to-noise ratio of at least 3 is required for high quality data. Another approach is to perform a *t*-test to determine whether the mean intensity of the feature pixels is significantly larger than the mean intensity of the background pixels (Yang, M. C. K. *et al.* 2001).

Once these low-signal spots have been identified there are a few different ways to treat them. Most commonly, the data are removed from further analysis. However, another common approach is to set these signals to a default value of either zero or some other value determined by the signal from negative controls on the array.

1.1.4.2 Normalization

Normalization can be performed within a single array or between multiple arrays. Performed within a single array, the goal is to balance the Cy3 and Cy5 signal intensities so that comparisons between the two samples can be made. Obviously, this application is needed when each channel contains a sample of interest. This will not always be the case, however. An alternative experimental design for comparing a large number of samples would be to place a sample in one channel and a hybridization control in the other. This hybridization control can be another RNA sample or a genomic DNA sample. Selection of hybridization controls is discussed further in Section 4.2. For this discussion, the important point is that comparisons

between the two channels do not need to be balanced in this case, but comparisons between multiple slides do. Across multiple arrays, the ratio of the Cy3 and Cy5 signals is typically balanced by normalization. Note that if a sample of interest is used as a hybridization control (*e.g.* a zero-time-point sample), normalizations within a single array and between multiple arrays should be performed.

One decision that must be made in normalization is which spots to use. Normalization can be performed on either a small subset of the spots on the microarray or all of them. The latter strategy is referred to as a global normalization. Either method involves a unique set of assumptions. Normalization with a small subset of genes might involve doping the labeling reaction with known quantities of a foreign transcript, which should bind only to corresponding spots of foreign DNA that have been spotted on the microarray. In this case, it must be assumed that there is minimal cross-hybridization between the labeled cDNA from the organism of interest and the foreign spots, and vice versa. These assumptions can easily be checked. Normalization may also be performed on a set of "housekeeping" genes that are assumed to maintain constant expression levels throughout a variety of experimental conditions.

Global normalization methods make two assumptions. First, they assume that, although expression of many genes may be changing, the overall changes balance one another such that overall gene expression remains constant. Second, they assume that constant overall gene expression translates to constant overall signal on the microarrays. It is important to recognize that each probe on the array has different binding characteristics (*i.e.* GC content, melting temperature, *etc.*); therefore, each spot has a unique signal : concentration calibration. While some samples may violate this first assumption (such as samples from hypoxic cultures), the global normalization may still be applied with the caveat that the resulting expression values are relative to total expression. It should be noted that global normalizations are only valid for full-genome microarrays. Microarrays that only contain a small number of genes should never be normalized globally—especially when those genes are selected based on a similar empirical response—because the assumptions are unlikely to hold.

Even after the set of genes has been defined, several options exist for performing the normalization. For example, in the case of a single-array comparison, the signal values for each channel can be adjusted so that the mean or total signals for each channel are equivalent. The signals from the two channels are commonly compared by plotting the Cy3 signal (G) and Cy5

signal (R) on two separate axes, as a log(R) vs. log(G) scatter plot. An alternative representation is a scatter plot of M (log(R/G)) vs. A ($\log(\sqrt{RG})$), which is essentially the previous scatter plot rotated by 45° (Dudoit *et al.* 2002). This visualization tool more clearly accounts for the effects of overall signal intensity, A. A more complex normalization option uses regression techniques such as LOWESS smoothing to fit the two signals so that they have a smooth linear relationship. Additionally, Analysis of Variance (ANOVA) techniques can be used to account for various experimental parameters, such as array effects, and correct the data accordingly (Tseng *et al.* 2001; Yang, Y. H. *et al.* 2001 review these methods for normalization.).

Normalization can potentially have a large impact because it can skew the data set in a way that can either clarify or obscure the apparent differential expression.

1.1.4.3 Selection of Differentially Expressed Genes

For many, the ultimate goal of microarray analysis is to produce a list of genes that are up-regulated and down-regulated in response to a particular stimulus. An overarching problem throughout this portion of the data analysis is the lack of replication. These experiments are difficult to perform well, and reagents and materials for these experiments are expensive. Much of the focus in this area has been to determine how significant expression differences can be selected with only two or three replicates.

The simplest method for selecting differentially expressed genes is to apply a fold-cutoff to the normalized data. Typically, a two-fold cutoff is used since this is generally assumed to be the limit of detection as well as the minimal bound for biological significance (DeRisi *et al.* 1997). However, this approach is generally recognized as being flawed because each gene will have a unique variance in its expression. For instance, a two-fold cutoff may not be significant to genes with naturally large variance.

Another approach has been to use statistical tests, such as *t*-tests, to evaluate each gene independently and define gene-specific cutoffs (Chen *et al.* 1997; Tusher *et al.* 2001). An analogous method applying Bayesian probabilities has also been developed (Baldi and Long 2001; Long *et al.* 2001). ANOVA methods can also be applied to determine the variances in expression in response to a particular treatment. Gene-by-gene comparisons of this treatment variance to the variance in the random (residual) error reveal whether the expression change is significant (Wolfinger *et al.* 2001; Cui and Churchill 2003). Some of these methods assume that

the signal values are normally and identically distributed, which is a good assumption for log-transformed data.

Another set of methods improves upon the above models by not assuming normal distributions in microarray signals from repeated experiments (Baggerly *et al.* 2001). Yet another method uses a bootstrap, combined with an ANOVA model, to generate simulated data sets with the same distribution as the original (Kerr *et al.* 2000). Based on these simulated data sets, confidence intervals are obtained for each of the expression changes.

Methods for identifying differential expression have advanced beyond the simple two-fold cutoffs. Several methods are available to choose from, each with a unique set of assumptions.

1.1.4.4 Pattern Discovery

The next level of microarray data analysis is using these quantitative measurements of gene expression to draw conclusions about a larger system, *e.g.* a particular pathway or regulon, or even the cell as a whole (Ideker *et al.* 2001). The options for further analysis of microarray data are many. In general, there are two types of methods: unsupervised methods, in which each gene is treated identically, and supervised methods in which expression data are combined with *a priori* knowledge of the cellular physiology of the organism. The most commonly used unsupervised method for pattern discovery is hierarchical clustering (Eisen *et al.* 1998). Based on some metric of similarity (typically a Pearson correlation coefficient), this method identifies genes that exhibit similar expression patterns.

1.1.4.5 Evaluation of Experimental Error

Experimental error can be quite large in DNA-microarray data sets. This error unavoidably affects the conclusions that can be drawn from these data sets. For example, a data set with severe error may only be able to identify a 4-fold change in expression. Therefore, it is important to estimate this error and understand its sources. Several approaches have been taken to accomplish this.

One study decomposed an Affymetrix GeneChip® experiment into two parts: sample preparation and hybridization and designed experiments to evaluate the noise in each step (Tu *et al.* 2002). The hybridization noise was found to dominate, and genes with the lowest expression were found to exhibit the highest noise. Another study used identical samples to compare signal

from spots that were (1) on the same array, (2) from the same labeling reaction on different arrays, and (3) from different labeling reactions on different arrays (Loos *et al.* 2001). On average, the coefficient of variation between spots was found to be 7.1% across all conditions. Another study found that the identifying differential expression from a single microarray experiment was poorly reproducible and could lead to false positives and false negatives accounting for as much as 10% of the total number of genes (Lee *et al.* 2000). The frequency of misclassification decreased significantly with increased replication. The authors recommend that each experiment be performed at least three times.

1.1.4.6 Summary of DNA Microarray Data Analysis

Software packages like SpotFire DecisionSite™ perform most of the data analysis steps described here. While many options exist in the literature, methods used by these software packages will naturally become the most common methods in practice. As use of these packages increases, these methods will become increasingly standardized. Many of the major issues in microarray data analysis have been addressed over the past five years; therefore, future improvement on the methods described here will likely be incremental.

1.1.5 Summary of DNA Microarray Methods

The new technology of DNA microarrays allows investigators to examine gene expression of several thousand genes in a single experiment. The high-throughput nature of this assay makes it an extremely valuable tool in microbiology. While the experimental methods described here are fairly standard, they are expensive and time consuming. Today, the cost of reagents and supplies costs well over $30 for analysis on a single spotted DNA microarray (ignoring capital costs). Double that cost for analysis with Affymetrix GeneChips®. Considering that this only yields a "snapshot" transcriptional profile, there is a great deal to be gained by improvement to the current microarray methods. In the future, use of nanotechnology and silicon nanowires will help to improve both the speed and accuracy of transcriptional analysis (Hahm and Lieber 2004). In addition, nanowires arrays may also allow for real-time analysis of transcript pools. While protocols for experimental and data analysis are now well established, improvements continue to be made.

1.2 Elucidation of Microbial Stress Responses through Expression Analysis

DNA microarrays can be used with *E. coli* to define genes involved in stress responses and observe gene expression during changes in environmental conditions. Of most interest in this work are the effects of protein misfolding as well as the effects of both high and low oxygen levels. This section discusses the current knowledge of each of these cellular states and illustrates how microarrays have been applied to both answer current questions and raise new ones.

1.2.1 Heat Shock Response

Conformational changes in protein structure can cause a great deal of damage inside the cell, including decreased enzymatic activities and altered metabolic fluxes, and may ultimately result in an inability to completely metabolize nutrients. The general response to misfolded proteins includes both an increase in the levels of chaperones to help proteins fold properly as well as an increase in the turnover of proteins by proteolysis. Several genes encoding these chaperones and proteases are regulated by the RNA polymerase sigma factor σ^{32}, which is encoded by the gene *rpoH* (Grossman *et al.* 1987). The individual genes involved in this regulon are reviewed in (Gross 1996).

The σ^{32} regulon is known to be stimulated by stresses such as heat shock, expression of a misfolded protein, and exposure to ethanol. These stresses stimulate the σ^{32} regulon by increasing both the rate of translation of the *rpoH* transcript as well as the stability of σ^{32} (Figure 1.3). The *rpoH* transcript has been shown to have a secondary structure, which limits accessibility of the Shine-Delgarno sequence and the start codon, thereby preventing translation (Morita *et al.* 1999). Exposure to high temperatures leads to increased kinetic motion and tends to remove these secondary structures, which makes more transcripts available for translation and increases the rate of translation. Because the transcript is already present, the only delay in activation of the regulon is in translation.

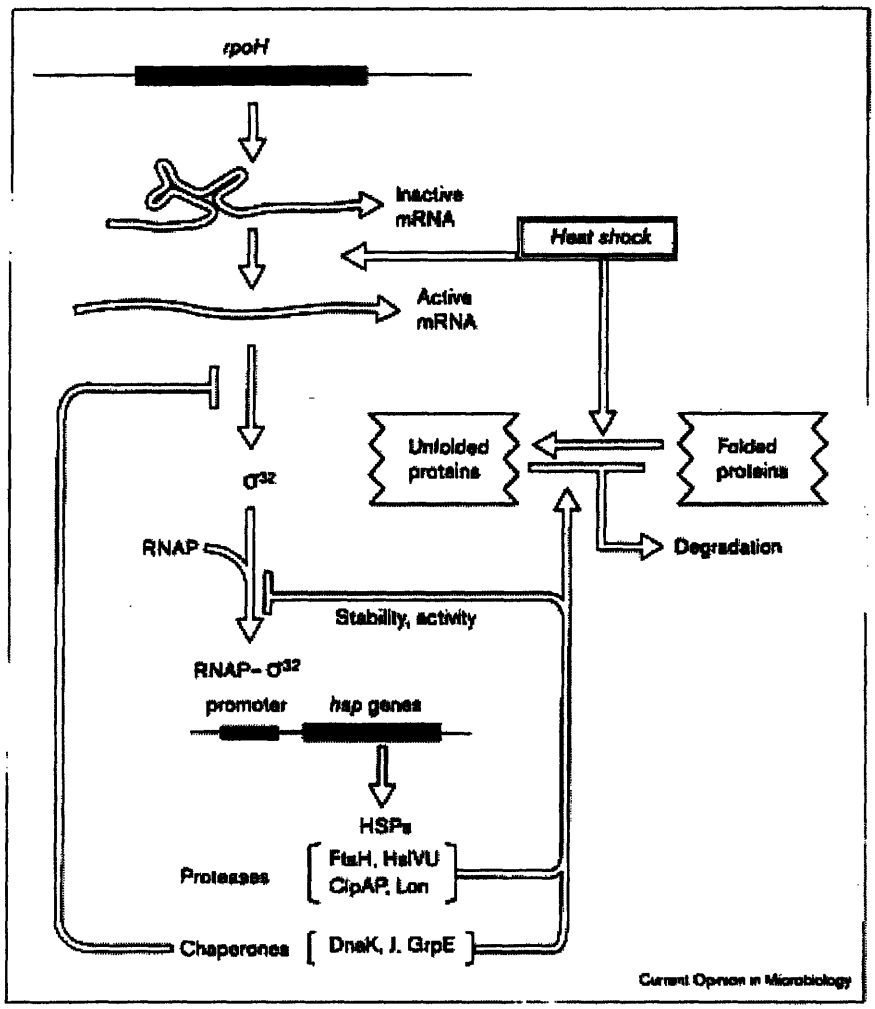


Figure 1.3: Mechanism of *E. coli* Heat Shock Response
Heat shock activates the σ^{32} regulon in two ways. First, it removes secondary structures in the *rpoH* transcript, allowing it to be translated. Second, it increases the level of unfolded proteins, which draw the DnaK/DnaJ/GrpE chaperone complex away from σ^{32}. Without this chaperone complex, σ^{32} is free to bind the RNA polymerase apoenzyme (RNAP) and transcribe its regulon of heat shock proteins (HSPs). Both σ^{32} and unfolded proteins can be degraded by heat shock proteases. Production of unfolded proteins does not affect translation of the *rpoH* transcript (Yura and Nakahigashi 1999 - reproduced with permission).

The chaperone complex of DnaK/DnaJ/GrpE serves as a negative regulator of σ^{32} activation (Straus *et al.* 1990; Bukau 1993). Binding of this complex to σ^{32} likely prevents its association with the RNA polymerase apoenzyme (Tatsuta *et al.* 1998). This chaperone complex also indirectly increases the susceptibility of σ^{32} to degradation by the FtsH (HflB) protease (Tatsuta *et al.* 1998), although this mechanism is not entirely understood. FtsH is the major protease involved in this degradation (Herman *et al.* 1995). Misfolded proteins inside the cell

presumably compete with σ^{32} for access to this chaperone complex (Parsell and Sauer 1989). Therefore, an increase in the levels of misfolded proteins may result in higher levels of free σ^{32}, which could bind to RNA polymerase and activate its regulon. Since all of the proteins involved in inactivation of σ^{32} are heat shock proteins, the activation of σ^{32} is somewhat self-defeating. As the regulon is activated, σ^{32} activity decreases until a new steady state is reached. For a 30°C to 42°C temperature shift, this process was found to occur over a period of 15 min, with σ^{32} levels peaking 4-5 min after the perturbation (Straus *et al.* 1987).

1.2.1.1 Heat Shock

Regulation of the heat shock response has been shown to occur at the levels of both σ^{32} translation and activation. Because heat shock is the best characterized microbial stress response, one of the first microbial high-density microarray experiments used this response to validate the experimental technique in comparison to arrays on a nylon substrate (Richmond *et al.* 1999). Comparing a culture grown at 37°C with another just minutes after a shift to 50°C, this study identified 77 genes induced by heat shock, 23 of which had been previously identified. These experiments also identified the genes *rseA*, *clpA*, and *prlC* as showing temperature-dependent expression. This study further validated the members of the heat-shock stimulon identified in previous work (Chuang and Blattner 1993; Gross 1996). Another 42 genes were found to be repressed by the temperature shift. Overall, 35 ORF's with unknown function were identified as being affected by this temperature change (Richmond *et al.* 1999).

1.2.1.2 Expression of Misfolded Proteins

Expression of misfolded proteins is known to stimulate the heat shock response (Parsell and Sauer 1989; Wild *et al.* 1993). The mechanism by which this occurs is illustrated in Figure 1.3. As this figure indicates, misfolded proteins do not affect the translation rate of the *rpoH* transcript (Kanemori *et al.* 1994).

One DNA microarray study compared the transcriptional profiles from temperature-shifted cultures (Richmond *et al.* 1999) to those from cultures producing both folded (soluble) and unfolded (insoluble) recombinant proteins (Lesley *et al.* 2002). While both cultures showed significant changes compared to the pre-induction sample, it was production of misfolded recombinant protein that showed overlap with the heat-shock expression profile. 21 genes identified as heat shock genes also showed large increases in expression when misfolded proteins

were produced. Leading the list of induced genes were *ibpA* and *ibpB*. These investigators also observed differential expression of several genes encoding ribosomal-associated proteins. A hypothesis was proposed whereby the ribosome senses misfolded protein and stalls translation until chaperones are available. Another study examined transcriptional profiles during production of a heterologous protein and also found strong induction of heat-shock response genes (Rohlin *et al.* 2002). Again, *ibpA* topped the list of genes with significantly increased expression, followed by *dnaK* and *dnaJ*. This study also observed decreased expression of genes involved in glycolytic metabolism and glucose transport, associated with protein production. Genes repressed by protein production included 29 genes involved in amino acid biosynthesis and metabolism, 30 genes involved in central metabolism, and 14 genes involved in nucleotide biosynthesis and metabolism.

1.2.2 Hyperoxic Stress

Exposure to strongly aerobic conditions is known to generate two reactive oxygen species: hydrogen peroxide (H_2O_2) and superoxide ($O_2^{\bullet -}$). Both of these reactive oxygen species are capable of oxidizing macromolecules, but are most damaging in their ability to generate the hydroxide radical ($HO^{\bullet}$) and the hydroperoxide radical ($HOO^{\bullet}$). These two radical species react much more rapidly and with a wide range of molecules, and are therefore highly toxic to the cell. One goal of the hyperoxic stress responses is to consume H_2O_2 and $O_2^{\bullet -}$ and prevent formation of the detrimental reactive oxygen species $HO^{\bullet}$ and $HOO^{\bullet}$.

1.2.2.1 Hyperoxic Stress due to Superoxide

It is well known that superoxide oxidizes iron-sulfur clusters. Iron-sulfur clusters are a protein prosthetic group, consisting of iron atoms coordinated to free sulfur atoms as well as sulfur atoms from cysteine residues. The two most common types of iron-sulfur clusters are [2Fe-2S], and [4Fe-4S], as shown in Figure 1.4. Oxidation of these clusters ultimately results in loss of an iron ion from the cluster. One example of this oxidation reaction is as follows (Imlay 2002):

$$O_2^{\bullet-} + 2\,H^+ + [4Fe\text{-}4S]^{2+} \rightarrow [3Fe\text{-}4S]^+ + Fe^{2+} + H_2O_2 \qquad (1.2)$$

This oxidation can inactivate iron-sulfur enzymes such as dihydroxyacid dehydratase (Kuo *et al.* 1987) which is involved in branched-chain amino acid biosynthesis. Aconitases have also been found to have superoxide-labile iron-sulfur clusters (Varghese *et al.* 2003).

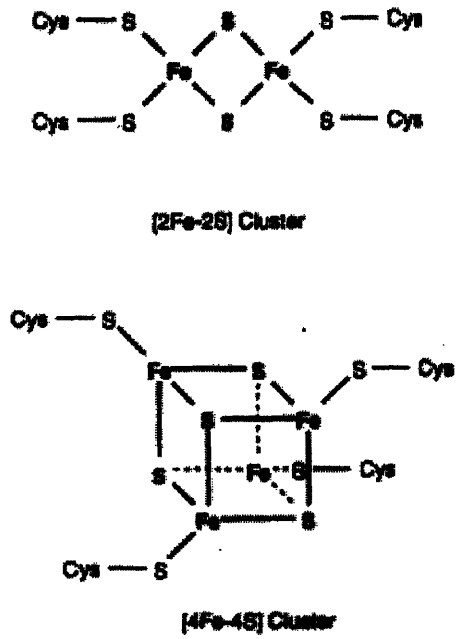


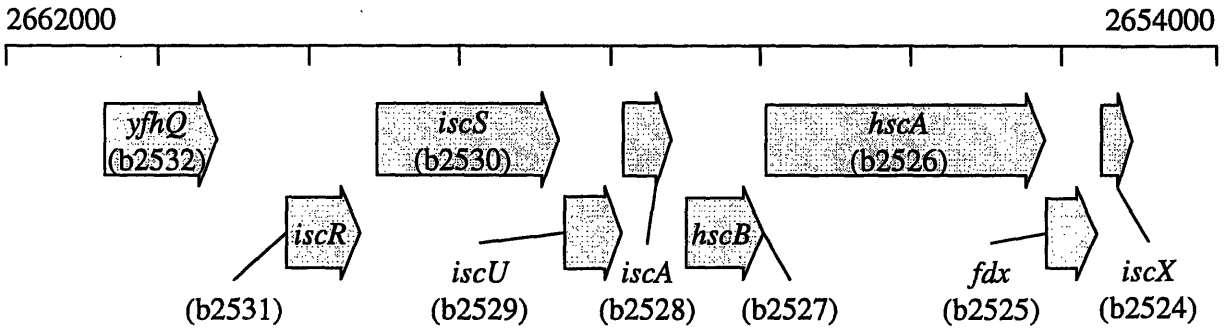
Figure 1.4: Two Iron-Sulfur Clusters in Their Reduced States
[2Fe-2S] and [4Fe-4S] clusters are the most common types of iron-sulfur clusters. Both clusters are composed of free iron and sulfur ions coordinated to four cysteine residues, represented here as Cys—S (Figure from (Gennis and Stewart 1996 - reproduced with permission).

The two-component regulator, SoxR/SoxS, controls the cellular response to superoxide. The homodimeric protein SoxR is the sensor; each of its subunits contains one [2Fe-2S] cluster that is susceptible to oxidation by superoxide (Gaudu *et al.* 1997). This oxidation has been shown to be complete within 2-3 min following exposure to superoxide-generating compounds (Ding and Demple 1997). SoxR regulates transcription of the *soxS* gene and has been shown to bind the DNA upstream of this gene regardless of the oxidation state of its iron-sulfur clusters (Gaudu and Weiss 1996). This observation has led to a model in which only the oxidized form of SoxR is able to activate transcription of *soxS*. The SoxS protein then activates the transcription of genes involved in the superoxide stress response.

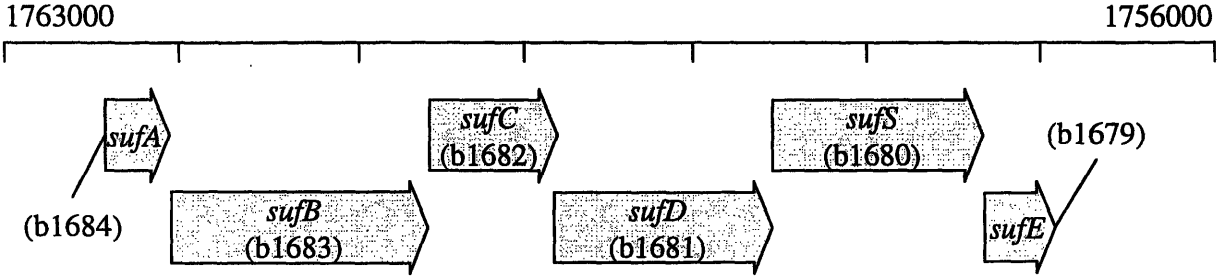
The superoxide response includes genes in the SoxRS regulon such as superoxide dismutase (SOD), which catalyzes the reaction

$$2\,O_2^{\bullet-} + 2\,H^+ \rightarrow H_2O_2 + O_2 \qquad (1.3)$$

Also involved in the superoxide response are two systems for repair of oxidized iron-sulfur clusters. In *E. coli*, these two repair systems have only recently been discovered. The Isc system is thought to consist of 10 genes in multiple, adjacent transcription units. The protein IscS was discovered to have activity as a cysteine desulfurase. *In vitro*, this protein is able to extract a sulfur atom from cysteine and transfer it to the iron-sulfur cluster of dihydroxy-acid dehydratase (IlvD) (Flint 1996). It is also known that HscA and HscB are chaperones that interact with IscU and are specific to iron-sulfur proteins (Hoff *et al.* 2000). Both IscU and ferredoxin (Fdx) have been proposed to act as scaffold proteins for formation of iron-sulfur clusters before transferring them to the target proteins (Takahashi and Nakamura 1999).



The Suf system involves six genes. In this system, the SufSE protein has been shown to act as a cysteine desulfurase. It extracts a sulfur atom from cysteine to generate an iron-sulfur cluster in the protein SufA, which likely serves as a scaffold for cluster formation (Loiseau *et al.* 2003). The homologous protein in *Erwinia chrysanthemi* has been shown to act as a scaffold (Ollagnier-de Choudens *et al.* 2003). The SufBCD complex has been shown to increase the cysteine desulfurase activity of SufSE, and may act as a modulator of desulfurase activity in order to limit generation of excess sulfur (Outten *et al.* 2003).



Additionally, a third cysteine desulfurase is known to exist in *E. coli*. CSD (cysteine sulfinate desulfinase), encoded by *csdA*, has been shown to form molybdopterin, a molybdenum-

containing prosthetic group, but its role in the cell is not entirely understood (Leimkühler and Rajagopala 2001).

DNA microarray studies have helped to elucidate genes regulated by SoxRS as well as those regulated by the closely related MarRA and Rob regulators. Upon addition of paraquat, a superoxide generating reagent, to *E. coli* cultures, 66 genes were found to increase in expression, while another 46 were found to decrease in expression (Pomposiello *et al.* 2001). This list contained, not only the *soxS* genes, but also eight genes known to be regulated by SoxS, including *acrA*, *fldA*, *fpr*, *fumC*, *fur*, *inaA*, *sodA*, and *zwf*. Other notable effects of paraquat addition found in this study include increased expression of the sets of genes listed below.

- Central metabolism and sugar transport genes, presumably to regenerate the reducing power of NADH, NADPH, and $FADH_2$
- Genes involved in the electron transport chain, possibly to consume excess oxygen
- Genes encoding ribosomal proteins as well as a gene involved in translational initiation (*fmt*)
- Genes involved in repair of macromolecules, which may be oxidized by superoxide

This study also measured gene expression changes that occur due to artificial induction of *soxS* from a plasmid. A smaller set of genes was affected (37 increased, 58 decreased) by this directed perturbation. Together, these two experiments identified seven genes that were not previously reported as part of the SoxRS regulon, including *cysK*, *lpxC*, *map*, *nfnB*, *ptsG*, *ybjC*, and *yggX*. Because mild perturbations were purposely used, none of the genes from the heat-shock response were identified as having differential expression and only two OxyR-regulated genes (*ahpC* and *dps*) were identified as being up-regulated.

DNA microarray studies of two other stress-inducible regulons, MarRA and Rob, confirmed an overlap with the SoxRS regulon. The SoxRS-regulated genes *inaA*, *fumC*, *ompF*, *sodA*, and *zwf* were identified as being differentially expressed in a strain that constitutively expresses MarA (Barbosa and Levy 2000). Addition of sodium salicylate to wild-type cultures is known to trigger the MarRA regulon and was also found to stimulate expression of several SoxRS-regulated genes (Pomposiello *et al.* 2001). Additionally, salicylate treatment resulted in increased expression of sugar transport genes, similar to that observed with paraquat treatment. Based on the results of these two studies, seven genes were identified as being members of the

combined marA/soxS/rob regulon, including *aldA*, *yncE*, *map*, *mdaB*, *nfnB*, *pgi*, and *yhbW* (Martin and Rosner 2002).

1.2.2.2 Hyperoxic Stress due to Peroxide

Hydrogen peroxide is generated by oxidation of iron-sulfur clusters (**1.2**), as well as by the cellular defense to superoxide (**1.3**). According to the Fenton reaction, further reduction of hydrogen peroxide by free iron(II) forms hydroxyl radicals.

$$Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + HO^{\bullet} + HO^{-} \qquad (1.4)$$

These radicals are well known to oxidize DNA (Levin *et al.* 1982; Imlay and Linn 1986). As if this potentially mutagenic activity were not damaging enough, hydroxyl radicals also have the potential to damage proteins and membrane lipids (Imlay 2002).

Hydrogen peroxide has also been shown to damage biological molecules without formation of hydroxyl radicals. Oxidation of iron-sulfur clusters, non-cysteinyl residues of proteins, and lipids are reviewed in (Imlay 2002). However, the peroxide effect that is best understood is that of disulfide bond formation. Typically, disulfide bonds do not form in the reducing cytoplasm of the *E. coli* cell. Hydrogen peroxide is known to oxidize cysteine residues to stimulate formation of disulfide bonds that otherwise would not exist. This conformational change can alter the metabolism of the cell by inactivating enzymes. Some of the enzymes that are known to be inactivated by formation of disulfide bonds include glyceraldehyde-3-phosphate dehydrogenase (Lind *et al.* 1998) and tyrosine phosphatase (Denu and Tanner 1998).

OxyR is also known to react with peroxides to form a disulfide bond between Cys199 and Cys208 (Zheng *et al.* 1998), which activates this regulator protein. Upon activation, OxyR induces transcription of genes encoding hydroperoxidase I (*katG*) and alkylhydroperoxidase (*ahpC* and *ahpF*). These hyperoxic defense enzymes remove peroxides via the following reactions.

$$2\,H_2O_2 \rightarrow 2\,H_2O + O_2 \qquad (1.5)$$

$$ROOH + NADH + H+ \rightarrow ROH + H_2O + NAD^{+} \qquad (1.6)$$

Thus, OxyR senses both hydrogen peroxide as well as the redox state of the cell in order to activate this branch of the hyperoxic stress response (Åslund *et al.* 1999).

An excellent DNA-microarray study was carried out to elucidate the genes involved in the OxyR regulon (Zheng, Wang, Templeton *et al.* 2001). Samples were taken from wild-type *E. coli* as well as a Δ*oxyR* mutant, both with and without hydrogen peroxide exposure. This study confirmed increased transcription of the genes *ahpC*, *ahpF*, *dps*, *fur*, *gor*, *grxA*, *katG*, and *trxC* upon exposure to hydrogen peroxide in the wild-type strain, but not in the Δ*oxyR* mutant. Seven additional genes (*hemH*, *sufA*, *sufB*, *sufC*, *yaiA*, *yaaA*, and *yljA*) were selected as being OxyR dependent based on the microarray data. DNase I footprinting and primer extension confirmed OxyR binding sites upstream of each of these genes. The *suf* operon is likely under the control of other regulators in addition to OxyR. This same expression analysis study found that genes in the SoxRS regulon and the Isc operon were up-regulated by hydrogen peroxide, independent of OxyR. This study provides what is so far the best example of how DNA microarrays have been applied to elucidate microbial stress responses.

1.2.3 Hypoxic Conditions

A wide array of metabolic changes is known to occur in environments with little or no oxygen. Such environments are sensed by two regulators, FNR and the ArcAB two-component system. FNR contains a [4Fe-4S] cluster that becomes oxidized to a [2Fe-2S] cluster in the presence of oxygen (Jordan *et al.* 1997; Khoroshilova *et al.* 1997). Unlike SoxR, which becomes active upon oxidation of its iron-sulfur cluster, FNR is active when its cluster is in the reduced form. FNR can act as both an inducer and a repressor of transcription. The ArcAB system consists of a membrane-bound sensor protein, ArcB, and a cytoplasmic transcriptional regulator, ArcA. Under anaerobic conditions, ArcB transfers phosphates through a cascade of residues within its cytoplasmic domain and ultimately to ArcA (Kwon *et al.* 2000). Phosphorylated ArcA subsequently activates transcription of genes involved in anaerobic metabolism and represses that of genes involved in aerobic metabolism. Oxidized quinones (ubiquinone and menadione), which are present during aerobic respiration, have recently been shown to inhibit the autophosphorylation of ArcB, thereby regulating this signal of anaerobic respiration (Georgellis *et al.* 2001). These two regulators work in concert to provide two levels of respiratory control. The ArcAB system exerts its control at low oxygen levels (10-20% of air saturation), while FNR is active in the absence or near-absence of oxygen (Tseng *et al.* 1996; Alexeeva *et al.* 2003).

These regulators are known to control the metabolic changes that occur in the transition from aerobic to fermentative metabolism. FNR, ArcAB, or both have been shown to decrease expression of genes involved in the TCA and glyoxylate cycles (*aceB/EF*, *acnA*, *fumA/C*, *gltA*, *icdA*, *mdh*, *sdhCDAB*, and *sucABCD*) and increase expression of those involved in glycolysis and fermentation (*adhE* and *pflB*). These regulators also repress genes encoding aerobic respiratory enzymes that use the electron donors NADH (*ndh* and *nuoABCEFGHIJKLMN*), lactate (*lldD*), and succinate (*sdhCDAB*). Genes encoding anaerobic respiratory enzymes that use formate (*fdnGHI*), glycerol-3-phosphate (*glpD*), and hydrogen (*hyaABC* and *hybC/O*) as electron donors are induced by ArcAB and FNR. Interestingly, the genes *glpABC*, which code for an anaerobic respiratory enzyme, are induced under anaerobic conditions (by FNR) and repressed under microaerobic conditions (by ArcAB).

The anaerobic regulators are also known to alter the expression of aerobic cytochromes that use oxygen as the ultimate electron acceptor. Cytochrome *o* oxidase (*cyoABCD*) transfers electrons from reduced ubiquinol to oxygen. As the major aerobic cytochrome, it is repressed by both FNR and ArcAB. Cytochrome *d* oxidase (*cydAB*) performs the same function, but has a higher affinity for oxygen. Therefore, this enzyme is most valuable in low oxygen environments and, while repressed by FNR, is induced by ArcAB. Under micro- and anaerobic conditions, enzymes that use alternative terminal electron acceptors are activated by FNR. These anaerobic respiratory enzymes use DMSO (*dmsABC*), fumarate (*frdABCD*), nitrate (*narGH/I*), and nitrite (*nirBD* and *nrfABCD*) as terminal electron acceptors.

The aconitase and fumarase enzymes are involved in the TCA cycle. The transcriptional regulation of these genes in response to oxygen levels in *E. coli* is particularly interesting. *E. coli* contains three fumarase genes that are regulated as described in Table 1.1. The FumA and FumC isozymes are the two dominant aerobic fumarases. Lacking an iron-sulfur cluster, FumC is oxygen resistant. Under highly aerobic conditions, the iron-sulfur cluster in FumA can become oxidized; therefore, *fumC* is induced as part of the SoxRS regulon in order to compensate. FumA is the most active isozyme under low oxygen (microaerobic) conditions. FumB is most active during anaerobic conditions and is correspondingly induced by FNR. There are also two aconitase isozymes, which are regulated as described in Table 1.2. Both aconitases contain iron-sulfur clusters, but the cluster in AcnA is resistant to oxygen. Therefore, AcnA is preferred under aerobic conditions. Consistent with this are the observations that *acnA* is

induced by SoxRS and repressed by FNR and ArcAB. The oxygen sensitive *acnB* gene is active under anaerobic conditions, but is also transcribed under highly aerobic conditions, possibly to act as a sensor of iron depletion (Varghese *et al.* 2003).

Table 1.1: Regulation of Fumarase Genes in *E. coli*
Positive regulation is indicated by (+) and negative regulation is indicated by (—)

Gene		*fumA*	*fumB*	*fumC*
Regulator	FNR	—	+	
	ArcAB	—		—
	SoxRS			+
Fe-S Cluster		Yes	Yes	No

Table 1.2: Regulation of Aconitase Genes in *E. coli*
Positive regulation is indicated by (+) and negative regulation is indicated by (—)

Gene		*acnA*	*acnB*
Regulator	FNR	—	
	ArcAB	—	—
	SoxRS	+	
Fe-S Cluster		Yes, but oxygen resistant	Yes

Although the transcriptional regulation of FNR and ArcAB are well understood, many questions remain, and because of the large number of genes affected, answering them is an excellent application for DNA microarrays. Two microarray studies identified additional genes in the FNR (Salmon *et al.* 2003) and ArcAB (Liu and Wulf 2004) regulons and found that each regulator, either directly or indirectly, affected 17% and 9% of the genome, respectively. Thus, these regulons may be even larger than currently believed.

The FNR study examined transcriptional changes in wild-type cultures grown aerobically and anaerobically as well as in Δ*fnr* mutant cultures grown anaerobically (Salmon *et al.* 2003). Based on these results, 94 genes showed FNR-dependent expression, only five of which had been previously reported as being regulated by FNR. As expected, the genes *ndh* and *nuoE* were identified in this category as repressed under anaerobic conditions. Also in this category were 23 anaerobically activated genes known to be regulated by Lrp (leucine-responsive regulatory protein) under aerobic conditions, suggesting some overlap with this regulon. The 5-fold increase in expression of *lrp* under anaerobic conditions indicates that its product does indeed play a role in anaerobic transcriptional regulation. Another 57 genes were found to have

anaerobic differential expression that was independent of FNR. These genes are attributed to the ArcAB regulator. The genes *appB* and *appC*, which encode for a third cytochrome oxidase, appear in this category as induced by anaerobiosis. It appears that this enzyme is synthesized to prepare for the possibility of increased oxygen availability.

The ArcAB study compared samples from wild-type and Δ*arcA* mutant cultures grown anaerobically (Liu and Wulf 2004). Combining microarray data with a matrix of potential ArcA binding sites led to a list of 58 operons affected by ArcA. While seven of these operons had known respiratory function (*e.g. cydAB*, *ndh*, and *sdhCDAB*), the remaining 51 had apparently unrelated functions such as osmoprotection (*caiT*), flagellar biosynthesis (*fliE/MN*), cell division (*ftsZ*), and nickel transport (*nikABCDE*).

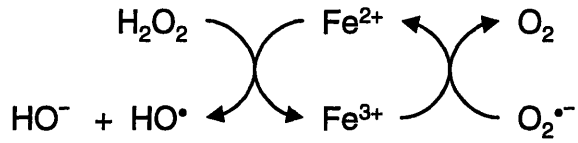
1.2.4 Iron Homeostasis

Iron plays several roles in the metabolism of oxygen and the defense against reactive oxygen species. The electron transport chain transfers electrons to oxygen, ultimately reducing it to water as the endpoint of respiration. This system consists of several cytochromes that contain both iron-sulfur clusters and heme groups. In most cases, these iron prosthetic groups are directly involved in either electron transport or oxygen activation. The hydroperoxidase proteins (encoded by *katE* and *katG*), which are involved in defense against hydrogen peroxide, both contain heme groups that are critical to their function. In the defense against superoxide, Fe-SOD (encoded by *sodB*), one of *E. coli*'s three superoxide dismutases, contains an iron cofactor as well. While other SOD's contain different metals (*e.g.* Mn, Cu), a metal cofactor is required at the active site. Integration of iron into the cell's metabolism is a requirement of life in an aerobic environment (Earhart 1996; Pierre and Fontecave 1999 for reviews on iron metabolism).

The damaging effects of iron have been mentioned in previous sections, but the various reactions must be considered together to gain a complete picture. As shown in **(1.4)**, iron(II) ions can generate damaging reactive oxygen species via the Fenton reaction. In addition, iron(III) ions can remove superoxide species.

$$Fe^{3+} + O_2^{\bullet-} \rightarrow Fe^{2+} + O_2 \qquad (1.7)$$

These two reactions form a cycle called the Haber-Weiss reaction.



(1.8)

Based on this iron-catalyzed cycle, the toxic effects of superoxide would result from turnover of iron ions that would generate destructive hydroxyl radicals. However, it is unlikely that reaction **(1.7)** will occur *in vivo*. The superoxide concentrations required for this reaction to proceed *in vitro* are much higher than the estimated intracellular concentrations (Keyer *et al.* 1995). Furthermore, other intracellular reducing agents are more effective at reducing iron(III), *e.g.* NADH (Imlay and Linn 1988) and glutathione.

Alternatively, the toxic effects of superoxide may result from its oxidation of iron-sulfur clusters, *e.g.* **(1.2)**. This reaction would create both iron(II) ions and hydrogen peroxide, the two reactants in the Fenton reaction. Therefore, an overall picture of the interplay between reactive oxygen species and iron looks like the reaction network in Figure 1.5.

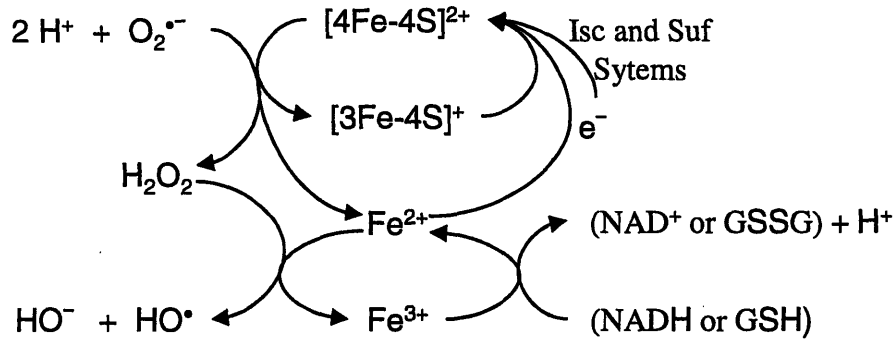


Figure 1.5: Iron-Catalyzed Generation of Reactive Oxygen Species
The reaction network above represents current knowledge of the *in vivo* reactions with free iron, iron-sulfur clusters, and reactive oxygen species. Although reactions with [4Fe-4S] clusters are shown here, reactions with [2Fe-2S] clusters are expected to be similar. Regeneration of these clusters by the Isc and Suf systems is not entirely understood.

Regulated mainly by Fur, iron metabolism in *E. coli* is well understood. In the presence of iron, Fur, along with an iron(II) cofactor, represses transcription of a set of genes involved in iron transport and storage. Genes involved in synthesis of enterobactin (*entCEBA/D/F*), a molecule that binds extracellular iron(III), all appear in the Fur regulon. In addition, genes involved in transporting enterobactin back into the cell are also regulated by Fur. These genes include products of the genes *cirA*, *exbB/D*, *fepA/B/DGC/E*, and *tonB*. Once inside the cell, the

protein Fes is responsible for release of iron from enterobactin. Other iron-transport complexes, such as ferridicitrate, are transported by the products of the genes *fecABCDE*. It is worth noting that the genes *feoAB*, which encode a transport protein for iron(II), are not regulated by Fur.

Although the Fur regulon is well defined, a recent DNA-microarray study found some surprising results (McHugh *et al.* 2003). Samples from Δ*fur* cultures and cultures grown under iron-limiting conditions were compared with wild-type cultures and were analyzed using DNA microarrays. As expected, genes repressed by Fur showed increased expression in the *fur* mutant as well as during iron-limiting conditions. For example, genes involved in enterobactin synthesis and transport showed strong activation. Several unknown genes were also identified to have the same expression pattern. Genes in the *suf* operon were also found to be up-regulated in these cultures. A surprising result from this study was the repression of many genes whose products have heme and iron-sulfur cofactors. Several cytochrome genes and genes involved in anaerobic respiration showed decreased expression under iron-limiting conditions. This is apparently an attempt to reduce the demand for iron in these stressed cultures. It is not clear whether expression of these genes is normally activated by Fur or if another regulator is involved.

1.3 Transcriptional Analysis during Fermentation Scale-Up

The ability to monitor physiological changes during fermentation or cell culture has the potential to greatly improve our understanding of bioprocess engineering. This section discusses the effects of fermentation scale-up, along with transcriptional analysis work that has been done to better understand these effects.

Although *E. coli* has both aerobic and anaerobic metabolic pathways, aerobic growth is preferred for recombinant protein production because of faster growth and energy production. Stoichiometrically, ATP generation is 19 times more efficient during aerobic growth than during anaerobic growth (Voet and Voet 1995). If available, oxygen is used as the ultimate electron acceptor in a series of reactions that withdraws electrons from NADH, NADPH, and $FADH_2$. However, if oxygen is not supplied in sufficient quantity, microbial growth can become slow. If supplied in excess, it can lead to problems in both natural and industrial systems, as will be shown. Aeration of bioprocesses is a delicate balance.

Transition from a lab-scale fermentor to a production-scale fermentor can result in slower growth, lower product yield, and reduced yield coefficients of biomass per substrate (Riesenberg

et al. 1990; Bylund *et al.* 1998; George *et al.* 1998). Scale-up difficulties have been attributed to heterogeneities in large fermentors such as oxygen gradients (Sweere, Janse *et al.* 1988). Such gradients have frequently been observed in large fermentors (Carilli *et al.* 1961; Steel and Maxon 1966; Manfredini *et al.* 1983; Oosterhuis and Kossen 1984). Studies that induced DO oscillations in small-scale fermentors, intending to simulate the environment of a large-scale reactor, showed decreased product formation (Vardar and Lilly 1982; Yegneswaran *et al.* 1991). Another series of studies with DO oscillations showed that biomass production decreases as the frequency of the oscillation decreases (Sweere, Janse *et al.* 1988; Sweere, Mesters *et al.* 1988). These studies indicate that biomass and product formation upon scale-up are sensitive to both the rate of aeration and the level of mixing in the fermentor.

Previously, microbial responses to environmental changes have been observed by measurement of nucleotide (Neubauer *et al.* 1995) and mixed-acid (Xu *et al.* 1999) production. Authors of recent papers investigating substrate gradients have pointed out the need for understanding the microbial response to rapid environmental fluctuations on a genetic scale (Larsson *et al.* 1996; Xu *et al.* 1999). Toward this goal, one group has transformed *E. coli* with green fluorescent protein (GFP) promoter probes, which were constructed by cloning the promoters of several hyperoxic stress genes upstream of GFP sequences. Induction of these genes can be measured by the fluorescence intensity in the broth (Albano *et al.* 1998). Another study followed the expression of several *E. coli* stress response genes in a two-compartment reactor, consisting of both a stirred-tank reactor (STR) and a plug-flow reactor (PFR). Included in the list of studied genes were two genes known to respond to hypoxia and two heat-shock genes (*clpB* and *dnaK*). With glucose feed addition at the PFR inlet and no air or oxygen addition, mRNA levels of these four genes increased by two- to three-fold over the 54 s residence time in the PFR, confirming that the culture experiences hypoxia. Similar analysis of the top, middle, and bottom of a 30,000-L production-scale fermentor showed that *frd* and *dnaK* were expressed in larger amounts at the top and middle of the fermentor (Schweder *et al.* 1999). Examination of stress responses indicates that *E. coli* are exposed to hypoxic and hyperoxic conditions during fermentation.

1.4 α1-Antitrypsin

Human α1-antitrypsin (also known as α_1-proteinase inhibitor) belongs to the family of serine protease inhibitors (serpins). While α1-antitrypsin (α_1AT) is an inhibitor of the protease trypsin, its physiological role is to regulate the activity of neutrophil elastase, a protease involved in degradation of connective tissue, particularly in the lungs. Without functional α_1AT, unchecked degradation by elastase can lead to lesions in lung tissue characteristic of emphysema. The oxidants present in cigarette smoke have been shown to oxidize the active-site methionine residues (Met351 or Met358) of α1-antitryspin, thereby inactivating it (Johnson and Travis 1979; Taggart *et al.* 2000). This temporary loss in α_1AT activity contributes to the pathology of emphysema in smokers (Evans and Pryor 1994).

Individuals lacking functional α_1AT are also at risk for hereditary emphysema as well as chronic obstructive pulmonary disease. As a therapeutic, α_1AT is used to treat these individuals, thereby preventing any further damage to their lung tissue. α_1AT deficiency is prevalent in people of Northern European descent, and there are an estimated 150-200 thousand individuals with α_1AT deficiency in Europe and North America (Bayer Healthcare LLC 2004). α_1AT has been produced by Bayer under the name Prolastin® since 1987. Recently, other versions of this therapeutic protein have been approved for production by Baxter under the name Aralast in 2002 and by Aventis under the name Zemaira™ in 2003. All three of these products are produced by purification from human plasma. A collaboration between Bayer and PPL Therapeutics to produce α_1AT by transgenic means has recently ended unsuccessfully.

High-quality manufacturing of a therapeutic recombinant protein that is susceptible to oxidation, like α_1AT, can be difficult. However, the reducing environment of the *E. coli* cell makes it an attractive expression system for producing this protein. In previous work, a peptide mapping procedure was developed to observe oxidation of Met351 and Met358 (Griffiths and Cooney 2002). When applied to recombinant α_1AT produced by *E. coli*, none of the product was found to be oxidized (Griffiths 2002). Under normal aerobic conditions, the cytoplasmic levels of reactive oxygen species, such as hydrogen peroxide and superoxide, are too low to oxidize cysteine and methionine residues. Therefore, α_1AT is not oxidized at the active-site methionine residues during the recombinant production process. For some proteins, *E. coli*'s reducing environment can be a disadvantage because it prevents disulfide bonds from forming in the

cytoplasm. α_1AT, however, has only one cysteine residue, does not form any disulfide bonds, and is therefore unaffected.

Despite the advantage of a reducing cytoplasm, *E. coli* is not an ideal expression system for α_1AT. *E. coli*'s heat shock response leads to degradation of α_1AT produced at 30°C in minimal medium cultures. Using pulse-chase labeling, previous work has found that 25% of the protein produced 60 min after induction is lost to degradation (Laska 2000). The heat-shock protease ClpP is responsible for most of this degradation. α_1AT produced in a ClpP$^-$ mutant showed a significant decrease in both the rate and extent of proteolysis. Unfortunately, the ClpP$^-$ mutant produced low overall yields of α_1AT and was a poor production strain.

An interesting feature of α_1AT degradation is that it increases with the partial pressure of oxygen. In cultures grown in pure oxygen, 35% of α_1AT is degraded, while in cultures grown in pure nitrogen, only 18% was degraded (Laska 2000). The oxygen dependence of degradation is surprising considering the lack of α_1AT oxidation. Previous work has shown that α_1AT mutants lacking the oxygen-sensitive C232, M351, and M358 residues are as susceptible to degradation as the original recombinant α_1AT (Laska 2000). While α_1AT degradation is oxygen dependent, it does not appear to result directly from oxygen. There are several examples in the literature of proteins that are degraded after being oxidized (Stadtman and Wittenberger 1985; Davies *et al.* 1987); however, this is not the case for α_1AT. Since the oxygen dependence of degradation does not appear to be a direct effect of oxygen, it must instead be a result of *E. coli*'s response to oxygen.

The link between oxygen and degradation of a recombinant protein product is particularly problematic because oxygen is difficult to control in large-scale fermentations. Gradients in oxygen have been well documented (Carilli *et al.* 1961; Steel and Maxon 1966; Manfredini *et al.* 1983; Oosterhuis and Kossen 1984) and have been associated with scale-up difficulties (Sweere, Janse *et al.* 1988).

1.5 Summary of Literature Review

DNA microarrays provide an excellent platform for studying the global responses of *E. coli* cultures to both recombinant protein production and oxygen environment. Production of recombinant α_1AT is known to stimulate the heat-shock response. Furthermore, induction in a high-oxygen environment stimulates the hyperoxic stress response. These stress responses may

act together to degrade recombinant α_1AT. DNA microarrays may prove to be an excellent tool for monitoring activation of these stress responses as well as diagnosing fermentation difficulties.

2 Goals and Objectives

"You've got to be careful if you don't know where you're going 'cause you might not get there!"

—Yogi Berra

The overarching goal of this work was to apply the technology of DNA microarrays toward improved understanding in bioprocess engineering. The specific problem of interest is production of a therapeutic recombinant protein with *Escherichia coli* fermentation. The protein α1-antitrypsin (α_1AT) exhibits oxygen-dependent degradation, which is known to result from the heat-shock stress response. Using α_1AT as a model system, the effects of microbial stress responses on *E. coli* fermentation were explored. Specifically the objectives of this thesis were as follows.

- Produce DNA microarrays for global gene expression analysis of *E. coli.*
- Develop and validate techniques for hybridization of samples and analysis of data.
- Identify *E. coli* genes with significantly altered expression under hypoxic and hyperoxic conditions.
- Identify *E. coli* genes with significantly altered expression upon production of a recombinant protein.
- Elucidate the role of oxygen and the hyperoxic stress responses in the mechanism of oxygen-dependent degradation of recombinant α_1AT in *E. coli, i.e.* determine which genes and gene products contribute to this degradation.
- Determine a strategy to improve the yield and quality of recombinant α_1AT.

3 Materials and Methods

"They give you a round bat, and they throw you a round ball. And they tell you to hit it square."

—Willie Stargell

3.1 Expression Strains and Plasmids

For most experiments, the *E. coli* strain BL21 (DE3) from Novagen was used. All B strains of *E. coli* are deficient in the *lon* protease; BL21 also lacks the *ompT* outer membrane protease. With these two proteases missing, this strain is likely to exhibit greater protein stability during production. This strain is also a lysogen of the bacteriophage DE3, and therefore its chromosome has inserted into it a copy of the *lacI* repressor gene as well as a copy of the T7 RNA polymerase gene under control of the *lacUV5* promoter. Under normal conditions, the LacI repressor binds to the *lacUV5* promoter region. When isopropyl-β-D-thiogalactopyranoside (IPTG) is added to this strain, it binds the repressor, thereby removing it from the promoter and allowing transcription of the T7 RNA polymerase gene to proceed.

This work also made use of a ClpA-deficient strain, SG1147, which was generously donated by Dr. Susan Gottesman (National Cancer Institute, Bethesda, MD). This mutant strain has the genotype BL21 (DE3) *clpA319*::*kan*.

The pEAT8 plasmid was generously provided by Dr. Myeong-Hee Yu of the Korea Research Institute of Bioscience and Biotechnology (Lee *et al.* 1993). This plasmid confers ampicillin resistance, contains a human α1-antitrypsin (α_1AT) insert, and was originally generated by inserting the human α_1AT gene into the pET3d plasmid (Novagen). The pEAT8-137 plasmid was later generated from pEAT8 to eliminate an internal start codon at amino acid position 137, thereby preventing production of an α_1AT fragment (Laska 2000). In this plasmid, expression of the α_1AT gene is controlled by a T7 promoter. When this vector is used in the DE3 strains described above, IPTG induces production of the T7 RNA polymerase, which in turn transcribes the α_1AT gene. The plasmids pEAT8-137 and pET3d were both used in this work, but pEAT8 never was.

3.2 Media

The defined M9 minimal medium, shown in Table 3.1, was preferred not only for reproducible results within this work, but also for reproducing previous observations (Laska 2000). The M9 medium was used throughout for the sake of consistency. As the starting point for glycolysis, glucose is preferred as a carbon source over other carbohydrates. The metals in the medium (both magnesium and the trace metals) are essential for activation of many enzymes required for growth. For experiments using strains transformed with either pEAT8-137 or pET3d, ampicillin was supplemented to the medium at a concentration of 100 μg/mL.

Table 3.1: M9 Minimal Medium Composition

Component	Molecular Weight	Stock Solution	Concentration in Stock		Stock Concentration Factor	Concenration in Medium	
Na_2HPO_4	142.0	M9 Salts	60. g/L	0.42 M	10 X	6.0 g/L	42 mM
KH_2PO_4	136.09	M9 Salts	30. g/L	0.22 M	10 X	3.0 g/L	22 mM
NH_4Cl	53.49	M9 Salts	10.0 g/L	0.187 M	10 X	1.00 g/L	18.7 mM
NaCl	58.44	M9 Salts	5.0 g/L	86 mM	10 X	0.50 g/L	8.6 mM
Glucose	180.16	Glucose	200. g/L	1.11 M	40 X	5.0 g/L	28 mM
$MgSO_4$	120.37	$MgSO_4$	120. g/L	1.00 M	1000 X	0.120 g/L	1.00 mM
$Na_2EDTA•2H_2O$	372.2	Trace Salts	20. g/L	55 mM	667 X	30. mg/L	82 μM
$CaCl_2•H_2O$	147.02	Trace Salts	510 mg/L	3.5 mM	667 X	0.76 mg/L	5.2 μM
$FeCl_3•6H_2O$	270.30	Trace Salts	16.9 g/L	63 mM	667 X	25 mg/L	94 μM
$CuSO_4•5H_2O$	249.68	Trace Salts	160. mg/L	0.64 mM	667 X	0.24 mg/L	0.96 μM
$MnSO_4•H_2O$	169.01	Trace Salts	130. mg/L	0.77 mM	667 X	0.194 mg/L	1.15 μM
$CoCl_2•6H_2O$	237.93	Trace Salts	182 mg/L	0.77 mM	667 X	0.27 mg/L	1.15 μM
$ZnSO_4•7H_2O$	287.56	Trace Salts	174 mg/L	0.61 mM	667 X	0.26 mg/L	0.91 μM

3.3 Cell Growth and Induction

The medium used for overnight cultures was exactly the same as that used for experiments, including additional supplements. Overnight cultures were grown at 37°C at 250 rpm, typically for 10 h. This relatively short duration helped to avoid the passage of the culture into stationary phase and was particularly important considering that the cultures were grown in minimal medium. To extend the growth phase of the overnight cultures, the temperature may be reduced to 30°C, but this was not done in this work.

Cell growth was monitored by measurement of the optical density at 600 nm (OD_{600}). OD_{600} was measured using disposable plastic cuvettes and a water blank. In order to account for differences between the blank and sample cuvettes, it was necessary to measure the absorbance at 600 nm (A_{600}) of each cuvette filled with water. No more than 1000 µL of culture was used to measure OD_{600}. When necessary, the following general rule for *E. coli* cultures (Winkler 1995) was used to convert OD_{600} to dry cell weight (DCW) of the culture:

$$1\ OD_{600} = 0.34\ g\ DCW/mL \qquad (3.1)$$

When inoculating experimental flasks, the OD_{600} of the overnight culture was measured and the proper volume of this culture was transferred to the experimental flask to produce an initial OD_{600} of 0.05. Typically, the overnight cultures had OD_{600} of 5-6, and approximately 1 mL of overnight culture was required to inoculate 100 mL of medium.

In order to achieve α_1AT production that was almost entirely soluble, the induction phase began at OD_{600} of 0.7 and was carried out at 30°C for all experiments. The induction OD_{600} of 0.7 was selected, because this was high enough to give reasonable α_1AT yields, but low enough to prevent aggregation of the recombinant product. The growth phase of the cultures was also carried out at 30°C to avoid any unnecessary perturbations during the transition from growth to induction. During the growth phase, the increase in OD_{600} from 0.05 to 0.7 required 5.5-6.5 h. At this point, the cultures were induced by addition of IPTG to 0.4 mM.

3.4 Preparation of Soluble and Insoluble Protein Extracts

At the end of the induction phase, cultures were immediately cooled on ice to arrest growth. The contents of shake flasks were transferred to conical centrifuge tubes and centrifuged for 20 min at 2,000 rpm (1,100 × *g*) at 4°C (IEC CRU-5000). The medium was decanted from the cell pellets, which were resuspended in 5 mL TE3 Buffer (100 mM Tris, pH 8.0, 5 mM EDTA) by vortexing. The samples were transferred to glass culture tubes and sonicated in a Branson Sonifier® 450 with microtip (Output Control set at Level 3, Duty Cycle set at 50%) for two 90-s periods to lyse the cells. To prevent excessive heating during sonication, samples were suspended in an ice/water bath during sonication and were provided a 60-s cooling period between sonication steps. Sonicated cells were centrifuged for 15 min at 10,500 rpm (9,000 × *g*) at 4°C (IEC Centra-4, Rotor 820). Supernatants were removed with

syringe and needle, sterile filtered with 0.2-μm Acrodisc® syringe filters (Pall 4192), and stored on ice. Volumes of these final soluble protein extracts were recorded.

The pellets were resuspended in Wash Buffer (50 mM Tris, pH 8.0, 1 mM EDTA, 0.5% Triton) by vortexing. Samples were centrifuged for 15 min at 5,000 rpm (2,000 × *g*) at 4°C (IEC Centra-4, Rotor 820). The supernatants, containing cell debris and membrane fragments, were removed using syringe and needle and discarded. The pellets, containing insoluble protein, were resuspended in a volume of TE3 Buffer equal to the volume of the corresponding soluble protein extract and were stored on ice for immediate analysis.

3.5 Analysis of Proteins

This section describes the techniques that were used to analyze the protein extracts, which were prepared as described in Section 3.4.

3.5.1 Total Protein Assay

The Bio-Rad Protein Assay (500-0006) was applied to each soluble extract prior to performing the α_1AT activity assay (Section 3.5.2). The manufacturer's low-concentration test-tube assay was used with 800 μL of sample and 200 μL of concentrated assay reagent. A calibration curve was generated with bovine serum albumin (BSA) (Pierce 23209) standard diluted to concentrations ranging from 4-13 μg/mL with Milli-Q water. In this range, the calibration curve was linear. The extracts were diluted by combining 4 μL of extract into 796 μL Milli-Q water.

3.5.2 α1-Antitrypsin Activity Assay: Elastase Inhibitory Capacity

Elastase Inhibitory Capacity was used to assay for the activity of α_1AT in soluble cell extracts. This protocol, which was developed and modified in previous work (Beatty *et al.* 1982; Konz 1998; Laska 2000), is essentially an elastase activity assay. The α_1AT activity in a particular sample is measured as the amount by which the sample reduces the activity of elastase compared with an α_1AT-free blank. This assay was performed on the same day that the soluble extracts were prepared because the extracts' activities can change over time.

Before performing the assay, a working elastase stock was prepared by combining 20-30 μL of porcine pancreatic elastase (Sigma E-1250) with 0.15-M NaCl to a total volume of

750 μL. The amount of elastase varied batch-to-batch and was adjusted to have activity of approximately 1 mAU/s in the α_1AT-free blank (1 mAU = one one-thousandth of an absorbance unit). The chosen substrate for elastase was *N*-succinyl-$(Ala)_3$-nitroanilide (Sigma S-4760). Substrate stock solutions were prepared at 40 mg/mL in dimethyl sulfoxide (DMSO) and were stored at -20°C. Before performing the assay, a working substrate stock solution was prepared by diluting the DMSO stock to 2 mg/mL in an aqueous 100 mM Tris buffer at pH 8.0. Keeping this aqueous DMSO solution well mixed was found to be critical to the reproducibility of the assay.

The amount of the soluble extract used was estimated based on the total protein content of the extracts, which had been determined prior to this assay. The general rule of thumb after a 90-min induction period at 30°C in minimal medium was:

$$\text{Volume of Extract for EIC Assay } (\mu L) = 560 \,/\, \text{Total Protein Content of Extract } (mg/mL) \quad \textbf{(3.2)}$$

The necessary volume of each soluble extract was diluted to 1000 μL with Tris8 (100 mM Tris at pH 8.0). A blank consisting of 1000 μL Tris8 was also prepared. To each sample, 50 μL of the working elastase stock was added, and the samples were incubated at room temperature for 30 min. Elastase activity was detected using the kinetics mode of a UV/Vis spectrophotometer (Agilent HP8452A). A 3-mL quartz cuvette was washed with a 30-mM hydrochloric acid solution before and after each sample. This step was found to greatly improve the reproducibility of the assay. The blank was set after combining 2 mL of Tris8 and 100 μL of substrate working stock in the cuvette and mixing well. All 1050 μL of incubated extract was added to the cuvette and mixed well. The rate of appearance of the cleavage product of the elastase substrate was detected at 410 nm over a 30-60 s range and was recorded in units of mAU/s. All incubations and assays, including those for the blank, were performed in triplicate to account for the relatively low reproducibility of the assay. In this work, relative standard deviations of the assay were about 5-10%. The elastase inhibitory capacity for a particular extract was calculated as the difference between the slopes of the α_1AT-containing extract and the α_1AT-free blanks.

Previous work used this assay in conjunction with the Trypsin Inhibitory Capacity (TIC) assay to determine the oxidation of α_1AT. However, the TIC assay is performed at 256 nm, a

wavelength at which there is a great deal of noise from host proteins in the extract. The relatively small amount of α_1AT produced from minimal medium cultures at 30°C cannot be distinguished from the noise; therefore, the TIC assay was not used in this work.

3.5.3 Polyacrylamide Gel Electrophoresis

For analysis of pulse-chase samples, gels were prepared in the lab. Further information on the preparation and use of these gels is given in Section 3.8.2.

For Western blots and all other applications, 10% Tris-HCl Ready Gel Precast Gels (Bio-Rad 161-1155) were used. Samples were added to 0.5 volumes of 3× Reducing SDS Sample Loading Buffer (187.5 mM Tris, pH 6.8, 6% SDS, 15% β-mercaptoethanol, 30% glycerol, 0.3% bromophenol blue). Standards of α_1AT purified using a previously developed FPLC method (Griffiths 2002; Griffiths and Cooney 2002) were also run on the gel. One lane of Low Molecular Weight (LMW) Markers (Amersham Biosciences 17-0446-01) was also run on the gel. These gels were run at room temperature at 100 V in SDS Running Buffer (25 mM Tris base, 190 mM glycine, 0.1% SDS) for 90 min in a Mini-Protean 3 Electrophoretic Cell (Bio-Rad 165-3301).

Staining was performed using two Coomassie stains in which the dye permeated the gel and bound to protein, and a destain step in which unbound Coomassie was removed from the gel. Coomassie Brilliant Blue R-250 (Bio-Rad 161-0400) was prepared as a 0.15% stock solution. Gels were immersed in Stain #1 (10% isopropanol, 10% acetic acid, 0.003% Coomassie) for approximately 30 min. Gels were placed in the Stain #2 (10% acetic acid, 0.003% Coomassie) for at least 90 min, but typically overnight. Finally, gels were placed in Destain (10% acetic acid) for approximately 60 min. Completing the destaining step in only 60 min was made possible by placing a rolled Kim-wipe in the Destain to absorb the Coomassie as it left the gel. Imaging and quantification were performed as described in Section 3.5.5.

3.5.4 Western Blotting

Western blotting was performed by using the ECL Plus™ detection reagents (Amersham Biosciences RPN-2132) and the manufacturer's standard protocol, with the changes described here.

Samples were run on electrophoresis gels as described in Section 3.5.3. Pure α_1AT was loaded on the gels, but the LMW markers were not used since they do not produce any signal on

the blot. Instead, one of the wells of the gel was filled with 10 μL Kaleidoscope Pre-Stained Markers (Bio-Rad 161-0324). Neither of these markers produce signal on the blot; but, they do give visual confirmation of transfer to the membrane.

No staining of the gel was carried out. Instead, the protein in the gel was transferred to a PVDF membrane (Bio-Rad 162-0184) using a Mini Trans-Blot Transfer Cell (Bio-Rad 170-3935). The transfer was carried out in Western Transfer Buffer (20 mM Tris base, 150 mM glycine, 20% methanol, and 0.05% SDS) for 60 min at 100 V. Tris-buffered saline (TBS) (50 mM Tris, pH 8.0, 138 mM NaCl, 2.7 mM KCl) was used for all washes and incubations. TBS-T was prepared by adding 33 drops Tween-20 to 1 L TBS buffer. Following the transfer, the membrane was placed in a freshly made blocking solution (1.5 g Carnation Nonfat Dry Milk in 30 mL TBS-T) for 1 h. Rabbit anti-human α_1AT IgG antibody (Sigma A-0409) was used as the primary antibody and goat anti-rabbit IgG horseradish peroxidase conjugate (Sigma A-0545) was used as the secondary antibody. ECL Detection solution was prepared by combining 5 mL reagent A and 125 μL reagent B. The blot was incubated in the detection solution for 5 min, was exposed to BioMax® film (Kodak 165-1454) for approximately 1 min, and was developed. Imaging and quantification were performed as described in Section 3.5.5.

3.5.5 Imaging and Quantification

Coomassie-stained polyacrylamide gels and film from chemiluminescent Western blots were scanned using a Molecular Dynamics Personal Laser Densitometer with ImageQuant® software (Molecular Dynamics). Protein bands were quantified using the Local Median background correction option.

3.6 General Nucleic Acid Protocols

The protocols described in this section are commonly used techniques for working with nucleic acids. These protocols will be referred to by the protocols in Section 3.7.

3.6.1 Precipitation of Nucleic Acids

Nucleic acids in solution were precipitated by first adding 0.1 volumes of sodium acetate (3 M, pH 5.2), followed by 3 volumes of 100% ethanol. For instance, when working with a 300 μL total RNA sample, 30 μL of sodium acetate was added, followed by 900 μL of ethanol.

The microtube was inverted to ensure mixing and was placed at -80°C for no more than 30 min. The microtube was centrifuged at 13,000 rpm (11,000 × *g*) at 4°C (IEC Centra-4, Rotor 817) for 15 min to pellet the precipitated nucleic acid. The supernatant was carefully removed from the microtube by flattened pipette tips (Sorenson BioScience 13760), and the pellet was washed—not resuspended—in cold 70% ethanol. Again, the microtube was centrifuged at 13,000 rpm (11,000 × *g*) at 4°C (IEC Centra-4, Rotor 817) for 10 min, and the supernatant was carefully removed. The pellet was dried by spinning for 2 min in a Speed Vac Concentrator (Savant).

3.6.2 Determination of Concentration and Purity (A_{260} and A_{280})

Absorbance measurements at 260 nm and 280 nm (A_{260} and A_{280}) were used to determine both the concentration and purity of nucleic acid solutions. Since nucleic acids absorb at 260 nm, the A_{260} measurement is proportional to the nucleic acid concentration in the sample. The extinction coefficient depends on the identity and state of the nucleic acid, as indicated by Table 3.2. Since proteins generally absorb near 280 nm, the ratio A_{260}/A_{280} is an indicator of the relative levels of nucleic acid and protein. Typically, a ratio lower than 1.8 indicates excessive contamination by proteins.

Table 3.2: Commonly Used Extinction Coefficients at 260 nm for Nucleic Acid Samples
The extinction coefficient is the conversion factor between A_{260} measurements and nucleic acid concentration (Ausubel *et al.* 1995).

Nucleic Acid	Examples in This Work	Extinction Coefficient ((mL/μg) cm^{-1})
Single-Stranded DNA	Cy3- and Cy5-labeled samples	0.027
Double-Stranded DNA	Genomic DNA, Plasmid DNA	0.020
RNA	Total RNA	0.025

All measurements used a UV/Vis spectrophotometer (Agilent HP8452A) and a black quartz cuvette with both Suprasil windows and a 1-cm path length (Fisher Scientific 14-385-928D). Measurements were made in a total volume of 150 μL, consisting of both sample and buffer. The buffer used for these measurements was TNE Buffer (10 mM Tris, pH 8.0, 0.2 M NaCl, 1 mM EDTA); this high-salt buffer was chosen because it will overwhelm any salt in the samples that might contribute to variability in the measurements. Buffer (typically 145 - 149 μL) was added to the cuvette, which was then used to set the blank for the absorbance

measurements. Next, approximately 1 μg of sample (typically 1 – 5 μL) was added to the cuvette such that the total volume reached 150 μL. The diluted sample was mixed in the cuvette, three times, with a 200-μL pipette tip to ensure homogeneity. Without this step, absorbance readings were inaccurate and poorly reproducible. Three quick scans (each measuring both A_{260} and A_{280}) were taken of each sample to show that there was no drift in the measurements. If the three scans showed any significant trend, the diluted sample was mixed again. The average A_{260} reading was used for concentration calculations. The average A_{260}/A_{280} ratio was also calculated to determine purity.

3.6.3 Determination of Label Incorporation (A_{550} and A_{650})

Details of the Cy3 and Cy5 fluorescent labeling procedure are discussed in Sections 3.7.7 and 3.7.8. The incorporation frequency of these labels was determined by taking additional absorbance measurements simultaneously with the A_{260} and A_{280} measurements. Characteristics of the two fluorescent labels are given in Table 3.3. The protocol for these measurements was the same as that described in Section 3.6.2, except that for each scan, absorbance readings for two additional wavelengths—550 nm and 650 nm[1]—were taken. Based on these measurements and the information in Table 3.3, the concentration of label was determined in both Cy3- and Cy5-labeled samples.

Table 3.3: Characteristics of Fluorescent Labels
The extinction coefficient is the conversion factor between absorbance measurements, at the wavelength of maximum absorbance, and label concentration (Amersham Pharmacia Biotech 2000)

Fluorescent Label	Wavelength of Maximum Absorbance (nm)	Extinction Coefficient (M^{-1} cm^{-1})	Wavelength of Maximum Emission (nm)
Cy3	550	150,000	570
Cy5	649	250,000	670

Incorporation frequency was calculated using the following formula

[1] This wavelength differs from that in Table 3.3 because the UV/Vis spectrophotometer only scans at even integer wavelengths

$$I = \frac{C_D}{MW_{nt} \cdot m_{amu}} \cdot \frac{1}{C_L \cdot N_A} \quad \textbf{(3.3a)}$$

where I is incorporation frequency in nucleotides per label molecule (nt/molecule), C_D is the concentration of single-stranded DNA in μg/mL, C_L is the concentration of label in μM, MW_{nt} is the average molecular weight of a deoxyribonucleotide residue (309 amu/nt), m_{amu} is the definition of an atomic mass unit (1.66×10^{-27} kg/amu), and N_A is Avogadro's Number (6.022×10^{23} molecules/mole). Inserting the numerical values into (**3.3**a) simplifies the formula to

$$I = 3.237 \frac{C_D}{C_L} \quad \textbf{(3.3b)}$$

3.6.4 Native Agarose Gels

1% agarose solutions were prepared in TAE Buffer (40 mM Tris acetate, pH 8.5, 2 mM EDTA). Gels were prepared by combining 25 mL of this agarose solution with ethidium bromide stock to a concentration of 0.5 μg/mL. This produced a gel that was 10 cm in length. The samples, including 10× Gel Loading Solution (Ambion 8556), occupied no more than 10 μL volume and were loaded into the ten wells of the gel. Gels were run for 45 min at 75 V in cold TAE Buffer and were immediately viewed and photographed on a UV lightbox.

3.6.5 Genomic DNA Isolation

Cultures of *E. coli* pEAT8-137 were grown in 100 mL LB broth at 37°C for the sole purpose of genomic DNA isolation. Cultures were grown to OD_{600} of 1.0 and a 12-mL sample was taken and centrifuged at 2,000 rpm (1,100 × *g*) at 4°C (IEC CRU-5000) for 20 min to pellet cells.

Isolation of *E. coli* genomic DNA was carried out using Midi-Prep 100/G Genomic-tips (Qiagen, 10243) and the manufacturer's *Bacteria* protocol. Briefly, the cell pellet was exposed to lysozyme (Sigma L-6876), RNase A (USB Scientific 78020Y), and Proteinase K (Qiagen 19131) to simultaneously lyse cells and digest RNA and protein. Lysed cells were loaded onto a Genomic-tip, the tip was washed, and the genomic DNA was eluted into a round-bottom centrifuge tube. DNA in the eluate was precipitated by adding 3.5 mL isopropanol and

centrifuging at 10,500 rpm (9,000 × *g*) at 4°C for 15 min, and the supernatant was discarded. The pellet was washed—not resuspended—with 4 mL cold 70% ethanol. The tube was centrifuged again at 10,500 rpm (9,000 × *g*) at 4°C for 10 min, and the supernatant was discarded. 780 µL TE Buffer (10 mM Tris, pH 8.0, 1 mM EDTA) was added to the DNA pellet and the centrifuge tube was heated to 55°C for 1 h to dissolve the pellet. At this point, yield and purity of DNA were determined as described in Section 3.6.2.

3.6.6 Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR) was used to generate DNA for additional spots on DNA microarrays. Kits from two different vendors: Stratagene (600250) and Qiagen (201203) were used in this work. Reaction volumes of 100 µL were prepared using both kits, but the compositions differed slightly. Reactions were prepared in Microseal™ 96 polypropylene plates (MJ Design, Inc. MSP-9621). Using the Stratagene kit, reactions were prepared to have the following composition:

- 1× Reaction Buffer (10 µL of 10× solution, Stratagene 600250)
- 200 µM of each dNTP (1 µL of 20 mM dNTP mix, sold as 100 mM solutions of each dNTP, Invitrogen 10297-018)
- 10 ng/µL genomic DNA template OR 1 ng/µL plasmid template
- 500 nM forward primer
- 500 nM reverse primer
- 0.025 U/µL of *PfuTurbo*® DNA Polymerase (1 µL of 2.5 U/µL solution, Stratagene 600250)

Using the Qiagen kit, reactions were prepared to have the following composition:

- 1× PCR Buffer (10 µL of 10× solution, Qiagen 201203)
- 3.0 mM total $MgCl_2$ (supplemented 1.5 mM $MgCl_2$ in 1× PCR Buffer by adding 6 µL of 25 mM solution, Qiagen 201203)[2]
- 200 µM of each dNTP (1 µL of 20 mM dNTP mix, sold as 100 mM solutions of each dNTP, Invitrogen 10297-018)

[2] This component was optional and was only used when specified. When not used, the reaction was at 1.5 mM total $MgCl_2$, and 6 µL additional water was added to the reaction instead.

- 10 ng/μL genomic DNA template OR 1 ng/μL plasmid template
- 500 nM forward primer
- 500 nM reverse primer
- 1× Q-Solution (20 μL 5× solution, Qiagen 201203)[3]
- 0.025 U/μL *Taq* DNA Polymerase (0.5 μL of 5 U/μL solution, Qiagen 201203)

Primers for each PCR performed in this work are given in Table 3.4.

After preparing the reactions, wells were covered with Microseal™ A Film (MJ Design, Inc. MSA-5001) and placed in a DNA Engine thermal cycler (MJ Research). Reactions were run with the following general program:

Heat to 95°C for 1 min

Heat to 95°C for 1 min Cool to *T* °C for 1 min Heat to 72°C for *t* min	Repeat *x* times

Heat to 72°C for 10 min

Hold at 4°C

where *T* is the annealing temperature estimated from the G/C content of the primers, *t* is the elongation time determined by the length of the target, and *x* is the number of rounds. Reaction conditions for each PCR performed in this work are given in Table 3.5.

Amplified DNA was cleaned by removing the enzyme and unincorporated dNTP's with a QIAquick™ PCR Purification Kit (Qiagen 28014). As indicated by the standard protocol for this kit, 5 volumes of Buffer PB (500 μL) were added to each microtube and mixed. The manufacturer's protocol for this kit was followed and DNA was eluted from the column using 50 μL Buffer EB.

3.6.7 *In Vitro* Transcription

In vitro transcription was used to generate RNA probes for doping into total RNA samples in known amounts. These riboprobes were intended to produce signal on corresponding spots, which would have been used for normalization, or comparison of one array to another. *In*

[3] This component was optional and was only used when specified. When not used, 20 μL additional water was added to the reaction instead.

vitro transcription was performed on the BS1, BS2, BS3, and BS4 PCR products using the Riboprobe® System-T7 (Promega, P1440) and the manufacturer's "Synthesis of Large Amounts of RNA" protocol.

Table 3.4: PCR Primers Used to Generate Normalization Controls and Additional Spots for DNA Microarrays

The sequences are divided in groups of three. These divisions are only a visual aid and do not necessarily correspond to the reading frame of the gene. The first four primers pairs (BS1-4) were generated as normalization controls as described in Section 4.1. Each of these primers contains a binding sequence for T7 RNA polymerase, which is shown in bold. The last seven primer pairs were for additional spots that were added to the DNA microarrays.

Gene	Direction	Primer DNA Sequence (5' → 3')
BS1 (*ynzD*)	FOR	GCG CGA **TAA TAC GAC TCA CTA TAG GGG** CTG CTG CAA ATT GTC ATG GCG
	REV	ATT CAT CAC CCG CTA CTG CTC GAA GGC
BS2 (*ypfB*)	FOR	GGT GCT CCA ATG AAA ACA TTT GAA CGG C
	REV	GCG ACA **CTA ATA CGA CTC ACT ATA GGG** CCT CCA TTT TAT CCA CCC C
BS3 (*yflC*)	FOR	GCG ATA **TAA TAC GAC TCA CTA TAG GGT** TAA GCG CGC CAT CAA AGA CGG GG
	REV	CCC GCG TTT TTC AAA ATA CGA TAG CCC
BS4 (*yqzI*)	FOR	GAA TGC TTA GAA GCT CTA GGG
	REV	GCG CGA **TAA TAC GAC TCA CTA TAG GGT** TCT TCC CAC TTG ATT CGT CCC C
α1-Antitrypsin	FOR	GCC CAG AAG ACA GAT ACA TCC C
	REV	GGG ATT CAC CAC TTT TCC CAT GAA GAG GGG
T7 RNA Polymerase	FOR	GCG TTT AGC TCG CGA ACA GTT GGC
	REV	CCG ACT CTA AGA TGT CAC GG
fumA	FOR	CGC CCA GAG CAT AAC CAA ACC AGG C
	REV	CGC CCC CAT CGA ATC TTT TTC GCT GCG
fumC	FOR	GCA GGT CAT GAA TAC AGT ACG CAG CG
	REV	CAT ACT GCC GAC CAT CTG TTC TGG CCG
sdhB	FOR	GCT CCG CGT ATG CAG GAT TAC ACC C
	REV	AGC CCC TTC GGA CAT ACA CTG ACG C
cydB	FOR	CCT ATC TGC AAT GCG TAC CGT GGG CG
	REV	TAC CGG CTG TCA GGA TGA TGC AGG C
tolC	FOR	TGG GTT CAG TTC GTT GAG CCA GGC C
	REV	ACT TTC CAG TTG CTC GCT GGC ACC G

Table 3.5: PCR Conditions Used to Generate Normalization Controls and Additional Spots for DNA Microarrays

Gene	Template	Annealing Temperature: T (°C)	Elongation Time: t (s)	Rounds: x	Reaction Conditions	Length of PCR Product (bp)
BS1 (*ynzD*)	*B. subtilis* Genomic DNA	50	60	35	Stratagene kit	151
BS2 (*ypfB*)	*B. subtilis* Genomic DNA	50	60	35	Stratagene kit	181
BS3 (*yflC*)	*B. subtilis* Genomic DNA	50	60	35	Stratagene kit	152
BS4 (*yqzI*)	*B. subtilis* Genomic DNA	42	60	35	Qiagen kit with 3 mM Total $MgCl_2$	148
α1-Antitrypsin	pEAT8-137 Plasmid	42	165	25	Stratagene kit	1152
T7 RNA Polymerase	*E. coli* BL21 (DE3) Genomic DNA	45	240	25	Qiagen kit with 1x Q-Solution	2547
fumA	*E. coli* BL21 (DE3) Genomic DNA	53	180	30	Qiagen kit with 1x Q-Solution	1869
fumC	*E. coli* BL21 (DE3) Genomic DNA	53	180	30	Qiagen kit with 3 mM Total $MgCl_2$	1411
sdhB	*E. coli* BL21 (DE3) Genomic DNA	53	180	30	Qiagen kit with 3 mM Total $MgCl_2$	611
cydB	*E. coli* BL21 (DE3) Genomic DNA	53	120	30	Qiagen kit with 3 mM Total $MgCl_2$	366
tolC	*E. coli* BL21 (DE3) Genomic DNA	53	180	30	Qiagen kit with 3 mM Total $MgCl_2$	979

3.7 DNA Microarrays

The protocols described here are generally based on those described by the Institute for Genomic Research (TIGR) (Hegde *et al.* 2000), with slight modifications due to the differences between prokaryotic and eukaryotic organisms.

These protocols require using small volumes of liquid. In order to obtain proper mixing and reproducibility in concentrations, each microtube was frequently centrifuged briefly to collect all liquid at the bottom of the tube—especially after heating steps.

3.7.1 Plate Preparation

PCR products used for printing microarrays were the generous gift of Dr. Susan Lovett (Brandeis University, Waltham, MA). The steps taken by her lab to generate these PCR products are described here. The primer set used for these reactions (Sigma-Genosys *E. coli* ORFmers) is

based on the University of Wisconsin annotation of the *E. coli* genome (Blattner *et al.* 1997) and consists of 4,290 primer pairs that amplify the open-reading frame (ORF) of each gene, *i.e.* everything between the start codon and stop codon. The PCR protocol was similar to that described in Section 3.6.6. The success of each PCR reaction was checked by running samples on native agarose gels, using a protocol similar to that described in Section 3.6.4. These gels confirmed that there was only one amplified DNA fragment and that the size of this fragment corresponded to the predicted size of the ORF. Unsuccessful reactions were repeated and checked again—some were found to be successful in this second round. Overall, 4,082 (95%) of the PCR reactions were successful.

These PCR products were supplied lyophilized in twelve plastic V-well 384-well plates. For improved reliability during printing, the PCR products were transferred to sturdier 384-well plates with conical profile wells for microarray application (Genetix X7022). To accomplish this transfer, 20 μL Milli-Q water was added to each well in the original plates, and the plates were agitated on an orbital shaker at room temperature overnight to dissolve the DNA. The right half of Plate 12 (Columns 13-24) contained controls from the Lovett Lab and were left untouched throughout this preparation. The contents from the twelve old plates were then transferred by pipetting to twelve new plates. The old and new plates had identical formats, *i.e.* the location of every PCR product remained the same. The new plates were then placed in a Speed-Vac centrifuge on high heat for 1 h, which removed almost all of the liquid from the new plates. To ensure that all DNA from the old plates had been transferred, a wash was performed by adding 15 μL 50% (v/v) dimethyl sulfoxide (DMSO) to each well of the old plates. The old plates were agitated on an orbital shaker at room temperature for 1 h to dissolve any residual DNA. Again, the contents from the twelve old plates were transferred by pipetting to the twelve new plates, using the same format. In the end, wells in the right half of the new Plate 12 were empty. The new plates were agitated on an orbital shaker at room temperature overnight to dissolve the dried DNA in the wells.

The PCR products were supplied in a quantity of approximately 1 μg (Susan Lovett, personal communication) and were dissolved in 15 μL 50% DMSO as recommended by TIGR (Hegde *et al.* 2000), producing concentrations of about 67 ng/μL.

3.7.2 Spotted Controls

Controls were obtained from the Lovett Lab and were prepared by dissolution in 15 µL 50% DMSO. These controls are listed in Table 3.6. These controls were used to fill the last 192 wells on the right half of Plate 12.

Table 3.6: Spotted Controls Used for DNA Microarrays
All controls, except 50% DMSO, were supplied by the Lovett Lab. *E. coli* K-12 Genomic DNA was supplied by the Lovett Lab, and digestion was performed as described in Section 3.7.4.

Label	Description	Dilutions (ng/µL)
50% DMSO	no DNA	N/A
Genom	*Hae* III-digested *E. coli* K-12 Genomic DNA	2.67, 1.33, 0.67 & 0.33
tRNA	*E. coli* tRNA (Sigma-Aldrich R 1753)	100, 50, 25, & 12.5
rRNA	*E. coli* rRNA (Sigma-Aldrich R 7628)	100, 50, 25, & 12.5
Yeast	Yeast Library	100, 50, 25 & 12.5
Calf	Calf Thymus DNA (Sigma-Aldrich D 8661)	100, 50, 25, & 12.5
Human	Human Library	100, 50, 25 & 12.5
Vir Or	Viral PCR Product 1 (supplied in orange tube)	~20
Vir Pur	Viral PCR Product 1 (supplied in purple tube)	~20
RAP17	RAP17 C-GlyGly PCR Product	100, 50, 25, & 12.5
D280	pWKS 130 recJ D281 Plasmid	~50
C551	pWSK2a C552 Plasmid	~20
pBSSK	pBSSK Plasmid	180 & 90
XSeA	pBSSK XSeA Plasmid	150 & 75

3.7.3 Slide Printing

DNA microarray slides consisted of PCR products corresponding to each *E. coli* gene spotted on Corning GAPS slides. In all full-genome prints, one of the twelve 384-well plates was printed twice. With the exception of these 384 spots, every gene appeared only once on each slide. Roughly 91% of the spots corresponding to *E. coli* genes had no replicate. The slides were essentially printed with thirteen 384-well plates, resulting in 4,992 spots per slide.

Two different arrayers were used to print full-genome arrays throughout the course of this work; both were quill-pin robotic arrayers. Using a Virtek ChipWriter™, spots were printed

with 16 pins in a 17 × 17 grid with 250-μm spacing. Using a Bio*Robotics* MicroGrid II arrayer (MIT BioMicro Center), spots were printed with 32 pins in a 13 × 12 grid with 375-μm spacing. Arrays from four full-genome prints (A, B, C, and E) were used in this work. Prints A, B, and C were performed on the Virtek ChipWriter™, while Print E was performed on the Bio*Robotics* MicroGrid II.

At the end of these prints, each slide was etched in the lower right corner with the print letter and a unique number, corresponding to the order in which the slides were printed. The slides printed first generally had larger spots and produced better signal. Slides were placed in a desiccator for at least 24 h to dry. Once dry, the slides were cross-linked in a UV StrataLinker® 2400 (Stratagene) with a dose of 150 mJ and were stored in the desiccator.

Gene Array List (GAL) files, which store the gene name for each spot position, were generated based on the twelve plate files. The software used for image analysis, GenePix® Pro 3.0 (Axon Laboratories), contains a GAL file generator, which was used for the Virtek slides. However, for Bio*Robotics* slides, the software used to operate the arrayer was used to generate the GAL files.

3.7.4 Genomic DNA Preparation

E. coli genomic DNA was used as a hybridization standard in microarray experiments. Development of this technique is explained in Section 4.2. Genomic DNA was isolated from cultures grown solely for that purpose, as described in Section 3.6.5; these cultures are not the same as those used for microarray experiments. Given below is the protocol for digestion of that DNA to produce smaller, more easily labeled fragments.

For the *Hae*III restriction digest, the DNA solution was split into four equal aliquots in order to reduce the volume and improve the homogeneity of the digestion mixture. To each 195-μL DNA sample, 22.5 μL of 10× REACT® 2 Buffer and 7.5 μL of *Hae*III restriction enzyme (Invitrogen 15205-016) were added and mixed well by inverting. The reaction microtubes were placed at 37°C for 3 h.

The digested DNA was isolated using the QIAquick™ PCR Purification Kit (Qiagen 28014). This kit removes DNA fragments smaller than 100 bp from the mixture. As indicated by the standard protocol for this kit, five volumes of Buffer PB (1125 μL) were added to each microtube and mixed. Each sample was loaded onto a separate QIAquick™ column in two steps

by first adding 700 μL of the sample, centrifuging, discarding flow-through, and repeating with the remaining 650 μL. The columns were washed with Buffer PE as described in the standard protocol. Finally, the columns were eluted with 30 μL elution buffer, and the four eluates were combined.

The yield and purity of the isolated product were determined as described in Section 3.6.2. The products were also analyzed by running on an agarose gel. Initial experiments indicated that after 3 h, no fragments larger than 3.0 kb were visible on the gel, as shown in Figure 3.1. For subsequent digests, gels were used to verify that there were no fragments larger than 3.0 kb and thereby ensure reproducibility of the digest. When prepared in this way, a genomic DNA sample from a single culture had concentrations near 200 μg/mL and was useful for nearly twenty-five microarray experiments. Typically, genomic DNA was isolated and prepared in parallel from two 12-mL culture samples.

3.7.5 Growth, Induction, and Sample Collection

Experiments were performed by growing cultures and inducing as described in Section 3.3. For most experiments, samples for microarray analysis were collected immediately before induction (t = 0 min), and 10, 30, 60, and 90 min after induction. For the experiments in which aeration conditions were changed, the change was made at the time of induction. The three gas mixtures that were used in this work were pure N_2, pure O_2, and air. These gases were introduced to the headspace of the cultures via glass tubing which was placed through a punctured metal cap. Upstream of the culture the gas line contained a sterile filter (Pall PN4210); and, upstream of that, a water bubbler that served to hydrate the gases. Gas flow (regulated at approximately 1 slpm) through the water bubbler was typically begun 30 min before the gas was introduced to the culture, to allow the gas to hydrate.

Most of the microarray experiments required splitting cultures. For these experiments, 400 mL of culture was grown in a 2-L shake flask to OD_{600} of 0.7. At this point, an appropriate volume of culture (100–120 mL) was transferred from the growth flask to a fresh 500-mL induction flask. Then, 1-M IPTG (40–48 μL) was added directly to the new culture to achieve a final concentration of 0.4 mM, and the new headspace gas was introduced. These steps were repeated for the second and third flasks such that induction for each flask was separated by 1 min. Sample collection for each flask was also separated by 1 min

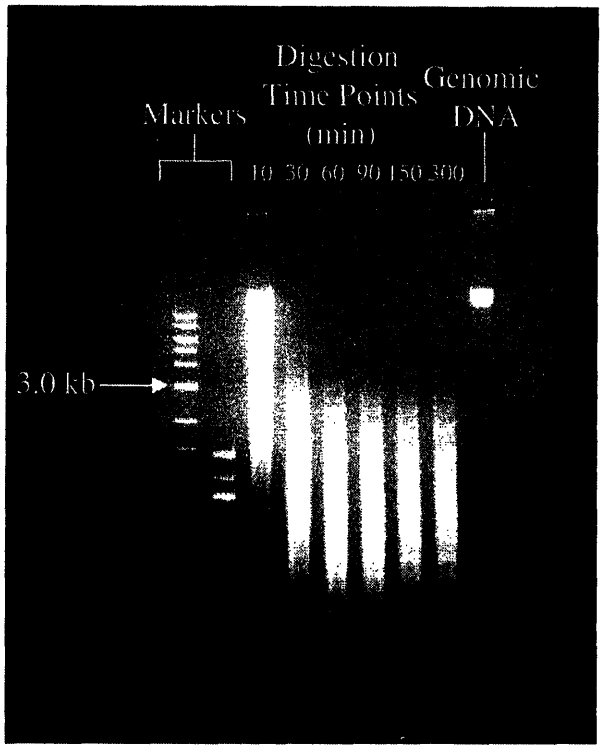


Figure 3.1: Native Agarose Gel Showing Progress of *Hae*III Digestion of Genomic DNA
A native agarose gel was run as described in Section 3.6.4. DNA Markers were loaded in the first two lanes of the gel. The middle lanes contained time points (labels are in minutes) showing the progress of the restriction digestion of *E. coli* genomic DNA by *Hae*III. The final lane contained the original genomic DNA sample before digestion. This initial experiment showed that the digestion was nearly complete after 60 min and, when complete, there were no visible DNA fragments larger than 3.0 kb.

Samples for both OD_{600} measurement and microarray analysis were taken simultaneously. Samples for microarray analysis were taken by transferring 5 mL of culture to each of two 14-mL sterile polypropylene culture tubes (Becton Dickinson 352059). Pre-induction (t = 0) samples taken for microarray analysis were sometimes increased to 7.5 mL because the cell density is low at this point and a higher volume helps to ensure an adequate amount of extracted RNA. These culture tubes were immediately frozen by immersion in liquid nitrogen and were stored at -80°C.

3.7.6 Total RNA Isolation

RNA is much more susceptible to degradation than DNA (Figure 3.2), and the enzymes that degrade RNA (RNases) are prevalent. In order to ensure isolation of intact RNA, all

procedures were performed using reagents (water, buffers, *etc.*) and supplies (microtubes, pipette tips, *etc.*) that are RNase free.

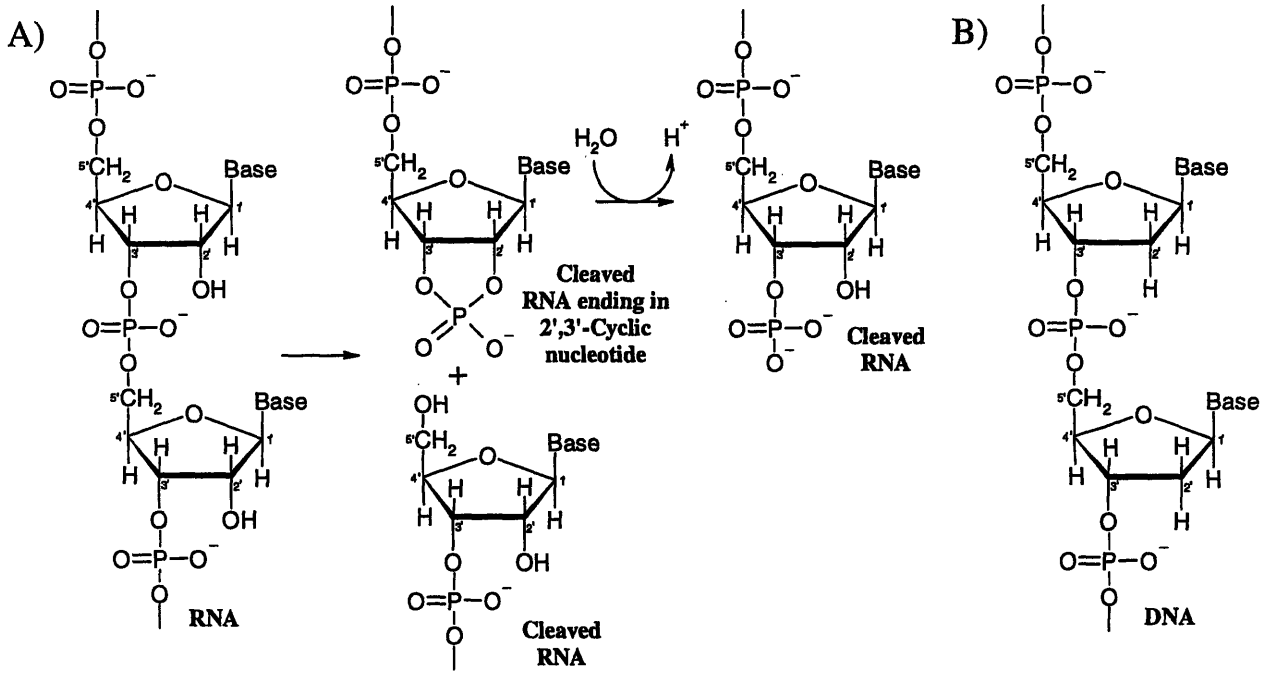


Figure 3.2: The 2'-Hydroxyl of RNA Makes It Susceptible to Degradation
A) The mechanism of RNA cleavage by RNase A. The nucleophilic 2'-hydroxyl attacks the adjacent phosphorous atom, cleaving the sugar-phosphate backbone and producing the 2',3'-cyclic intermediate. This intermediate is then hydrolyzed to regenerate the 3'-phosphate (Voet and Voet 1995). **B)** In contrast, DNA does not have a 2'-hydroxyl group, which explains, in part, why it is less susceptible to degradation.

Total RNA was extracted from frozen cell samples for analysis with DNA microarrays. No attempts were made to purify or amplify the mRNA from the total RNA. RNA isolation was performed using RNeasy® Midi-Prep Kits (Qiagen 75144). Qiagen gives the following summary of their kit.

> Biological samples are first lysed and homogenized in the presence of a highly denaturing guanidine isothiocyanate (GITC) containing buffer, which immediately inactivates RNases to ensure isolation of intact RNA. Ethanol is added to provide appropriate binding conditions, and the sample is then applied to the RNeasy® column where the total RNA binds and contaminants are efficiently washed away. High-quality RNA is then eluted in RNase-free water, ready for use in any downstream application.

The removal of RNases by the buffer component GITC allows all steps using the RNeasy® column, up to elution, to be performed at room temperature.

5-mL frozen cell samples in culture tubes were thawed in ice-water (2-3 h). Thawed cell samples were transferred to a 15-mL Falcon tube. The thawed cell sample was centrifuged at 2,000 rpm (1,100 × *g*) at 4°C for 20 min (IEC CRU-5000).

The manufacturer's *Bacteria* protocol (with $5 \times 10^8 - 5 \times 10^9$ cells) was followed with the observations and minor changes listed here. All centrifugation steps were performed in a swinging-bucket centrifuge (IEC CRU-5000) at 2,000 rpm (1,100 × *g*) at 15-20°C. Because GITC inactivates RNases, it is not necessary to keep the samples cold between the lysis and elution. Lysis buffer was prepared fresh by combining, per sample, 50 µL lysozyme (Sigma L-6876) stock (50 mg/mL) with 450 µL TE Buffer (10 mM Tris, pH 8.0, 1 mM EDTA) to make a 5 mg/mL solution. This lysis buffer is more concentrated than recommended, but was found to give higher RNA yields. Supernatant was removed from the Falcon tube and great care was taken to remove as much as possible, since remaining medium may degrade RNA in the sample. The pellet was resuspended in 500 µL of lysis buffer and incubated for 5 min at room temperature. Lysate was loaded onto the RNeasy® columns and washed as described by the manufacturer's protocol. RNA was eluted from the Midi-Prep column into a fresh tube by twice adding 150 µL of RNase-free water and centrifuging 3 min.

The 300 µL total RNA sample was treated with DNase to remove any remaining genomic DNA from the sample. 31 µL of a 10× DNase Buffer (400 mM Tris, pH 7.4, 60 mM $MgCl_2$, 20 mM $CaCl_2$, made with RNase-free water from the RNeasy® Kit) was added to the sample, followed by 1 µL of DNase I (Amersham Pharmacia Biotech 27-0514-01). The mixture was incubated for 30 min at 37°C. The total RNA was precipitated according to the protocol in Section 3.6.1. The RNA pellet was resuspended in 30 µL RNase-free water (from the RNeasy® Kit).

The concentration of each total RNA sample was determined by absorbance measurements as described in Section 3.6.2. Typically, the RNA concentration was 3-5 mg/mL. Quality of each sample was analyzed by running the samples on a native agarose gel as described in Section 3.6.4. A known high-quality *E. coli* Total RNA control sample (Ambion 7940) was run on each gel (1 µg) to help distinguish degradation during extraction from degradation in the gel. Figure 3.3 shows how total RNA samples appear on a gel.

3.7.7 Fluorescent Labeling of Total RNA Samples

Two-channel hybridization, a method that uses two different fluorescent labels, was used for all microarray experiments described in this work. Unless otherwise indicated, each hybridization was performed with a total RNA sample (labeled with Cy3) in one channel and a genomic DNA standard (labeled with Cy5) in the other channel. This section describes the labeling protocol for total RNA samples, while Section 3.7.8 describes the labeling protocol for genomic DNA.

Two different labeling methods were used for the arrays presented here. The initial method (used for slides from Prints A) employed SuperScript™ II reverse transcriptase, but was eventually replaced by a method using CyScript™ reverse transcriptase (used for slides from Prints B, C, and E). The CyScript™ method was chosen because the enzyme produces higher label incorporation.

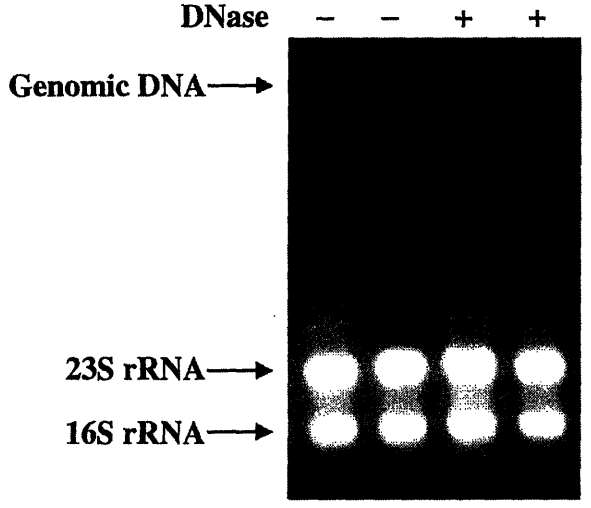


Figure 3.3: Native Agarose Gel with Total RNA Samples
A native agarose gel was run as described in Section 3.6.4. The first two lanes contain identically treated total RNA samples without DNase treatment, while the last two lanes contain identically treated total RNA samples with DNase treatment. The first two samples contain a high molecular weight band corresponding to *E. coli* genomic DNA; this bands is not present in the DNase-treated samples. The strongest bands in total RNA samples arise from the two ribosomal RNA molecules (23S and 16S), which are in high abundance in any *E. coli* total RNA sample. The existence of sharp, distinct rRNA bands on the gel indicates that the RNA in the sample is not degraded; degradation would cause smearing of these bands. The last two lanes of the gel are typical of total RNA isolated as described in Section 3.7.6.

3.7.7.1 CyScript™ Reverse Transcriptase Method

Labeling of total RNA samples was performed using reagents from the CyScribe™ First-Strand cDNA Labeling Kit (Amersham Biosciences, RPN-6200); however, the manufacturer's protocol was only loosely followed. 25 μg of a total RNA sample was placed in an RNase-free microtube. If the total RNA sample was taken from a N_2-induced culture, then 35-40 μg was typically used. Enough RNase-free water (from the CyScribe™ Kit) was added to bring the total volume to 10 μL and the RNA concentration to 2.5 mg/mL. Occasionally, the concentration of the original RNA sample was slightly lower than 2.5 mg/mL; in these cases 10 μL of the total RNA sample was added without water.

Next, 1 μL of Random Nonamers (from the CyScribe™ Kit) was added to the diluted RNA. The microtube was then incubated at 70°C for 5 min to denature the RNA, thereby increasing binding opportunities for the random nonamers. The microtube was placed at room temperature for 10 min to allow the random nonamers to anneal.

The following reagents were added to the microtube and were mixed well in the following order to bring the total volume in the microtube to 20 μL

- 4 μL 5× CyScript™ Buffer (from CyScribe™ Kit)
- 2 μL 0.1-M dithiothreitol (DTT) Solution (from CyScribe™ Kit)
- 1 μL dUTP nucleotide mixture (from CyScribe™ Kit). This mixture contains all four dNTP's, with levels of dTTP lower than those of the other three dNTP's.
- 1 μL 1-mM FluoroLink™ Cy3-dUTP (Amersham Biosciences PA-53022).
- 1 μL CyScript™ reverse transcriptase (from CyScribe™ Kit)

The microtube was incubated at 42°C for 1.5 h to carry out the enzymatic reaction. After addition of the fluorescent labels, microtubes were wrapped in foil and were only exposed to light briefly for pipetting steps to avoid inactivating the label. After the reaction was complete, the RNA template was degraded by addition of 2 μL 2.5-M NaOH and incubation at 37°C for 15 min. Addition of 10 μL 2-M HEPES free acid neutralized the pH.

The labeled DNA was then cleaned by removing unincorporated label and the enzyme with a QIAquick™ PCR Purification Kit (Qiagen 28014). As indicated by the standard protocol for this kit, 5 volumes of Buffer PB (160 μL) were added to each microtube and were mixed.

The manufacturer's protocol for this kit was followed and DNA was eluted from the column using 50 µL Buffer EB.

Modifications made to this protocol during experiments with Print E arrays included the use of Amber tubes (USA Scientific 1615-5507) to further protect the sample from light and a 2-min cooling step on ice before addition of 2-M HEPES free acid.

3.7.7.2 SuperScript™ Reverse Transcriptase Method

The SuperScript™ method was similar to that described for the CyScribe™ method above. The same mass of total RNA was used and it was diluted in the same way. 1 µL of random hexamers primers (Invitrogen 48190-011) was added to each microtube and the primer annealing step was carried out as described in the previous section.

The following reagents were added to the microtube and were mixed well in the following order to bring the total volume in the microtube to 25 µL

- 5 µL 5× First Strand Buffer (Invitrogen 18064-014)
- 2.5 µL 0.1-M dithiothreitol (DTT) Solution (Invitrogen 18064-014)
- 1.5 µL dNTP mix. This mixture was assembled from four separate solutions (Invitrogen 10297-018, dNTP mix sold as 100 mM solutions). The final mix contained 12.5 mM dATP, dCTP, and dGTP and 5 mM dTTP.
- 3 µL 1-mM FluoroLink™ Cy3-dUTP (Amersham Biosciences PA-53022).
- 2 µL SuperScript™ II reverse transcriptase (Invitrogen 18064-014)

The microtube was incubated at 42°C for 1.5 h to carry out the enzymatic reaction. After addition of the fluorescent labels, microtubes were wrapped in foil and were only exposed to light briefly for pipetting steps to avoid inactivating the label. After the reaction was complete, 1.5 µL EDTA was added to the microtube to quench the reaction. The RNA template was degraded by addition of 1.5 µL 500-mM NaOH and incubation at 70°C for 10 min. Addition of 1.5 µL 500-mM HCl neutralized the pH. Cleanup was carried out as described above.

3.7.8 Fluorescent Labeling of Genomic DNA Samples

The protocol followed for labeling of genomic DNA samples is very similar to those described in Section 3.7.7 for total RNA samples. This protocol is based on the Pollack protocol (Pollack *et al.* 1999). Genomic DNA isolated as described in Section 3.7.2 was used for this

protocol. Based on the absorbance readings taken after isolation, 1 μg (typically 5 μL) of this solution was transferred to a clean microtube. To this tube, the following was added:

- 2.50 μL random primers (hexamers) (Invitrogen 48190-011)
- 2.50 μL 10× *EcoPo*I Buffer (New England Biolabs M0210S)
- Sterile Milli-Q Water to bring the total volume in the microtube to 17 μL (typically 7 μL)

The microtube was heated at 95°C for 5 min to denature the genomic DNA fragments. The microtube was placed on ice for 5 min to allow the primers to anneal.

The following components were added to the microtube and mixed well to bring the total volume to 25 μL.

- 1 μL dNTP mixture (3.125 mM dATP, dCTP, dGTP, 0.781 mM dTTP, sold as 100 mM solutions of each dNTP, Invitrogen 10297-018)
- 2 μL 1-mM FluoroLink™ Cy5-dUTP (Amersham Biosciences PA-55022)
- 5 μL 5-U/μL Klenow Fragment (*E. coli* DNA Polymerase I Large Fragment, New England Biolabs M0210S)

The microtube was incubated at 37°C for 1.5 h to allow the labeling reaction to proceed. After addition of the fluorescent labels, microtubes were wrapped in foil and were only exposed to light briefly for pipetting steps to avoid inactivating the label. At the end of the reaction, 1.25 μL 0.5-M EDTA, pH 8.5-9.0 was added to the microtube to quench the reaction.

The labeled DNA was cleaned by removing unincorporated label and the enzyme with a QIAquick™ PCR Purification Kit (Qiagen 28014). As indicated by the standard protocol for this kit, 5 volumes of Buffer PB (130 μL) were added to each microtube and mixed. The manufacturer's protocol for this kit was followed and the DNA was eluted from the columns using 50 μL Buffer EB.

3.7.9 Prehybridization

Since the DNA microarray slides were not chemically treated to block non-specific binding to the surface, it was necessary to perform a prehybridization step to protect against non-specific binding. Incubation of the slide in a solution containing bovine serum albumin (BSA) allowed the albumin to occupy non-specific binding sites outside of the spots, thereby reducing the non-specific binding of labeled DNA and lowering the background signal.

Prehybridization buffer was prepared fresh the day of the experiment. For each slide, 450 mg of BSA was dissolved in 29.25 mL sterile Milli-Q water. The solution was supplemented with 11.25 mL of 20× SSC stock (3 M NaCl, 0.3 M sodium citrate, pH 7.0) and 4.5 mL 1% SDS, to give 45 mL of a 5× SSC, 0.1% SDS hybridization buffer. After the BSA had completely dissolved, the prehybridization buffer was sterile filtered, transferred to 50-mL Falcon tubes, and warmed to 45°C. The slides to be hybridized were placed in warm hybridization buffer in a 45°C water bath for 30 - 60 min.

3.7.10 Hybridization

After both the Cy3-labeled DNA generated from the sample and the Cy5-labeled DNA generated from the *E. coli* genomic DNA standard were isolated, 5 µL of each sample was used to determine concentration, purity, and label incorporation, as described in Sections 3.6.2 and 3.6.3. This step is very important for troubleshooting problems with the hybridization and can be used to rule out errors in the labeling reactions and subsequent isolations.

Next, each Cy3-labeled DNA solution was combined with one of the Cy5-labeled standard solutions to bring the total volume to about 90 µL. The DNA precipitation protocol in Section 3.6.1 was carried out, beginning with addition of 9 µL 3-M sodium acetate followed by 270 µL ethanol.

During the DNA precipitation, cover slips were washed with sterile Milli-Q water and ethanol and dried with compressed nitrogen.

To each DNA pellet, 1 µL 10-mg/mL sonicated salmon sperm DNA was added. This component is necessary for occupying non-specific binding sites within the DNA spots. This solution was then supplemented with concentrated buffer and pure formamide to produce a 5× SSC, 1% SDS, 25% formamide solution. For slides printed with 16 pins, a 25 mm × 25 mm cover slip was used and a 12-µL hybridization volume was found to be appropriate. For slides printed with 32 pins, a 24 mm × 40 mm cover slip was used with 20 µL hybridization volume. The microtube was heated at 95°C for 5 min.

The slide to be hybridized was removed from the prehybridization buffer, rinsed with sterile Milli-Q water, dried quickly with compressed nitrogen, and placed in a hybridization chamber (Corning Microarray Technology 2551). After the 95°C incubation had completed, the sample was immediately centrifuged for a few seconds to collect the condensate and was

pipetted onto the microarray printed on the slide. A clean cover slip was placed on top and every effort was made to eliminate bubbles, as they prevent the hybridization solution from contacting the slide and result in little or no signal. The chamber was sealed and placed in a 45°C water bath for 16 h.

3.7.11 Slide Washing

Wash buffers were prepared by filtering and autoclaving to prevent contamination of the slides. For each slide hybridized, three 50-mL Falcon tubes were filled with each of the three Array Wash Buffers:

- Array Wash #1 – low-stringency wash: 1× SSC, 0.2% SDS
- Array Wash #2 – high-stringency wash: 0.1× SSC, 0.2% SDS
- Array Wash #3 – final wash: 0.1× SSC

Tubes containing Array Wash #1 were preheated to 45°C before the hybridization chambers were removed from the water bath. Slides were quickly removed from the hybridization chambers and placed in Array Wash #1. The cover slip typically slid to the bottom of the slide; but occasionally, shaking was necessary to get the cover slip to move. After the cover slip had reached the bottom of the slide, the slide was picked up using forceps and replaced in the wash buffer so that the cover slip was on the back of the slide. This released the hybridization solution and exposed the entire microarray to the wash buffer. Falcon tubes were replaced in the 45°C water bath for 4 min. Slides were transferred to Array Wash #2 for 4 min at room temperature, and subsequently to Array Wash #3 for 4 min at room temperature. Next, each slide was placed in a 50-mL Falcon tube containing only a crumpled Kimwipe in the conical bottom to absorb the remaining liquid from the slide. Falcon tubes were centrifuged at 1,000 rpm (300 × *g*) at room temperature (IEC CRU-5000) for 2 min. Remaining liquid on the slides was blown away using compressed nitrogen.

3.7.12 Slide Scanning

Slides were scanned at 532 nm (for Cy3) and 635 nm (Cy5) on a GenePix® 4000B Scanner (Axon Instruments, Union City, CA). PMT Voltages were selected as described in Section 4.3, and the four images (low-resolution preview scan, 532 image, 635 image, and ratio image) were saved as TIFF files at a resolution of 10 μm/pixel.

3.7.13 Image Analysis

Image analysis is the link between the experimental protocol and data analysis. In this step, the output from the experimental protocol—the TIFF image—was used to generate the input to the data analysis procedure—a table of spots and their associated genes and statistics. The TIFF images were analyzed using the GenePix® Pro 3.0 software (Axon Instruments, Union City, CA).

Using the GAL file, GenePix® placed a grid of virtual spots, or features, over the image. The most difficult part of image analysis was aligning these virtual spots over the actual spots. Initially, the spot diameters ranged from 10-15 pixels; but, when necessary, these diameters were adjusted to visually fit the average spot size on the image. As a first pass, a manual, rough fit was made between the virtual spots and the actual spots on the image. This was followed by a fit using the Alignment algorithm in GenePix®, which simultaneously aligned all of the spots on the image. This algorithm was set to resize the virtual spots anywhere between twice and half their original size (200% maximum and 50% minimum). GenePix® also allows the user to input a threshold value to distinguish the feature pixels from background pixels; however, this threshold value was set to zero, thereby placing increased emphasis on spot filtering, as describe in Section 3.7.15.

The alignment was reviewed visually. Occasionally, tiny scratches or dust particles appeared on the slides and interfered with the alignment. When these imperfections covered a spot and made it impossible to quantify, the GenePix® flagging feature was used to mark the spot as "Bad." At other times, the alignment missed a perfectly good spot and instead chose a dust particle as the spot. In these cases, the virtual spot was aligned manually. As an estimate, changes were made to less than 1% of spots aligned by the GenePix® algorithm. With the alignment complete, the feature pixels were defined.

Defining the background pixels required additional work from GenePix®. For each spot, the software defined a concentric circle with three times the diameter of the spot (Figure 3.4). Any pixels within this circle and at least two pixels away from any feature were considered to be background pixels. With the background region defined, GenePix® calculated statistics such as mean, median, and standard deviation for each spot with both feature and background pixels.

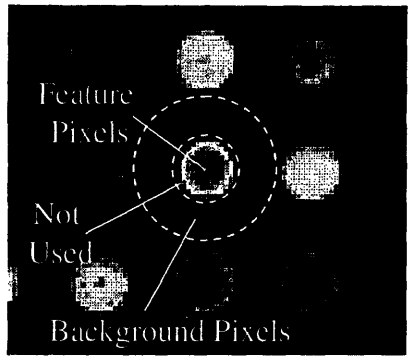


Figure 3.4: Definition of Background Pixels
A two-pixel annulus surrounding the feature was defined by GenePix® and was not used in the analysis. Next, GenePix® defined a region concentric to the feature with three times the diameter. All pixels within this region and outside of the two-pixel annulus of any adjacent spot were considered to be background.

3.7.14 Calculation of Signal and Log Ratios

For each spot, signal was calculated in each channel using the following formula

$$G = (I_{F,532})_{Med} - (I_{B,532})_{Med} \quad \textbf{(3.4a)}$$

$$R = (I_{F,635})_{Med} - (I_{B,635})_{Med} \quad \textbf{(3.4b)}$$

In the above equations, R is the Cy5 Signal, G is the Cy3 Signal, and $(I)_{Med}$ variables represent the median feature intensity for feature (F) and background (B) at wavelengths 635 nm and 532 nm. The signal ratio, y, was calculated as follows.

$$y = \log_2\left(\frac{G}{R}\right) \quad \textbf{(3.5)}$$

Notice that because the genomic DNA control was always labeled with Cy5, the Cy5 signal always appeared in the denominator of the signal ratio. This way, increased expression in the Cy3-labeled sample resulted in an increased signal ratio.

3.7.15 Spot Filtering

The first step in microarray data analysis was removing unwanted spots from further analysis. Unwanted spots were identified in several ways, as shown below

3.7.15.1 Failed PCR's

Based on analysis of the PCR products performed by the Lovett Lab, 208 of the 4,290 reactions failed after two attempts (see Section 3.7.1). Although some of these spots produced strong signal, they were all removed from further data analysis because they did not appear as expected on native agarose gels.

3.7.15.2 Control Spots

Control spots were also removed from further analysis. The controls are present to help interpret data from a single array. Beyond that, they have little use. Most of the control spots typically gave either very weak signal or very strong signal and, therefore, would tend to skew the results if included.

3.7.15.3 Spots Affected by Carryover

In one batch of arrays (Print E), spots following rRNA control spots exhibited unusually strong signal. These unnaturally high signals were attributed to carryover of rRNA during the printing process. Despite strong signals, the three spots following each of the sixteen rRNA control spot were eliminated from further data analysis.

The same phenomenon was observed in the first spots of several grids in the same print. Because a test print with food coloring had been printed immediately prior to starting the plates, these unnaturally high signals were also attributed to carryover. The first five spots of each of the sixteen affected grids were also eliminated from further analysis in this print.

3.7.15.4 Manually Flagged Spots

During the process of image analysis, some spots were manually flagged as "Bad," because data could not be reliably extracted from them. Typically, these spots were obstructed by a piece of dust or a scratch.

3.7.15.5 Spots with Low Signal

Spots with low signal were subjected to a *t*-test for both the Cy3 and Cy5 channels to determine whether the mean pixel intensity in the feature was greater than the mean pixel intensity in the background. However, before the *t*-test was performed, an *f*-test was performed to determine whether the pixel variances in the feature and background regions were equal.

The hypotheses for the *f*-test were as follows:

$$H_0 : \sigma_B^2 = \sigma_F^2$$
$$H_1 : \sigma_B^2 \neq \sigma_F^2 \tag{3.6}$$

where σ_B^2 is the background pixel variance and σ_F^2 is the feature pixel variance. The variables S_B and S_F are the standard deviation estimators for the background and feature regions, respectively. n_B and n_F are the number of pixels in each region. The values $f_P(DF1,DF2)$ are the values of the F distribution at probability P and degrees of freedom $DF1$ and $DF2$.

A 99% confidence level was chosen for this *f*-test. If $f_{0.01}(n_B - 1, n_F - 1) \leq \frac{S_B^2}{S_F^2} \leq f_{0.99}(n_B - 1, n_F - 1)$, then the null hypothesis in (3.6) could not be rejected and the variances were taken to be equal. Depending on the outcome of this *f*-test, two cases were considered.

Case 1 – Variances Equal. In this case, a pooled variance was calculated for both regions using the formula (Milton and Arnold 1995).

$$S_P^2 = \frac{(n_B - 1)S_B^2 + (n_F - 1)S_F^2}{n_B + n_F - 2}$$

Next, a one-tailed *t*-test was performed using the following hypotheses

$$H_0 : \mu_B = \mu_F$$
$$H_1 : \mu_B < \mu_F \tag{3.7}$$

$\bar{I}_B$ and $\bar{I}_F$ are the mean estimators of pixel intensity for the background and feature regions, respectively. $t_{0.95}(DF)$ is the value of the T distribution using a 95% confidence level at degrees of freedom DF. If $t_{0.95}(n_B + n_F - 2) \leq \frac{(\bar{I}_B - \bar{I}_F)}{\sqrt{S_P^2\left(1/n_B + 1/n_F\right)}}$ then the null hypothesis in (3.7) was rejected (Milton and Arnold 1995) because there was evidence that the feature mean was larger than the background mean.

Case 2 – Variances Unequal. In this case, an unpooled degrees of freedom was calculated as follows (Milton and Arnold 1995)

$$DF_{UP} = \frac{\left(S_B^2/n_B + S_F^2/n_F\right)}{\frac{\left(S_B^2/n_B\right)}{n_B - 1} + \frac{\left(S_F^2/n_F\right)}{n_F - 1}}$$

Again, a one-tailed *t*-test was performed using the same hypotheses in (3.7). The *t*-statistic was calculated and tested as follows: If $t_{0.95}(DF_{UP}) \leq \frac{(\bar{I}_B - \bar{I}_F)}{\sqrt{S_B^2/n_B + S_F^2/n_F}}$ then the null hypothesis in (3.7) was rejected (Milton and Arnold 1995) because there was evidence that the feature mean was larger than the background mean.

In either case, two criteria were tested for both channels of every spot. First, for a high-quality spot, the null hypothesis in (3.7) would be rejected. Second, the signal for the particular channel, calculated as in (3.5), should be a positive value. If a spot did not meet both of these criteria, it did not have significant signal in the channel being tested and was removed from further analysis. Notice that these two criteria are distinct, since the *t*-test is based on the *mean* pixel intensities and the signal is based on the *median* pixel intensities.

3.8 Pulse-Chase Analysis of Protein Degradation

Developed in previous work (Laska 2000), the pulse-chase protocol was used to quantify *in vivo* degradation of α_1AT. This protocol involved adding a pulse of radiolabeled methionine to a growing culture, followed 3 min later by a chase with excess unlabeled methionine. Analysis with SDS-PAGE allows one to follow the fate of all protein produced during the 3-min pulse period.

3.8.1 Growth, Induction, and Sample Collection

100-mL cultures were grown in air as described in Section 3.3 to OD_{600} of 0.7. At this point, a 150-µL pre-induction sample was taken and frozen, the culture was induced by adding IPTG to a concentration of 0.4 mM, and 6 mL of culture was transferred to a 30-mL Pyrex bubbler tube with cap and gas line. When experiments were performed at different aeration conditions (*i.e.* N_2, air, and O_2), a single culture was grown and split among three bubbler tubes immediately after induction. After 55 min of induction, 1000 µL of culture sample was

withdrawn from each bubbler tube. 150 µL of this sample was immediately frozen in liquid nitrogen as the pre-pulse sample, while the remaining 850 µL was used for OD_{600} measurement. After 60 min of induction, 15 µL ^{35}S-methionine (New England Nuclear NEG-709A EasyTag™ methionine) was pulsed into each culture. After a 3-min pulse period, labeling was quenched by the addition of 19.2 µL Chase Solution (40 mg/mL methionine, 10 mg/mL cysteine). 150-µL samples were taken at 0.5, 1, 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 50, and 60 min after the chase, and each was immediately frozen in liquid nitrogen. The time of each sample was recorded to the precision of 1 s. As the samples accumulated, they were placed at -80°C until the analysis was performed.

3.8.2 Analysis by SDS-PAGE

Samples were analyzed by loading the 17 samples on a polyacrylamide gel. Samples were thawed on ice for approximately 1 h and were centrifuged at 4°C for 5 min in a Sorvall MC-12V centrifuge. Supernatants were replaced with 150 µL 1× Reducing SDS Sample Loading Buffer (62.5 mM Tris, pH 6.8, 2% SDS, 5% β-mercaptoethanol, 10% glycerol, 0.1% bromophenol blue) and pellets were resuspended by pipetting.

Gels were prepared in the lab, using a protocol similar to that described previously (King and Laemmli 1971). Reagents for preparing gels include 30% Acrylamide/Bis solution (37.5:1) (Bio-Rad 161-0158), ammonium persulfate (APS) (Bio-Rad 161-0700), and *N,N,N',N'*-tetramethylethylenediamine (TEMED) (OmniPur 8920). Separating Gel Buffer (1.5 M Tris, pH 8.8, 0.4% SDS) and Stacking Gel Buffer (0.5 M Tris, pH 6.8, 0.4 % SDS) were prepared in advance, sterile-filtered, and stored at 4°C. For a single gel, the lower separating gel (7.5%) was prepared by combining 2.5 mL acrylamide, 5 mL Milli-Q water, 2.5 mL Separating Gel Buffer, 50 µL fresh 10% APS (<1 week old), and 5 µL TEMED in a flask. The solution was mixed well and immediately transferred to a Criterion Cassette (26-well, 1.0 mm, Bio-Rad 345-9903). Since exposure to oxygen inhibits polymerization, the gel solution was overlaid with a small amount of Milli-Q water. After 1 h, the separating gel had polymerized. The water was poured off, and excess water was absorbed with filter paper. The upper stacking gel was prepared by combining 1 mL acrylamide, 6.5 mL Milli-Q water, 2.5 mL Stacking Gel Buffer, 50 µL fresh 10% APS, and 10 µL TEMED. This solution was pipetted to fill the top of the cassette and the comb was inserted. Polymerization proceeded for 45 min.

After stacking gel polymerization was complete, gel cassettes were placed in the Criterion cell (Bio-Rad) and the top reservoir of the cassette was filled with 4°C SDS Running Buffer (Section 3.5.3) for loading. Samples were heated at 95°C for 5 min and were loaded using a Hamilton syringe. The syringe was rinsed between samples by pulling liquid up and down seven times. A standard of α_1AT purified using a previously developed method (Griffiths 2002; Griffiths and Cooney 2002) was placed in one lane of each gel. Samples from each culture were analyzed on separate gels. After all gels were loaded, the lower reservoirs were filled with 4°C SDS Running Buffer and the gels were run at 20 mA/gel until just before the dye front ran off the bottom of the gel (~2 h).

When complete, gels were removed from cassettes and stacking gels were cut off and discarded. A corner was cut from each gel to help orient it. Gels were stained as described in Section 3.5.3 and were then washed three times for 10 min in distilled water to remove excess acetic acid, because it can damage the gel drier.

Gels must be dried for exposure to the phosphor screen because the weak β-decay of ^{35}S can be blocked by water in the gels. Gels were laid out on a piece of Saran Wrap and were covered with a 20 cm × 25 cm piece of filter paper (Bio-Rad 165-0962) wetted with distilled water. This method was preferred over placing the gels on the filter paper, because the gels were much easier to arrange on plastic wrap than on filter paper. The gels and filter paper were inverted and the layer of plastic wrap was replaced by a plastic-wrapped piece of filter paper. This sandwich was placed in the gel drier (Bio-Rad Model 583) with the wet filter paper on the bottom. The drier was run at a constant 85°C under vacuum for 2 h.

Dried gels were exposed to a 20 cm × 25 cm phosphor screen (Molecular Dynamics) overnight. The screen was scanned on a Molecular Dynamics PhosphorImager® 445 SI using ImageQuant software. These images were analyzed in ImageQuant. Individual bands were quantified using the Local Median background correction option. Signals from entire lanes were quantified in the same way.

3.8.3 Pulse-Chase Data Analysis and Modeling

Pulse-chase data from individual α_1AT bands were scaled to the total signal from the lane, corresponding to all of the protein synthesized during the 3-min pulse period. This scaling was intended to account for inconsistencies in sample volume loaded on the gel. To these

corrected data, the pulse-chase model below was applied to calculate the rate constant of folding, k_f, and the pseudo-first-order rate constant of proteolysis, k_p. This model was adapted from previous work (Laska 2000):

$$\frac{AT}{AT_0} = \frac{K \cdot t_p + \frac{k_p}{k_f} \exp(-K \cdot t)\left[1 - \exp(-K \cdot t_p)\right]}{K \cdot t_p + \frac{k_p}{k_f}\left[1 - \exp(-K \cdot t_p)\right]} \tag{3.8}$$

In this model, t_p is a constant representing the duration of the pulse (3 min). AT is the corrected signal from the α_1AT band, t is the time after the chase at which the sample was taken, and $K = k_f + k_p$. One difference between the model used here and its original form is that the normalization constant AT_0 is taken to be the corrected α_1AT signal exactly at the time of the chase. This parameter was not treated as an input to the model, but rather as an output from the model, in addition to k_p and k_f. This allowed data from the first sample (30 s) to be treated exactly the same as data from all other samples.

Based on these model parameters, the apparent extent of degradation was calculated as follows.

$$X_{app} = \frac{\frac{k_p}{k_f}\left[1 - \exp(-K \cdot t_p)\right]}{K \cdot t_p + \frac{k_p}{k_f}\left[1 - \exp(-K \cdot t_p)\right]} \tag{3.9}$$

The ratio of proteolysis and folding rates was calculated as the ratio of the corresponding rate constants.

$$\frac{r_p}{r_f} = \frac{k_p}{k_f} \tag{3.10}$$

Confidence intervals for all of the output parameters were difficult to calculate since (3.8) cannot be transformed to a linear model. To estimate the error in these model parameters, a computational procedure was used to perturb the original data and recalculate the model parameters. After 1,000 repetitions of this procedure, a distribution was obtained for each parameter; the confidence interval was taken as the standard deviation of this distribution.

Values of t were perturbed according to a normal distribution with standard deviation of 1 s. Values of AT were perturbed according to a normal distribution with standard deviation corresponding to 2% of the original data value. The value of 2% was selected because it was representative of the deviation between individual data points and the model line.

3.9 Amino Acid Analysis

This protocol was used to analyze free intracellular amino acid levels in cultures immediately before and 60 min after induction.

3.9.1 Growth, Induction, and Sample Collection

Three 100-mL cultures were grown as described in Section 3.3. Two of these were cultures of *E. coli* BL21 (DE3) pEAT8-137, which contained the recombinant α_1AT gene. The third was a culture of *E. coli* BL21 (DE3) pET3d, which did not contain the α_1AT gene. At OD_{600} of 0.7, 20-mL samples were withdrawn from each culture. From the remaining volume, another 50-mL from each culture and was transferred to a sterile 250-mL shake flask. IPTG was added to the pET3d culture (empty-vector) and to one of the pEAT8-137 cultures (induced) to a concentration of 0.4 mM. Nothing was added to the third culture (uninduced). After 60 min of induction, another 20-mL sample was withdrawn from each of these cultures.

3.9.2 Sample Preparation

These 20-mL samples were placed in conical centrifuge tubes and were placed on ice for 10 min. The cells were separated in two stages. First, these tubes were centrifuged for 20 min at 2,000 rpm (1,100 × *g*) at 4°C (IEC CRU-5000). The medium was decanted from the cell pellets, which were then resuspended in 1 mL Milli-Q water and transferred to 1.5-mL microtubes. These tubes were centrifuged at 9,000 rpm (5,000 × g) at 4°C (IEC Centra-4) for 15 min. Supernatant was discarded and the pellet was resuspended in 270 µL Milli-Q water. To each tube, 30 µL 50% trichloroacetic acid (TCA) was added to create a 5% TCA solution, which lyses the cells and precipitates proteins. The tubes were placed on ice for 30 min and were then centrifuged at 9,000 rpm (5,000 × g) at 4°C (IEC Centra-4) for 15 min. The supernatants were transferred to fresh tubes and an extraction was performed with 300 µL diethyl ether, in order to

remove TCA from the samples. The remaining aqueous fraction was placed in a vacuum centrifuge overnight.

To the residue, 50 μL of 0.4 N borate buffer (Agilent 5061-3339) was added. Each sample was heated at 50°C for 2 min and removed from the tube using a 23-gage needle and a 5-mL syringe. This small volume of sample was carefully filtered into an amber HPLC vial using a 0.2 μm PVDF membrane (Pall PN-4450T).

3.9.3 HPLC Analysis

Samples were analyzed by HPLC (Agilent 1050) using a Hypersil AA-ODS column (Agilent). 2 μL of samples was injected with 4 μL Borate Buffer and 1 μL *o*-phthalaldehyde (OPA). A gradient was generated using Buffer A (20 mM sodium acetate, pH 7.2, 5 μM EDTA, 0.018% triethylamine, 0.3% tetrahydrofuran—added after filtration) and Buffer B (20 mM sodium acetate, pH 7.2, 40% methanol, 40% acetonitrile). The detector was operated at 338 nm. At a constant flow rate of 0.45 mL/min, a gradient was run as follows:

- Held at 0% B at 0-2 min
- Ramped up to 10% B at 2-7 min
- Ramped up to 15% B at 7-15 min
- Ramped up to 60% B at 15-30 min
- Held at 60% B at 30-40 min
- Ramped down to 0% B at 40-42 min
- Held at 0% B at 42-50 min

In addition to samples, an amino acid standard (Sigma AA-S-18) was also run periodically to account for changes that occurred over time. The peaks from this standard were used to identify amino acid peaks in samples.

4 Development of DNA Microarrays as a Quantitative Assay

> "Baseball in the only field of endeavor in which a man can succeed three times out of ten and be considered a good performer."
>
> —*Ted Williams*

Before high-throughput data were generated using DNA microarrays, a series of validation experiments was performed. These experiments were intended to answer several questions about this analytical technique.

- How can the experiments be designed to allow for comparisons of data from multiple arrays?
- How can the experiments be designed to account for experimental variation and allow for its correction?
- Is this assay quantifiable?
- How can microarray data be analyzed to select genes that show differential expression?
- Is this assay reproducible?
- Can expected gene expression patterns be revealed?

This chapter presents the development of the DNA microarray protocols used in this thesis and describes the experiments performed to answer these questions.

4.1 *Bacillus subtilis* ORF's as Internal Controls

Normalization is the process of scaling signal data to allow comparison of either two samples on a single array or multiple samples from multiple arrays. One way to normalize microarray data is to use internal controls, *i.e.* print spots of foreign DNA that show low homology with genes from the organism of interest (*E. coli* in this case). Corresponding RNA sequences can be doped into the labeling reaction in known quantities in order to generate targets that are complementary to the printed DNA probes. This would allow signals from the Cy3 and Cy5 channels to be easily related to one another by the signals from these labeling control spots. For instance, if the same amount of foreign RNA were added to each labeling reaction, then the Cy3 and Cy5 data could be adjusted so that the signals from these internal controls become equal.

An effort was initiated to use this method of internal controls for normalizing data on full-genome *E. coli* microarrays. Four *Bacillus subtilis* ORF's were selected as controls. Small (~150 bp) unknown genes were chosen, as these would be least likely to show homology with *E. coli* genes. Using primers that incorporated a T7 RNA polymerase binding site, PCR products (BS1, BS2, BS3, and BS4) were generated (as described in Section 3.6.6). The strategy for development of these internal controls is described in Figure 4.1. These small DNA probes were printed on *E. coli* microarrays and showed no signal when hybridized with *E. coli* RNA samples, indicating that the labeling controls do indeed show low homology with *E. coli* genes.

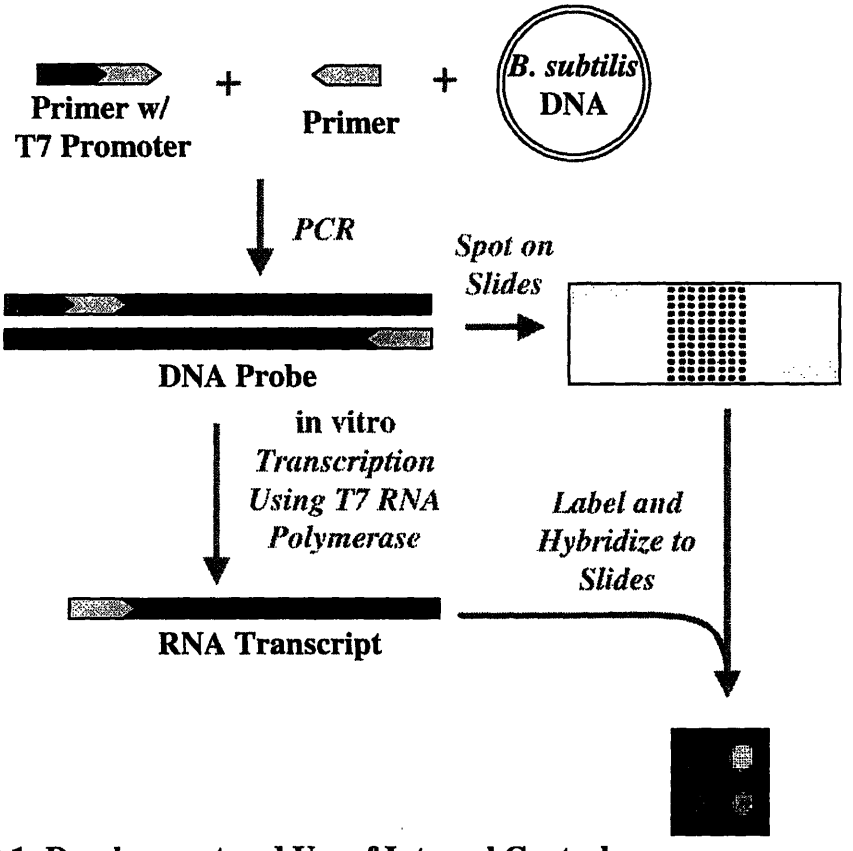


Figure 4.1: Development and Use of Internal Controls
Probes for use as internal controls can be generated by performing PCR with one primer that contains a T7 promoter region. Some of this probe would be spotted onto DNA microarrays as though it were an *E. coli* PCR product. The remaining probe would be used in an *in vitro* transcription reaction using T7 RNA polymerase to generate an RNA transcript. These transcripts would then be doped into both labeling reactions in known quantities with total RNA samples. These transcripts would generate signal in both channels at the corresponding spots on the array.

In vitro transcription, using T7 RNA polymerase, was performed to generate RNA fragments corresponding to the labeling controls. In order to isolate RNA from the reaction

mixtures, cleanup was attempted by using the same RNeasy Kit that was used for total RNA purification. Qiagen, the manufacturer, reports that the kit will only purify RNA fragments larger than 200 nucleotides (nt). As expected, high yields could only be achieved when RNA was precipitated without using the RNeasy kit. Yields from the reaction were acceptable; however, the small size of the RNA molecules in the samples also became an issue during the labeling protocol.

During labeling, small RNA controls are problematic because they are more susceptible to loss during the clean-up step. QIAquick™ PCR Purification Kits are reported to remove DNA fragments smaller than 100 bp. This step in the labeling protocol is needed to remove excess label and nucleotides as well as the reverse transcriptase enzyme. While the four labeling control RNA molecules are all larger than 100 nt, the labeled cDNA produced in the labeling step may not be. Because the labeling reaction is primed with random primers, the primer will just as likely bind to the far 5' end of the transcript as it will to the far 3' end of the transcript, as described in Figure 4.2A. Therefore, a large fraction of labeled cDNA molecules produced in the labeling step will be smaller than 100 nt. The loss of these small cDNA molecules during clean-up results in overall loss of the label and reduced signal on the arrays. In an effort to compensate for the lost controls, it is tempting to add more of the control RNA to the labeling mix. However, this strategy only increases the amount of label that is lost and further decreases the overall signal on the arrays.

One strategy that was attempted in order to improve labeling was to prime control transcripts with the same primers that were originally used to generate PCR products, as illustrated in Figure 4.2B. This would guarantee that labeled cDNA molecules would be full-length and therefore less prone to wash out during clean-up. Signals were no better during this hybridization. Most likely, there is still a significant loss of labeled cDNA molecules. The 100 nt cutoff is certainly not a hard and fast rule. Even molecules 150 nt in length may show significant loss.

Although these labeling controls would be helpful for normalization, they are not necessary on full-genome microarrays. With more than four thousand spots, it is safe to use a global normalization, *i.e.* assume that the total signal is constant across all microarrays. While further work may lead to a successful conclusion on this front, the labeling controls were put

aside. One benefit of eliminating these internal controls is that we were no longer constrained to comparing two RNA samples on the microarrays.

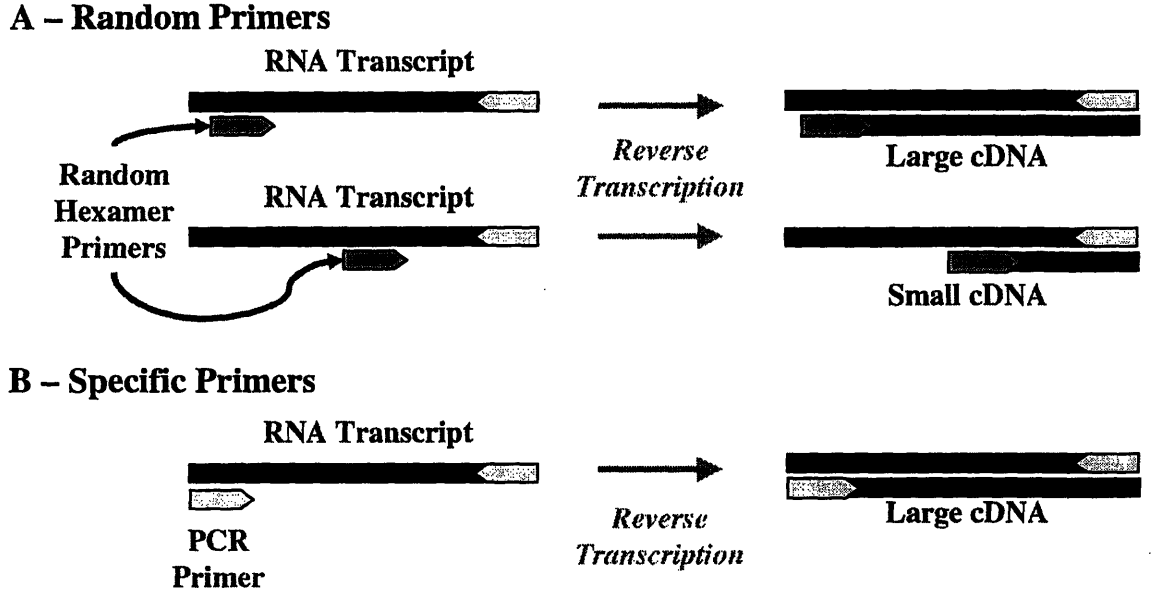


Figure 4.2: Reverse Transcription Labeling of RNA Molecules with Random Primers and Specific Primers
A) Since random primers may bind anywhere within the transcript, the product will be a mixture of large and small cDNA molecules. **B)** Primers specific to the transcript will produce only full-length cDNA molecules.

4.2 Genomic DNA as a Hybridization Control

Two-color hybridization using two different fluorescent dyes (typically Cy3 and Cy5) allows two samples to be compared on a single array. For comparisons of only two samples, (*e.g.* mutant A vs. mutant B) it is easiest to label one RNA sample using Cy3, label the other RNA sample with Cy5, and hybridize both to the same array. In contrast, time course experiments, which generate multiple samples, are not as straightforward to analyze and compare. One approach to analysis of time course experiments might be to compare every possible pairing of samples on a separate array (*e.g.* Sample1 vs. Sample2, Sample2 vs. Sample3, Sample3 vs. Sample1). However, the number of arrays needed for this approach would increase with the number of samples in a factorial relationship. Another standard approach is to hybridize each RNA sample to a separate array along with a common hybridization control sample. For example, this might involve labeling all samples with Cy3, labeling the hybridization control with Cy5, and comparing signal from each sample to that of the hybridization control on the

same array. With four or more samples, this approach uses fewer arrays and is therefore preferred.

One may question why hybridization controls are needed at all, *i.e.* why bother with two-color hybridization? If the experimental techniques are reproducible, should it not be easy to compare one-color signals from different slides to calculate expression ratios? For most spots, this will be true, but slight variations in spot morphology and background intensity, as illustrated in Figure 4.3 make complete reproducibility impossible. Differences in array print number can also reduce reproducibility. Arrays from the beginning of a print will have much larger spots than those printed toward the end, and spot size can affect signal. To help correct for these unavoidable variations, a hybridization control, or standard, is hybridized in the second channel. Any differences in spot morphology or background intensity would affect the signals from the sample and the standard in a similar manner such that a signal:standard ratio should be reproducible from array to array.

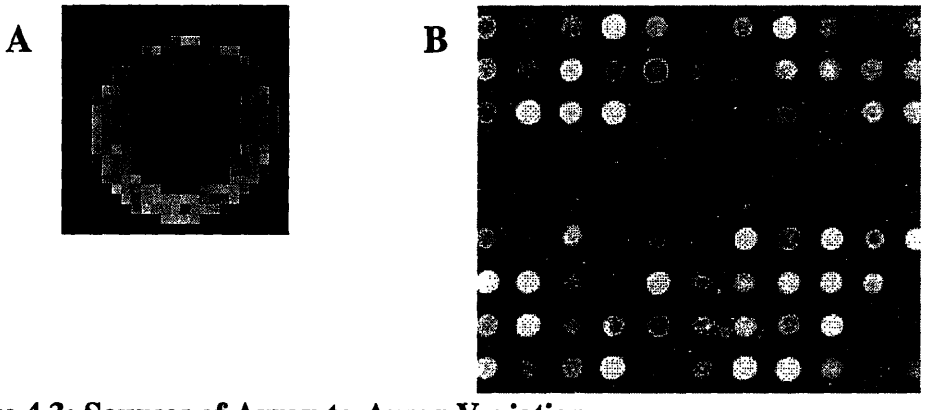


Figure 4.3: Sources of Array-to-Array Variation
A) Spot morphology can affect reproducibility. Donut shaped spots are common.
B) Differences in background intensity are another common source of signal variation

The next question is what to use as a hybridization control. One option is to use one of the RNA samples (typically the zero-time-point sample). This approach requires one fewer array, since one of the samples is used as a control. However, one drawback is that a higher volume of control sample is required. For microbial experiments where the zero-time-point sample occurs at low cell density, this issue becomes even more problematic because RNA yields from these samples are typically low. In addition, comparing arrays from separate experiments would not be possible since the control would be different for each experiment. Another drawback of this approach is that valid data can be lost if signal from a spot is

undetectable in the control sample and high in all other samples. In this case, log ratios could not be calculated and array-to-array comparisons would be impossible.

Several alternatives to this zero-time-point control have been used. One option is to combine equal amounts of each RNA sample in the experiment to generate a pooled control sample. Because this control is a composite of all the samples, spots that are detectable in any of the samples should also be detectable in the control (Eisen and Brown 1999). However, like the zero-time-point control, this pooled sample would not allow for comparisons of arrays from separate experiments. Another option for a control sample is a sample pooled from RNA isolated under a variety of stress conditions. Combining these RNA samples should give a standard with good representation of many genes in the genome. The drawback with this approach is that it would be impossible to exactly duplicate this standard should it be consumed or degraded.

The best option for a hybridization control is genomic DNA from the organism of interest. Because each DNA molecule contains one copy of each gene, this standard should theoretically produce detectable signal, at similar levels, in all spots. Furthermore, this standard can be regenerated at any time and can be used to compare arrays from different series of experiments. Genomic DNA is also more stable than any RNA sample. Another benefit of genomic DNA as a control is that it produces high signal-to-noise ratios. Because the coverage of protein-coding genes on the *E. coli* genome is 87.8% (Blattner *et al.* 1997), any fragment of genomic DNA has high odds of binding to one of the protein-coding ORF's spotted on the array. In contrast, total RNA samples contain roughly 80-90% rRNA, none of which should bind to spots on the arrays. However, labeled rRNA sequences can bind nonspecifically to produce high background signal. Therefore, a total RNA sample has a much higher likelihood of undetectable spots than does a genomic DNA sample. In the end, genomic DNA should produce a large number of spots with high signal.

The success of genomic DNA performance as a hybridization control must be determined. One study compared three different hybridization controls: genomic DNA, pooled RNA samples, and an RNA sample from a single time point (Kim *et al.* 2002). In self-versus-self hybridizations, all controls performed extremely well and were reproducible. A comparison of the Cy3-labeled genomic DNA and. Cy5-labeled genomic DNA signals revealed a correlation coefficient of 0.98-0.99. When used as hybridization controls with other RNA samples, a pooled

RNA control identified 176 differentially expressed genes, 127 of which (72%) were also identified by a direct comparison. In contrast, indirect comparisons using a genomic DNA control identified 168 differentially expressed genes, 78 of which (46%) were also identified by a direct comparison. Therefore, genomic DNA appears to be inferior as a control. However, the authors also point out that genes that were both highly and significantly differentially expressed were identified similarly using all methods.

In order to test the reproducibility of a genomic DNA control for our *E. coli* system, genomic DNA was isolated using a Genomic-tip kit as described in Section 3.6.5 and was digested with the restriction endonuclease *Hae*III as described in Section 3.7.4. Labeling of genomic DNA was performed by randomly priming 1 μg of DNA and performing a replication with DNA Polymerase I (Klenow) in the presence of Cy5-dUTP, as described in Section 3.7.8. *E. coli* total RNA was purified from a culture grown in minimal M9 Medium at 30°C to OD_{600} of 0.90. In two separate labeling reactions, Cy3-labeled cDNA was generated from this total RNA sample. Also in two separate labeling reactions, Cy5-labeled cDNA was generated using genomic DNA. Each Cy3-labeled sample was combined with a Cy5-labeled sample and hybridized onto a full-genome microarray.

The genomic DNA sample was always labeled with Cy5 in order to achieve similar signal from both channels. The incorporation of labeled nucleotides was much higher with the Klenow enzyme than with reverse transcriptase. However, Cy5 is less stable than Cy3 and, under identical labeling conditions, usually produced weaker signal. By allowing Klenow to incorporate Cy5, these two effects offset one another to give approximately equal values for both Cy3 and Cy5 label incorporation.

The Cy3 and Cy5 images from one of these microarrays are shown in Figure 4.4. It is immediately obvious that many more spots are visible in the Cy5 (genomic DNA) channel than in the Cy3 (total RNA) channel. As expected, genomic DNA produced signal from nearly all spots. On the Cy5 image, the column of blank spots in the middle of each grid was a string of negative controls that were not expected to show signal. Images from both of the duplicate arrays were analyzed and unwanted spots were removed as described in Sections 3.7.13 - 3.7.15. A small number of spots were eliminated due to low signal from only the genomic DNA hybridization control alone (Figure 4.5). As a hybridization standard, genomic DNA produced strong signal from a high fraction of spots on the microarray.

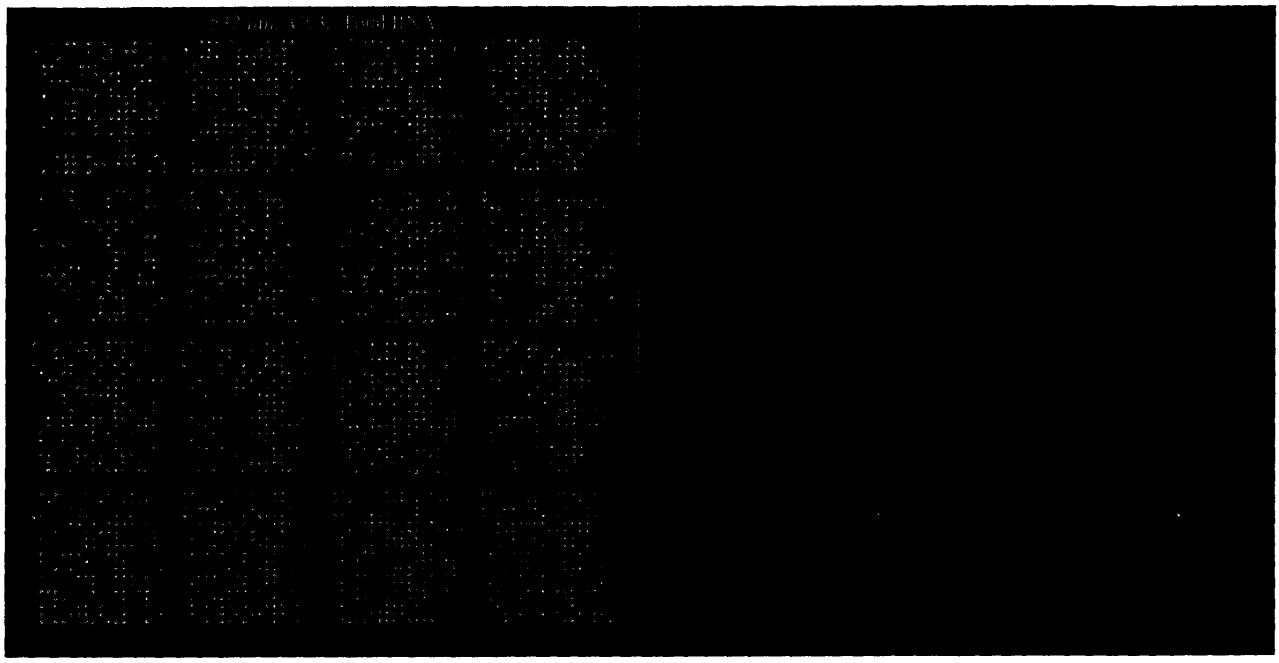


Figure 4.4: Images from Cy3 and Cy5 Channels
An *E. coli* total RNA sample was labeled using Cy3-dUTP and an *E. coli Hae*III-digested genomic DNA sample was labeled using Cy5-dUTP. The samples were hybridized to a full-genome array and the array was scanned at 532 nm (to detect Cy3) and 635 nm (to detect Cy5).

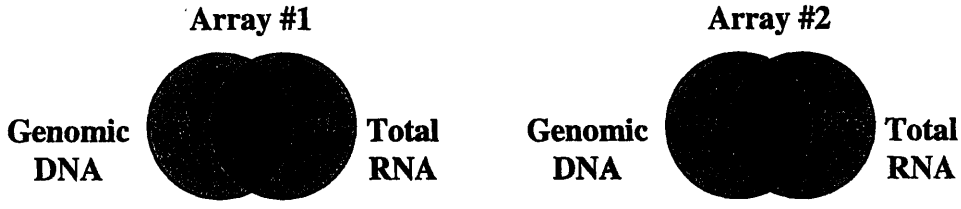


Figure 4.5: Number of Genes Removed Due to Low Signal
Array #1 had overall higher signal and had fewer spots eliminated from it. These data show that a relatively small number of spots were eliminated due to low signal from the genomic DNA hybridization control alone. The number of spots eliminated due to low signal from total RNA can be quite high since rRNA in the sample binds nonspecifically to increase the background signal.

Comparisons of both of the duplicate microarrays showed that signals in both Cy3 and Cy5 channels have high reproducibility. Cy3 and Cy5 signals from both arrays were compared by scatter plots (Figure 4.6). It is clear that data correlate well in both channels.

Correlation coefficients for both the genomic DNA and total RNA samples are shown in Figure 4.7. The correlation coefficients were equivalent, based on a 95% confidence interval. Therefore, data from identical genomic DNA samples was as reproducible as data from identical

RNA samples. Alternative standards that use RNA instead of genomic DNA would provide no significant improvement with regard to reproducibility. The correlation between the signal ratios from each array was just as strong as that from either of the two signals alone. Because the signals were already highly correlated, the hybridization control added nothing in terms of reproducibility.

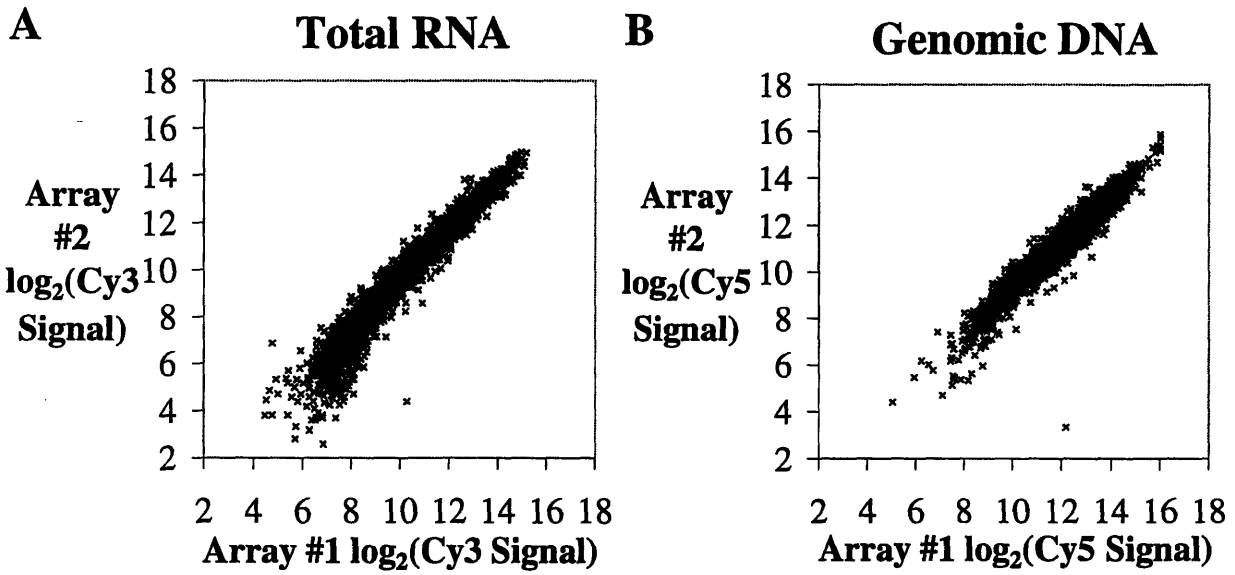


Figure 4.6: Comparisons of Two Microarrays with Identical Samples
Identical samples were labeled in separate reactions and were hybridized to separate arrays. **A)** Cy3 Signal generated from total RNA **B)** Cy5 Signal generated from genomic DNA. Note that signals have been log transformed but have not been normalized.

A second set of duplicate experiments was performed similar to the one above. The same total RNA sample was hybridized to different arrays. However, unlike the above experiment, the two genomic DNA samples were from two different preparations extracted from different cultures grown one month apart. Figure 4.8 shows correlation coefficients for these two arrays as a measure of signal reproducibility. Since the intervals overlap, it cannot be said that the correlation coefficients were distinct, at a 95% confidence level. This is further evidence supporting the claim that genomic DNA provides comparable reproducibility to total RNA. Furthermore, the fact that different DNA samples were used for these two arrays highlights an advantage of genomic DNA as a standard—it can be easily regenerated with minimal loss in reproducibility. This pair of arrays also illustrates the utility of a hybridization control. Use of genomic DNA as a hybridization control improved the reproducibility of the experiment from *R*

= 0.926 for total RNA alone to R = 0.945 for the ratio signal. Some of the variability that existed in the total RNA signal was removed by taking the ratio of total RNA signal : genomic DNA signal.

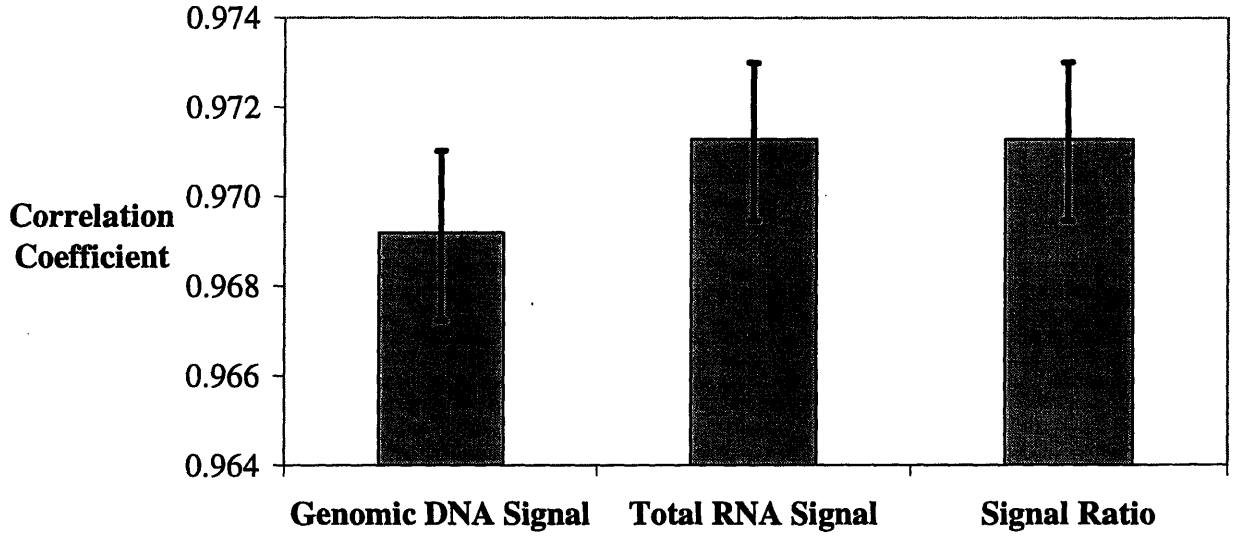


Figure 4.7: Correlation Coefficients from Duplicate Arrays
Two arrays with the same genomic DNA sample labeled with Cy5 and the same total RNA sample labeled with Cy3 were compared. Correlation coefficients (R) were calculated for the genomic DNA signal, the total RNA signal, and the ratio of these two signals (all on a base-2 log scale). Error bars correspond to bounds set by a 95% confidence interval and were defined as follows: lower and upper bounds are $\frac{(1+R)-(1-R)\exp\left(2z_{0.025}/\sqrt{n-3}\right)}{(1+R)+(1-R)\exp\left(2z_{0.025}/\sqrt{n-3}\right)}$ and $\frac{(1+R)-(1-R)\exp\left(-2z_{0.025}/\sqrt{n-3}\right)}{(1+R)+(1-R)\exp\left(-2z_{0.025}/\sqrt{n-3}\right)}$, respectively, where n is the number of data, and $z_{0.025}$ is the value of the normal distribution at probability 0.025 (Milton and Arnold 1995). The same spots were used for both comparisons (n = 3,959).

Correlation coefficients from the first set of duplicate experiments (Figure 4.7) are larger than those from the second set of duplicates (Figure 4.8) probably because the arrays used in the first experiment were from the middle and end of the print (arrays A59 and A107—48 arrays apart) whereas arrays used in the second experiment were from the beginning and end of the print (arrays A13 and A113—100 arrays apart). As stated previously, array number impacts spot size, which in turn affects the reproducibility of the signal. Arrays that are farthest apart in print number will be least reproducible. Notice also that when the variability between arrays is already low (as in the first set of duplicates), the hybridization control provides the least benefit.

In contrast, when the variability between arrays is large (as in the second set of duplicates), the hybridization control provides the greatest benefit.

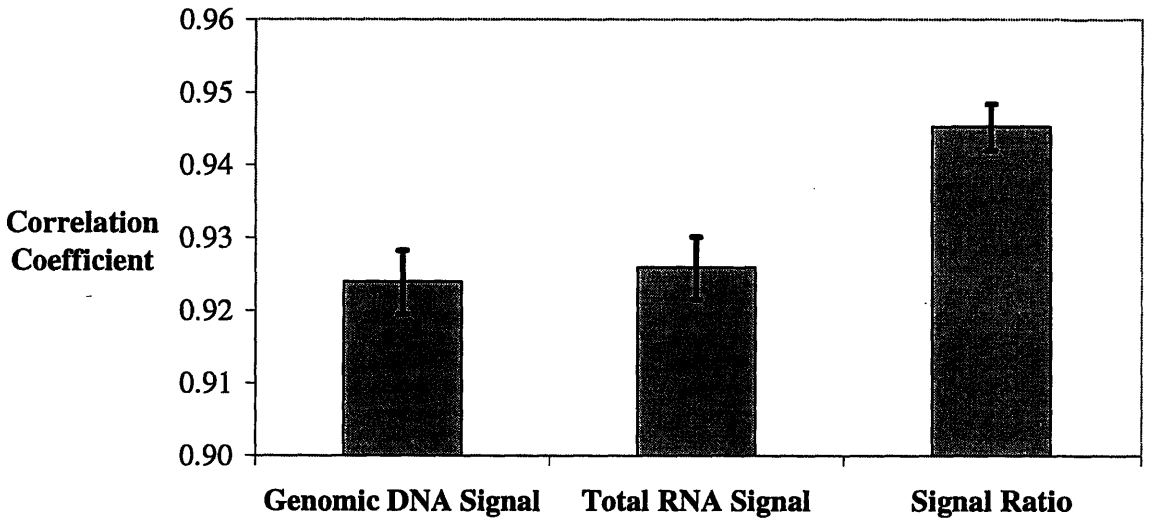


Figure 4.8: Correlation Coefficients from Duplicate Arrays with Genomic DNA from Different Preparations
A total RNA sample was labeled with Cy3, and different genomic DNA samples (extracted from two different cultures) were labeled with Cy5. Samples were hybridized to two separate arrays. Correlation coefficients (R) were calculated for the genomic DNA signal, and the total RNA signal, as well as the ratio of these two signals (all on a base-2 log scale). Error bars correspond to bounds set by a 95% confidence interval, as described in Figure 4.7.

Although the signal from a microarray is often consistent from array to array, imperfections in spot morphology and background signal lead to unavoidable inconsistencies. Hybridization controls are needed to account for array-to-array variability. Genomic DNA performs well as a hybridization control, in that it is as reproducible as any duplicated RNA sample. Genomic DNA has several advantages over an RNA hybridization control, including improved stability, low background signal, and high reproducibility even from samples prepared from different cultures. In the ultimate test, scaling the total RNA signal with the genomic DNA signal was found to produce more consistent data than using the total RNA signal alone.

4.3 Analysis of Detector

Data produced by microarrays is only as good as the detector used to scan them. Before the first spot was analyzed, it was critical to confirm that the GenePix 4000B scanner was being operated in the linear range of detection. Furthermore, it was necessary to determine the range of

signals over which the scanner was optimal as well as the optimal scanner settings that will maximize this range, such as photo multiplier tube (PMT) voltage. Scanner settings were studied by directly printing a set of arrays with the labels Cy3-dUTP and Cy5-dUTP.

Serial dilutions of unincorporated Cy3-dUTP and Cy5-dUTP were prepared in 10-mM phosphate buffer in a 384-well plate. For each label, each set of dilutions was repeated four times (Dilution Series A-D) on the plate. Concentrations ranged from 0.1 fM to 21 μM. Arrays were printed using the Virtek ChipWriter in a dimly-lit environment to limit degradation of the label signal. The pins used for this print are reported to deliver 0.6 nL in each spot. Based on this, the number of label molecules in each spot can easily be calculated.

These arrays were scanned immediately after printing and signals from the spots were related to the number of label molecules (Figure 4.9). These plots show that signal is proportional to the molecules of label in the spot over a concentration range spanning four orders of magnitude. At the two PMT voltages shown (650 V for Cy3 and 750 V for Cy5) the data for Cy3 and Cy5 labels are similar. Because Cy5 is less stable than Cy3, it is not surprising that a higher PMT voltage is needed to achieve similar intensities.

Based on signal vs. molecules data, values such as dynamic range, saturation level, limit of detection, and sensitivity were calculated. At the PMT voltages used in Figure 4.9, the sensitivities, calculated as $\frac{\Delta\log(Signal)}{\Delta\log(Molecules)}$, were found to be 0.87 for Cy3 and 0.90 for Cy5. Other scanning characteristics were calculated by scanning arrays at different PMT voltages. Figure 4.10 shows how these values vary with PMT voltage. At low PMT voltages, dynamic range was small due to a high limit of detection. At high PMT voltages, dynamic range was also small due to low saturation level. The optimal scanner settings were at intermediate PMT voltages, where the dynamic range peaked. Although unincorporated label may behave differently than labeled cDNAs, optimal scanner setting should be similar for both. Trends in sensitivity and signal-to-noise ratios were also considered in determining optimal scanner settings. The optimal ranges for PMT voltages were found to be 500-700 V for Cy3 and 600-800 V for Cy5. The dye dilution experiments carried out here were similar to those performed by the manufacturer (Pickett *et al.* 2001). That report suggested broad PMT voltage ranges of 500-900 V for both channels.

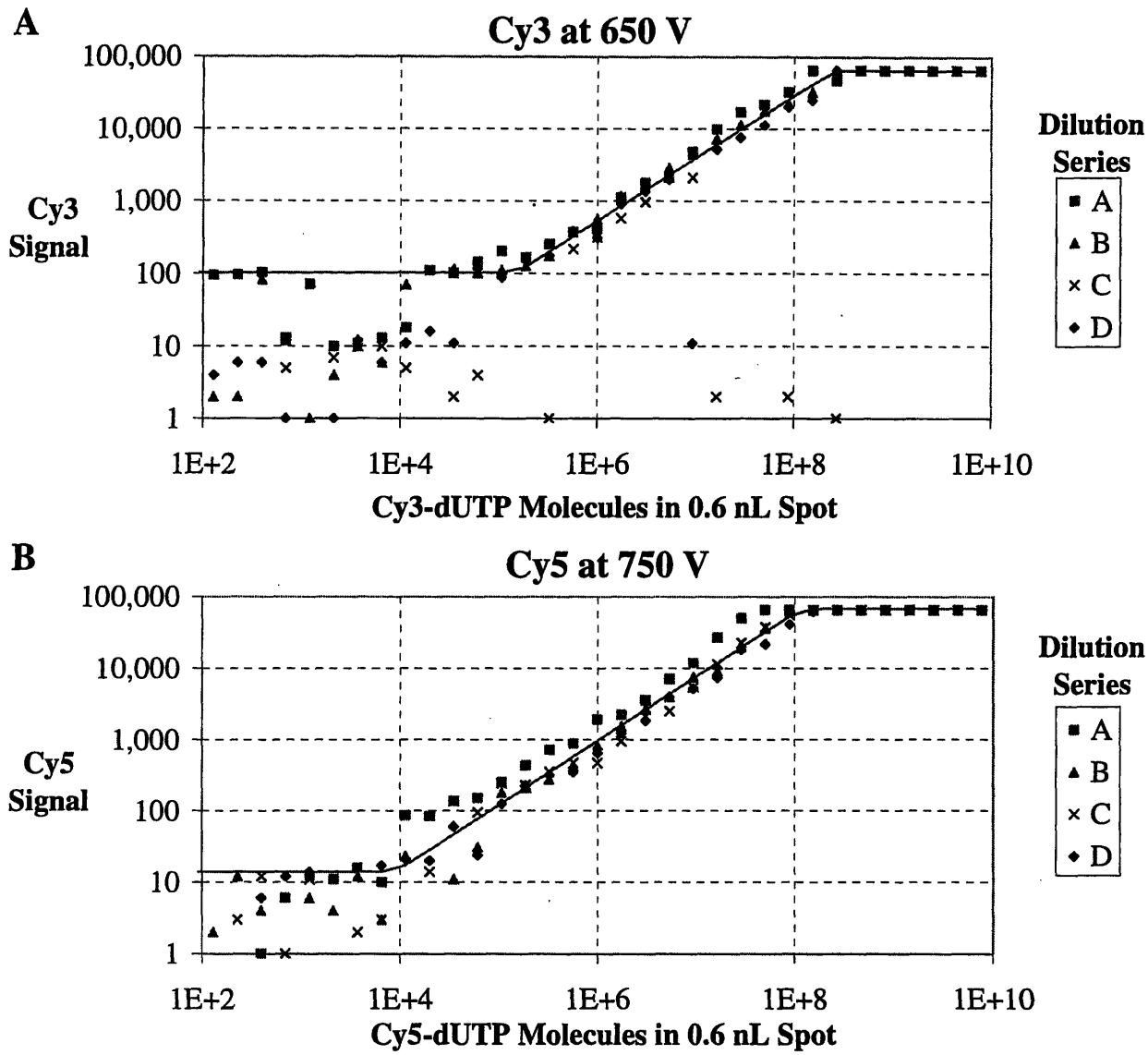


Figure 4.9: Signals from Cy3-dUTP and Cy5-dUTP Serial Dilutions
Slides were arrayed with the labels Cy3-dUTP and Cy5-dUTP in four series of serial dilutions (A-D) ranging from 0.1 fM to 21 μM. The arrays were scanned at different PMT voltages. Shown here are **A)** Cy3 signals from the scan at 650 V and **B)** Cy5 signals from the scan at 750 V.

Histograms of pixel intensities have a low-signal peak, corresponding to pixels in the background region. Most investigators comparing two total RNA samples will typically set the PMT voltages such that each channel has the same peak intensity. However, with a genomic DNA standard in the Cy5 channel, it is not reasonable to expect the background intensities to be equivalent. As mentioned previously, genomic DNA produced very low background signal. For the work presented here, PMT voltages were selected within the ranges above, such that the peak

intensity on the 532-nm (Cy3) histogram was roughly 10-15× that on the 635-nm (Cy5) histogram. This produced images with quantifiable spots over a wide range of intensities.

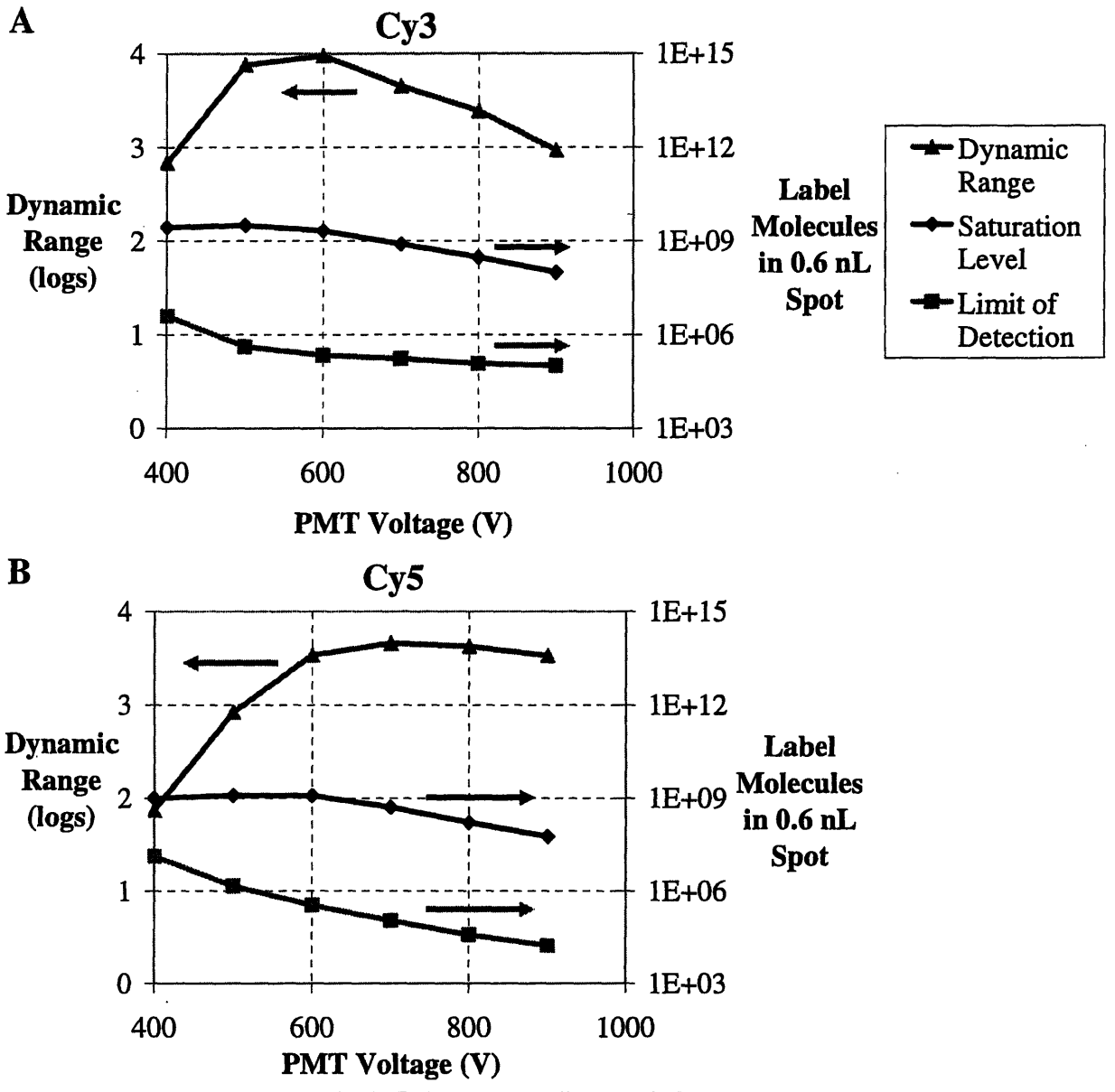


Figure 4.10: Effect of PMT Voltage on Dynamic Range
Serial dilutions of Cy3-dUTP and Cy5-dUTP fluorescent dyes at known concentrations were directly spotted onto glass slides. These slides were scanned at various PMT voltages. For every scan, the saturation level, limit of detection, and dynamic range were calculated. To maximize the dynamic range, the scanner should be operated at 500-700 V for Cy3 and 600-800 V for Cy5.

As a side note, one interesting observation was made regarding the Cy3 and Cy5 fluorescent dyes. Cy5 was found to produce signal, not only at 635 nm—its wavelength of

maximum emission—but also at 532 nm, the wavelength at which Cy3 was scanned. Similarly, Cy3 also yields signal at 635 nm. This observation did not result from cross-contamination of dyes, since the plate was prepared such that each pin printed only one of the dyes. These effects were slight; the 532/635 signal ratios were on the order of 100 for spots containing Cy3 and 0.01 for spots containing Cy5, and would probably not be observed over the background intensity from a hybridization.

4.4 Saturation of Probe DNA during Hybridization

The basic assumption of any hybridization technique is that the amount of probe is much larger than the amount of the sample being quantified. Without enough probe, the spot may become saturated by the sample, causing no signal increase, no matter much sample were added. Because of the small amount of DNA in the original twelve plates from the Lovett Lab, saturation of probe DNA during hybridization is a concern. To determine whether probes were saturated, the technique was carried out with increasing amounts of sample. If the signals were linearly related to the amounts of sample, then no saturation occurred in the range studied.

To confirm that saturation was not occurring with the full-genome *E. coli* microarrays used in this work, five hybridizations were performed. Two large labeling reactions (total RNA labeled with Cy3 and genomic DNA labeled with Cy5) were carried out, as described in the SuperScript procedure in Section 3.7.7.2 and the standard procedure in Section 3.7.8, except at five times the normal volume. A batch of Cy3-labeled cDNA was prepared from 125 μg total RNA, and a batch of Cy5-labeled DNA was prepared from 5 μg genomic DNA. These reaction mixtures were split into five equal-volume reactions just before addition of label and enzyme, both of which contain solvents that would have made equal distribution difficult after their addition. Following addition of the respective labels and enzymes, the reactions were incubated and purified as normal. After the QIAquick™ purification step, the five labeled samples in each set were combined and absorbance measurements were taken. Rather than combine the Cy3 and Cy5 samples, as was usually done at this point, each sample was precipitated in parallel, and each pellet was resuspended in 1× Hybridization Buffer. Table 4.1 shows how these two labeled samples were distributed among the five microarrays. The Cy5-labeled sample was split equally among the five arrays, but the Cy3-labeled sample was split in varying amounts. The Cy3 hybridization solution and the Cy5 hybridization solution were mixed in varying ratios, with

addition of 1× Hybridization Buffer as necessary, to produce five 20-μL hybridization solutions with the distribution described in Table 4.1.

Table 4.1: Distribution of Cy3 and Cy5 Batches among Five Microarrays
A large volume of Cy3-labeled cDNA and Cy5-labeled genomic DNA was generated and hybridized to five microarrays

Array Name	MID1	LOW	MID2	HIGH	MID3
Slide Number	A11	A61	A62	A64	A117
Percentage of Cy5-Labeled Sample	20%	20%	20%	20%	20%
Percentage of Cy3-Labeled Sample	20%	10%	20%	30%	20%

The three MID arrays were hybridized with equal sample amounts in an effort to estimate the reproducibility of the microarray technique. These three arrays were purposely taken from the beginning, middle, and end of the 128-array batch, in an effort to account for the effects of array number (*i.e.* print order) on signal variation.

Signal ratios from all five hybridizations are presented in scatterplots in Figure 4.11. These signal ratios have been filtered and log transformed as described in Section 3.7.14 but have not been normalized, *i.e.* no correction has been performed to account for experimental differences between arrays. Using unnormalized data to examine differences between arrays is a crude method of comparison. Nevertheless, this is the best method for observing signal intensity as a function of sample quantity. The three MID arrays will allow the array-to-array variance to be quantified and accounted for. The differences between the LOW, MID, and HIGH arrays are subtle. When compared against one another, the three MID arrays should ideally produce signal ratios that fall on the diagonal, as in Figure 4.11A & Figure 4.11C. In comparison to the MID arrays, data from the LOW array should fall below the diagonal, as in Figure 4.11B. Inversely, data from the HIGH array should fall above the diagonal, as in Figure 4.11D. Although it may be difficult to see on the scatterplots, the trends in the data agree with these theoretical predictions.

A more descriptive scatterplot is shown in Figure 4.12. Here, the average signal ratios from the three MID arrays are compared against the signal ratios from the HIGH and LOW arrays. This scatterplot shows, as expected, that signal ratios were consistently higher in the HIGH array and consistently lower in the LOW array.

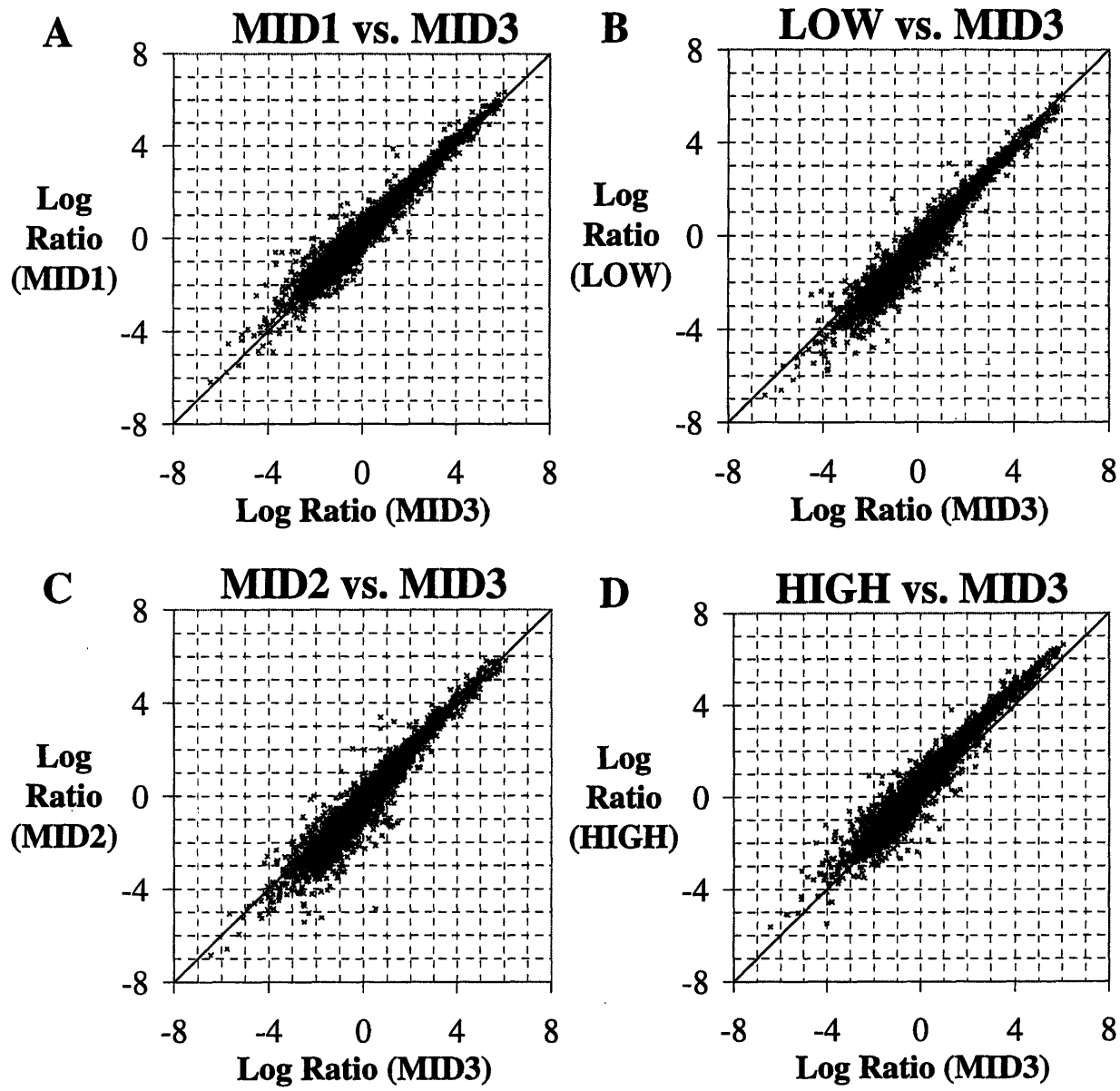


Figure 4.11: Scatterplots for Arrays with Varying Sample Volumes – MID3 vs. All Others

Large amounts of labeled cDNA and genomic DNA were generated and were placed on five arrays as described in Table 4.1. Signal ratios (Cy3/Cy5) were log transformed for plotting. In these scatterplots, the MID3 array is compared against the other four arrays. The MID arrays (**A & B**) produce signal ratios that fall on the diagonal. On average, data for the LOW array (**C**) fall below the diagonal, while data for the HIGH array (**D**) fall above the diagonal.

Using this data set, two approaches were taken to determine whether the probe DNA was saturated. One approach was to examine the difference in signal ratios between each pair of arrays and determine whether they were proportional to differences in the amount of sample

added to the array. A second approach was to perform a linear regression for each gene, across all arrays. If these hybridizations were performed in the linear detection range, then the y-intercept of this regression should be zero, indicating that signal is proportional to sample quantity. Both of these approaches indicated that signals were not saturated.

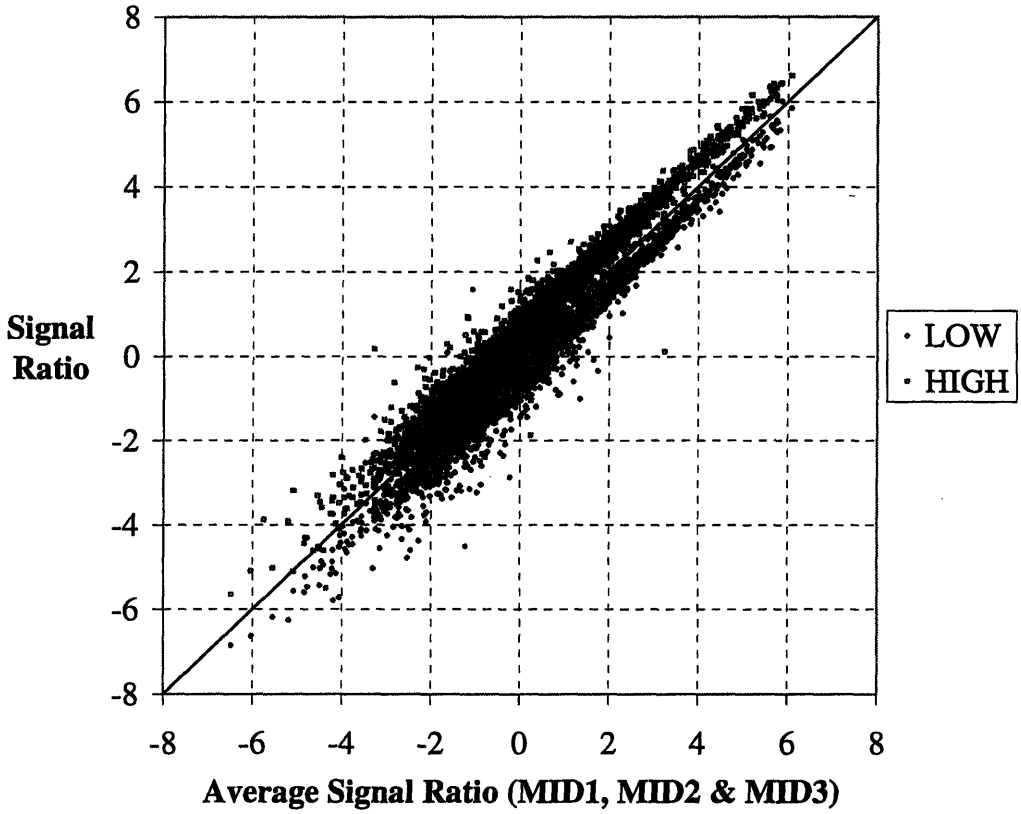


Figure 4.12: Scatterplot for Arrays with Varying Sample Volumes – MID vs. HIGH & LOW
Large amounts of labeled cDNA and genomic DNA were generated and were placed on five arrays as described in Table 4.1. Signal ratio (Cy3/Cy5) values were log transformed for plotting. The average signal ratios from the three MID arrays is compared to signal rations from HIGH and LOW arrays.

4.4.1 Comparison of Signal Differences

The MID arrays were hybridized with twice as much Cy3 sample as the LOW array. If signal were proportional to sample amount, then the three identical arrays should have twice the signal as array A61. On a base-2 log scale, this would correspond to a difference of 1, and the LOW array data in Figure 4.12 should be 1 unit below the diagonal, on average. A similar comparison with the HIGH array shows that it was hybridized with 1.5× the volume of the MID arrays. Data for this array in Figure 4.12 should be 0.58 units above the diagonal, on average.

To determine whether the data matched these predictions, the signal differences between arrays were calculated. For every pair of arrays in the data set, signal differences were calculated and averaged across all genes. These offset values are plotted in Figure 4.13. As expected, the differences between in the three MID arrays did not significantly differ from zero. The LOW vs. MID comparisons showed offsets that were lower than expected; however, the HIGH vs. MID comparisons were consistent with the expected difference.

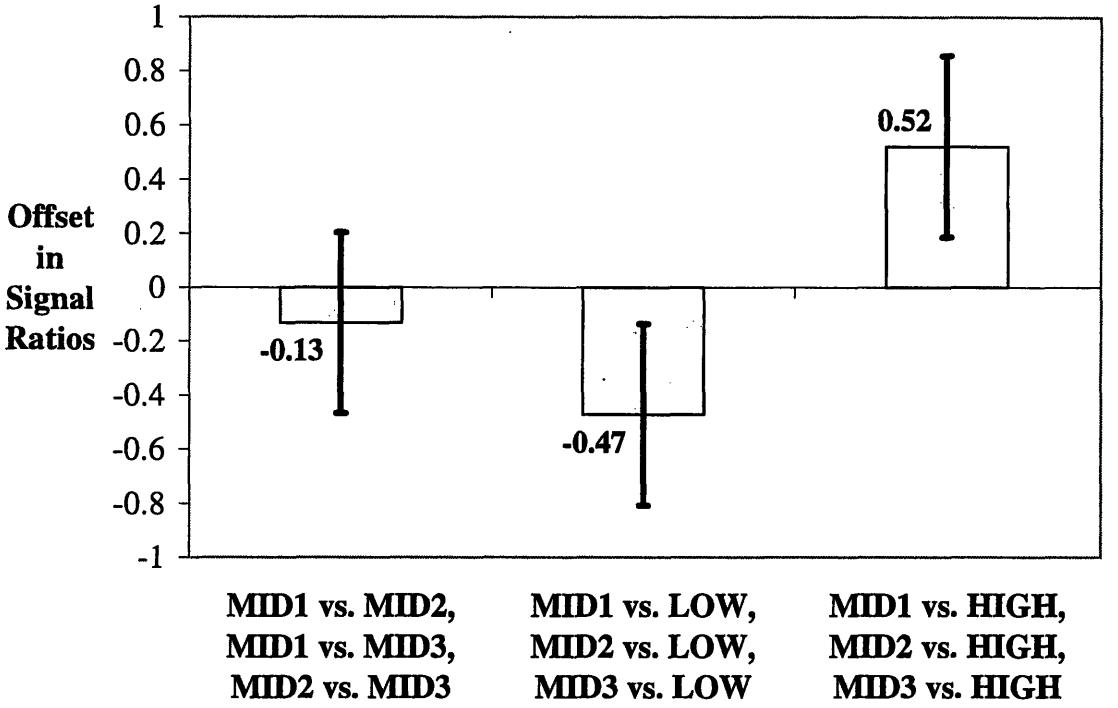


Figure 4.13: Average Offset in Signal Ratios
For every pair of arrays in the data set, signal differences were calculated and averaged across all genes. Each bar represents the average offset from the three comparisons. The error bars represent the standard deviation between the offsets from the three MID comparisons. These error values are assumed to apply to the other two LOW vs. MID and HIGH vs. MID sets.

If the inconsistency between the expected and measured offset values were a result of saturation, it would most likely appear in the HIGH vs. MID comparison. The appearance of this inconsistency in the LOW vs. MID comparison suggests that it may be due to experimental error. As mentioned previously, comparison of unnormalized microarray data is not entirely valid. The large error in the offset between identical arrays in (MID vs. MID in Figure 4.13) certainly indicates that array-to-array differences are present. It is certainly possible that experimental factors, such as particularly low background or low-stringent washes, might cause the unnormalized signal ratios from the LOW array to be higher than expected.

4.4.2 Linear Regression of Signal

To show that the signal was linearly related to the amount of sample, correlation coefficients, between the signal ratio (Cy3/Cy5) and the relative amount of sample used, were calculated for each gene. 80% of the spots had correlation coefficients greater than 0.7. Therefore, a large fraction of spots had a strong positive correlation between sample amount and concentration, which would not be the case if spots were saturated.

Linear regression was also performed on these data, as indicated from Figure 4.14. This plot gives examples of linear regression fits using the first ten genes in the data set that show signal. The slope and intercept were determined for each gene. If the probe DNA were not saturated, then the signal would be linearly related to the amount of sample, with y-intercept of zero. Practically, however, this will not always be the case. These intercept values can be further analyzed by scaling them relative to the average signal ratio and plotting them as in Figure 4.15.

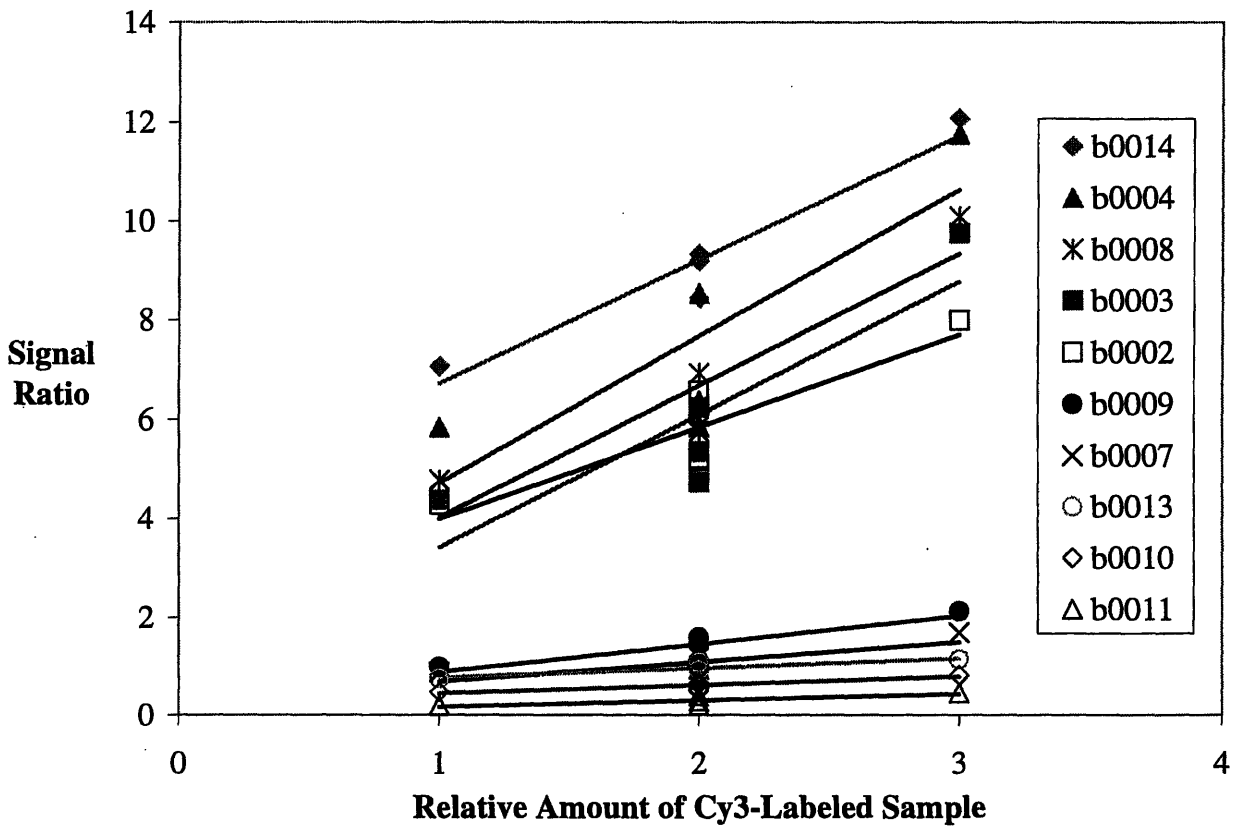


Figure 4.14: Linearity of Signal Ratios with Amount of Cy3 Label
Signal ratios (Cy3/Cy5) are plotted against the relative amount of Cy3-labeled sample added to the arrays. Only the first ten genes with signal are shown.

The scaled intercept values were calculated for every gene in the data set and the average value was found to be 0.314 ± 0.007, at a confidence level of 95%. This means that the intercept value is, on average, 31% of the average signal ratio. Clearly, the intercept values are biased such that they are greater than zero. Although the signal ratio is the statistic of interest for subsequent calculations, the individual signals can also be used to perform the same scaled intercept calculations. The average scaled intercept for the Cy5 signal was found to be 0.954 ± 0.008, which is close to the expected value of 1 (since the same quantity of Cy5 was used on all five arrays). The average scaled intercept for Cy3 was found to be 0.290 ± 0.010. This demonstrates that the bias originates from the Cy3 signal, and is only weakly, if at all, affected by the Cy5 signal.

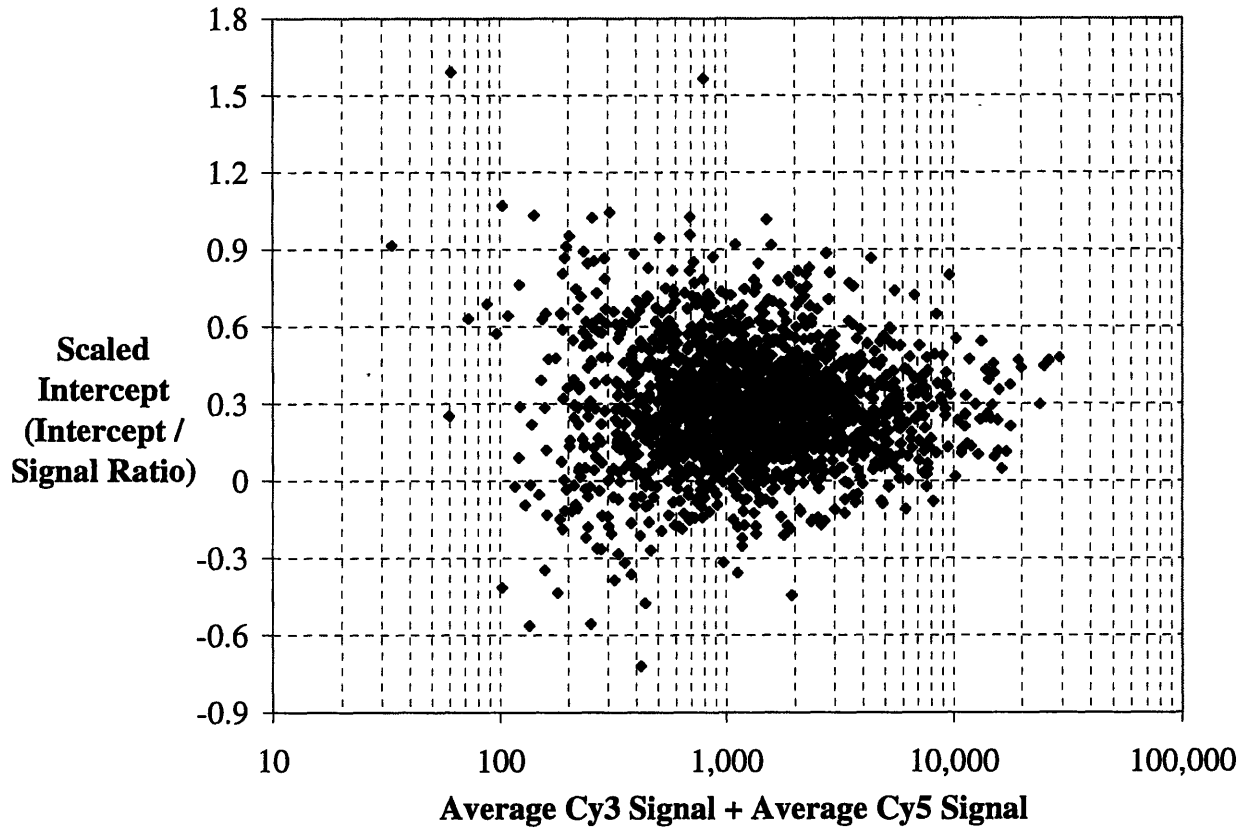


Figure 4.15: Scaled Intercept Values
Intercepts were calculated from linear regressions performed on signal ratios (Cy3/Cy5), like those in Figure 4.14. Intercepts were scaled relative to the average signal ratios and were plotted against the sum of the average signals (Cy3 + Cy5). An intercept value of zero indicates that data were collected in the linear detection range. On average, the intercept value is 31% of the average signal ratio.

The 31% offset in the intercept is of concern, because it demonstrates that the microarray signal is not proportional to the concentration of labeled cDNA. This could be an indication that, while not saturated at these levels, the probes are nearing saturation. However, if this offset were an effect of near-saturation, we would expect the problem to generally become worse as the signal increased. Figure 4.15 reveals just the opposite—these scaled intercepts neither increase nor decrease with signal. A handful of genes showed scaled intercept values near 1, which might indicate that these spots were saturated and did not change signal at all in response to increasing sample amounts. However, these spots occur red most frequently at low signal (below ~1,000) and become more sparse at higher signal values. It is commonly accepted that data with low signal are less reliable than data with high signal. Therefore, scaled intercept values of 1 or larger appeared to be an effect of noise in the data. Examination of Figure 4.15 suggests that probe DNA is not being saturated and is not nearing saturation.

Although saturation does not appear to be a problem, steps were taken to further improve the label incorporation and decrease the yield of the labeling reactions in an effort to reduce the possibility of saturation.

4.4.3 Conclusions on Microarray Sensitivity

Saturation of probe DNA does not appear to be a problem. DNA microarray analysis is performed in the linear detection range of both the DNA probes and the scanner. However, the sensitivity of the microarray analysis is not as strong as it should be. While scatterplots show that lower cDNA amounts leads to lower signal, the effect was not as strong as expected. In addition, regressions on signal data indicated that there is a linear relationship between signal and the quantity of cDNA added to the array. However, this is not a simple proportional relationship. The low sensitivity of the scanner, as measured in Section 4.3 certainly contributes to the low sensitivity of the microarrays. However, the scanner sensitivity is not low enough to completely account for the offset observed here. Another possible explanation is that labeled cDNA nonspecifically bound to the background area of the slide and therefore contributed to the background intensity, rather than the feature intensity. Certainly, this would have resulted in decreased sensitivity of the analysis. Since the background area was the same for each array and is presumably saturated long before the features, the amount of labeled cDNA bound to the background would be constant across all arrays and proportional to the amount of each initial

cDNA sequence. This effect could account for a constant 31% offset in the scaled intercept values. Yet another explanation would be the converse of the above. Labeled cDNA generated from rRNA sequences might nonspecifically bind to the features, thereby artificially increasing the signal producing the same effect.

In an effort to reduce nonspecific binding to the background region, a chemical blocking step was explored; but, it did not produce improved signal-to-noise ratios and was not repeated.

4.5 Microarray Data Analysis

In choosing a data analysis method to apply to our data, we searched for methods that selected differentially expressed genes based on gene-specific expression cutoffs. Not all genes exhibit the same variance in expression; one gene may naturally exhibit four-fold changes in expression, while another may only vary within 10% of the basal level. Therefore, a two-fold change in expression may be significant for the latter gene, but not the former. Microarray methods are sensitive enough to distinguish moderate gene expression changes from large gene expression changes, so analysis of the data from these experiments should enable us to do the same.

Analysis of variance (ANOVA) methods have been widely used and fulfill the criterion above. As an added benefit, it is possible to use ANOVA methods to perform both the normalization and selection of differentially expressed genes. The method described here is very similar to that described in the literature (Kerr *et al.* 2000). The equations and calculations involved in this method are described in detail in Chapter 10. A brief description is given here.

4.5.1 Analysis of Variance (ANOVA) Modeling of Microarray Data

The first step in performing ANOVA was creating the model. Analysis of variance uses data from experiments, which are often well designed and replicated, to determine the effects of different sources of variation. Typically, the sources of variation are blocks (repeated experiments used to observe "normal" variation) and treatments (the experimental conditions being tested). The model includes a mean value of all measurements, several terms that represent deviations from the mean, and a residual error term that accounts for random variation.

For microarray analysis in this work, a two-way ANOVA model (Model 1) was applied

$$y_{agr} = \mu + A_a + G_g + AG_{ag} + \varepsilon'_{agr} \quad \textbf{(4.1)}$$

The variables used throughout this section are described in Table 4.4. In **(4.1)**, y_{agr} represents the signal ratio (on a base-2 log scale) from a particular array (a), gene (g) and replicate spot (r), μ represents the mean value of all signal ratios in the data set, A_a represents array effects, G_g represents gene effects, AG_{ag} represents array-gene interaction effects, and ε'_{agr} represents the residual error for Model 1.

In order to fulfill the assumptions of the ANOVA method, the original data set must meet a few simple requirements.

- The data set must be normally distributed.
- The variation across arrays and genes must be constant throughout the data set.

The signal ratio was chosen as the measurement to analyze because these values are very close to being normally distributed.

Model 1 **(4.1)** has four sources of variation:

- Array effects: The overall signal from a particular array can vary widely due to variation in the experimental technique, biological variation in gene expression, and changes in gene expression that are expected based on the experimental design. One example of the latter case is that the fraction of mRNA in the total RNA sample is known to decrease in nitrogen-grown cultures; therefore, arrays with those samples generally show lower Cy3 signal. Experimental factors that might lead to increased variation in signal include the incorporations and yields of the Cy3- and Cy5-labeled samples as well as the overall binding efficiency of both samples to the probes on the slide.
- Gene effects: Each gene is expressed at a different level, and due to varying binding properties of DNA sequences, the signal from a particular spot will depend largely on the gene identity and sequence.
- Array-gene interaction effects: The above effects may combine to produce variation specific to a particular array and gene.
- Residual error 1: If spots for the same gene have been replicated on the arrays, then the residual error will account for the random variation between these replicated spots.

Once these model parameters were determined, the data set was normalized by subtracting the array effects from each signal ratio to obtain normalized signal ratios, $\hat{y}_{agr}$.

$$\hat{y}_{agr} = y_{agr} - A_a \tag{4.2a}$$

Applying this normalization to Model 1 produces

$$\hat{y}_{agr} = \mu + G_g + AG_{ag} + \varepsilon'_{agr} \tag{4.2b}$$

This is considered to be a global normalization since it assumes that treatments will not affect the overall (or average) signal ratio values. This approach assumes that some genes will decrease in expression while others will increase, and overall levels of gene expression will remain the same. As a cautionary note, global normalization should never be applied to a partial genome data set, especially when the set of genes has been selected as having similar regulation or expression responses.

At this point, it should be noted that each sample represents exactly one block and exactly one treatment. In addition, because of the way experiments were performed (total RNA labeled with Cy3 and genomic DNA labeled with Cy5), each array represents only one RNA sample. Therefore, each array represents one block and one treatment. This relationship can be used to further analyze the AG_{ag} parameters with a second two-way ANOVA model (Model 2).

$$AG_{btgn} = BG_{bg} + TG_{tg} + \varepsilon''_{btgn} \tag{4.3}$$

In this model, BG_{bg} represents block-gene interaction effects, TG_{tg} represents treatment-gene interaction effects, and ε''_{btgn} represents the residual error for Model 2. **(4.3)** distributes the array-gene variation between three additional sources of variation:

- Block-gene interactions: Repeated experiments, or blocks, are intended to be highly reproducible. However, variation will always exist. These repeated experiments are performed in an effort to quantify this "normal" variation so that abnormal changes can be identified. Experimental factors, such as variations in label incorporation, blocking, hybridization, washing, and scanning, may all contribute to overall block-specific variation. However, these experimental factors would typically affect all genes (or spots) in a consistent manner and therefore would not be accounted for in this term. Variation that is specific to both the block and gene is usually attributed to biological variation.

- Treatment-gene interactions: This is the variation we are interested in studying. We would like to identify genes that change expression in response to a particular treatment, in a consistent manner across all blocks. For differentially expressed genes, the magnitude of the TG_{tg} term will be large.
- Residual error 2: This error term accounted for random variation that could not be accounted for by the other terms in this model.

Because each term in Model 2 depends on the gene, the model can be applied to each gene independently.

A similar analysis can be performed on the A_a terms using a third two-way ANOVA model (Model 3).

$$A_{btn} = B_b + T_t + BT_{bt} + \varepsilon'''_{btn} \tag{4.4}$$

In this model, B_b represents block effects, T_t represents treatment effects, BT_{bt} represents block-treatment interaction effects, and ε'''_{btn} represents the residual error for Model 3. For most experiments performed in this work, Model 3 adds nothing to the remainder of the analysis, because the normalization step has removed all of the A_a or A_{btn} terms **(4.2)**. However, in the event that a single sample is repeated on multiple arrays, the residual error term in this model would be significant and must be accounted for.

Combining all of these two-way ANOVA models, produces the following three-way ANOVA model for signal ratio values.

$$y_{btgrn} = \mu + B_b + T_t + BT_{bt} + G_g + BG_{bg} + TG_{tg} + \varepsilon_{btgrn} \tag{4.5}$$

In the above model, ε_{btgrn} represents the overall residual error and is calculated by making the following substitution.

$$\varepsilon'_{btgrn} + \varepsilon''_{btgn} + \varepsilon'''_{btn} = \varepsilon_{btgrn} \tag{4.6}$$

The normalized signal ratios take the form

$$\hat{y}_{btgrn} = \mu + G_g + BG_{bg} + TG_{tg} + \varepsilon_{btgrn} \tag{4.7}$$

For purposes of observing treatment effects, the effects of random variation and repeated measurements are eliminated to produce treatment-averaged normalized signal ratios, which are referred to in this work as expression values, $\bar{\hat{y}}_{\bullet tg\bullet\bullet}$.

$$\bar{\hat{y}}_{\bullet tg\bullet\bullet} = \mu + G_g + TG_{tg} \tag{4.8}$$

These are the values that are plotted throughout this thesis. The error bars in those plots are calculated as follows to account for variation in blocks, replicate spots, and repeated sample analysis, where ρ_{btgn} is the number of replicate spots.

$$\frac{\sum_{brn} \left| \varepsilon_{btgrn} \right|}{\sum_{bn} \rho_{btgn}} \tag{4.9}$$

This section has described how a series of two-way ANOVA models can be used to normalize microarray data and estimate treatment effects. Section 4.5.2 will describe how each of these effects was determined.

4.5.2 Balanced vs. Unbalanced Data Sets

The exact equations used to calculate the parameters in the above models are given in Chapter 10. These calculations are easily done for a balanced data set, *i.e.* a complete data set with data for every possible combination of factors. Unfortunately, DNA microarray experiments are imperfect and will not generate a balanced data set. Several factors that can lead to an unbalanced data set are listed below.

- Spots with low signal are removed from the analysis. Some investigators choose to artificially set these values to zero in order to keep the data set balanced. However, zero signal implies zero expression, and although we cannot measure the expression, we cannot say that it is zero. In addition, signals of zero cannot be transformed to the log scale.
- Not all spots are replicated. In order to save space on the slides and save the DNA material, spots were sparingly replicated in this work. Most genes have only one data point per array, while some others have two. This also makes the data set unbalanced.

- Some samples may not be analyzed. Experimental design should always be carefully considered, because imperfect designs can lead to unbalanced data sets. However, experimental factors may prevent samples from being analyzed, *e.g.* hybridizations may fail and unstable RNA samples may become degraded before analysis can be performed.

Although a balanced analysis may be applied to an unbalanced data set, it will only be an approximation. In order to account for these imperfections, a more complicated unbalanced ANOVA method was applied and was found to produce better results.

4.5.3 Selection of Differentially Expressed Genes

The method for selecting differentially expressed genes was a compromise between using gene-specific cutoffs and a global cutoff. Neither method alone was found to capture all of the genes with interesting expression changes. Genes with large expression changes but high variance (typically due to low expression under one of the conditions) were not captured using gene-specific cutoffs. Inversely, a global cutoff tends to neglect genes that show slight, but highly reproducible, expression differences.

The ANOVA model described in Section 4.5.1 quantifies the variation in gene expression between different treatments (TG_{tg} terms) as well as the random (residual) variation. Comparing these two variances reveals whether the observed expression differences are random noise or represent a unique effect. Differentially expressed genes were identified by performing a multiple comparison on every pair of TG_{tg} values for every gene in the data set. The number of comparisons per gene is represented by the expression $\frac{\tau!}{2!(\tau-2)!}$ where τ is the number of treatments in the data set.

The first step in this procedure is to calculate log ratios (LR_{ijg}) for each pair of treatmtents i and j (with $i < j$).

$$LR_{ijg} = TG_{ig} - TG_{jg} \tag{4.10}$$

The magnitude of this log ratio is a measure of the observed differential expression.

Next, parameters were calculated to quantify random variation as well as global variation. The gene-specific residual degrees of freedom (DOF_{Rg}) was calculated as the sum of the degrees of freedom from Model 1 (DOF_{R1g}) and Model 2 (DOF_{R2g}).

$$DOF_{Rg} = DOF_{R1g} + DOF_{R2g} \tag{4.11a}$$

$$DOF_{Rg} = \left(\sum_{a} \rho_{ag} - \sum_{a} \zeta_{ag}\right) + \left(\sum_{btn} \zeta_{btgn} - \beta_g - \tau_g + 1\right) \tag{4.11b}$$

$$DOF_{Rg} = \sum_{btn} \rho_{btgn} - \beta_g - \tau_g + 1 \tag{4.11c}$$

In **(4.11b)**, the variable ζ_{ag} or ζ_{btgn} represents the number of data collected for a particular array and gene. The global residual degrees of freedom (DOF_R) was similarly calculated as the sum of the degrees of freedom from Model 1 (DOF_{R1}) and Model 2 (DOF_{R2}).

$$DOF_R = DOF_{R1} + DOF_{R2} \tag{4.12a}$$

$$DOF_R = \left(\sum_{ag} \rho_{ag} - \sum_{ag} \zeta_{ag}\right) + \left(\sum_{btgn} \zeta_{btgn} + \gamma_{Real} - \alpha + \beta + \tau - \sum_{g} \beta_g - \sum_{g} \tau_g - 1\right) \tag{4.12b}$$

$$DOF_R = \sum_{btgn} \rho_{btgn} + \gamma_{Real} - \alpha + \beta + \tau - \sum_{g} \beta_g - \sum_{g} \tau_g - 1 \tag{4.12c}$$

These equations use the variable γ_{Real}, which is the number of genes for which data were collected. The sums of squares were also calculated for both the gene-specific (SS_{Rg}) and global (SS_R) cases. For the gene-specific case, the formula was

$$SS_{Rg} = \sum_{btrn} \left(\varepsilon'_{btgrn} + \varepsilon''_{btgn}\right)^2 \tag{4.13}$$

For the global case, the formula was

$$SS_R = \sum_{btgrn} \left(\varepsilon'_{btgrn} + \varepsilon''_{btgn}\right)^2 \tag{4.14}$$

For the purposes of comparison, these sums-of-squares values were scaled based on the degrees of freedom. If $DOF_{Rg} > 0$, the mean square (MS_{Rg}) is calculated as the ratio of the two previous statistics.

$$MS_{Rg} = \frac{SS_{Rg}}{DOF_{Rg}} \tag{4.15}$$

The same calculation is performed to calculate the global mean square (MS_R).

$$MS_R = \frac{SS_R}{DOF_R} \tag{4.16}$$

These mean square values were used to perform statistical tests for identifying differentially expressed genes.

The next step in the multiple comparison was two *t*-tests on LR_{ijg} values. One test used the gene-specific variance (MS_{Rg}), while the other used the global variance (MS_R). The hypotheses for these tests were as follows.

$$\begin{aligned} H_0 &: LR_{ijg} = 0 \\ H_1 &: LR_{ijg} \neq 0 \end{aligned} \tag{4.17}$$

To evaluate these hypotheses, the probability of H_1 was calculated for each case. For the gene-specific case, this probability was $P_{ijg}^{Gene-Specific} = P\left[t(DOF_{Rg}) \leq \frac{|LR_{ijg}|}{MS_{Rg}\left(\frac{1}{\sum_{bn} \rho_{bign}} + \frac{1}{\sum_{bn} \rho_{bjgn}} \right)} \right]$. If this probability was greater than some defined cutoff ($P_{Cutoff}^{Gene-Specific} = 0.95$ for most cases in this work), then the null hypothesis was rejected. Thus, the differential expression between treatments *i* and *j* would be significant based on this test.

In a similar global test, the probability of H_1 was calculated as $P_{ijg}^{Global} = P\left[t(DOF_R) \leq \frac{|LR_{ijg}|}{MS_R\left(\frac{1}{\sum_{bn} \rho_{bign}} + \frac{1}{\sum_{bn} \rho_{bjgn}} \right)} \right]$. One major difference between this test and

the gene-specific test is that different probability cutoffs are required. Applying a 95% significance level to the gene-specific test means that each gene has a probability of 5% or lower of being selected as differentially expressed, *i.e.* for each gene there is a 5% probability of a false positive. For the sake of consistency, we would like to apply a 95% significance level to the global test as well, thereby forcing a 5% probability of a false positive on the entire data set. If a 95% probability cutoff were applied to the roughly 4,000 comparisons in this global test, the probability of every comparison being correct would be $(0.95)^{4,000} = 7.84 \times 10^{-90}$. An error would be virtually guaranteed! Therefore, this cutoff value must be increased in order to maintain 95% confidence in the global system. The Bonferroni correction is a simple method for accounting for this change in scale. For γ genes, the new cutoff is estimated as follows:

$$P_{Cutoff}^{Global} = 1 - \frac{1 - P_{Cutoff}^{Gene-Specific}}{\gamma} \tag{4.18}$$

For the case of 4,000 genes ($\gamma = 4{,}000$), $P_{Cutoff}^{Global} = 0.9999875$. However, this is only an estimate; the cutoff value is calculated for each data set based on the exact number of genes.

To this point, each gene has been given a probability of selection from the gene-specific test and a probability of selection from the global test. Based on these probabilities and the defined cutoffs for each test, it is possible to generate two lists of differentially expressed genes based on the results from each test. Rather than simply selecting genes that appear on both lists as differentially expressed, the results of the two tests are combined to determine the probability that each gene would not have been incorrectly selected by both tests.

$$P_{ijg}^{Combined} = 1 - \left(1 - P_{ijg}^{Gene-Specific}\right) \cdot \left(1 - P_{ijg}^{Global}\right) \tag{4.19}$$

Cutoff values from the two tests were similarly combined

$$P_{Cutoff}^{Combined} = 1 - \left(1 - P_{Cutoff}^{Gene-Specific}\right) \cdot \left(1 - P_{Cutoff}^{Global}\right) \tag{4.20}$$

For the example above with $P_{Cutoff}^{Gene-Specific} = 0.95$ and $\gamma = 4{,}000$, the combined cutoff is calculated to be $P_{Cutoff}^{Combined} = 0.999999375$. Genes were selected as differentially expressed when $P_{ijg}^{Combined} > P_{Cutoff}^{Combined}$.

It should be noted that when more than two treatments are compared, the same tests are performed for each two-treatment comparison. In this case, genes are selected as differentially expressed when at least one treatment comparison passes the above test. For the case of multiple treatments, others have reported using *f*-tests, which would use the variance of all TG_{tg} values, rather than the difference between each pair of TG_{tg} values (Cui and Churchill 2003). While the multiple comparison *t*-tests used in this work indicate which pair(s) of treatments show differential expression, *f*-tests do not and have been avoided for this reason.

4.5.4 Grouping Genes into Functional Categories

The next level of data analysis is to observe trends not at the level of individual genes, but at the level of gene groups. Ultimately, we would like to extrapolate what is observed at the level of transcripts to draw conclusions about metabolism, energy consumption, and stress responses within in the cell. In order to draw conclusions at a higher level, it is necessary to connect individual genes into categories. Genes can be categorized both by *a priori* knowledge and by empirical evidence.

4.5.4.1 EcoCyc Database

The function and regulation of *E. coli* genes and gene products, as defined by volumes of *E. coli* literature, is collected by EcoCyc (Karp *et al.* 2002), an online ***E. coli*** **encyclopedia.** Using the information from this database, four categories were identified, in which a gene might make meaningful connections with other genes:

- **Protein Complexes:** Although microarray data give no information about the levels of various proteins inside the cell, increased gene expression, at the very least, indicates an attempt to increase the level of the protein product. Therefore, classifying genes according to complexes formed by their products may identify meaningful trends. EcoCyc contains data for 178 protein complexes that are composed of products from at least two different genes.
- **Pathways:** The metabolic pathways inside the cell are well understood. Genes involved in these pathways are often regulated in similar ways and can provide more insight into the metabolism changes occurring inside the cell. EcoCyc contains data for 196 pathways that use products from at least two different genes.

- **Transcription Units:** Bacterial genes are often transcribed in groups. Genes from the same transcription unit that show similar expression provide further validation of the analytical technique. EcoCyc contains data for 338 transcription units involving at least two different genes.
- **Regulons:** Transcription units can also be grouped according to their regulation. Regulators within the cell such as Crp, ArcA, OxyR, and RNA polymerase σ factors are strong indicators of the cell's response to its environment. EcoCyc contains data for 98 regulons involving at least two transcription units. One of these regulons, the exponential phase σ^{70} RNA polymerase, was intentionally omitted. With 725 genes (17% of the *E. coli* genome) listed in this regulon, the probability of pairing two genes is so high that these groups are not considered to be significant.

4.5.4.2 Hierarchical Clustering

Hierarchical clustering was made possible by the Cluster and TreeView software packages (Eisen *et al.* 1998). Clustering was performed on genes only (not arrays) by calculating centered correlation coefficients. The average-linkage clustering method was also used; this method combines clusters (or genes) based on the correlation between the clusters' average profiles.

The data set for clustering was generated by eliminating genes with too few data, so that any pair of genes would have at least two common measurements. This guaranteed that a valid correlation coefficient would be calculated for any pair of genes. To meet this requirement, a maximum number of missing data was defined by rounding the quantity $(n\text{-}2)/2$ down to the nearest integer, where n was the number of treatments in the data set. Thus, for a thirteen-gene data set, genes with more than five missing measurements were removed. The data set was also scaled such that the signal ratio of the zero-time-point measurement was zero (when there was no zero-time-point measurement, the average signal ratio value was set to zero). Since scaling does not affect the calculation of correlation coefficients, this step may seem unnecessary. However, for average-linkage clustering, scaling can affect the average cluster profiles when data are missing. In turn, this can affect the correlation coefficients between this average profile and individual genes.

Table 4.2: Variables Used in Description of Microarray Data Analysis
(continued on next page)

Variable	Description
a	Array index (equivalent to the combined index btn)
$\dot{a}$	Index for arrays not involved in internal summations (used only in unbalanced case)
A_a or A_{btn}	Array effects
$\overline{AG}_{\bullet g\bullet}, \overline{AG}_{b\bullet g\bullet}$	AG_{ag} values averaged over missing ($\bullet$) indices
AG_{ag} or AG_{btgn}	Array-gene interaction effects
α	Number of arrays in the data set
b	Block index. Blocks are repeated experiments.
B_b	Block effects. Blocks are repeated experiments.
BG_{bg}	Block-gene interaction effects
BT_{bt}	Block-treatment interaction effects. Sample effects.
β	Number of blocks in the data set
β_g	Number of blocks with valid data for a particular gene
DOF_{R1g}	Residual degrees of freedom for gene g - Model 1 only (ϵ'_{agr})
DOF_{R2g}	Residual degrees of freedom for gene g - Model 2 only (ϵ''_{btgn})
DOF_{Rg}	Residual degrees of freedom for gene g - Final Model (ϵ_{btgrn})
DOF_{R1}	Overall residual degrees of freedom - Model 1 only (ϵ'_{agr})
DOF_{R2}	Overall residual degrees of freedom - Model 2 only (ϵ''_{btgn})
DOF_R	Overall residual degrees of freedom - Final Model (ϵ_{btgrn})
ε'_{agr} or ε'_{btgrn}	Model 1 residual error
ε''_{ag} or ε''_{btgn}	Model 2 residual error
ε'''_{a} or ε'''_{btn}	Model 3 residual error
ε_{agr} or ε_{btgrn}	final error, residual error in final three-way ANOVA model
f	symbol for f-tests
g	Gene index
G_g	Gene effects
γ	Number of genes in the data set
γ_{Real}	Number of genes in the data set for which $\rho_{ag} > 0$
i	treatment index for treatment comparisons, $i < j$
j	treatment index for treatment comparisons, $i < j$
$J_{\dot{a}a}$	Expression used during matrix calculation for unbalanced data sets
$K_{\dot{a}}$	Expression used during matrix calculation for unbalanced data sets
LR_{ijg}	Log ratios (on a base-2 log scale)
MS_{Rg}	Residual means square for gene g
MS_R	Residual means square
μ	Grand average. Mean of all signal ratio values
μ_g	Gene average. Mean of normalized signal ratios for gene g

Variable	Description
n	Repeated sample index. A single sample can be analyzed mutliple on multiple arrays. This index accounts for the number of times the sample was analyzed.
v	Number of replicate analyses for a particular sample
r	Replicate spot index. Some genes are represented by two spots on a single array. This index indicates the number of the replicate spot.
$P_{Cutoff}^{Combined}$	Probability cutoff for combined test
$P_{ijg}^{Combined}$	Probability of significance from combined test
$P_{Cutoff}^{Gene-Specific}$	Probability cutoff for gene-specific test
$P_{ijg}^{Gene-Specific}$	Probability of significance from gene-specific test
P_{Cutoff}^{Global}	Probability cutoff for global test
P_{ijg}^{Global}	Probability of significance from global test
ρ	Number of replicated spots in a balanced data set
ρ_{ag} or ρ_{btgn}	Number of replicated spots with valid data
SSE_a	Sum of squares of Model 1 residual error for a given array
SSE_{ag}	Sum of squares of Model 1 residual error for a given array and gene
SSE_{bg}	Sum of squares of Model 2 residual error for a given block and gene
SSE_g	Sum of squares of Model 1 residual error for a given gene
SSE_μ	Sum of squares of Model 1 residual error
SSE_{tg}	Sum of squares of Model 2 residual error for a given treatment and gene
SS_{Rg}	Residual sum of squares for gene g
SS_R	Residual sum of squares
t	Treatment index. Treatments are the conditions being studied. Also used for t-tests.
T_t	Treatment effects. Treatments are the conditions being studied.
TG_{tg}	Treatment-gene interaction effects
τ	Number of treatments in the data set
τ_g	Number of treatments with valid data for a particular gene
$\bar{y}_{\bullet\bullet\bullet}$, $\bar{y}_{a\bullet\bullet}$, $\bar{y}_{\bullet g\bullet}$, $\bar{y}_{ag\bullet}$	Signal ratios (y_{agr}) averaged over missing ($\bullet$) indices
y_{agr} or y_{btgrn}	Signal ratios (on a base-2 log scale)
$\hat{y}_{agr}$ or $\hat{y}_{btgrn}$	Normalized signal ratios (on a base-2 log scale)
$\bar{\hat{y}}_{\bullet tg\bullet\bullet}$	Expression values OR Treatment-averaged normalized signal ratios (on a base-2 log scale)
ζ_{ag}	A binary matrix relating whether a particular AG_{ag} value is present. (0 if $\rho_{ag} = 0$, 1 if $\rho_{ag} > 0$)

4.5.5 Summary of Microarray Data Analysis

Using ANOVA to perform microarray data analysis fulfills the requirement of providing gene-specific cutoffs for differential expression. The method described above uses the parameters calculated from the ANOVA model, namely the TG_{ig} parameters, to quantify differential expression. Similarly, the residual error, also calculated from the ANOVA model, was used to quantify the normal variation for the gene as a whole. Finally, a *t*-test was performed to evaluate the significance of that differential expression, based on the normal variation in the gene.

4.6 Reproducibility of Microarray Analysis

The microarray data set generated in Section 4.4 served a dual purpose. In addition to allowing saturation of probes to be evaluated, this data set also allowed the reproducibility of microarray hybridizations to be estimated. Estimation of the reproducibility is important to microarray analysis because large experimental variation may falsely appear as a change in gene expression. A 2-fold change in expression, which corresponds to a value of 1 on a log ratio scale, is typically regarded as the standard limit of detection for this technique. Using the five repeated microarrays, we can determine how often a comparison between two spots falsely appears as differential expression.

Using the data analysis procedure described in Section 4.5, the data from all five arrays were normalized. Although all of these arrays were used to analyze the same total RNA and genomic DNA samples, each array was treated as though it were used to analyze a unique sample from a unique treatment. Log ratios (LR_{ijg}) were calculated for every possible comparison between the five slides (roughly ten comparisons per gene). Theoretically, these identical samples should exhibit no differential expression. Practically, however, 2.6% of the log ratio values were found to have magnitude larger than 1, which would indicate a 2-fold change in expression.

These data were further analyzed by calculating standard deviations in the normalized signal ratios for each gene. These values are displayed in the histogram in Figure 4.16. This histogram shows that most of the signal ratios vary by 0.1 - 0.2. In this data set, less than 0.9% of the genes exhibit variation larger than 1. However, to analyze the probability of differential

expression falsely rising above a 2-fold cutoff, we must consider propagation of error from *two* signal ratios. Toward this end, it was found that 2.8% of the genes exhibit variation larger than 0.707 (or $\sqrt{1/2}$). To illustrate the importance of these values, consider two normalized signal ratios that are from the same gene and have measurement error of 0.707:

$$x_1 \pm 0.707 \quad \text{and} \quad x_2 \pm 0.707$$

The log ratio is simply the difference of these two values, and its error would be calculated by propagation of error as follows:

$$LR_{ijg} = (x_1 - x_2) \pm \sqrt{(0.707)^2 + (0.707)^2} = (x_1 - x_2) \pm 1 \qquad \textbf{(4.21)}$$

Based on this analysis, 2.8% of the genes in the data set will exhibit enough measurement variation to falsely appear as being differentially expressed, based on a 2-fold cutoff.

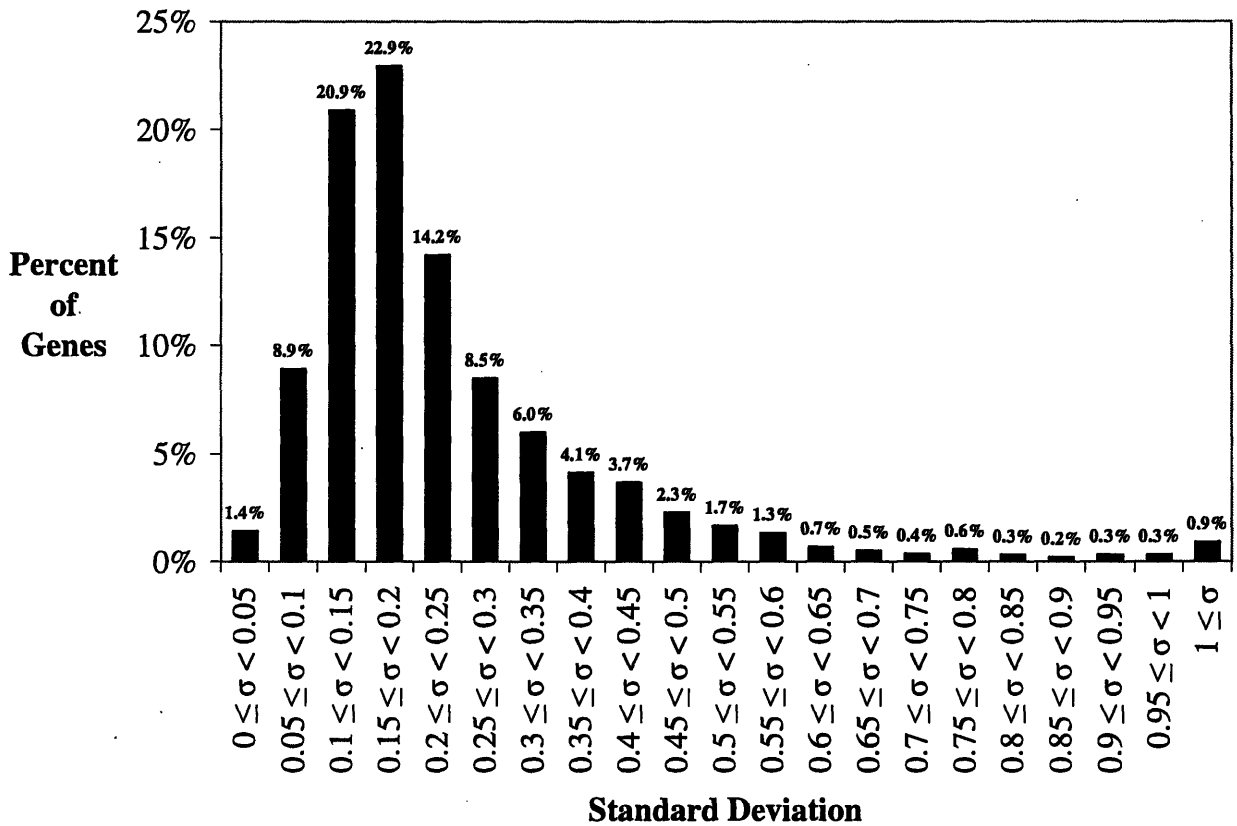


Figure 4.16: Histogram of Gene Standard Deviations for Five Repeated Arrays
Data from the arrays used in Section 4.4 (LOW, MID1, MID2, MID3, and HIGH) were normalized and the standard deviation in the normalized signal ratios was calculated for each gene.

Based on the analysis of these five identical slides, 2.6 – 2.8% of log ratios can be expected to falsely appear as being differentially expressed. Fortunately, this false prediction rate can be reduced by performing replicate experiments and applying the statistical tests for differential expression described in Section 4.5.3.

4.7 Final Validation of Genes Differentially Expressed upon Induction

Thus far, this chapter has presented the development of the experimental methods for performing DNA microarray experiments and statistical methods for analyzing and interpreting these results. The final test of all of these methods was a small experiment involving conditions that have been well studied. This trial run was intended to show that differential expression could be observed and to generate real microarray data that could be used to validate data analysis techniques. The conditions that were chosen for comparison were samples taken immediately before induction and 60 minutes following induction. Induction has been well studied, as described in Section 1.2.1.2 and the expected expression changes are well known. Furthermore, the effects of induction will be present throughout our experiments, but may not necessarily be the conditions intended for study. Therefore, analysis of the effects of induction during the validation phase allowed for correction of expression data in later experiments.

4.7.1 Experimental Details

On four different days, four cultures of *E. coli* BL21 pEAT8-137 were grown in M9 minimal medium at 30°C to an OD_{600} of 0.7, at which point the pre-induction sample was collected for microarray analysis. The cultures were immediately induced by addition of IPTG and were kept under air. A second sample was collected 60 min after induction. The experimental conditions were carefully monitored in order to maintain consistency across the four cultures. As shown in Figure 4.17, the growth curves were highly reproducible. In total, eight samples were collected and all were analyzed with DNA microarrays from Print A using the SuperScript method described in Section 3.7.7.2. Microarray images were analyzed and the resulting data were filtered to remove low-signal spots and control spots.

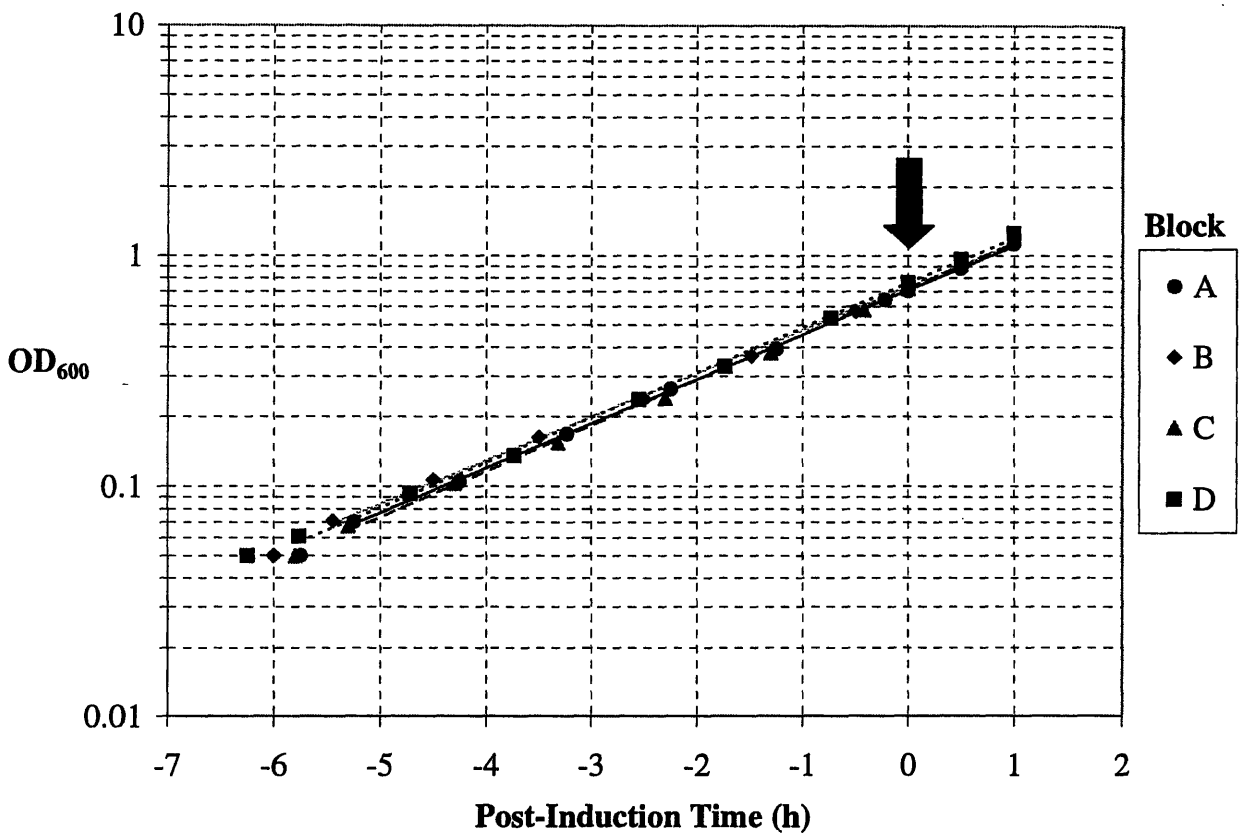


Figure 4.17: Growth Curves from Four Cultures in Validation Experiment
On four different days, four cultures were grown for approximately 6 h and induced at OD_{600} of 0.7 (arrow). Samples were collected for microarray analysis immediately before and 60 min after induction. Growth curves were strongly reproducible for each repeated experiment (Block).

4.7.2 Selection of Differentially Expressed Genes

Prior to normalization and further analysis with the ANOVA models, the assumptions for the ANOVA model were checked. First, the data must be normally distributed. Figure 4.8 shows the original signal ratios for each array and confirms that these data are indeed close to normal. The second criterion is that the variances must be constant across all experimental factors. The overall variance in the data set is 3.98. Figure 4.19 confirms that each array, block, and treatment has variances near this overall value. The variance within each gene is much smaller, since each gene contains much less data (typically 8 data points, 16 at the most).

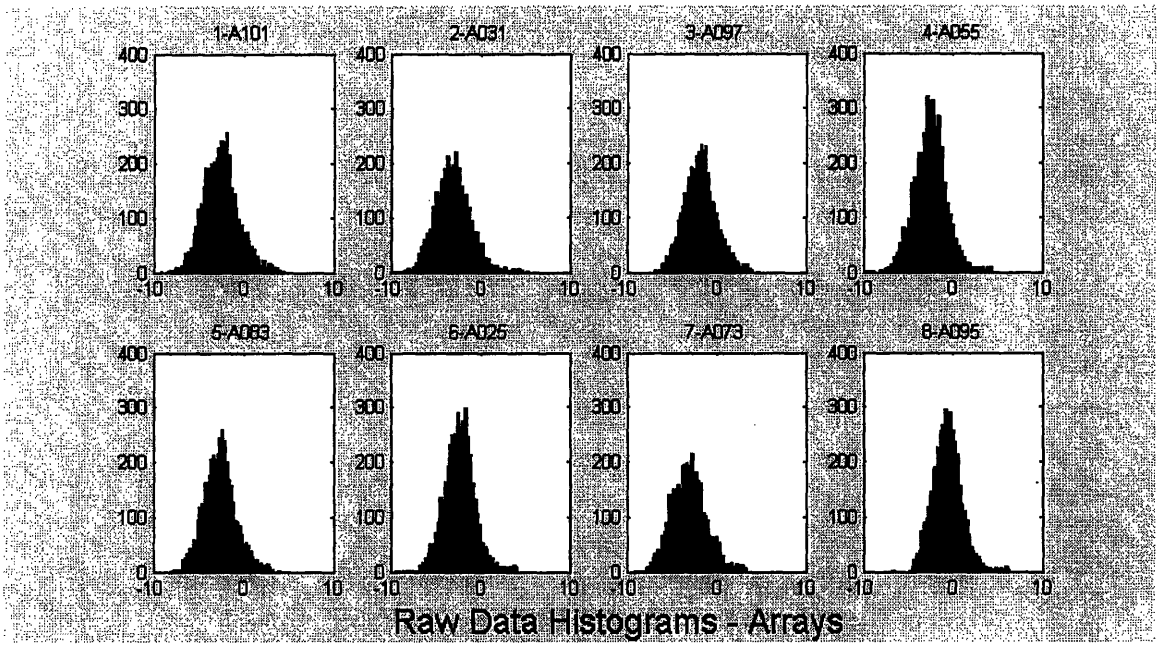


Figure 4.18: Histograms of Signal Ratios
Histograms of signal ratios for each array in the data set. Signal ratio distributions are approximately normal. Since these data are shown before normalization, the peak signal ratios are different for each array.

Analysis of variance was performed as described in Section 4.5 and model parameters were calculated. To confirm that each model parameter was significant, ANOVA Tables were generated. These tables, shown in Table 4.3, display the variances for each parameter and compare them with the residual variance. Note that Model 3 in **(4.4)** was found to be invalid, since none of the samples were repeated on separate arrays. Therefore, the BT_{bt} interaction terms take the place of the ε'''_{btn} residuals. Neither the B_b nor T_t terms were found to be significant at a confidence level of 0.95; therefore, the A_a parameters cannot be further analyzed. The ratio values for all other values are large, indicating that the residual variance is extremely small compared to all other factors. All parameters were found to be significant, which confirms that the ANOVA model is valid.

As a final confirmation that the ANOVA model is appropriate for this data set, the distributions of the residual values were examined. Across all parameters, the distributions were approximately normal with similar variances. Figure 4.20 shows histograms of residual values for all arrays in this data set.

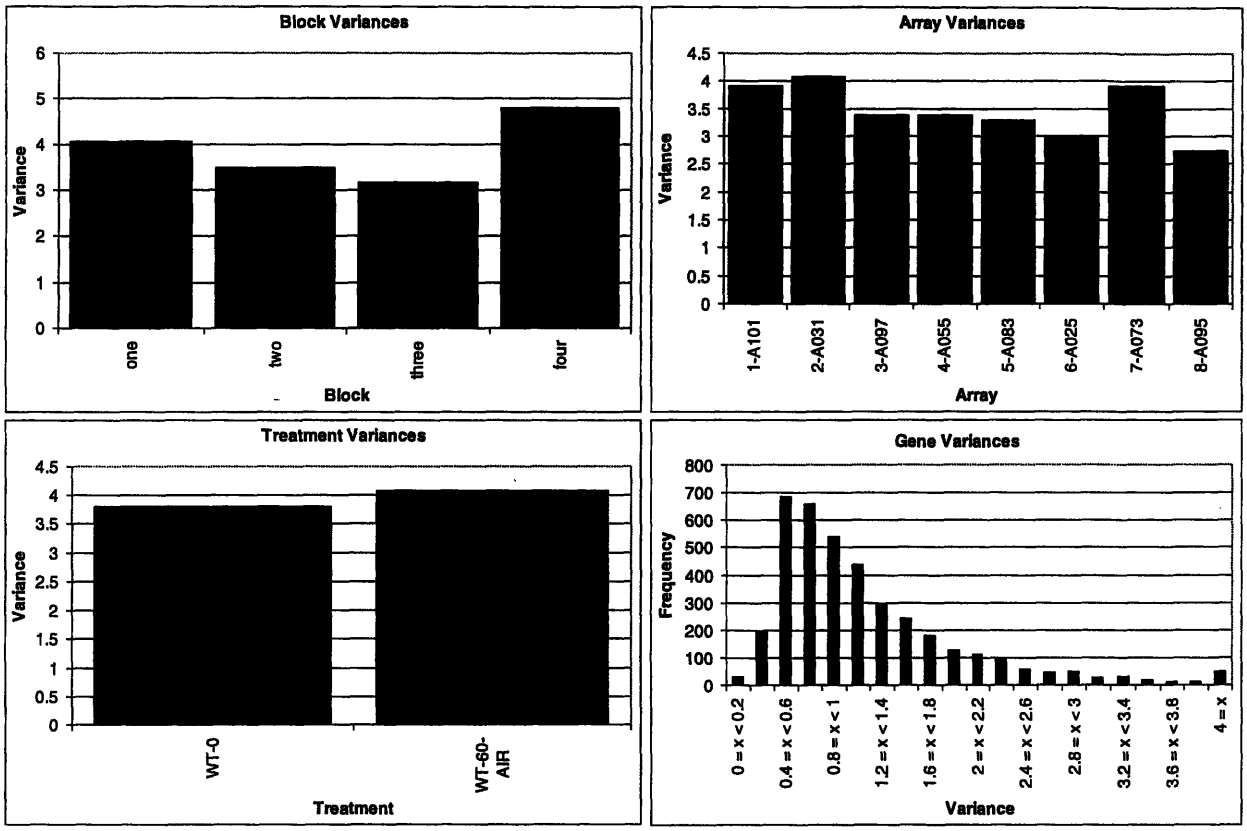


Figure 4.19: Variances in Raw Data Signal Ratios across Each Experimental Factor

Variances for each block, array, and treatment are plotted and are all similar in magnitude. Gene variances are represented in histogram format.

Next, a set of differentially expressed genes was selected as described in Section 4.5.3. Results of this selection process are displayed in various scatter plots in Figure 4.21, Figure 4.22, and Figure 4.23. Log ratios and probabilities of significance are plotted in the volcano plot in Figure 4.21 to demonstrate that neither a high log ratio nor high significance are sufficient for differential expression by this method (Cui and Churchill 2003). Figure 4.22 plots signal ratio data from the two treatments against one another. Differentially expressed genes appear far from the diagonal on this plot; however, this criterion alone is insufficient for selection. Finally, Figure 4.23 uses probabilities of significance from both the gene-specific and global tests to illustrate exactly how differentially expressed genes are selected.

Table 4.3: ANOVA Tables for Microarray Data Analysis
ANOVA Tables are displayed for two two-way ANOVA models and the overall three-way ANOVA model. Probability values of 1 indicate that the probability is greater than 0.99999. All model parameters are significant

Model 1

Source of Variation	Sum Squares	Degrees of Freedom	Mean Squares	Ratio	Probability of Significance
μ	159474	1			
$A_{a\,(btn)}$	19654	7	2808	13786	1
G_g	91927	3853	23.9	117.1	1
$AG_{ag\,(btgn)}$	11010	24645	0.447	2.19	1
$\varepsilon'_{agr\,(btgnr)}$	493	2421	0.204		
$y_{agr\,(btgnr)}$	282558	30927			

Model 2

Source of Variation	Sum Squares	Degrees of Freedom	Mean Squares	Ratio	Probability of Significance
TG_{tg}	4902	3757	1.305	6.48	1
BG_{bg}	4148	11155	0.372	1.85	1
$\varepsilon''_{ag\,(btgn)}$	1960	9733	0.201		
$AG_{ag\,(btgn)}$	11010	24645			

Overall Model

Source of Variation	Sum Squares	Degrees of Freedom	Mean Squares	Ratio	Probability of Significance
μ	159474	1			
$A_{a\,(btn)}$	19654	7	2808	13786	1
G_g	91927	3853	23.9	118.2	1
TG_{tg}	4902	3757	1.305	6.46	1
BG_{bg}	4148	11155	0.372	1.84	1
$\varepsilon_{agr\,(btgnr)}$	2453	12154	0.202		
$y_{agr\,(btgnr)}$	282558	30927			

Of the 3,854 genes in this data set, 384 (10%) were selected as differentially expressed: 156 genes (4%) showed increased expression, while 228 genes (6%) showed decreased expression. These genes are displayed in Table 4.4. Since cultures were induced with IPTG, the *lac* operon should be stimulated. This activation should also increase transcription of the T7 RNA polymerase gene. T7 RNA polymerase should then transcribe the α_1AT gene to produce

the recombinant protein. As expected, three of the top four genes on the list of increased expression are genes directly related to induction. Topping the list with a log ratio of 4.77 (27-fold increase) is the *lacY* gene. The other *lac* genes do not appear on the list because there were few data for them. Signals for these spots were low, particularly in pre-induction samples, where the *lac* operon is repressed. IPTG is also known to stimulate the melibiose operon (Richmond *et al.* 1999; Wei *et al.* 2000). Indeed, the *melB* gene was found to be significant with a log ratio of 1.09 (2.1-fold increase). The genes that were expected to change upon IPTG induction did so and were selected as differentially expressed.

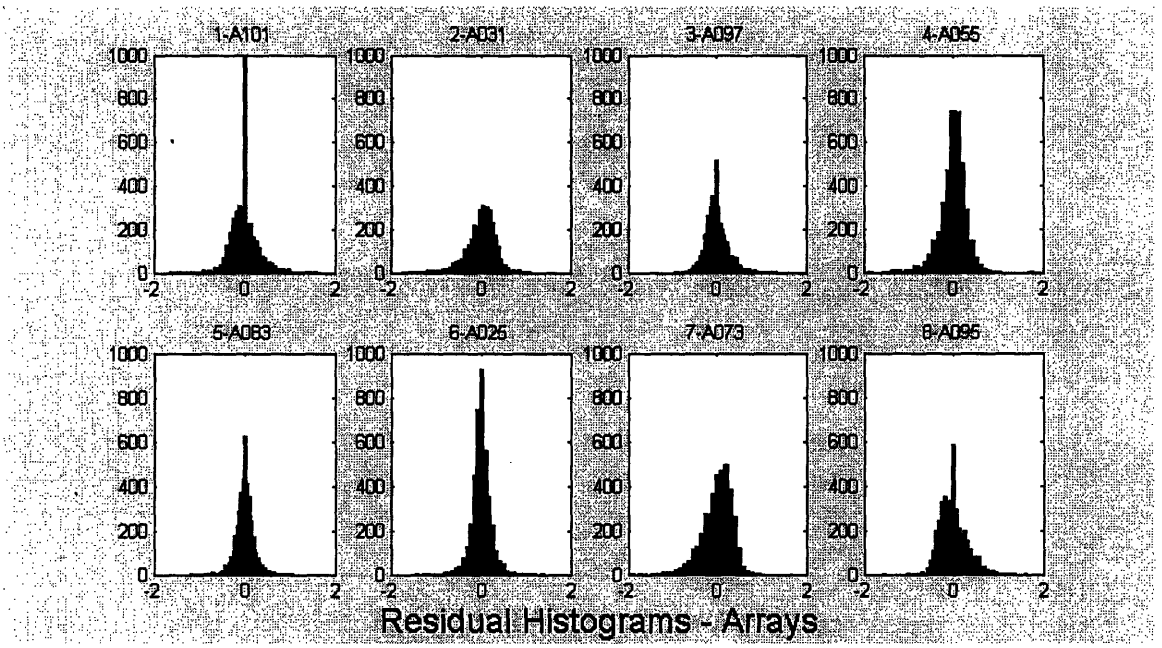


Figure 4.20: Histograms of Residual Values
Histograms of residuals for each array in the data set. These distributions are approximately normal with similar variances.

In addition, it is also well known that production of a heterologous protein will activate the heat-shock response in *E. coli*. Several genes involved in this response that are regulated by the σ^{32} RNA polymerase, were also selected as differentially expressed and appear in Table 4.4. These genes include *dnaK*, *dnaJ*, *hslS*, *hslT*, *htpG*, and *topA*. The results of this validation experiment met all of the expectations and confirmed that differentially expressed genes can indeed be identified using this analytical technique. Further analysis of the genes listed in Table 4.4 is reserved for Chapter 7.

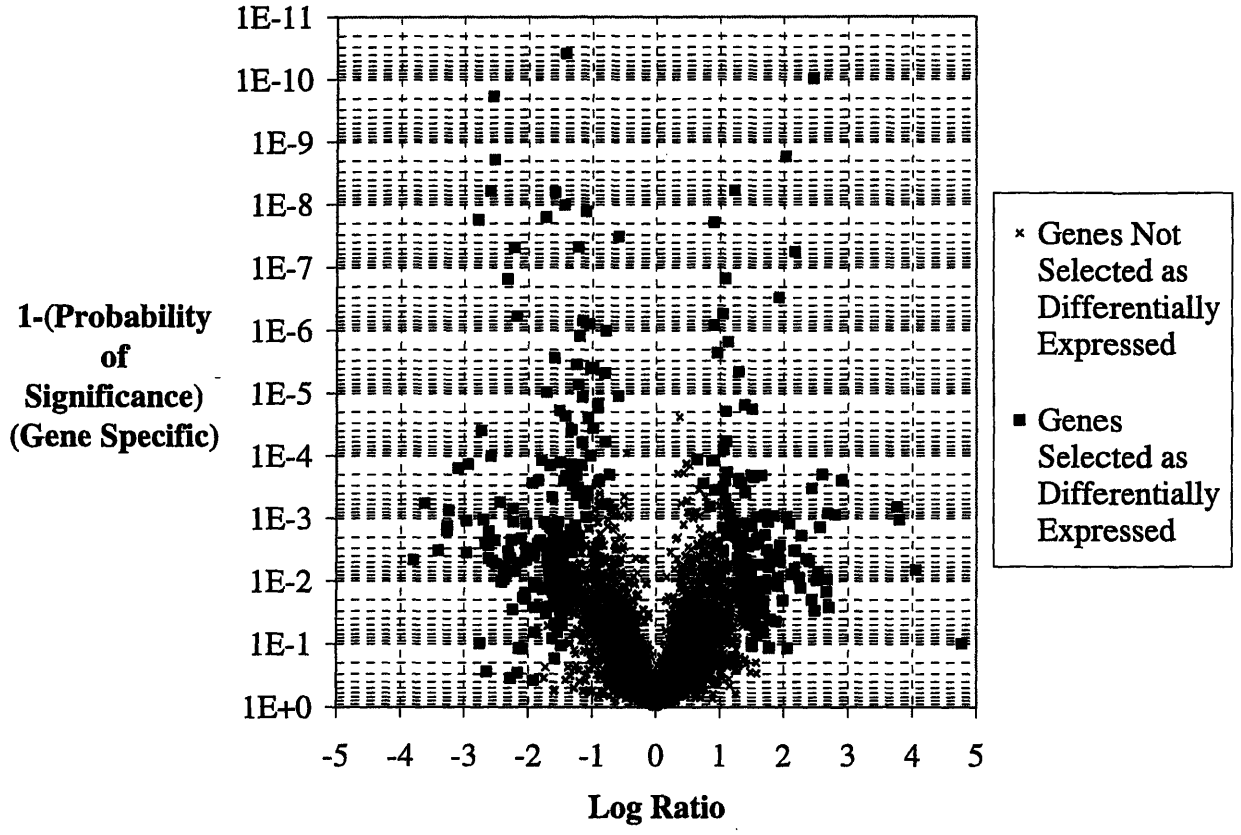


Figure 4.21: Volcano Plot for Selection of Differentially Expressed Genes
Each gene in the data set is plotted based on its log ratio and the probability of its significance (from the gene-specific test). Genes were selected as differentially expressed based on a combined probability accounting for both gene-specific and global variance.

4.7.3 ANOVA with Cy5 Signals

The same procedure for selecting differentially expressed genes was carried out for the $\log_2$(Cy5 Signals). Since the Cy5-labeled genomic DNA is used as a hybridization control, minimal variation would be expected in these signals, across the different experimental conditions. Indeed, using the same cutoff levels, only 21 (0.5%) of genes were selected as having apparent differential expression. This not only validates the reproducibility of the genomic DNA labeling and hybridization, but also serves as a negative control for the data analysis procedure.

When the ANOVA model is applied to the Cy5 signals, array effects (A_a) and gene effects (G_g) are expected to be significant. However, the BG_{bg} and TG_{tg} terms are not expected to be significant, since the genomic DNA sample is always the same, regardless of the total RNA

sample that is used. By comparing ANOVA tables for the signal ratios and the $\log_2$(Cy5 Signals), A_a and G_g terms were found to be significant. Surprisingly, the BG_{bg} and TG_{tg} terms were also significant with probability of > 0.99999. While the mean squares for these two terms decrease in the Cy5 signals table, the mean square for the overall residual error also decreases. Therefore, the ratios remain significantly larger than 1. Because of the large degrees of freedom for these parameters, even a small deviation from 1 will be significant.

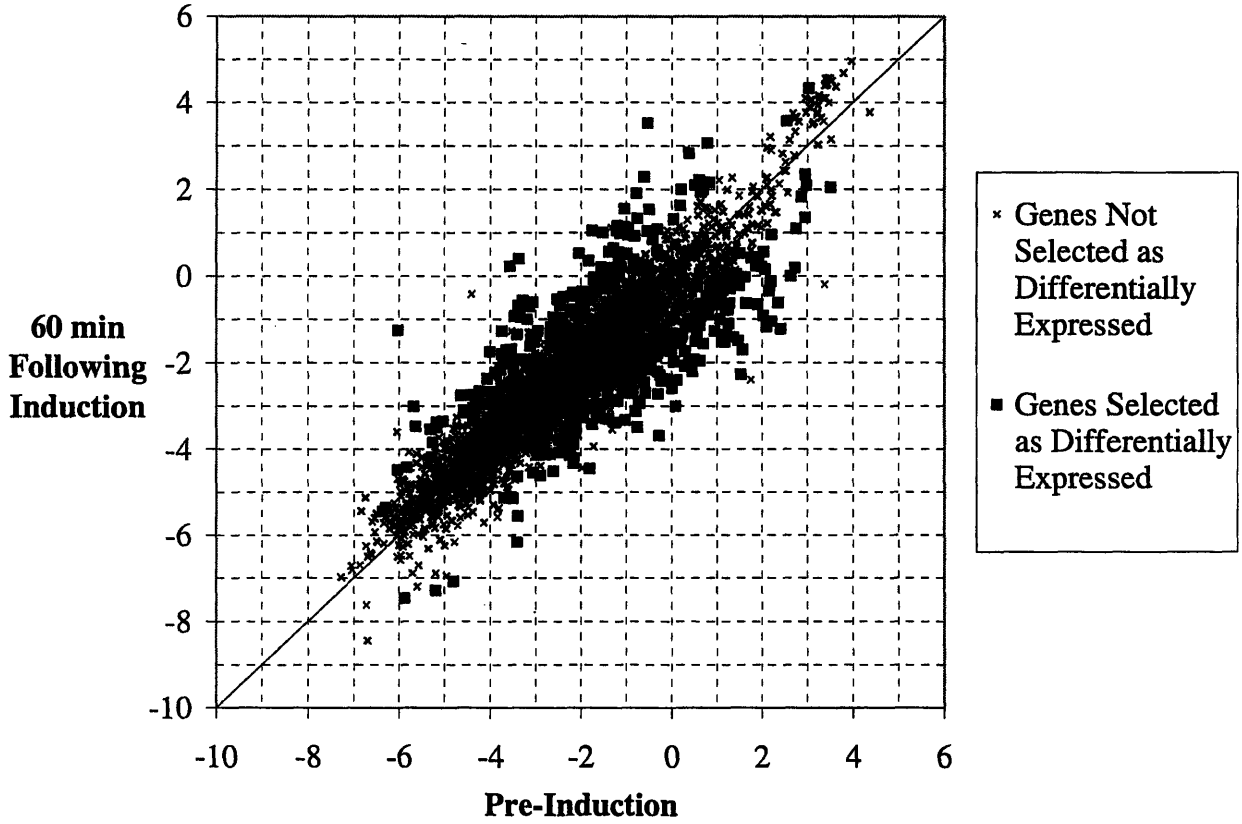


Figure 4.22: Scatterplot of Expression Values
Samples for expression analysis were taken from four identical cultures immediately before induction and 60 min following induction. Expression values are plotted for each treatment. The criteria used in this work for selecting differentially expressed genes is more complex than choosing genes that fall far from the diagonal line. The reproducibility and number of replicates also play a role in this selection.

4.7.4 Summary of Validation Experiment

Microarray analysis was performed on *E. coli* samples taken immediately before and 60 min after induction of recombinant protein production. A set of differentially expressed genes was selected. This set of genes will be important to recognize and understand since they

will serve as background in all of the planned experiments. Genes that respond to IPTG, genes involved in the induction process, and heat shock genes validated the experimental and statistical methods used. When the same statistical methods were applied to the Cy5 signals, a small number of genes were found to show apparent differential expression. This observation confirmed the reproducibility of the genomic DNA signal and further validated the data analysis protocol.

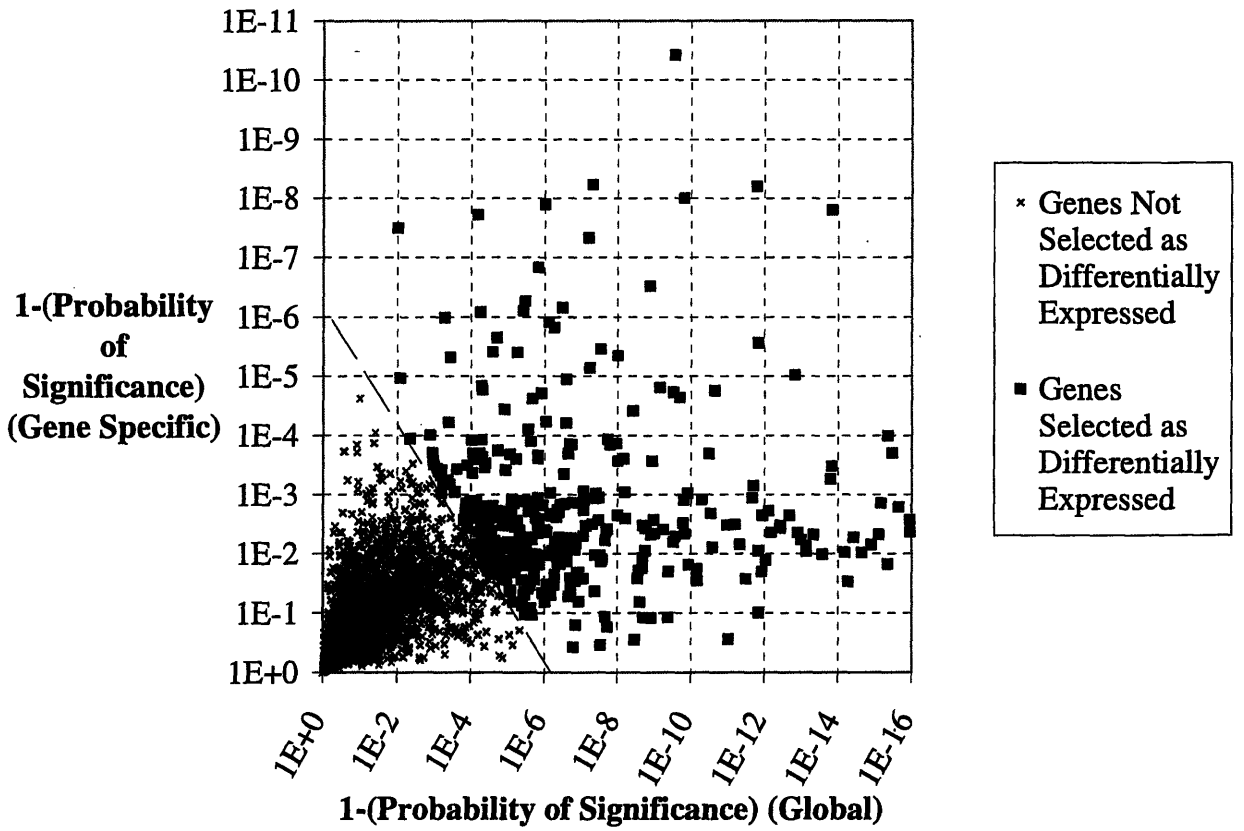


Figure 4.23: Scatterplot of Results from Global and Gene-Specific Tests
Each gene in the data set was plotted based on probabilities calculated from significance tests on the log ratios using gene-specific and global variances. This plot is analogous to the Volcano Plot in Figure 4.21 folded in half. These two probabilities are the criteria used to select genes as differentially expressed. A line, which crosses both axes at 6.5×10^{-7}, is the dividing line for differential expression.

Table 4.4: Genes Showing Increased and Decreased Expression 60 min following Induction

156 genes show increased expression and 228 genes showed decreased expression. Genes marked with a "*" showed consistent changes in (Lesley *et al.* 2002), while those with a "^" showed consistent changes in (Richmond *et al.* 1999).

Increased Expression 60 min Following Induction

Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio
lacY	b0343	4.8	yggA	b2923	1.7	xerD	b2894	1.4
AT	Antitr	4.1	yrfH*^	b3400	1.7	b1228	b1228	1.3
proV	b2677	3.8	b2433	b2433	1.7	rpoB	b3987	1.3
T7	T7Pol	3.8	yjcZ	b4110	1.7	cspG	b0990	1.3
b1266	b1266	2.9	atoD	b2221	1.7	abc	b0199	1.3
yhbZ	b3183	2.8	phoR^	b0400	1.7	b1478	b1478	1.3
b1362	b1362	2.7	spoT	b3650	1.7	ytfL	b4218	1.3
yi81_2^	b0582	2.7	b2562	b2562	1.7	mcrB	b4346	1.3
deaD	b3162	2.7	ybiA	b0798	1.7	gppA	b3779	1.3
mbhA	b0230	2.7	fdoG	b3894	1.7	infB	b3168	1.3
rnhB	b0183	2.6	yehR	b2123	1.6	vacB	b4179	1.3
proX	b2679	2.6	hemK	b1212	1.6	yagV	b0289	1.3
b1361	b1361	2.5	fdoI	b3892	1.6	prlA	b3300	1.3
hybB	b2995	2.5	yjfJ	b4182	1.6	nusA	b3169	1.3
cynX	b0341	2.5	yigW_1	b3840	1.6	hflK	b4174	1.3
proP*	b4111	2.5	fabH	b1091	1.6	lpxA	b0181	1.3
yhbM	b3163	2.4	ycfJ	b1110	1.6	queA	b0405	1.2
b0556	b0556	2.4	phoQ	b1129	1.6	b0949	b0949	1.2
yi82_1^	b0017	2.4	b0788	b0788	1.6	yehX	b2129	1.2
truB	b3166	2.4	yhbC	b3170	1.6	recF	b3700	1.2
plsX	b1090	2.3	dnaA	b3702	1.6	yhhS	b3473	1.2
hybC	b2994	2.3	yjfM	b4185	1.5	b2970	b2970	1.2
b0555	b0555	2.2	prfA	b1211	1.5	htpX*^	b1829	1.2
hybE	b2992	2.2	yijE	b3943	1.5	b1631	b1631	1.1
ptxA	b4195	2.2	intD	b0537	1.5	htpG*^	b0473	1.1
frwD	b3953	2.2	ychH	b1205	1.5	proW	b2678	1.1
rbfA	b3167	2.2	codA	b0337	1.5	yjcX	b4088	1.1
b2997	b2997	2.1	b0557	b0557	1.5	b3486	b3486	1.1
yhhI	b3484	2.1	b1592	b1592	1.5	yohK	b2142	1.1
b2680	b2680	2.1	rffE	b3786	1.5	melB	b4120	1.1
yhbE	b3184	2.0	b1445	b1445	1.5	ybaL	b0478	1.1
dnaJ*^	b0015	2.0	b2681	b2681	1.5	b0715	b0715	1.1
hslT*^	b3687	2.0	b0818	b0818	1.5	rhaT	b3907	1.1
fdoH	b3893	1.9	b1628	b1628	1.5	b0842	b0842	1.1
Int	b0657	1.9	cdsA	b0175	1.5	pitB	b2987	1.0
hslS*^	b3686	1.9	b2983	b2983	1.5	b0491^	b0491	1.0
hemA	b1210	1.9	clpB*^	b2592	1.5	rplR	b3304	1.0
hybG	b2990	1.9	b0868	b0868	1.5	b0879^	b0879	1.0
lpxB	b0182	1.9	b2474	b2474	1.5	uhpB	b3668	1.0
rffD	b3787	1.9	pcm	b2743	1.4	fecA	b4291	1.0
b1560	b1560	1.9	lpxD	b0179	1.4	ykgG	b0308	1.0
b0815	b0815	1.8	rffH	b3789	1.4	b1442	b1442	1.0
b2971	b2971	1.8	b2473	b2473	1.4	yjeM	b4156	0.9
topA^	b1274	1.8	b2881	b2881	1.4	yadM	b0138	0.9
yceC	b1086	1.8	yhfW	b3380	1.4	yadQ	b0155	0.9
yhiV	b3514	1.8	era	b2566	1.4	ydeA	b1528	0.9
codB	b0336	1.8	b2063	b2063	1.4	ftsA	b0094	0.9
yheL^	b3343	1.8	hflC	b4175	1.4	yhbG	b3201	0.9
mrdA	b0635	1.7	b2974	b2974	1.4	ftsQ	b0093	0.9
rffG	b3788	1.7	dnaK*^	b0014	1.4	b0934	b0934	0.8
dniR	b0211	1.7	yifJ	b3792	1.4	uhpC	b3667	0.7
b1791	b1791	1.7	b1523	b1523	1.4	dppB	b3543	0.6

Decreased Expression 60 min Following Induction

Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio
secB	b3609	-0.6	pheA	b2599	-1.3	b0838	b0838	-1.6	ilvD	b3771	-1.9
flgE^	b1076	-0.6	gpmA	b0755	-1.3	thiC	b3994	-1.6	cysI	b2763	-2.0
yacC	b0122	-0.7	rmf	b0953	-1.3	ppiA	b3363	-1.6	b1780	b1780	-2.0
flgL	b1083	-0.8	b1490	b1490	-1.4	cadB	b4132	-1.6	cysN	b2751	-2.0
yicL	b3660	-0.8	yebL	b1857	-1.4	cspC	b1823	-1.6	yciI	b1251	-2.0
flgG	b1078	-0.8	b1782^	b1782	-1.4	tpx	b1324	-1.6	ilvA	b3772	-2.0
feoA*	b3408	-0.8	pgk	b2926	-1.4	ptsH	b2415	-1.6	talB	b0008	-2.0
rbsD	b3748	-0.9	b2007	b2007	-1.4	b0346	b0346	-1.6	argI	b4254	-2.1
flgD	b1075	-0.9	b1777	b1777	-1.4	crp	b3357	-1.6	trpC	b1262	-2.1
tehB	b1430	-0.9	b1338	b1338	-1.4	hupA	b4000	-1.6	yjhB	b4279	-2.1
nfnB	b0578	-0.9	ydfG	b1539	-1.4	sdhD	b0722	-1.6	guaC	b0104	-2.1
glpF*	b3927	-1.0	bioC	b0777	-1.4	ilvN	b3670	-1.6	acs	b4069	-2.1
cysU	b2424	-1.0	gltK	b0653	-1.4	hisB	b2022	-1.6	b2844	b2844	-2.2
ptsI	b2416	-1.0	cfa	b1661	-1.4	pgi	b4025	-1.6	phnJ	b4098	-2.2
b1729	b1729	-1.0	argR	b3237	-1.4	htrA	b0161	-1.6	malG	b4032	-2.2
b1757	b1757	-1.0	trpD	b1263	-1.4	yiaE	b3553	-1.6	bioB	b0775	-2.2
flgF	b1077	-1.0	mdoG	b1048	-1.4	cysJ	b2764	-1.6	malF	b4033	-2.2
minD	b1175	-1.1	ybeK	b0651	-1.4	pfkB	b1723	-1.6	b1044	b1044	-2.2
b2529	b2529	-1.1	bioA	b0774	-1.4	kdpB	b0697	-1.6	b1178	b1178	-2.2
b1820	b1820	-1.1	cysP	b2425	-1.4	ybjC*	b0850	-1.6	nadE	b1740	-2.2
gltL	b0652	-1.1	adk	b0474	-1.4	ndk	b2518	-1.6	cysD	b2752	-2.3
yceH	b1067	-1.1	hisF	b2025	-1.4	fldA	b0684	-1.6	b1697	b1697	-2.3
cysA	b2422	-1.1	b1498	b1498	-1.4	yigL	b3826	-1.6	sucB	b0727	-2.3
gapC_2	b1416	-1.1	yhjE	b3523	-1.4	mdh	b3236	-1.6	tyrB	b4054	-2.3
crr	b2417	-1.1	sdhA	b0723	-1.4	yhjS	b3536	-1.6	trpB	b1261	-2.3
b0572^	b0572	-1.2	yibD	b3615	-1.4	asd	b3433	-1.6	yccA*	b0970	-2.3
nupC	b2393	-1.2	purU	b1232	-1.4	b1650	b1650	-1.6	cysH	b2762	-2.4
dps	b0812	-1.2	aspC	b0928	-1.5	grxB	b1064	-1.6	panD	b0131	-2.4
nrdA	b2234	-1.2	hisC	b2021	-1.5	arcA	b4401	-1.6	cybC	b4236	-2.4
minE	b1174	-1.2	sdhC	b0721	-1.5	manX	b1817	-1.7	fumA	b1612	-2.4
b2301	b2301	-1.2	gapA	b1779	-1.5	fpr	b3924	-1.7	yfiA	b2597	-2.4
dgt	b0160	-1.2	gatY	b2096	-1.5	cysC	b2750	-1.7	icdA	b1136	-2.5
b1448	b1448	-1.2	b3995	b3995	-1.5	b1973	b1973	-1.7	ilvB	b3671	-2.5
ygiM	b3055	-1.2	trpS	b3384	-1.5	pepD	b0237	-1.7	lamB*	b4036	-2.5
b0574	b0574	-1.2	b1724	b1724	-1.5	livJ	b3460	-1.7	malK	b4035	-2.6
glpK	b3926	-1.2	thiE	b3993	-1.5	aroP	b0112	-1.7	trpA	b1260	-2.6
ahpC	b0605	-1.2	ilvG_1	b3767	-1.5	b0573	b0573	-1.7	argG	b3172	-2.6
purH	b4006	-1.2	ycgC	b1198	-1.5	nrdB*	b2235	-1.7	argT	b2310	-2.6
hemL	b0154	-1.3	aceA	b4015	-1.5	gnd	b2029	-1.7	sodA	b3908	-2.6
yaiE	b0391	-1.3	pflB^	b0903	-1.5	ydaH	b1336	-1.7	cysK	b2414	-2.6
lipA	b0628	-1.3	udhA	b3962	-1.5	b1337	b1337	-1.7	hisJ	b2309	-2.6
cysW	b2423	-1.3	uspA	b3495	-1.5	sbp	b3917	-1.7	tdcA	b3118	-2.7
hisD	b2020	-1.3	pflA	b0902	-1.5	adhE	b1241	-1.7	tyrA	b2600	-2.7
serC	b0907	-1.3	gltA	b0720	-1.5	purD	b4005	-1.8	fba^	b2925	-2.7
gltJ	b0654	-1.3	mdaA	b0851	-1.5	manY	b1818	-1.8	yceI	b1056	-2.7
atpC	b3731	-1.3	b0332	b0332	-1.5	glnH	b0811	-1.8	b2537	b2537	-2.8
artP	b0864	-1.3	glyA^	b2551	-1.5	b3004	b3004	-1.8	malE	b4034	-2.8
glnA	b3870	-1.3	dapD	b0166	-1.5	hisM	b2307	-1.8	sucD	b0729	-2.9
tus	b1610	-1.3	b1431	b1431	-1.5	b1043	b1043	-1.8	aroF	b2601	-3.0
hupB	b0440	-1.3	slyD	b3349	-1.5	yacK	b0123	-1.8	sucC	b0728	-3.0
yhbW	b3160	-1.3	yadF	b0126	-1.5	tpiA	b3919	-1.9	artJ	b0860	-3.1
aroL	b0388	-1.3	glnB	b2553	-1.5	yjgD	b4255	-1.9	gatB	b2093	-3.2
bcp	b2480	-1.3	yccJ	b1003	-1.5	agp	b1002	-1.9	sucA	b0726	-3.3
ilvM	b3769	-1.3	dsbA	b3860	-1.5	manZ	b1819	-1.9	b0725	b0725	-3.3
kgtP	b2587	-1.3	dacA	b0632	-1.5	asnA	b3744	-1.9	aldA	b1415	-3.4
argA	b2818	-1.3	b1163	b1163	-1.6	trpE	b1264	-1.9	gatC	b2092	-3.6
acnB	b0118	-1.3	b1059	b1059	-1.6	b2853	b2853	-1.9	ilvC	b3774	-3.8

4.8 Summary of Microarray Development

Techniques were developed for performing microarray analysis and interpreting the volumes of data generated from these experiments. These techniques were validated to confirm that data were meaningful and reproducible. High-throughput data can now be generated with confidence.

5 Effects of Aeration on Induced Cultures

"There are three types of baseball players: those who make it happen, those who watch it happen, and those who wonder what happens."
—Tommy Lasorda

In order to elucidate the effects of varying aeration conditions, it is helpful to understand the players involved. The DNA-microarray techniques developed in the previous chapter were applied to identify genes that show altered expression in response to varying aeration conditions. Some of these genes are directly involved in stress responses and adaptation to the new aeration environments. Others are indirect, and sometimes unexpected, consequences of these stress responses. Therefore, DNA microarrays reveal not only the conditions to which the culture was exposed, but also how it responds at a global level.

The effect of oxygen-dependent α1-antitrypsin (α_1AT) degradation was reproduced in this work using the pulse-chase protocol described in Section 3.8 (Figure 5.1). Immediately following induction, a culture grown in air at 30°C in minimal medium was split between three bubbler tubes and each was sparged with a different gas: pure nitrogen, air, and pure oxygen. Newly synthesized α_1AT was pulse-chase labeled starting 60 min after induction.

α_1AT degradation has its origin in the heat-shock response (Laska 2000), which is known to be activated by production of misfolded proteins. Furthermore, this degradation is not due to *in vivo* oxidation of the oxygen sensitive methionine residues (Laska 2000; Griffiths 2002). Therefore, the oxygen dependence of this degradation is likely due to the hyperoxic stress response in *E. coli*. In order to understand the combined effects of both the heat-shock and hyperoxic stress responses, DNA-microarray analyses were carried out on *E. coli* cultures producing recombinant α_1AT in defined aeration environments.

5.1 Experimental Details

As described in Section 3.7.5, 400-mL cultures of *E. coli* BL21 (DE3) pEAT8-137 were grown at 30°C to OD_{600} of 0.7. At this point, the large culture was split into three smaller cultures, each exposed to a different headspace gas: pure nitrogen, air, and pure oxygen (these three cultures will be referred to as N2, AIR, and O2 throughout this thesis). Simultaneous to the split, the cultures were induced to produce recombinant α_1AT by addition of IPTG. Cultures

were grown for 90 min and samples for both OD_{600} measurement and microarray analysis were collected at specified time points. This experiment was repeated three times on three different days. For all of these cultures, t = 0 min refers to both the time at which IPTG was added to begin induction, as well as the time at which the culture was split and exposed to different aeration conditions. Figure 5.2 shows the growth curve for a typical culture. Growth rates during growth and induction phases from these three experimental sets are given in Table 5.1. Cultures grown in pure nitrogen show much slower growth than those in air and pure oxygen. In fact, these cultures show linear, *not* exponential, growth. Although the exponential growth rate for the N2 culture in Table 5.1 is useful for the purpose of comparison, a linear model in which OD_{600} increases by 0.075 units/h is more appropriate.

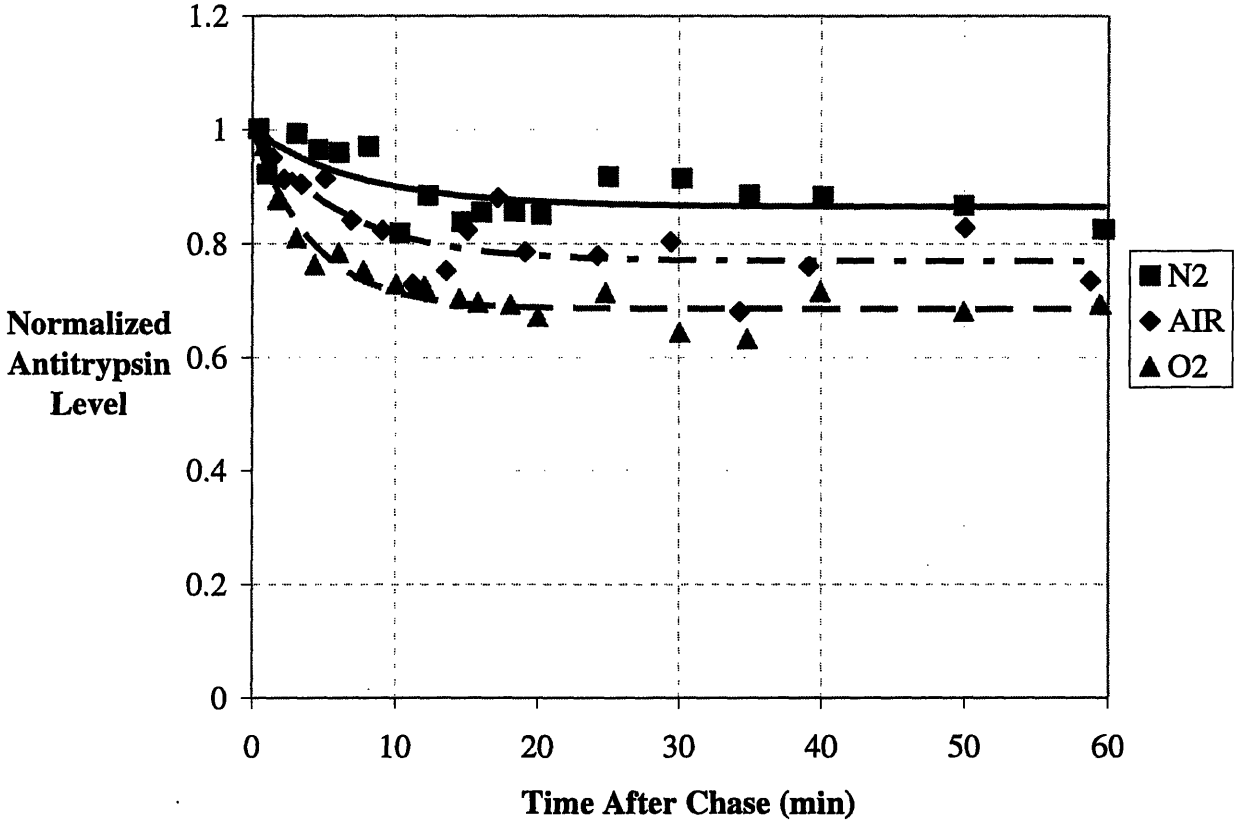


Figure 5.1: Degradation Profile of α1-Antitrypsin from Cultures Induced in Pure N_2, Air, and Pure O_2
A culture at OD_{600} of 0.7 was simultaneously split and induced in three different aeration environments: pure nitrogen, air, and pure oxygen. After 60 min of induction, newly synthesized protein was pulse-chase labeled with ^{35}S-methionine. Degradation of recombinant α_1AT was monitored for the next 60 min.

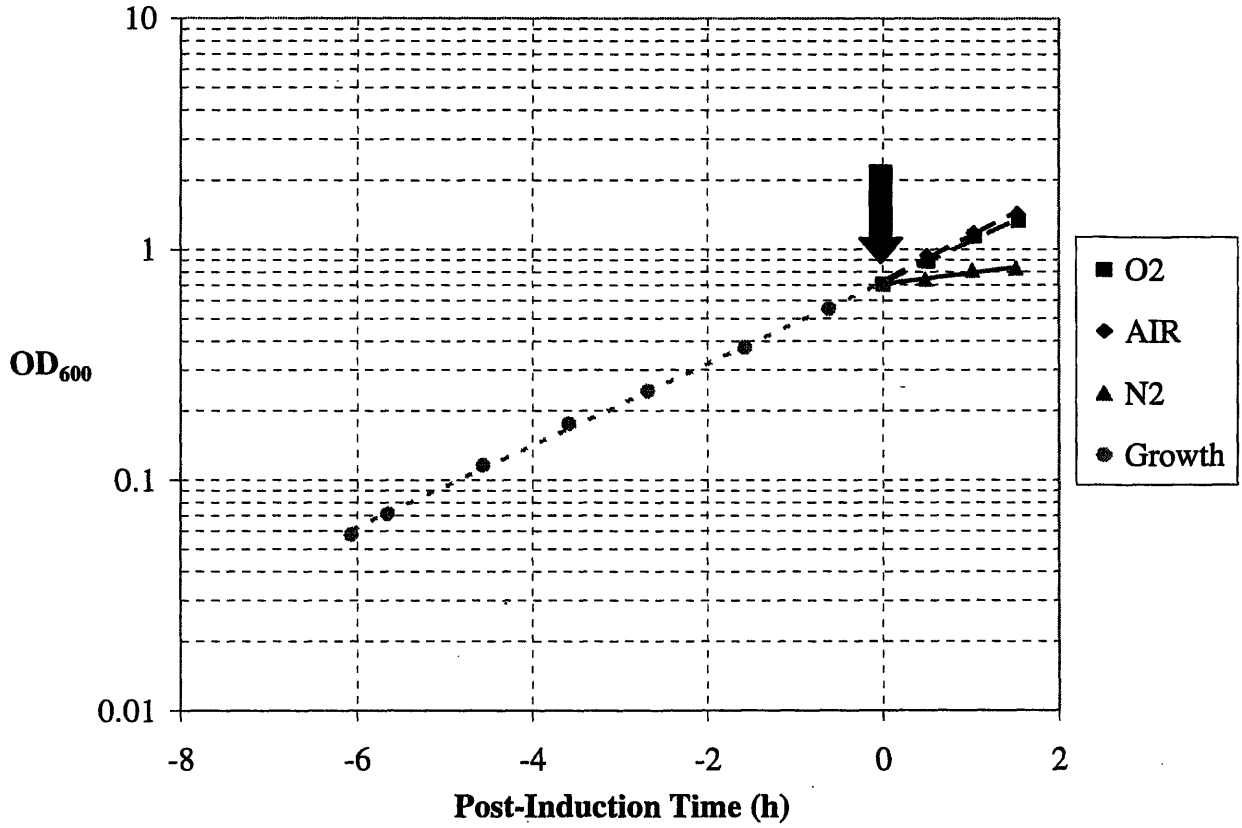


Figure 5.2: Growth Curves for Cultures Induced in N_2, Air, and O_2
A culture was grown to OD_{600} of 0.7. At this point ($t = 0$), the culture was split into three smaller cultures, each of which was induced (arrow). Induction occurred under pure nitrogen, air, and pure oxygen for the three cultures. The entire experiment was repeated three times; these results are typical of the repeated cultures.

Table 5.1: Specific Growth Rates for Cultures Induced in N_2, Air, and O_2
Specific growth rates (μ) were calculated by applying the standard model for exponential growth: $\frac{dOD_{600}}{dt} = \mu \cdot OD_{600}$. Values were calculated by applying a linear regression to $\log(OD_{600})$ vs. t data (*e.g.* Figure 5.2). 95% confidence intervals are given.

	Growth	Induction		
		N2	AIR	O2
Specific Growth Rate (h^{-1})	0.43 ± 0.01	0.10 ± 0.05	0.51 ± 0.08	0.44 ± 0.08

After the 90-min induction period, protein extracts were prepared and the α_1AT activity assay was performed. Figure 5.3 shows the activities of α_1AT from each of the cultures on a per cell basis. Production of recombinant α_1AT was highest in AIR cultures. The O2 and N2

cultures showed production rates that were lower, but comparable to one another. For AIR and O2 cultures, specific activity values were roughly 50% higher than those obtained in previous work after 60 min of induction (Laska 2000), indicating a constant production rate on a per cell basis. For N2 cultures, specific activity values were almost twice as high as those after 60 min. This difference could be due to experimental variance. For example, oxygen entering the N2 cultures during sampling is a large source of error. Alternatively, the differing results might indicate that the α_1AT production rate increased between 60 min and 90 min in the N2 culture.

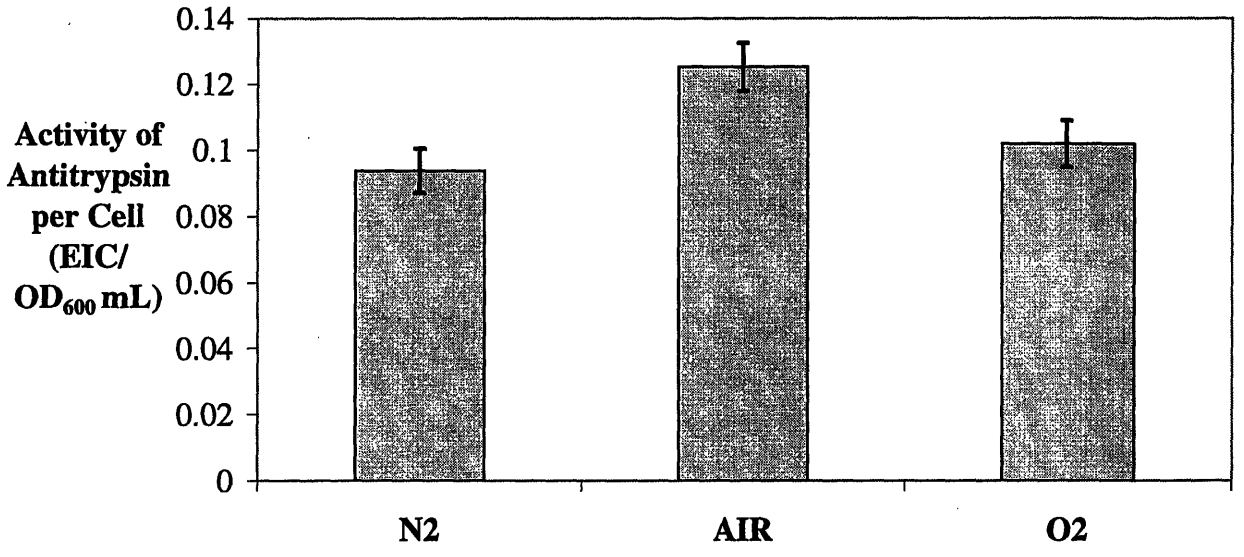


Figure 5.3: α1-Antitrypsin Activity per Cell from Cultures Induced in N_2, Air, and O_2
Cultures were induced to produce recombinant α_1AT for a 90-min induction period in pure nitrogen, air, and pure oxygen. This experiment was repeated three times. These results are typical of the three repeated experiments.

5.2 Analysis of Expression Data

The three repeated growth experiments will be referred to as Blocks A, B, and C. In addition to pre-induction ($t = 0$) samples, samples were collected at $t = 10$, 30, 60, and 90 min for each culture of each block. As described in Section 3.7, RNA was purified from these samples and was used to analyze global gene expression. Samples from the Block A experiment were analyzed on slides from Print B, while those from Blocks B and C were analyzed on slides from Print E. For Blocks B and C, samples at 30 min and 90 min were not analyzed; therefore, analysis of the gene expression data was divided into two categories: (1) Block A at all time

points, and (2) All blocks at 0-, 10-, and 60-min time points. Data generated from these experiments were analyzed using unbalanced ANOVA models as appropriate.

5.2.1 Analysis of Block A at All Time Points

5.2.1.1 Identification of Differentially Expressed Genes

With the data from Block A alone, it was not possible to use the combined probability method described in Section 4.5.3 to identify differentially expressed genes. Except for those genes that were duplicated on the same array, there was only one measurement for each gene under each treatment. Therefore, there were not enough degrees of freedom to calculate gene-specific probabilities of significance. Instead, the global significance test was used as the selection criterion for differential expression.

This single block contained thirteen treatments, which produced 78 unique treatment comparisons. Of the 4,013 genes in the data set, 1,787 genes (45%) showed significant differential expression in at least one comparison. The large number of genes showing expression changes is partly due to the large number of comparisons. The N2 samples were also responsible for a large fraction of these expression changes. The length of this list cannot be attributed to the global significance test, since it is generally more stringent than either the gene-specific or combined tests. To make the data set more meaningful, comparisons involving different aeration gases at different times (*e.g.* 10-N2 vs. 30-AIR) were eliminated. The 42 remaining comparisons included all comparisons with the pre-induction sample and still contained 1,465 genes (37%) with significant changes.

Using the expression values from this analysis, hierarchical clustering was carried out as described in Section 4.5.4.2.

5.2.1.2 Aeration-vs.-Time ANOVA

In the ANOVA in Section 5.2.1.1, each sample belonged to one of 13 treatments. Each gene was analyzed as a 13 × 1 data set. However, the design of the experiment permits further analysis of the experimental parameters, namely aeration (N_2, air, or O_2) and time after induction. By subtracting the zero-time-point sample from all other samples, the 13-treatment data set can be recast as a 12-sample data set with the following design:

		Post-Induction Time (min)			
		10	30	60	90
Aeration	N2	×	×	×	×
	AIR	×	×	×	×
	O2	×	×	×	×

In this form, it would be possible to apply an additional two-way ANOVA to determine the effects of aeration and time after induction on every gene in the data set.

To perform this second-round analysis, the expression values from the first round of ANOVA, calculated as in **(5.1)**, were used.

$$\bar{\hat{y}}_{\bullet tg\bullet\bullet} = \mu + G_g + TG_{tg} \tag{5.1}$$

For those genes with duplicate spots on the same array, the benefits of replication were lost, since these values were averaged across all replicates. However, this shortcut made the analysis easier since subtraction from only one zero-time-point sample was required. For each gene, the expression value for the zero-time-point sample was subtracted from the expression values for every other sample. These expression values were essentially converted to log ratios for comparisons with the pre-induction sample. A new ANOVA model was applied to each gene as follows:

$$\left(\bar{\hat{y}}_{\bullet pqg\bullet\bullet} - \bar{\hat{y}}_{\bullet(AIR)0g\bullet\bullet}\right) = \mu_g + P_{pg} + Q_{qg} + \dot{\varepsilon}_{pqg} \tag{5.2}$$

This model is similar to **(5.1)**, except the t index has been replaced by pq and the TG_{tg} interaction term has been further analyzed to give P_{pg}, Q_{qg}, and a residual error term $\dot{\varepsilon}_{pqg}$. The term μ_g represents the average expression difference for each gene, P_{pg} represents aeration effects, and Q_{qg} represents transient effects. Although this data set has a balanced design and replicate data were eliminated, this model was treated as unbalanced since some data were missing. Since each combination of parameters contained either one value or none at all, there were not enough degrees of freedom to analyze interaction effects.

This analysis identified 350 genes that changed expression with aeration in a consistent manner across all time points. Only 27 of these genes were found to change in the comparison between the AIR and O2 cultures; therefore, most of identified changes involved the N2 culture.

Additional analysis of this data set identified 115 genes that showed consistent changes between time points regardless of aeration. Most of these genes were also identified by the induction validation experiment described in Section 4.7. Since the expression changes observed

from these genes occurred regardless of aeration, they are not relevant to the goals of this chapter and will be discussed further in Chapter 7.

5.2.2 All Blocks at 0-, 10-, and 60-min Time Points

Addition of data from the repeated experiments allowed reproducible differential expression to be identified. With a data set consisting of seven treatments and three blocks, differential expression was determined by using the ANOVA model described in Section 4.5. Out of 3,973 genes in this data set, 1,097 (28%) were found to be differentially expressed in at least one treatment comparison. Of the 21 comparisons among these seven treatments, six were eliminated as meaningless, since they involved samples from both different times and different aeration gases. The remaining comparisons still contained 959 genes (24%) with significant expression changes.

5.2.3 Induction in Air vs. Oxygen

Analysis of the Block A data revealed a large number of expression changes involving the N2 culture. When compared with the pre-induction (0-min) sample, the 30-N2 and 60-N2 samples have more than four hundred genes changing expression (Figure 5.4). In contrast, the AIR and O2 cultures have, at most, 122 differentially expressed genes. Induction in N_2 clearly leads to a dramatic shift in the metabolism of the cell, which accounts for many of these expression changes. A similarly large number of changes in gene expression were observed in other anaerobic microarray studies (Salmon *et al.* 2003; Liu and Wulf 2004). Genes differentially expressed in N2 cultures are examined in more detail in Section 5.6.

Throughout this thesis, experiments and data analysis were carried out identically for the N2, AIR, and O2 cultures. However, the interpretation focuses on gene expression changes between the AIR and O2 cultures. These changes are more subtle and involve a more manageable set of genes. Moreover, the changes between AIR and O2 are most relevant toward understanding the oxygen-dependent degradation of α_1AT.

Based on the above analyses, a set of 133 genes was identified that showed differential expression between the AIR and O2 cultures at the same time. This list was seeded with genes selected from the three tests described in the previous sections. The global probability test in the analysis of Block A identified 22 genes that showed differential expression in at least two time points. From the second-round ANOVA analysis of the Block-A data, 27 genes showed

consistent differential expression between the AIR and O2 cultures across all time points. The combined probability test using all three blocks selected 96 genes that show significant differential expression in either the 10-min or 60-min time points. These genes were combined to generate a list of O2-AIR differentially expressed genes.

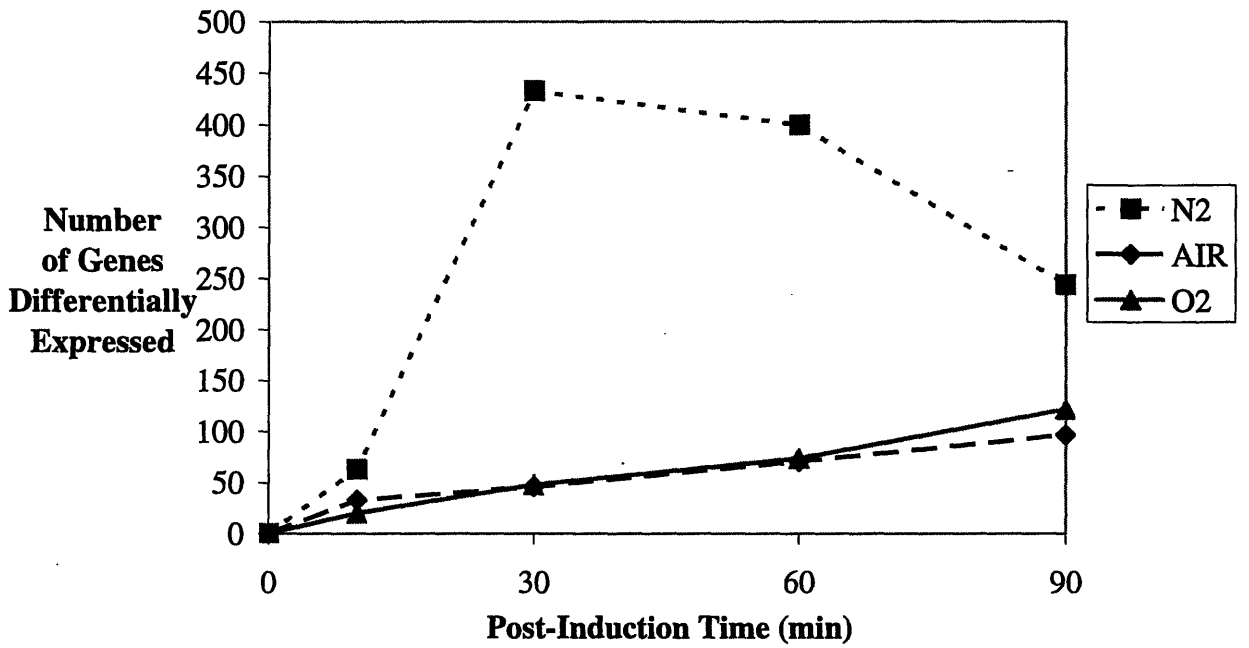


Figure 5.4: Cultures Induced in N_2 Have the Largest Number of Expression Changes
Considering only comparisons with the pre-induction (0-min) sample, the numbers of genes differentially expressed in the Block-A experiment are plotted for each culture.

Cluster analysis of the Block-A data found three clusters with significant representation on this list. Each cluster was named based on the functions of the clustered genes. The Suf Cluster, the Fur Cluster, and the SoxRS Cluster were represented by five, eight, and 20 genes on the list. Pictorial representations of the data from these three clusters are presented in Figure 5.5 and the average expression profiles of each cluster are plotted in Figure 5.6. Because these clusters had stronger representation in this list than any other cluster and showed strong oxygen dependence, the remaining genes from these clusters were also added to the list. These additional genes included two from the Suf Cluster, eight from the Fur Cluster, and 18 from the SoxRS Cluster. Although significant differential expression was not observed in these genes, their patterns of expression showed oxygen dependence. The final list of O2-AIR differentially expressed genes is shown in Table 5.2.

5.2.4 Summary of Analysis of Expression Data

This section has described the approaches taken to analyzing microarray data from experiments involving changing aeration in induced cultures. The remaining sections in this chapter interpret the results of these analyses.

5.3 Hyperoxic Stress Responses

In order to understand the state of the culture at each time point, hyperoxic stress response genes were examined. Trends observed in the expression profiles of peroxide stress response genes were consistent across several genes in the OxyR regulon. Despite the similarity in these genes, none were found to cluster together. While there were consistent trends throughout the O2 and AIR cultures, expression values from the N2 samples showed high variance, which likely prevented some of these genes from clustering together. Genes involved in the superoxide stress (SoxRS) response also showed profiles that were similar to one another and formed one cluster consisting of the genes *acrA*, *fur*, *nfo*, *sodA*, and *soxS*. Note that the gene *fur*, which encodes the ferric uptake regulator, appears in both the OxyR regulon and the SoxRS regulon. However, in this experiment, *fur* behaved like the other SoxRS genes.

The hyperoxic stress genes from both regulons were grouped and plotted, as shown in Figure 5.7. The responses to the change in aeration condition were distinct in each regulon. Genes in the OxyR regulon showed increased expression in the O2 culture relative to the AIR culture, but this difference was observed only at 10 and 30 min. At longer times, these genes were expressed at similar levels in both cultures. The peroxide response appeared to be active only at short times, indicating that species like H_2O_2, $HOO^{\bullet}$, and $HO^{\bullet}$ did not pose a threat at longer times. In contrast, genes involved in the SoxRS regulon showed increased expression in the O2 culture at 30, 60, and 90 min, when compared with the AIR culture. The superoxide response began as soon as 10 min following exposure to pure O_2, as indicated by the *soxS* gene, which showed a nearly 9-fold increase in expression between the pre-induction sample and the 10-O2 sample. Overall, the SoxRS response appeared to be most active at longer times and perhaps would have remained active had the culture times been extended. Clearly, the activation of each of these stress responses is distinct.

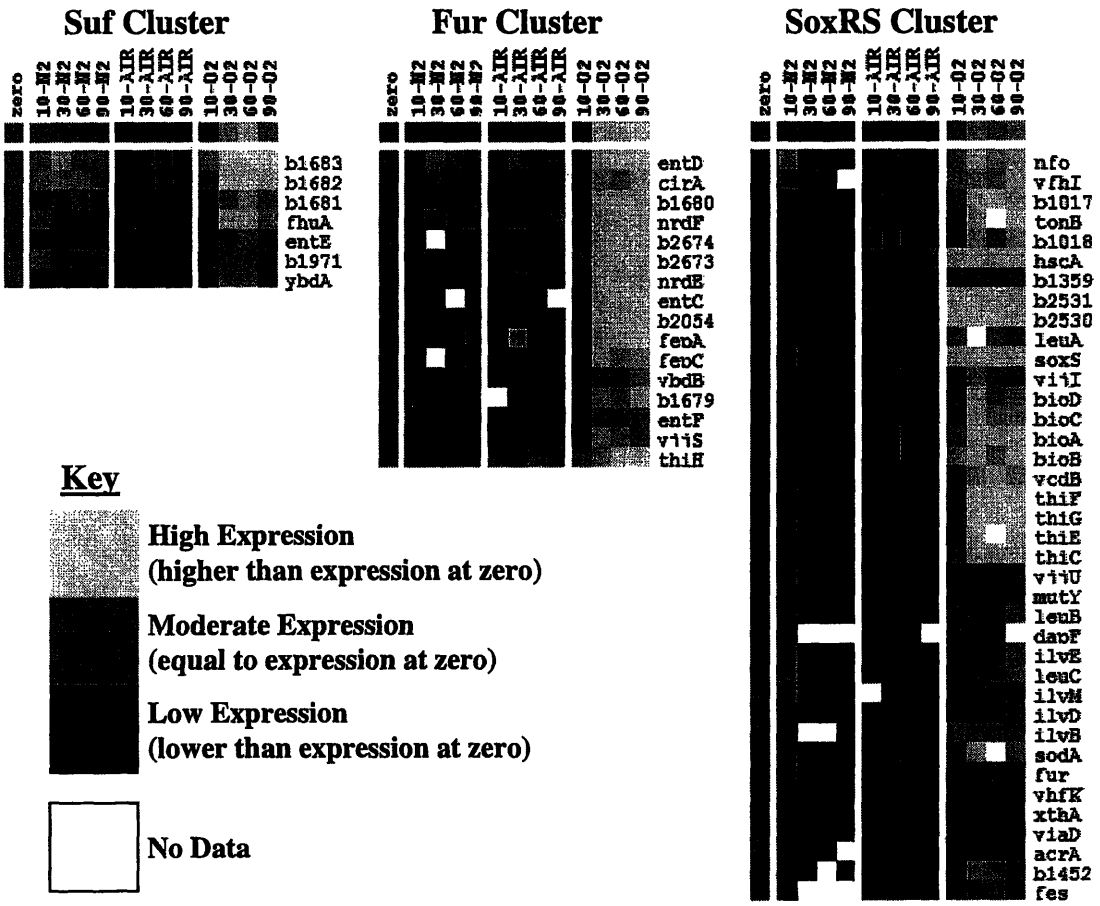


Figure 5.5: Oxygen-Dependent Clusters
Hierarchical clustering was performed on the entire Block-A data set of expression values as described in Section 4.5.4.2. Correlation coefficients for these clusters are 0.831 (Suf Cluster), 0.816 (Fur Cluster), and 0.829 (SoxRS Cluster). Each column represents one sample and each row represents a gene. The color of each square represents the magnitude of the expression value as indicated by the key. Notice that all expression data are shown relative to that in the zero sample. The top row of each cluster shows the average expression profile of genes in the cluster. In these clusters, O2 samples have higher expression than AIR samples.

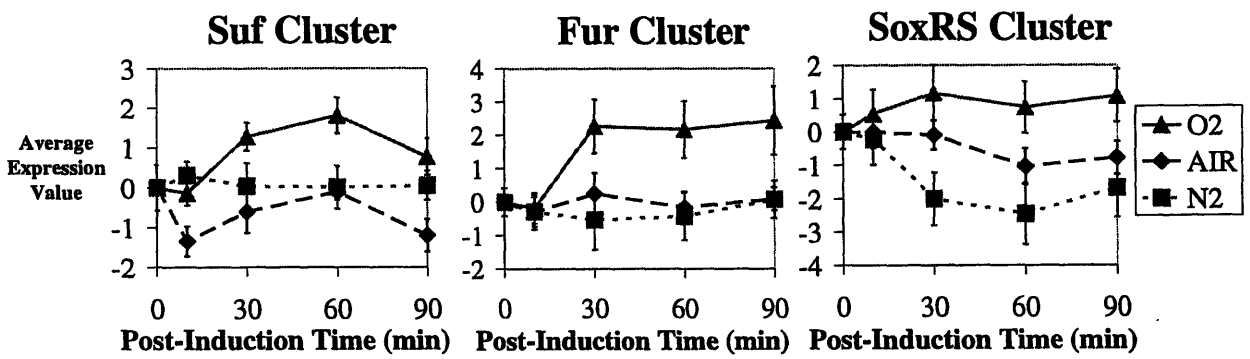


Table 5.2: Genes Differentially Expressed by Induction in Air and O_2
(on next two pages) 133 genes were selected as being differentially expressed between the AIR and O2 cultures. Genes were selected by any one of four criteria: **(1)** significance in at least two time points of the Block-A analysis, **(2)** significance in Block-A Aeration-vs.-Time analysis, **(3)** significance in at least one time point of the analysis of all blocks, **(4)** inclusion in one of the three major oxygen-dependent clusters. Information on gene products was obtained from the EcoCyc database. Log ratios are presented for comparisons between the AIR and O2 cultures. 112 of these genes showed increased expression in the O2 culture (positive log ratio), while 21 showed decreased expression. Significant log ratios are presented in gray. Cluster information is given only for those genes in one of the three major oxygen-dependent clusters. Gene names of genes identified in (Zheng, Wang, Templeton *et al.* 2001) are marked with a "*" and those identified in (Pomposiello *et al.* 2001) are marked with a "^".

Although not shown in Figure 5.7B, the SoxRS data from the N2 culture were consistently lower than those from the AIR culture. Even in AIR cultures, the SoxRS response appears to be moderately active.

The observations in Figure 5.7 were confirmed by examination of the individual genes in these two regulons. Based on a global significance test using only Block-A data, between the O2 and AIR cultures, the genes *grxA* and *dps*, of the OxyR regulon, showed significant O2-AIR differences only at the 10-min time point. In contrast, the gene *sodA*, of the SoxRS regulon, showed significant differences at 30, 60, and 90 min. Moreover, the gene *soxS* was the only gene with significant O2-AIR differences at all four time points. At later time points, the culture appeared to struggle with the effects of superoxide, but did not mount a response against the effects of peroxide and associated species.

Figure 5.6: Average Expression Profiles from Oxygen-Dependent Clusters
(opposite) Three oxygen-dependent clusters were well represented in the list of oxygen-dependent genes. Scaled expression values for each gene in the cluster were averaged. These values were scaled in order to minimize their variance. Error bars represent standard deviations in the expression values. All three of these clusters show significant differential expression between the AIR and O2 cultures at 30, 60, and 90 min.

Gene Name	Gene ID	Gene Product	Log Ratio (O2:AIR)							Cluster
			Block A				Block A Aeration vs.	All Blocks		
			10'	30'	60'	90'	Time ANOVA	10'	60'	
yaaI	b0013	conserved hypothetical protein	1.5	-0.2	-0.3	-0.2	0.2	2.2	-0.9	
leuC	b0072	isopropylmalate isomerase	0.5	0.8	1.7	1.6	1.2	0.6	2.1	SoxRS Cluster
leuB	b0073	3-isopropylmalate dehydrogenase	0.4	0.8	1.5	1.7	1.1	-0.3	1.5	SoxRS Cluster
leuA	b0074	2-isopropylmalate synthase	1.3	0.0	1.5	1.1	1.3	1.1	1.1	SoxRS Cluster
dksA	b0145	protein involved in translational regulation of RpoS	1.6	-0.3	-0.1	0.4	0.4	1.8	-0.9	
fhuA	b0150	outer membrane protein receptor for ferrichrome, colicin M, and phages T1, T5, and phi80	1.4	2.0	2.4	2.0	1.9	1.0	1.7	Suf Cluster
tra8_1	b0256	transposase of IS30	-3.5	-1.3	-0.4	0.0	-1.6	2.4	0.6	
codA	b0337	cytosine deaminase	0.3	-0.7	-1.6	-1.6	-0.9	-0.3	-2.3	
yaiL	b0354	nucleoprotein/polynucleotide-associated enzyme	1.1	-0.5	-0.5	0.2	0.1	1.7	-0.7	
psiF	b0384	induced by phosphate starvation	1.7	1.0	0.6	-0.4	0.7	2.0	-0.2	
ylaC	b0458	putative membrane protein	-0.3	-0.4	-1.4	-0.5	-0.6	0.8	-2.9	
*acrA^*	b0463	component of multidrug efflux system	0.0	0.8	0.5	1.3	0.6	0.0	0.5	SoxRS Cluster
entD	b0583	phosphopantetheinyl transferase	-0.8	1.8	1.9	2.1	1.3	-1.6	1.9	Fur Cluster
fepA	b0584	outer membrane receptor for ferric enterobactin (enterochelin) and colicins B and D	0.8	2.5	4.7	3.6	2.9	-0.7	3.7	Fur Cluster
fes	b0585	enterochelin esterase	0.1	1.1	2.1	0.6	1.0	0.1	1.4	SoxRS Cluster
entF	b0586	serine activating enzyme	0.2	1.2	0.9	1.4	0.9	0.1	0.9	Fur Cluster
fepC	b0588	ferric enterobactin ABC transporter	-0.5	2.4	1.3	1.8	1.2	-0.8	1.2	Fur Cluster
ybdA	b0591	YbdA MFS transporter	0.8	1.3	1.7	2.0	1.4	0.0	1.7	Suf Cluster
entC	b0593	isochorismate synthase, enterobactin specific	0.4	2.3	3.0	0.0	1.9	1.1	2.8	Fur Cluster
entE	b0594	enterobactin synthase multienzyme complex	0.8	1.7	1.8	1.6	1.5	0.8	1.8	Suf Cluster
ybdB	b0597	conserved hypothetical protein	0.0	0.8	1.5	1.0	0.8	-0.4	0.8	Fur Cluster
ahpC^*	b0605	component of alkylhydroperoxide reductase	1.3	1.2	0.6	1.1	1.1	1.4	0.5	
*ahpF**	b0606	component of alkylhydroperoxide reductase	1.7	0.7	0.6	0.7	0.9	1.5	0.3	
*fur^*	b0683	Fur transcriptional dual regulator	0.8	1.0	2.0	1.4	1.3	0.0	1.5	SoxRS Cluster
*fldA^*	b0684	oxidized flavodoxin 1	0.4	1.1	0.8	1.0	0.8	1.7	1.0	
nei	b0714	endonuclease VIII (VII?)	0.9	-0.3	-0.1	0.0	0.1	1.2	-0.3	
zitB	b0752	ZitB zinc CDF transporter	-0.7	0.2	0.7	1.2	0.4	-0.4	1.1	
gpmA	b0755	phosphoglycerate mutase 1	1.0	0.0	0.2	0.4	0.4	1.2	-0.4	
bioA	b0774	adenosylmethionine-8-amino-7-oxononanoate aminotransferase monomer	0.8	1.9	3.6	3.4	2.4	-0.5	3.6	SoxRS Cluster
bioB	b0775	biotin synthase monomer	0.7	2.0	3.4	3.6	2.4	-0.1	4.1	SoxRS Cluster
bioC	b0777	biotin biosynthesis; reaction prior to pimeloyl CoA	0.2	2.2	3.0	2.4	1.9	-0.5	2.2	SoxRS Cluster
bioD	b0778	dethiobiotin synthase monomer	0.4	1.8	2.9	2.2	1.8	-0.6	2.2	SoxRS Cluster
dps^*	b0812	stationary phase nucleoid protein that sequesters iron and protects DNA from damage	2.7	0.7	0.3	0.5	1.1	2.1	-0.2	
ybjK	b0846	putative DEOR-type transcriptional regulator	3.8	0.0	-0.3	0.0	1.8	3.0	0.0	
ycaJ	b0892	putative polynucleotide enzyme	1.3	0.2	-0.1	0.2	0.4	1.4	-0.4	
dmsA	b0894	component of dimethyl sulfoxide reductase	0.6	2.1	0.0	0.0	0.7	1.6	-0.2	
torD	b0998	chaperone protein for trimethylamine-N-oxide oxidoreductase I	1.6	-0.7	0.1	0.0	0.3	1.5	-0.8	
b1017	b1017	putative cytochrome	0.9	0.9	1.3	1.5	1.2	0.4	1.3	SoxRS Cluster
ycdO	b1018	conserved hypothetical protein	0.9	0.9	2.0	2.2	1.5	-1.2	1.9	SoxRS Cluster
ycdB	b1019	conserved hypothetical protein	0.7	0.8	1.6	1.9	1.2	0.8	1.0	SoxRS Cluster
rluE	b1135	23S rRNA pseudouridine synthase	1.6	0.4	0.1	0.7	0.7	2.1	0.0	
umuD	b1183	SOS mutagenesis; error-prone repair; processed to UmuD'; forms complex with UmuC	0.5	0.2	-0.2	-0.3	0.0	1.9	-0.7	
tonB	b1252	energy transducer; uptake of iron, cyanocobalimin; sensitivity to phages, colicins	0.0	0.9	0.0	1.7	0.8	-1.2	1.7	SoxRS Cluster
trpC	b1262	indole-3-glycerol phosphate synthase / phosphoribosylanthranilate isomerase	2.3	0.9	0.3	0.4	1.0	2.0	0.3	
topA	b1274	DNA topoisomerase type I, omega protein	-0.3	-0.3	0.2	0.3	0.0	-1.3	0.3	
lar	b1348	restriction alleviation and modification enhancement	-1.2	0.4	0.5	0.1	-0.1	-1.6	0.1	
ydaU	b1359	hypothetical protein	0.1	0.3	0.3	0.2	0.2	0.4	0.3	SoxRS Cluster
*ydbK^*	b1378	putative oxidoreductase, Fe-S subunit	0.9	1.9	1.6	2.2	1.7	0.5	2.1	
gapC_2	b1416	glyceraldehyde 3-phosphate dehydrogenase C, interrupted	0.3	0.3	-0.5	0.0	0.0	1.4	-0.5	
ydcZ	b1447	putative transport protein	1.5	1.0	-0.2	0.3	0.7	2.1	-0.2	
*yncE^*	b1452	putative receptor	0.7	2.0	1.9	1.1	1.4	0.0	0.6	SoxRS Cluster
nhoA	b1463	N-hydroxyarylamine O-acetyltransferase	1.8	-0.7	0.1	-0.8	0.1	2.1	-0.3	
ydeA	b1528	YdeA MFS transporter	0.6	0.6	-0.2	-0.2	0.2	1.6	0.1	
yleD	b1533	O-acetylserine/cysteine export protein	1.0	0.7	-0.4	0.6	0.5	2.1	-0.4	
tus	b1610	DNA-binding protein; inhibition of replication at Ter sites	1.6	2.8	1.9	0.8	1.8	0.8	2.3	
nemA	b1650	N-ethylmaleimide reductase	1.1	0.4	0.0	0.1	0.4	1.2	-0.2	
*sodB^*	b1656	superoxide dismutase (Fe)	1.5	0.2	0.0	-1.6	-0.1	0.4	-1.5	
sufE	b1679	sulfur acceptor that activates SufS cysteine desulfurase	0.0	1.3	1.5	1.9	1.3	-0.6	1.3	Fur Cluster
sufS	b1680	L-selenocysteine lyase (and L-cysteine desulfurase) monomer	0.5	1.7	2.0	1.8	1.5	0.1	2.1	Fur Cluster
sufD	b1681	component of SufB-SufC-SufD cysteine desulfurase (SufS) activator complex	1.0	1.7	2.0	1.8	1.6	1.3	1.0	Suf Cluster
*sufC**	b1682	ATPase component of SufB-SufC-SufD cysteine desulfurase (SufS) activator complex	1.4	2.9	2.5	2.8	2.4	1.5	1.7	Suf Cluster
*sufB**	b1683	component of SufB-SufC-SufD cysteine desulfurase (SufS) activator complex	1.9	2.1	1.9	2.6	2.1	1.7	0.8	Suf Cluster
katE	b1732	hydroperoxidase II	0.7	1.0	0.0	-0.4	0.3	2.0	-0.8	
xthA	b1749	exonuclease III	0.2	0.3	-0.2	0.4	0.2	-0.9	-0.1	SoxRS Cluster

Gene Name	Gene ID	Gene Product	Log Ratio (O2:AIR)							Cluster
			Block A				Block A Aeration vs. Time ANOVA	All Blocks		
			10'	30'	60'	90'		10'	60'	
yeaM	b1790	putative ARAC-type regulatory protein	1.5	-0.6	-0.4	-0.2	0.1	2.0	-0.8	
ftn	b1905	cytoplasmic ferritin, an iron storage protein)	0.1	-1.3	-3.3	-2.8	-1.8	1.3	-2.4	
b1964	b1964	putative outer membrane protein	1.9	0.2	-2.5	-0.2	-0.2	2.6	-0.1	
yedY	b1971	putative reductase	1.1	1.5	1.2	1.1	1.2	0.4	0.8	Suf Cluster
hisC	b2021	histidine-phosphate aminotransferase	-0.4	-0.2	-0.6	-0.4	-0.4	0.0	-2.4	
rfbA	b2039	dTDP-glucose pyrophosphorylase	-2.0	-0.4	-0.3	0.0	-0.7	-2.5	-0.3	
wcaF	b2054	putative transferase	0.9	3.1	3.5	2.8	2.6	0.6	2.3	Fur Cluster
gatC	b2092	galactitol PTS permease	0.0	-1.3	-1.6	-0.5	-0.9	0.7	-1.6	
cirA	b2155	outer membrane receptor for iron-regulated colicin I receptor; porin; requires tonB gene product	-0.9	0.9	1.2	1.1	0.6	-0.4	1.0	Fur Cluster
nfo	b2159	endonuclease IV	0.7	1.6	1.4	1.3	1.3	-0.3	1.3	SoxRS Cluster
yejG	b2181	conserved hypothetical protein	2.2	0.6	0.2	0.4	0.9	2.4	-0.3	
sseB	b2522	overproduction causes enhanced serine sensitivity	1.6	0.5	0.2	0.7	0.7	2.1	-0.1	
*pepB^*	b2523	putative peptidase	0.7	1.5	1.4	1.4	1.2	0.2	2.4	SoxRS Cluster
fdx	b2525	oxidized ferredoxin	2.7	1.6	1.0	1.0	1.6	1.5	1.1	
hscA	b2526	heat shock protein, chaperone, member of Hsp70 protein family	1.2	1.2	1.8	1.6	1.4	0.6	1.8	SoxRS Cluster
hscB	b2527	Hsc20 co-chaperone that acts with Hsc66 in IscU iron-sulfur cluster assembly	3.3	1.6	1.1	1.8	1.9	3.3	1.0	
iscA	b2528	putative regulator	1.1	1.6	1.6	2.4	1.7	1.9	2.0	
iscU	b2529	scaffold protein involved in iron-sulfur cluster assembly	0.6	1.5	1.8	1.9	1.4	1.7	2.2	
iscS	b2530	cysteine desulfurase	2.0	1.9	2.0	2.2	2.0	2.4	1.9	SoxRS Cluster
iscR	b2531	IscR transcriptional regulator	2.2	1.8	2.0	1.8	1.9	1.7	2.2	SoxRS Cluster
*trxC**	b2582	oxidized thioredoxin 2	1.0	0.3	0.1	0.5	0.5	1.2	0.1	
tyrA	b2600	chorismate mutase / prephenate dehydrogenase	1.3	1.1	-0.2	0.2	0.6	2.4	-0.1	
grpE	b2614	phage lambda replication; host DNA synthesis; heat shock protein; protein repair	2.1	0.5	0.6	0.7	1.0	2.1	-0.7	
yqaE	b2666	hypothetical protein	1.3	0.2	0.0	1.0	0.6	3.2	-1.1	
nrdH	b2673	glutaredoxin-like protein; hydrogen donor	0.5	3.3	4.3	5.0	3.3	0.6	3.9	Fur Cluster
nrdI	b2674	stimulates ribonucleotide reduction	-0.1	3.3	2.9	3.4	2.4	-0.1	3.2	Fur Cluster
nrdE	b2675	ribonucleoside-diphosphate reductase II	0.2	2.3	3.2	3.6	2.3	-0.7	3.9	Fur Cluster
nrdF	b2676	ribonucleoside-diphosphate reductase II	-0.1	2.1	2.7	2.5	1.8	0.5	3.3	Fur Cluster
emrA	b2685	accessory transport protein	-1.1	-0.8	-0.2	-1.3	-0.9	1.4	-0.2	
hypE	b2730	plays structural role in maturation of all 3 hydrogenases	0.0	0.0	0.9	1.1	0.5	-2.0	0.6	
eno	b2779	enolase	1.3	0.3	0.4	0.3	0.6	1.6	-0.2	
lysA	b2838	diaminopimelate decarboxylase	1.6	0.5	0.4	0.1	0.7	1.7	0.4	
mutY	b2961	adenine glycosylase; G.C --> T.A transversions	0.2	0.4	0.5	0.4	0.4	-0.3	0.6	SoxRS Cluster
*glgS**	b3049	glycogen biosynthesis, rpoS dependent	-1.7	0.0	-2.6	0.0	-2.4	-0.8	-1.7	
mdh	b3236	malate dehydrogenase	0.9	0.0	-0.6	1.1	0.3	1.7	-0.2	
yhfK	b3358	hypothetical protein	0.3	0.2	0.2	0.5	0.3	-0.1	0.1	SoxRS Cluster
ftsX	b3462	FtsE/FtsX ABC transporter	0.8	0.4	-0.8	0.4	0.2	1.4	-0.5	
arsR	b3501	ArsR transcriptional regulator	0.0	0.3	-0.6	0.6	0.1	1.4	-0.7	
yiaD	b3552	putative outer membrane protein	-0.3	0.5	0.2	1.0	0.3	-0.7	0.1	SoxRS Cluster
yicL	b3660	putative permease transporter	2.3	0.1	0.1	-0.1	0.6	1.6	-0.1	
ilvN	b3670	acetohydroxybutanoate synthase I / acetolactate synthase I	0.7	2.3	1.3	2.4	1.7	0.1	1.9	
ilvB	b3671	acetohydroxybutanoate synthase I / acetolactate synthase I	0.7	1.5	2.2	2.2	1.7	0.7	3.1	SoxRS Cluster
tnaB	b3709	TnaB tryptophan ArAAP transporter	0.9	-1.1	0.4	-0.5	-0.1	1.7	0.0	
rbsR	b3753	RbsR-ribose	-5.1	0.0	-0.2	0.0	-2.7	-5.1	-0.2	
ilvG_1	b3767	IlvG_1	0.6	0.2	2.2	2.2	1.3	-0.7	2.6	
ilvM	b3769	acetohydroxybutanoate synthase II / acetolactate synthase II	0.0	1.3	1.0	1.7	1.0	-0.5	1.8	SoxRS Cluster
ilvE	b3770	branched chain amino acid aminotransferase	0.0	0.6	1.6	1.7	1.0	-0.8	1.6	SoxRS Cluster
ilvD	b3771	dihydroxy-acid dehydratase	-0.5	0.4	1.2	1.5	0.7	-0.5	2.0	SoxRS Cluster
dapF	b3809	diaminopimelate epimerase	0.2	0.5	2.1	0.0	0.9	0.2	2.1	SoxRS Cluster
*sodA^*	b3908	superoxide dismutase (Mn)	1.2	2.5	0.0	3.6	2.1	-0.5	2.9	SoxRS Cluster
fpr^*	b3924	flavodoxin NADP+ reductase	1.2	2.7	2.5	3.2	2.4	1.3	2.7	
ftsN	b3933	essential cell division protein	-0.3	0.3	0.6	0.0	0.3	-1.5	0.3	
yijJ	b3948	hypothetical protein	0.6	0.8	1.7	0.8	1.0	0.5	0.9	SoxRS Cluster
thiH	b3990	thiH protein	0.1	1.0	1.1	1.9	1.1	0.8	1.4	Fur Cluster
thiG	b3991	thiG protein	0.0	1.7	3.5	4.0	2.3	-0.5	3.5	SoxRS Cluster
thiF	b3992	ThiF protein / thiF protein	-0.1	1.8	3.3	3.8	2.2	-0.6	3.2	SoxRS Cluster
thiE	b3993	thiamin phosphate synthase	-0.6	2.0	0.0	3.8	1.8	0.0	3.4	SoxRS Cluster
thiC	b3994	thiC protein	0.0	2.0	3.5	3.8	2.3	0.5	3.9	SoxRS Cluster
malF	b4033	maltose ABC transporter	-1.3	-2.0	-0.7	-0.3	-1.1	-1.7	-0.9	
*lamB^*	b4036	phage lambda receptor protein; maltose high-affinity receptor	0.3	-1.4	-1.6	-2.1	-1.2	0.0	-1.7	
dgkA	b4042	diacylglycerol kinase	1.7	0.1	-1.2	-0.5	0.0	1.9	-0.8	
soxS^*	b4062	SoxS transcriptional activator	2.9	3.6	3.0	3.1	3.1	1.7	3.1	SoxRS Cluster
frdD	b4151	fumarate reductase membrane protein	0.6	-0.9	-0.3	-0.2	-0.2	1.2	-0.4	
ytfA	b4205	hypothetical protein	-1.1	0.5	0.1	0.7	0.1	3.0	0.2	
ytfJ	b4216	conserved hypothetical protein with possible extracytoplasmic function	0.8	0.1	-0.3	0.4	0.2	2.2	-0.9	
yjgN	b4257	putative membrane protein possible involved in transport	-1.4	1.5	0.0	-1.2	-0.3	3.0	0.4	
hsdM	b4349	host modification; DNA methylase M	-0.9	-0.1	-0.2	0.3	-0.2	-1.4	-0.4	
fhuF	b4367	acts in reduction of ferrioxamine B iron	0.4	2.2	1.5	1.6	1.4	1.5	0.7	Fur Cluster
yjjU	b4377	putative transcriptional regulator	0.1	0.1	0.5	0.5	0.3	-0.4	0.2	SoxRS Cluster

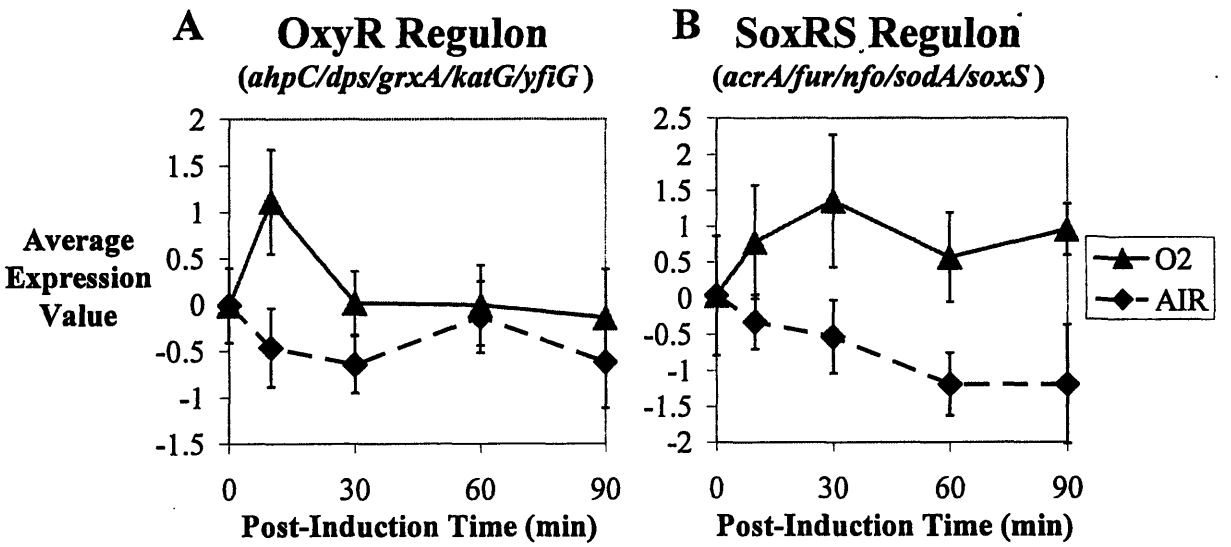


Figure 5.7: Hyperoxic Stress Responses in AIR and O2 Cultures – Block A
Expression values were averaged across the listed genes in **A)** the OxyR regulon and **B)** the SoxRS regulon. Scaling was performed by minimizing the sum of the variances across all samples. Error bars represent the standard deviation in the expression values. Samples from the N2 culture were not included in these plots because the variation between genes was much larger than that observed for the O2 and AIR cultures.

5.4 Chaperones and Proteases

In an effort to clarify the oxygen dependent degradation of α_1AT, expression data for genes known to encode chaperones and proteases were examined. While it has been shown that that the heat-shock protein ClpP plays a role in the degradation of α_1AT (Laska 2000), it is not the only protease involved. Oxygen-dependent expression of either *clpP* or other genes that code for proteases or folding chaperones may help to explain why α_1AT degradation is oxygen dependent.

Throughout this section, it is important to remember that significant changes in gene expression observed in these experiments do not necessarily lead to significant changes in protein levels over the course of the experiment (roughly one doubling time for the O2 and AIR cultures). Nevertheless, this analysis was critical to understanding how the cultures responded to changes in aeration.

5.4.1 Clp Proteases

E. coli contains three Clp complexes: ClpAP, ClpXP, and ClpYQ (HslUV). Each of these complexes consists of a central protease (ClpP or ClpQ) and an ATP-dependent chaperone (ClpA, ClpX, or ClpY). These chaperones recognize various motifs in unfolded proteins and bind them. These unfolded proteins can then be transferred to the protease component for degradation (Hoskins *et al.* 2002 for review). Except for the gene *clpA*, all of these genes are known to be regulated by the heat-shock sigma factor σ^{32} and, therefore, can be stimulated by protein overexpression. In other bacteria, mutations in genes encoding Clp ATPases have been found to cause sensitivity to oxidative stress (Rouquette *et al.* 1996; Ekaza *et al.* 2001), suggesting that these proteins may be responsible for the oxygen dependence of α_1AT degradation.

Because ClpP is known to play a role in α_1AT degradataion (Laska 2000), expression profiles of *clpP* as well as genes encoding other proteins in ClpP complexes were examined (Figure 5.8). As expected of genes encoding heat-shock proteases, *clpP* showed increased expression following induction. The genes *clpA* and *clpP* showed dramatic increases in expression in the N2 culture, particularly at 30 and 60 min. In general, the AIR and O2 cultures showed expression values that were similar to one another throughout the experiment. Expression of *clpX* differed between the AIR and O2 cultures at 10 min, and expression of *clpA* differed, if only slightly, throughout the experiment. However, inclusion of data from Blocks B and C revealed that none of these three genes showed significant and reproducible differential expression between the AIR and O2 cultures.

5.4.2 Recombinant α1-Antitrypsin Degradation in a ClpA$^-$ Mutant

Based on the above expression profiles, a hypothesis for oxygen-dependent degradation was developed. Although *clpP* expression did not show differential expression between the AIR and O2 cultures, the slightly increased expression of *clpA* in oxygen may have led to higher levels of the ClpAP complex in O2 cultures, and ultimately to increased degradation. To test this hypothesis, pulse-chase experiments were performed to quantify degradation of α_1AT in a ClpA$^-$ mutant with a BL21 background (Figure 5.9). When compared with BL21 cultures (Figure 5.1), the aerobic ClpA$^-$ cultures showed rate constants for proteolysis (k_p) and proteolysis to folding ratios (r_p/r_f) that were either lower or equivalent (Figure 5.10). These results suggested that

ClpA plays a role in α_1AT folding and proteolysis since the rate constants of both decreased in the ClpA$^-$ mutant. However, in the N2 ClpA$^-$ cultures, proteolysis was unexpectedly found to proceed with a higher rate constant. The increased degradation of α_1AT in the N2 culture suggested that ClpA plays a protective role in α_1AT degradation in hypoxic cultures.

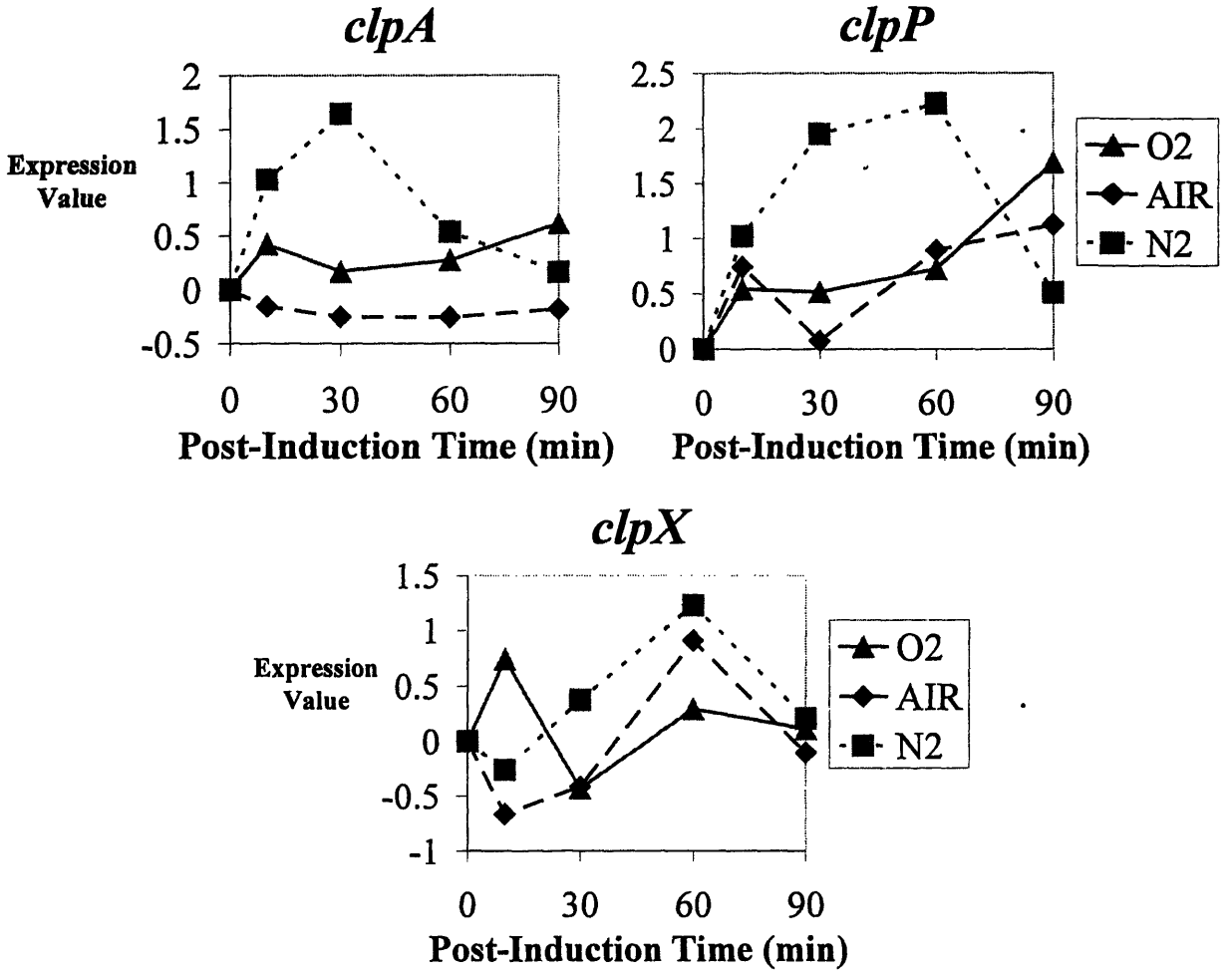


Figure 5.8: Expression Profiles of Clp Proteins
Data from the Block-A experiment. Cultures were simultaneously induced and exposed to different aeration environments at 0 min.

Specific activity measurements appeared to be inconsistent with degradation results. For example, despite decreased degradation in the aerobic ClpA$^-$ cultures, the specific activities of α_1AT showed no significant change (Figure 5.11). Moreover, pulse-chase data for N2 ClpA$^-$ cultures showed increased degradation, but specific activities from these cultures were larger than those from BL21 cultures. One explanation for these inconsistent results is that degradation rates 60 min after induction may not be representative of the degradation that occurs during the

entire 90 min induction period. As an example, consider expression of *clpA* in BL21 N2 cultures, which appears to peak around 30 min (Figure 5.8). If ClpA were largely responsible for α_1AT degradation, a pulse-chase analysis performed at 30 min may have revealed decreased expression in the ClpA$^-$ mutant, which would have been consistent with overall yields.

While some of the results with this ClpA$^-$ mutant appear promising, this set of experiments was insufficient to elucidate the role of ClpA in oxygen dependent α_1AT degradation.

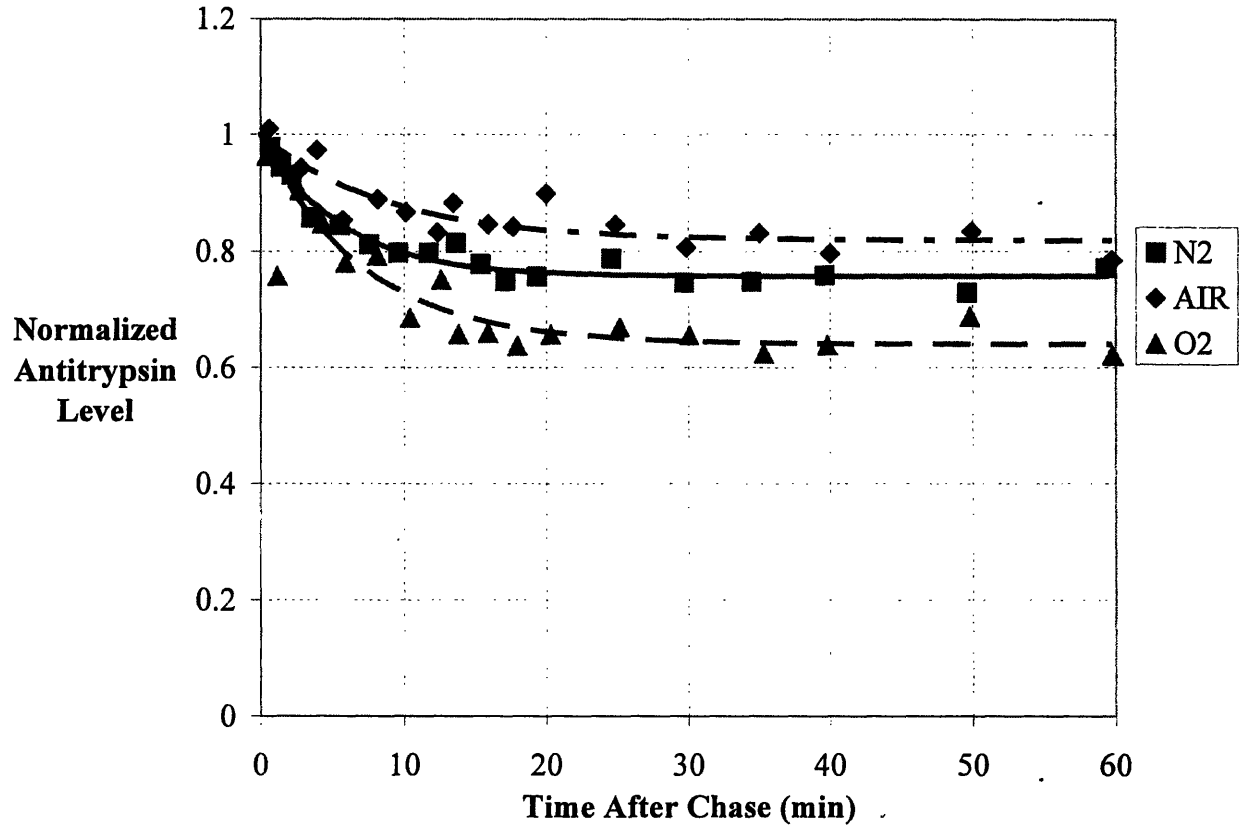


Figure 5.9: Degradation Profile of α1-Antitrypsin in a ClpA$^-$ Mutant
A culture of a ClpA$^-$ mutant strain at OD_{600} of 0.7 was simultaneously split and induced in three different aeration environments: pure nitrogen, air, and pure oxygen. After 60 min of induction, newly synthesized protein was pulse-chase labeled with ^{35}S-methionine. Degradation of recombinant α_1AT was monitored for the next 60 min.

5.4.3 Oxygen-Dependent Proteases and Chaperones

Using a list of 37 chaperones and 48 proteases obtained from the Swiss-Prot Protein Database (Gasteiger *et al.* 2003), the list of differentially expressed O2-AIR genes was queried. The genes that were found to appear on both lists are discussed below:

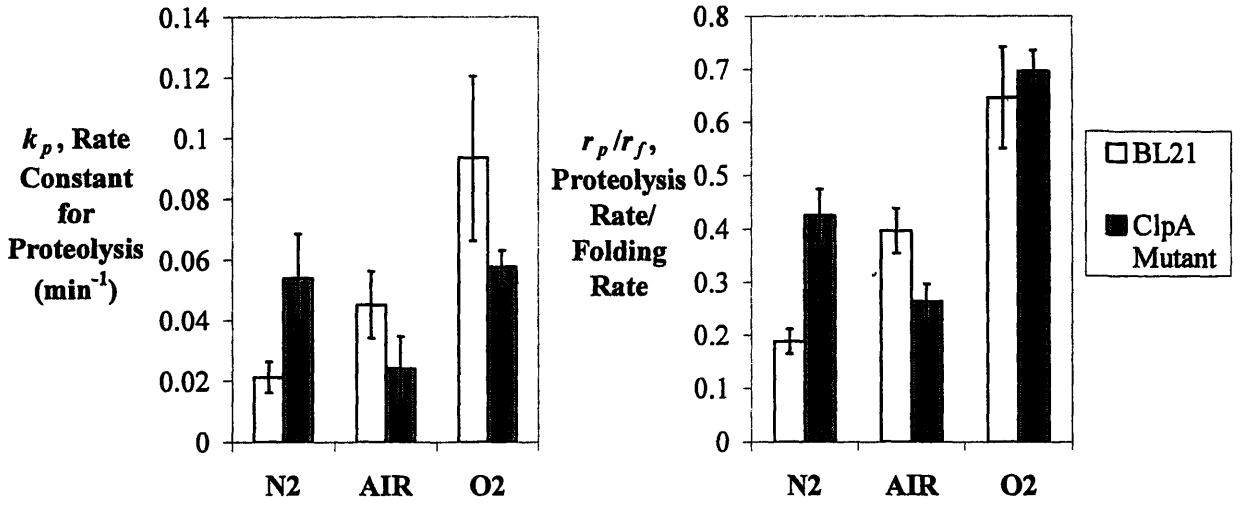


Figure 5.10: Kinetic Parameters for α1-Antitrypsin Degradation in BL21 and ClpA⁻ Cultures

Model parameters for data in Figure 5.1 and Figure 5.9. k_p is the pseudo-first-order rate constant for α_1AT proteolysis and r_p/r_f is the ratio of the rate of proteolysis and the rate of folding. Parameters and confidence intervals were calculated as described in Section 3.8.3.

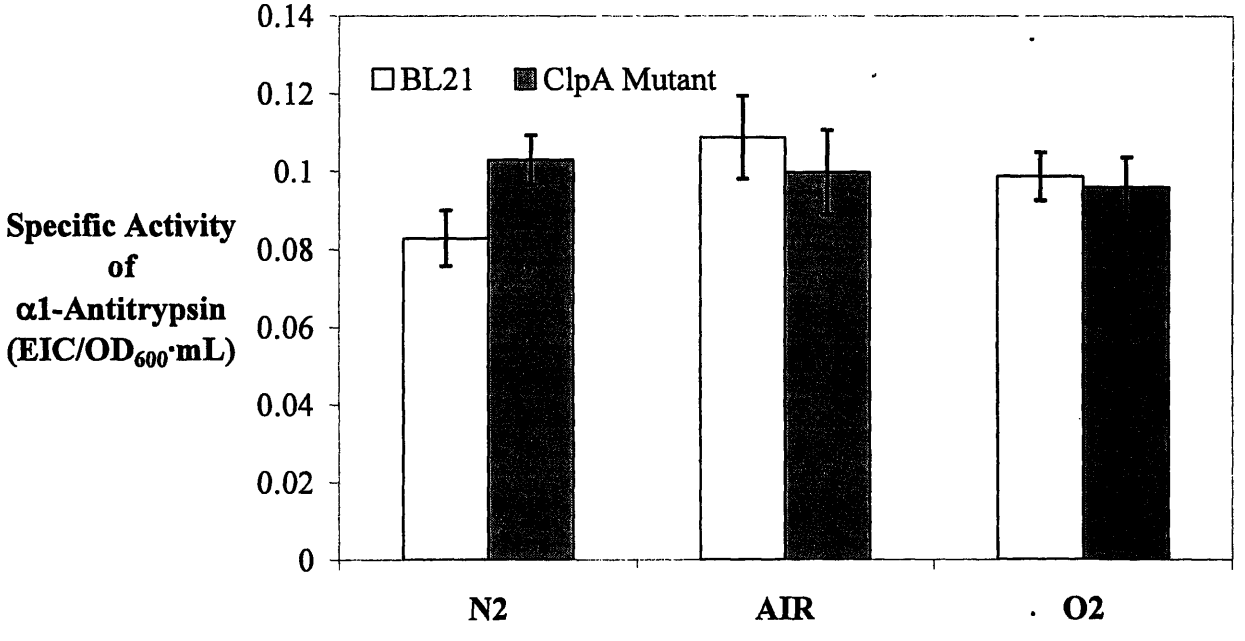


Figure 5.11: Specific Activities of α1-Antitrypsin from BL21 and ClpA⁻ Cultures

Cultures of *E. coli* BL21 and a ClpA-deficient mutant with a BL21 background were grown, split into defined aeration environments (pure nitrogen, air, and pure oxygen), and induced to produce recombinant α_1AT for 90 min. Activity of the recombinant protein was measured as Elastase Inhibitory Capacity (EIC) and scaled by the OD_{600} and volume of the cultures at harvest.

- The genes *hscA* and *hscB* (*yfhE*) were both found to exhibit significantly increased expression in the O2 culture, based on data from all three replicate experiments. These two genes encode chaperones that are specific to proteins containing iron-sulfur clusters and assist in assembly of those clusters. These chaperones are likely induced in response to superoxide exposure and are not expected to interact with α_1AT. These chaperones are discussed further in Section 5.5.
- The gene *umuD* was found to have increased expression in the O2 culture at 10 min, based on data from all three replicate experiments. This gene codes for a DNA polymerase subunit that serves to repair DNA damage. Although the UmuD protein is listed as a peptidase, it only performs this function on itself to generate the active UmuD' protein. Nevertheless, it is interesting to see that its expression pattern resembles those of genes in the OxyR regulon. Interestingly, UmuD has been found to present its cleaved partner UmuD' to the ClpXP protease for degradation (Neher *et al.* 2003). While UmuD may not have protease activity, it is a factor in a proteolysis pathway. A mechanism whereby UmuD presents α_1AT to the ClpXP complex would be purely speculation at this point.
- The gene *torD* codes for a chaperone specific to the TorA protein, which is involved in anaerobic respiration. TorA is a molybdoprotein, but has iron cofactors as well. Recently, TorD has been found to play a role in maturation of TorA, prior to addition of its molybdenum cofactor (Ilbert *et al.* 2003). The oxygen-dependence of *torD* might indicate that the molybdenum cofactor is oxygen sensitive. The specificity of this chaperone makes interaction with α_1AT unlikely.
- The gene *grpE* showed significantly increased expression in the O2 culture 10 min after induction, according to all three replicate experiments (Figure 5.12). The product of this gene, along with the chaperones DnaK and DnaJ, binds to σ^{32} and inactivates the heat-shock response. This complex may also bind to misfolded proteins and direct them to proteases.

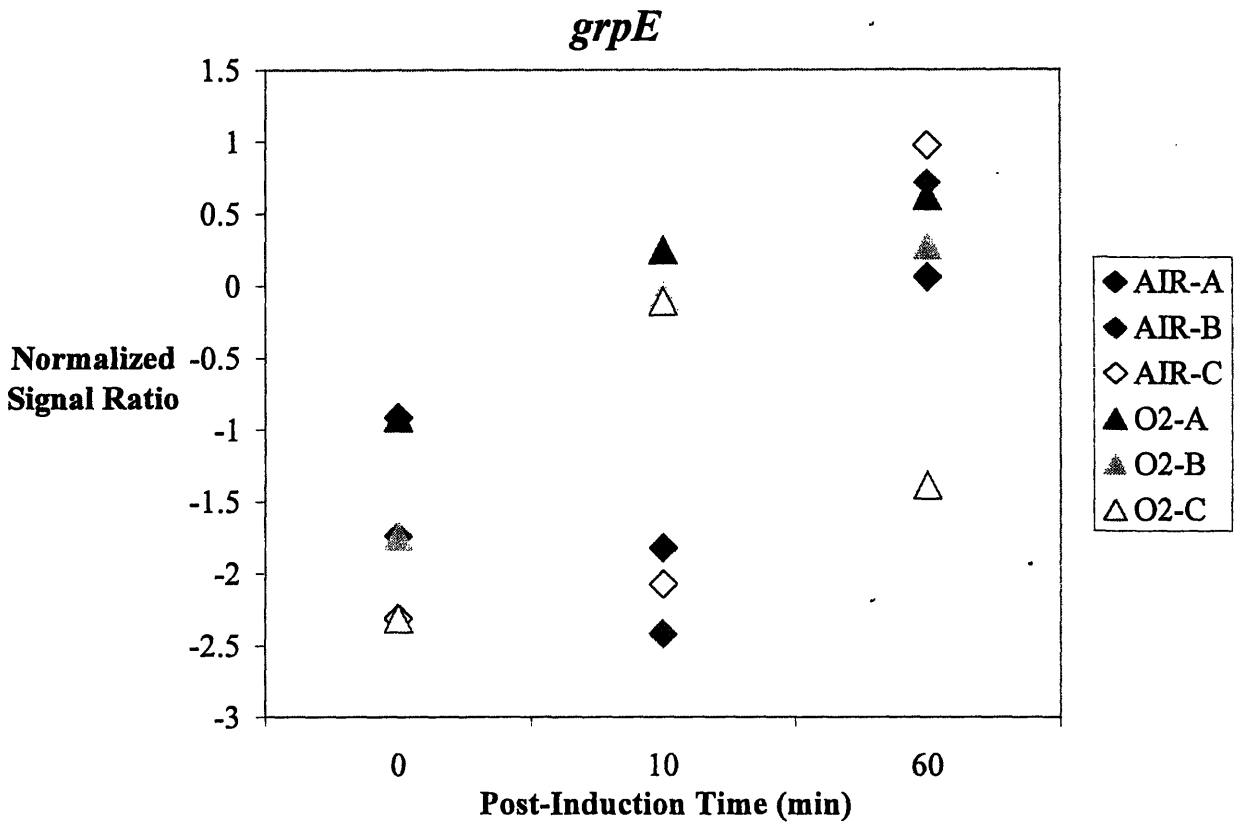


Figure 5.12: Oxygen-Dependent Expression of *grpE*
Normalized signal ratio data are plotted for the AIR and O2 cultures from each of the repeated experiments (A, B, and C). Cultures were simultaneously induced and exposed to different aeration environments at 0 min. Results from each experiment showed consistent O2-AIR differential expression at 10 min.

The chaperones HscA, HscB, and TorD and the peptidase UmuD are all specific to particular classes of proteins, none of which fit the description of α_1AT. It is unlikely that any of them would interact with a recombinant protein to either fold or degrade it. In contrast, GrpE is a global chaperone, which may very well interact with α_1AT.

5.4.4 Oxygen-Dependent Expression of *grpE*

Because DnaK is a high abundance protein (VanBogelen *et al.* 1987), it is possible that GrpE is the limiting component of the DnaK/DnaJ/GrpE chaperone complex. Therefore, levels of this complex may increase upon an increase in transcription of *grpE*, possibly resulting in increased proteolysis of unfolded proteins. The DnaK/DnaJ/GrpE chaperone complex inactivates σ^{32} by increasing its susceptibility to degradation by FtsH. DnaJ, DnaK, and GrpE were also found to be essential for proteolytic degradation of a heterologous protein by both Lon

and ClpAP proteases (Sherman and Goldberg 1992; Huang *et al.* 2001). This chaperone complex likely functions, not by shuttling proteins to a particular protease, but rather by maintaining proteins in a soluble state where they will be more easily degraded. Increased expression of *grpE* in O2 cultures only at 10 min suggests regulation similar to that of OxyR-regulated genes; however, no such link has been reported. If GrpE is stable, this increase at 10 min may certainly result in increased protein levels at 60 min, the point at which degradation rates were measured. Although this was not explored further in this work, oxygen dependent expression of *grpE* may provide a link between hyperoxic and heat-shock stress responses, and may contribute to the oxygen dependence of α_1AT degradation.

5.5 Oxygen-Dependent Genes

Although the oxygen-dependent expression of *grpE* was promising, it was only increased in O2 cultures 10 min after induction. It was also of interest to find genes that showed an oxygen-dependent expression pattern at all times. Examination of superoxide stress response genes showed that the *soxS* gene was strongly oxygen dependent, with low expression in N2 cultures and high expression in O2 cultures. Furthermore, comparison of the hyperoxic stress responses suggested that superoxide elicited a sustained response. In order to better understand oxygen dependent degradation of α_1AT, the effects of superoxide were explored by examining genes with similar expression profiles.

5.5.1 Superoxide Stress Response Genes

In addition to *soxS* itself, several genes in Table 5.2 are part of the SoxRS regulon. This regulon is activated by superoxide and the SoxR protein is the sensor. Some of the genes that show oxygen dependence in induced cultures include *sodA*, *fur*, *acrA*, and *nfo*. The gene *sodA* encodes the Mn-binding superoxide dismutase, a protein with a major role in superoxide defense. The gene *fur* encodes a transcriptional regulator that represses iron uptake. *acrA* codes for a multidrug resistance protein, which acts in concert with ArcB and TolC to export antimicrobial agents from the cell. Finally, the gene *nfo* encodes endonuclease IV, which repairs DNA at sites where glycosylases (such as MutY below) have removed oxidized or otherwise altered bases.

5.5.2 Oxidation of Iron-Sulfur Clusters

It is well known that superoxide oxidizes iron-sulfur clusters. A search of the Swiss-Prot database revealed 103 *E. coli* proteins with known or putative iron-sulfur clusters. Thus, there is a large potential for damage to the cell by superoxide. Eight genes in Table 5.2 encode iron-sulfur proteins—*bioB*, *ilvD*, *leuC*, *mutY*, *fdx*, *fhuF*, *dmsA*, and *ydbK*.

- Biotin synthase, encoded by *bioB*, is a homodimeric protein consisting of two [2Fe-2S] and two [4Fe-4S] clusters (Ugulava *et al.* 2001; Ugulava *et al.* 2002). One of these clusters is likely the source of the sulfur atom in biotin (Bui *et al.* 1998; Begley *et al.* 1999; Gibson *et al.* 1999; Jameson *et al.* 2004); therefore, this iron-sulfur cluster must be continuously regenerated.
- The enzyme dihydroxy-acid dehydratase, encoded by *ilvD*, generates a precursor for synthesis of all three of the branched-chain amino acids. This enzyme loses its activity upon oxidation of its [4Fe-4S] cluster (Flint *et al.* 1993).
- The enzyme isopropylmalate isomerase, encoded by *leuC*, catalyzes a step in the leucine biosynthesis pathway. Although no iron-sulfur cluster has been observed for this protein, this enzyme is assumed to have a [4Fe-4S] cluster due to homology with *E. coli* aconitase (Gasteiger *et al.* 2003).
- The gene *mutY* appears to play a role in oxidative defense, as it codes for the protein adenine glycosylase, which is involved in repair of oxidatively damaged A:G base pairs. This enzyme contains a [4Fe-4S] cluster, which is not necessary for catalytic activity, but is critical to positioning on DNA (Porello *et al.* 1998). Because of the presence of an iron-sulfur cluster and its role in repair of oxidative damage, oxygen-dependent transcription of the *mutY* gene certainly provides an advantage to the cell. This enzyme may work in concert with endonuclease IV (Nfo) to repair bases removed by MutY.
- In general, ferredoxins are iron-sulfur proteins involved in electron transfer reactions. *E. coli* ferredoxin, encoded by the gene *fdx*, has been found to play a role in biosynthesis of iron-sulfur clusters *in vivo*. This ferredoxin contains one [2Fe-2S] cluster (Kakuta *et al.* 2001).
- In addition to containing a [2Fe-2S] cluster, FhuF is also involved in increasing iron availability in the cell. The expression of *fhuF* is extremely sensitive to decreases in iron availability and responds to the Fur regulatory protein. As a member of the Fur Cluster

(Figure 5.6) *fhuF* shows increased expression in oxygen at 30, 60, and 90 min. Interestingly, *fhuF* was found to be repressed by OxyR (Zheng, Wang, Doan *et al.* 2001), which provides further evidence that the peroxide stress response was not active during later time points in the experiment.

- Dimethyl sulfoxide (DMSO) reductase is an enzyme that catalyzes the transfer of electrons to DMSO in the absence of oxygen. While its activity is not essential to the cell in highly aerobic conditions, expression of one of its genes, *dmsA*, was activated after 10 min in pure oxygen. DmsA has a putative [4Fe-4S] cluster.
- The gene *ydbK* is a putative oxidoreductase that appears to perform the same function as pyruvate dehydrogenase. It is similar to [4Fe-4S] ferredoxins and its regulation is unknown.

Most likely, these genes likely showed differential expression for the same reason: superoxide oxidized an iron-sulfur cluster in the protein product, thereby causing lost enzyme activity. Transcription of the gene was increased in an attempt to recover that activity.

Although their genes do not appear in Table 5.2, the enzymes fumarase and aconitase, both of which are involved in the TCA cycle, contain iron-sulfur clusters that make them sensitive to oxygen. The exception is *fumC*, which codes for a fumarase that does not contain an iron-sulfur cluster and is therefore resistant to superoxide. Interestingly, *fumC* transcription is known to be activated by the SoxS protein. Although *fumC* did not cluster with the oxygen dependent genes, its expression showed a slight increase in the O2 culture at 10 and 30 min, similar to that observed for *soxS*. In the case of aconitase, neither *acnA* nor *acnC* showed significant differential expression between the O2 and AIR cultures.

5.5.3 Pathways Dependent upon Iron-Sulfur Clusters

The list of O2-AIR differentially expressed genes (Table 5.2) included genes involved in three metabolic pathways: biotin biosynthesis, branched-chain amino acid biosynthesis, and thiamin biosynthesis. The first two of these pathways contain iron-sulfur proteins as discussed in the previous section. The appearance of all or most of the genes involved in these pathways is surprising. One explanation is that oxidation of the iron-sulfur cluster inactivated one step of the pathway, resulting in decreased levels of the metabolic product. In response, the expression

levels of all genes in the pathway were indiscriminately increased in order to alleviate the bottleneck.

This model is consistent with our knowledge of transcriptional regulation for these pathways. Transcription of genes involved in biotin biosynthesis is regulated by the biotin repressor protein (BirA), which requires biotinyl-5'-adenylate as a cofactor (Eisenberg *et al.* 1982). If biotin levels drop due to inactivation of BioB by superoxide, the levels of the cofactor would also drop, resulting in lower repression, *i.e.* higher transcription, of all genes in the *bio* regulon. In an analogous mechanism, the leucine-responsive regulatory protein (Lrp), when bound to leucine, can represses synthesis of the *ilvL(G1)(G2)MEDA* transcription unit. Inversely, if leucine levels drop, due to inactivation of the Fe-S clusters in IlvD, LeuC, or both, expression of genes in this transcription unit becomes activated. In the absence of leucine, Lrp can also act as an activator for the *leuABC* transcription unit. Increased transcription of the biotin and branched-chain amino acid pathways is easily explained; however, the appearance of four genes from the thiamin biosynthesis pathway in this list is not entirely clear.

Since thiamin may serve to scavenge reactive oxygen species inside the cell (Jung and Kim 2003), one explanation for its oxygen dependent expression may be that thiamin biosynthesis is activated by the superoxide stress response to generate thiamin to act as a superoxide sink. Although the presence of an iron-sulfur cluster in this pathway has been predicted, no evidence for this claim exists. In fact, there is evidence to the contrary (Mueller *et al.* 2001). Current understanding of the pathway indicates that the ThiI protein contains a catalytic disulfide bond. As with any disulfide bond, it has the potential to be strongly affected by the redox state of the cell. Highly oxidative conditions could potentially inactivate this protein by preventing reduction of the disulfide bond. Although it may not have an iron-sulfur cluster, ThiI has been shown to interact with a protein from the iron-sulfur repair locus, IscS, in order to produce thiamin (Lauhon and Kambampati 2000). If ThiI must compete with iron-sulfur proteins for copies of IscS, exposure to superoxide would shift the balance in favor of the iron-sulfur proteins, which would ultimately lead to lower production of thiamin. Thiamin and oxidative stress are known to be linked in several ways, but competition for IscS seems to be the best explanation for superoxide-related expression of thiamin biosynthesis genes.

5.5.4 Repair of Iron-Sulfur Clusters

The Isc system and the Suf system both show increased expression in response to induction in high oxygen conditions. Except for *iscX*, *yfhQ*, and *sufA* all of the genes in these two systems showed increased expression in the O2 culture, compared with the AIR culture. This expression difference was particularly evident at 30-, 60-, and 90-min time points. While the *suf* operon has recently been found to be regulated by OxyR (Zheng, Wang, Templeton *et al.* 2001), the data presented here suggest that the Suf and Isc systems may also be regulated by SoxRS. These data indicate that, as expected, these systems are stimulated by superoxide stress and play a role in the defense against superoxide.

5.5.5 Iron Uptake Genes

Another interesting observation from the list of O2-AIR differentially expressed genes (Table 5.2) is that several are involved in iron uptake. Iron is known to contribute to both the damaging effects of reactive oxygen species (via the Fenton reaction), as well as *E. coli*'s defense against these species (in enzymes such as catalase and superoxide dismutase).

As a member of both the SoxRS regulon and the OxyR regulon (Zheng *et al.* 1999), the gene *fur* showed increased expression between the O2 and AIR cultures. This observation is consistent with results from another microarray analysis (Pomposiello *et al.* 2001). With increased expression of this repressor protein, iron metabolism genes might be expected to show decreased expression. However, just the opposite was observed.

Of the 40 genes involved in the Fur regulon, 18 were found to be induced in the O2 culture (Table 5.2). Enterobactin is a ferrisiderophore, an iron transport molecule, which binds extracellular iron(III). Genes involved in enterobactin biosynthesis include *entC/E/D/F*. The genes *fepA*, *cirA*, and *fhuA* encode outer membrane receptors for enterobactin, while *tonB* encodes a periplasm-spanning gate for these receptors (Earhart 1996). The gene *fepC* encodes one subunit of the ferric-enterobactin ABC transporter, which transports the complex into the cytoplasm. Once in the cytoplasm, enterobactin esterase, the product of the *fes* gene, helps to digest enterobactin. This protein is necessary for removal of iron from enterobactin. All of these Fur-regulated genes showed increased expression upon induction in oxygen.

The increased expression of iron uptake genes might be explained by differing growth media. Cultures in previous microarray studies of hyperoxic stress were grown in LB medium

(Pomposiello *et al.* 2001; Zheng, Wang, Templeton *et al.* 2001), whereas cultures in this work were grown in minimal medium. This minimal medium contained 94 μM $FeCl_3$; therefore, conditions were not iron limiting. However, it is possible that these low iron levels might result in activation of the Fur regulon, despite increased expression of the *fur* gene. This is consistent with the model in which Fur acts as a transcriptional repressor only upon binding an iron cofactor (Hantke 2001). Iron limited conditions would result in derepression of the Fur regulon.

However, this model does not explain why iron uptake genes increase expression upon exposure to oxygen. The explanation of this result is further complicated by the observation that exposure to oxygen is known to increase the levels of free iron in the cell (Keyer and Imlay 1996; Srinivasan *et al.* 2000). According to these results, iron uptake genes might be expected to decrease upon oxygen exposure. However, just the opposite was observed. Therefore, the signal for stimulation of the Fur regulon by oxygen is complicated and is not simply based on limited iron availability. One explanation may lie in further understanding of the above results. According to (Keyer and Imlay 1996), the free iron that is released upon oxygen exposure is in the form iron(II). This shift in the overall redox state of free iron is likely the oxygen-dependent signal for increased expression of iron uptake genes. Whether this signal is transmitted through increased Fur levels is unknown.

Another major difference between this work and previous studies of the detrimental effects of iron (Imlay and Linn 1987; McCormick *et al.* 1998) is that those studies used SOD mutants, which have no defense against superoxide. These mutants, even when grown in air, are likely under more severe peroxide stress than the BL21 cultures grown in pure oxygen in this work, for the following reason. Of the two defense mechanisms against superoxide—SOD and iron-sulfur clusters—SOD generates fewer peroxide molecules stoichiometrically. Therefore, without SOD, the oxidation of iron-sulfur clusters becomes the major source of superoxide defense, generating not only iron(II), but peroxide as well. It is possible that iron is only damaging under these conditions because hydrogen peroxide, which participates in the damaging Fenton reaction, is also generated at high levels.

Based on the results from iron-sulfur cluster damage and repair, it appears that iron actually plays a protective role during superoxide stress. Its uptake is likely up-regulated in order to regenerate oxidized iron-sulfur clusters. It should be noted that genes involved in cysteine biosynthesis do not show significant oxygen-dependent expression. In fact, many of

these genes show identical expression between O2 and AIR cultures. Despite the activation of genes encoding cysteine desulfurases (*iscS* and *sufS*) in the O2 cultures, which certainly increase the demand for this amino acid, cysteine biosynthesis does not appear to be affected. Therefore, repair of iron-sulfur clusters is likely limited by iron, not by sulfur.

5.5.6 Summary of Oxygen-Dependent Genes

The differential expression of SoxRS-regulated genes between the O2 and AIR cultures demonstrated the widespread effects of superoxide on the *E. coli* cell and also led to insight as to how the aerobic cultures responded to this challenge. Oxygen exposure was found to activate the superoxide stress response. The main effect of superoxide appeared to be oxidation of iron-sulfur clusters, as indicated by oxygen dependent expression of genes encoding iron-sulfur proteins. In an effort to regenerate these clusters, the Isc and Suf repair systems were activated. Intracellular iron levels appeared to be the limiting component in this repair mechanism, because iron uptake was also stimulated by oxygen. The peroxide response was found to be activated within the first 10 min of oxygen exposure, but was less active at longer times.

Based on the results presented here and knowledge of the interactions between oxygen and iron (Liochev and Fridovich 1994), the following mechanism was developed to explain the above observations. Upon initial exposure to highly aerobic conditions, hydrogen peroxide was rapidly generated by oxidation of iron-sulfur clusters by superoxide. Both the released iron and peroxide participated in the Fenton reaction to generate damaging hydroxyl radicals and stimulated the OxyR-regulated stress response. This reaction eventually slowed as the pool of reduced iron-sulfur clusters became depleted, resulting in decreased generation of peroxide and an accumulation of superoxide. Induction of the iron(III) uptake regulon indicates that free iron may be more prevalent in its reduced form. Repair of the iron-sulfur clusters was limiting after long-term exposure, due to a lack of either free iron or effective reducing agents. This mechanism is consistent with the above observations and confirms that oxidation of iron-sulfur clusters is the main route of both superoxide toxicity and hydrogen peroxide formation during oxygen exposure.

Of most interest is whether iron-sulfur-cluster oxidation has an impact on recombinant protein production and degradation of α1AT. It is conceivable that the biosynthetic rates of some amino acids such as isoleucine, leucine, and valine may be reduced by oxygen exposure.

Superoxide dismutase mutants of *E. coli* are known to be auxotrophic for the branched-chain amino acids (Boehme *et al.* 1976). While it is tempting to speculate that low amino acid levels may have led to translational errors in α_1AT, the stringent response was not found to be activated. The data are inconsistent with this hypothesis. It is also clear that the cell is attempting to obtain additional iron in order to defend itself against reactive oxygen species and possibly to regenerate iron-sulfur clusters. Therefore, the cultures may grow better and produce protein at a higher level in the presence of an iron-rich medium. Experiments along these lines are described in Chapter 6.

5.6 Gene Expression in N_2-Induced Cultures

Although most of the discussion in this chapter has focused on differences between cultures induced in oxygen and air, there are also some interesting changes that occur in response to induction in nitrogen.

Because of the global normalization of the microarray data, comparisons between samples induced in nitrogen and those induced under air or oxygen come with a caveat. The expression values presented here are shown relative to overall expression in that sample. For most sample comparisons, it is reasonable to assume that overall expression is equivalent and thus the absolute expression values are directly comparable. For instance, if expression values in an AIR sample were higher than those in an O2 sample by a value of 1, then it would be assumed that cells in the AIR culture have twice as many copies of the transcript. However, this will not always be the case.

In N2 samples it is clear that the ratio of mRNA:rRNA is lower than in other samples because the background signal from the microarrays (Cy3 channel) is frequently higher than in other samples. This background signal is attributed to nonspecific binding of rRNA molecules. Therefore, the assumption of equal overall expression likely does not hold. As a consequence, expression values given relative to overall expression reveal nothing about the absolute expression values (copies per cell). To illustrate this point, consider an example in which a N2 sample and an O2 sample have equivalent absolute expression of a particular gene (*i.e.* the same number of copies per cell). However, if the overall expression of all genes in the O2 sample were twice that of the N2 sample, then the relative expression in the N2 sample would be twice that in the O2 sample. All of the observations in this section are based on comparisons between

N2 cultures and either AIR or O2 cultures; therefore, all changes in expression are relative to total expression, and no assumptions can be made about absolute expression levels.

5.6.1 Analysis of Expression Data

Block-A data were queried for genes with significant expression changes that gave the N2 culture either higher or lower expression than both the AIR and O2 cultures at the same time point, based on the global significance test. The same query was performed on the data from all blocks, using the combined significance test. The genes identified from these searches were combined with all the genes that showed significant differences between N2-vs.-AIR and N2-vs.-O2 comparisons in the Block-A Aeration-vs.-Time ANOVA. This set of genes was grouped according to gene groups from the EcoCyc database (Table 5.3, Table 5.4, and Table 5.5). Some of the interesting groups are discussed in the following sections.

5.6.2 Respiratory Enzymes

5.6.2.1 Complex I (NADH Dehydrogenase)

Complex I functions under aerobic conditions to transfer electrons from NADH to Coenzyme Q (ubiquinone). Two complexes are available in *E. coli* to carry out this role. NADH dehydrogenase I involves products from 13 different genes and, at 530 kDa, is one of the largest protein complexes in the cell (Spehr *et al.* 1999). This complex also has the ability to couple the electron transfer with proton pumping across the membrane. All thirteen of these proteins appear in the complex in equal proportions and all appear in the same transcription unit. With the exception of *nuoI*, all of the genes in the complex showed decreased expression in the N2 culture. In fact, these twelve genes clustered together with a correlation coefficient of 0.88.

Table 5.3: Gene Groups with Decreased Expression during Induction in N_2 (continued on next page) Genes were grouped according to information from the EcoCyc database. The 104 groups shown here exhibited decreased expression for multiple genes. γ_{DOWN} is the number of genes that show significantly decreased expression in N2 cultures. γ_{TOTAL} is the total number of genes in the group. The remaining genes did not show significant changes (PC=protein complex, PW=pathway, RG=regulon, TU=transcription unit).

EcoCyc Gene Group	γ_{DOWN}	γ_{TOTAL}
PC-acetyl CoA carboxylase	3	4
PC-apo RNA polymerase	3	3
PC-ATP synthase	3	8
PC-cytochrome o ubiquinol oxidase	4	4
PC-flavin reductase / sulfite reductase-(NADPH)	2	2
PC-F-O complex of ATP synthase	2	3
PC-leucine ABC transporter	2	5
PC-membrane-bound subcomplex of succinate dehydrogenase	2	2
PC-NADH dehydrogenase I	12	13
PC-proline ABC transporter	2	3
PC-ribonucleoside-diphosphate reductase II	2	2
PC-RNA polymerase sigma 28	3	4
PC-RNA polymerase sigma 38	3	4
PC-RNA Polymerase sigma19	3	4
PC-RNA polymerase sigma32	3	4
PC-RNA polymerase sigma54	3	4
PC-RNA polymerase sigma70	3	4
PC-secd-secf-yajc-yidc-cplx	3	4
PC-sece/secg/secy-cplx	2	3
PC-sec-secretion-cplx	4	8
PC-succinate dehydrogenase	3	4
PC-thiosulfate ABC transporter	2	4
PC-Trans-202-Cplx	2	3
PW-alanine biosynthesis I	2	3
PW-biotin biosynthesis I	3	4
PW-de novo biosynthesis of purine nucleotides I	12	22
PW-de novo biosynthesis of pyrimidine deoxyribonucleotides	3	10
PW-enterobacterial common antigen biosynthesis	4	11
PW-fatty acid biosynthesis -- initial steps	5	11
PW-fatty acid elongation -- saturated	4	6
PW-fatty acid elongation -- unsaturated	3	5
PW-formylTHF biosynthesis	2	12
PW-glyoxylate cycle	2	6
PW-isopentenyl diphosphate biosynthesis - mevalonate-independent	2	7
PW-leucine biosynthesis	4	6
PW-lipid-A-precursor biosynthesis	4	8
PW-methionine biosynthesis I	2	5
PW-ppGpp biosynthesis	2	3
PW-sulfate assimilation	4	6
PW-superpathway of glyoxylate bypass and TCA	7	20
PW-superpathway of KDO2-lipid A and peptidoglycan biosynthesis	4	28
PW-superpathway of KDO2-lipid A biosynthesis	4	14
PW-superpathway of peptidoglycan and lipid A precursor biosynthesis	4	22
PW-superpathway of saturated and unsaturated fatty acid elongation	4	6
PW-superpathway of sulfate assimilation and cysteine biosynthesis	4	8
PW-TCA cycle -- aerobic respiration	7	18
PW-thiamine biosynthesis	4	8
PW-tryptophan biosynthesis	2	5
PW-UhpBA Two-Component Signal Transduction System	2	3
RG-BirA-bio-5'-AMP transcriptional repressor	4	5
RG-DnaA-ATP transcriptional dual regulator	2	9
RG-Lrp-Leucine transcriptional activator	2	14
RG-MetJ-S-adenosylmethionine transcriptional repressor	5	8
RG-ModE-Molybdate transcriptional dual regulator	3	36
RG-NarP-Phosphorylated transcriptional regulator	3	21
RG-PurR-Hypoxanthine transcriptional repressor	14	28
RG-TrpR-Tryptophan transcriptional repressor	3	12
RG-TyrR-Tyrosine transcriptional repressor	2	9
TU-abc	2	3
TU-accBC	2	2
TU-atpIBEFHAGDC	3	9
TU-bioBFCD	3	4
TU-cyoABCDE	5	5
TU-cysDNC	2	3
TU-cysJIH	2	3

EcoCyc Gene Group	γ_{DOWN}	γ_{TOTAL}
TU-cysPUWAM	3	5
TU-dusB-fis	2	2
TU-emrRAB	2	3
TU-fabHDG	2	3
TU-g30K-rpmF	2	2
TU-glpEGR	3	3
TU-glpGR	2	2
TU-hscBA-fdx	2	3
TU-hyb0ABCDEFG	3	8
TU-ilvLG_1G_2MEDA	2	7
TU-iscRSUA	4	4
TU-leuLABCD	4	5
TU-livKHMGF	2	5
TU-metBL	2	2
TU-metY-yhbC-nusA-infB	2	3
TU-metY-yhbC-nusA-infB-rbfA-truB-rpsO-pnp	4	7
TU-napFDAGHBC-ccmABCDEF-dsbE-ccmH	3	15
TU-nrdHIEF	4	4
TU-nuoABCEFGHIJKLMN	12	13
TU-proU	2	3
TU-purEK	2	2
TU-rplJL-rpoBC	3	4
TU-rplM-rpsI	2	2
TU-rplNXE-rpsNH-rplFR-rpsE-rpmD-rplO-prlA-rpmJ	10	12
TU-rpmBG	2	2
TU-rpmH-rnpA	2	2
TU-rpoBC	2	2
TU-rpsF-priB-rpsR-rplI	3	4
TU-rpsJ-rplCDWB-rpsS-rplV-rpsC-rplP-rpmC-rpsQ	10	11
TU-rpsLG-fusA-tufA	3	4
TU-rpsMKD-rpoA-rplQ	4	5
TU-rpsP-rimM-trmD-rplS	3	4
TU-sdhCDAB-b0725-sucABCD	4	9
TU-secE-nusG	2	2
TU-thiCEFGH	5	5
TU-trpLEDCBA	2	6
TU-uhpABC	2	3
TU-yajC-secD-secF	3	3
TU-ycfC-purB	2	2

EcoCyc Gene Group	γ_{UP}	γ_{TOTAL}
PC-anaerobic nucleoside-triphosphate reductase activating system	2	3
PC-aspartate-carbamoyltransferase	2	2
PC-Copper Transporting Efflux System	3	4
PC-cytochrome bd-II terminal oxidase	2	2
PC-cytochrome d ubiquinol oxidase	2	2
PC-glutamate ABC transporter	3	3
PC-glutamine ABC transporter	2	3
PC-YhdW/YhdX/YhdY/YhdZ ABC transporter	2	4
PW-(deoxy)ribose phosphate degradation	4	7
PW-asparagine biosynthesis III	2	2
PW-glycerol degradation I	3	15
PW-glycogen biosynthesis	2	3
PW-glycolysis I	6	14
PW-histidine biosynthesis I	5	8
PW-lysine biosynthesis I	3	7
PW-mannitol degradation	3	5
PW-mixed acid fermentation	4	25
PW-Nitrogen Regulation Two-Component System	4	6
PW-non-oxidative branch of the pentose phosphate pathway	3	7
PW-sorbitol degradation	3	5
PW-superpathway of glycolysis and Entner-Doudoroff	6	17
PW-superpathway of hexitol degradation	3	6
PW-superpathway of oxidative and non-oxidative branches of pentose phosphate pathway	3	9
PW-trehalose biosynthesis I	2	2
RG-AppY transcriptional activator	2	9
RG-AsnC transcriptional dual regulator	2	2
TU-appCBA	2	3
TU-argT-hisJQMP	2	5
TU-cydAB	2	2
TU-deoCABD	2	4
TU-epd-pgk	2	2
TU-focA-pflB	2	2
TU-fruBKA	2	3
TU-glnALG	3	3
TU-glnHPQ	2	3
TU-glnK-amtB	2	2
TU-glnLG	2	2
TU-gltIJKL	3	4
TU-hisGDCBHAFI	5	8
TU-otsBA	2	2
TU-pspABCDE	2	5
TU-pyrBI	2	2
TU-rpoE-rseABC	2	4
TU-yhdWXYZ	2	4
TU-ylcA-ybcZ	2	2
TU-ylcBCD-ybdE	3	4

Table 5.4: Gene Groups with Increased Expression during Induction in N_2
(opposite) Genes were grouped according to information from the EcoCyc database. The 46 groups shown here exhibited increased expression for multiple genes. γ_{UP} is the number of genes that showed significantly increased expression in N2 cultures. γ_{TOTAL} is the total number of genes in the group. The remaining genes did not show significant changes (PC=protein complex, PW=pathway, RG=regulon, TU=transcription unit).

Table 5.5: Regulons with Mixed Expression during Induction in N_2
Genes were grouped according to information from the EcoCyc database. The 26 regulons shown here exhibited mixed expression, with some genes showing increased expression and others showing decreased expression. γ_{MIXED} is the number of genes that show significant expression changes in N2 cultures. γ_{TOTAL} is the total number of genes in the regulon. The remaining genes did not show significant changes.

EcoCyc Gene Group	γ_{MIXED}	γ_{TOTAL}
RG-ArcA-Phosphorylated transcriptional dual regulator	34	79
RG-ArgR-L-arginine transcriptional repressor	2	10
RG-CRP transcriptional dual regulator	5	11
RG-CRP-cAMP transcriptional dual regulator	38	241
RG-CysB transcriptional dual regulator	8	18
RG-CytR transcriptional dual regulator	5	11
RG-DeoR transcriptional repressor	3	6
RG-Fis transcriptional dual regulator	25	59
RG-FlhD transcriptional dual regulator	2	28
RG-FNR transcriptional dual regulator	38	119
RG-FruR transcriptional dual regulator	7	26
RG-Fur transcriptional dual regulator	10	40
RG-Hns transcriptional dual regulator	7	40
RG-IHF transcriptional dual regulator	36	159
RG-Lrp transcriptional dual regulator	15	53
RG-MarA transcriptional activator	3	15
RG-Nac transcriptional activator	5	13
RG-NarL-Phosphorylated transcriptional dual regulator	23	79
RG-NtrC-Phosphorylated transcriptional dual regulator	16	43
RG-PhoB-Phosphorylated transcriptional dual regulator	4	29
RG-RNAP32-CPLX	4	24
RG-RNAP54-CPLX	21	85
RG-RNAPE-CPLX	3	18
RG-RNAPS-CPLX	16	69
RG-SoxS transcriptional activator	3	18

NADH dehydrogenase II is a single protein, encoded by the gene *ndh*, which lacks the proton-translocating ability of NADH dehydrogenase I. For this reason, Ndh is assumed to be active during aerobic conditions when NADH is in high abundance. The gene *ndh* is known to be repressed by FNR and both repressed and activated by Fis (Factor for Inversion Stimulation) (Green *et al.* 1996). Moderate levels of the Fis protein activate *ndh* transcription and high levels of Fis repress *ndh* transcription. Expression data show that the *ndh* gene was strongly activated after switching to nitrogen, which suggests that FNR levels were low and that Fis levels were moderate. Incidentally, the *fis* gene was found to be repressed somewhat by exposure to nitrogen. A corresponding decrease in Fis levels would certainly explain the sudden increase in *ndh* expression.

In N2 cultures, the genes encoding NADH dehydrogenase I were strongly repressed, while the gene encoding NADH dehydrogenase II showed equally strong induction. Since neither of these proteins were expected to be present at appreciable levels in completely anaerobic cultures, induction of *ndh* indicates that the N2 cultures were actually grown in microaerobic conditions.

5.6.2.2 Cytochrome Oxidases

Cytochrome oxidases transfer electrons from Coenzyme Q to the terminal electron acceptor. The enzymes cytochrome *d* oxidase, cytochrome *o* oxidase, and cytochrome *bd*-II oxidase all use oxygen as the terminal electron acceptor. Cytochrome *o* oxidase has a low affinity for oxygen and is most abundant during aerobic conditions (Rice and Hempfling 1978). As expected, the four genes involved in producing this enzyme (*cyoA*, *cyoB*, *cyoC*, and *cyoD*) all showed higher expression in the AIR and O2 cultures than in the N2 culture. The other two oxidases, however, are more active during anaerobiosis. Cytochrome *d* oxidase has a higher affinity for oxygen and is therefore induced in a low-oxygen environment in order to make the most efficient use of the available oxygen (Rice and Hempfling 1978). Cytochrome *bd*-II oxidase expression is known to be activated by the anaerobic regulator AppY (no data were collected for expression of the *appY* gene) (Atlung and Brøndsted 1994). Indeed, the Block-A data revealed that the four genes encoding these oxidases (*appB*, *appC*, *cydA*, and *cydB*) all showed increased expression by induction in nitrogen. These data are consistent with a shift from aerobic to anaerobic respiration.

5.6.2.3 Other Aerobic Respiratory Enzymes

Complex II couples the oxidation of succinate to fumarate with the reduction of Coenzyme Q. This reaction plays a role in the TCA cycle as well as the electron transport chain. The genes encoding this complex (*sdhCDAB*) are all on the same transcription unit, which is repressed by both ArcA and FNR. With the exception of *sdhB*, these genes showed decreased expression upon induction in nitrogen, as expected.

5.6.2.4 Anaerobic Respiratory Enzymes

Anaerobic respiratory enzymes transfer electrons from NADH, glycerol-3-phosphate, formate, and H_2 to menaquinone, and ultimately to alternative terminal electron acceptors such as nitrate, nitrite, dimethyl sulfoxide (DMSO), trimethylamine-*N*-oxide (TMAO), and fumarate. Based on the list of differentially expressed genes in the N2 cultures, there was no clear trend in the anaerobic respiratory enzymes. Most of the genes involved in these pathways did not show significant expression changes in N2 cultures, and those that did, showed decreased expression. Only three of these genes showed increased expression under the experimental conditions: *frdB*, which codes for a component of fumarate reductase; *torY*, which codes for a component of TMAO reductase III; and *glpA*. Curiously, the genes *glpA* and *glpC*, which encode glycerol-3-phosphate dehydrogenase and appear on the same transcription unit, show opposing effects in N2 cultures. The gene *glpA* showed increased expression in the N2 culture and *glpC* showed decreased expression. If the N2 culture were anaerobic, all of these genes would have increased in expression, and if the N2 culture were microaerobic, these genes would be expected to remain the same. Using *lacZ* fusions, it has been shown that cytochrome *d* oxidase expression increases below 15% air saturation, while expression of anaerobic respiratory enzymes only increases below 10% air saturation (Tseng *et al.* 1996). Since *cyd* genes showed increased expression, but anaerobic-enzyme genes did not, the N2 cultures were likely operated at microaerobic conditions, with oxygen levels somewhere between 15% and 10% air saturation.

5.6.3 Protein Synthesis

Machinery for protein synthesis showed decreased expression in N2 cultures. Of the 49 ribosomal proteins for which data were available, 43 were on the list of genes differentially expressed in the N2 cultures. On average, in the N2 cultures, these genes showed signal ratios

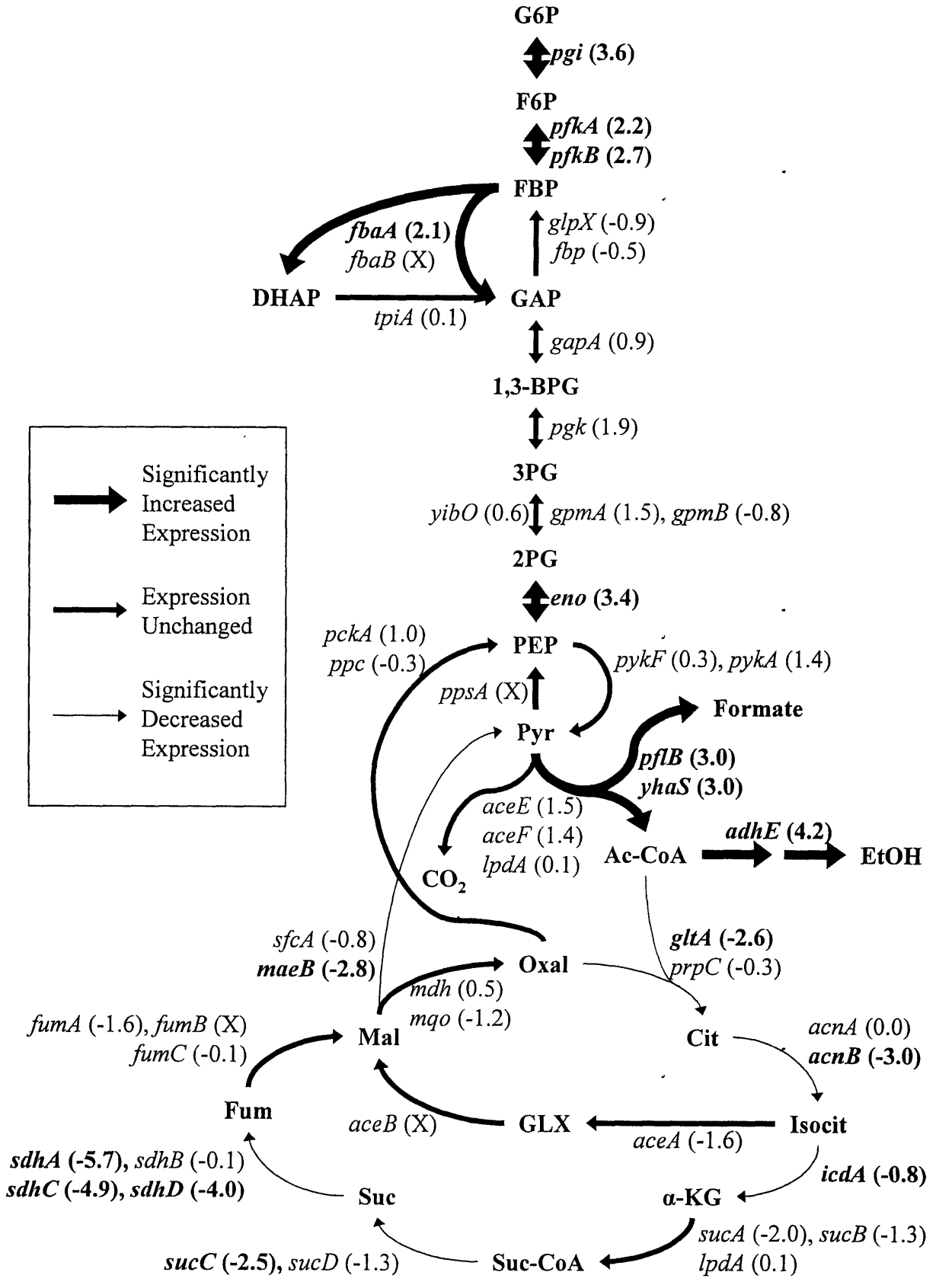
that were lower than those in aerobic cultures by a value of 1.5. This corresponds to a 2.9-fold decrease in expression of these genes. In addition, the genes encoding the subunits of the RNA polymerase core (*rpoA*, *rpoB*, and *rpoC*) all showed decreased expression. Overall, protein synthesis appeared to be slower in cultures induced in N_2.

5.6.4 Shift to Anaerobic Metabolism

The metabolic changes that occur upon shift from an aerobic environment to an anaerobic environment have been well studied. The metabolic flux through the TCA cycle is known to decrease as fermentation pathways are activated to provide an alternative route for oxidizing NADH and metabolizing pyruvate. These same trends are observed in data from the N2 cultures (Figure 5.13).

After 30 min in nitrogen, the culture made expression changes to increase the flux through the glycolytic pathway. Of the 13 glycolysis genes for which data were available, six appeared on the list of differentially expressed N2 genes. Typically, these genes had increased expression at 30 min and 60 min post-induction. Also showing increased expression were several genes involved in fermentation pathways. Genes encoding pyruvate-formate lyase (*pflB*) and pyruvate-formate lyase 4 (*yhaS*) both showed increased expression in the N2 cultures. The gene encoding alcohol dehydrogenase (*adhE*) also showed increased expression. Together the enzymes encoded by these genes are responsible for converting pyruvate to ethanol. Induction in nitrogen led to increased gene expression for six enzymes in the pathway between glucose and ethanol.

Figure 5.13: Anaerobic vs. Aerobic Gene Expression Changes in Central Metabolism
(opposite) The reactions of central metabolism, including glycolysis, gluconeogenesis, mixed acid fermentation, the TCA cycle, and the glyoxylate bypass are shown. Gene names are shown next to the reactions their products catalyze, and log ratios from Block-A N2-AIR comparisons at 30 min are shown in parentheses. Positive values indicate increased expression in the N2 culture. An X indicates that no data were collected for that gene. Genes showing significant expression changes are shown in bold. Arrows increase and decrease in size if at least one gene in the reaction showed differential expression.


G6P
pgi (3.6)
F6P
pfkA (2.2)
pfkB (2.7)
FBP
fbaA (2.1)
fbaB (X)
glpX (-0.9)
fbp (-0.5)
DHAP
tpiA (0.1)
GAP
gapA (0.9)
1,3-BPG
pgk (1.9)
3PG
yibO (0.6)
gpmA (1.5), gpmB (-0.8)
2PG
eno (3.4)
pckA (1.0)
ppc (-0.3)
PEP
pykF (0.3), pykA (1.4)
ppsA (X)
Pyr
Formate
pflB (3.0)
yhaS (3.0)
aceE (1.5)
aceF (1.4)
lpdA (0.1)
CO2
Ac-CoA
adhE (4.2)
EtOH
sfcA (-0.8)
maeB (-2.8)
Oxal
gltA (-2.6)
prpC (-0.3)
mdh (0.5)
mqo (-1.2)
fumA (-1.6), fumB (X)
fumC (-0.1)
Mal
Cit
acnA (0.0)
acnB (-3.0)
Fum
aceB (X)
GLX
aceA (-1.6)
Isocit
sdhA (-5.7), sdhB (-0.1)
sdhC (-4.9), sdhD (-4.0)
Suc
icdA (-0.8)
α-KG
sucA (-2.0), sucB (-1.3)
lpdA (0.1)
sucC (-2.5), sucD (-1.3)
Suc-CoA
Significantly Increased Expression
Expression Unchanged
Significantly Decreased Expression

Several genes involved in the TCA cycle showed decreased expression in N2 cultures. Genes encoding the enzymes citrate synthase (*gltA*) aconitase (*acnB*), isocitrate dehydrogenase (*icdA*), and succinyl-CoA synthetase (*sucC*) all showed decreased expression. As mentioned previously, several genes that code for succinate dehydrogenase also showed decreased expression in N2 cultures. These changes presumably resulted in increased flux through glycolytic and fermentative pathways as well as decreased flux through the TCA cycle. Induction in nitrogen redirects pyruvate away from reactions that generate NADH in the TCA cycle, toward fermentation reactions that will consume it.

5.6.5 Summary of Gene Expression Changes in N_2-Induced Cultures

The changes in gene expression discussed here only scratch the surface in terms of interpretation of this data set. 646 (16%) genes in the *E. coli* genome showed differential expression upon induction in nitrogen. Some of these changes are consistent with known regulation of gene expression, but many more are the result of unknown regulatory mechanisms. Those changes that were more easily explained were those involved in the shift from aerobic to anaerobic respiration. These transcriptional changes were consistent with changes in the ArcA regulon, but not entirely with changes in the FNR regulon. Coupled with the unaltered expression of anaerobic respiratory genes, these results indicated that the N2 cultures were in a microaerobic regime. This data set also showed that genes involved in protein synthesis were down-regulated, relative to levels of all other transcripts. Expected metabolism changes involved in anaerobic fermentation were also observed.

5.7 Summary and Conclusions

The effects of aeration extremes on cultures producing a recombinant protein were examined using DNA microarray analysis. The dominant effects were those of superoxide. The SoxRS response was activated by exposure to oxygen and, to a lesser extent, air. Consistent with oxidation of iron-sulfur clusters by superoxide, several genes and pathways involving iron-sulfur proteins showed increased expression. In addition, the two iron-sulfur repair systems, Isc and Suf, both showed increased expression. Despite activation of the iron uptake repressor, *fur*, iron uptake actually showed a dramatic increase. It is proposed that this increase is in response to increased intracellular iron(II) levels from oxidation of iron-sulfur clusters. Because the changes

described here constitute the dominant oxygen-dependent gene expression changes, this is the most likely path to follow to further elucidate the oxygen dependence of α_1AT degradation.

6 Effects of Iron Supplementation on Induced Cultures

"Good pitching will stop good hitting and vice versa."
—Casey Stengel

The roles of iron and oxygen in cellular damage are ambiguous. Both are known to be detrimental to the cell, but it is unclear whether the effects are complementary or oppose one another. It is similarly unclear whether iron supplementation will aggravate or mitigate the effects of hyperoxic stress. One of the negative consequences of iron on *E. coli* cultures is the Fenton reaction **(1.4)**, which is known to generate hydroxyl radicals that damage membrane lipids, proteins, DNA, and other molecules in the cell. Other evidence of damage by iron has been found in SOD mutants, which have decreased defense against superoxide. In these strains, iron contributes to the toxicity of superoxide (Keyer *et al.* 1995; McCormick *et al.* 1998). However, addition of iron to the growth medium has also been shown to protect against superoxide stress. SOD mutants of both *E. coli* (Benov and Fridovich 1998) and yeast (De Freitas *et al.* 2000) were shown to have improved aerobic growth in the presence of iron. This *E. coli* study reported decreased leakage of sulfite in iron-supplemented cultures (Benov and Fridovich 1998). In addition, catalase levels were found to be equivalent, indicating that peroxide stress was not aggravated by iron addition.

The different observations in these two sets of SOD-mutant studies may have resulted from differences in medium. The first two studies were performed in rich media, while the latter two were performed in defined media. These defined media may have been iron-limiting, even with supplementation of iron. Thus, an optimal iron level might exist at which superoxide stress is alleviated and the toxic Fenton reaction proceeds only slowly.

Results form the previous chapter demonstrated that cultures induced under pure oxygen experienced superoxide stress and damage to iron-sulfur clusters throughout the cell. In response to this damage, the superoxide stress response was activated as were systems for repair of iron-sulfur clusters.

Under hyperoxic conditions, transcription of *fur*, the gene encoding the ferric uptake repressor, was found to increase significantly. This gene appears in both the OxyR regulon for peroxide defense as well as the SoxRS regulon for superoxide defense. Iron uptake genes were

expected to show decreased expression in order to limit the toxic Fenton reaction; however, just the opposite was observed. 18 genes in the Fur regulon showed increased expression in oxygen-induced cultures. Genes involved in synthesis of enterobactin, an iron transporter, as well as those involved in transporting enterobactin and iron across the membrane, showed significantly increased expression. This observation suggests that iron transport plays a role in protecting the culture from superoxide stress, perhaps by supplying iron for the systems of iron-sulfur repair.

To investigate whether iron would help to improve the defense against superoxide, iron was supplemented to cultures producing α1-antitrypsin (α_1AT). The effects on degradation of recombinant α_1AT as well as gene expression were observed.

6.1 Observations on Iron Supplementation of E. coli Cultures

Cultures were grown at different iron levels to observe the growth response of *E. coli* cultures to varying doses of the iron species Fe^{2+} and Fe^{3+}. Seven cultures of *E. coli* BL21 were grown in the following seven media: M9 minimal medium without iron, M9 with supplemented $FeCl_2$ (0.25 mM, 0.50 mM, and 0.75 mM), and M9 with supplemented $FeCl_3$ (0.25 mM, 0.50 mM, and 0.75 mM). Note that the concentrations listed here are for the supplemental iron. Since M9 medium contains close to 100 μM $FeCl_3$, these are not the total iron concentrations in the medium. Sterile filtered iron solutions were used for both of these experiments. $FeCl_2$-supplemented media were prepared by adding 7.5 mL, 15 mL, and 22.5 mL of a 5 mM stock. $FeCl_3$-supplemented media were prepared by adding 375 μL, 750 μL, and 1,125 μL of a 100 mM stock.

Optical density (OD_{600}) measurements were strongly affected by the level of iron in the cultures. To correct for the effects of iron on A_{600}, a second flask containing identical medium was prepared and treated the same as the cultures, but this flask was never inoculated. The corrected OD_{600} of the culture was calculated by subtracting the A_{600} of the medium alone at the same time. For each culture, 150 mL of medium was prepared. 50 mL was placed in a 250-mL shake flask, while the remaining 100 mL was added to a 500-mL shake flask and was inoculated. All fourteen flasks were shaken at 250 rpm in air. These extra medium flasks allowed for OD_{600} correction based on the level of iron in the medium.

Growth curves from this experiment are shown in Figure 6.1. Figure 6.1A shows the uncorrected A_{600} measurements for each of the seven cultures over a 6.5 h period. A_{600} measurements were taken from the seven medium flasks and the seven cultures at the same time points throughout this 6.5 h period. A_{600} measurements from the medium flasks are shown in Figure 6.1B. Some interesting trends were observed. First, the A_{600} values increased significantly between 0 h and 0.5 h. Additionally, the magnitude of this increase depended on the amount of supplemental iron. These features were indicative of a reaction occurring upon addition of iron to the medium. This A_{600} increase was not observed without heat (30°C) and agitation, which suggests that oxygen may play a role in this reaction. The second notable trend was that the A_{600} value depended not only on the amount of iron, but also the oxidation state. $FeCl_2$ generally had a lower A_{600} than did $FeCl_3$.

C was generated by subtracting the results from Figure 6.1B from those in Figure 6.1A. These corrected OD_{600} values all appear to fall on top of one another, indicating that the effects of iron were minimal, and that the correction with separate medium flasks was appropriate. The growth rates from these cultures are plotted in Figure 6.2. All of these growth rates appear to be equivalent.

When iron was added to M9 minimal medium, a white precipitate formed. This precipitate is most likely $Fe(PO_4)$ and/or $Fe_2(PO_4)_3$, both of which are white in color. Phosphate is a major component of the medium and would be readily available to form this precipitate. In retrospect, use of iron citrate may have been more appropriate, since it may have prevented formation of this precipitate. However, as a TCA-cycle intermediate, citrate may have also had undesired consequences on the metabolism of the cell.

Experiments were performed using both sterile-filtered iron and autoclaved iron. The autoclaved iron had a noticeable red color to it (even after addition to M9 medium) that was absent from the sterile-filtered iron. This red color is attributed to hematite (Fe_2O_3) formation during autoclaving.

In the range of concentrations used here, supplementation of iron(II) and iron(III) has no effect on the growth rate of *E. coli* cultures.

Figure 6.1: Dose Response of *E. coli* to $FeCl_2$ and $FeCl_3$
(opposite) Cultures were grown in seven different media containing either $FeCl_2$ (0.25 mM – 0.75 mM), $FeCl_3$ (0.25 mM – 0.75 mM), or no supplement. .Separate flasks were filled with the same seven media and were treated identically. **A)** A_{600} from each culture and **B)** A_{600} from the medium flasks were monitored periodically. **C)** Corrected OD_{600} of each culture was calculated as the difference of these two measurements.

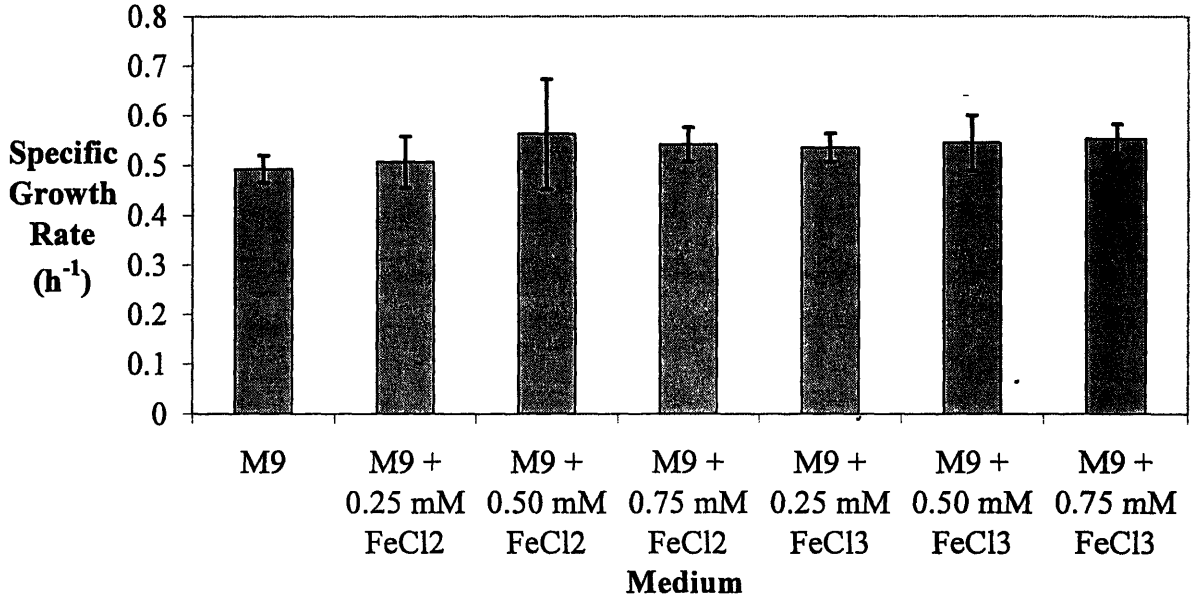
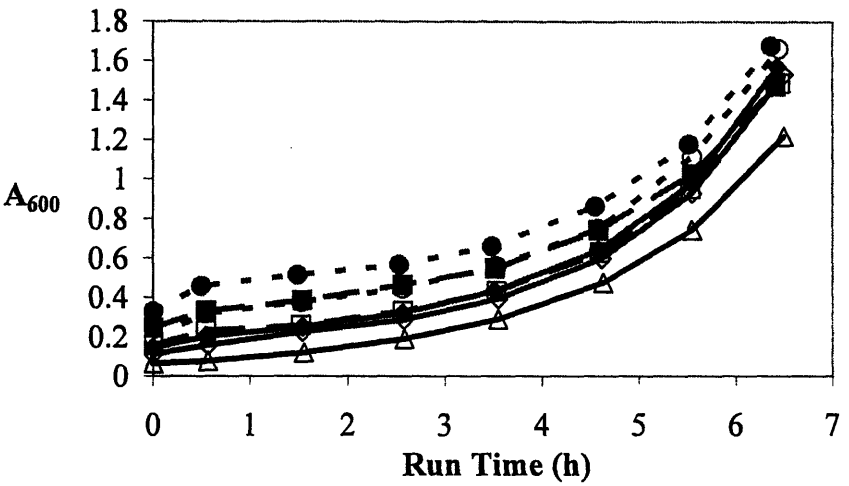


Figure 6.2: Specific Growth Rates of *E. coli* in Iron-Supplemented Media
Cultures were grown in seven different media containing either $FeCl_2$ (0.25 mM – 0.75 mM), $FeCl_3$ (0.25 mM – 0.75 mM), or no supplement. Specific growth rates were calculated by performing a linear regression on the log-transformed, corrected OD_{600} values (Figure 6.1C). Error bars represent a 99% confidence interval.

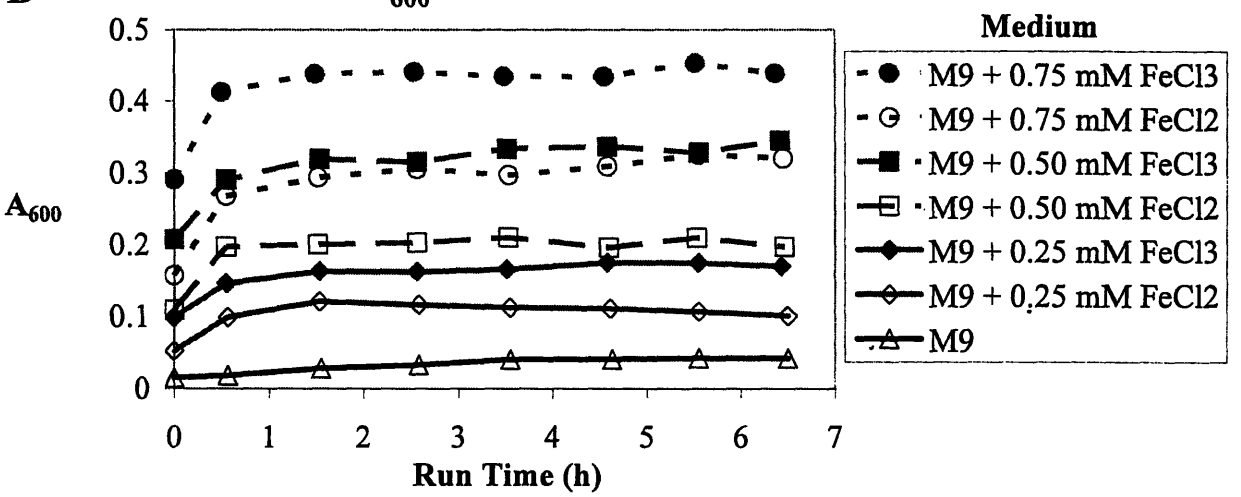
6.2 *Effects of Iron on α1-Antitrypsin Degradation*

The effects of superoxide stress dominate the changes in gene expression observed under hyperoxic conditions. It is likely that these same effects led to increased degradation of α_1AT under these conditions. Supplementation of iron is expected to alleviate the damaging effects of superoxide stress and remove the oxygen dependence of α_1AT degradation. In order to observe the effects of iron supplementation on the quality of the recombinant product, pulse-chase experiments were performed to monitor α_1AT degradation.

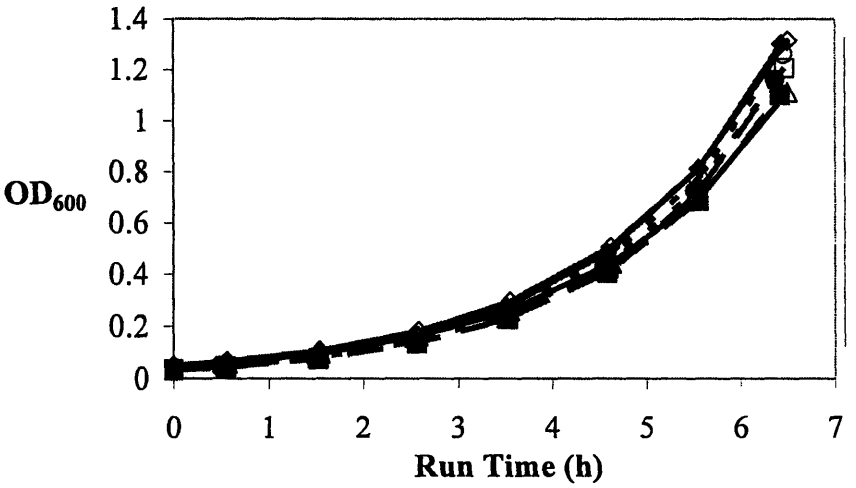
A


A600 of Cultures
A600
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7
Run Time (h)

B


A600 of Medium
A600
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7
Run Time (h)
Medium
M9 + 0.75 mM FeCl3
M9 + 0.75 mM FeCl2
M9 + 0.50 mM FeCl3
M9 + 0.50 mM FeCl2
M9 + 0.25 mM FeCl3
M9 + 0.25 mM FeCl2
M9

C


OD600 of Cultures - Corrected
OD600
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7
Run Time (h)

6.2.1 Autoclaved $FeCl_2$ and $FeCl_3$ Supplementation

Examples of iron supplementation in the literature used both Fe^{2+} and Fe^{3+} in various concentrations. The initial experiment described here used iron in both oxidation states. A concentration of 500 μM supplemental iron was chosen, since this amount was found to counter the effects of paraquat addition to an *E. coli* SOD mutant (Benov and Fridovich 1998). This amount was also shown to have little effect on growth (Section 6.1). Stock solutions of both $FeCl_2$ and $FeCl_3$ were prepared at 5 mM concentration and were autoclaved.

Pulse-chase experiments were performed as described in Section 3.8. Three separate 100-mL cultures were grown, one in $FeCl_2$-supplemented M9, another in $FeCl_3$-supplemented M9, and the third was a control in unsupplemented M9. Three 250-mL shake flasks were also filled with 50 mL of each medium and were treated identically to account for the A_{600} of the medium. In bubbler tubes, all three cultures were bubbled, via manifold, with pure O_2. Figure 6.3 shows the degradation profiles of α_1AT starting 60 min after induction. Both iron-supplemented cultures showed degradation that was slower and proceeded to a lower extent when compared with the control. These cultures also showed significantly lower rate constants for α_1AT proteolysis (k_p) as well as r_p/r_f ratios that were either lower or equivalent (Figure 6.4). Regardless of oxidation state, autoclaved iron alleviated the degradation of α_1AT observed under hyperoxic conditions.

6.2.2 Sterile-Filtered $FeCl_2$ and $FeCl_3$ Supplementation

Oxidation of Fe^{2+} to Fe^{3+} is a concern with autoclaved iron solutions. In an attempt to confirm that autoclaving the iron had no effect on the degradation of the recombinant product, the experiment performed in Section 6.2.1 was repeated with sterile-filtered $FeCl_2$ (using 5 mM stock solution) and sterile-filtered $FeCl_3$ (using 100 mM stock solution). Figure 6.5 shows the degradation profiles in each medium. Both iron-supplemented cultures showed proteolysis rate constants (k_p) and r_p/r_f ratios that were either equivalent or higher than those in the control culture (Figure 6.6). Unlike supplementation with autoclaved iron, sterile-filtered iron did not alleviate degradation of α_1AT.

This surprising result was explained by the formation of hematite (Fe_2O_3) during autoclaving. Typically, addition of iron to the phosphate-buffered M9 medium results in formation of an iron phosphate precipitate. However, autoclaved iron also contains hematite.

Based on the red color in the iron-supplemented M9 medium, this hematite is somewhat resistant to reaction with phosphate ions. It is proposed that the iron in hematite is more readily accessible to cells than the iron in the iron phosphate precipitate. Therefore, it is hematite that alleviates the degradation of α_1AT under hyperoxic conditions. This model also explains why data in Figure 6.4 are similar for both $FeCl_2$ and $FeCl_3$, since both would be expected to form hematite.

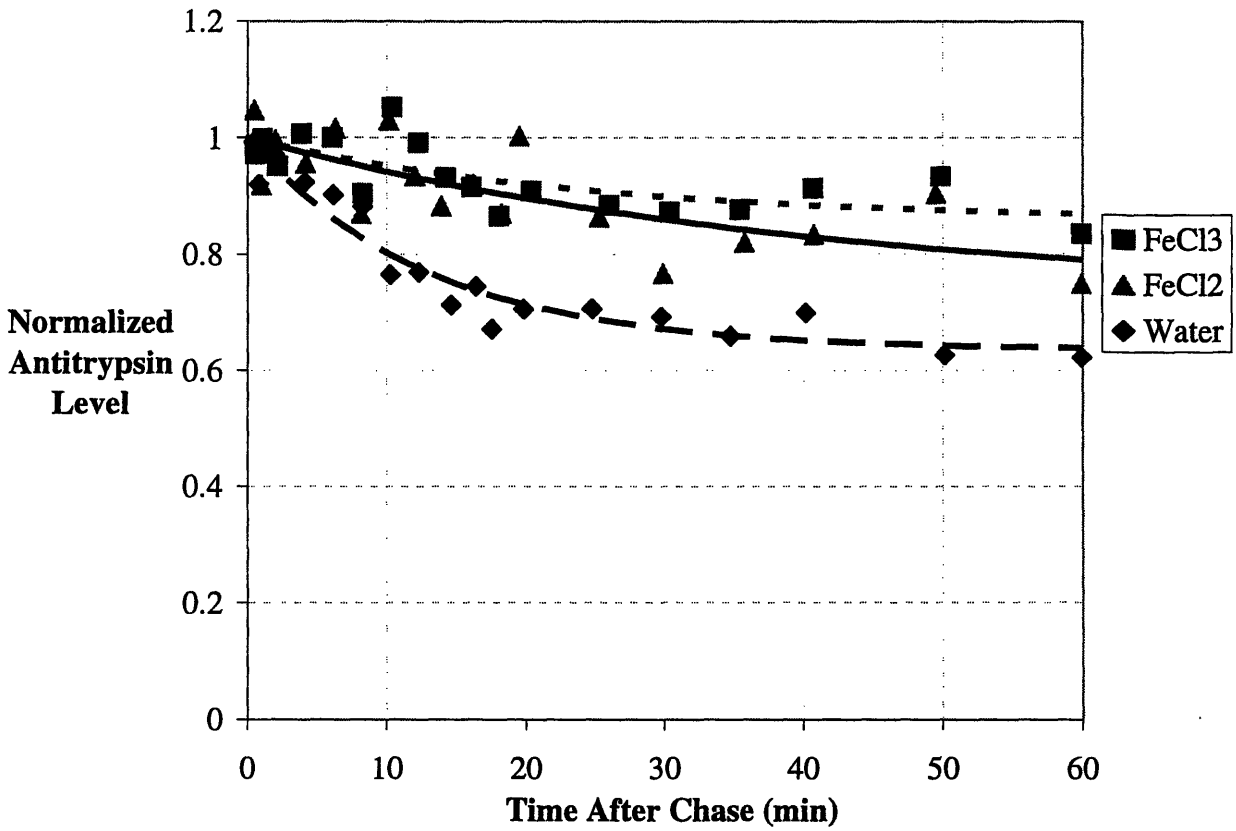


Figure 6.3: α1-Antitrypsin Degradation in Oxygen-Induced Cultures Supplemented with Autoclaved Iron
Three cultures were grown in M9 medium supplemented with either 500 μM autoclaved $FeCl_3$, 500 μM autoclaved $FeCl_2$, or water (control culture). Cultures were induced in oxygen. After 60 min of induction, recombinant α_1AT was pulse-chase labeled with ^{35}S-methionine. Degradation was monitored for the next 60 min.

One experimental source of error may have had an impact on the experiments presented in this section and in Section 6.2.2 and must be discussed. The rates of methionine uptake and utilization are very fast. Based on the growth rate of cultures in minimal medium at 30°C, a rough estimate of the rate of translation would be 1.7 mg/(mL·OD_{600}·min). Assuming

methionine constitutes 3% of the total cellular protein, these bubbler cultures would incorporate methionine at an approximate rate of 2.3 μmoles/min. Based on experimentally determined values of methionine uptake in the absence of protein synthesis (Kadner 1974), an *E. coli* bubbler culture at 30°C in minimal medium is expected to transport methionine at a rate of approximately 4.2 μmoles/min. Based on these rough estimates, it is clear that the 1.4-μmole pulse of methionine would have been consumed long before the end of the 3-min period. As a result, significant degradation of labeled protein would have occurred before the chase, and would not have been observed. The effect of this excessively long pulse period would be an underestimation of the proteolysis rate.

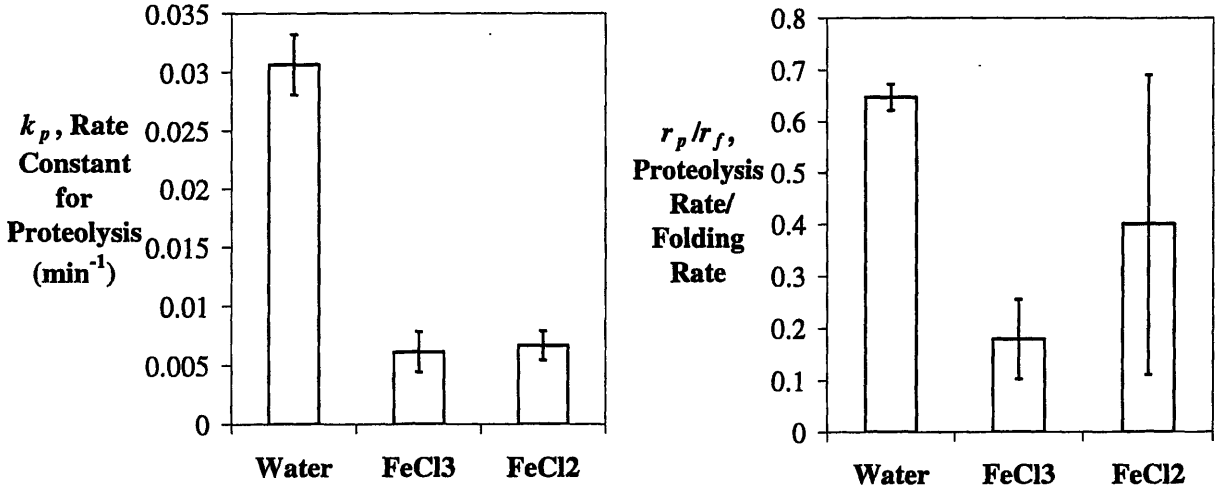


Figure 6.4: Kinetic Parameters for α1-Antitrypsin Degradation in Cultures Supplemented with Autoclaved Iron
Model parameters for data in Figure 6.3. Cultures were supplemented with Water (control), 500 μM autoclaved $FeCl_3$, and 500 μM autoclaved $FeCl_2$, and were induced in oxygen. *kp* is the pseudo-first-order rate constant for α_1AT proteolysis and r_p/r_f is the ratio of the rate of proteolysis and the rate of folding. Parameters and confidence intervals were calculated as described in Section 3.8.3.

Despite this source of inaccuracy, experiments were performed with a consistent 3-min pulse period. Therefore, all experiments in this work should be directly comparable. However, data from the experiments in this section and Section 6.2.2 appeared to be inconsistent with those from other experiments. The ^{35}S-methionine used in labeling for these experiments was more than two months old. Although the reagent was kept at 4°C, it is expected to degrade at a rate of 2% per week (this methionine degradation is separate from its radioactive decay). At the time of these experiments, less than 80% of the original methionine remained. This loss of reagent

appeared to worsen the effect described above, such that the measured proteolysis rates were significantly lower than those from all other experiments in this work. All other experiments described here used ^{35}S-methionine that was less than 2 months from the assay date.

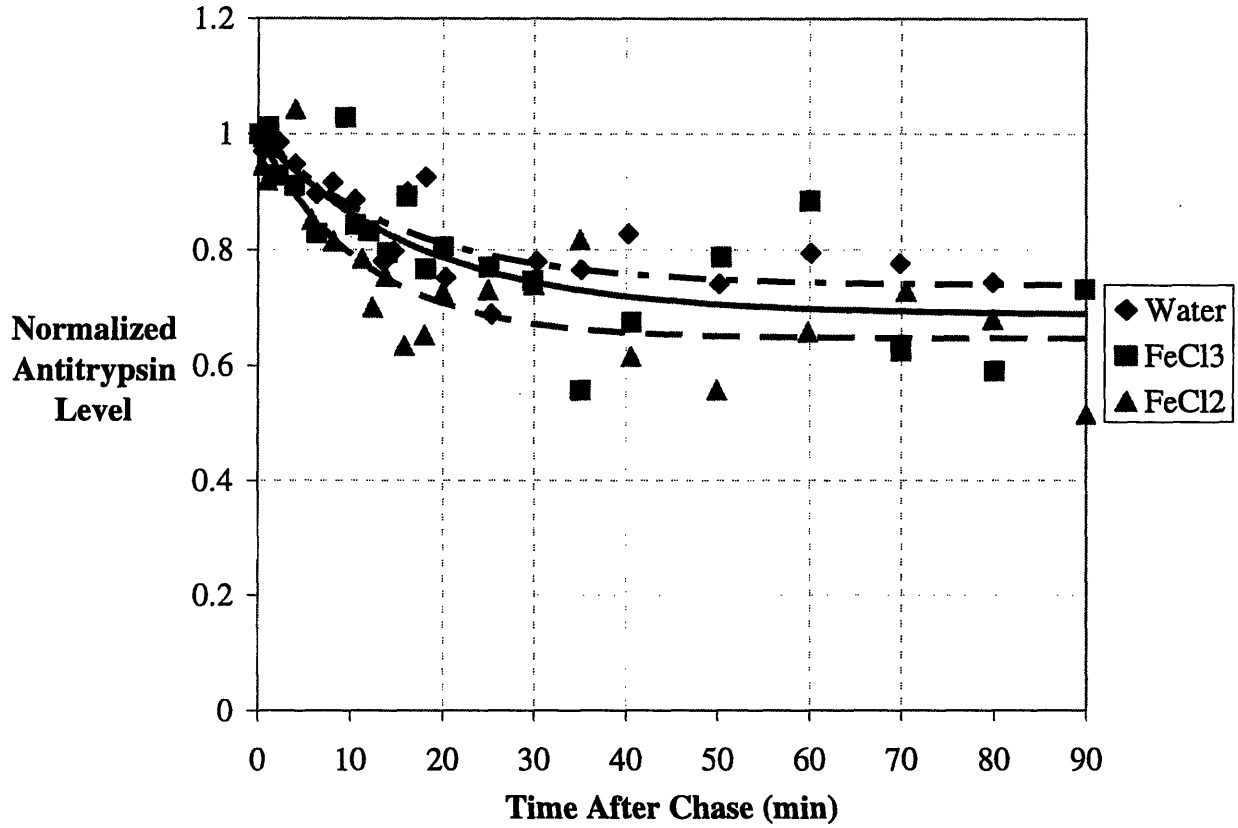


Figure 6.5: α1-Antitrypsin Degradation in Oxygen-Induced Cultures Supplemented with Sterile-Filtered Iron
Three cultures were grown in M9 medium supplemented with either 500 μM sterile-filtered $FeCl_3$, 500 μM sterile-filtered $FeCl_2$, or water (control culture). Cultures were induced in oxygen. After 60 min of induction, recombinant α_1AT was pulse-chase labeled with ^{35}S-methionine. Degradation was monitored for the next 90 min.

6.2.3 Iron Supplementation with Varying Aeration

In order to provide a direct comparison with the results of α_1AT degradation in BL21 cultures presented in Figure 5.1, cultures were supplemented with iron and grown in pure nitrogen, air, and pure oxygen. The pulse-chase protocol was performed by growing one 100-mL culture in M9 supplemented with 500 μM autoclaved $FeCl_2$ and splitting it into three bubbler tubes, each with different aeration. A 250-mL shake flask was also filled with 50 mL of iron-supplemented medium to account for the absorbance of the medium at 600 nm. Figure 6.7

shows the degradation profiles from each of these cultures. All three iron-supplemented cultures showed degradation with similar rates (Figure 6.8) and extents. While the r_p/r_f ratio showed a slightly increasing trend with oxygen level, the oxygen dependence was much smaller than that observed without iron supplementation. While these cultures still showed a moderate level of degradation, supplemental iron largely removed the oxygen dependence of this degradation.

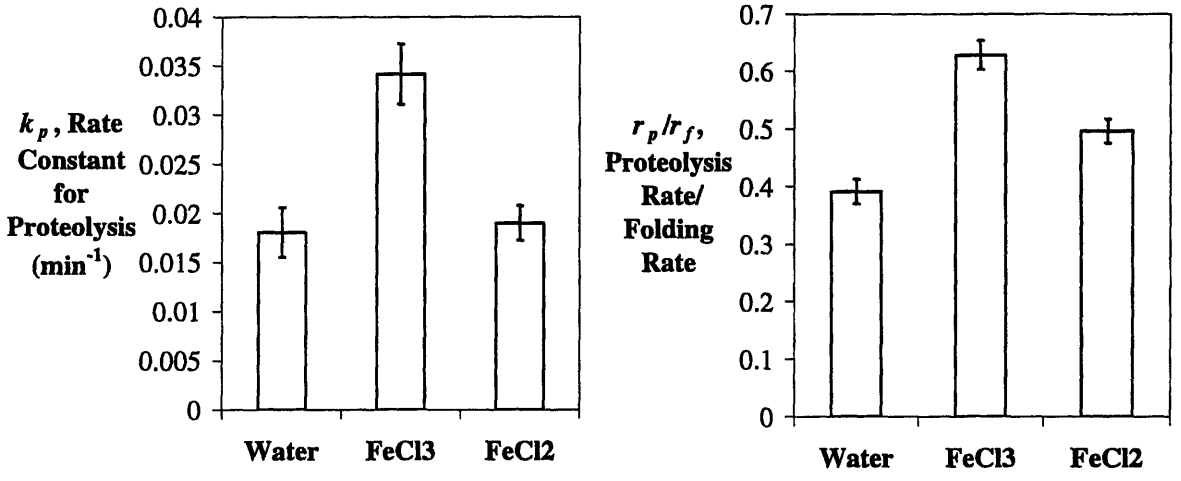


Figure 6.6: Kinetic Parameters for α1-Antitrypsin Degradation in Cultures Supplemented with Sterile-Filtered Iron
Model parameters for data in Figure 6.5. Cultures were supplemented with Water (control), 500 μM sterile-filtered $FeCl_3$, and 500 μM sterile-filtered $FeCl_2$, and were induced in oxygen. *kp* is the pseudo-first-order rate constant for α_1AT proteolysis and r_p/r_f is the ratio of the rate of proteolysis and the rate of folding. Parameters and confidence intervals were calculated as described in Section 3.8.3.

Compared with the O2 culture in the BL21 experiment, the iron-supplemented O2 culture had a significantly lower rate constant for proteolysis, k_p. As expected, iron had the largest effect on the O2 culture and the least effect on the N2 culture.

6.2.4 Additional Validation of the Effects of Iron

Because the cultures that show significant differences in α_1AT degradation are those supplemented with autoclaved iron, it must be determined whether the presence of iron in the samples impacts the sample analysis. During the analysis, each sample was centrifuged in order to separate cells and medium; however, the iron in the medium consistently settled with the cells and was present throughout the sample analysis. This additional iron complicated sample loading and increased variability in loading volumes. An additional validation experiment was performed in order to determine whether this component altered the SDS-PAGE analysis.

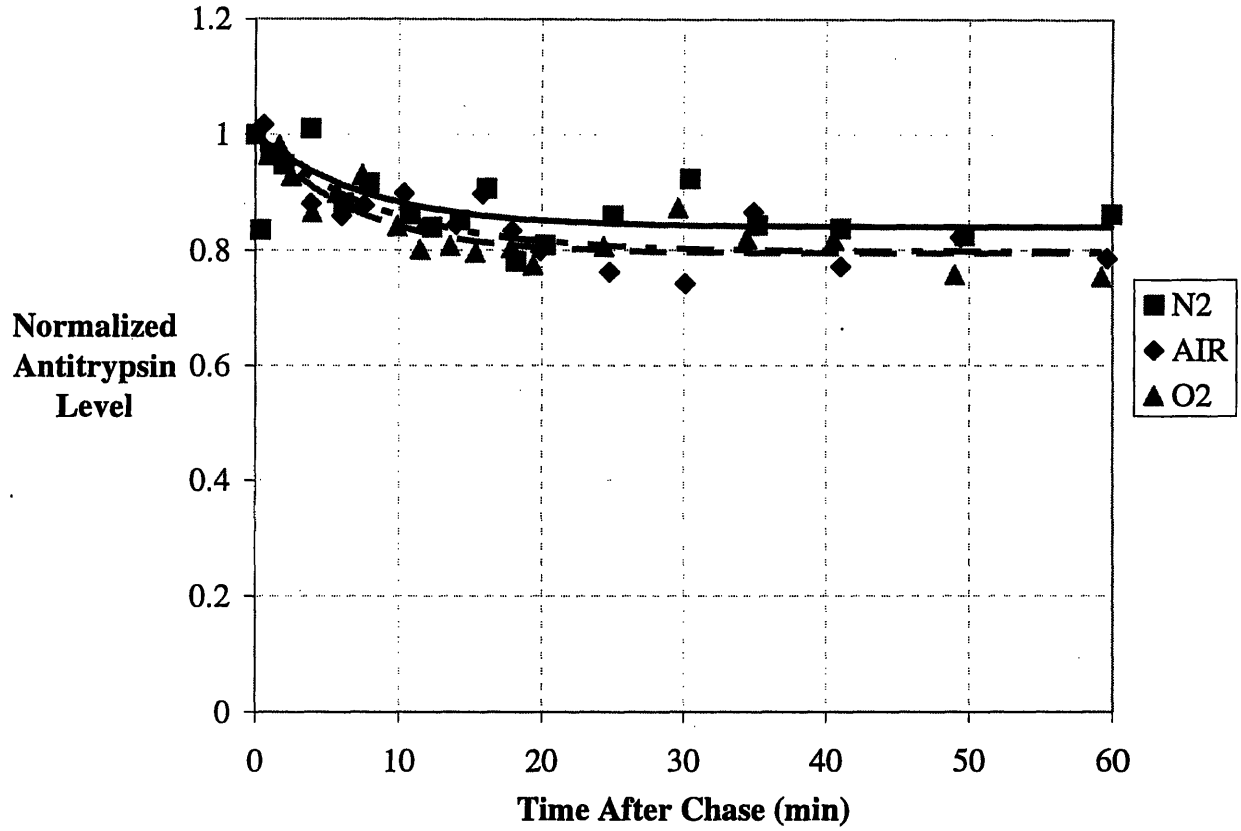


Figure 6.7: α1-Antitrypsin Degradation in Iron-Supplemented Cultures with Varying Aeration
One culture was grown in M9 medium supplemented with 500 μM autoclaved $FeCl_2$. The culture was split and induced in nitrogen, air, and oxygen. After 60 min of induction, recombinant α_1AT was pulse-chase labeled with ^{35}S-methionine. Degradation was monitored for the next 60 min.

One 100-mL culture was grown in unsupplemented M9 medium, and 6 mL of each culture was placed in two bubbler tubes. Both cultures were induced and bubbled with pure O_2. Samples were taken as normal from both tubes, except that the sample tubes for one of the cultures already contained 15 μL of 5-mM autoclaved $FeCl_2$ stock solution. While neither of these cultures was induced in iron, the sample analysis for one of these cultures was performed in the presence of iron. Otherwise, the samples were identical. Figure 6.9 shows the degradation profiles of these two cultures. Both profiles were identical in the rate constants of degradation (k_p) and r_p/r_f ratios (Figure 6.10). Therefore, degradation of α_1AT is not an artifact of iron in the SDS-PAGE analysis. However, it should be noted that the presence of iron clearly increased the variability of the sample analysis, as evidenced both by the scatter of data around the degradation curve in Figure 6.9 and the size of the confidence intervals in Figure 6.10. The iron pellet in the

samples made consistent sample loading difficult and likely increased the variability of the SDS-PAGE analysis.

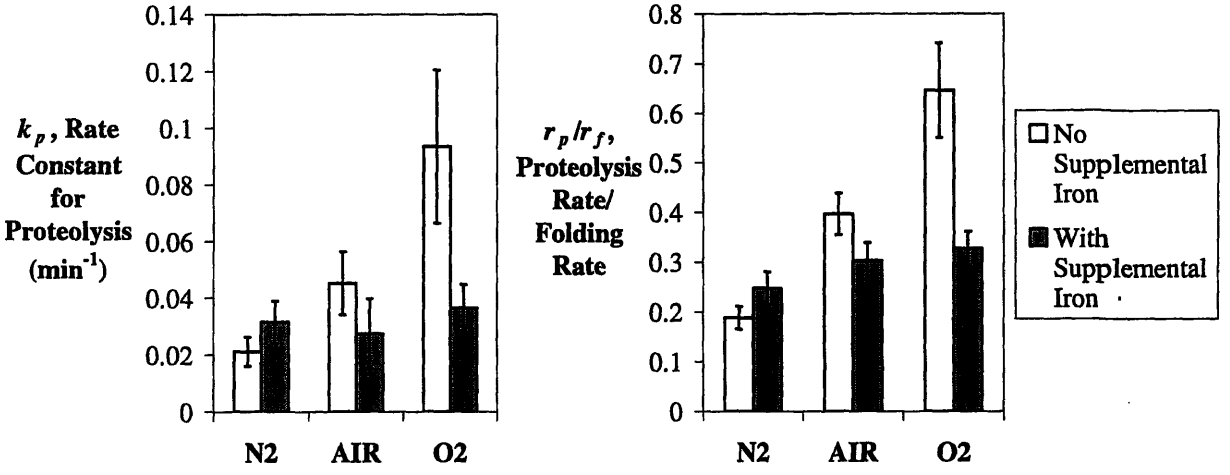


Figure 6.8: Kinetic Parameters for α1-Antitrypsin Degradation in Cultures Supplemented with $FeCl_2$
Model parameters for data in Figure 6.7. k_p is the pseudo-first-order rate constant for α_1AT proteolysis and r_p/r_f is the ratio of the rate of proteolysis and the rate of folding. Parameters and confidence intervals were calculated as described in Section 3.8.3.

6.2.5 Production of Recombinant α1-Antitrypsin in Iron-Supplemented Medium

Based on the pulse-chase experiments in the preceding chapters, iron supplementation was shown to alleviate the degradation of α_1AT and remove its oxygen dependence. However, these data were collected for α_1AT produced 60 min after induction, and degradation was not examined at other time points. The ultimate test of the benefit of iron supplementation was to observe the total production of α_1AT.

As described in Section 5.1, a 400-mL culture was grown from OD_{600} of 0.05 to 0.7, except the medium was supplemented with autoclaved $FeCl_2$ stock to a final concentration of 500 μM. To account for the absorbance of the medium, a flask of medium was also grown under the same conditions but was not inoculated. All other experimental steps were performed as described in Section 5.1. The culture was simultaneously induced and split into three smaller cultures under pure nitrogen, air, and pure oxygen. Samples were collected for microarray analysis over the next 90 min and after that time, the cultures were harvested and the α_1AT

activity assay was performed on the soluble extract. The microarray analysis of these samples is reserved for Section 6.3.

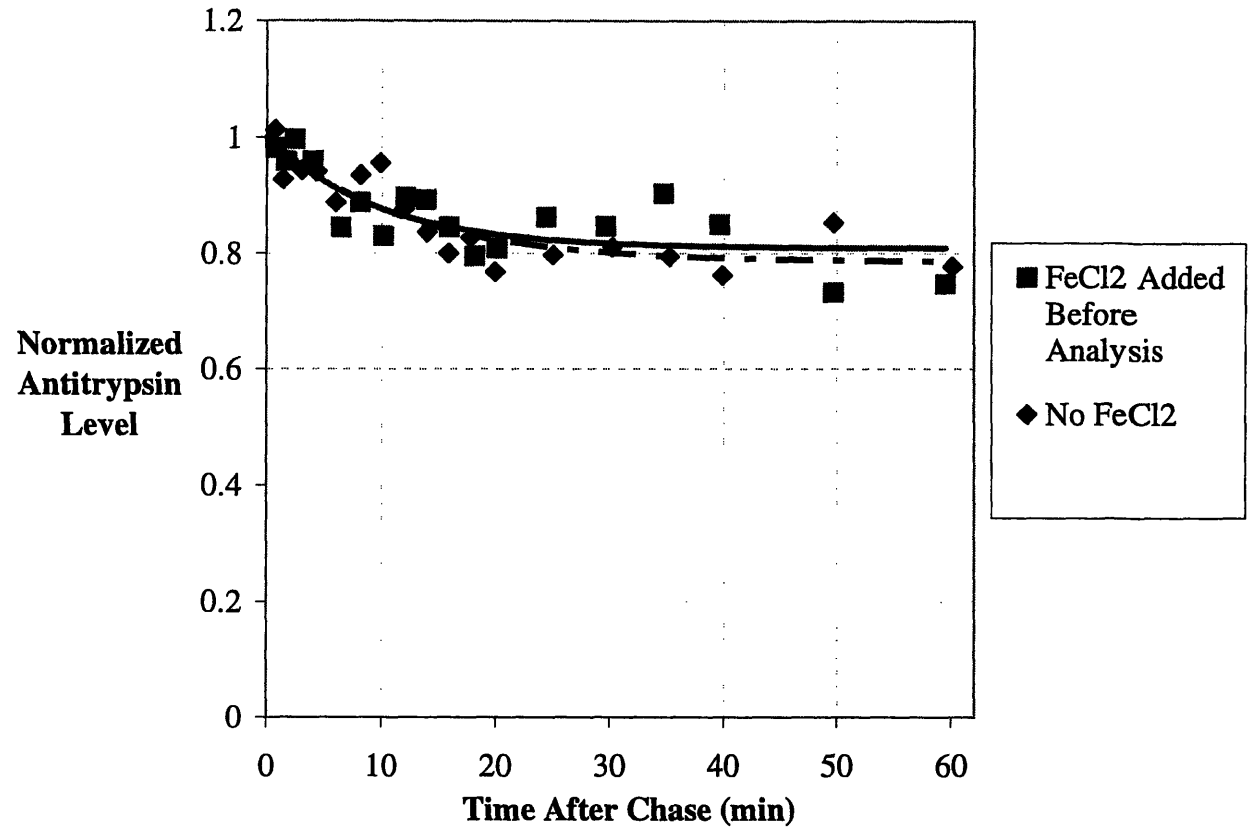


Figure 6.9: Effect of Iron during Analysis of α1-Antitrypsin Degradation
One culture was grown in M9 medium. The culture was split in two and induced in oxygen. After 60 min of induction, recombinant α_1AT was pulse-chase labeled with ^{35}S-methionine. Degradation was monitored for the next 60 min. Samples from one culture were collected in a fresh 1.5-mL tube, while samples from the other culture were collected in a 1.5-mL tube containing $FeCl_2$.

Figure 6.11 compares the specific activity values (on a per cell basis) of the unsupplemented and iron-supplemented cultures. Surprisingly, there was no difference in activity levels between the two conditions. An explanation for this observation can be found in the model of α_1AT degradation. The model in Figure 6.12 can be described in mathematical terms as follows.

$$\frac{dX}{dt} = \mu X \tag{6.1a}$$

$$\frac{dI}{dt} = r_t X - k_f I - k_p I \tag{6.1b}$$

$$\frac{dN}{dt} = k_f I \tag{6.1c}$$

In this model, t represents time, X represents the biomass concentration, I represents the concentration of a folding intermediate of α_1AT, N represents the concentration of native α_1AT, μ is the specific growth rate of the culture, r_t is the rate of α_1AT translation, k_f is the rate constant for α_1AT folding, and k_p is the pseudo-first order rate constant for α_1AT proteolysis. This is the same model that was applied to analyze the pulse-chase experiments (Laska 2000), except for the addition of the biomass model in **(6.1a)** and the biomass correction to the rate of translation in **(6.1b)**. These additional features were not needed for the pulse-chase model, because α_1AT translation only occurred during the 3-min pulse period. *dX/dt* was essentially zero during that short period of time.

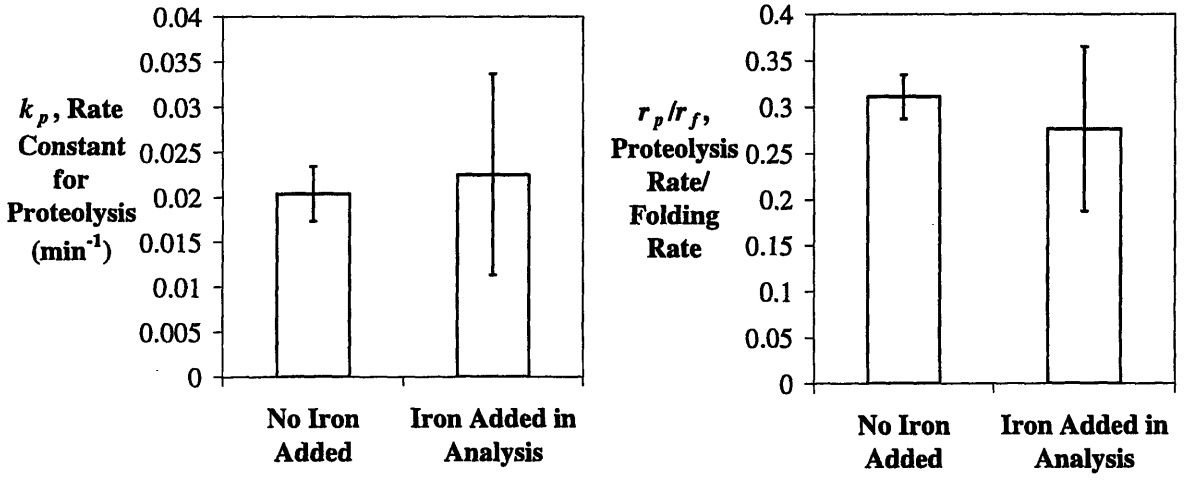


Figure 6.10: Effects of Iron during Analysis - Kinetic Parameters for α1-Antitrypsin Degradation
Model parameters for data in Figure 6.9. k_p is the pseudo-first-order rate constant for α_1AT proteolysis and r_p/r_f is the ratio of the rate of proteolysis and the rate of folding. Parameters and confidence intervals were calculated as described in Section 3.8.3.

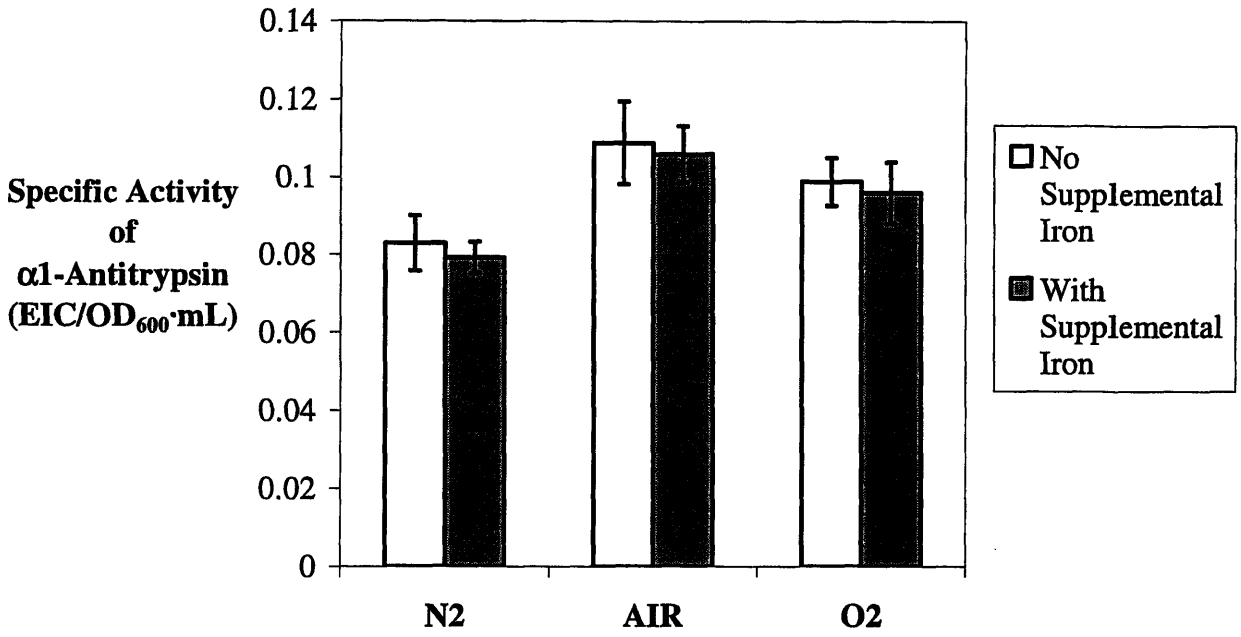


Figure 6.11: Specific Activities of α1-Antitrypsin with and without Supplemental Iron

Cultures were grown in media with and without iron supplementation, split into defined aeration environments (pure nitrogen, air, and pure oxygen), and induced to produce recombinant α_1AT for 90 min. Activity of the recombinant protein was measured as Elastase Inhibitory Capacity (EIC) and scaled by the OD_{600} and volume of the cultures at harvest.

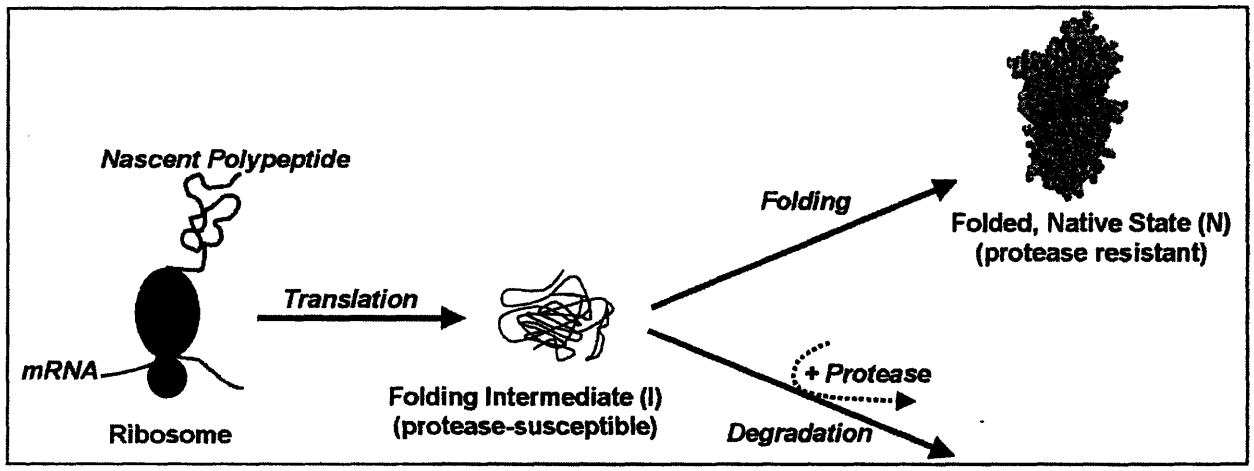


Figure 6.12: Model for α1-Antitrypsin Degradation

In this model, translation of α_1AT proceeds at rate r_t and produces a folding intermediate (I). This intermediate is susceptible to degradation, which proceeds with pseudo-first-order rate constant k_p. Alternatively, the folding intermediate can fold to form native α_1AT (N), which is resistant to degradation. This folding step proceeds with first-order rate constant k_f. Activity measurements presumably account for the levels of this native species (Laska 2000 - reprinted with permission).

This model of α_1AT production was applied over a range of k_f and k_p values. The predicted levels of both the N and I species after 90 min of induction are displayed in contour form in Figure 6.13. The values of k_f and k_p obtained from pulse-chase experiments, performed in oxygen-induced cultures both with and without supplemental iron (Figure 6.8), are also plotted in Figure 6.13. Combining the experimental and model results, it was found that levels of the N species would be relatively unchanged by addition of supplemental iron. Since this is the species presumably measured by the activity assay, this result is consistent with the data in Figure 6.11. Additionally, the model showed that supplemental iron would increase the levels of the I species. Because both k_p and k_f dropped upon addition of supplemental iron, levels of the folding intermediate, I, would accumulate according to **(6.1b)**. However, the increase in I was balanced by the decrease in k_f. According to **(6.1c)**, dN/dt would change very little, and consequently, N would remain unchanged. Thus, the model was consistent with the experimental observations.

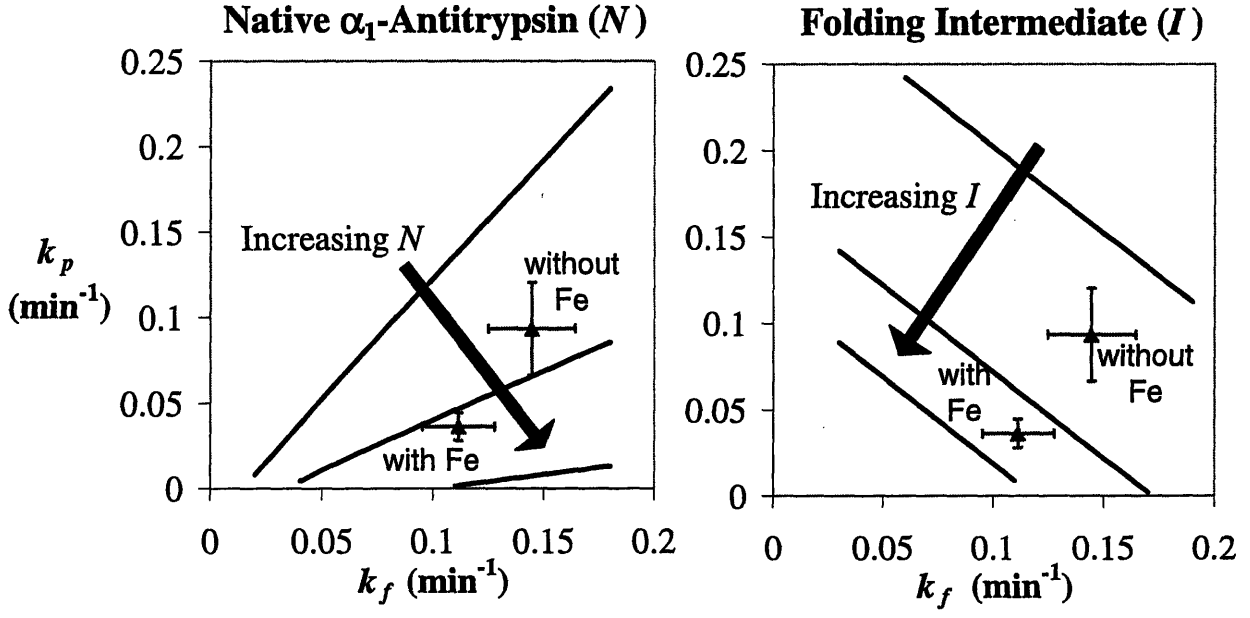


Figure 6.13: Contour Plots in k_f-k_p Space of α1-Antitrypsin Species after 90 min

Results from the α_1AT production model **(6.1)** are shown at different values of k_f and k_p. The following model parameters were used: μ = 0.007 min^{-1}, X_0 = 0.7 OD_{600}, I_0 = 0 EIC/mL, N_0 = 0 EIC/mL. Rate of translation, r_t, was chosen to be 0.0026 EIC/(OD_{600}·mL·min), which fit experimental observations at both 60 min and 90 min after induction. For I, contour lines are 0.8 EIC/mL apart and for N, contour lines are 5 EIC/mL apart. Plotted points represent the k_f and k_p values from O2 cultures with and without supplemented $FeCl_2$ (Figure 6.8).

Although supplementation of iron did not improve overall α_1AT yields under the conditions used here, it may be a useful strategy under other conditions. Iron supplementation was found to reduce degradation of α_1AT produced at 60 min, indicating that supplemented cultures are more robust to hyperoxic conditions. This strategy may also improve resistance to oscillations in oxygen level, such as would be experienced during scale up.

The ideal solution to α_1AT degradation would not only decrease k_p, but would also increase k_f. Overall α_1AT activity after 90 min of induction was independent of supplemental iron. This result was explained by the model of α_1AT production. Based on the pulse-chase results in Section 6.2.3, addition of iron decreased the rate constants for both folding and proteolysis. Modeling results showed that, together, this combination resulted in higher levels of the α_1AT folding intermediate (I), but little change in the levels of the native folded α_1AT species (N).

6.2.6 Supplementation of Iron-Sulfur Dependent Metabolites

The beneficial effects of supplemental iron on recombinant α_1AT production likely result from its ability to counter the effects of superoxide. For instance, iron-sulfur clusters that are degraded by superoxide may be regenerated using supplemental iron. This mechanism would certainly improve production of branched-chain amino acids (BCAA's) and biotin (and possibly thiamin as well) under highly aerobic conditions, since production of these metabolites depends on iron-sulfur clusters. In an attempt to link iron-sulfur clusters to the degradation of α_1AT, biotin, thiamin, isoleucine, leucine, and valine were supplemented to an induced culture. Adding the final product of these pathways should down-regulate production of iron-sulfur clusters, thereby reducing the sensitivity of the culture to superoxide. An associated decrease in α_1AT degradation would clearly establish a link between iron-sulfur clusters and α_1AT degradation. The pulse-chase protocol was carried out, and, at the time of induction, 60 µL of a 100× stock solution was added to bubbler-tube cultures, such that the final concentrations were 40 µg/mL each of isoleucine, leucine, and valine and 0.5 µg/mL each of biotin and thiamin.

The α_1AT degradation profiles for this culture and the unsupplemented control culture are shown in Figure 6.14. While the supplemented culture showed a decrease in the rate constant of proteolysis (k_p) (Figure 6.15), this was balanced by a decrease in the rate constant of folding (k_f). Overall, r_p/r_f ratios were unchanged by supplementation of iron-sulfur dependent metabolites.

This observation suggests that eliminating iron-sulfur proteins has no effect on α_1AT degradation. More likely, this result indicates that the four iron-sulfur proteins involved in production of these supplemented metabolites comprise only a small fraction of the total iron-sulfur protein content of the cell. Therefore, eliminating only these four does little to alter the culture's sensitivity to superoxide.

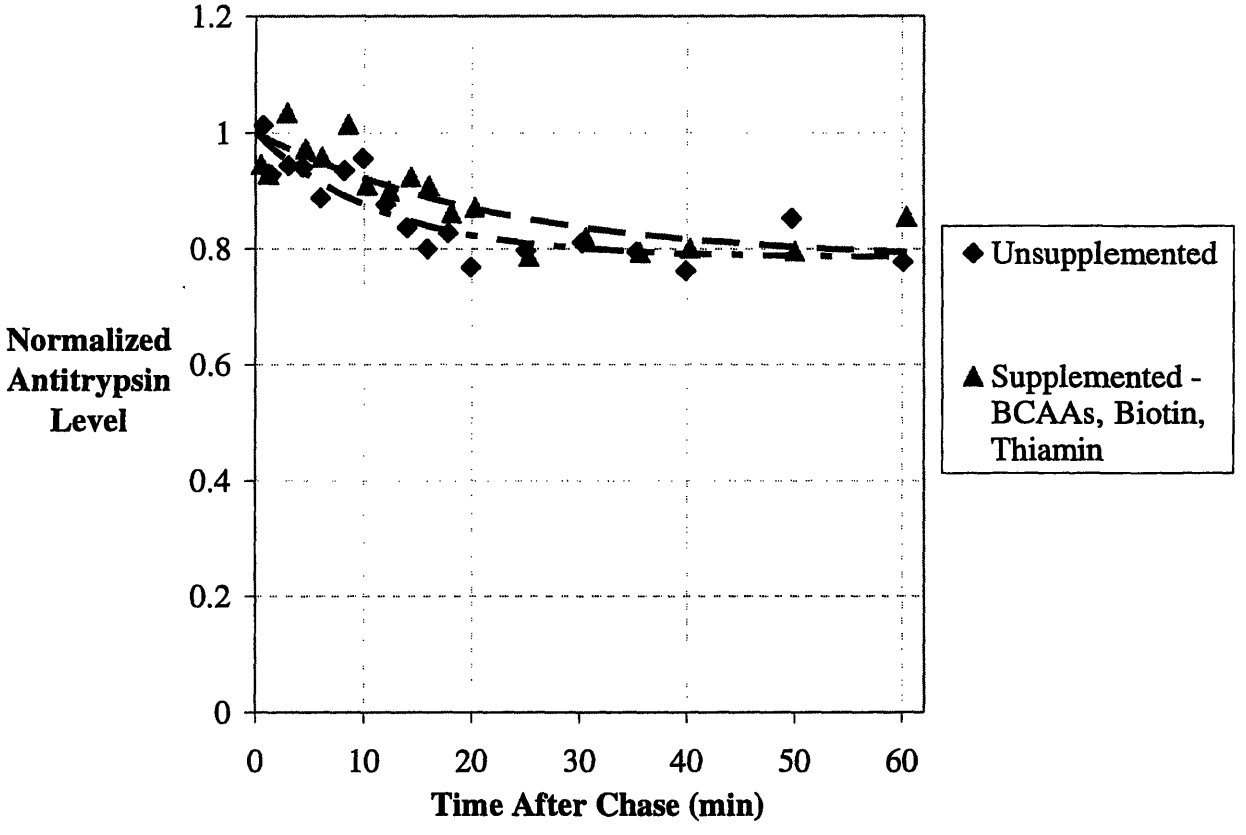


Figure 6.14: Effect of Iron-Sulfur Dependent Metabolites on α1-Antitrypsin Degradation
One culture was grown in M9 medium. The culture was split in two and induced in pure oxygen. One of these cultures was supplemented with branched-chain amino acids (to 40 μg/mL each), biotin (to 0.5 μg/mL), and thiamin (to 0.5 μg/mL) at the time of induction. After 60 min of induction, recombinant α_1AT was pulse-chase labeled with ^{35}S-methionine. Degradation was monitored for the next 60 min.

6.2.7 Summary of α1-Antitrypsin Production upon Iron Supplementation

Pulse-chase analysis of recombinant α_1AT degradation was used to determine rate constants for folding and proteolysis of α_1AT produced 60 min after induction. Supplementation of both iron(II) and iron(III) (sterilized by autoclaving) was found to alleviate degradation of

recombinant α_1AT in hyperoxic cultures. Moreover, supplementation of autoclaved iron(II) dramatically decreased the oxygen dependence of α_1AT degradation. Suprisingly, proteolysis was found to remain the same or worsen in cultures supplemented with sterile-filtered iron. In all of these cultures, an iron phosphate precipitate formed; however, the hematite (Fe_2O_3) formed by autoclaving appeared to be resistant to reaction with phosphate ions. It is proposed that iron from the iron-phosphate precipitate is unavailable to cells, while the iron in hematite is. Therefore, autoclaved iron(II) and iron(III) solutions, both of which contain hematite, alleviate α_1AT degradation.

Although iron-supplemented cultures did not show increased α_1AT yields, this observation was explained by applying the model of α_1AT folding and proteolysis developed previously (Laska 2000). Because iron supplementation decreased the rates of both folding and proteolysis, levels of the α_1AT folding intermediate increased, but the overall production of native α_1AT remained relatively unchanged. Yields of recombinant α_1AT may not improved upon iron supplementation, but this strategy may still be useful in improving the robustness of cultures to hyperoxic conditions.

Gene expression analysis was carried out on iron supplemented cultures in order to better understand the effects of iron and develop a hypothesis for how iron alleviates α_1AT degradation.

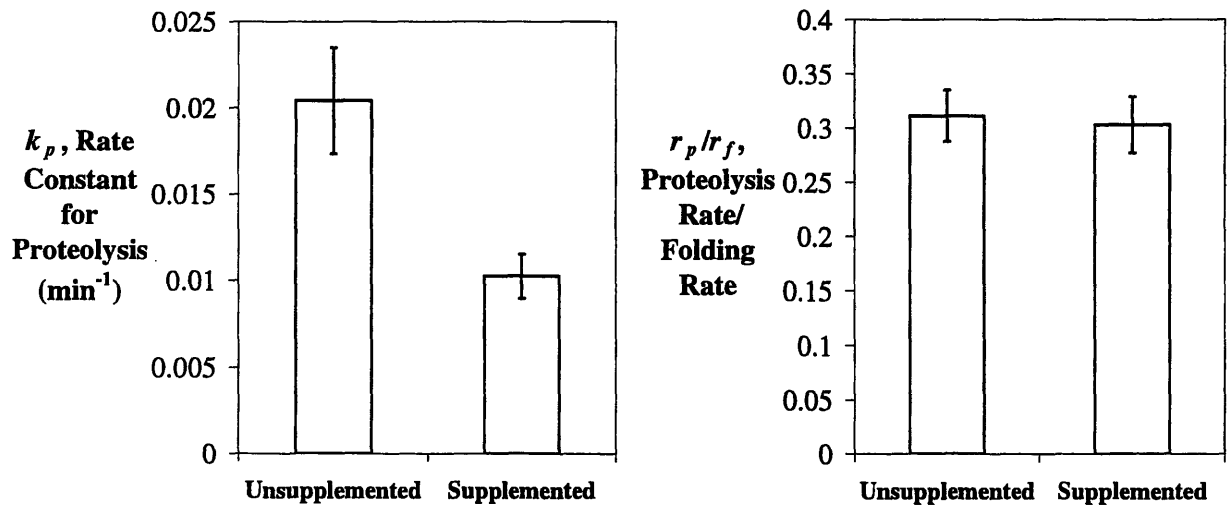


Figure 6.15: Effects of Iron-Sulfur Dependent Metabolites - Kinetic Parameters for α1-Antitrypsin Degradation

Model parameters for data in Figure 6.14. k_p is the pseudo-first-order rate constant for α_1AT proteolysis and r_p/r_f is the ratio of the proteolysis and folding rates. Parameters and confidence intervals were calculated as described in Section 3.8.3.

6.3 Global Effects of Iron Supplementation

The samples collected in the experiment described in Section 6.2.5 were analyzed using DNA microarrays. Samples were taken immediately before induction as well as at 10, 30, 60, and 90 min following induction from the N2, AIR, and O2 cultures. This data set (referred to as Fe) is directly comparable to that presented in Block A of Chapter 5, which is the same experiment performed without iron (referred to here as NoFe). DNA microarrays from Print E were used to analyze the Fe samples.

6.3.1 Analysis of Expression Data

Data from these thirteen arrays were analyzed in two different ways using the ANOVA method. Both analyses focused on differences between cultures induced without iron and those supplemented with iron. The first analysis ignored the effects of time and aeration (N_2, air, and O_2), while the second analysis considered them.

Table 6.1: Genes Showing Significant Overall Expression Changes in Response to Iron Supplementation
(opposite) 427 genes were selected as having significant and high expression changes in response to iron supplementation. The log ratio values given here can be considered to be the overall effect of iron supplementation on the signal ratio (log transformed on the base-2 scale), regardless of time or aeration.

Decreased Expression with Supplemental Iron												Increased Expression with Supplemental Iron								
Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio	Gene Name	Gene ID	Log Ratio
chaC	b1218	-3.9	fbp	b4232	-1.8	accB	b3255	-1.5	osmC	b1482	-1.3	yhcD	b3216	6.3	rfaL	b3622	2.2	b4273	b4273	1.5
acrA	b0463	-3.9	b0789	b0789	-1.8	thiG	b3991	-1.5	ybbF	b0524	-1.3	yrfA	b3392	5.5	yhbV	b3159	2.1	b3534	b3534	1.5
ytfR	b4228	-3.0	b0329	b0329	-1.8	cydC	b0886	-1.5	flgM	b1071	-1.3	yrfB	b3393	5.4	b3524	b3524	2.1	b1697	b1697	1.5
bioA	b0774	-2.9	glvB	b3682	-1.8	b2962	b2962	-1.5	bioH	b3412	-1.3	yhcE	b3217	4.8	b3226	b3226	2.1	arcA	b4401	1.4
rbfA	b3167	-2.7	aroA	b0908	-1.7	yjiK	b4333	-1.5	flgE	b1076	-1.3	srlA_2	b2703	4.3	b0447	b0447	2.1	b1373	b1373	1.4
rbsB	b3751	-2.6	pcm	b2743	-1.7	ybgF	b0742	-1.5	folE	b2153	-1.3	kdpD	b0695	4.1	rspA	b1581	2.1	ompC	b2215	1.4
b2342	b2342	-2.6	b1678	b1678	-1.7	yggZ	b2963	-1.5	yrfI	b3401	-1.3	b0853	b0853	4.0	pspE	b1308	2.1	sugE	b4148	1.4
atoB	b2223	-2.6	topA	b1274	-1.7	ycfD	b1128	-1.5	yehV	b2127	-1.3	b3515	b3515	4.0	b0689	b0689	2.1	ykgA	b0300	1.4
iadA	b4328	-2.6	def	b3287	-1.7	ybaM	b0466	-1.5	yebG	b1848	-1.3	yhgA	b3411	3.9	yhaK	b3106	2.0	b3472	b3472	1.4
yjhH	b4298	-2.4	b1437	b1437	-1.7	b0491	b0491	-1.5	hydN	b2713	-1.3	gusR	b1618	3.9	rfaJ	b3626	2.0	yccA	b0970	1.4
truB	b3166	-2.4	entC	b0593	-1.7	deoC	b4381	-1.5	hemC	b3805	-1.3	fimI	b4315	3.8	b0316	b0316	2.0	b1802	b1802	1.4
gcvH	b2904	-2.4	damX	b3388	-1.7	b1981	b1981	-1.5	ftsL	b0083	-1.3	b3046	b3046	3.7	fnr	b1334	2.0	yeiN	b2165	1.4
b2966	b2966	-2.4	speC	b2965	-1.7	b1679	b1679	-1.5	hemH	b0475	-1.3	b1119	b1119	3.7	yjgZ	b4277	2.0	yeiC	b2166	1.4
b0824	b0824	-2.4	ptsA	b3947	-1.7	yigN	b3832	-1.5	hemX	b3803	-1.3	yheI	b3331	3.6	yafX	b0248	2.0	yhhG	b3481	1.4
kbl	b3617	-2.4	metR	b3828	-1.7	aceE	b0114	-1.5	sppA	b1766	-1.3	b0539	b0539	3.5	focA	b0904	2.0	yhgK	b3420	1.4
acs	b4069	-2.4	xerD	b2894	-1.7	yjiA	b4352	-1.5	sdhD	b0722	-1.3	b1772	b1772	3.4	b0510	b0510	2.0	cyoA	b0432	1.4
entA	b0596	-2.3	leuB	b0073	-1.7	yebK	b1853	-1.5	yabB	b0081	-1.3	b1328	b1328	3.4	celB	b1737	1.9	yhcM	b3232	1.4
snpF	b1290	-2.3	yjbK	b4046	-1.7	hisF	b2025	-1.5	yecD	b1867	-1.3	uvrY	b1914	3.4	yjdH	b4125	1.9	ybeF	b0629	1.3
ygiG	b3058	-2.3	aroC	b2329	-1.7	b2745	b2745	-1.5	rffH	b3789	-1.3	mrcB	b0149	3.3	rfc	b2035	1.9	yjhP	b4306	1.3
nrdG	b4237	-2.3	b0711	b0711	-1.7	b3575	b3575	-1.5	uhpA	b3669	-1.3	yjhJ	b4300	3.3	yiiE	b3889	1.9	yjcB	b4060	1.3
cybC	b4236	-2.2	acpD	b1412	-1.7	b3713	b3713	-1.5	yejH	b2184	-1.3	kdpE	b0694	3.2	b4250	b4250	1.9	b0263	b0263	1.3
ilvG_1	b3767	-2.2	b1018	b1018	-1.7	b4212	b4212	-1.5	tnaB	b3709	-1.3	ygjK	b3080	3.2	csgG	b1037	1.9	yjfG	b4233	1.3
b2513	b2513	-2.2	accA	b0185	-1.7	ygcA	b2785	-1.5	pyrC	b1062	-1.3	yiaM	b3577	3.2	b0874	b0874	1.9	b1843	b1843	1.3
cynS	b0340	-2.2	yigK	b3824	-1.7	b0662	b0662	-1.5	glnD	b0167	-1.3	tdcA	b3118	3.1	b0514	b0514	1.9	yihE	b3859	1.3
gntR	b3438	-2.1	nikE	b3480	-1.7	yieE	b3712	-1.5	amn	b1982	-1.2	rfaY	b3625	3.1	yeeC	b2010	1.9	b2174	b2174	1.3
hisA	b2024	-2.1	yjfO	b4189	-1.7	yadP	b0147	-1.5	ruvA	b1861	-1.2	manA	b1613	3.1	tra5_2	b0541	1.8	yhjJ	b3527	1.3
potF	b0854	-2.1	b1604	b1604	-1.7	b4144	b4144	-1.4	b1452	b1452	-1.2	yraJ	b3144	3.1	yjcG	b4067	1.8	b2689	b2689	1.3
topB	b1763	-2.1	yeiH	b2158	-1.7	ygaG	b2687	-1.4	himA	b1712	-1.2	tra5_4	b2089	3.1	xylF	b3566	1.8	b2876	b2876	1.3
b2047	b2047	-2.1	truA	b2318	-1.7	b2898	b2898	-1.4	lhr	b1653	-1.2	yhcC	b3211	3.0	yjiD	b4326	1.8	secE	b3981	1.3
agaR	b3131	-2.1	lplA	b4386	-1.7	b0868	b0868	-1.4	b1837	b1837	-1.2	b1471	b1471	3.0	yhgL	b3421	1.8	arsC	b3503	1.2
b1438	b1438	-2.1	b2882	b2882	-1.7	malF	b4033	-1.4	rhaT	b3907	-1.2	yi21_5	b3044	3.0	mreD	b3249	1.7	b1411	b1411	1.2
aceK	b4016	-2.1	pdxH	b1638	-1.7	yheU	b3354	-1.4	htrC	b3989	-1.2	malZ	b0403	2.8	hofF	b3327	1.7	fic	b3361	1.2
yohF	b2137	-2.1	rffD	b3787	-1.7	sdhC	b0721	-1.4	yabJ	b0066	-1.1	b2003	b2003	2.8	hybG	b2990	1.7	yjbO	b4050	1.2
panB	b0134	-2.1	fabH	b1091	-1.7	b0212	b0212	-1.4	b1199	b1199	-1.1	trkG	b1363	2.8	atpE	b3737	1.7	b2886	b2886	1.2
iclR	b4018	-2.1	apt	b0469	-1.6	ppdD	b0108	-1.4	b0420	b0420	-1.1	fimF	b4318	2.8	rpmI	b1717	1.7	b0561	b0561	1.2
rffG	b3788	-2.0	pepT	b1127	-1.6	potG	b0855	-1.4	lpdA	b0116	-1.1	hipB	b1508	2.8	yiaA	b3562	1.7	cyoB	b0431	1.2
yffH	b2467	-2.0	ilvN	b3670	-1.6	b2463	b2463	-1.4	b0949	b0949	-1.1	yhhQ	b3471	2.7	b0346	b0346	1.7	b1691	b1691	1.2
ychP	b1220	-2.0	yjfI	b4181	-1.6	ftsN	b3933	-1.4	mfd	b1114	-1.1	yhaL	b3107	2.7	yiaR	b3582	1.7	caiB	b0038	1.2
b0516	b0516	-2.0	hflX	b4173	-1.6	rimL	b1427	-1.4	b2681	b2681	-1.1	b2817	b2817	2.7	tolR	b0738	1.7	aldA	b1415	1.2
gcvR	b2479	-2.0	aspS	b1866	-1.6	b2968	b2968	-1.4	b1758	b1758	-1.1	insB_3	b0274	2.7	b1775	b1775	1.7	ycaD	b0898	1.1
yceC	b1086	-2.0	mukB	b0924	-1.6	b1340	b1340	-1.4	b2488	b2488	-1.0	b2335	b2335	2.6	yaeG	b0162	1.7	hynA	b0972	1.1
ntpA	b1865	-2.0	b0848	b0848	-1.6	ydiE	b1705	-1.4	b0574	b0574	-1.0	fimC	b4316	2.6	b4286	b4286	1.7	nuoJ	b2280	1.1
yggD	b2941	-1.9	b1115	b1115	-1.6	yhbM	b3163	-1.4	fixX	b0044	-1.0	yehA	b2108	2.6	yhjA	b3518	1.7	tap	b1885	1.1
rimI	b4373	-1.9	cysG	b3368	-1.6	yjiH	b4330	-1.4	b0487	b0487	-1.0	glgS	b3049	2.5	tra8_1	b0256	1.6	b2955	b2955	1.1
accD	b2316	-1.9	nfo	b2159	-1.6	b1436	b1436	-1.4	recR	b0472	-1.0	glgP	b3428	2.5	b2299	b2299	1.6	glpB	b2242	1.1
b1499	b1499	-1.9	hypE	b2730	-1.6	hslJ	b1379	-1.4	glpR	b3423	-1.0	b3047	b3047	2.4	atpI	b3739	1.6	yehI	b2118	1.1
eda	b1850	-1.9	yceB	b1063	-1.6	b2494	b2494	-1.4	hycH	b2718	-1.0	csgE	b1039	2.4	atpB	b3738	1.6	b0794	b0794	1.1
ydjA	b1765	-1.9	phoA	b0383	-1.6	narL	b1221	-1.4	zwf	b1852	-1.0	ykgD	b0305	2.4	yciC	b1255	1.6	yggB	b2924	1.1
bioB	b0775	-1.9	b1833	b1833	-1.6	b2510	b2510	-1.4	aceF	b0115	-1.0	b2004	b2004	2.4	nuoK	b2279	1.6	atpD	b3732	1.1
b1298	b1298	-1.9	xseA	b2509	-1.6	yijC	b3963	-1.4	sdhA	b0723	-1.0	b2921	b2921	2.4	yejB	b2178	1.6	b2870	b2870	1.1
dcp	b1538	-1.9	hfq	b4172	-1.6	yheT	b3353	-1.4	b1312	b1312	-1.0	b2088	b2088	2.4	yjhB	b4279	1.6	secG	b3175	1.1
sbcD	b0398	-1.9	aslB	b3800	-1.6	plsX	b1090	-1.4	dnaX	b0470	-0.9	leuL	b0075	2.3	yihQ	b3878	1.6	dmsA	b0894	1.1
ftsA	b0094	-1.8	ydiD	b1701	-1.6	b0609	b0609	-1.3	acnB	b0118	-0.9	acrE	b3265	2.3	b2059	b2059	1.6	fadL	b2344	1.0
dadX	b1190	-1.8	pepA	b4260	-1.5	metF	b3941	-1.3	dnaJ	b0015	-0.9	yrhB	b3446	2.3	cyoC	b0430	1.6	fsr	b0479	1.0
ybiK	b0828	-1.8	pepN	b0932	-1.5	b2672	b2672	-1.3	sucB	b0727	-0.9	yrfD	b3395	2.3	b1320	b1320	1.6	dppF	b3540	1.0
gltA	b0720	-1.8	rpiA	b2914	-1.5	b2054	b2054	-1.3	pstB	b3725	-0.9	yrbG	b3196	2.3	ybgE	b0735	1.6	b1880	b1880	1.0
cytR	b3934	-1.8	yjfR	b4192	-1.5	argC	b3958	-1.3	flgH	b1079	-0.9	yraH	b3142	2.3	slp	b3506	1.6	b1567	b1567	1.0
b4045	b4045	-1.8	ispA	b0421	-1.5	b1013	b1013	-1.3	flgA	b1072	-0.9	galS	b2151	2.3	yhhZ	b3442	1.5	fruK	b2168	0.8
ygbD	b2711	-1.8	ybbA	b0495	-1.5	purT	b1849	-1.3	holA	b0640	-0.7	feoB	b3409	2.3	yhhP	b3470	1.5	ftsZ	b0095	0.8
b2673	b2673	-1.8	b3034	b3034	-1.5							yigW_2	b3841	2.2	yzgL	b3427	1.5	proP	b4111	0.8
												gatR_1	b2087	2.2	b1626	b1626	1.5	appC	b0978	0.8
												fimG	b4319	2.2	b3856	b3856	1.5	bacA	b3057	0.8
												tdcC	b3116	2.2	cheA	b1888	1.5	pstC	b3727	0.8

6.3.1.1 Two-Treatment ANOVA

In the first analysis, data from iron-supplemented cultures were combined with data from unsupplemented cultures, and all were normalized together. The presence and absence of iron supplementation were regarded as the only two treatments. Samples with different aeration conditions and at different time points were simply considered to be repeated measurements of one another. Thus, the data set had two treatments with 13 blocks each. This two-treatment ANOVA identified genes that showed consistent expression differences with and without iron supplementation, regardless of aeration and time. Because the 13 samples within each data set (Fe and NoFe) were treated as repeats, the overall significance of observed differential expression increased dramatically. As a result, a large number of genes were identified as having differential expression. Only at a significance level of 0.999, was a manageable number of genes selected. At this level, 427 genes were found to show differential expression; 239 showed decreased expression, while 189 showed increased expression (Table 6.1).

6.3.1.2 26-Treatment ANOVA

The second analysis regarded each of the 26 arrays as a separate treatment, and the entire data set as a single block. Only Fe vs. NoFe treatment comparisons with the same aeration and at the same time point (13 comparisons) were considered in this 26-treatment ANOVA. Based on the global significance test at a level of 0.95, 1,166 genes were found to show significant differential expression in at least one comparison. Comparing only the zero time point, a group of 114 genes was found to have significant iron-dependent differential expression (Table 6.2 and Table 6.3). Since iron was present throughout the growth phases of these cultures, this list represents genes that were affected purely by iron supplementation, without the effects of α_1AT production and changing aeration conditions.

Table 6.2: Genes with Decreased Expression before Induction in Response to Iron Supplementation
(opposite) 48 genes were selected as having significantly decreased expression in response to iron supplementation. Descriptions of the gene product were taken from the EcoCyc database. The log ratio values given here are for the Fe-0 vs. NoFe-0 comparison.

Gene Name	Gene ID	Gene Product	Expression Value
allC	b0516	allantoate amidohydrolase	-4.5
phnL	b4096	phosphonates transport ATP-binding protein PhnL	-3.8
wcaJ	b2047	putative colanic acid biosynthsis UDP-glucose lipid carrier transferase	-3.6
kdpA	b0698	potassium ion P-type ATPase transporter	-3.5
ygbD	b2711	flavorubredoxin reductase	-3.4
pflE	b0824	putative pyruvate formate-lyase 2 activating enzyme	-3.2
phnJ	b4098	phosphonate metabolism	-3.1
yjiH	b4330	putative membrane protein	-3.1
ytfR	b4228	YtfQ/YtfR/YtfS/YtfT/YjfF ABC transporter	-3.1
nrdG	b4237	anaerobic nucleoside-triphosphate reductase activating system	-3.0
ybhO	b0789	cardiolipin synthase 2	-3.0
yohF	b2137	putative oxidoreductase	-2.9
fhuC	b0151	iron (III) hydroxamate ABC transporter	-2.8
ppdD	b0108	prelipin peptidase dependent protein	-2.8
ppdB	b2825	prepilin peptidase dependent protein B	-2.8
acrA	b0463	trans-cplx-201	-2.8
yjgK	b4252	conserved hypothetical protein	-2.7
entA	b0596	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase	-2.7
yedS	b1964	putative outer membrane protein	-2.7
gcvH	b2904	dihydrolipoyl-GcvH-protein	-2.7
fhuD	b0152	iron (III) hydroxamate ABC transporter	-2.7
fbp	b4232	fructose 1,6 bisphosphatase monomer	-2.7
cysG	b3368	uroporphyrinogen methyltransferase / 1,3-dimethyluroporphyrinogen III dehydrogenase / siroheme ferrochelatase	-2.7
ygiZ	b3027	conserved hypothetical protein	-2.6
sapF	b1290	peptide uptake ABC transporter	-2.6
ycjL	b1298	probable amidotransferase subunit	-2.6
truB	b3166	tRNA pseudouridine synthase	-2.6
miaA	b4171	delta(2)-isopentenylpyrophosphate tRNA-adenosine transferase	-2.6
rbfA	b3167	ribosome-binding factor A	-2.6
cynS	b0340	cyanase monomer	-2.6
kbl	b3617	2-amino-3-ketobutyrate CoA ligase	-2.5
yeiH	b2158	putative membrane protein	-2.5
ydjA	b1765	conserved protein	-2.5
cybC	b4236	cytochrome b562 (soluble)	-2.5
gntR	b3438	HTH-type transcriptional regulator gntR	-2.5
ybbA	b0495	YbbA/YbbP ABC transporter	-2.5
nadD	b0639	nicotinate-mononucleotide adenylyltransferase	-2.5
ecfC	b2968	putative secretion protein for export	-2.5
nirB	b3365	large subunit of nitrite reductase	-2.4
ytfT	b4230	YtfQ/YtfR/YtfS/YtfT/YjfF ABC transporter	-2.4
ytfS	b4229	YtfQ/YtfR/YtfS/YtfT/YjfF ABC transporter	-2.4
uhpA	b3669	UhpA-Phosphorylated transcriptional activator	-2.4
ftsA	b0094	cell division protein, complexes with FtsZ	-2.4
nikE	b3480	nickel ABC transporter / Transporters	-2.3
rpiA	b2914	ribose-5-phosphate isomerase A	-2.3
gltA	b0720	citrate synthase monomer	-2.2
yjfF	b4231	YtfQ/YtfR/YtfS/YtfT/YjfF ABC transporter	-2.2
yqgA	b2966	putative transport protein	-2.2

Gene Name	Gene ID	Gene Product	Expression Value
yhcD	b3216	putative outer membrane protein	6.3
yidL	b3680	putative ARAC-type regulatory protein	4.8
yrfB	b3393	conserved hypothetical protein	4.7
ybjN	b0853	putative sensory transduction regulator	4.7
ydjH	b1772	putative kinase	4.3
yhcE	b3217	hypothetical protein	4.2
kdpD	b0695	Sensor protein KdpD	3.9
yqiG	b3046	putative membrane protein	3.9
yddK	b1471	putative glycoportein	3.7
ynfA	b1582	inner membrane protein	3.7
ybcC	b0539	putative exonuclease (EC 3.1.11.3) of lambda	3.6
yhgA	b3411	hypothetical protein	3.6
hypF	b2712	HypF transcriptional regulator	3.6
yqiH	b3047	putative membrane protein	3.5
yheI	b3331	putative protein secretion protein for export	3.4
leuL	b0075	leu operon leader peptide	3.3
yhhQ	b3471	hypothetical protein; gene is a predicted member of the purine regulon	3.2
rfaY	b3625	lipopolysaccharide core biosynthesis	3.2
pspD	b1307	phage shock protein localized to the peripheral inner membrane	3.2
yehA	b2108	putative type-1 fimbrial protein	3.2
tra5_3	b1026	putative transposase for insertion sequence IS3	3.2
tra5_4	b2089	putative transposase for insertion sequence IS3	3.2
yrfA	b3392	conserved hypothetical protein	3.2
insB_3	b0274	IS1 protein InsB	3.1
yagU	b0287	conserved protein	3.0
malY	b1622	enzyme that may degrade or block biosynthesis of endogenous mal inducer, probably aminotrasferase	3.0
yeeT	b2003	hypothetical protein	3.0
fimI	b4315	fimbrial protein	3.0
yi21_4	b2861	IS21 protein	2.9
ycgE	b1162	putative transcriptional regulator	2.9
nanR	b3226	NanR transcriptional regulator	2.9
yhaL	b3107	hypothetical protein	2.9
hybG	b2990	hydrogenase-2 operon protein: may effect maturation of large subunit of hydrogenase-2	2.8
uvrY	b1914	UvrY- Phosphorylated transcriptional regulator	2.8
pphB	b2734	protein phosphatase 2 / protein-tyrosine-phosphatase / phosphoprotein phosphatase	2.8
yhhM	b3467	putative receptor	2.8
yagL	b0278	DNA-binding protein	2.8
yi21_5	b3044	IS21 protein	2.7
yjfQ	b4191	putative DEOR-type transcriptional regulator	2.7
yi52_1	b0259	IS5 protein	2.7
yjgD	b4255	conserved hypothetical protein	2.7
ycjZ	b1328	putative transcriptional regulator LYSR-type	2.7
ydgK	b1626	putative oxidoreductase	2.7
fruA	b2167	EIIFru	2.7
kdpE	b0694	KdpE-Phosphorylated transcriptional activator	2.7
atpI	b3739	membrane-bound ATP synthase , dispensable protein, affects expression of atpB	2.7
yfiA	b2597	stationary phase translation inhibitor and ribosome stability factor	2.7
ybdO	b0603	putative transcriptional regulator LYSR-type	2.6
dacB	b3182	D-alanyl-D-alanine carboxypeptidase, fraction B; penicillin-binding protein 4	2.6

Table 6.3: Genes with Increased Expression before Induction in Response to Iron Supplementation

(opposite and below) 66 genes were selected as having significantly increased expression in response to iron supplementation. Descriptions of the gene product were taken from the EcoCyc database. The log ratio values given here are for the Fe-0 vs. NoFe-0 comparison.

Gene Name	Gene ID	Gene Product	Expression Value
ybgE	b0735	conserved hypothetical protein	2.6
rplP	b3313	50S ribosomal subunit protein L16	2.6
gutM	b2706	GutM transcriptional activator	2.6
srlA_2	b2703	EIIGut glucitol PTS permease subunit	2.6
tdcA	b3118	TdcA transcriptional activator	2.6
yrhB	b3446	hypothetical protein	2.5
srlD	b2705	sorbitol-6-phosphate dehydrogenase	2.5
atpE	b3737	ATP synthase c subunit	2.5
yrbG	b3196	YrbG CacA transporter	2.5
rplS	b2606	50S ribosomal subunit protein L19	2.5
intF	b0281	putative phage integrase	2.5
glgS	b3049	glycogen biosynthesis, rpoS dependent	2.5
fabA	b0954	beta-hydroxydecanoyl-ACP dehydrase / trans-2-decenoyl-ACP isomerase	2.5
secE	b3981	sec-secretion-cplx	2.5
hyaB	b0973	hydrogenase I	2.4
atpB	b3738	ATP synthase a subunit	2.2
hslT	b3687	heat shock protein	2.2

Consideration of expression differences in the AIR and O2 cultures expanded the lists in Table 6.2 and Table 6.3 to 714 genes. As mentioned previously, expression differences in these aerobic cultures are of most interest in this work. Because of the size of this list and the number of expression values involved, the genes are not listed here. Instead, an analysis of these genes based on the EcoCyc gene groups is presented in Table 6.4, Table 6.5, and Table 6.6.

Table 6.4: Gene Groups with Increased Expression upon Iron Supplementation

Genes that showed increased expression in either the zero-time-point, the AIR culture, or the O2 culture in response to iron supplementation were grouped according to information from the EcoCyc database. The 40 groups shown here exhibited increased expression for multiple genes. γ_{UP} is the number of genes that showed significantly increased expression with supplemental iron. γ_{TOTAL} is the total number of genes in the group. The remaining genes did not show significant changes (PC=protein complex, PW=pathway, RG=regulon, TU=transcription unit).

EcoCyc Gene Group	γ_{UP}	γ_{TOTAL}
PC-ATP synthase	2	8
PC-cytochrome o ubiquinol oxidase	3	4
PC-F-O complex of ATP synthase	2	3
PC-galactose ABC transporter	2	3
PC-glycerol-3-phosphate-dehydrogenase, anaerobic	2	3
PC-sece/secg/secy-cplx	2	3
PC-sec-secretion-cplx	2	8
PW-ArcAB Two-Component Signal Transduction System	2	3
PW-Chemotactic Signal Transduction System	2	4
PW-CreCB Two-Component Signal Transduction System	2	3
PW-fatty acid elongation -- unsaturated	2	5
PW-glycogen biosynthesis	2	3
PW-glycogen degradation	2	7
PW-KdpDE Two-Component Signal Transduction System	2	2
RG-CpxR transcriptional dual regulator	3	10
RG-CsgD transcriptional activator	2	6.
RG-GalS transcriptional repressor	3	4
RG-GlpR transcriptional repressor	2	8
RG-OmpR-Phosphorylated transcriptional dual regulator	3	9
TU-atpIBEFHAGDC	3	9
TU-celABCDF-ydjC	3	6
TU-creABCD	2	4
TU-csgDEFG	2	4
TU-cyoABCDE	3	5
TU-feoAB	2	2
TU-fimAICDFGH	5	7
TU-fimBEAICDFGH	5	9
TU-fimEAICDFGH	5	8
TU-glgCAP	2	3
TU-glpABC	2	3
TU-kdpDE	2	2
TU-mglBAC	2	3
TU-rfaQGPSBIJYZK	3	10
TU-rpsJ-rplCDWB-rpsS-rplV-rpsC-rplP-rpmC-rpsQ	2	11
TU-rpsP-rimM-trmD-rplS	2	4
TU-rtcBA	2	2
TU-srlAEBD-gutM-srlR-gutQ	4	7
TU-tdcABCDEFG	2	8
TU-ybgC-tolQRA	3	4
TU-yiaKLMNOPQRS	3	9

Table 6.5: Gene Groups with Decreased Expression upon Iron Supplementation

(opposite) Genes that showed decreased expression in either the zero-time-point, the AIR culture, or the O2 culture in response to iron supplementation were grouped according to information from the EcoCyc database. The 65 groups shown here exhibited decreased expression for multiple genes. γ_{DOWN} is the number of genes that showed significantly decreased expression with supplemental iron. γ_{TOTAL} is the total number of genes in the group. The remaining genes did not show significant changes (PC=protein complex, PW=pathway, RG=regulon, TU=transcription unit).

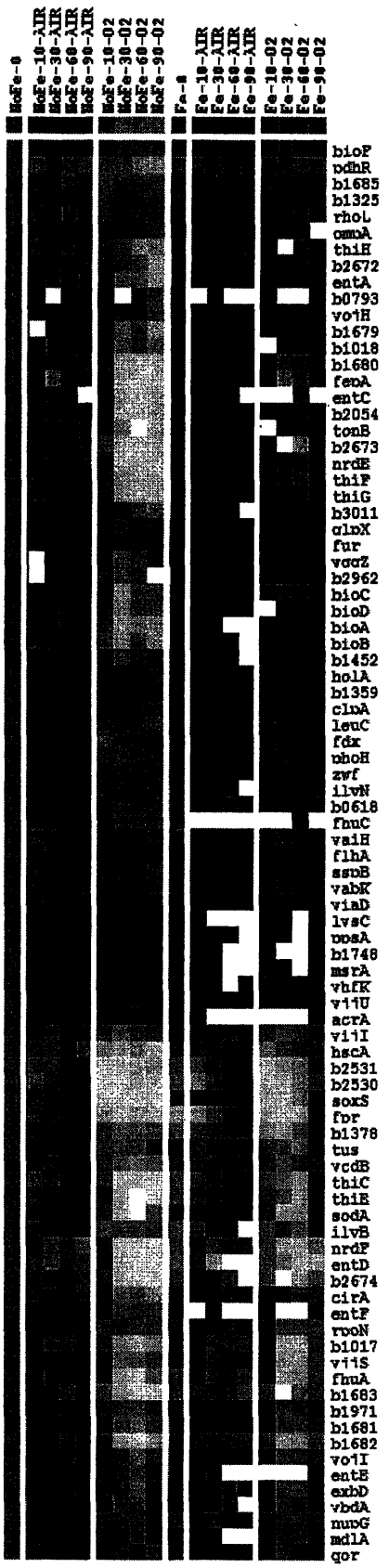
EcoCyc Gene Group	γ_{DOWN}	γ_{TOTAL}
PC-acetyl CoA carboxylase	2	4
PC-acetyl-CoA carboxyltransferase	2	2
PC-anaerobic nucleoside-triphosphate reductase activating system	2	3
PC-EIIABCFrv	2	4
PC-iron (III) hydroxamate ABC transporter	2	3
PC-maltose ABC transporter	2	4
PC-membrane-bound subcomplex of succinate dehydrogenase	2	2
PC-nitrite reductase	2	2
PC-succinate dehydrogenase	2	4
PC-YtfQ/YtfR/YtfS/YtfT/YjfF ABC transporter	4	5
PW-arginine biosynthesis II	2	11
PW-biotin biosynthesis I	4	4
PW-conversion of succinate to propionate	2	3
PW-de novo biosynthesis of purine nucleotides I	2	22
PW-de novo biosynthesis of pyrimidine ribonucleotides	2	11
PW-Entner-Doudoroff pathway	2	3
PW-fatty acid biosynthesis -- initial steps	3	11
PW-fatty acid oxidation pathway	3	8
PW-folate biosynthesis	2	11
PW-gluconeogenesis	2	16
PW-glyoxylate degradation	2	6
PW-histidine biosynthesis I	3	8
PW-isoleucine biosynthesis I	3	10
PW-leucine biosynthesis	2	6
PW-mixed acid fermentation	2	25
PW-pyridine nucleotide biosynthesis	2	5
PW-respiration (anaerobic)	4	26
PW-superpathway of arginine and polyamine biosynthesis	3	15
PW-superpathway of gluconate degradation	3	7
PW-superpathway of glycolysis and Entner-Doudoroff	3	17
PW-superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass	6	36
PW-superpathway of glyoxylate bypass and TCA	4	20
PW-superpathway of histidine, purine, and pyrimidine biosynthesis	7	43
PW-superpathway of leucine, valine, and isoleucine biosynthesis	4	15
PW-superpathway of oxidative and non-oxidative branches of pentose phosphate pathway	2	9
PW-superpathway of ribose and deoxyribose phosphate metabolism	2	10
PW-TCA cycle -- aerobic respiration	4	18
PW-thiamine biosynthesis	3	8
PW-UhpBA Two-Component Signal Transduction System	2	3
PW-valine biosynthesis	3	8
RG-ArgR-L-arginine transcriptional repressor	3	10
RG-BirA-bio-5'-AMP transcriptional repressor	5	5
RG-CRP transcriptional dual regulator	2	11
RG-CytR transcriptional dual regulator	2	11
RG-DeoR transcriptional repressor	2	6
RG-Fur transcriptional dual regulator	11	40
RG-MalT-Maltotriose-ATP transcriptional activator	2	9
RG-RNAPE-CPLX	2	18
RG-sigma19 factor	2	7
TU-atoDAE	2	3
TU-bioBFCD	4	4
TU-cynTSX	2	3
TU-deoCABD	2	4
TU-fecABCDE	2	5
TU-fhuACDB	2	4
TU-hisGDCBHAFI	3	8
TU-ilvLG_1G_2MEDA	4	7
TU-malEFG	2	3
TU-nirBDC-cysG	4	4
TU-nrdHIEF	3	4
TU-sdhCDAB-b0725-sucABCD	2	9
TU-thiCEFGH	4	5
TU-uhpABC	2	3
TU-ydjA-selD-topB	2	3
TU-yjeFE-amiB-mutL-miaA-hfq-hflXKC	3	9

Table 6.6: Regulons with Mixed Expression upon Iron Supplementation
Genes that showed differential expression in either the zero-time-point, the AIR culture, or the O2 culture in response to iron supplementation were grouped according to information from the EcoCyc database. The 30 regulons shown here exhibited mixed expression, with some genes showing increased expression and others showing decreased expression. γ_{MIXED} is the number of genes that showed differential expression with supplemental iron. γ_{TOTAL} is the total number of genes in the regulon. The remaining genes did not show significant changes.

EcoCyc Gene Group	γ_{MIXED}	γ_{TOTAL}
RG-AppY transcriptional activator	3	9
RG-ArcA-Phosphorylated transcriptional dual regulator	17	79
RG-CaiF transcriptional activator	2	10
RG-CPLX0-222	4	28
RG-CRP-cAMP transcriptional dual regulator	43	241
RG-FhlA-Formate transcriptional activator	4	16
RG-Fis transcriptional dual regulator	15	59
RG-FlhD transcriptional dual regulator	10	28
RG-FNR transcriptional dual regulator	32	119
RG-FruR transcriptional dual regulator	7	26
RG-Hns transcriptional dual regulator	11	40
RG-IHF transcriptional dual regulator	39	159
RG-LexA transcriptional repressor	3	14
RG-Lrp transcriptional dual regulator	15	53
RG-Lrp-Leucine transcriptional activator	7	14
RG-MarA transcriptional activator	6	15
RG-ModE-Molybdate transcriptional dual regulator	7	36
RG-NarL transcriptional dual regulator	3	7
RG-NarL-Phosphorylated transcriptional dual regulator	21	79
RG-NarP-Phosphorylated transcriptional regulator	6	21
RG-NtrC-Phosphorylated transcriptional dual regulator	5	43
RG-PhoB-Phosphorylated transcriptional dual regulator	6	29
RG-PurR-Hypoxanthine transcriptional repressor	4	28
RG-RcsB transcriptional activator	2	8
RG-RNAP32-CPLX	4	24
RG-RNAP54-CPLX	17	85
RG-RNAPS-CPLX	10	69
RG-Rob transcriptional activator	4	8
RG-SoxS transcriptional activator	6	18
RG-TrpR-Tryptophan transcriptional repressor	2	12

6.3.1.3 An Iron-Dependent Cluster

Cluster analysis was performed on the combined Fe-NoFe data set with all N2 samples omitted. Eliminating these highly variable samples allowed examination of the effects of iron specifically on aerobic cultures. One particularly interesting cluster, with correlation of 0.66 is shown in Figure 6.16 In general, the genes in this cluster showed decreased expression in Fe cultures, when compared with NoFe cultures. These genes are also characterized by a sharp drop in expression from the Fe-60-O2 sample to the Fe-90-O2 sample. The 87 genes in this cluster are listed in Table 6.7.



Key

High Expression (higher than expression at zero for unsupplemented culture)

Moderate Expression (equal to expression at zero for unsupplemented culture)

Low Expression (lower than expression at zero for unsupplemented culture)

No Data

Figure 6.16: Cluster of Iron-Dependent Genes with and without Iron Supplementation

The Fe and NoFe data sets were combined and normalized together. The N2 samples were eliminated and hierarchical clustering was performed as described in Section 4.5.4.2. The correlation coefficient for this cluster is 0.66. Each column represents one sample and each row represents a gene. The color of each square represents the magnitude of the expression value as indicated by the key. Notice that all expression data are shown relative to that in the zero sample for the unsupplemented culture. The top row of the cluster shows the average expression profile of genes in the cluster. Genes in this cluster have lower expression in Fe cultures, particularly for the 90-O2 time point.

Table 6.7: Iron-Dependent Genes from Iron-Dependent Cluster
This cluster contained 87 genes with a correlation coefficient of 0.66.

Gene Name	Gene ID	Gene Name	Gene ID	Gene Name	Gene ID
yabK	b0067	*phoH*	b1020	*nrdE*	b2675
leuC	b0072	*tonB*	b1252	*nrdF*	b2676
pdhR	b0113	*ycjG*	b1325	*yggX*	b2962
fhuA	b0150	*ydaU*	b1359	*yggZ*	b2963
fhuC	b0151	*ydbK*	b1378	*nupG*	b2964
yaiH	b0376	*yncE*	b1452	*exbD*	b3005
mdlA	b0448	*tus*	b1610	*yqhD*	b3011
acrA	b0463	*sufE*	b1679	*rpoN*	b3202
entD	b0583	*csdB*	b1680	*sspB*	b3228
fepA	b0584	*sufD*	b1681	*yhfK*	b3358
entF	b0586	*sufC*	b1682	*yiaD*	b3552
ybdA	b0591	*sufB*	b1683	*ilvN*	b3670
entC	b0593	*ydiH*	b1685	*ilvB*	b3671
entE	b0594	*ppsA*	b1702	*rhoL*	b3782
entA	b0596	*astC*	b1748	*sodA*	b3908
citC	b0618	*zwf*	b1852	*fpr*	b3924
holA	b0640	*flhA*	b1879	*glpX*	b3925
fur	b0683	*yedY*	b1971	*yijI*	b3948
bioA	b0774	*wcaF*	b2054	*thiH*	b3990
bioB	b0775	*cirA*	b2155	*thiG*	b3991
bioF	b0776	*yojH*	b2210	*thiF*	b3992
bioC	b0777	*yojI*	b2211	*thiE*	b3993
bioD	b0778	*fdx*	b2525	*thiC*	b3994
ybhS	b0793	*hscA*	b2526	*lysC*	b4024
clpA	b0882	*iscS*	b2530	*qor*	b4051
ompA	b0957	*iscR*	b2531	*soxS*	b4062
b1017	b1017	*ygaM*	b2672	*msrA*	b4219
ycdO	b1018	*nrdH*	b2673	*yjjS*	b4367
ycdB	b1019	*nrdI*	b2674	*yjjU*	b4377

The cluster presented in Figure 6.16 contained many of the oxygen-sensitive genes that were identified in Chapter 5, including genes involved in branched-chain amino acid (BCAA) biosynthesis, biotin biosynthesis, and thiamin biosynthesis, as well as genes from the SoxRS and Fur regulons. Interestingly, genes from both the Isc and Suf Fe-S repair systems appeared in this cluster as well. The common expression pattern shared by all of these genes further supports

the claim that iron supplementation countered the effects of superoxide, particularly at 90 min after induction.

Based on the two ANOVA analyses as well as the iron-dependent cluster described here, the following interpretation considers the general effects of iron supplementation on induced cultures, with focus on the specific effects in the aerobic cultures.

6.3.2 Iron Metabolism

In the presence of iron, Fur should be active and would be expected to repress genes involved in iron transport. It was no surprise to find that many genes in the Fur regulon showed decreased transcription. Comparisons of the zero-time-point samples Fe-0 and NoFe-0 identified three repressed iron-uptake genes: *entA*, *fhuC*, and *fhuD* (Table 6.2). Fe-vs.-NoFe comparisons in all samples from the aerobic (AIR and O2) cultures, identified eight more Fur genes that were repressed in at least one comparison: *fecA/B*, *nrdE/H/I*, *sodA*, and *tonB* (Table 6.5). Down-regulation of the superoxide stress gene, *sodA*, in the O2 culture also indicated that iron alleviated superoxide stress. If the superoxide response were active, it would have offset the Fur repression. The iron-dependent cluster identified another set of repressed genes from the Fur regulon: *cirA*, *entA/C/D/E/F*, *exbD*, *fepA*, *fhuA/C*, *nrdHIEF*, *sodA*, and *tonB* (Table 6.7). While the repression of all of these genes indicated activation of Fur, it was interesting to see that *fur* itself also appeared in this cluster. Finally, the two-treatment analysis identified eight Fur genes that were repressed overall by iron supplementation: *entA/C*, *fecD*, *fepA/E*, *nrdE/H*, and *tonB*. (Table 6.1). In samples from the Fe experiment, three enterobactin biosynthesis genes (*entB/E/F*) frequently displayed low-signal spots that were removed by the data filter. This observation was not made in the NoFe experiment. Partially, this difference was due to experimental effects such as high background and low signal-to-noise. But, this may also have been an indication that synthesis of these transcripts was turned off upon iron supplementation. Aside from Fur genes, another six genes involved in biosynthesis of proto- and siroheme were identified as being repressed, overall, by iron supplementation: *ccmE*, *cysG* and *hemB/C/E/H*. Each of the analyses described in Section 6.3.1 lends support to the claim that supplemental iron repressed the cellular systems for iron uptake.

Interestingly, the iron(II) transport gene *feoB* showed increased expression upon iron supplementation (Table 6.4). While *feoB* expression was consistently increased in the iron

cultures, the effect was strongest at later times (60 and 90 min) in all three aeration conditions (Figure 6.17). In the O2 culture, its neighboring gene, *feoA*, also showed increased expression at 10 min. These genes most likely showed increased expression due to a lack of iron(II). Although iron was supplemented in the form iron(II), it appears to have been oxidized to iron(III). There are two possible explanations for this. As described in Section 6.2, hematite formed upon autoclaving is the most likely source of iron for supplemented cultures. This hypothesis is consistent with the expression results presented here, because the iron in hematite is in the form of iron(III). Alternatively, iron(III) may be generated by increased flux through the Fenton reaction. Regardless of the explanation, there is a need to balance the two forms of iron in order to regulate the redox state of the cell.

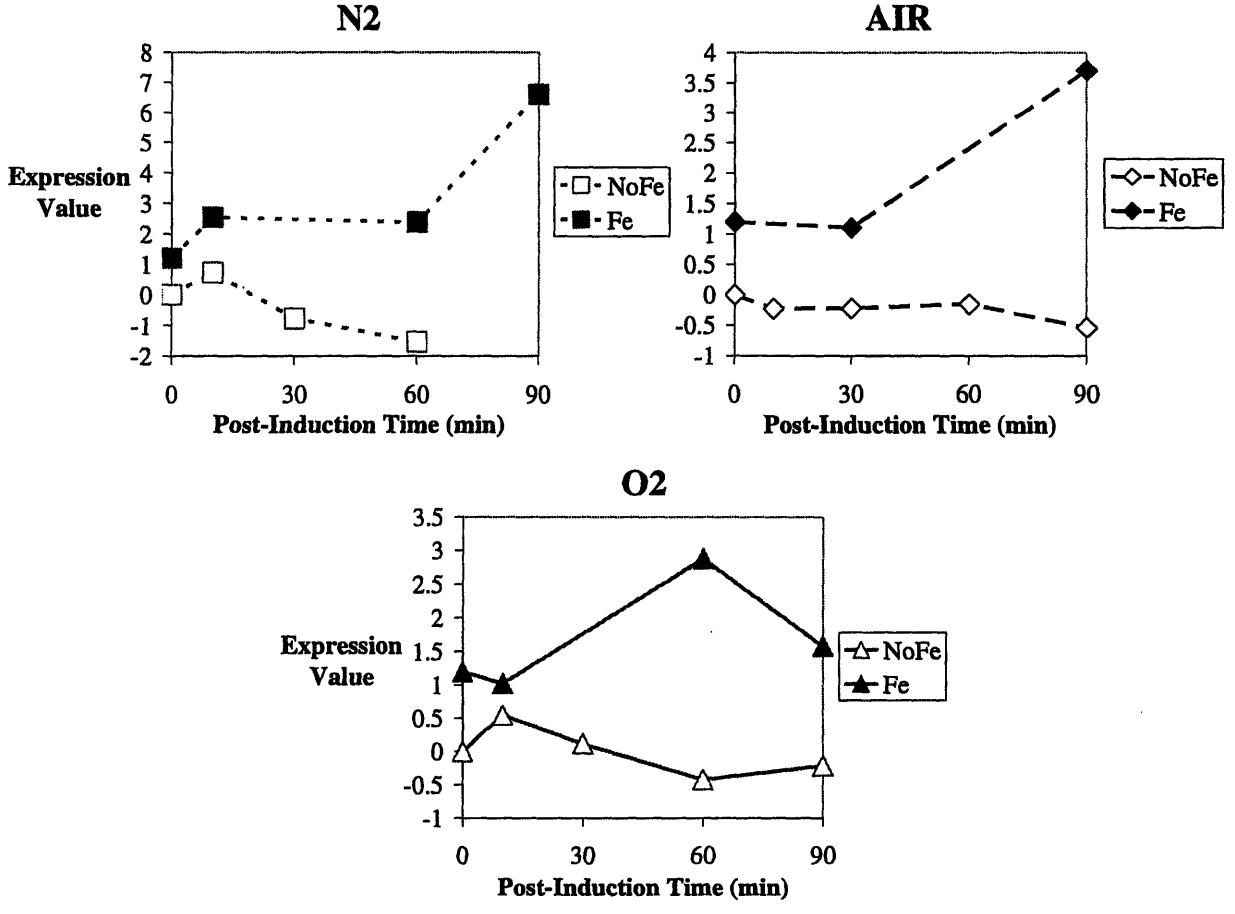


Figure 6.17: Expression Profile of *feoB* with and without Iron Supplementation

Expression values (on a base-2 log scale) for this iron(II) transport gene in N2, AIR, and O2 cultures with and without $FeCl_2$ supp (lementation. Cultures were simultaneously induced and exposed to different aeration environments at 0 min.

6.3.3 Hyperoxic Stress Responses

To better understand the effects of level and type of hyperoxic stress experienced, the hyperoxic stress genes examined in Figure Figure 5.7 were reexamined in the presence of iron. The average expression values of these genes are plotted in Figure 6.18 and Figure 6.19 both with and without iron supplementation.

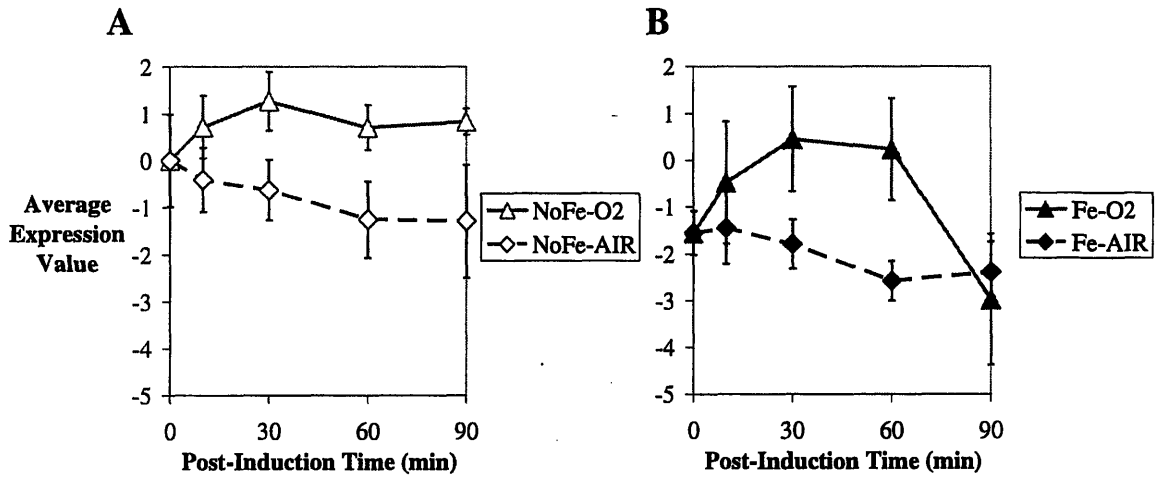


Figure 6.18: Superoxide Stress Response (SoxRS-Regulated) Genes with and without Iron Supplementation
Expression values (on a base-2 log scale) for the genes *acrA*, *fur*, *nfo*, *sodA*, and *soxS* from AIR and O2 cultures were scaled such that **(1)** the sum of the variance across the 18 time points was minimized and **(2)** the average expression value of the NoFe-0 time point was zero. The averaged scaled values are plotted for **A)** cultures without iron supplementation and **B)** cultures with iron supplementation. Error bars represent the standard deviation of these scaled values.

The superoxide response showed a similar pattern both with and without iron supplementation, *i.e.* a roughly 3-fold induction in O2 cultures compared with AIR cultures. Overall, the response was noticeably lower in the Fe experiment. The two-treatment analysis identified the SoxRS genes *acrA*, *nfo*, and *zwf* as being repressed overall by iron supplementation. In addition to *soxS* itself, several SoxRS genes appeared in the iron-dependent cluster in Figure 6.16: *acrA*, *fpr*, *fur*, *sodA*, and *zwf*. As a side note, the SoxRS gene *inaA* showed an overall increase in expression and was not included in Figure 6.18. Overall, the SoxRS response showed an approximate 2-fold repression with iron supplementation, indicating that iron reduced the damaging effects of superoxide.

The superoxide response in the iron-supplementation experiment dropped even more in the 90-O2 time point. At 90 min, expression values from superoxide stress genes were

indistinguishable between the Fe-90-O2 and Fe-90-AIR time points. Further confirmation of this effect came from the genes *acrA*, *fpr*, *nfo*, and *sodA*, all of which showed significantly decreased expression at 90 min in response to iron supplementation in the O2 culture (Table 6.6). Iron supplementation reduced the superoxide stress in aerobic cultures and allowed the O2 culture to further recover from superoxide stress between 60 and 90 min.

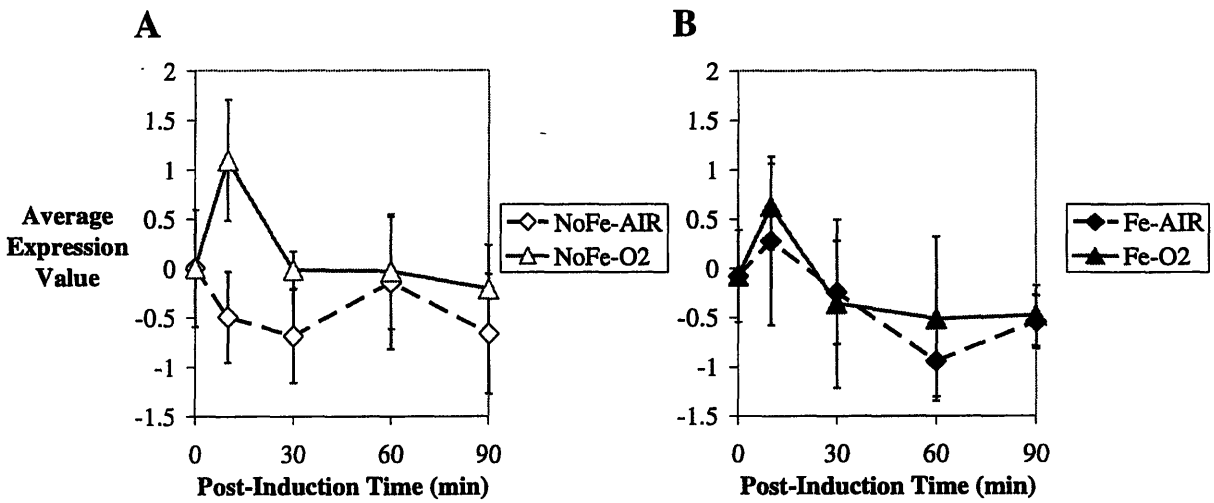


Figure 6.19: Peroxide Stress Response (OxyR-Regulated) Genes with and without Iron Supplementation

Expression values (on a base-2 log scale) for the genes *ahpC*, *dps*, *grxA*, *katG*, and *yfiG* from AIR and O2 cultures were scaled such that (1) the sum of the variance across the 18 time points was minimized and (2) the average expression value of the NoFe-0 time point was zero. Note that the average value for the Fe-0 time point was not set to zero. The averaged scaled values are plotted for **A)** cultures without iron supplementation and **B)** cultures with iron supplementation. Error bars represent the standard deviation of the scaled values.

The peroxide response is also affected by iron supplementation. Genes involved in this response showed similar expression in the O2 culture, with and without iron supplementation. However, with iron supplementation, these genes showed increased expression in the AIR culture at 10 min. At that time point, in the AIR culture, expression of the gene *ahpC* was 5.5-fold higher in response to iron supplementation. Although iron supplementation appears to have had little effect on the O2 culture, it apparently exacerbated the peroxide stress experienced in the AIR culture, such that peroxide stress in the AIR and O2 cultures was indistinguishable.

Supplementation of iron alleviated superoxide stress at the expense of increased peroxide stress. These data are consistent with the model from Figure 1.5 in which iron-sulfur clusters act as a superoxide sink and generate hydrogen peroxide as a byproduct. Supplemental iron is likely

used to reduce these iron-sulfur clusters via the enzymes of the Isc and Suf systems. With continuously regenerated clusters, the cultures clear superoxide and form hydrogen peroxide at a faster rate.

6.3.4 Iron-Sulfur Clusters

Several genes that code for iron-sulfur proteins showed interesting changes in response to iron supplementation. As expected, these changes are most apparent in the O2 cultures. From the list of genes in Table 6.8, 12 genes encoding iron-sulfur proteins were identified. An additional 11 iron-sulfur genes were identified as having significant iron-dependent expression in either the AIR or O2 cultures. This list is presented in Table 6.8.

Of the 23 genes in Table 6.8, 13 were down-regulated in iron-supplemented cultures. In the previous chapter, several iron-sulfur genes were found to show significantly increased expression in O2 cultures. It was hypothesized that this increased expression countered the oxidation of iron-sulfur clusters in order to minimize the lost activity of these proteins. In the presence of iron, however, these iron-sulfur clusters can be regenerated, and increased transcription is not necessary. Particularly in aerobic cultures, iron supplementation may counter the effects of superoxide, assisting in the regeneration of these iron-sulfur clusters.

One gene that encodes an iron-sulfur protein, *bioB*, showed decreased expression overall and particularly at 90 min in the O2 culture (Table 6.8). This expression pattern is similar to that observed for SoxRS regulated genes. Additional genes in the biotin biosynthesis pathway also showed iron-dependent expression. In the two-treatment ANOVA, the gene *bioA* was also identified as having decreased expression. Just like *bioB*, the genes *bioA/D/F* all showed significantly decreased expression at 90 min in the O2 culture.

Also on the list of iron-dependent iron-sulfur genes are *acnB* and *fumB*. While neither of these genes showed significant oxygen dependence, particularly between AIR and O2 cultures, their overall iron dependence was found to be significant.

Conspicuously absent from the list in Table 6.8 are the genes *ilvD* and *leuC*, one or both of which were implicated in the oxygen-dependent expression of the branched-chain amino acid biosynthesis pathway. While these genes may not show iron-dependence, other genes in the pathway do. Four genes from this pathway were identified as having decreased expression in at

least one time point: *ilvE/M/N* and *leuB* (Table 6.5). An overlapping group of four genes was identified from the two-treatment ANOVA: *ilvE/G_2/N* and *leuB* (Table 6.1).

Table 6.8: Expression Differences of Genes Encoding Iron-Sulfur Proteins with and without Iron Supplementation

23 iron-sulfur genes were identified as having iron-dependent expression. Base-2 log ratios from Fe vs. NoFe comparisons are shown for the 26-treatment ANOVA at all time points in the AIR and O2 cultures. Overall log ratios were calculated from the two-treatment ANOVA. Significant log ratios are in bold.

	Gene Name	Gene ID	Log Ratio (log_2 (Fe/NoFe))								
			AIR				O2				
			10	30	60	90	10	30	60	90	Overall
Increasing Expression with Iron Supplementation	*prpD*	b0334	0.2	-0.5	0.2	-0.6	-0.5	-0.4		**2.8**	0.1
	dmsA	b0894	0.7	2.2	1.6	0.9	0.3	1.0	**2.2**	1.0	**1.1**
	hyaA	b0972	0.2	0.5	0.6	1.6	0.5	0.1	1.7		**1.1**
	fnr	b1334	2.1	1.5	0.7	1.2	2.0	**3.5**	**3.1**	0.4	**2.0**
	yeaW	b1802	1.2	1.3	1.4	2.2	1.2	1.5	0.9	1.6	**1.4**
	napG	b2205	**-2.3**		0.5		-2.2	**3.0**	1.8		0.3
	napF	b2208	-1.5	0.9	0.7	0.9	0.6	**2.3**	1.1	1.3	0.8
	ygfS	b2886	0.6	0.2	0.9	1.0	1.7	1.2	1.6	0.8	**1.2**
	yheA	b3337	0.3	0.0	**2.8**	0.6	2.4	-1.0	-0.6	-1.6	0.4
	nrfC	b4072			**3.4**		0.7	1.1	1.1		1.1
Decreasing Expression with Iron Supplementation	*fixX*	b0044	-1.6	-1.4	-0.1	-1.5	-1.2	-0.4	-0.2	-1.0	**-1.0**
	acnB	b0118	-0.8	-0.8	-1.7	-1.8	-0.8	-0.8	-0.9	-1.8	**-0.9**
	ykgJ	b0288	-0.4	0.4	0.9	0.9	1.1	**-2.4**	0.2	-0.3	-0.1
	bioB	b0775	-1.8	-2.4	-1.2		-1.8	-2.5	-1.5	**-5.4**	**-1.9**
	pflE	b0824	-2.1	**-4.1**	-1.3	**-2.9**	**-2.5**	**-4.5**	-1.8	-2.0	**-2.4**
	hcr	b0872			**-3.6**						-0.6
	nuoG	b2283	-0.6	-1.8	-2.0	-1.3	-0.7	-1.5	**-2.3**	0.0	-1.0
	hyfH	b2488	-1.3	-1.1	-1.0	-1.8	-0.9	**-2.5**	-1.1	-1.4	**-1.0**
	hyfI	b2489	**-2.7**		0.8		**-2.6**			-1.0	-1.3
	hydN	b2713	-2.0	-1.7	-1.4	**-2.4**	-1.4	-1.6	-1.1	-1.2	**-1.3**
	hycG	b2719	-1.1	-2.4	-1.6		-0.9	1.3	**-2.7**		-0.4
	fumB	b4122	0.5	0.5	-0.6	0.3	**-2.5**		-0.1	1.2	-0.1
	nrdG	b4237	-0.4	-1.7	**-3.1**		**-3.0**			-0.2	**-2.3**

Although the presence of iron-sulfur clusters in thiamin biosynthesis enzymes appears unlikely based on experimental evidence, several genes in this pathway showed oxygen dependence in Chapter 5, and some also showed iron dependence. The genes *thiC/E/G* were identified as having decreased expression in at least one time point (Table 6.5). In addition, *thiG* and *dxs* showed overall decreased expression with iron supplementation (Table 6.1).

Genes involved in the repair of iron-sulfur clusters did not show significant iron-dependence overall. Of all of the genes in the Isc and Suf systems, only *sufE* showed significantly decreased expression in iron. However, nine of these genes appeared in the iron-dependent cluster: *fdx*, *hscA*, *iscR/S*, *sufB/C/D/E/S*. Like most genes in this cluster, the expression of these genes is characterized by a sharp drop at 90 min in the O2 culture. The expression of genes in these two systems is more closely linked to the superoxide-stress response than iron uptake.

Iron-sulfur genes showing increased expression in Fe cultures would seem to be inconsistent with the hypothesis of supplemental iron regenerating iron-sulfur clusters. Of the ten iron-sulfur genes showing increased expression with iron supplementation, the best characterized are *dmsA*, *napG*, *napF*, *nrfC*, and *fnr*. Transcription of the first four genes is known to be activated by FNR, which is encoded by the last gene. Some or all of the five remaining iron-sulfur genes on this list may have similar regulation. The increased activation of FNR is discussed in the next section and is consistent with regeneration of iron-sulfur clusters by supplemental iron.

Several genes encoding iron-sulfur proteins showed decreased expression with iron supplementation, even in the O2 culture. This observation is consistent with supplemental iron being used to regenerate oxidized iron-sulfur clusters. With functional regeneration systems, it was not necessary for these genes to be overexpressed, as they were in the NoFe cultures. While the systems for iron-sulfur-cluster regeneration did not show significant iron dependence overall, their expression dropped at 90 min in the Fe-O2 culture, indicating that the demand for regeneration of iron-sulfur clusters had declined due to iron supplementation.

6.3.5 FNR Activation

Surprisingly, many of the genes showing differential expression between Fe and NoFe cultures were genes involved in anaerobic respiration—even in the O2 cultures. Genes encoding both of the anaerobic regulatory proteins, *fnr* and *arcA* appeared in Table 6.4 with significantly increased expression in the iron-supplemented cultures. Both of these genes, along with *arcB* showed significantly increased expression in the Fe-O2 culture, compared with the NoFe-O2 culture. Additionally, the genes *dmsA/B*, *focA*, *glpA/B*, *napF*, and *nrfB/C*, which are anaerobic respiratory genes known to be activated by FNR, showed increased expression in Table 6.4.

Furthermore, *narL*, *nuoB/F/G*, *sdhA/C/D*, and *sucB*, which are known to be repressed by FNR, showed decreased expression in Table 6.5.

These results are consistent with FNR activation by iron supplementation, which seems reasonable considering that FNR is activated when its iron-sulfur cluster is in the reduced form. These results indicate that regeneration of iron-sulfur clusters by supplemental iron was so effective that anaerobic respiration was activated even under highly aerobic conditions.

There were, however, several additional genes that showed expression opposite of that expected by FNR activation, *e.g. caiF*, *cydC*, *cyoA/B/C/D*, *hypE*, *nikE*, and *nirB/D*. Regulation of these and other anaerobic respiratory genes is complicated by multiple transcription factors. While not all of the observed changes were consistent with a shift to anaerobic respiration, the culture's response to iron appears to overlap with the FNR regulon.

6.3.6 Heat-Shock Response

Several heat-shock genes were identified in the data analysis in Section 6.3.1. The gene *dnaJ* was found to have significantly lower expression in the iron culture. Although none of the individual aerobic time points showed significant differential expression, the overall trend was clear (Figure 6.20). This consistent expression difference indicates a link between iron supplementation and protein folding. Results from previous sections have provided convincing evidence that iron supplementation assisted in the regeneration of oxidized iron-sulfur clusters. The oxidation state of these clusters can certainly affect the structure of the proteins that bear them. When the clusters are reduced, the proteins are in a stable, active conformation. It is certainly conceivable that oxidation of these clusters would force the proteins into a less stable conformation that would be more susceptible to chaperone binding and eventual proteolysis. As a result, the heat shock response would be activated.

Iron supplementation, which counters the damaging effects of superoxide, also appears to improve the folding of iron-sulfur proteins and inhibit the activation of the heat-shock response. The lack of differential *dnaJ* expression between the AIR and O2 cultures, would seem to refute the above argument. However, an alternative explanation would be that the levels of superoxide present in air are enough to saturate the transmission of this protein-folding signal to *dnaJ* expression. In support of this hypothesis, the *dnaJ* expression difference between the N2 and

AIR cultures was significant at 30 min. This system is sensitive to protein folding changes at low oxygen levels.

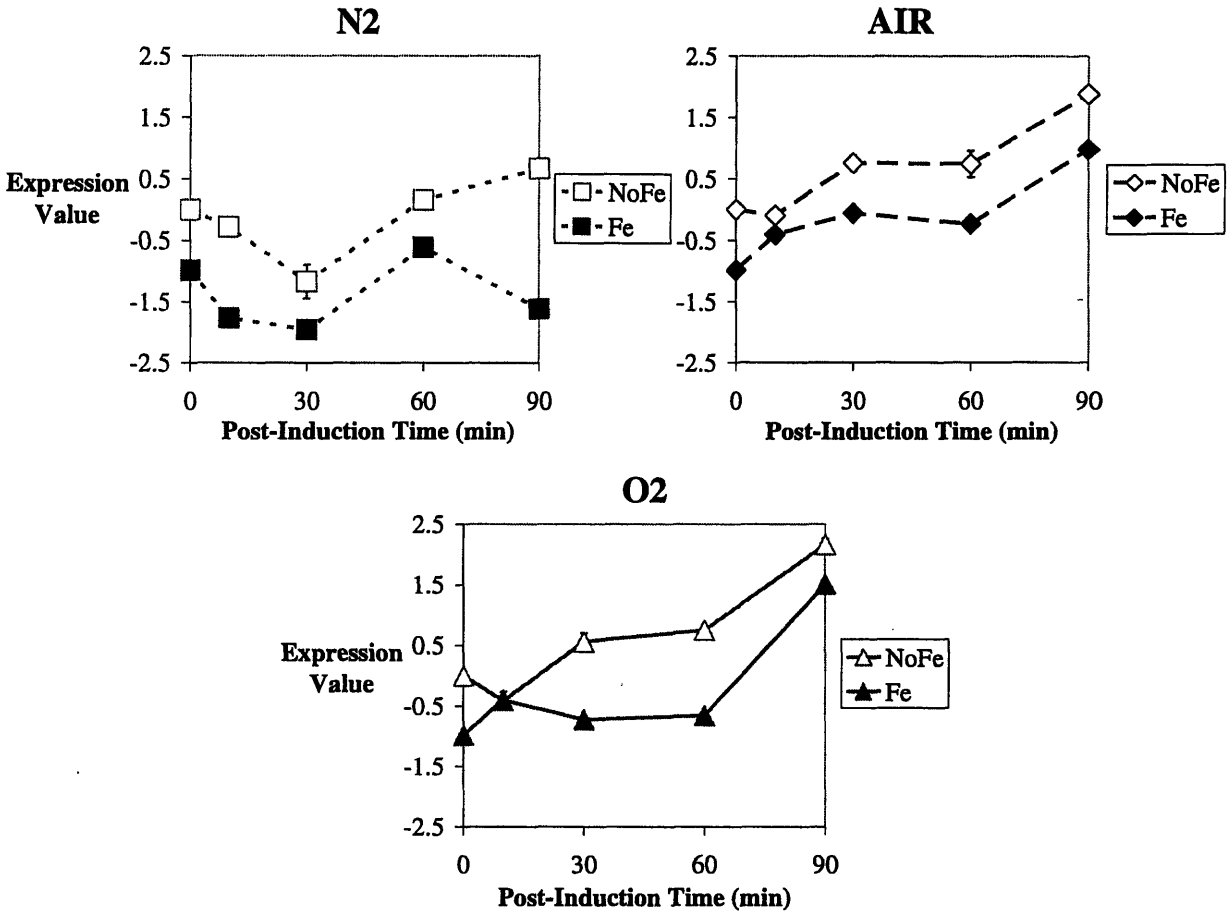


Figure 6.20: Expression Profile of *dnaJ* with and without Iron Supplementation
Expression values (on a base-2 log scale) in N2, AIR, and O2 cultures with and without $FeCl_2$ supplementation. Cultures were simultaneously induced and exposed to different aeration environments at 0 min.

Other heat-shock genes (*htrC* and *topA*) showed a similar expression pattern. However, two other σ^{32}-regulated genes (*rfaL* and *hslT*) showed just the opposite—increased expression upon iron supplementation.

Although not considered a heat-shock gene, *clpA* was found to be a member of the iron-dependent cluster in Figure 6.16. Its oxygen-dependent expression was slight and was never found to be significant from the analyses in Chapter 5 (Figure 6.21). In addition, none of the Fe-vs.-NoFe expression differences were found to be significant. However, its appearance in this

cluster and its response to iron supplementation suggest that *clpA* expression is moderately sensitive to changes in superoxide stress.

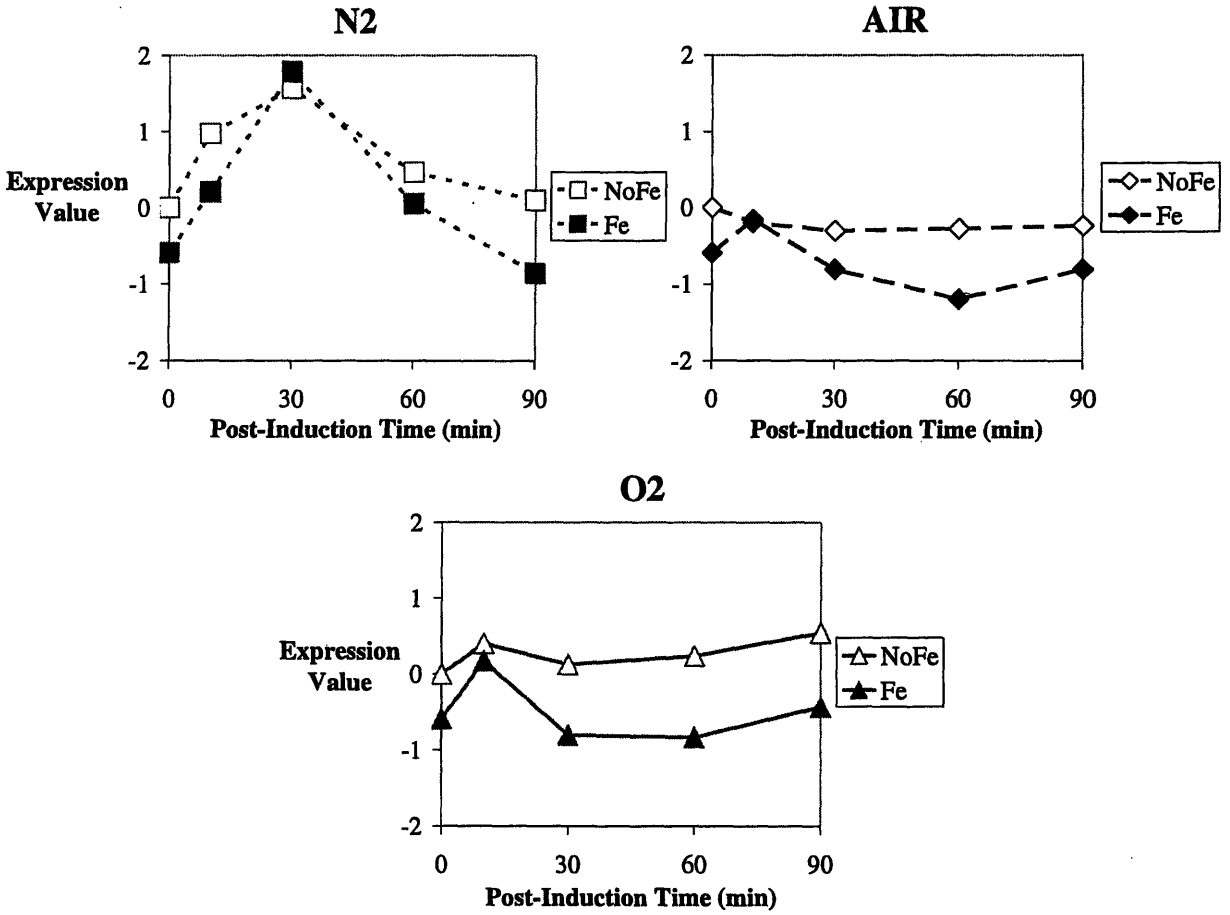


Figure 6.21: Expression Profile of *clpA* with and without Iron Supplementation
Expression values (on a base-2 log scale) in N2, AIR, and O2 cultures with and without $FeCl_2$ supplementation. Cultures were simultaneously induced and exposed to different aeration environments at 0 min.

The implications of this apparent activation of heat-shock proteins extend beyond iron-sulfur proteins. Because the heat-shock protease ClpP is know to degrade α_1AT, further activation of the heat-shock response by oxygen would explain the oxygen dependence of α_1AT proteolysis.

6.3.7 Unknown Genes with Iron-Dependent Expression

A powerful feature of DNA-microarray experiments is that they allow characterization of unknown genes. Some of the unknown genes showing the strongest differential expression in iron-supplemented cultures are mentioned here.

The genes *ydcS/V* showed increased expression upon iron addition. The products of these genes are suspected to function with YdcT and YdcU to form a spermidine/putrecine transporter (Saurin *et al.* 1999).

Several unknown genes in the same transcription unit *yiaK* (b3575) and *yiaM/N/R* all appeared to have iron-dependent expression. Interestingly, *yiaK* showed strongly decreased expression in iron, while the other three showed strongly increased expression. This expression pattern might indicate a separate regulatory region for *yiaM/N/R* for which YiaK acts as a repressor.

For the gene *yhcD* very few spots had significant signal in the NoFe experiment. In contrast, spot signals for this gene in the Fe experiment were strong. While this is largely a qualitative observation, this gene's log ratio was calculated to be 6.3, based on the few comparisons that could be made. *yhcD* belongs to the family of fimbrial transport proteins. Two adjacent genes encoding hypothetical proteins *yhcC/E* also showed similar expression. Another noteworthy unknown gene, *yraJ*, which showed increased expression with iron supplementation, also showed similarity to fimbrial transport proteins.

Additional unknown genes included *yagU*, *yeeT*, *yddK*, *ydjH*, *ygjK*, *yhcC/E*, *yhgA*, *yhhQ*, *ynfA*, *yqiG/H*, and *yrfA/B*, all of which showed increased expression with iron supplementation.

6.4 Summary of Iron Supplementation

This chapter presents work that is of interest in the contexts of both microbial stress responses as well as bioprocess engineering.

DNA-microarray analysis of iron-supplemented cultures revealed that, as expected, genes involved in iron uptake were repressed. Iron supplementation was found to reduce superoxide stress over the long term and increase peroxide stress over the short term. In addition, several genes encoding iron-sulfur proteins and pathways involving those proteins were found to have reduced expression upon iron supplementation. These results are consistent with a model in which the main route of peroxide formation is through oxidation of iron-sulfur clusters by superoxide. Increased iron levels, which result in increased regeneration of iron-sulfur clusters, relieve superoxide stress, while aggravating peroxide stress.

Pulse-chase labeling studies showed that supplementation of iron in the culture medium alleviated the *in vivo* degradation of α_1AT. Furthermore, supplementation with 500 μM $FeCl_2$

was found to dramatically reduce the oxygen dependence of α_1AT degradation. Gene expression analysis of cultures supplemented with iron showed increased expression of three heat-shock genes. It is proposed that oxidation of iron-sulfur clusters forces the proteins that bear them into a less stable conformation. This increased degree of protein misfolding triggers the heat-shock response, which activates proteases, like ClpP, to degrade not only iron-sulfur proteins but others as well. Supplementation of iron allows these iron-sulfur clusters to regenerate, thereby blocking the signal of misfolded proteins and preventing activation of the heat-shock response. This would explain why iron supplementation alleviated the degradation of recombinant α_1AT.

7 Effects of Recombinant Protein Production

"To know for sure, I'd have to throw with a normal hand, and I've never tried it."

*—**Mordecai "Three-Finger" Brown** on whether his curveball was helped by the absence of an index finger*

Without a control experiment, no conclusions can be drawn on the transcriptional effects of recombinant protein production. The validation experiment in Section 4.7 identified genes that showed differential expression in samples taken immediately before and 60-min after induction. However, this experimental design does not distinguish the effects of induction from the effects of time. Therefore, a control experiment performed without recombinant protein production is needed to definitively identify genes affected by induction. This chapter describes two additional control experiments performed for exactly this purpose. This chapter also interprets the results from the initial validation experiment as well as the temporal part of the Aeration-vs.-Time analysis from Chapter 5.

As described in Section 1.2.1.2, the heat shock response becomes activated upon overexpression of a recombinant protein. The set of proteases induced as part of this response can potentially degrade the recombinant protein. In the case of recombinant α1-antitrypsin (α_1AT), degradation results, in part, from the ClpP heat-shock protease (Laska 2000). While this response to recombinant protein production has been studied and well characterized, one of the goals in this work was to characterize the expression pattern of these heat-shock genes under different aeration conditions.

7.1 Induction Control Systems

As described in Section 3.1, the pET expression system was activated by addition of IPTG. The *E. coli* BL21 (DE3) strain used in this work contained a copy of the T7 RNA polymerase gene under the control of the *lacUV5* promoter. This promoter was activated by IPTG to produce T7 RNA polymerase, which, in turn, specifically transcribed the plasmid-borne α_1AT gene in large quantities.

Because of the complexity of this expression system, multiple controls were required. The first control experiment was performed as described in Sections 3.7.5 and 5.1, except without addition of IPTG. The experiment was performed on a culture of *E. coli* with the

pEAT8-137 (α_1AT) plasmid, and cultures were split at OD_{600} of ~0.7. The three smaller cultures were exposed to pure nitrogen, air, and pure oxygen for a period of 90 min. This uninduced control should ideally have no production of either T7 RNA polymerase or recombinant α_1AT; however, the *lac* repressor allowed some "leaky" expression of both of these proteins. The second control experiment was performed in the same manner, but on a culture of *E. coli* containing the empty vector, *i.e.* the pET3d plasmid with no α_1AT insert. For the empty-vector control, IPTG was added concurrently with the split, at an OD_{600} of ~0.7, and each of the three cultures was induced for 90 min. This culture only produced T7 RNA polymerase and allowed the effects of T7-induced recombinant α_1AT production to be distinguished from those of T7 RNA polymerase alone.

Comparison of production from the induced (from Chapter 5), uninduced, and Empty-Vector cultures revealed that α_1AT activity was highest in the Induced culture (Figure 7.1). The empty-vector control showed α_1AT levels near zero, while the Uninduced culture showed α_1AT levels that were roughly 10% of those observed from the Induced culture. As mentioned previously, the non-zero production from the Uninduced culture was attributed to leaky expression of T7 polymerase from the *lacUV5* promoter.

7.2 Analysis of Expression Data

Samples were taken from each of the two induction control cultures and were analyzed using DNA microarrays as described in Section 3.7. Samples from the Uninduced culture were analyzed on slides from Print C, while samples from the Empty-Vector cultures were analyzed on slides from Print E. These data sets were combined with the data set analyzed in Chapter 5. Analysis of the combined data set is described in this section. To gain a more complete understanding of the effects of induction, two additional data sets were also considered.

7.2.1 Induction Validation and Transient ANOVA Data Sets

Two data sets described in previous chapters are considered for analysis in this chapter. The results from the induction validation experiment were only briefly analyzed in Section 4.7 The list of genes identified as being differentially expressed is shown in Table 4.4. The transient ANOVA, in which the effects of aeration and time were simultaneously analyzed, is also

interpreted in this section. A list of the 115 genes identified from that analysis are presented in Table 7.1, Table 7.2, and Table 7.3.

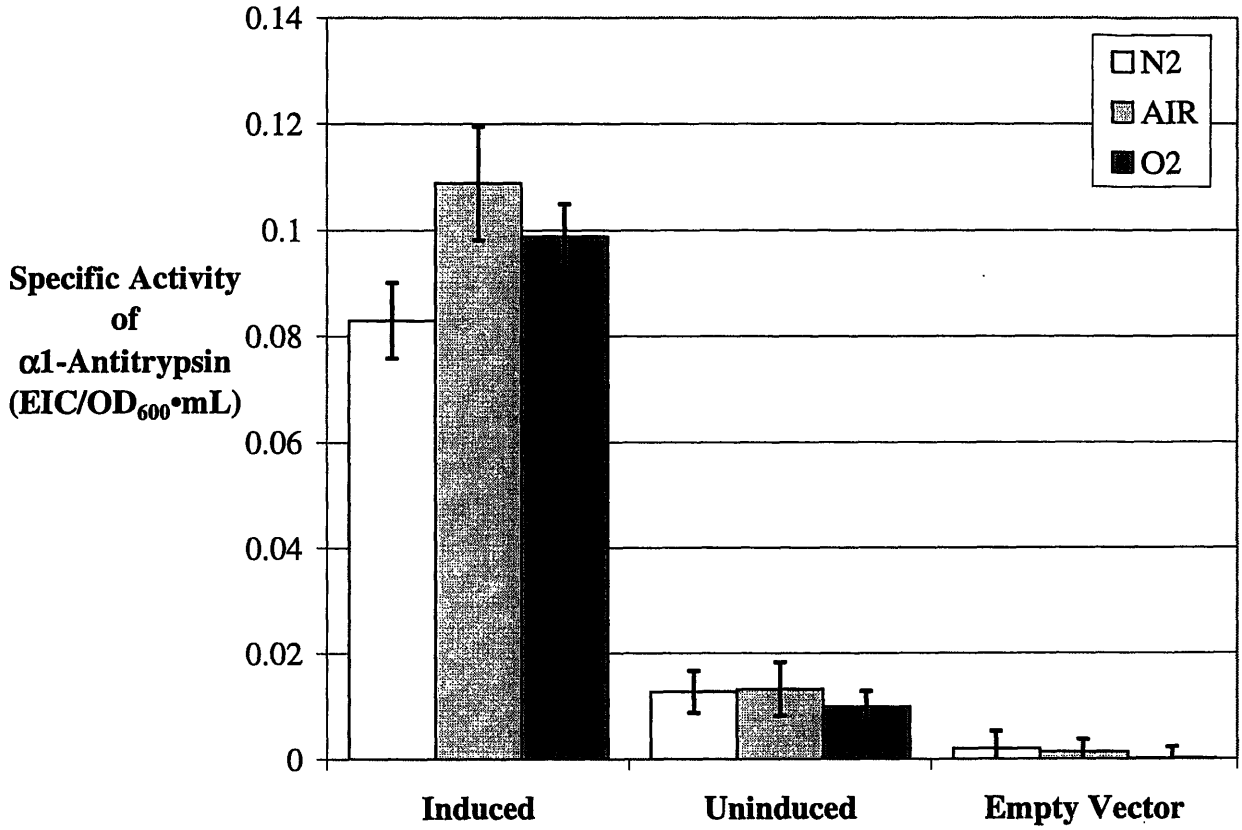


Figure 7.1: Activity of Recombinant α1-Antitrypsin per Cell with Different Levels of Induction
Cultures were grown in minimal medium at 30°C to OD_{600} of 0.7. Two cultures of *E. coli* BL21 (DE3) with the pEAT8-137 α_1AT plasmid and a third culture of *E. coli* BL21 (DE3) pET3d with no α_1AT insert were grown in minimal medium at 30°C to OD_{600} of 0.7. One of the pEAT8-137 cultures was supplemented with IPTG to 0.4 mM (Induced), the other was not supplemented (Uninduced). The pET3d culture was supplemented with IPTG to 0.4 mM (Empty Vector). α_1AT was quantified using the EIC assay as described in Section 3.5.2.

7.2.2 Induced vs. Uninduced vs. Empty-Vector Cultures

Analysis of the effects of induction regarded each of the 39 arrays from Section 7.1 as a separate treatment, and the entire data set as a single block. Only Induced vs. Uninduced, Induced vs. Empty Vector, and Uninduced vs. Empty Vector treatment comparisons were made with the same aeration and at the same time point (39 comparisons). Based on the global significance test at a level of 0.95, 1,865 genes showed significant differential expression in at least one comparison. With only the Induced vs. Uninduced comparisons, 648 genes were

identified as being differentially expressed. With only the Induced vs. Empty Vector comparisons, 1,061 genes were selected.

Table 7.1: Genes Showing Increased Expression during Induction
The transient ANOVA from Section 5.2.1.2 identified 26 genes that showed increased expression during production of a recombinant protein. Descriptions of gene products were taken from the EcoCyc database. At each of the time points (10, 30, 60, and 90 min), the average log ratio in comparison with the zero-time-point sample is shown. These log ratios are independent of aeration.

Gene Name	Gene ID	Gene Product	10	30	60	90
AT	Antitryp	Recombinant human antitrypsin	1.8	5.2	5.3	5.5
trpH	b1266	putative enzyme	0.6	2.6	2.7	3.5
atoB	b2223	putative transferase	0.0	1.2	1.1	3.2
hslT	b3687	heat shock protein	0.5	2.2	2.6	3.0
yhhI	b3484	conserved protein similar to H-repeat-associated proteins	0.5	2.1	2.5	2.9
gip	b0508	hydroxypyruvate isomerase	0.0	0.0	-1.0	2.6
proV	b2677	proline ABC transporter	0.4	1.1	2.2	2.5
lacY	b0343	LacY lactose MFS transporter	4.4	4.6	4.3	2.4
ptxA	b4195	eiisga	0.3	1.4	1.6	2.2
mbhA	b0230	putative motility protein	-0.4	-1.0	-0.7	2.2
lnt	b0657	apolipoprotein N-acyltransferase	-0.3	1.3	2.1	1.9
cspA	b3556	CspA transcriptional activator	-0.1	1.0	2.4	1.8
hybB	b2995	probable cytochrome Ni/Fe component of hydrogenase-2	-0.3	0.4	0.6	1.8
dnaJ	b0015	chaperone with DnaK; heat shock protein	-0.2	0.1	0.6	1.6
intD	b0537	prophage DLP12 integrase	0.7	-1.1	1.3	1.3
yi81_1	b0016	IS186 protein	-0.7	0.2	0.4	1.3
deaD	b3162	inducible ATP-independent RNA helicase	-0.4	0.8	1.9	1.2
kdpE	b0694	KdpE-Phosphorylated transcriptional activator	0.8	-2.0	1.1	1.2
yhjG	b3524	conserved protein	0.2	-1.7	0.7	1.1
fepA	b0584	outer membrane receptor for ferric enterobactin (enterochelin) and colicins B and D	-1.0	1.6	0.3	1.0
b1371	b1371	hypothetical protein	0.1	-2.1	0.5	0.6
ilvG_1	b3767	IlvG_1	-0.1	-1.3	-1.8	0.5
ydjY	b1751	conserved protein, 4Fe-4S protein	0.0	-0.7	1.2	0.0
rplD	b3319	50S ribosomal subunit protein L4, regulates expression of S10 operon	-0.1	-1.8	0.5	-0.5
rplW	b3318	50S ribosomal subunit protein L23	-0.3	-2.2	0.3	-0.8
rbsR	b3753	RbsR-ribose	-1.1	2.3	0.3	

Table 7.2: Genes Showing Mixed Expression during Induction
The transient ANOVA from Section 5.2.1.2 identified two genes that showed mixed expression during production of a recombinant protein. Descriptions of gene products were taken from the EcoCyc database. At each of the time points (10, 30, 60, and 90 min), the average log ratio in comparison with the zero-time-point sample is shown. These log ratios are independent of aeration.

Gene Name	Gene ID	Gene Product	10	30	60	90
ompG	b1319	outer membrane protein	0.3	-2.6	-0.1	1.2
ydcR	b1439	multi modular; putative transcriptional regulator ; also putative ATP-binding component of a transport system	0.2	-2.7	0.3	0.5

7.3 Global Transcriptional Effects of Induction

The analysis of data sets described in the previous section examined induction from several different angles. The induction validation experiment examined gene expression differences in the same culture before and after induction. The Aeration-vs.-Time ANOVA identified genes with transient expression in induced cultures and, therefore, examined the cumulative effects of induction. Finally, the analyses between the Induced cultures and the induction controls (Uninduced and Empty-Vector cultures) examined gene expression under identical conditions, with varying induction. This thorough analysis of the effects of induction revealed interesting gene expression changes, as described here.

7.3.1 Direct Effects of Induction

The induction validation experiment confirmed that genes for T7 RNA polymerase and recombinant α_1AT as well as *lacY* showed significantly increased expression following induction. These genes are expected to show significant increases based on the pET expression system. An understanding of the expression for each of these genes in the Induced, Uninduced, and Empty-Vector cultures lends insight into the other changes occurring in each of these cultures.

Table 7.3: Genes Showing Decreased Expression during Induction
(on next two pages) The transient ANOVA from Section 5.2.1.2 identified 87 genes that showed decreased expression during production of a recombinant protein. Descriptions of gene products were taken from the EcoCyc database. At each of the time points (10, 30, 60, and 90 min), the average log ratio in comparison with the zero-time-point sample is shown. These log ratios are independent of aeration.

Gene Name	Gene ID	Gene Product	10	30	60	90
lamB	b4036	phage lambda receptor protein; maltose high-affinity receptor	-0.4	-2.4	-5.1	-6.5
malM	b4037	periplasmic protein of mal regulon	-0.8	-3.4	-4.8	-5.8
acs	b4069	acetyl-CoA synthetase	-0.7	-0.4	-1.8	-5.2
gatC	b2092	EIIGat	-0.1	-1.4	-3.6	-4.9
malK	b4035	maltose ABC transporter	-0.4	-2.0	-3.9	-4.6
malF	b4033	maltose ABC transporter	-0.4	-2.6	-4.5	-4.4
ilvC	b3774	acetohydroxy acid isomeroreductase	0.1	-1.9	-3.6	-4.3
gatZ	b2095	tagatose-1,6-bisphosphate aldolase 2	-0.4	-1.0	-3.5	-4.2
malG	b4032	maltose ABC transporter	0.0	-1.7	-3.6	-4.2
malE	b4034	maltose ABC transporter	-0.4	-2.0	-3.4	-4.1
fba	b2925	fructose bisphosphate aldolase monomer	-0.7	-2.2	-2.5	-3.9
sapC	b1292	peptide uptake ABC transporter	-0.2	0.2	0.2	-3.9
sucC	b0728	succinyl-CoA synthetase	-0.2	-1.3	-3.3	-3.7
sucA	b0726	subunit of E1(0) component of 2-oxoglutarate dehydrogenase	0.5	-1.3	-2.7	-3.6
gatB	b2093	EIIGat	-0.3	-1.2	-2.8	-3.5
nuoA	b2288	NADH dehydrogenase I	-1.5	-2.1	-2.4	-3.5
nuoB	b2287	NADH dehydrogenase I	-1.3	-2.4	-2.6	-3.4
sdhC	b0721	succinate dehydrogenase membrane protein	-0.4	-1.9	-2.1	-3.0
sucD	b0729	succinyl-CoA synthetase	0.4	-0.6	-2.2	-3.0
nuoF	b2284	NADH dehydrogenase I	-0.4	-2.3	-3.2	-2.8
nuoC	b2286	NADH dehydrogenase I	-0.5	-2.1	-3.0	-2.8
nuoM	b2277	NADH dehydrogenase I	0.0	-1.0	-2.4	-2.7
b0725	b0725	hypothetical protein	0.6	-0.4	-1.8	-2.7
trpC	b1262	indole-3-glycerol phosphate synthase / phosphoribosylanthranilate isomerase	0.0	-2.3	-1.8	-2.7
yibD	b3615	putative glycosyltransferase	-0.1	-2.0	-2.3	-2.6
abgB	b1337	hypothetical protein	-0.2	-2.4	-2.1	-2.6
sdhD	b0722	succinate dehydrogenase membrane protein	-0.3	-1.5	-1.9	-2.6
aceA	b4015	isocitrate lyase monomer	-0.3	-2.1	-2.6	-2.6
nuoE	b2285	NADH dehydrogenase I	-0.3	-2.6	-3.3	-2.6
tyrA	b2600	chorismate mutase / prephenate dehydrogenase	-0.1	-1.6	-2.0	-2.5
manZ	b1819	EIIMan	0.3	-0.5	-0.7	-2.5
aroP	b0112	AroP phenylalanine/tyrosine/tryptophan APC transporter	-0.8	-1.7	-1.8	-2.5
trpB	b1261	tryptophan synthase B protein	0.2	-1.1	-2.5	-2.5
cysD	b2752	sulfate adenylyltransferase	1.1	-2.6	-2.8	-2.5
aroF	b2601	2-dehydro-3-deoxyphosphoheptonate aldolase	-0.4	-1.4	-2.6	-2.5
sucB	b0727	SucB-lipoate	0.4	-0.5	-2.1	-2.5
manY	b1818	EIIMan	-0.1	-0.2	-1.0	-2.5
nuoL	b2278	NADH dehydrogenase I	0.2	-0.6	-1.5	-2.4
manX	b1817	EIIMan	-0.1	0.3	-1.3	-2.4
trpA	b1260	tryptophan synthase A protein	0.6	-0.6	-2.0	-2.4
cysI	b2763	sulfite reductase hemoprotein subunit	0.1	-2.0	-3.1	-2.3
icdA	b1136	isocitrate dehydrogenase	0.7	-0.9	-1.3	-2.2
sodA	b3908	superoxide dismutase (Mn)	-0.2	-0.5	-3.0	-2.2
fumA	b1612	fumarase A monomer	-0.2	-1.2	-1.5	-2.1
sdhA	b0723	succinate dehydrogenase flavoprotein	0.3	-1.5	-2.3	-2.1
nuoH	b2282	NADH dehydrogenase I	0.0	-1.2	-1.7	-2.1

Gene Name	Gene ID	Gene Product	10	30	60	90
cysH	b2762	3'-phospho-adenylylsulfate reductase	0.3	-1.3	-2.8	-2.1
cyoC	b0430	subunit III	0.0	-1.8	-1.4	-2.1
cybC	b4236	cytochrome b562 (soluble)	0.1	-0.8	-1.9	-2.1
nuoK	b2279	NADH dehydrogenase I	0.4	-0.4	-1.8	-2.0
ycjJ	b1296	YcjJ APC transporter	-0.3	-1.6	-0.8	-2.0
tyrB	b4054	aromatic-amino-acid transaminase	-0.1	-1.8	-2.1	-2.0
purD	b4005	phosphoribosylamine-glycine ligase	-0.1	-0.9	-1.5	-1.9
folE	b2153	GTP cyclohydrolase I	-0.1	-1.8	-2.5	-1.9
trpE	b1264	anthranilate synthase component I	0.0	-1.3	-2.6	-1.9
cysJ	b2764	sulfite reductase flavoprotein subunit	-0.1	-2.3	-3.3	-1.8
hisG	b2019	ATP phosphoribosyltransferase monomer	1.0	-1.0	-2.0	-1.8
ydaH	b1336	AbgT Transporter	0.5	-0.8	-0.4	-1.8
yiaJ	b3574	YiaJ transcriptional repressor	-0.4	-0.9	-1.4	-1.7
cysK	b2414	cysteine synthase	0.8	-0.6	-1.2	-1.7
gnd	b2029	6-phosphogluconate dehydrogenase (decarboxylating)	-0.3	-0.7	-2.0	-1.7
abgA	b1338	putative aminohydrolase (EC 3.5.1.14)	0.0	-1.6	-2.1	-1.7
yceI	b1056	periplasmic protein; possibly secreted	0.5	0.3	-1.6	-1.7
pckA	b3403	phosphoenolpyruvate carboxykinase (ATP)	0.1	-0.1	-0.4	-1.6
pepD	b0237	peptidase D, a dipeptidase where amino-terminal residue is histidine	0.3	-0.4	-0.7	-1.5
trpS	b3384	tryptophanyl-tRNA synthetase	0.1	-1.0	-0.3	-1.5
trpD	b1263	anthranilate synthase component II monomer	0.3	-1.1	-1.5	-1.5
yobF	b1824	hypothetical protein	1.1	0.2	-1.2	-1.4
cysC	b2750	adenylylsulfate kinase	-0.3	-2.0	-2.6	-1.4
yjbI	b4038	conserved protein	0.2	-0.9	-1.5	-1.4
yccA	b0970	putative carrier/transport protein; substrate or modulator of FtsH-mediated proteolysis	0.8	0.3	-0.1	-1.3
livJ	b3460	branched chain amino acids ABC transporter	0.1	-1.4	-1.9	-1.3
tauB	b0366	TauA/TauB/TauC ABC transporter	0.5	-0.8	-1.5	-1.3
hisF	b2025	hisF subunit	0.3	-0.2	-2.1	-1.3
cysN	b2751	sulfate adenylyltransferase	1.1	-1.0	-1.6	-1.2
hisD	b2020	histidinal dehydrogenase / histidinol dehydrogenase	1.4	-0.5	-1.7	-1.2
hisC	b2021	histidine-phosphate aminotransferase	1.4	0.1	-1.0	-1.1
adhE	b1241	PFL-deactivase / alcohol dehydrogenase / acetaldehyde dehydrogenase	0.7	0.9	0.3	-1.1
cysA	b2422	sulfate ABC transporter	0.4	-1.4	-2.4	-1.1
yciW	b1287	putative oxidoreductase	0.8	-0.6	-1.1	-1.0
hisA	b2024	phosphoribosylformimino-5-amino-1-phosphoribosyl-4-imidazole carboxamide isomerase	-0.1	-0.5	-2.3	-1.0
msrA	b4219	methionine sulfoxide reductase A / protein-methionine-S-oxide reductase / methionine sulfoxide reductase	0.4	-1.0	-1.6	-0.8
leuB	b0073	3-isopropylmalate dehydrogenase	0.1	-1.4	-2.2	-0.8
ftn	b1905	cytoplasmic ferritin, an iron storage protein)	1.7	0.2	0.3	-0.7
tauA	b0365	TauA/TauB/TauC ABC transporter	0.8	-0.2	-1.2	-0.5
ilvE	b3770	branched chain amino acid aminotransferase	0.4	-0.9	-1.7	-0.4
amn	b1982	AMP nucleosidase	0.6	0.3	-0.7	-0.2

It should be noted that spots corresponding to T7 RNA polymerase and α_1AT genes were often saturated, because the signal was beyond the maximum detection limit of the scanner. This saturation was expected; it is assumed that the cultures devote a great deal of energy to making

both of these transcripts and, consequently, proteins. Although these values were treated as quantitative measurements for the purposes of data analysis, the following interpretation of these values treats them qualitatively.

As expected, expression of the *lacY* gene remained constant in the Uninduced cultures since no IPTG was added. However, the Induced cultures and the Empty-Vector cultures both showed increased expression. In both of these cultures, *lacY* expression dropped again at later times. For the Induced culture, a slight drop occurred between 60 and 90 min, whereas for the Empty-Vector culture, expression returned to the basal level between 30 and 60 min (Expression values at 0, 60, and 90 min are likely too low to be detectable in the empty-vector N2 culture). Profiles for this gene are shown in Figure 7.2. The drop in *lacY* expression suggests that either IPTG was consumed or that levels of the LacI repressor had increased enough to halt transcription of the *lac* operon. In either case, the burden of producing recombinant α_1AT (Induced culture) appears to slow the process.

Addition of IPTG not only stimulated the *lac* operon, but also activated transcription of the T7 RNA polymerase gene (Figure 7.3). As expected, the profiles for T7 were similar to those observed for *lacY*. The Induced and Empty-Vector cultures, to which IPTG was added, showed higher T7 expression than the Uninduced culture. Interestingly, the Empty-Vector culture showed an initial increase in expression, followed by a slight drop between 10 min and 30 min. This drop may result from the same effect that lowered *lacY* expression. At later times, expression of the T7 transcript in the Empty-Vector culture is somewhere between that in the Uninduced and Induced cultures. This trend is also apparent in the O2 and N2 cultures.

The only culture producing recombinant α_1AT is the Induced culture. As expected, expression of the α_1AT transcript in this culture is consistently highest in N2, AIR, and O2 cultures (Figure 7.4). Since neither of the other two cultures were induced to produce α_1AT, they might be expected to show similar levels of α_1AT expression. Practically, however, this was not the case. Although the data from the Empty-Vector culture are quite noisy, levels of the α_1AT transcript are generally higher in the Uninduced culture than in the Empty-Vector culture. Despite IPTG addition, the Uninduced culture contains plasmid-borne copies of the α_1AT gene, and some basal expression occurs. In contrast, the Empty-Vector culture has no copies of the α_1AT gene and the expression should be zero. Ideally, these spots should not have had any signal and their data should have been removed in the filter for low expression.

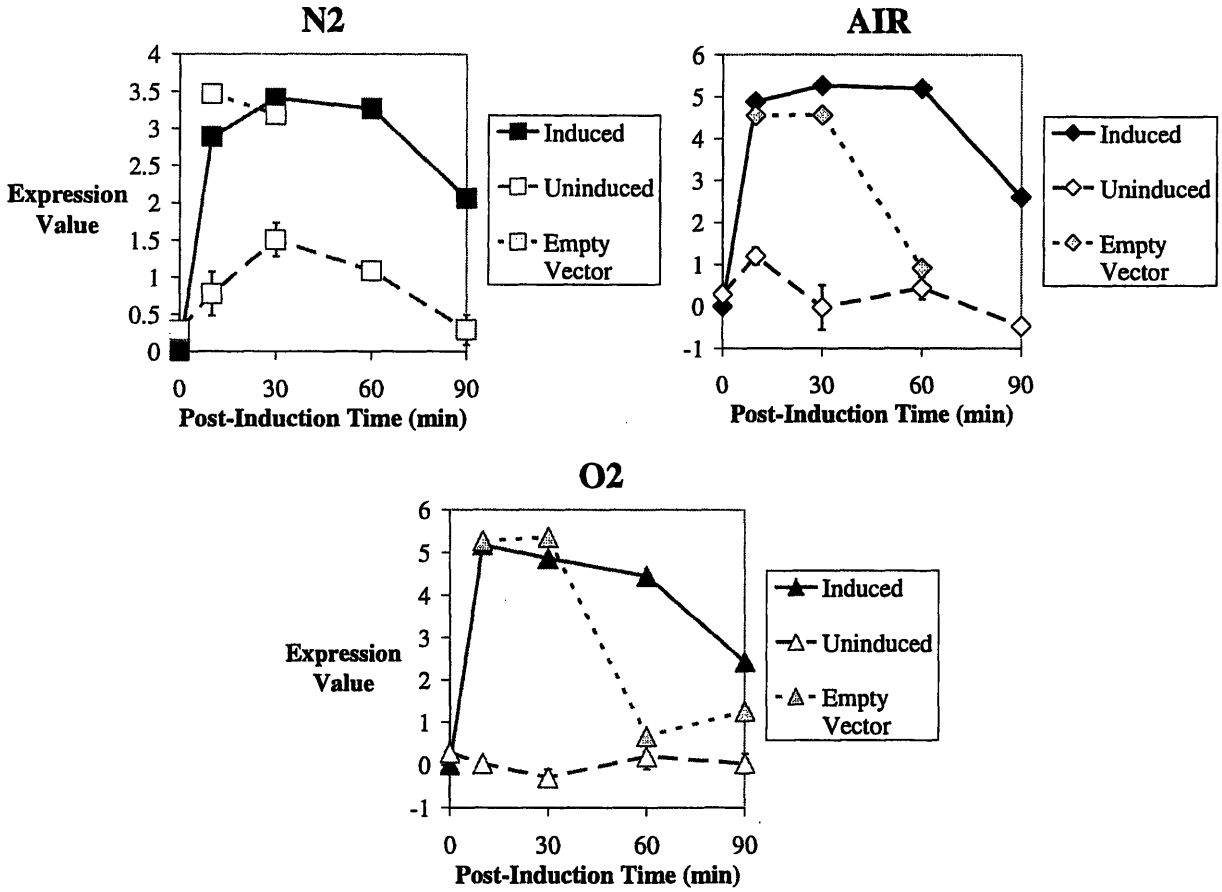


Figure 7.2: Expression Profile of *lacY* upon Induction
Expression values for Induced, Uninduced, and Empty-Vector cultures induced in N_2, air, and O_2. At 0 min, IPTG was added to the Induced and Empty-Vector cultures simultaneously with exposure to different aeration environments.

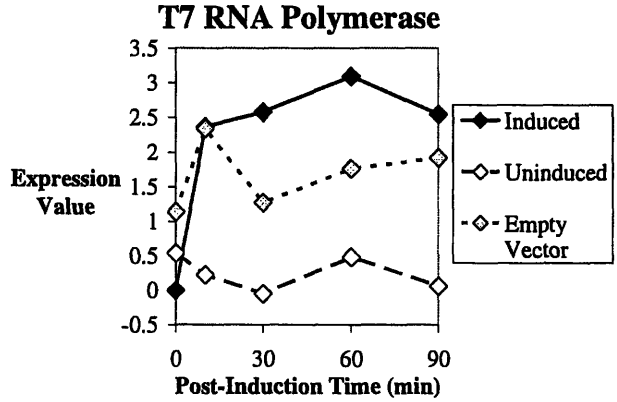


Figure 7.3: Effects of Induction on Transcription of T7 RNA Polymerase Gene in AIR Cultures
Expression values for Induced, Uninduced, and Empty-Vector cultures induced in air. At 0 min, IPTG was added to the Induced and Empty-Vector cultures.

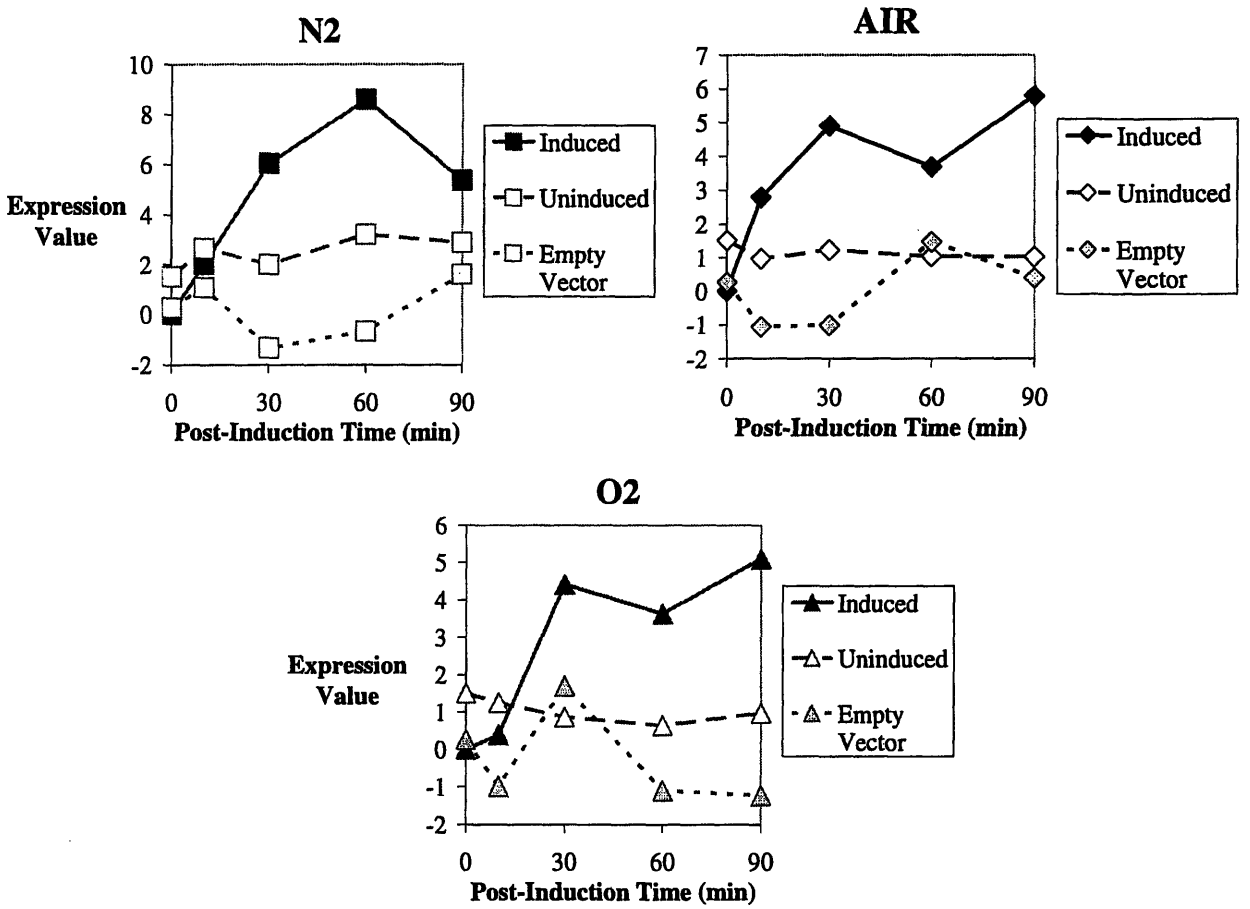


Figure 7.4: Effects of Induction on Transcription of Recombinant α1-Antitrypsin Gene
Expression values for Induced, Uninduced, and Empty-Vector cultures induced in N_2, air, and O_2. At 0 min, IPTG was added to the Induced and Empty-Vector cultures simultaneously with exposure to different aeration environments.

7.3.2 Heat-Shock Response

It is well known that the heat-shock response in *E. coli* becomes activated upon production of a misfolded protein. Between pulse-chase labeling experiments in this work and previous work, there is evidence to indicate that an α_1AT intermediate is misfolded. Additionally, because the heat-shock protease ClpP has been implicated in the degradation of α_1AT, it appears likely that the heat-shock response is active in cultures producing this recombinant protein. This unfolded α_1AT intermediate is likely sufficient to stimulate the heat-shock response in *E. coli.*

Activation of the heat-shock response was first noted in the induction validation experiment. The list of differentially expressed genes (Table 4.4) contained eight genes in the

σ^{32} regulon: *clpB*, *dnaK*, *dnaJ*, *hslS*, *hslT*, *htpG*, *htpX*, and *topA*. In comparisons of samples taken immediately before induction as well as 60 min after induction, these genes exhibited log ratio value ranging from 1.1 to 2.0. The first seven of these genes were also identified in response to temperature increase (Richmond *et al.* 1999) and unfolded protein production (Lesley *et al.* 2002). The gene *topA* was identified only in the former. The gene *hslR* (*yrfH*), which encodes a ribosomal heat-shock protein, was also identified by all three studies. Based on results from the second-round ANOVA for Block-A data in nitrogen, air, and oxygen, the genes *dnaJ* and *hslT* were both selected as showing significant transient effects regardless of aeration. Between 10 min and 90 min, these genes showed increased expression values of 1.9 and 2.4, respectively (Figure 7.5).

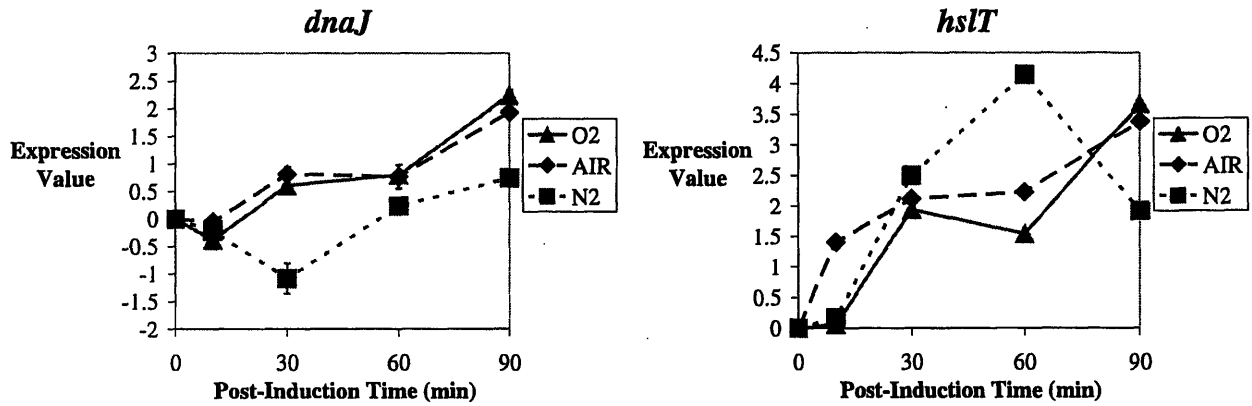


Figure 7.5: Expression of Two Heat-Shock Genes during Induction
Expression values for two genes from the Block-A data set from Chapter 5. Cultures were simultaneously induced and exposed to different aeration environments at 0 min. Both of these genes showed increasing expression, regardless of aeration.

Comparisons between Induced, Uninduced, and Empty-Vector cultures confirmed this activation of the heat-shock response. The difference in expression of these genes between Induced and Uninduced cultures was obvious. Most heat-shock genes showed increasing expression in the Induced culture, but remained unchanged in the Uninduced culture. A subset of these genes also showed decreased expression in the Empty-Vector culture (Figure 7.6). The general pattern of expression for these genes—increasing expression from empty-vector to uninduced to induced—is similar to that observed for the α_1AT gene, which indicates that recombinant α_1AT is the stimulus for the heat-shock response. Even production of α_1AT at basal levels, as in the Uninduced culture, is sufficient to maintain levels of heat-shock transcripts at a constant level. In contrast, without any α_1AT production, as in the Empty-Vector culture,

expression of these heat-shock genes drops. Production of T7 RNA polymerase appears to have less of an effect on expression of these heat-shock genes, perhaps because it folds more effectively than α_1AT.

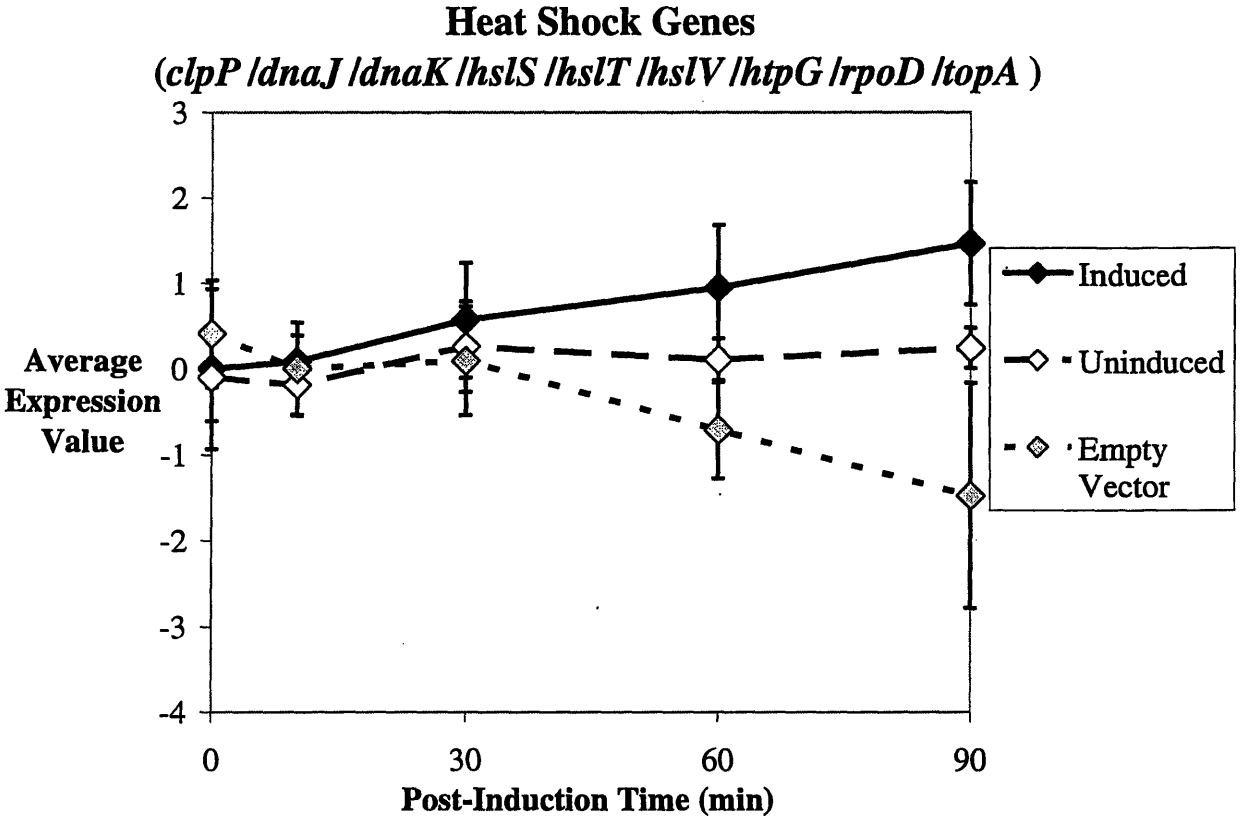


Figure 7.6: Average Expression Profile of Heat Shock Genes
Expression values for Induced, Uninduced, and Empty-Vector cultures induced in air. At 0 min, IPTG was added to the Induced and Empty-Vector cultures.

7.3.3 Catabolite Repression

The signal molecule cyclic AMP (cAMP) is produced inversely with glucose levels. It is produced by the enzyme adenylate cyclase. The activity of this enzyme is activated by the phosphorylated form of the glucose-specific transport protein $EIIA^{Glc}$. This protein is part of the PTS (phospotransferase system), which activates glucose permease. As glucose enters the cell, it becomes phosphorylated and $EIIA^{Glc}$ loses its phosphoryl group. Thus, when extracellular glucose levels are low, the phosphorylated $EIIA^{Glc}$ is the dominant form, and adenylate cyclase is activated to produce cAMP. As glucose levels increase, the $EIIA^{Glc}$ phosphate is continuously transferred to glucose and cAMP production drops. The latter situation is referred to as catabolite repression.

cAMP is an important factor in transcriptional regulation during carbon-source starvation. cAMP stimulates breakdown of the stored carbohydrate reserves in glycogen. cAMP also binds to CRP (cAMP regulatory protein), which activates transcription of several genes involved in central metabolism in an effort to optimally utilize all of the sugars available to the cell. The CRP-cAMP complex also activates the expression of several sugar transport genes to expand its metabolism to a variety of different carbon sources. Without cAMP, CRP cannot regulate transcription and its regulon is repressed.

Throughout the data sets analyzed as described in Section 7.2, several groups of genes were consistently identified as having differential expression. Most, if not all of the genes in the *sdhCDAB*/b0725/*sucABCD* transcription unit were identified in all analyses. In addition, transcription units involved in maltose transport, *malEFG* and *malK*/*lamB*/*malM*, were commonly observed. Finally, genes involved in mannose transport, *manXYZ*, were also identified in these analyses. Each of these transcription units is known to be activated by the cAMP-CRP regulator. The induction validation data set found these genes to be down-regulated after induction, the Aeration-vs.-Time analysis found them to be decreasingly expressed in Induced cultures, and the induction control experiments identified them as having decreased expression in Induced over Uninduced cultures.

Further analysis of each of these data sets strengthened the link between induction and the cAMP-CRP regulator. In the induction validation experiment, 38 genes activated by the cAMP-CRP regulon were found to show decreased expression following induction. The transient analysis identified a total of 18 cAMP-CRP genes with decreased expression, mostly from the above transcription units. Finally, comparison of Induced vs. Uninduced cultures found 39 cAMP-CRP genes with decreased expression in the Induced cultures. Although 241 genes are currently known to be regulated by cAMP-CRP, the relatively small number of changes described here were consistent across the different data sets.

The above observations suggest that induction reduced the activity of this regulator, most likely by reducing cAMP levels. One explanation is that addition of IPTG indirectly inactivated adenylate cyclase. As shown in Section 7.3.1, IPTG increased the transcription of the *lacY* gene, which certainly led to increased levels of LacY, the lactose permease. This lactose transporter is known to be inhibited by $EIIA^{Glc}$, the same component of the PTS system that is responsible for adenylate cyclase activation. The dramatic increase in levels of lactose permease may be

sufficient to occupy $EIIA^{Glc}$ and thereby inhibit cAMP formation. Other microarray analyses on IPTG-induced cultures have also found decreased expression of genes involved in central metabolism and nucleotide biosynthesis (Rohlin *et al.* 2002). Interestingly, these observations were not made from microarray analyses on arabinose-induced cultures (Lesley *et al.* 2002). This effect could potentially explain other work in which recombinant protein production (induced by IPTG) was found to inhibit glucose uptake (Neubauer *et al.* 2003).

A similar catabolite repression effect has been observed upon production of a protease-sensitive β-galactosidase (Grossman 1984). It was proposed that either cAMP or a cofactor for adenylate cyclase was required by the proteolysis machinery. In either case, the effect of catabolite repression is likely not a general effect of recombinant protein production. It occurs in this system either because of proteolysis of α_1AT or IPTG induction.

7.3.4 Amino Acid Biosynthesis

7.3.4.1 Gene Expression Data

A decrease in amino acid biosynthesis genes upon induction was a common observation throughout this work (Figure 7.7). Table 7.4 lists amino acid biosynthesis genes identified from both the induction validation (0 min-vs.-60 min) experiment and transient gene analysis. These genes affect the synthesis of 18 of the common amino acids and four uncommon amino acids. The only common amino acids not represented on this list are glutamate and proline. Because the cultures in this induction validation experiment were grown in minimal medium, amino acid biosynthesis pathways were, in general, very active prior to induction. In fact, the rate of amino acid synthesis was probably the limiting factor for growth, and fluxes through these pathways were likely maximal. Table 7.5 lists several genes from the induction validation experiment that were involved in amino acid transport and were also significantly down-regulated 60 min after induction. Before induction, amino acid availability was low; therefore, these transport genes were likely induced in an attempt to maximize amino acid transport into the cell. Another set of amino acid transporter genes (*proP*, *proV*, *proW*, and *proX*) was also present in Table 4.4, but these genes showed *increased* expression. These genes encode two transport proteins that, in addition to proline, also transport molecules such as betaine, taurine, and ectoine, which contribute to defense against osmotic shock (Barron *et al.* 1987). Of these amino acid transport

genes, only *livJ* (decreased) and *proV* (increased) appear in the list of transient (Aeration-vs.-Time) genes.

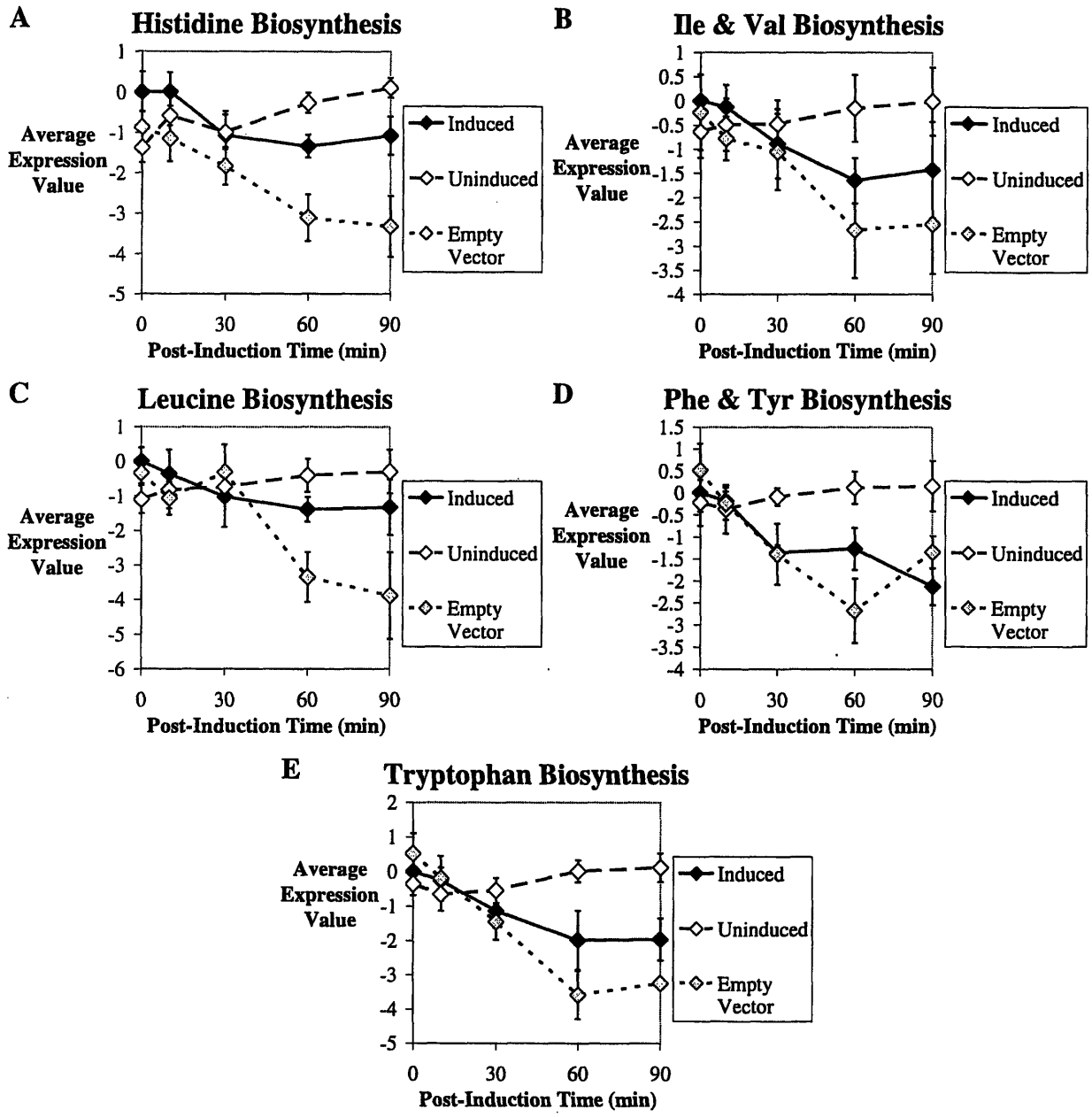


Figure 7.7: Amino Acid Biosynthesis Pathways in AIR Cultures
Average expression values from several amino acid biosynthesis pathways are plotted for cultures induced in air under various induction conditions. At 0 min, IPTG was added to the Induced and Empty-Vector cultures. The genes included in each plot are as follows: **A)** Hisidine Biosynthesis (*hisA/B/C/D/F/G/I*), **B)** Ile & Val Biosynthesis (*ilvA/B/C/D/E/G_2/H/I/M/N*), **C)** Leucine Biosynthesis (*ilvE, leuA/B/C/D, tyrB*), **D)** Phe & Tyr Biosynthesis (*pheA, tyrA/B*), **E)** Tryptophan Biosynthesis (*trpA/B/C/D/E*).

Table 7.4: Amino Acid Biosynthesis Genes with Decreased Expression 60 min following Induction

This table was compiled using amino acid biosynthesis genes from the 0 min-vs.-60 min experiment (Table 4.4) and the Aeration-vs.-Time experiment (Table 7.1, Table 7.2, and Table 7.3).

Genes			Amino Acids Affected
0 min-vs.-60 min Only	**Aeration-vs.-Time Only**	**Both**	
argA, argG, argI			Arg
aroL, pheA		*aroF, trpA, trpB, trpC, trpD, trpE, tyrA, tyrB*	Leu, Phe, Trp, Tyr
asd, asnA, aspC, dapD, panD			Asp, βAla, Lys, Hcy, Hse, Met, SAM, Thr
		cysC, cysD, cysH, cysI, cysJ, cysK, cysN	Cys
glyA, serC			Gly, Ser
hisB	*hisA, hisG*	*hisC, hisD, hisF*	His
ilvA, ilvB, ilvD, ilvM, ilvN	*ilvE, leuB*	*ilvC*	Ala, Ile, Leu, Val
glnA			Gln

Table 7.5: Amino Acid Transport Genes with Decreased Expression 60 min following Induction

This table was compiled from amino acid biosynthesis genes from Table 4.4.

Genes	Amino Acids Affected
argT, hisJ, hisM	Arg, His, Lys, Orn
artJ, artP	Arg
glnH	Gln
gltJ, gltK, gltL	Glu
livJ	Ile, Leu, Val

Examination of amino acid biosynthesis genes in the Uninduced and Empty-Vector cultures revealed an interesting trend across these cultures. As expected, many of the same amino acid biosynthesis and transport genes identified in Table 7.4 and Table 7.5 showed lower expression in the Induced culture than they did in the Uninduced culture. Surprisingly, the expression of these genes was even lower in the Empty-Vector cultures than in the Induced cultures.

7.3.4.2 Amino Acid Analysis

The decrease in expression of amino acid biosynthesis genes has two explanations. First, it may be an effect of catabolite repression. According to the EcoCyc database, there are 241 genes in the cAMP-CRP regulon. Five of these genes (*ansB*, *aspA*, *ilvB*, *ilvN*, and *serA*) are known to be amino acid biosynthesis genes that are activated by cAMP-CRP (Friden *et al.* 1982; Jennings and Beacham 1993; Golby *et al.* 1998; Yang *et al.* 2002). It is possible that the drop in cAMP levels led to lower expression of these five genes. Furthermore, secondary effects of cAMP-CRP regulation might have served to repress many other amino acid biosynthesis genes. Another explanation for decreased amino acid biosynthesis may be an increase in intracellular amino acid levels. Increased amino acid levels inside the cell would feedback to inhibit synthesis of the biosynthesis genes, as observed. Some of the regulators that are known to facilitate this feedback control include CysB, Lrp, TrpR, and TyrR. In an attempt to distinguish between these two effects, free amino acid levels were measured in cultures producing recombinant α_1AT. Increased amino acid levels would not necessarily eliminate the first explanation, but would strongly favor the second.

Amino acid analysis was performed as described in Section 3.9. Measurements of free amino acid levels were taken at 0 min (OD_{600} of 0.7) and 60 min from an Induced culture, an Uninduced culture, and an Empty-Vector culture. Data from the five amino acids with the clearest peaks are shown in Figure 7.8. At 60 min, for every amino acid shown here, the level in the Induced culture was higher than the level for either of the other cultures. Clearly induction resulted in increased amino acid levels 60 min after induction. Additionally, with the exception of valine, all of the 60-min levels in the Empty-Vector culture were higher than those in the Uninduced culture. Since the difference between thse two cultures is expression of the T7 RNA polymerase, it is concluded that production of this protein alone is sufficient to increase amino acid levels.

Figure 7.9 presents the increase in amino acid level from 0 min to 60 min. These data indicated a moderate increase in amino acid levels in the Empty-Vector culture and a strong increase in the Induced culture. The Uninduced culture theoretically produced no proteins, while the Empty-Vector culture produced one protein (T7), and the Induced culture produced two (T7 and α_1AT). Thus, the increase in amino acid level seemed to correlate with the number of heterologous proteins produced.

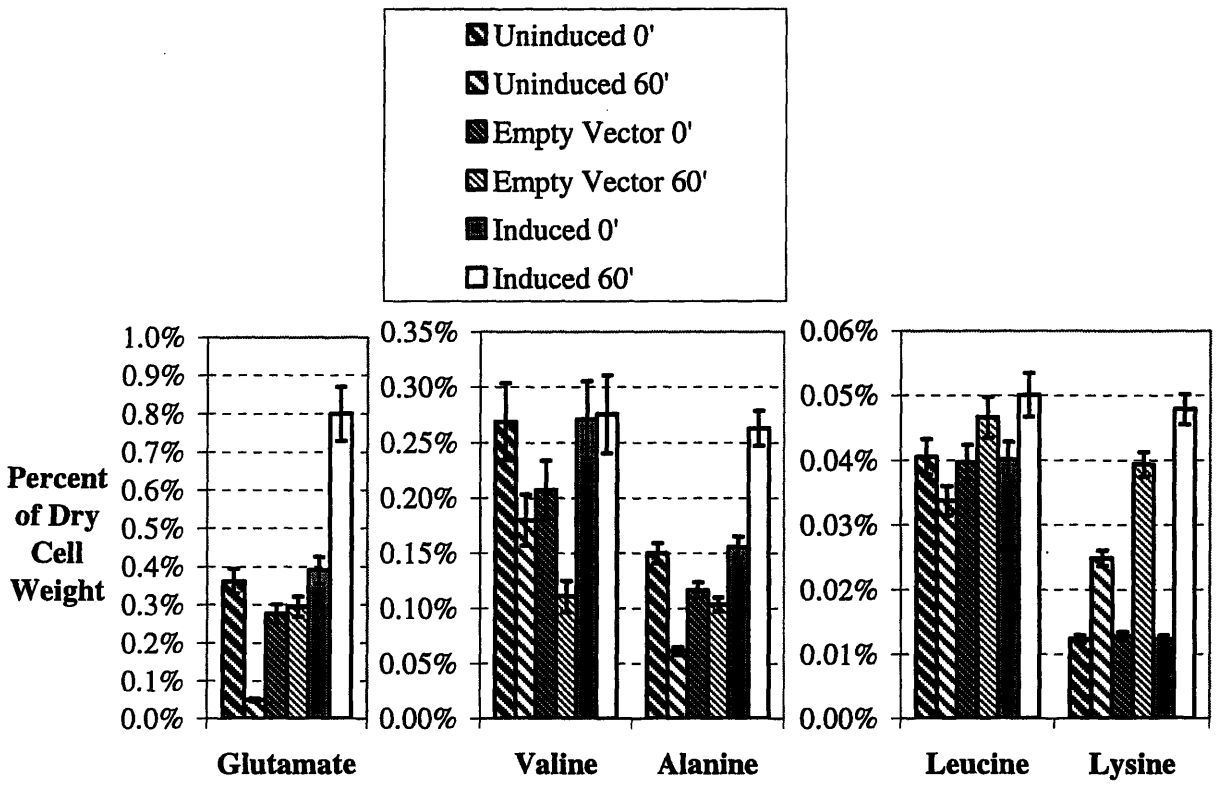


Figure 7.8: Free Amino Acid Analysis before and after Induction
Three cultures, two with the pEAT8-137 α_1AT vector and one with the pET3d empty vector, were grown to OD_{600} of 0.7 (0'). At this point, IPTG was added to one of the pEAT8-137 cultures (Induced) and the pET3d culture (Empty Vector), while nothing was added to the third culture (Uninduced). A second sample was taken at 60' and all six samples were analyzed by HPLC as described in Section 3.9. Data are shown here for five of the clearest peaks. Error bars represent the relative standard deviation from duplicate analyses of a standard solution with known concentrations.

7.3.4.3 Conclusions on Expression of Amino Acid Biosynthesis Genes

Two possible explanations were proposed to explain the increase in amino acid levels in response to induction. One explanation is that proteolysis has released amino acids. Production of an unfolded recombinant protein, like α_1AT, is known to stimulate the heat-shock response and associated proteolysis. The observation that both heat-shock genes and amino acid biosynthesis genes remain unchanged in Uninduced cultures is consistent with this hypothesis. However, heat-shock related proteolysis cannot explain the Empty-Vector cultures, in which heat-shock genes were repressed and amino acid levels increased.

A second explanation is that induction caused a decrease in overall translational rates. Immediately following induction, amino acid biosynthesis would have continued at the same

rate, causing the cell to accumulate amino acids faster than they could be incorporated into newly synthesized proteins. The T7 polymerase may have caused reduced translation by interfering with the normal transcription within the cell, which would have ultimately reduced transcript levels. Although T7 polymerase is known to be specific to its promoter, the cell likely contains many more polymerase molecules than copies of the promoter sequence. Overexpression of T7 RNA polymerases appears to be the most likely cause for the increase in amino acid levels.

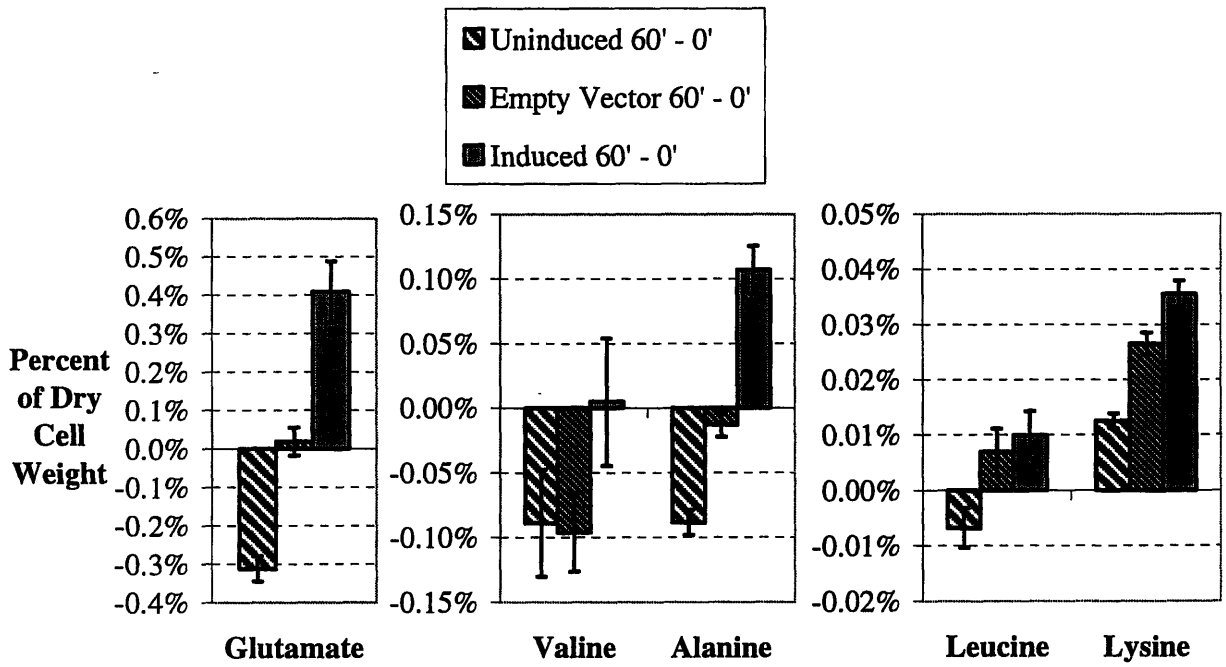


Figure 7.9: Free Amino Acid Increases after 60 min of Induction
Data shown here are differences between 60' and 0' samples from Figure 7.8. Error bars were calculated by propagation of error.

7.4 Conclusions on Effects of Recombinant Protein Production

Global transcriptional analysis of *E. coli* cultures producing a recombinant protein revealed both general effects and effects that are likely specific to this protein and/or expression system. As expected, heat-shock proteins were activated in response to α_1AT production. This effect is a result of the poorly folded recombinant α_1AT. Catabolite repression was observed following induction. It was proposed that this is an effect of IPTG addition. The resulting overproduction of lactose permease (LacY) may inhibit adenylate cyclase, thereby reducing cAMP levels. Decreased expression of amino acid biosynthesis genes was also observed. Analysis of free amino acids confirmed that amino acid levels increased upon induction.

Presumably, the biosynthesis pathways experienced feedback inhibition. The increased levels of free amino acids were thought to result from proteolysis, but this hypothesis was not consistent with expression data. Alternatively, the observed accumulation of amino acids likely resulted from decreased translational rates caused by transcriptional interference from overexpressed T7 RNA polymerase.

8 Contributions and Conclusions

"Bob Gibson is the luckiest pitcher I ever saw. He always pitches when the other team doesn't score any runs."

—Tim McCarver

Overall, this work demonstrated the utility of DNA microarrays in elucidating effects of process conditions on a bioprocess. The problem of α_1AT degradation was identified as a process issue that results from the negative effects of stress. This problem was further studied through both conventional microbiological techniques as well as global expression analysis. The combined results established links between the transcriptional response and process issues. More specific contributions and conclusions are as follows:

- Protocols for global gene expression analysis of *E. coli* using DNA microarrays were implemented and validated through a series of experiments that determined optimal scanning parameters, confirmed that DNA probes were not saturated, and quantified the reproducibility of the assay.
- Throughout this work, DNA microarray analysis was carried out using *E. coli* genomic DNA as a hybridization control. When consistently labeled in the Cy5 channel, genomic DNA was found to produce strong, reproducible signal in a high number of spots, with the added advantage of low non-specific binding to background regions. This approach allowed for direct comparison of all samples within this work. When compared with any total RNA sample, genomic DNA is more stable and reproducible across multiple preparations. This work confirmed that meaningful microarray data can be generated using genomic DNA as a hybridization control.
- Because of the combinations of factors studied throughout this work, an ANOVA model was chosen for analysis of microarray data. This model considered gene-specific effects as well as the effects of experimental treatments and blocks (repeated experiments). Considering imperfections in microarray data sets, an unbalanced solution was appropriate. A computationally efficient unbalanced solution to this three-way ANOVA model was developed by decomposing it into two-way ANOVA models, which were much easier to solve. ANOVA allowed for simultaneous normalization and selection of differentially expressed genes.

- Two genes encoding chaperones that are known to play a role in proteolytic mechanisms were identified as having oxygen-dependent expression.
    - The gene *clpA*, which encodes a protein that functions in concert with ClpP to degrade proteins, showed slight oxygen-dependent expression. However, oxygen-induced production of α_1AT in a ClpA$^-$ strain did not affect overall yields, and the experiments were inconclusive. *clpA* was found to show lower expression in iron-supplemented cultures, particularly when induced in oxygen. This is consistent with a model in which iron improved protein folding, which in turn reduced *clpA* expression.
    - The gene *grpE*, which encodes a heat-shock protease, was shown to have oxygen-dependent expression. Consistently, at the 10-min time point, expression in the O2 culture was found to be higher than that in the AIR culture. In addition, *dnaJ* was found to show repression upon iron supplementation, which indicates that free iron reduced the activity of the heat-shock response, possibly by countering its activation by superoxide. The products of these genes work in concert with the heat-shock chaperone DnaK to increase proteolytic degradation of misfolded proteins. Oxygen-dependent expression of one or all of these components could be responsible for the increased degradation of α_1AT in oxygen-induced cultures.
- Global gene expression data supported a model of oxygen toxicity in which superoxide oxidizes iron-sulfur clusters. This reaction not only generates free iron, but is also the major source of hydrogen peroxide. Both of these products participate in the Fenton reaction to create another reactive oxygen species, HO$^\bullet$. The pool of reduced iron-sulfur clusters quickly became depleted such that their regeneration limited peroxide formation and damage by HO$^\bullet$. With fewer reduced iron-sulfur clusters, superoxide accumulated to extreme levels, which may have damaged even highly resistant iron-sulfur clusters.
- Based on gene expression data, supplementation of iron was found to regenerate iron-sulfur clusters, thereby reducing the long-term effects of superoxide stress. A consequence of iron supplementation was increased short-term formation of hydrogen peroxide, which is consistent with the model of oxygen toxicity described above. Regeneration of iron-sulfur clusters was so effective that the anaerobic regulator FNR was activated, even under hyperoxic conditions.

- Supplementation of iron (in the form of $FeCl_2$) was found to reduce expression of genes involved in iron(III) uptake and increase expression of genes involved in iron(II) uptake. This result indicated that the original iron(II) was oxidized to iron(III), possibly via the Fenton reaction. During iron supplementation, reducing agents for both formation of iron(II) and formation of iron-sulfur clusters appeared to be limiting.
- A potential mechanism was identified whereby increased oxygen levels led to higher levels of unfolded protein, thereby stimulating the heat-shock response. Oxidation of iron-sulfur clusters may force the proteins that bear them into unstable conformations that are more susceptible to degradation. The resulting increase in the levels of unfolded proteins would stimulate the heat-shock response, leading to degradation of, not only the iron-sulfur proteins, but of other proteins as well, including the recombinant product.
- Proteins and pathways that were particularly susceptible to oxidation by superoxide were identified. This list included enzymes in the biosynthesis pathways for branched-chain amino acids (isoleucine, leucine, and valine), biotin, and thiamin. The thiamin biosynthesis pathway behaved similarly to those for both biotin and branched-chain amino acids, which are known to contain iron-sulfur enzymes. Unlike these pathways, no iron-sulfur clusters have been found in the thiamin biosynthesis pathway. An alternative explanation for the iron-sulfur dependence of this pathway came from the observation that one of the iron-sulfur repair enzymes also plays a role in thiamin biosynthesis.
- Two unexpected consequences of recombinant protein production were observed.
    - Induced cultures were found to display catabolite repression. A large number of genes in the cAMP-CRP regulon were consistently down-regulated as a result of induction. It was proposed that this was an effect of IPTG addition. The same component of the glucose PTS ($EIIA^{Glc}$) is known to both activate adenylate cyclase to generate cAMP and inhibit lactose permease. Increased levels of lactose permease, which result from IPTG addition, may occupy $EIIA^{Glc}$ so that adenylate cyclase cannot be activated. As a result, cAMP levels would drop.
    - Induced cultures showed decreased expression of amino acid biosynthesis genes. Furthermore, amino acid levels were found to increase in these cultures. While decreased expression of biosynthesis genes was explained by feedback inhibition, the cause of the amino acid accumulation was not determined. It was proposed

that this was a result of heat-shock induced proteolysis. However, Empty-Vector cultures, which did not exhibit a heat-shock response but did show increased amino acid levels, were inconsistent with this explanation. Another explanation was a decrease in translation rates, possibly due to non-specific interactions with T7 RNA polymerase.

9 Recommended Future Work

"The best thing about baseball is that you can do something about yesterday tomorrow."

—Manny Trillo

9.1 Improved Iron Supplementation

Because iron supplementation alleviated superoxide stress and removed the oxygen dependence of α_1AT degradation, it is a promising strategy for improving the scale-up and robutness of recombinant protein processes. It was found that supplemental iron did not improve the overall yields of α_1AT because the decreased rate of proteolysis was accompanied by a decreased rate of α_1AT folding. Overexpression of cellular chaperones in combination with iron supplementation may prove to maintain or increase the rate of proper folding, while decreasing the rate of proteolysis. According to modeling results presented in Chapter 6, this combination should result in improved α_1AT activity.

9.2 Potential Links Between Oxygen and Heat Shock

The heat-shock and hyperoxic responses are already known to overlap (Farr and Kogoma 1991), but the oxygen dependent expression of *grpE* could provide a strong link between these two stress responses. Because the expression profile of this gene is similar to those of OxyR-regulated genes, there is evidence to indicate the mechanism of this oxygen-dependent expression. The next step in exploring this potential regulation is to apply computational methods to identify an OxyR binding site in the promoter region of *grpE*. If the results of this work are promising, conventional microbiology techniques should then be used to confirm that its expression is oxygen dependent, and even further, OxyR dependent.

Although the heat-shock genes *clpP* and *clpX* did not show oxygen-dependent expression, *clpA* showed slight oxygen dependence. However, further work with a ClpA mutant strain was inconclusive. Careful analysis of this chaperone may elucidate its role, if any, in oxygen dependent degradation of recombinant α_1AT.

9.3 Recombinant Protein Production in Hypoxic Cultures

This work confirmed prior observations that hypoxic cultures exhibited surprisingly large α_1AT production (Laska 2000). Expression analysis of these hypoxic cultures served to further elucidate the metabolic changes that occurred during induction in pure nitrogen. Initial analysis of the data presented here found very little that was surprising in these hypoxic cultures. However, further analysis focused on why high production was achieved may find more in this data set.

Analysis of hypoxic cultures may elucidate strategies for improving production yields beyond those of aerobic cultures. Hypoxic recombinant protein fermentations with high yields would be incredibly valuable because they could be operated at high cell densities without the concern of oxygen transfer during scale-up. Such a fermentation strategy might consist of aerobic growth to a high cell density, followed by induction in a microaerobic or anaerobic environment. In following this path, it would be important to distinguish microaerobic cultures from anaerobic cultures, because the two may have different production yields.

9.4 Generality of Recombinant Protein Effects

It is unknown whether the effects observed with α_1AT are generally applicable. Work with other recombinant protein systems with different stability as well as different induction systems will help to determine the generality of each of the observed effects

9.4.1 Oxygen-Dependent Degradation

There are several examples of recombinant proteins that become oxidized and are subsequently degraded. However, this does not appear to be the case with α_1AT. It displays oxygen dependent degradation, despite the reducing environment of the *E. coli* cell. It should be determined whether this effect is unique to α_1AT.

9.4.2 Catabolite Repression

The observation of catabolite repression upon induction of recombinant protein production is quite interesting. It remains to be determined whether this is a general effect that would occur with any recombinant protein and any induction system. This effect may be specific to protease-sensitive products, or to systems in which IPTG is used to induce. The

global analysis presented here identified the transcription units *sdhCDAB*/b0725/*sucABCD*, *malEFG*, *malK*/*lamB*/*malM*, and *manXYZ* as the strongest signals of catabolite repression. Therefore, future analyses need only examine these genes in other recombinant protein production systems to narrow the causes of this effect.

9.4.3 Amino Acid Accumulation

The increase in amino acid levels after induction is poorly understood. It is not known whether this is a general effect that will occur in any recombinant protein system, or if it only occurs when T7 RNA polymerase is used to transcribe the heterologous gene. Examination of this effect, simply by analysis of free amino acid levels, in other induction systems may help to elucidate its cause. Additionally, measurement of translational rates by pulse-chase labeling both before and after induction would help to determine whether the effect was caused by elevated amino acid biosynthesis or reduced amino acid utilization.

9.5 Simulation of Oxygen Gradients at the Lab Scale

This work elucidated the effects of discrete changes in extreme oxygen environments on *E. coli*. The next step is to explore the effects of continual interchange between those extremes. In the past, these effects have been explored by running parallel fermentors at individual dissolved oxygen levels and pumping fluid between them. An improvement on this approach would be to use microreactors to grow many small-volume cultures, each at a different dissolved oxygen level. Allowing for two way flow between every adjacent reactor would simulate the effect of growth in a large-scale fermentor with an oxygen gradient. Applying expression analysis to such a system would help to further elucidate the effects of oxygen gradients as well as unexpected consequences of scale-up.

10 Appendix: ANOVA Modeling of DNA Microarray Data Sets

"It ain't over 'til it's over!"
—Yogi Berra

The Analysis of Variance (ANOVA) model described in this section is used for both normalization of DNA microarray data from different arrays as well as identification of differentially expressed genes. As with any ANOVA model, this model attributes statistical variation in microarray data to different factors. For example, variation in data might arise because the data come from different arrays, or because the data represent different genes. Data might also vary because of biological factors, such that data from a duplicated experiment would show differences in expression. Such duplicated experiments will be referred to as experimental blocks and should be distinguished from a repeated analysis of the same sample. For many, the ultimate goal of microarray analysis is to observe the variation that occurs as a result of different experimental treatments. Examples of treatments might be adding hydrogen peroxide to one culture and water to another, performing the same experiment on two different strains, or time-points taken from the same culture. The ANOVA model assigns statistical variation to the sources mentioned here (arrays, genes, blocks, and treatments), as well as combinations thereof.

To illustrate the use of this model, the equations will be developed and applied to the hypothetical DNA microarray data in Table 10.1. Both data sets consist of signal ratios for three genes (*genA*, *genB*, and *genC*) investigated on multiple arrays. The indices a, b, t, and g are used to refer to array, blocks, treatments, and genes, respectively. The index r refers to the number of duplicate spots on a particular array. Another feature of the model that was never applied in this work is the ability to analyze multiple slides hybridized with the same sample. The index n refers to the number of replicate analyses of the same sample. All variable used in this appendix are listed in Table 4.2.

Table 10.1: Example DNA Microarray Signal Ratio Data Sets (y_{agr} or y_{btgrn})
A) A balanced three-gene data set. B) An unbalanced three-gene data set.

A

	a	1	2	3	4	5	6
	b	1	1	1	2	2	2
	t	1	2	3	1	2	3
g	r						
1 - *genA*	1	-2	0.5	5.5	-5.5	0	1.5
	2	1	1.5	2.5	-3.5	0	4.5
2 - *genB*	1	-4	-1	-7.5	-5	4	-3.5
	2	-5	-2	-11	-2	5	-4.5
3 - *genC*	1	-9.5	-3.5	-1	-1.5	1	-4.5
	2	-11	-4.5	-4	-0.5	-1	-2.5

B

	a	1	2	4	5	6
	b	1	1	2	2	2
	t	1	2	1	2	3
g	r					
1 - *genA*	1	-0.5	-1.5	-5	0	2
	2		0.5		0	5
2 - *genB*	1	-4.5	-1	-5	4.5	
3 - *genC*	1	-9	-3.5	-2.5	0	-5
	2	-10		-1.5		-3

Every sample in the Table 10.1 data sets and those throughout this thesis belong to only one block and only one treatment. Because most of the experiments in this thesis were performed using an RNA sample in the Cy3 channel and a DNA standard in the Cy5 channel, only one sample is analyzed per array. Therefore, each array represents one block and one treatment, and selection of a particular array automatically selects a single sample as well as the block and treatment associated with that sample. However, the inverse is not necessarily true. Any sample can be placed on multiple arrays, this replicate analysis is embodied in the index n. At times, the index a will be used (*e.g.* y_{agr}), while at other times, the indices btn are more appropriate (*e.g.* y_{btgrn}). Keep in mind that both are equivalent.

Throughout this derivation, the variables α, γ, β_g, τ_g, and ρ_{ag} will be used to define the number of arrays, genes, blocks, treatments, and replicates-per-array, respectively. Three of these variables have indices, suggesting that they are not always constant values. As an example, consider ρ_{ag}. As the a and g indices indicate, the number of replicate measurements may be both

array- and gene-dependent. Some genes or arrays may have fewer spots that give quantifiable data; and would therefore have a smaller ρ_{ag} value.

Table 10.1A shows a balanced data set, while Table 10.1B shows an unbalanced data set. For our purposes, a balanced data set is one for which there are no missing observations. In mathematical terms, this constraint can be stated as $\rho_{btgn} = \rho \ \forall \ b,t,g,n$. The data set in Table 10.1A is balanced because there are two replicate measurements for each set of conditions (ρ = 2).

The data set in Table 10.1B is unbalanced for several reasons. First, some conditions have only one measurement, while others have two. For example, $\rho_{1111} = \rho_{2111} = \rho_{1231} = \rho_{2231} = 1$, while $\rho_{1211} = \rho_{2211} = \rho_{2311} = \rho_{1131} = \rho_{2131} = \rho_{2331} = 2$. In addition, the gene *genB* never had two measurements. This scenario would result when a particular gene is represented by only one spot on the array. Furthermore, there are no data for *genB* at $b = 2$, $t = 3$ ($\rho_{2321} = 0$). This data set is also unbalanced due to poor experimental design; there are no data for $b = 1$, $t = 3$, which means that no array analysis was performed for that sample.

Section 10.1 describes the solution to this ANOVA model assuming a balanced data set like that shown in Table 10.1A. However, this model solution is not used in this thesis, since actual data sets are unbalanced. Like the data set in Table 10.1B, actual data sets have two replicates for some genes and only a single replicate for others. In addition, actual data sets often have missing values that are filtered before analysis. Other experiments presented in this work also have experimental designs ($\{b, t\}$ combinations) that are unbalanced. Section 10.2 describes the modifications that are made to the balanced ANOVA model to apply it toward an unbalanced data set.

10.1 Balanced Data Sets

10.1.1 The Balanced Array-Gene ANOVA Model

The data set in Table 10.1A can be viewed as a two-way design (arrays × genes) with replicates. Therefore, a two-way ANOVA model is appropriate.

$$y_{agr} = \mu + A_a + G_g + AG_{ag} + \varepsilon'_{agr} \quad \forall a, g, r \tag{10.1}$$

This model is referred to as the Array-Gene ANOVA Model or Model 1. **(10.1)** attributes the variance between signal ratios (y_{agr}) to different effects. The signal ratio is modeled as a mean value (μ) with four sources of variation about that mean (A_a, G_g, AG_{ag}, and ε'_{agr}). The term A_a accounts for variation due to array-specific effects. This term will have the same value for all samples on the same array. The term G_g accounts for variation due to gene-specific effects. The AG_{ag} term, also known as the interaction term, accounts for effects of a particular gene on a particular array. For instance, if the spot for the gene *genA* has an above-average signal ratio on one array and a below-average signal ratio on another array, then this variation would be accounted for in the AG_{ag} term because it depends on both the array and the gene. All other variation falls into the ε'_{agr}, or residual, term. This term accounts for supposedly random variation among data that cannot be accounted for by the first three terms. This model equation is the basis for the Final ANOVA model.

10.1.1.1 Model Constraints

The Array-Gene ANOVA Model comes with several constraints that must be discussed. Since μ is the mean of all the signal ratios, it can be expressed as

$$\mu = \frac{\sum_{a=1}^{\alpha}\left[\sum_{g=1}^{\gamma}\left(\sum_{r=1}^{\rho} y_{agr}\right)\right]}{\alpha\gamma\rho} \tag{10.2a}$$

This equation can be abbreviated as

$$\mu = \frac{\sum_{agr} y_{agr}}{\alpha\gamma\rho} \tag{10.2b}$$

Writing one or more indices below the summation sign indicates a sum over the entire range of each index. **(10.2b)** can be rearranged as

$$\sum_{agr} y_{agr} = \alpha\gamma\rho\mu \tag{10.2c}$$

If all of the equations represented by **(10.1)** were summed, the result would be

$$\begin{aligned}\sum_{agr} y_{agr} &= \sum_{agr}\left[\mu + A_a + G_g + AG_{ag} + \varepsilon'_{agr}\right] \\ &= \alpha\gamma p\mu + \sum_{agr}\left[A_a + G_g + AG_{ag} + \varepsilon'_{agr}\right]\end{aligned} \tag{10.3}$$

Applying **(10.2c)** to this equation shows that the overall sum of the four remaining terms must be zero.

$$\sum_{agr}\left[A_a + G_g + AG_{ag} + \varepsilon'_{agr}\right] = 0 \tag{10.4}$$

Next, **(10.4)** is divided into individual constraints for each of the four sources of variation.

$$\sum_{agr} A_a = 0,\ \sum_{agr} G_g = 0,\ \sum_{agr} AG_{ag} = 0,\ \sum_{agr} \varepsilon'_{agr} = 0 \tag{10.5}$$

While for any particular signal ratio these individual sources of variation may be nonzero, the overall effect of these terms must be zero. For example, there may be significant variation between different arrays; however, these deviations will balance such that, overall, there will be no deviation from the mean due to array effects. The first two constraints in **(10.5)** can be easily simplified since the A_a parameters depend on neither g nor r. Similarly, the G_g parameters depend on neither a nor r.

$$\sum_{a} A_a = 0\ ,\ \sum_{g} G_g = 0 \tag{10.6}$$

In addition, the AG_{ag} parameters do not depend on r and the third constraint in **(10.5)** can be simplified to

$$\sum_{ag} AG_{ag} = 0 \tag{10.7}$$

The constraint in **(10.7)** can be simplified further. Consider, for example, what would happen if the constraint were met, but the statement $\sum_{a} AG_{a1} \neq 0$ were also true—in other words, for gene 1, the AG_{ag} terms sum to a nonzero value. If this were true, this variation would actually be gene-dependent, and should be accounted for as part of the G_1 parameter rather than the AG_{a1}

parameters. To avoid this scenario, **(10.7)** must be simplified such that the individual sums across all arrays and genes are zero.

$$\sum_{a} AG_{ag} = 0 \quad \forall g, \quad \sum_{g} AG_{ag} = 0 \quad \forall a \tag{10.8}$$

According to these constraints, if the AG_{ag} terms were written in matrix form, the sum of all rows and columns must be zero.

$$\begin{array}{ccccccc} AG_{11} & \cdots & AG_{1g} & \cdots & AG_{1\gamma} & 0 \\ \vdots & \ddots & & & \vdots & \vdots \\ AG_{a1} & & AG_{ag} & & AG_{a\gamma} & 0 \\ \vdots & & & \ddots & \vdots & \vdots \\ AG_{\alpha 1} & \cdots & AG_{\alpha g} & \cdots & AG_{\alpha\gamma} & 0 \\ 0 & \cdots & 0 & \cdots & 0 \end{array}$$

Without the constraints in **(10.8)**, it would be possible for gene- or array-specific effects to enter into the interaction term.

The final constraint in **(10.5)** can also be further simplified. Consider whether the constraint could be simplified to $\sum_{ag} \varepsilon'_{agr} = 0$. Expanding the expression $\sum_{ag} \varepsilon'_{agr}$ and applying the constraints in **(10.6)** and **(10.7)** would lead to the following.

$$\sum_{ag} \varepsilon'_{agr} = \sum_{ag} \left(y_{agr} - \mu - A_a - G_g - AG_{ag} \right) = \sum_{ag} \left(y_{agr} - \mu \right) = \sum_{ag} y_{agr} - \alpha\gamma\mu \tag{10.9}$$

If $\sum_{ag} \varepsilon'_{agr} = 0$, then according to the above equation, $\mu = \dfrac{\sum_{ag} y_{agr}}{\alpha\gamma}$, which is incorrect because it would require ρ different μ values—one for each r value—when there is really only one. Therefore the sum of the residuals over all a and g cannot be zero and, in order to satisfy the last constraint of **(10.5)**, the following must be true.

$$\sum_{r} \varepsilon'_{agr} = 0 \quad \forall a, g \tag{10.10}$$

The constraints in **(10.6)**, **(10.8)**, and **(10.10)** will be used to develop equations for determining the model parameters of the Array-Gene ANOVA model.

10.1.1.2 Degree-of-Freedom Analysis

In order to determine whether the model presented in **(10.1)** is a viable model, a degree-of-freedom (DOF) analysis must be performed. In analysis of variance, the variance from each source is compared to the residual variance. If the residual DOF were zero, then this comparison would be impossible and the analysis would be useless. Such a scenario would indicate that there are too few data to estimate all of the model parameters. Therefore, a viable model will have residual degrees of freedom greater than zero.

Table 10.2 shows the DOF analysis. In total, there are $\alpha\gamma\rho$ degrees of freedom, one for each data point. The single μ value always occupies one of these degrees of freedom. There are α A_a values, but because of the first constraint in **(10.6)**, one of these values can be determined if all the others are known. Therefore, there are only α - 1 degrees of freedom for A_a. Similarly, based on the second constraint of **(10.6)** there are $\gamma - 1$ degrees of freedom for G_g, and based on **(10.8)** there are $(\alpha - 1)(\gamma - 1)$ degrees of freedom for AG_{ag}. The residual degrees of freedom (DOF_{R1}) is the DOF for the ε'_{agr} parameters. This value is calculated by subtracting the sum of all other DOF's from the total DOF as shown below.

$$DOF_{R1} = \alpha\gamma\rho - [1 + (\alpha - 1) + (\gamma - 1) + (\alpha - 1)(\gamma - 1)] = \alpha\gamma(\rho - 1) \qquad \textbf{(10.11)}$$

For a single gene, the degrees of freedom for G_g and AG_{ag} need not be considered, and the gene-specific value, DOF_{R1g}, is calculated as follows

$$DOF_{R1g} = \alpha\rho - [1 + (\alpha - 1)] = \alpha(\rho - 1) \qquad \textbf{(10.12)}$$

If $\rho = 1$, both the overall and gene-specific residual DOF's become zero, and this model becomes unusable. However, for $\rho > 1$, the model is viable, and, with γ of approximately 4,000 for *E. coli*, DOF_{R1} for a real data set would be large.

Table 10.2: Degree-of-Freedom Analysis for the Balanced Array-Gene ANOVA Model

This table lists the model parameters, or sources of variation, from the balanced form of the ANOVA model in **(10.1)**. The general DOF for a balanced data set and the DOF for the example balanced data set in Table 10.1A are given. **(10.1)** is a viable model only when $\rho > 1$, since this makes the residual degrees of freedom greater than zero.

Parameter (Source of Variation)	Degrees of Freedom (DOF)	DOF for Example Balanced Data Set
μ	1	1
A_a	α-1	5
G_g	γ-1	2
AG_{ag}	$(\alpha$-1$)(\gamma$-1$)$	10
ε'_{agr}	$\alpha\gamma(\rho$-1$)$	18
y_{agr}	$\alpha\gamma\rho$	36

10.1.1.3 Determination of Model Parameters

The parameter μ in **(10.1)** is the mean of the signal ratios, as shown in **(10.2)**. This can be confirmed by minimizing the sum of the squares of the error or residual term, ε'_{agr} as

$$SSE_{\mu} = \sum_{agr} \varepsilon'^{2}_{agr} = \sum_{agr} \left(y_{agr} - \mu - A_a - G_g - AG_{ag}\right)^2 \qquad \textbf{(10.13)}$$

μ is determined by calculating the derivative of SSE_{μ} with respect to μ and setting that value to zero, thereby minimizing SSE_{μ}.

$$\frac{\partial SSE_{\mu}}{\partial \mu} = -2\sum_{agr}\left(y_{agr} - \mu - A_a - G_g - AG_{ag}\right) = 0 \qquad \textbf{(10.14a)}$$

Applying the constraints in **(10.6)** and **(10.8)** greatly simplifies this equation by eliminating the last three terms in the sum.

$$\frac{\partial SSE_{\mu}}{\partial \mu} = -2\sum_{agr}\left(y_{agr} - \mu\right) = 0 \qquad \textbf{(10.14b)}$$

Solving for μ produces

$$\sum_{agr}\left(y_{agr}-\mu\right)=0 \tag{10.15a}$$

$$\sum_{agr}\mu=\sum_{agr}y_{agr} \tag{10.15b}$$

$$\alpha\gamma p\mu=\sum_{agr}y_{agr} \tag{10.15c}$$

$$\mu=\frac{\sum_{agr}y_{agr}}{\alpha\gamma p}=\bar{y}_{\bullet\bullet\bullet} \tag{10.15d}$$

In the above equation, the notation of replacing the subscripts with bullets $(\bullet)$ indicates averaging over that index. **(10.15d)** confirms that μ is simply the mean of all the signal ratios.

To determine the parameters A_a, SSE_a is calculated as

$$SSE_a=\sum_{gr}\varepsilon'^2_{agr}=\sum_{gr}\left(y_{agr}-\mu-A_a-G_g-AG_{ag}\right)^2\ \forall a \tag{10.16}$$

Taking the derivative with respect to A_a, setting that value to zero, and eliminating extraneous terms produces

$$\frac{\partial SSE_a}{\partial A_a}=-2\sum_{gr}\left(y_{agr}-\mu-A_a\right)=0\ \forall a \tag{10.17}$$

Solving for A_a leads to

$$A_a=\frac{\sum_{gr}y_{agr}}{\gamma p}-\mu=\bar{y}_{a\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet}\ \forall a \tag{10.18}$$

Similar sets of equations can be used to determine the parameters G_g, and AG_{ag}. G_g parameters are determined by

$$SSE_g = \sum_{ar} \varepsilon'^2_{agr} = \sum_{ar} \left(y_{agr} - \mu - A_a - G_g - AG_{ag} \right)^2 \quad \forall\, g \tag{10.19}$$

$$\frac{\partial SSE_g}{\partial G_g} = -2 \sum_{ar} \left(y_{agr} - \mu - G_g \right) = 0 \quad \forall\, g \tag{10.20}$$

$$G_g = \frac{\sum_{ar} y_{agr}}{\alpha\rho} - \mu = \bar{y}_{\bullet g \bullet} - \bar{y}_{\bullet\bullet\bullet} \quad \forall\, g \tag{10.21}$$

AG_{ag} parameters are determined by

$$SSE_{ag} = \sum_{r} \varepsilon'^2_{agr} = \sum_{r} \left(y_{agr} - \mu - A_a - G_g - AG_{ag} \right)^2 \quad \forall\, a, g \tag{10.22}$$

$$\frac{\partial SSE_{ag}}{\partial AG_{ag}} = -2 \sum_{r} \left(y_{agr} - \mu - A_a - G_g - AG_{ag} \right) = 0 \quad \forall\, a, g \tag{10.23}$$

$$AG_{ag} = \frac{\sum_{r} y_{agr}}{\rho} - \mu - A_a - G_g = \bar{y}_{ag\bullet} - \bar{y}_{a\bullet\bullet} - \bar{y}_{\bullet g \bullet} + \bar{y}_{\bullet\bullet\bullet} \quad \forall\, a, g \tag{10.24}$$

Finally, rearrangement of **(10.1)** shows that the model parameter ε'_{agr} can easily be calculated once the other four parameters have been determined.

$$\varepsilon'_{agr} = y_{agr} - \mu - A_a - G_g - AG_{ag} \quad \forall\, a, g, r \tag{10.25}$$

Least squares minimization leads to the equations **(10.15)**, **(10.18)**, **(10.21)**, **(10.24)**, and **(10.25)**, which will provide the values for the model parameters as described in the following section.

10.1.1.4 Solving the Model

The solution of the Balanced Array-Gene ANOVA Model is trivial. First, the signal ratio averages $\bar{y}_{\bullet\bullet\bullet}, \bar{y}_{a\bullet\bullet}, \bar{y}_{\bullet g\bullet}, \bar{y}_{ag\bullet}$ must be calculated from the original data. Then, the equations derived above can be used to directly calculate the model parameters from these averages. A summary of the equations used to calculate the model parameters is shown in Figure 10.1. Note that there are many equally valid combinations of equations that can be used. This figure only

describes one combination. Note also that this solution consists of $(1 + \alpha + \gamma + \alpha\gamma + \alpha\gamma\rho)$ equations for calculating the same number of parameters.

$$\mu = \bar{y}_{...} \qquad (1)$$

$$A_a = \bar{y}_{a..} - \bar{y}_{...} \quad \forall a = 1 \ldots \alpha - 1 \qquad (\alpha\text{-}1)$$

$$\sum_a A_a = 0 \qquad (1)$$

$$G_g = \bar{y}_{.g.} - \bar{y}_{...} \quad \forall g = 1 \ldots \gamma - 1 \qquad (\gamma\text{-}1)$$

$$\sum_g G_g = 0 \qquad (1)$$

$$AG_{ag} = \bar{y}_{ag.} - \bar{y}_{a..} - \bar{y}_{.g.} + \bar{y}_{...} \quad \forall a = 1 \ldots \alpha - 1, g = 1 \ldots \gamma - 1 \qquad (\alpha\text{-}1)(\gamma\text{-}1)$$

$$\sum_a AG_{a\gamma} = 0 \quad \forall a = 1 \ldots \alpha - 1 \qquad (\alpha\text{-}1)$$

$$\sum_g AG_{\alpha g} = 0 \quad \forall g \qquad (\gamma)$$

$$\varepsilon'_{agr} = y_{agr} - \mu - A_a - G_g - AG_{ag} \quad \forall a, g, r \qquad (\alpha\gamma\rho)$$

Figure 10.1: Summary of Equations for Solving the Balanced Array-Gene ANOVA Model
Numbers in parentheses indicate the number of equations represented

Even for a real data set with approximately 4,000 genes, the calculations can easily be done in a spreadsheet. Performing these calculations for the balanced data set in Table 10.1A produces the values in Figure 10.2. To this point, the solution strategy for the Balanced Array-Gene ANOVA model is equivalent to that described in other sources (Box *et al.* 1978; Kerr *et al.* 2000).

10.1.2 Normalization of Balanced Data Sets

Before data from different arrays can be compared, the data must be normalized to account for array-to-array variation. Up to this point, the difficult task of identifying and

quantifying this variation has already been done. Array-to-array variation is represented in the Array-Gene ANOVA Model by the A_a term. A balanced data set can be normalized by subtracting this term from each signal ratio (Kerr *et al.* 2000). The normalized signal ratios would be expressed as

$$\hat{y}_{agr} = y_{agr} - A_a = \mu + G_g + AG_{ag} + \varepsilon'_{agr} \quad \forall a, g, r \tag{10.26}$$

Applying this normalization to the data in Table 10.1A produces the data in Table 10.3B.

μ

-2

G_g

g	
1	2.5
2	-1
3	-1.5

A_a

a	
1	-3
2	0.5
3	-0.5
4	-1
5	3.5
6	0.5

AG_{ag}

g \ a	1	2	3	4	5	6
1	2	0	4	-4	-4	2
2	1.5	1	-5.5	0.5	4	-1.5
3	-3.5	-1	1.5	3.5	0	-0.5

ε'_{agr}

g	r \ a	1	2	3	4	5	6
1	1	-1.5	-0.5	1.5	-1	0	-1.5
	2	1.5	0.5	-1.5	1	0	1.5
2	1	0.5	0.5	1.5	-1.5	-0.5	0.5
	2	-0.5	-0.5	-1.5	1.5	0.5	-0.5
3	1	0.5	0.5	1.5	-0.5	1	-1
	2	-0.5	-0.5	-1.5	0.5	-1	1

Figure 10.2: Solution to the Array-Gene ANOVA Model for the Balanced Example Data Set

This method is called a global normalization because, for each array, it uses all of the data for all of the genes. Every signal ratio plays a role in the calculations of the parameters μ and A_a. Such normalizations assume that during any experiment, changes in gene expression will balance. Some genes increase in expression level, while others decrease; but overall, there is no significant change in either direction. Therefore, the genes that change in expression level will balance out such that the average signal ratio should be the same on all arrays. For a balanced data set, this normalization method is equivalent to setting the average or total signal ratio to the same value for each array. Notice that each array in the normalized data set in Table 10.3B has a total signal ratio of -12 and an average signal ratio of -2.

Table 10.3: Original and Normalized Balanced Data Sets
A) Original signal ratio data (reproduced from Table 10.1A). **B)** Normalized signal ratio data

A

	a	1	2	3	4	5	6
	b	1	1	1	2	2	2
	t	1	2	3	1	2	3
g	*r*						
1 - *genA*	1	-2	0.5	5.5	-5.5	0	1.5
	2	1	1.5	2.5	-3.5	0	4.5
2 - *genB*	1	-4	-1	-7.5	-5	4	-3.5
	2	-5	-2	-11	-2	5	-4.5
3 - *genC*	1	-9.5	-3.5	-1	-1.5	1	-4.5
	2	-11	-4.5	-4	-0.5	-1	-2.5

B

	a	1	2	3	4	5	6
	b	1	1	1	2	2	2
	t	1	2	3	1	2	3
g	*r*						
1 - *genA*	1	1	0	6	-4.5	-3.5	1
	2	4	1	3	-2.5	-3.5	4
2 - *genB*	1	-1	-1.5	-7	-4	0.5	-4
	2	-2	-2.5	-10	-1	1.5	-5
3 - *genC*	1	-6.5	-4	-0.5	-0.5	-2.5	-5
	2	-7.5	-5	-3.5	0.5	-4.5	-3

If the data set consists of a small number of genes, it may contain an disproportionately large number of differentially expressed genes; therefore, a global normalization may not be appropriate. Although the sample data set from Table 10.1A falls in this category, the global normalization is applied for illustrative purposes only.

10.1.3 The Balanced Block-Treatment ANOVA Model

Once the global normalization is balanced, there is no longer any need to view the data set as a whole. Instead, each gene can be examined individually, and the experimental treatments that produce signal ratios that significantly differ from those of other treatments can be performed. However, to this point, variation due to experimental treatments has not entered into the model. The Block-Treatment ANOVA Model is a second two-way model that will introduce variation not only due to experimental treatments, but to experimental blocks as well.

Since the μ term is a global term, we can combine it with the term G_g to produce a new gene-specific term.

$$\mu + G_g = \mu_g \quad \forall\, g \tag{10.27}$$

μ_g is simply the mean normalized signal ratio for gene g. Substituting this expression into the normalized Array-Gene ANOVA Model in **(10.26)** produces

$$\hat{y}_{agr} = \mu_g + AG_{ag} + \varepsilon'_{agr} \quad \forall\, a, g, r \tag{10.28}$$

For a given gene, the first term of this model is the mean value for that gene, while the second term accounts for the variation between arrays. The third term is, of course, the residual error

term. Since every term has gene dependence, **(10.28)** makes it clear that the model parameters for each gene are completely independent of those for other genes. The remainder of the discussion in Section 10.1 will focus on individual genes, but is applicable to all genes.

The AG_{ag} term can be further analyzed to account for variation between different blocks and treatments. The experimental layout is a two-way design (blocks × treatments), where each array represents exactly one block and one treatment. Table 10.4 shows the experimental design for the Table 10.1A data. To incorporate this into the model, the variation due to array effects (AG_{ag}) is divided into variation due to experimental blocks (BG_{bg}) and treatments (TG_{tg}) . To illustrate this change, the subscripts a will now change to btn (*e.g.* y_{btgrn} and AG_{btgn}). To implement this design into the ANOVA model, a second two-way ANOVA model is created.

$$AG_{btgn} = BG_{bg} + TG_{tg} + \varepsilon''_{btgn} \quad \forall b,t,g,n \tag{10.29}$$

Note that the above model does not contain a mean term, since—according to the constraints in **(10.8)**—the mean is zero for all a and g.

Table 10.4: Experimental Design for Example Balanced Data Set
Since each array represents exactly one block and one treatment, each array represents an element in a (blocks × treatments) matrix. Each number in the center of the table represents an array.

b \ *t*	1	2	3
1	1	2	3
2	4	5	6

Note that a similar model (Model 3) can be developed to further analyze the A_a term into B_b, T_t, and BT_{bt} terms.

$$A_{btn} = B_b + T_t + BT_{bt} + \varepsilon'''_{btn} \quad \forall b,t,n \tag{10.30}$$

However, applying this model would be pointless since the A_a term is eliminated during the normalization step.

The same steps that were taken for the Array-Gene ANOVA Model (determination of constraints, DOF analysis, and derivation of parameters) must also be taken for the model in **(10.29)** for every gene.

10.1.3.1 Model Constraints

The constraints for the Block-Treatment ANOVA Model in **(10.29)**, are developed based on the constraints in **(10.8)**, which can be rewritten as

$$\sum_{g} AG_{btgn} = 0 \quad \forall b,t,n, \sum_{btn} AG_{btgn} = 0 \quad \forall g \tag{10.31}$$

The first of these constraints is expanded to

$$\sum_{g} AG_{btgn} = \sum_{g} \left(BG_{bg} + TG_{tg} + \varepsilon''_{btgn}\right) = 0 \quad \forall b,t,n \tag{10.32}$$

As before, each term of this constraint becomes a separate constraint.

$$\sum_{g} BG_{bg} = 0 \quad \forall b, \sum_{g} TG_{tg} = 0 \quad \forall t, \sum_{g} \varepsilon''_{btgn} = 0 \quad \forall b,t,n \tag{10.33}$$

A similar procedure can be followed for the second constraint of **(10.31)**.

$$\sum_{btn} AG_{btgn} = \sum_{btn} \left(BG_{bg} + TG_{tg} + \varepsilon''_{btgn}\right) = 0 \quad \forall g \tag{10.34}$$

$$\sum_{btn} BG_{bg} = 0 \quad \forall g, \sum_{btn} TG_{tg} = 0 \quad \forall g, \sum_{btn} \varepsilon''_{btgn} = 0 \quad \forall g \tag{10.35}$$

Since BG_{bg} is independent of both t and n, and since TG_{tg} is independent of both b and n, the first two of these equations can be simplified to

$$\sum_{b} BG_{bg} = 0 \quad \forall g, \sum_{t} TG_{tg} = 0 \quad \forall g \tag{10.36}$$

Finally, for reasons similar to those explained previously for the AG_{btgn} constraints in **(10.8)**, the last constraint of **(10.35)** must also be simplified to

$$\sum_{bn} \varepsilon''_{btgn} = 0 \quad \forall t,g, \sum_{tn} \varepsilon''_{btgn} = 0 \quad \forall b,g \tag{10.37}$$

10.1.3.2 Degree-of-Freedom Analysis

A degree-of-freedom analysis must be performed on this model to determine whether it is valid. For any given gene, there are α—or $\beta\tau\nu$—AG_{btgn} values, which are constrained by the following equation from **(10.31)**.

$$\sum_{bt} AG_{btgn} = 0 \quad \forall\, g \tag{10.38}$$

For a given gene, this constraint occupies one degree of freedom; therefore, only $\beta\tau\nu$-1 degrees of freedom remain. Due to the constraints in **(10.36)**, there are β-1 DOF for the β BG_{bg} terms and τ-1 DOF for the τ TG_{tg} terms. As before, the residual degrees-of-freedom (DOF_{R2g}), the DOF for ε''_{btgn}, is calculated by subtracting the DOF for all other parameters from the total DOF.

$$DOF_{R2g} = (\beta\tau\nu - 1) - [(\beta - 1) + (\tau - 1)] = (\beta - 1)(\tau - 1) + \beta\tau(\nu - 1) \tag{10.39}$$

Table 10.5 summarizes the results of this DOF analysis for the general case as well as for the Table 10.1A data. According to the DOF analysis, $DOF_{R2g} > 0$ when either of two conditions is met: **1)** $\beta > 1$ and $\tau > 1$, or **2)** $\nu > 1$. Therefore, the Block-Treatment ANOVA Model is valid only for genes that satisfy these conditions. In this work, $\nu = 1$ for all data sets; therefore, the first condition must be met in order for this model to be valid. For the overall data set, the degrees-of-freedom for AG_{ag}, BG_{bg}, and TG_{tg} are calculated by multiplying by a factor of (γ-1), because of the constraints in **(10.31)** and **(10.33)**. DOF_{R2} is calculated as follows.

$$\begin{aligned} DOF_{R2} &= (\beta\tau\gamma\nu - 1)(\gamma - 1) - [(\beta - 1)(\gamma - 1) + (\tau - 1)(\gamma - 1)] \\ &= (\gamma - 1)[(\beta - 1)(\tau - 1) + \beta\tau(\nu - 1)] \end{aligned} \tag{10.40}$$

10.1.3.3 Determination of Model Parameters

The parameters in **(10.29)**, BG_{bg} and TG_{tg}, are calculated in a manner analogous to that applied to the parameters in the Array-Gene ANOVA Model. The parameter BG_{bg} is calculated by

$$SSE_{bg} = \sum_{tn} \varepsilon''^{2}_{btgn} = \sum_{tn} \left(AG_{btgn} - BG_{bg} - TG_{tg}\right)^2 \quad \forall\, b, g \tag{10.41a}$$

Applying the constraint in **(10.36)**, the above simplifies to

$$SSE_{bg} = \sum_{tn} \varepsilon''^{2}_{btgn} = \sum_{tn} \left(AG_{btgn} - BG_{bg}\right)^2 \quad \forall\, b, g \tag{10.41b}$$

The minimum value of SSE_{bg} is calculated by taking the derivative with respect to BG_{bg} and setting that to zero, as shown below.

$$\frac{\partial SSE_{bg}}{\partial BG_{bg}} = -2\sum_{tn}\left(AG_{btgn} - BG_{bg}\right) = 0 \quad \forall b, g \tag{10.42}$$

Solving for BG_{bg} produces the following.

Table 10.5: Degree-of-Freedom Analysis for the Balanced Block-Treatent ANOVA Model

This table lists the model parameters from the balanced form of the ANOVA model in **(10.29)**. The DOF for a single gene in a balanced data set and the DOF for any gene in Table 10.1A are given. **(10.29)** is a valid model when either **1)** $\beta > 1$ and $\tau > 1$ or **2)** $v > 1$, since the residual degrees of freedom would be greater than zero.

Parameter (Source of Variation)	Degrees of Freedom (DOF)	DOF for Example Balanced Data Set
BG_{bg}	$(\beta\text{-}1)$	1
TG_{tg}	$(\tau\text{-}1)$	2
ε''_{btgn}	$(\beta\text{-}1)(\tau\text{-}1)+\beta\tau(v\text{-}1)$	2
AG_{btgn}	$\beta\tau v\text{-}1$ OR $\alpha\text{-}1$	5

$$BG_{bg} = \frac{\sum_{tn} AG_{btgn}}{\tau v} = \overline{A}\overline{G}_{b\bullet g\bullet} \quad \forall b, g \tag{10.43}$$

A similar procedure can be carried out for TG_{tg}.

$$SSE_{tg} = \left[\sum_{bn} \varepsilon''_{btgr}\right]^2 = \left[\sum_{bn}\left(AG_{btgn} - TG_{tg}\right)\right]^2 \quad \forall t, g \tag{10.44}$$

$$\frac{\partial SSE_{tg}}{\partial TG_{tg}} = -2\left[\sum_{bn}\left(AG_{btgn} - TG_{tg}\right)\right] = 0 \quad \forall t, g \tag{10.45}$$

$$TG_{tg} = \frac{\sum_{bn} AG_{btgn}}{\beta v} = \overline{A}\overline{G}_{\bullet tg\bullet} \quad \forall t, g \tag{10.46}$$

Rearranging **(10.29)** shows that the residual error can be calculated using the following formula.

$$\varepsilon''_{btgn} = AG_{btgn} - BG_{bg} - TG_{tg} \quad \forall b,t,g,n \tag{10.47}$$

10.1.3.4 Solving the Model

The equations for calculating the balanced Block-Treatment ANOVA Model parameters are summarized in Figure 10.3. Even when this set of equations is applied to every gene, the calculations are easily done in a spreadsheet. Applying these equations to the Table 10.1A data produces the model parameters in Figure 10.4. Notice that each of the constraints in **(10.33)**, **(10.36)**, and **(10.37)** is met by these data.

$$BG_{bg} = \overline{A}\,\overline{G}_{b\bullet g\bullet} \quad \forall b = 1\ldots\beta-1, g = 1\ldots\gamma-1 \qquad (\beta\text{-}1)(\gamma\text{-}1)$$

$$\sum_b BG_{bg} = 0 \quad \forall g = 1\ldots\gamma-1 \qquad (\gamma\text{-}1)$$

$$\sum_g BG_{bg} = 0 \quad \forall b \qquad (\beta)$$

$$TG_{tg} = \overline{A}\,\overline{G}_{\bullet tg\bullet} \quad \forall t = 1\ldots\tau-1, g = 1\ldots\gamma-1 \qquad (\tau\text{-}1)(\gamma\text{-}1)$$

$$\sum_t TG_{tg} = 0 \quad \forall g = 1\ldots\gamma-1 \qquad (\gamma\text{-}1)$$

$$\sum_g TG_{tg} = 0 \quad \forall t \qquad (\tau)$$

$$\varepsilon''_{btgn} = AG_{btgn} - BG_{bg} - TG_{tg} \quad \forall b,t,g,n \qquad (\beta\tau\gamma\nu)$$

Figure 10.3: Summary of Equations for Solving the Balanced Block-Treatment ANOVA Model
Numbers in parentheses indicate the number of equations represented

Combining the Block-Treatment ANOVA Model with the Array-Gene ANOVA Model—along with the following substitution of error terms

$$\varepsilon'_{btgrn} + \varepsilon''_{btgn} = \varepsilon_{btgrn} \tag{10.48}$$

produces the final model for normalized signal ratios.

$$\hat{y}_{btgrn} = \mu_g + BG_{bg} + TG_{tg} + \varepsilon_{btgrn} \quad \forall b,t,g,r,n \tag{10.49}$$

BG_{bg}

g \ b	1	2
1	2	-2
2	-1	1
3	-1	1

TG_{tg}

g \ t	1	2	3
1	-1	-2	3
2	1	2.5	-3.5
3	0	-0.5	0.5

ε''_{btgr}

	a	b	t	g	1	2	3	4	5	6
	1	1	1	2	2	2				
	1	2	3	1	2	3				
1	1	0	-1	-1	0	1				
2	1.5	-0.5	-1	-1.5	0.5	1				
3	-2.5	0.5	2	2.5	-0.5	-2				

Figure 10.4: Solution to the Block-Treatment ANOVA Model for the Example Balanced Data Set

10.1.4 Identifying Differentially Expressed Genes Using Balanced ANOVA Models

Based on the parameter values calculated from this balanced ANOVA model, expression values are calculated (on a base-2 log scale) as the treatment-averaged normalized signal ratios.

$$\bar{\hat{y}}_{\bullet tg\bullet\bullet} = \frac{\sum_{brn} \hat{y}_{btgrn}}{\beta\rho\nu} = \mu_g + TG_{tg} \quad \forall t,g \tag{10.50}$$

These expression values are displayed in Table 10.6. To determine the effect of a particular treatment comparison, log ratios are calculated as the difference between two expression values.

$$LR_{ijg} = \bar{\hat{y}}_{\bullet jg\bullet\bullet} - \bar{\hat{y}}_{\bullet ig\bullet\bullet} = TG_{jg} - TG_{ig} \quad \forall g, i < j \tag{10.51}$$

Based on these log ratios, the gene-specific and global significance tests can be applied as described in Section 4.5.3.

Within two genes (1 - *genA* and 2 - *genB*), four comparisons were found to be significant ($t = 1$ vs. 3, $t = 2$ vs. 3 for each gene).

Table 10.6: Expression Values for Example Balanced Data Set

$\bar{\hat{y}}_{\bullet tg\bullet\bullet}$

g \ t	1	2	3
1 - *genA*	-0.5	-1.5	3.5
2 - *genB*	-2	-0.5	-6.5
3 - *genC*	-3.5	-4	-3

10.2 Unbalanced Data Sets

Because the balanced solution described in the previous section is relatively simple to derive and solve, it is tempting to apply it to unbalanced data sets as well, such as those given in Table 10.1B. However, this would only be an estimated solution. A more appropriate solution for the unbalanced data set requires consideration of the number of data available for each possible combination of conditions. This section discusses how the derivation and solution of the balanced ANOVA change for an unbalanced data set. The equations derived here are the ones that were actually applied to the real data sets in this work. This section ends with a comparison with the estimated solution using the balanced equations from Section 10.1.

10.2.1 The Unbalanced Array-Gene ANOVA Model

For the case of an unbalanced data set, the Array-Gene ANOVA Model in **(10.1)** remains the same; however, the constraints, parameter calculations, and solution of this model all change for the unbalanced ANOVA.

10.2.1.1 Model Constraints

In **(10.2)**, we were able to make the assumption that $\sum_{ag} \rho_{ag} = \alpha\gamma p$, which no longer holds for an unbalanced data set. This equation is adjusted accordingly.

$$\mu = \frac{\sum_{agr} y_{agr}}{\sum_{ag} \rho_{ag}} \qquad \textbf{(10.52a)}$$

Rearranging **(10.52a)** produces

$$\sum_{agr} y_{agr} = \mu \sum_{ag} \rho_{ag} \tag{10.52b}$$

Summing all of the equations represented by **(10.1)** produces

$$\begin{aligned} \sum_{agr} y_{agr} &= \sum_{agr} \left(\mu + A_a + G_g + AG_{ag} + \varepsilon'_{agr}\right) \\ &= \mu \sum_{ag} \rho_{ag} + \sum_{a} \left(A_a \sum_{g} \rho_{ag} \right) + \sum_{g} \left(G_g \sum_{a} \rho_{ag} \right) + \sum_{ag} \left(AG_{ag} \rho_{ag}\right) + \sum_{agr} \varepsilon'_{agr} \end{aligned} \tag{10.53}$$

Applying **(10.52b)** to **(10.53)** reveals that the last four terms sum to zero.

$$\sum_{a} \left(A_a \sum_{g} \rho_{ag} \right) + \sum_{g} \left(G_g \sum_{a} \rho_{ag} \right) + \sum_{ag} \left(AG_{ag} \rho_{ag}\right) + \sum_{agr} \varepsilon'_{agr} = 0 \tag{10.54}$$

Each of the terms in this equation can be considered to be a separate constraint. For instance, the first two terms produce the constraints

$$\sum_{a} \left(A_a \sum_{g} \rho_{ag} \right) = 0, \quad \sum_{g} \left(G_g \sum_{a} \rho_{ag} \right) = 0 \tag{10.55}$$

The third term in **(10.54)** produces the following constraint.

$$\sum_{ag} \left(AG_{ag} \rho_{ag}\right) = 0 \tag{10.56}$$

As was done with the balanced ANOVA, this constraint can be further divided.

$$\sum_{a} \left(AG_{ag} \rho_{ag}\right) = 0 \ \ \forall g, \quad \sum_{g} \left(AG_{ag} \rho_{ag}\right) = 0 \ \ \forall a \tag{10.57}$$

The constraint involving ε'_{agr} in **(10.10)** remains the same. The constraints defined in **(10.55)**, **(10.57)**, and **(10.10)** are used to determine each of the parameters in the Unbalanced Array-Gene ANOVA Model.

10.2.1.2 Degree-of-Freedom Analysis

The model solution derived here allows for genes to be included for which no data are available. While these genes would be included in the γ count, they would not be counted in the γ_{Real} parameter. The degree of freedom analysis applied in the unbalanced ANOVA differs from

that for the balanced ANOVA in that γ is replaced by γ_{Real} and $\alpha\gamma\rho$ is replaced by $\sum_{ag}\rho_{ag}$. In addition, the number of AG_{ag} values is not necessarily represented by $\alpha\gamma_{Real}$, but depends on the number of conditions for which valid data were collected. The variable ζ_{ag} is used to represent the number of AG_{ag} values for which $\rho_{ag} = 0$. The degrees-of-freedom for each parameter are listed in Table 10.7. The residual degrees of freedom is calculated as follows

$$\begin{aligned} DOF_{R1} &= \sum_{ag}\rho_{ag} - \left[1+(\alpha-1)+(\gamma_{Real}-1)+\left(\sum_{ag}\zeta_{ag}-\gamma_{Real}-\alpha+1\right)\right] \\ &= \sum_{ag}\rho_{ag} - \sum_{ag}\zeta_{ag} \end{aligned} \qquad \textbf{(10.58)}$$

In order for this model to be valid, the degrees of freedom for both AG_{ag} as well as ε'_{agr} must be greater than zero. This creates two criteria that the data set must pass before this model can be applied: **1)** $\sum_{ag}\zeta_{ag} > (\gamma_{Real} + \alpha - 1)$ and **2)** $\sum_{ag}\rho_{ag} > \sum_{ag}\zeta_{ag}$.

Table 10.7: Degree-of-Freedom Analysis for the Unbalanced Array-Gene ANOVA Model

This table lists the model parameters, or sources of variation, from the unbalanced form of the ANOVA model in **(10.1)**. The general DOF for an unbalanced data set and the DOF for the example unbalanced data set in Table 10.1B are given. This model is valid when the following constraints are met $\sum_{ag}\rho_{ag} > \sum_{ag}\zeta_{ag} > (\gamma_{Real} + \alpha - 1)$.

Parameter (Source of Variation)	Degrees of Freedom (DOF)	DOF for Example Unbalanced Data Set
μ	1	1
A_a	α-1	4
G_g	γ_{Real}-1	2
AG_{ag}	$\sum_{ag}\zeta_{ag} - \gamma_{Real} - \alpha + 1$	7
ε'_{agr}	$\sum_{ag}\rho_{ag} - \sum_{ag}\zeta_{ag}$	6
y_{agr}	$\sum_{ag}\rho_{ag}$	20

This DOF analysis can also be applied to individual genes to produce the following.

$$DOF_{R1g} = \sum_{a} \rho_{ag} - \sum_{a} \zeta_{ag} \tag{10.59}$$

10.2.1.3 Determination of Model Parameters

As with the balanced data set, it can be shown that the term μ is simply the mean of all of the signal ratios. The sum of the squares of all error terms is calculated and minimized as in **(10.13)** and **(10.14a)**, resulting in

$$\sum_{agr} \left(y_{agr} - \mu - A_a - G_g - AG_{ag} \right) = 0 \tag{10.60a}$$

Applying this summation to all terms individually produces

$$\sum_{agr} y_{agr} - \mu \sum_{ag} \rho_{ag} - \sum_{a} \left(A_a \sum_{g} \rho_{ag} \right) - \sum_{g} \left(G_g \sum_{a} \rho_{ag} \right) - \sum_{ag} \left(A_{ag} \rho_{ag} \right) = 0 \tag{10.60b}$$

Applying the constraints from **(10.55)** and **(10.57)** simplifies this equation to

$$\sum_{agr} y_{agr} - \mu \sum_{ag} \rho_{ag} = 0 \tag{10.60c}$$

$$\mu \sum_{ag} \rho_{ag} = \sum_{agr} y_{agr} \tag{10.60d}$$

$$\mu = \frac{\sum_{agr} y_{agr}}{\sum_{ag} \rho_{ag}} = \frac{\sum_{agr} y_{agr}}{\alpha \gamma p} = \bar{y}_{\bullet\bullet\bullet} \tag{10.60e}$$

Although the constraints are slightly different, the end result here is the same as that in the balanced ANOVA derivation. The term μ represents the mean of all signal ratios.

Equations for other parameters in the model can be calculated as follows. For the parameters A_a, the sum of squares is calculated as in **(10.16)**.

$$SSE_a = \sum_{gr} \varepsilon'^2_{agr} = \sum_{gr} \left(y_{agr} - \mu - A_a - G_g - AG_{ag} \right)^2 \quad \forall a \tag{10.61}$$

In order to minimize SSE_a, its derivative with respect to A_a is taken as follows.

$$\frac{\partial SSE_a}{\partial A_a} = -2\sum_{gr}\left(y_{agr} - \mu - A_a - G_g - AG_{ag}\right) = 0 \;\; \forall a \tag{10.62}$$

Expanding this summation and rearranging the equation produces

$$\sum_{gr} y_{agr} - \mu\sum_{g}\rho_{ag} - A_a\sum_{g}\rho_{ag} - \sum_{g}\left(G_g\rho_{ag}\right) - \sum_{g}\left(AG_{ag}\rho_{ag}\right) = 0 \;\; \forall a \tag{10.63a}$$

Applying the constraint in **(10.57)** produces

$$\sum_{gr} y_{agr} - \mu\sum_{g}\rho_{ag} - A_a\sum_{g}\rho_{ag} - \sum_{g}\left(G_g\rho_{ag}\right) = 0 \;\; \forall a \tag{10.63b}$$

This produces an equation for A_a and G_g that cannot be simplified any further.

$$A_a + \frac{\sum_{g}\left(G_g\rho_{ag}\right)}{\sum_{g}\rho_{ag}} = \bar{y}_{a\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \;\; \forall a \tag{10.64}$$

An analogous procedure can be carried out for G_g and AG_{ag}.

$$SSE_g = \sum_{ar}\varepsilon'^2_{agr} = \sum_{ar}\left(y_{agr} - \mu - A_a - G_g - AG_{ag}\right)^2 \;\; \forall g \tag{10.65}$$

$$\frac{\partial SSE_g}{\partial G_g} = -2\sum_{ar}\left(y_{agr} - \mu - A_a - G_g - AG_{ag}\right) = 0 \;\; \forall g \tag{10.66}$$

$$\sum_{ar} y_{agr} - \mu\sum_{a}\rho_{ag} - \sum_{a}\left(A_a\rho_{ag}\right) - G_g\sum_{a}\rho_{ag} - \sum_{a}\left(AG_{ag}\rho_{ag}\right) = 0 \;\; \forall g \tag{10.67a}$$

$$\sum_{ar} y_{agr} - \mu\sum_{a}\rho_{ag} - \sum_{a}\left(A_a\rho_{ag}\right) - G_g\sum_{a}\rho_{ag} = 0 \;\; \forall g \tag{10.67b}$$

$$\frac{\sum_{a}\left(A_a\rho_{ag}\right)}{\sum_{a}\rho_{ag}} + G_g = \bar{y}_{\bullet g\bullet} - \bar{y}_{\bullet\bullet\bullet} \;\; \forall g \tag{10.68}$$

(10.68) presents a second relationship, which relates A_a and G_g. Performing this procedure with the AG_{ag} terms reveals that no simplification can be made based on the constraints for this model equation and all terms must be kept.

$$SSE_{ag} = \sum_r \varepsilon'^2_{agr} = \sum_r \left(y_{agr} - \mu - A_a - G_g - AG_{ag}\right)^2 \quad \forall \, a, g \tag{10.69}$$

$$\frac{\partial SSE_{ag}}{\partial AG_{ag}} = -2\sum_r \left(y_{agr} - \mu - A_a - G_g - AG_{ag}\right) = 0 \quad \forall \, a, g \tag{10.70}$$

$$\sum_r y_{agr} - \mu\rho_{ag} - A_a\rho_{ag} - G_g\rho_{ag} - AG_{ag}\rho_{ag} = 0 \quad \forall \, a, g \tag{10.71}$$

$$AG_{ag} = \bar{y}_{ag\bullet} - A_a - G_g - \bar{y}_{\bullet\bullet\bullet} \quad \forall \, a, g \tag{10.72}$$

(10.72) shows that the parameters AG_{ag} can be calculated once the values of A_a and G_g are known. Once all of these parameter values have been determined, ε'_{agr} can be calculated as in **(10.25)**. To this point, the unbalanced ANOVA equations described here are equivalent to that presented in (Lindman 1992).

10.2.1.4 Solving the Model

The equations derived in Section 10.2.1.3 are summarized in Figure 10.5. This is only one of many possible ways of solving this system of equations. Based on this set of equations, it is apparent that equations for the parameters A_a and G_g must be solved simultaneously. One option for doing this would be to solve a $[(\alpha + \gamma_{Real}) \times (\alpha + \gamma_{Real})]$ matrix. This is the solution recommended by (Lindman 1992). Although this would be a sparse matrix, with γ_{Real} on the order of 4,000 for *E. coli*, it would have close to 16 million elements. However, **(10.64)** and **(10.68)** can be simplified further to reduce the size of this matrix calculation. With some algebra, these equations can be combined by substitution to produce a single equation involving A_a.

$$\mu = \bar{y}_{\bullet\bullet\bullet} \qquad (1)$$

$$A_a + \frac{\sum_g (G_g \rho_{ag})}{\sum_g \rho_{ag}} = \bar{y}_{a\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \quad \forall\, a = 1 \ldots \alpha - 1 \qquad (\alpha\text{-}1)$$

$$\sum_a \left(A_a \sum_g \rho_{ag} \right) = 0 \qquad (1)$$

$$\frac{\sum_a (A_a \rho_{ag})}{\sum_a \rho_{ag}} + G_g = \bar{y}_{\bullet g\bullet} - \bar{y}_{\bullet\bullet\bullet} \quad \forall\, g = 1 \ldots \gamma - 1 \qquad (\gamma_{Real}\text{-}1)$$

$$\sum_g \left(G_g \sum_a \rho_{ag} \right) = 0 \qquad (1)$$

$$AG_{ag} = \bar{y}_{ag\bullet} - A_a - G_g - \bar{y}_{\bullet\bullet\bullet} \quad \forall\, a = 1 \ldots \alpha - 1, g = 1 \ldots \gamma - 1 \qquad (\alpha\text{-}1)(\gamma_{Real}\text{-}1)$$

$$\sum_g (AG_{ag} \rho_{ag}) = 0 \quad \forall\, a = 1 \ldots \alpha - 1 \qquad (\alpha\text{-}1)$$

$$\sum_a (AG_{ag} \rho_{ag}) = 0 \quad \forall\, g \qquad (\gamma_{Real})$$

$$\varepsilon'_{agr} = y_{agr} - \mu - A_a - G_g - AG_{ag} \quad \forall\, a, g, r \qquad \left(\sum_{ag} \rho_{ag} \right)$$

Figure 10.5: Summary of Equations for Solving the Unbalanced Array-Gene ANOVA Model

Numbers in parentheses indicate the number of equations represented. For a balanced data set, these equations degenerate into the balanced equations displayed in Figure 10.1.

Rearranging **(10.68)** to solve for G_g results in the following.

$$G_g = \bar{y}_{\bullet g \bullet} - \bar{y}_{\bullet\bullet\bullet} - \frac{\sum_a \left(A_a \rho_{ag}\right)}{\sum_a \rho_{ag}} \quad \forall\, g = 1 \ldots \gamma - 1 \tag{10.73}$$

Applying the second constraint in **(10.55)** results in

$$G_\gamma = \frac{-\sum_{g=1}^{\gamma-1}\left(G_g \sum_a \rho_{ag}\right)}{\sum_a \rho_{a\gamma}} \tag{10.74a}$$

(10.74a) can be rewritten by applying a substitution from **(10.73)**.

$$G_\gamma = \frac{-\sum_{g=1}^{\gamma-1}\left[\bar{y}_{\bullet g \bullet}\sum_a \rho_{ag} - \bar{y}_{\bullet\bullet\bullet}\sum_a \rho_{ag} - \sum_a \left(A_a \rho_{ag}\right)\right]}{\sum_a \rho_{a\gamma}} \tag{10.74b}$$

At this point, it is necessary to distinguish array indices that are involved in summations (a) and those that are not ($\dot{a}$). **(10.64)** can also be rewritten by dividing the second term in two and substituting **(10.73)** and **(10.74b)** into each of these terms.

$$A_{\dot{a}} + \frac{\sum_{g=1}^{\gamma-1}\left(\rho_{\dot{a}g} G_g\right) + \rho_{\dot{a}\gamma} G_\gamma}{\sum_g \rho_{\dot{a}g}} = \bar{y}_{\dot{a}\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \quad \forall\, \dot{a} = 1 \ldots \alpha - 1 \tag{10.75a}$$

$$\begin{aligned} A_{\dot{a}} &+ \frac{\sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\left(\bar{y}_{\bullet g \bullet} - \bar{y}_{\bullet\bullet\bullet} - \frac{\sum_a \left(A_a \rho_{ag}\right)}{\sum_a \rho_{ag}}\right)\right]}{\sum_g \rho_{\dot{a}g}} \\ &- \frac{\rho_{\dot{a}\gamma}\sum_{g=1}^{\gamma-1}\left[\bar{y}_{\bullet g \bullet}\sum_a \rho_{ag} - \bar{y}_{\bullet\bullet\bullet}\sum_a \rho_{ag} - \sum_a \left(A_a \rho_{ag}\right)\right]}{\sum_a \rho_{a\gamma}\sum_g \rho_{\dot{a}g}} \\ &= \bar{y}_{\dot{a}\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \quad \forall\, \dot{a} = 1 \ldots \alpha - 1 \end{aligned} \tag{10.75b}$$

The result is an equation in which the G_g parameters have been eliminated. Expanding these summations results in

$$A_{\dot{a}} + \frac{\sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\,\bar{y}_{\bullet g\bullet}\right]}{\sum_{g}\rho_{\dot{a}g}} - \bar{y}_{\bullet\bullet\bullet}\frac{\sum_{g=1}^{\gamma-1}\rho_{\dot{a}g}}{\sum_{g}\rho_{\dot{a}g}} - \frac{\sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\left(\frac{\sum_{a}\left(A_a\rho_{ag}\right)}{\sum_{a}\rho_{ag}}\right)\right]}{\sum_{g}\rho_{\dot{a}g}}$$
$$- \rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\bar{y}_{\bullet g\bullet}\sum_{a}\rho_{ag}\right)}{\sum_{a}\rho_{a\gamma}\sum_{g}\rho_{\dot{a}g}} + \bar{y}_{\bullet\bullet\bullet}\rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{a}\rho_{ag}\right)}{\sum_{a}\rho_{a\gamma}\sum_{g}\rho_{\dot{a}g}} + \rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{a}\left(A_a\rho_{ag}\right)\right)}{\sum_{a}\rho_{a\gamma}\sum_{g}\rho_{\dot{a}g}} \quad \textbf{(10.75c)}$$
$$= \bar{y}_{\dot{a}\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \quad \forall\,\dot{a} = 1\ldots\alpha-1$$

Collecting terms involving A_a on the left and terms involving $\bar{y}_{a\bullet\bullet}$, $\bar{y}_{\bullet g\bullet}$, and $\bar{y}_{\bullet\bullet\bullet}$ and μ on the right produces

$$A_{\dot{a}} - \frac{\sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\left(\frac{\sum_{a}\left(A_a\rho_{ag}\right)}{\sum_{a}\rho_{ag}}\right)\right]}{\sum_{g}\rho_{\dot{a}g}} + \rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{a}\left(A_a\rho_{ag}\right)\right)}{\sum_{a}\rho_{a\gamma}\sum_{g}\rho_{\dot{a}g}}$$
$$= \bar{y}_{\dot{a}\bullet\bullet} - \frac{\sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\,\bar{y}_{\bullet g\bullet}\right]}{\sum_{g}\rho_{\dot{a}g}} + \rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\bar{y}_{\bullet g\bullet}\sum_{a}\rho_{ag}\right)}{\sum_{a}\rho_{a\gamma}\sum_{g}\rho_{\dot{a}g}} \quad \textbf{(10.75d)}$$
$$- \bar{y}_{\bullet\bullet\bullet} + \bar{y}_{\bullet\bullet\bullet}\frac{\sum_{g=1}^{\gamma-1}\rho_{\dot{a}g}}{\sum_{g}\rho_{\dot{a}g}} - \bar{y}_{\bullet\bullet\bullet}\rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{a}\rho_{ag}\right)}{\sum_{a}\rho_{a\gamma}\sum_{g}\rho_{\dot{a}g}} \quad \forall\,\dot{a} = 1\ldots\alpha-1$$

Simplifying this equation results in the following.

$$A_{\dot{a}} - \frac{\sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\left(\frac{\sum_{a}\left(A_a\rho_{ag}\right)}{\sum_{a}\rho_{ag}}\right)\right]}{\sum_{g}\rho_{\dot{a}g}} + \rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{a}\left(A_a\rho_{ag}\right)\right)}{\sum_{a}\rho_{a\gamma}\sum_{g}\rho_{\dot{a}g}}$$

$$= \bar{y}_{\dot{a}\bullet\bullet} - \frac{\sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\bar{y}_{\bullet g\bullet}\right]}{\sum_{g}\rho_{\dot{a}g}} + \rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{ar}y_{agr}\right)}{\sum_{a}\rho_{a\gamma}\sum_{g}\rho_{\dot{a}g}} \qquad \textbf{(10.75e)}$$

$$-\frac{\bar{y}_{\bullet\bullet\bullet}\rho_{\dot{a}\gamma}}{\sum_{g}\rho_{\dot{a}g}}\left[1+\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{a}\rho_{ag}\right)}{\sum_{a}\rho_{a\gamma}}\right] \quad \forall\, \dot{a}=1\ldots\alpha-1$$

Finally, multiplying all terms by $\sum_{g}\rho_{\dot{a}g}$ results in

$$A_{\dot{a}}\sum_{g}\rho_{\dot{a}g} - \sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\left(\frac{\sum_{a}\left(A_a\rho_{ag}\right)}{\sum_{a}\rho_{ag}}\right)\right] + \rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{a}\left(A_a\rho_{ag}\right)\right)}{\sum_{a}\rho_{a\gamma}}$$

$$= \sum_{gr}y_{\dot{a}gr} - \sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\bar{y}_{\bullet g\bullet}\right] + \rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{ar}y_{agr}\right)}{\sum_{a}\rho_{a\gamma}} \qquad \textbf{(10.75f)}$$

$$-\bar{y}_{\bullet\bullet\bullet}\rho_{\dot{a}\gamma}\left[1+\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{a}\rho_{ag}\right)}{\sum_{a}\rho_{a\gamma}}\right] \quad \forall\, \dot{a}=1\ldots\alpha-1$$

Based on this equation, the following definitions can be made:

$$J_{\dot{a}a} \equiv -\sum_{g=1}^{\gamma-1}\left(\frac{\rho_{\dot{a}g}\rho_{ag}}{\sum_{a}\rho_{ag}}\right)+\rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\rho_{ag}}{\sum_{a}\rho_{a\gamma}} \quad \forall\,\dot{a}=1\ldots\alpha-1, a=1\ldots\alpha \tag{10.76}$$

$$\begin{aligned} K_{\dot{a}} \equiv & \sum_{gr} y_{\dot{a}gr} - \sum_{g=1}^{\gamma-1}\left[\rho_{\dot{a}g}\bar{y}_{\bullet g\bullet}\right]+\rho_{\dot{a}\gamma}\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{ar} y_{agr}\right)}{\sum_{a}\rho_{a\gamma}} \\ & -\bar{y}_{\bullet\bullet\bullet}\rho_{\dot{a}\gamma}\left[1+\frac{\sum_{g=1}^{\gamma-1}\left(\sum_{a}\rho_{ag}\right)}{\sum_{a}\rho_{a\gamma}}\right] \quad \forall\,\dot{a}=1\ldots\alpha-1 \end{aligned} \tag{10.77}$$

To calculate A_a, **(10.75f)** is applied (α-1) times, and the first constraint in **(10.55)** is applied once. Thus, a dense ($\alpha \times \alpha$) matrix, as shown in Figure 10.6, is generated. This matrix calculation has been reduced from 16 million elements to 64, in the case of the induction validation experiment in Section 4.7. Next, the G_g parameters can be calculated using **(10.73)** and **(10.74b)**. The equations **(10.70)** and **(10.25)** can be used to solve for the remaining parameters. The solution to the unbalanced data set in Table 10.1B is displayed in Figure 10.7. As discussed in Section 10.2.3, this is only an intermediate solution, and although all of the constraints and equations in Figure 10.5 are satisfied, this is not the final solution for the unbalanced data set. Notice that for this solution, the sums of these parameters are not necessarily zero; however, the *weighted sums* are zero. For example, in the case of the G_g parameters, (1.4667 + 1 – 1.9667) ≠ 0. However, when these values are weighted by the corresponding $\sum_{a}\rho_{ag}$ values, the weighted sum becomes [8 · (1.4667) + 4 · (1) + 8 · (-1.9667)] = 0. Another difference between this solution and the solution in Figure 10.2 is that no parameters are calculated for conditions where no data are available (*e.g.* AG_{62}). In addition, ε'_{agr} is only relevant for conditions with multiple measurements.

$$
\begin{array}{c|c}
a & \begin{array}{cccc} 1 & \cdots & (\alpha-1) & \alpha \end{array} \\
\hline
\dot{a} & \\
1 & \left[\begin{array}{cccc} \sum_g \rho_{1g} + J_{11} & \cdots & J_{1(\alpha-1)} & J_{1\alpha} \end{array}\right. \left| \begin{array}{c} K_1 \end{array}\right. \\
\vdots & \left[\begin{array}{cccc} \vdots & \ddots & \vdots & \vdots \end{array}\right. \left| \begin{array}{c} \vdots \end{array}\right. \\
(\alpha-1) & \left[\begin{array}{cccc} J_{(\alpha-1)1} & \cdots & \sum_g \rho_{(\alpha-1)g} + J_{(\alpha-1)(\alpha-1)} & J_{(\alpha-1)\alpha} \end{array}\right. \left| \begin{array}{c} K_{(\alpha-1)} \end{array}\right. \\
\alpha & \left[\begin{array}{cccc} \sum_g \rho_{1g} & \cdots & \sum_g \rho_{(\alpha-1)g} & \sum_g \rho_{\alpha g} \end{array}\right. \left| \begin{array}{c} 0 \end{array}\right.
\end{array}
$$

Figure 10.6: Matrix for Calculation of A_a Parameters

μ

-2

G_g

g	
1	1.4667
2	1
3	-1.9667

A_a

a	
1	-3.6333
2	0.1333
4	-1.1333
5	2.6333
6	2

AG_{ag}

g \ a	1	2	4	5	6
1	3.6667	-0.1	-3.3333	-2.1	2.0333
2	0.1333	-0.1333	-2.8667	2.8667	
3	-1.9	0.3333	3.1	1.3333	-2.0333

ε'_{agr}

g	r \ a	1	2	4	5	6
1	1		-1		0	-1.5
	2		1		0	1.5
2	1					
3	1	0.5		-0.5		-1
	2	-0.5		0.5		1

Figure 10.7: Intermediate Solution to the Array-Gene ANOVA Model for the Example Unbalanced Data Set

10.2.2 The Unbalanced Block-Treatment ANOVA Model

The Block-Treatment ANOVA Model given in **(10.29)** is the same for both the balanced and unbalanced cases. However, its solution differs between these two cases. These differences are discussed here.

10.2.2.1 Model Constraints

Although the mean of the individual AG_{ag} values alone may not be zero, when each value is weighted by the number of measurements, the mean becomes zero. This is simply a restatement of the constraint in **(10.56)**. Therefore, the Block-Treatment Model contains no mean term, because the weighted mean is zero.

The constraints for this model begin with a restatement of the constraints in **(10.57)**.

$$\sum_{g}\left(AG_{btgn}\rho_{btgn}\right)=0 \quad \forall b,t,n, \quad \sum_{btn}\left(AG_{btgn}\rho_{btgn}\right)=0 \quad \forall g \tag{10.78}$$

The second of these constraints can be expanded to produce the following.

$$\sum_{btn}\left(AG_{btgn}\rho_{btgn}\right)=\sum_{btn}\left(BG_{bg}\rho_{btgn}+TG_{tg}\rho_{btgn}+\varepsilon''_{btgn}\rho_{btgn}\right)=0 \quad \forall g \tag{10.79}$$

Each of these terms becomes a separate constraint equation.

$$\sum_{btn}\left(BG_{bg}\rho_{btgn}\right)=0 \quad \forall g, \quad \sum_{btn}\left(TG_{tg}\rho_{btgn}\right)=0 \quad \forall g, \quad \sum_{btn}\left(\varepsilon''_{btgn}\rho_{btgn}\right)=0 \quad \forall g \tag{10.80}$$

The first two of these constraints can be easily simplified since BG_{bg} is independent of both t and n, and TG_{tg} is independent of both b and n.

$$\sum_{b}\left(BG_{bg}\sum_{tn}\rho_{btgn}\right)=0 \ \forall g, \quad \sum_{t}\left(TG_{tg}\sum_{bn}\rho_{btgn}\right)=0 \ \forall g \tag{10.81}$$

The last constraint in **(10.80)** is expanded to eliminate any block-gene- or treatment-gene-dependent variation, since these should be accounted for by the BG_{bg} and TG_{tg} terms.

$$\sum_{bn}\left(\varepsilon''_{btgn}\rho_{btgn}\right)=0 \quad \forall t,g, \quad \sum_{tn}\left(\varepsilon''_{btgn}\rho_{btgn}\right)=0 \quad \forall b,g \tag{10.82}$$

The first constraint in **(10.78)** can also be expanded to

$$\sum_{g}\left(AG_{btgn}\rho_{btgn}\right)=\sum_{g}\left(BG_{bg}\rho_{btgn}+TG_{tg}\rho_{btgn}+\varepsilon''_{btgn}\rho_{btgn}\right)=0 \quad \forall b,t,n \tag{10.83}$$

Again, each of the terms in this equation can be considered to be separate constraints, as follows.

$$\begin{aligned}&\sum_{g}\left(BG_{bg}\rho_{btgn}\right)=0 \quad \forall b,t,n, \quad \sum_{g}\left(TG_{tg}\rho_{btgn}\right)=0 \quad \forall b,t,n,\\ &\sum_{g}\left(\varepsilon''_{btgn}\rho_{btgn}\right)=0 \quad \forall b,t,n\end{aligned} \tag{10.84}$$

These constraints cannot be simplified any further.

For reasons described in the following section, the first constraint of **(10.78)** does not hold for the final solution of unbalanced data sets. Therefore, the constraints derived from it, in **(10.84)**, are also invalid. In order to ensure a valid solution to the Unbalanced Block-Treatment ANOVA Model, a new set of constraints must be introduced to replace the first two in **(10.84)**

$$\sum_{g}\left(BG_{bg}\sum_{tn}\rho_{btgn}\right)=0 \quad \forall b, \quad \sum_{g}\left(TG_{tg}\sum_{bn}\rho_{btgn}\right)=0 \quad \forall t \tag{10.85}$$

These constraints are more stringent than those they are replacing and are analogous to those in **(10.33)** from the Balanced Block-Treatment ANOVA Model. The third constraint in **(10.84)** is not required for solving the model; therefore it is removed and is not replaced. Although we would like this constraint to hold for the final solution, it cannot be applied here.

10.2.2.2 Degree-of-Freedom Analysis

A degree-of-freedom analysis must also be performed on the Block-Treatment ANOVA Model to determine whether it is valid. For any given gene, there are $\sum_{a}\zeta_{ag}$ AG_{btgn} values, which are constrained by the following equation from **(10.79)**.

$$\sum_{btn}\left(AG_{btgn}\rho_{btgn}\right)=0 \quad \forall g \tag{10.86}$$

For a given gene, this constraint occupies one degree of freedom; therefore, only $\sum_{a}\zeta_{ag}-1$ degrees of freedom remain. Due to the constraints in **(10.81)**, there are (β_g – 1) DOF for the β_g BG_{bg} terms and (τ_g – 1) DOF for the τ_g TG_{tg} terms. As before, the residual degrees-of-freedom (DOF_{R2g}) is calculated by subtracting the DOF for all other parameters from the total DOF.

$$DOF_{R2g}=\left(\sum_{a}\zeta_{ag}-1\right)-\left[\left(\beta_g-1\right)+\left(\tau_g-1\right)\right]=\sum_{a}\zeta_{ag}-\beta_g-\tau_g+1 \tag{10.87}$$

Table 10.8 summarizes the results of this DOF analysis for the general case as well as for each gene of the Table 10.1A data. According to the DOF analysis, $DOF_{R2g} > 0$ when $\sum_{a}\zeta_{ag}>\beta_g+\tau_g-1$. Therefore, the Block-Treatment ANOVA Model is valid only for genes that meet this criterion. The validity of this condition must be checked for every gene in the data set. Any genes that do not pass should be eliminated. Upon elimination, the new data set should be reanalyzed with the Array-Gene ANOVA Model.

Table 10.8: Degree-of-Freedom Analysis for the Unbalanced Block-Treatent ANOVA Model

This table lists the model parameters from the ANOVA model in **(10.29)**. The DOF for a single gene in an unbalanced data set and the DOF for each gene in Table 10.1A are given. **(10.29)** is a valid model for a given gene when $\sum_a \zeta_{ag} > \beta_g + \tau_g - 1$, since the residual degrees of freedom would be greater than zero.

Parameter (Source of Variation)	Degrees of Freedom (DOF)	DOF for Example Unbalanced Data Set		
		$g = 1$	$g = 2$	$g = 3$
BG_{bg}	β_g-1	1	1	1
TG_{tg}	τ_g-1	2	1	2
ε''_{btgn}	$\sum_a \zeta_{ag} - \beta_g - \tau_g + 1$	1	1	1
AG_{btgn}	$\sum_a \zeta_{ag} - 1$	4	3	4

The same DOF analysis can also be performed on the system as a whole to determine the overall residual degrees of freedom (DOF_{R2}). This value can be determined by summing all of the DOF_{R2g} values and subtracting the value of DOF_{R2g} from a single gene to account for the third constraint in **(10.84)**. The value of DOF_{R2g} for a single gene can be calculated as $(\alpha - \beta - \tau + 1)$.

$$\begin{aligned} DOF_{R2} &= \sum_g DOF_{R2g} - (\alpha - \beta - \tau - + 1) \\ &= \sum_g \left(\sum_a \zeta_{ag} - \beta_g - \tau_g + 1 \right) - (\alpha - \beta - \tau + 1) \\ &= \sum_{ag} \zeta_{ag} - \sum_g \beta_g - \sum_g \tau_g + \gamma_{Real} - \alpha + \beta + \tau - 1 \end{aligned} \qquad \textbf{(10.88)}$$

This residual calculation shows that the Block-Treatment ANOVA Model is valid for the data set as a whole only when $\sum_{ag} \zeta_{ag} + \gamma_{Real} + \beta + \tau > \sum_g \beta_g + \sum_g \tau_g + \alpha + 1$. This DOF analysis is summarized in Table 10.9.

Table 10.9: Degree-of-Freedom Analysis for All Genes of the Unbalanced Block-Treatent ANOVA Model

This table lists the model parameters from the ANOVA model in **(10.29)**. The DOF for a general unbalanced data set and the DOF for the data set in Table 10.1A are given. **(10.29)** is a valid model when $\sum_{ag}\zeta_{ag}+\gamma_{Real}+\beta+\tau>\sum_{g}\beta_{g}+\sum_{g}\tau_{g}+\alpha+1$, since the residual degrees of freedom would be greater than zero.

Parameter (Source of Variation)	Degrees of Freedom (DOF)	DOF for Example Unbalanced Data Set
BG_{bg}	$\sum_{g}\beta_{g}-\beta-\gamma_{Real}+1$	2
TG_{tg}	$\sum_{g}\tau_{g}-\tau-\gamma_{Real}+1$	3
ε''_{btgn}	$\sum_{ag}\zeta_{ag}-\sum_{g}\beta_{g}-\sum_{g}\tau_{g}+\gamma_{Real}-\alpha+\beta+\tau-1$	2
AG_{btgn}	$\sum_{ag}\zeta_{ag}-\alpha-\gamma_{Real}+1$	7

10.2.2.3 Determination of Model Parameters

The BG_{bg} parameter values are calculated by minimizing the sum of the squares of the residual error for the Block-Treatment ANOVA Model. First, SSE_{bg} is calculated as follows

$$SSE_{bg}=\sum_{tn}\varepsilon''^{2}_{btgn}=\sum_{tn}\left(AG_{btgn}-BG_{bg}-TG_{tg}\right)^{2}\quad\forall b,g \tag{10.89}$$

Next, the derivative of this value with respect to BG_{bg} is calculated.

$$\frac{\partial SSE_{bg}}{\partial BG_{bg}}=-2\sum_{tn}\left(AG_{btgn}-BG_{bg}-TG_{tg}\right)=0\quad\forall b,g \tag{10.90}$$

Expanding the summation to all of these terms results in

$$\sum_{tn}\left(AG_{btgn}\rho_{btgn}\right)-BG_{bg}\sum_{tn}\rho_{btgn}-\sum_{tn}\left(TG_{tg}\rho_{btgn}\right)=0\quad\forall b,g \tag{10.91}$$

Rearranging this equation produces

$$BG_{bg}+\frac{\sum_{tn}\left(TG_{tg}\rho_{btgn}\right)}{\sum_{tn}\rho_{btgn}}=\frac{\sum_{tn}\left(AG_{btgn}\rho_{btgn}\right)}{\sum_{tn}\rho_{btgn}}\quad\forall b,g \tag{10.92}$$

Following the same protocol to determine a relationship for TG_{tg} produces the following.

$$SSE_{tg} = \sum_{bn} \varepsilon''^{2}_{btgn} = \sum_{bn} \left(AG_{btgn} - BG_{bg} - TG_{tg}\right)^2 \quad \forall t, g \tag{10.93}$$

$$\frac{\partial SSE_{tg}}{\partial TG_{tg}} = -2\sum_{bn} \left(AG_{btgn} - BG_{bg} - TG_{tg}\right) = 0 \quad \forall t, g \tag{10.94}$$

$$\sum_{bn} \left(AG_{btgn}\rho_{btgn}\right) - \sum_{bn} \left(BG_{bg}\rho_{btgn}\right) - TG_{tg}\sum_{bn} \rho_{btgn} = 0 \quad \forall t, g \tag{10.95}$$

$$\frac{\sum_{bn} \left(BG_{bg}\rho_{btgn}\right)}{\sum_{bn} \rho_{btgn}} + TG_{tg} = \frac{\sum_{bn} \left(AG_{btgn}\rho_{btgn}\right)}{\sum_{bn} \rho_{btgn}} \quad \forall t, g \tag{10.96}$$

The residual error is calculated as described in (**10.47**).

10.2.2.4 Solving the Model

These equations in Figure 10.8 can be applied to each gene individually to simplify the calculations. However, like the Unbalanced Array-Gene ANOVA Model, equations (**10.92**) and (**10.96**) must be solved simultaneously. The same substitutions applied to derive (**10.75f**) can also be applied to solve these two equations. Instead of one large matrix calculation, this model requires many small matrix calculations. A solution to the Block-Treatment ANOVA Model for the unbalanced data set in Table 10.1B is shown in Figure 10.9. For reasons discussed in Section 10.2.3, this solution is not the final solution to the overall Unbalanced ANOVA Model.

10.2.3 Solving the Unbalanced ANOVA Model

While the solution to the unbalanced ANOVA model presented in Figure 10.7 and Figure 10.9 appears to meet all of the necessary constraints, this solution is inadequate. The columns of ε''_{btgn} have weighted sums that are nonzero. In other words, the following constraint does not hold.

$$\sum_{g} \left(\varepsilon''_{btgn}\rho_{btgn}\right) = 0 \quad \forall b, t, n \tag{10.97}$$

This is the third constraint in (**10.84**), which was never applied to solve the model. In the solution to the balanced ANOVA model, an analogous constraint was derived (**10.33**), and was

found to hold in the solution in Figure 10.4. Therefore, something unique to the unbalanced data set caused this constraint to fail.

The problems originate from the second constraint on AG_{ag} in **(10.57)**. While this constraint is necessary for solving the model, it has no value in the final solution. Since AG_{ag} is an intermediate value that does not appear in the final model **(10.49)**, it makes no difference whether this constraint holds or not. In fact, forcing this constraint to hold for an unbalanced data set results in violation of the constraint in **(10.97)**.

$$BG_{bg} + \frac{\sum_{tn}\left(TG_{tg}\rho_{btgn}\right)}{\sum_{tn}\rho_{btgn}} = \frac{\sum_{tn}\left(AG_{btgn}\rho_{btgn}\right)}{\sum_{tn}\rho_{btgn}} \quad \forall \begin{matrix} b=1\ldots\beta-1, \\ g=1\ldots\gamma-1 \end{matrix} \quad \begin{pmatrix} \sum_{g}\beta_g - \gamma_{Real} \\ -\beta+1 \end{pmatrix}$$

$$\sum_{b}\left(BG_{bg}\sum_{tn}\rho_{btgn}\right)=0 \quad \forall g=1\ldots\gamma-1 \qquad (\gamma_{Real}\text{-}1)$$

$$\sum_{g}\left(BG_{bg}\sum_{tn}\rho_{btgn}\right)=0 \quad \forall b \qquad (\beta)$$

$$\frac{\sum_{bn}\left(BG_{bg}\rho_{btgn}\right)}{\sum_{bn}\rho_{btgn}} + TG_{tg} = \frac{\sum_{bn}\left(AG_{btgn}\rho_{btgn}\right)}{\sum_{bn}\rho_{btgn}} \quad \forall \begin{matrix} t=1\ldots\tau-1, \\ g=1\ldots\gamma-1 \end{matrix} \quad \begin{pmatrix} \sum_{g}\tau_g - \gamma_{Real} \\ -\tau+1 \end{pmatrix}$$

$$\sum_{t}\left(TG_{tg}\sum_{bn}\rho_{btgn}\right)=0 \quad \forall g=1\ldots\gamma-1 \qquad (\gamma_{Real}\text{-}1)$$

$$\sum_{g}\left(TG_{tg}\sum_{bn}\rho_{btgn}\right)=0 \quad \forall t \qquad (\tau)$$

$$\varepsilon''_{btgn} = AG_{btgn} - BG_{bg} - TG_{tg} \quad \forall b,t,g,n \qquad \left(\sum_{ag}\zeta_{ag}\right)$$

Figure 10.8: Summary of Equations for the Block-Treatment ANOVA Model
Numbers in parentheses indicate the number of equations represented. For a balanced data set, these equations degenerate into the balanced equations displayed in Figure 10.1.

BG_{bg}

g \ b	1	2
1	2.2917	-1.3750
2	0	0
3	-2.2917	1.3750

TG_{tg}

g \ t	1	2	3
1	-0.2917	-1.5583	3.4083
2	-1.3667	1.3667	
3	0.8292	1.75	-3.4083

ε''_{btgn}

	a	1	2	4	5	6
	b	1	1	2	2	2
	t	1	2	1	2	3
g						
1		1.6667	-0.8333	-1.6667	0.8333	0
2		1.5	-1.5	-1.5	1.5	
3		-0.4375	0.8750	0.8958	-1.7917	0

Figure 10.9: Intermediate Solution to the Block-Treatment ANOVA Model for the Example Unbalanced Data Set

The solution to this problem is to perform an iterative procedure, which essentially redistributes the unwanted variation in the ε''_{btgn} parameters to all of the other parameters, including AG_{ag}. The first step is to solve the equations in Figure 10.5 and Figure 10.8 to obtain solutions like those in Figure 10.7 and Figure 10.9. The output values of ε''_{btgn} are then treated as input values of y_{btgr} for the next iteration. The same equations are applied to ε''_{btgn}. At the end of each iteration, the resulting parameter values (A_a, G_g, BG_{bg}, and TG_{tg}) are added to the values from the previous iteration. It should be noted that, after the first iteration, the values of μ and ε'_{agr} are always zero; therefore, these values do not change. This procedure is repeated until the values of ε''_{btgn} meet the constraint in **(10.97)**, within some tolerance. At this point, the accumulated values of the calculated parameters are the final values.

With the final values of μ, A_a, G_g, ε'_{agr}, BG_{bg}, TG_{tg}, and ε''_{btgn} determined, the final values of AG_{ag} can also be calculated using **(10.72)**. Although not necessary, these values demonstrate that the second constraint of **(10.57)** has been sacrificed so that the constraint in **(10.97)** can be met. The final values to both the Array-Gene Model and the Block-Treatment Model for the unbalanced data set in Table 10.1B is displayed in Figure 10.10.

μ

-2

G_g

g	
1	1.5
2	1
3	-2

A_a

a	
1	-3
2	-0.5
4	-1.5
5	3
6	2

AG_{ag}

g \ a	1	2	4	5	6
1	3	0.5	-3	-2.5	2
2	-0.5	0.5	-2.5	2.5	
3	-2.5	1	3.5	1	-2

ε'_{agr}

g	r \ a	1	2	4	5	6
1	1		-1		0	-1.5
	2		1		0	1.5
2	1					
3	1	0.5		-0.5		-1
	2	-0.5		0.5		1

BG_{bg}

g \ b	1	2
1	2.5	-1.5
2	0	0
3	-2.5	1.5

TG_{tg}

g \ r	1	2	3
1	-0.5	-1.5	3.5
2	-1.5	1.5	
3	1	1.5	-3.5

ε''_{btgn}

g	a	1	2	4	5	6
	b	1	1	2	2	2
	r	1	2	1	2	3
1		1	-0.5	-1	0.5	0
2		1	-1	-1	1	
3		-1	2	1	-2	0

Figure 10.10: Final Solution to the Array-Gene ANOVA Model and the Block-Treatment ANOVA Model for the Example Unbalanced Data Set

Using the calculated values of A_a, the original signal ratios can be normalized exactly as described in Section 10.1.2. At this point, it is also possible to apply Model 3, another two-way ANOVA model described by **(10.30)**. However, since the A_a terms are eliminated during the normalization step, this model would add nothing to the analysis. The identification of differentially expressed genes uses only the residual error values from Model 1, ε'_{agr}, and Model 2, ε''_{btgn}.

10.2.4 Identifying Differentially Expressed Genes Using Unbalanced ANOVA Models

Expression values for unbalanced data are calculated in a manner similar to that applied for balanced data sets.

$$\bar{\hat{y}}_{\bullet tg\bullet\bullet} = \frac{\sum_{brn} \hat{y}_{btgrn}}{\sum_{brn} \rho_{btgrn}} = \mu_g + TG_{tg} \quad \forall t, g \tag{10.98}$$

The calculated expression values for the example unbalanced data set are displayed in Table 10.10.

Table 10.10: Expression Values for Example Balanced Data Set

$\bar{\hat{y}}_{\bullet tg\bullet\bullet}$

g \ t	1	2	3
1	-0.5	-1.5	3.5
2	-2	-0.5	-6.5
3	-3.5	-4	-3

Log ratio values were calculated and the same significance tests described in Section 4.5.3 were performed. Within two genes (1 - *genA* and 3 - *genC*), two comparisons were found to be significant ($t = 2$ vs. 3 for *genA* and $t = 1$ vs. 3 for *genC*).

For the purposes of comparison, the balanced equation developed in Section 10.1 can be applied to the unbalanced data set to provide an estimate of gene expression. The resulting parameter values and expression values are shown in Figure 10.11 and Table 10.11. Although the parameter values are similar to those calculated from the unbalanced ANOVA model, the differences can cause significantly different conclusions to be drawn from the data set. For example, none of the genes were found to have significant change when this balanced estimate was applied to the unbalanced data set.

μ

-2

G_g

g	
1	2.0625
2	0.5
3	-2.3125

A_a

a	
1	-4
2	0.63
4	-1.5
5	3.13
6	1.75

AG_{ag}

g \ a	1	2	4	5	6
1	3.4375	-1.1875	-3.5625	-3.1875	1.6875
2	1	-0.125	-2	2.875	
3	-1.1875	0.1875	3.8125	1.1875	-1.4375

BG_{bg}

g \ b	1	2
1	1.1250	-0.7875
2	0.4375	0.1250
3	-1.5	0.7875

ε'_{agr}

g	r \ a	1	2	4	5	6
1	1		-1		0	-1.5
	2		1		0	1.5
2	1					
3	1	0.5		-0.5		-1
	2	-0.5		0.5		1

TG_{tg}

g \ t	1	2	3
1	-0.0625	-2.1875	1.6875
2	-0.5	1.375	
3	1.3125	0.6875	-1.4375

ε''_{btgn}

g	a	b	r				
	1	1	1	2	4	5	6
	1	1	2	2	2		
	1	2	1	2	3		
1				2.375	-0.125	-2.7125	-0.2125	0.7875
2				1.0625	-1.9375	-1.625	1.375	
3				-1	1	1.7125	-0.2875	-0.7875

Figure 10.11: Final Solution to the Array-Gene ANOVA Model and the Block-Treatment ANOVA Model for the Example Unbalanced Data Set

Table 10.11: Expression Values for Example Balanced Data Set

$\bar{\hat{y}}_{\bullet tg\bullet\bullet}$

g \ t	1	2	3
1 - *genA*	0	-2.125	1.75
2 - *genB*	-2	-0.125	
3 - *genC*	-3	-3.625	-5.75

10.3 Summary of ANOVA Model

Data sets produced by DNA microarrays are far from perfect. Different genes may be represented by a different number of spots. In addition, these data sets often include missing data. Unbalanced data sets are unavoidable when working with DNA microarrays. ANOVA models can be applied to perform both normalization as well as identification of differentially

expressed genes. It is possible to apply a balanced ANOVA to an unbalanced data set in order to estimate the model parameters. However, only an unbalanced solution will provide an exact answer. Unbalanced solutions typically require global matrix calculations in order to simultaneously determine all of the parameter values. Such solutions are not feasible for the large data sets generated by DNA microarrays. The novel solution applied here divided a three-way ANOVA model into three two-way ANOVA models. This approach allowed much smaller matrix calculations to be performed, but required an iterative solution. Overall, this algorithm is more efficient than the standard global matrix calculations.

11 Bibliography

Affymetrix® (2004). Array Manufacturing. http://www.affymetrix.com/technology/manufacturing/index.affx, Affymetrix®.

Albano, C. R., L. Randers-Eichhorn, W. E. Bentley, and G. Rao (1998). "Detection of oxidative stress induction using green fluorescent protein promoter probes." Free Radical Bio Med **25**(Suppl. 1): S121.

Alexeeva, S., K. J. Hellingwerf, and M. J. Teixeira de Mattos (2003). "Requirement of ArcA for redox regulation in *Escherichia coli* under microaerobic but not anaerobic or aerobic conditions." J Bacteriol **185**(1): 204-209.

Alland, D., I. Kramnik, T. R. Weisbrod, L. Otsubo, R. Cerny, L. P. Miller, W. R. Jacobs, Jr., and B. R. Bloom (1998). "Identification of differentially expressed mRNA in prokaryotic organisms by customized amplification libraries (DECAL): The effect of isoniazid on gene expression in *Mycobacterium tuberculosis*." P Natl Acad Sci USA **95**(22): 13227-13232.

Amersham Pharmacia Biotech (2000). Product Specification: FluoroLink™ Cy3-dUTP PA 53022 & FluoroLink™ Cy5-dUTP PA 55022.

Åslund, F., M. Zheng, J. Beckwith, and G. Storz (1999). "Regulation of the OxyR transcription factor by hydrogen peroxide and the cellular thiol-disulfide status." P Natl Acad Sci USA **96**(11): 6161-6165.

Atlung, T. and L. Brøndsted (1994). "Role of the transcriptional activator AppY in regulation of the *cyx appA* operon of *Escherichia coli* by anaerobiosis, phosphate starvation, and growth phase." J Bacteriol **176**(17): 5414-5422.

Ausubel, F., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl, Eds. (1995). Short Protocols in Molecular Biology. Current Protocols in Molecular Biology, John Wiley & Sons, Inc.

Baggerly, K. A., K. R. Coombes, K. R. Hess, D. N. Stivers, L. V. Abruzzo, and W. Zhang (2001). "Identifying differentially expressed genes in cDNA microarray experiments." J Comput Biol **8**(6): 639-659.

Baldi, P. and A. D. Long (2001). "A Bayesian framework for the analysis of microarray expression data: Regularized *t*-test and statistical inferences of gene changes." Bioinformatics **17**(6): 509-519.

Barbosa, T. M. and S. B. Levy (2000). "Differential expression of over 60 chromosomal genes in *Escherichia coli* by constitutive expression of MarA." J Bacteriol **182**(12): 3467-3474.

Barron, A., J. U. Jung, and M. Villarejo (1987). "Purification and characterization of a glycine betaine binding protein from *Escherichia coli*." J Biol Chem **262**(24): 11841-11846.

Bayer Healthcare LLC (2004). Alpha-1 Antitrypsin Deficiency. http://www.bayerbiologicals.com/Products/Therapeutic/Proteinase/Alpha-1.asp, Bayer Healthcare LLC.

Beatty, K., P. Robertie, R. M. Senior, and J. Travis (1982). "Determination of oxidized α1-proteinase inhibitor in serum." J Lab Clin Med **100**(2): 186-192.

Begley, T. P., J. Xi, C. Kinsland, S. Taylor, and F. McLafferty (1999). "The enzymology of sulfur activation during thiamin and biotin biosynthesis." Curr Opin Chem Biol **3**(5): 623-629.

Benov, L. and I. Fridovich (1998). "Growth in iron-enriched medium partially compensates *Escherichia coli* for the lack of manganese and iron superoxide dismutase." J Biol Chem **273**(17): 10313-10316.

Blattner, F. R., G. Plunkett, III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao (1997). "The complete genome sequence of *Escherichia coli* K-12." Science **277**(5331): 1453-1474.

Boehme, D. E., K. Vincent, and O. R. Brown (1976). "Oxygen and toxicity: Inhibition of amino acid biosynthesis." Nature **262**(5567): 418-420.

Box, G. E. P., W. G. Hunter, and J. S. Hunter (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. New York NY, John Wiley & Sons.

Bui, B. T. S., D. Florentin, F. Fournier, O. Ploux, A. Méjean, and A. Marquet (1998). "Biotin synthase mechanism: On the origin of sulphur." FEBS Lett **440**(1-2): 226-230.

Bukau, B. (1993). "Regulation of the *Escherichia coli* heat-shock response." Mol Microbiol **9**(4): 671-680.

Bylund, F., E. Collet, S.-O. Enfors, and G. Larsson (1998). "Substrate gradient formation in the large-scale bioreactor lowers cell yield and increases by-product formation." Bioprocess Eng **18**(3): 171-180.

Carilli, A., E. B. Chain, G. Gualandi, and G. Morisi (1961). "Aeration studies III. Continuous measurement of dissolved oxygen during fermentation in large fermenters." Sci Repts Ist Super Sanita **1**: 177-189.

Chee, M., R. Yang, E. Hubbell, A. Berno, X. C. Huang, D. Stern, J. Winkler, D. J. Lockhart, M. S. Morris, and S. P. A. Fodor (1996). "Accessing genetic information with high-density DNA arrays." Science **274**(5287): 610-614.

Chen, Y., E. R. Dougherty, and M. L. Bittner (1997). "Ratio-based decisions and the quantitative analysis of cDNA microarray images." J Biomed Opt **2**(4): 364-374.

Chen, Y., V. Kamat, E. R. Dougherty, M. L. Bittner, P. S. Meltzer, and J. M. Trent (2002). "Ratio statistics of gene expression levels and apllications to microarray data analysis." Bioinformatics **18**(9): 1207-1215.

Chuang, S.-E. and F. R. Blattner (1993). "Characterization of twenty-six new heat shock genes of *Escherichia coli*." J Bacteriol **175**(16): 5242-5252.

Cui, X. and G. A. Churchill (2003). "Statistical tests for differential expression in cDNA microarray experiments." Genome Biol **4**(4): Art. No. 210.

Davies, K. J. A., S. W. Lin, and R. E. Pacifici (1987). "Protein damage and degradation by oxygen radicals IV. Degradation of denatured proteins." J Biol Chem **262**(20): 9914-9920.

De Freitas, J. M., A. Liba, R. Meneghini, J. S. Valentine, and E. B. Gralla (2000). "Yeast lacking Cu-Zn superoxide dismutase show altered iron homeostasis." J Biol Chem **275**(16): 11645-11649.

de Saizieu, A., U. Certa, J. Warrington, C. Gray, W. Keck, and J. Mous (1998). "Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays." Nat Biotechnol **16**(1): 45-48.

DeLisa, M. P., C.-F. Wu, L. Wang, J. J. Valdes, and W. E. Bentley (2001). "DNA microarray-base identification of genes controlled by autoinducer 2-stimulated quorum sensing in *Escherichia coli*." J Bacteriol **183**(18): 5239-5247.

Denu, J. M. and K. G. Tanner (1998). "Specific and reversible inactivation of protein tyrosine phosphatases by hydrogen peroxide: Evidence for a sulfenic acid intermediate and implications for redox regulation." Biochemistry-US **37**(16): 5633-5642.

DeRisi, J., L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent (1996). "Use of a cDNA microarray to analyse gene expression patterns in human cancer." Nat Genet **14**(4): 457-460.

DeRisi, J. L., V. R. Iyer, and P. O. Brown (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-686.

Ding, H. and B. Demple (1997). "*In vivo* kinetics of a redox-regulated transcriptional switch." P Natl Acad Sci USA **94**(16): 8445-8449.

Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments." Stat Sinica **12**(1): 111-139.

Earhart, C. F. (1996). "Uptake and Metabolism of Iron and Molybdenum." *Escherichia coli and Salmonella typhimurium*. 2nd Edition, F. C. Neidhardt, Ed. Washington, D.C., American Society for Microbiology Press: 1075-1090.

Eisen, M. B. and P. O. Brown (1999). "DNA arrays for analysis of gene expression." Method Enzymol **303**: 179-205.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998). "Cluster analysis and display of genome-wide expression patterns." P Natl Acad Sci USA **95**(25): 14863-14868.

Eisenberg, M. A., O. Prakash, and S.-C. Hsiung (1982). "Purification and properties of the biotin repressor." J Biol Chem **257**(24): 15167-15173.

Ekaza, E., J. Teyssier, S. Ouahrani-Bettache, J.-P. Liautard, and S. Köhler (2001). "Characterization of *Brucells suis clpB* and *clpAB* mutants and participation of the genes in stress reponses." J Bacteriol **183**(8): 2677-2681.

Evans, M. D. and W. A. Pryor (1994). "Cigarette smoking, emphysema, and damage to α_1-proteinase inhibitor." Am J Physiol **266**(6): L593-L611.

Farr, S. B. and T. Kogoma (1991). "Oxidative stress responses in *Escherichia coli* and *Salmonella typhimurium*." Microbiological Reviews **55**(4): 561-585.

Fislage, R., M. Berceanu, Y. Humboldt, M. Wendt, and H. Oberender (1997). "Primer design for a prokaryotic differential display." Nucleic Acids Res **25**(9): 1830-1835.

Flint, D. H. (1996). "*Escherichia coli* contains a protein that is homologous in function and N-terminal sequence to the protein encoded by the *nifS* gene of *Azotobacter vinelandii* and that can participate in the synthesis of the Fe-S cluster of dihydroxy-acid dehydratase." J Biol Chem **271**(27): 16068-16074.

Flint, D. H., E. Smyk-Randall, J. F. Tuminello, B. Draczynska-Lusiak, and O. R. Brown (1993). "The inactivation of dihydroxy-acid dehydratase in *Escherichia coli* treated with hyperbaric oxygen occurs because of the destruction of its Fe-S cluster, but the enzyme remains in the cell in a form that can be reactivated." J Biol Chem **268**(34): 25547-25552.

Friden, P., t. Newman, and M. Freundlich (1982). "Nucleotide sequence of the *ilvB* promoter-regulatory region: A biosynthetic operon controlled by attentuation and cyclic AMP." P Natl Acad Sci USA **79**(20): 6156-6160.

Gasteiger, E., A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, and A. Bairoch (2003). "ExPASy: The proteomics server for in-depth protein knowledge and analysis." Nucleic Acids Res **31**(13): 3784-3788.

Gaudu, P., N. Moon, and B. Weiss (1997). "Regulation of the *soxRS* oxidative stress regulon: Reversible oxidation of the Fe-S centers of SoxR *in vivo*." J Biol Chem **272**(8): 5082-5086.

Gaudu, P. and B. Weiss (1996). "SoxR, a [2Fe-2S] transcription factor, is active only in its oxidized form." P Natl Acad Sci USA **93**(19): 10094-10098.

Gennis, R. B. and V. Stewart (1996). "Respiration." *Escherichia coli* and *Salmonella typhimurium*. 2nd Edition, F. C. Neidhardt, Ed. Washington, D.C., American Society for Microbiology Press: 217-261.

George, S., G. Larsson, K. Olsson, and S.-O. Enfors (1998). "Comparison of the baker's yeast process performance in laboratory and production scale." Bioprocess Eng **18**(2): 135-142.

Georgellis, D., O. Kwon, and E. C. C. Lin (2001). "Quinones as the redox signal for the Arc two-component system of bacteria." Science **292**(5525): 2314-2316.

Gibson, K. J., D. A. Pelletier, and I. M. Turner, Sr. (1999). "Transfer of sulfur to biotin from biotin synthase (BioB protein)." Biochem Bioph Res Co **254**(3): 632-635.

Golby, P., D. J. Kelly, J. R. Guest, and S. C. Andrews (1998). "Transcriptional regulation and organization of the *dcuA* and *dcuB* genes encoding homologous anaerobic C4-dicarboxylate transporters in *Escherichia coli*." J Bacteriol **180**(24): 6586-6596.

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

Green, J., M. F. Anjum, and J. R. Guest (1996). "The *ndh*-binding protein (Nbp) regulates the *ndh* gene of *Escherichia coli* in response to growth phase and is identical to Fis." Mol Microbiol **20**(5): 1043-1055.

Griffiths, S. W. (2002). Oxidation of the Sulfur-Containing Amino Acids in Recombinant Human α1-Antitrypsin. Ph.D. Thesis, Department of Chemical Engineering. Cambridge, MA, Massachusetts Institute of Technology: 190.

Griffiths, S. W. and C. L. Cooney (2002). "Development of a peptide mapping procedure to identify and quantify methionine oxidation in recombinant human α1-antitrypsin." J Chromatogr A **942**(1-2): 133-143.

Gross, C. A. (1996). "Function and Regulation of the Heat Shock Proteins." *Escherichia coli* and *Salmonella typhimurium*. 2nd Edition, F. C. Neidhardt, Ed. Washington, D.C., American Society for Microbiology Press: 1389-1399.

Grossman, A. D. (1984). Studies on the Function and Regulation of Two Sigma Subunits of *Escherichia coli* RNA Polymerase. Ph.D. Thesis, Department of Molecular Biology. Madison, WI, University of Wisconsin.

Grossman, A. D., D. B. Straus, W. A. Walter, and C. A. Gross (1987). "σ^{32} synthesis can regulate the synthesis of heat shock protein in *Escherichia coli*." Gene Dev **1**(2): 179-184.

Hahm, J.-i. and C. M. Lieber (2004). "Direct ultrasensitive electrical detection of DNA and DNA sequence variations using nanowire nanosensors." Nano Lett **4**(1): 51-54.

Hantke, K. (2001). "Iron and metal regulation in bacteria." Current Opinions in Microbiology **4**(2): 172-177.

Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. E. Hughes, E. Snesrud, N. Lee, and J. Quackenbush (2000). "A concise guide to cDNA microarray analysis." BioTechniques **29**(3): 548-562.

Herman, C., D. Thévenet, R. D'Ari, and P. Bouloc (1995). "Degradation of σ^{32}, the heat shock regulator in *Escherichia coli*, is governed by HflB." P Natl Acad Sci USA **92**(8): 3516-3520.

Hoff, K. G., J. J. Silberg, and L. E. Vickery (2000). "Interaction of the iron-sulfur cluster assembly protein IscU with the Hsc66/Hsc20 molecular chaperone system of *Escherichia coli*." P Natl Acad Sci USA **97**(14): 7790-7795.

Hoskins, J. R., S. Sharma, B. K. Sathyanarayana, and S. Wickner (2002). "Clp ATPases and thier role in protein unfolding and degradation." Adv Protein Chem **59**: 413-429.

Huang, H.-C., M. Y. Sherman, O. Kandror, and A. L. Goldberg (2001). "The molecular chaperon DnaJ is required for the degradation of a soluble abnormal protein in *Escherichia coli*." J Biol Chem **276**(6): 3920-3928.

Ideker, T., T. Galitski, and L. Hood (2001). "A new approach to decoding life: Systems biology." Annu Rev Genom Hum G **2**: 343-372.

Ilbert, M., V. Méjean, M.-T. Giudici-Oritconi, J.-P. Samama, and C. Iobbi-Nivol (2003). "Involvement of a mate chaperone (TorD) in the maturation pathway of molybdoenzyme TorA." J Biol Chem **278**(31): 28787-28792.

Imlay, J. A. (2002). "How oxygen damages microbes: Oxygen tolerance and obligate anaerobiosis." Adv Microb Physiol **46**: 111-153.

Imlay, J. A. and S. Linn (1986). "Bimodal pattern of killing of DNA-repair-defective or anoxically grown *Escherichia coli* by hydrogen peroxide." J Bacteriol **166**(2): 519-527.

Imlay, J. A. and S. Linn (1987). "Mutagenesis and stress responses induced in *Escherichia coli* by hydrogen peroxide." J Bacteriol **169**(7): 2967-2976.

Imlay, J. A. and S. Linn (1988). "DNA damage and oxygen radical toxicity." Science **240**(4857): 1302-1309.

Jameson, G., M. Cosper, H. Hernandez, M. Johnson, and B. Huynh (2004). "Role of the [2Fe-2S] cluster in recombinant *Escherichia coli* biotin synthase." Biochemistry-US **43**(7): 2022-2031.

Jennings, M. P. and I. R. Beacham (1993). "Co-dependent positive regulation of the *ansB* promoter of *Escherichia coli* by CRP and the FNR protein: A molecular analysis." Mol Microbiol **9**(1): 155-164.

Johnson, D. and J. Travis (1979). "The oxidative inactivation of human α-1-proteinase inhibitor." J Biol Chem **254**(10): 4022-4026.

Jordan, P. A., A. J. Thomson, E. T. Ralph, J. R. Guest, and J. Green (1997). "FNR is a direct oxygen sensor having a biphasic response curve." FEBS Lett **416**(3): 349-352.

Jung, I. L. and I. G. Kim (2003). "Thiamine protects against paraquat-induced damage: scavenging activity of reactive oxygen species." Environ Toxicol Phar **15**(1): 19-26.

Kadner, R. J. (1974). "Transport systems for L-methionine in *Escherichia coli*." J Bacteriol **117**(1): 232-241.

Kakuta, Y., T. Horio, Y. Takahashi, and K. Fukuyama (2001). "Crystal structure of *Escherichia coli* Fdx, an adrenodoxin-type ferredoxin involved in the assembly of iron-sulfur clusters." Biochemistry-US **40**(37): 11007-11012.

Kane, M. D., T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore (2000). "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays." Nucleic Acids Res **28**(22): 4552-4557.

Kanemori, M., H. Mori, and T. Yura (1994). "Induction of heat shock proteins by abnormal proteins results from stabilization and not increased synthesis of σ^{32} in *Escherichia coli*." J Bacteriol **176**(18): 5648-5653.

Karp, P. D., M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S. M. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro (2002). "The EcoCyc Database." Nucleic Acids Res **30**(1): 56-58.

Kerr, M. K., M. Martin, and G. A. Churchill (2000). "Analysis of variance for gene expression microarray data." J Comput Biol 7(6): 819-837.

Keyer, K., A. S. Gort, and J. A. Imlay (1995). "Superoxide and the production of oxidative DNA damage." J Bacteriol **177**(23): 6782-6790.

Keyer, K. and J. A. Imlay (1996). "Superoxide accelerates DNA damage by elevating free-iron levels." P Natl Acad Sci USA **93**(24): 13635-13640.

Khoroshilova, N., C. Popescu, E. Münck, H. Beinert, and P. J. Kiley (1997). "Iron-sulfur cluster disassembly in the FNR protein of *Escherichia coli* by O_2: [4Fe-4S] to [2Fe-2S] conversion with loss of biological activity." P Natl Acad Sci USA **94**(12): 6087-6092.

Kim, H., B. Zhao, E. C. Snesrud, B. J. Haas, C. D. Town, and J. Quackenbush (2002). "Use of RNA and genomic DNA references for inferred comparisons in DNA microarray analyses." BioTechniques **33**(4): 924-930.

King, J. and U. K. Laemmli (1971). "Polypeptides of the tail fibres of bacteriophage T4." J Mol Biol **62**(3): 465-477.

Konz, J. O. (1998). Oxidative Damage to Recombinant Proteins During Production. Ph.D. Thesis, Department of Chemical Engineering. Cambridge, MA, Massachusetts Institute of Technology: 232.

Kuo, C. F., T. Mashino, and I. Fridovich (1987). "α,β-dihydroxyisovalerate dehydratase - a superoxide-sensitivie enzyme." J Biol Chem **262**(10): 4724-4727.

Kwon, O., D. Georgellis, and E. C. C. Lin (2000). "Phosphorelay as the sole physiological route of signal transmission by the Arc two-component system of *Escherichia coli*." J Bacteriol **182**(13): 3858-3862.

Larsson, G., M. Tornkvist, E. S. Wernersson, C. Tragardh, H. Noorman, and S.-O. Enfors (1996). "Substrate gradients in bioreactors: Origins and consequences." Bioprocess Eng **14**(6): 281-289.

Laska, M. E. (2000). The Effect of Dissolved Oxygen on Recombinant Protein Degradation in *Escherichia coli*. Ph.D. Thesis, Department of Chemical Engineering. Cambridge, MA, Massachusetts Institute of Technology: 276.

Lauhon, C. T. and R. Kambampati (2000). "The *iscS* gene in *Escherichia coli* is required for the biosynthesis of 4-thiouridine, thiamin, and NAD." J Biol Chem **275**(26): 20096-20103.

Lee, K. N., H. S. Shin, K.-S. Kwon, S. D. Park, and M.-H. Yu (1993). "Molecular properties of recombinant human α1-antitrypsin produced in *Escherichia coli* and *in vitro* translation system." Mol Cells **3**: 71-74.

Lee, M.-L. T., F. C. Kuo, G. A. Whitmore, and J. Sklar (2000). "Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations." P Natl Acad Sci USA **97**(18): 9834-9839.

Leimkühler, S. and K. V. Rajagopalan (2001). "A sulfurtransferase is required in the transfer of cysteine sulfur in the *in vitro* synthesis of molybdopterin from precursor Z in *Escherichia coli*." J Biol Chem **276**(25): 22024-22031.

Lesley, S. A., J. Graziano, C. Y. Cho, M. W. Knuth, and H. E. Klock (2002). "Gene expression response to misfolded protein as a screen for soluble recombinant protein." Protein Eng **15**(2): 153-160.

Levin, D. E., M. Hollstein, M. F. Christman, E. A. Schwiers, and B. N. Ames (1982). "A new *Salmonella* tester strain (TA102) with A·T base pairs at the site of mutation detects oxidative mutagens." P Natl Acad Sci USA **79**(23): 7445-7449.

Lind, C., R. Gerdes, I. Schuppe-Koistinen, and I. A. Cotgreave (1998). "Studies on the mechanism of oxidative modification of human glyceraldehyde-3-phosphate dehydrogenase by glutathione: Catalysis by glutaredoxin." Biochem Bioph Res Co **247**(2): 481-486.

Lindman, H. R. (1992). Analysis of Variance in Experimental Design. New York, NY, Springer-Verlag.

Liochev, S. I. and I. Fridovich (1994). "The role of $O2^{\cdot -}$ in the production of HO· in vitro and in vivo." Free Radical Biology & Medicine **16**(1): 29-33.

Liu, X. and P. D. Wulf (2004). "Probing the ArcA-P modulon of *Escherichia coli* by whole genome transcriptional analysis and sequence recognition profiling." J Biol Chem **279**(13): 12588-12597.

Loiseau, L., S. Ollagnier-de-Choudens, L. Nachin, M. Fontecave, and F. Barras (2003). "Biogenesis of Fe-S cluster by bacterial Suf system: SufS and SufE form a new type of cysteine desulfurase." J Biol Chem **278**(40): 38352-38359.

Long, A. D., H. J. Mangalam, B. Y. P. Chan, L. Tolleri, G. W. Hatfield, and P. Baldi (2001). "Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework." J Biol Chem **276**(23): 19937-19944.

Loos, A., C. Glanemann, L. B. Willis, X. M. O'Brien, P. A. Lessard, R. Gerstmeir, S. Guillouet, and A. J. Sinskey (2001). "Development and validation of *Corynebacterium* DNA microarrays." Appl Environ Microb **67**(5): 2310-2318.

Manfredini, R., V. Cavallera, L. Marini, and G. Donati (1983). "Mixing and oxygen transfer in conventional stirred fermentors." Biotechnol Bioeng **25**(12): 3115-3131.

Martin, R. G. and J. L. Rosner (2002). "Genomics of the *marA/soxS/rob* regulon of *Escherichia coli*: Identification of directly activated promoters by application of molecular genetics and informatics to microarray data." Mol Microbiol **44**(6): 1611-1624.

McCormick, M. L., G. R. Buettner, and B. E. Britigan (1998). "Endogenous superoxide dismutase levels regulate iron-dependent hydroxyl radical formation in *Escherichia coli* exposed to hydrogen peroxide." J Bacteriol **180**(3): 622-625.

McHugh, J. P., F. Rodriguez-Quinones, H. Abdul-Tehrani, D. A. Svistunenko, R. K. Poole, C. E. Cooper, and S. C. Andrews (2003). "Global iron-dependent gene regulation in *Escherichia coli*." J Biol Chem **278**(32): 29478-29486.

Milton, J. S. and J. C. Arnold (1995). Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences. New York, NY, McGraw-Hill, Inc.

Morita, M., M. Kanemori, H. Yanagi, and T. Yura (1999). "Heat-induced synthesis of σ^{32} in *Escherichia coli*: Structural and functional dissection of *rpoH* mRNA secondary structure." J Bacteriol **181**(2): 401-410.

Mueller, E. G., P. M. Palenchar, and C. J. Buck (2001). "The role of the cysteine residues in ThiI in the generation of 4-thiouridine in tRNA." J Biol Chem **276**(36): 33588-33595.

Neher, S. B., R. T. Sauer, and T. A. Baker (2003). "Distinct peptide signals in the UmuD and UmuD' subunits of UmuD/D' mediate tethering and substrate processing by the ClpXP protease." P Natl Acad Sci USA **100**(23): 13129-13224.

Neubauer, P., M. Ahman, M. Tornkvist, G. Larsson, and S.-O. Enfors (1995). "Response of guanosine tetraphosphate to glucose fluctuations in fed-batch cultivations of *Escherichia coli*." J Biotechnol **43**(3): 195-204.

Neubauer, P., H. Y. Lin, and B. Mathiszik (2003). "Metabolic load of recombinant protein production: Inhibition of cellular capacities for glucose uptake and respiration after induction of a heterologous gene in *Escherichia coli*." Biotechnol Bioeng **83**(1): 53-64.

Ollagnier-de Choudens, S., L. Nachin, Y. Sanakis, L. Loiseau, F. Barras, and M. Fontecave (2003). "SufA from *Erwinia chrysanthemi*." J Biol Chem **278**(20): 17993-18001.

Oosterhuis, N. M. G. and N. W. F. Kossen (1984). "Dissolved oxygen concentration profiles in a production-scale bioreactor." Biotechnol Bioeng **26**(5): 546-550.

Outten, F. W., M. J. Wood, F. M. Muñoz, and G. Storz (2003). "The SufE protein and the SufBCD complex enhance SufS cysteine desulfurase activity as part of a sulfur transfer pathway for Fe-S cluster assembly in *Escherichia coli*." J Biol Chem **278**(14): 45713-45719.

Parsell, D. A. and R. T. Sauer (1989). "Induction of a heat shock-like response by unfolded protein in *Escherichia coli*: Dependence on protein level not protein degradation." Gene Dev **3**(8): 1226-1232.

Pickett, S., S. Carriedo, and C. Wang (2001). Determining the signal-to-noise ratio and optimal photomultiplier gain setting in the GenePix 4000B. www.axon.com/genomics/SNR_and_PMT_Gain.pdf, Axon Instruments, Inc.

Pierre, J. L. and M. Fontecave (1999). "Iron and activated oxygen species in biology: The basic chemistry." BioMetals **12**(3): 195-199.

Plum, G. and J. E. Clark-Curtiss (1994). "Induction of *Mycobacterium avium* gene expression following phagocytosis by human macrophages." Infect Immun **62**(2): 476-486.

Pollack, J. R., C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown (1999). "Genome-wide analysis of DNA copy-number changes using cDNA microarrays." Nat Genet **23**(1): 41-46.

Pomposiello, P. J., M. H. J. Bennik, and B. Demple (2001). "Genome-wide transcriptional profiling of the *Escherichia coli* responses to superoxide stress and sodium salicylate." J Bacteriol **183**(13): 3890-3902.

Porello, S. L., M. J. Cannon, and S. S. David (1998). "A substrate recognition role for the [4Fe-4S]$^{2+}$ cluster of the DNA repair glycosylase MutY." Biochemistry-US **37**(18): 6465-6475.

Rice, C. W. and W. P. Hempfling (1978). "Oxygen-limited continuous culture in respiratory energy conservation in *Escherichia coli*." J Bacteriol **134**(1): 115-124.

Richmond, C. S., J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner (1999). "Genome-wide expression profiling in *Escherichia coli* K-12." Nucleic Acids Res **27**(19): 3821-3835.

Riesenberg, D., K. Menzel, V. Schulz, K. Schumann, G. Veith, G. Zuber, and W. A. Knorre (1990). "High cell density fermentation of recombinant *Escherichia coli* expressing human interferon alpha 1." Appl Microbiol Biot **34**(1): 77-82.

Rohlin, L., M.-K. Oh, and J. C. Liao (2002). "DNA microarray for microbial biotechnology: Gene expression profiles in *Escherichia coli* during protein overexpression." J Chin Inst Chem Eng **33**(1): 103-112.

Rosenow, C., R. M. Saxena, M. Durst, and T. R. Gingeras (2001). "Prokaryotic RNA preparation methods useful for high density array analysis: Comparison of two approaches." Nucleic Acids Res **29**(22): e112.

Rouquette, C., M.-T. Ripio, E. Pellegrini, J.-M. Bolla, R. I. Tascon, J.-A. Vázquez-Boland, and P. Berche (1996). "Identification of a ClpC ATPase required for stress tolerance and *in vivo* survival of *Listeria monocytogenes*." Mol Microbiol **21**(5): 977-987.

Salmon, K., S.-p. Hung, K. Mekjian, P. Baldi, G. W. Hatfield, and R. P. Gunsalus (2003). "Global gene expression profiling in *Escherichia coli* K12." J Biol Chem **278**(32): 29837-29855.

Saurin, W., M. Hofnung, and E. Dassa (1999). "Getting in or out: Early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters." J Mol Evol **48**(1): 22-41.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Schweder, T., E. Kruger, B. Xu, B. Jurgen, G. Blomsten, S.-O. Enfors, and M. Hecker (1999). "Monitoring of genes that respond to process-related stress in large-scale bioprocesses." Biotechnol Bioeng **65**(2): 151-159.

Sherman, M. Y. and A. L. Goldberg (1992). "Involvement of the chaperonin dnaK in the rapid degradation of a mutant protein in *Escherichia coli*." The EMBO Journal **11**(1): 71-77.

Spehr, V., A. Schlitt, D. Scheide, V. Guénebaut, and T. Friedrich (1999). "Overexpression of the *Escherichia coli nuo*-operon and isolation of the overproduced NADH: ubiquinone oxidoreductase (Complex I)." Biochemistry-US **38**(49): 16261-16267.

Srinivasan, C., A. Liba, J. A. Imlay, J. S. Valentine, and E. B. Gralla (2000). "Yeast lacking superoxide dismutase(s) show elevated levels of "free iron" as measured by whole cell electron paramagnetic resonance." J Biol Chem **275**(38): 29187-29192.

Stadtman, E. R. and M. E. Wittenberger (1985). "Inactivation of *Escherichia coli* glutamine synthetase by xanthine oxidase, nicotinate hydroxylase, horseradish peroxidase, or glucose oxidase: effects of ferredoxin, putidaredoxin, and menadione." Arch Biochem Biophys **239**(2): 379-387.

Steel, R. and W. D. Maxon (1966). "Dissolved oxygen measurement in pilot- and production-scale novobiocin fermentations." Biotechnol Bioeng **8**: 97-108.

Straus, D., W. Walter, and C. A. Gross (1990). "DnaK, DnaJ, and GrpE heat shock proteins negatively regulate heat shock gene expression by controlling the synthesis and stability of σ^{32}." Gene Dev **4**(12A): 2202-2209.

Straus, D. B., W. A. Walter, and C. A. Gross (1987). "The heat shock response of *Escherichia coli* is regulated by changes in the concentration of σ^{32}." Nature **329**(6137): 348-351.

Sweere, A. P. J., L. Janse, K. C. A. M. Luyben, and N. W. F. Kossen (1988). "Experimental simulation of oxygen profiles and their influence on baker's yeast production: II. Two-fermentor system." Biotechnol Bioeng **31**(6): 579-586.

Sweere, A. P. J., J. R. Mesters, L. Janse, K. C. A. M. Luyben, and N. W. F. Kossen (1988). "Experimental simulation of oxygen profiles and their influence on baker's yeast production: I. One-fermentor system." Biotechnol Bioeng **31**(6): 567-578.

Taggart, C., D. Cervantes-Laurean, G. Kim, N. G. McElvaney, N. Wehr, J. Moss, and R. L. Levine (2000). "Oxidation of either Methionine 351 or Methionine 358 in α1-antitrypsin causes loss of anti-neutrophil elastase activity." J Biol Chem **275**(35): 27258-27265.

Takahashi, Y. and M. Nakamura (1999). "Functional assignment of the ORF2-*iscS-iscU-iscA-hscB-hscA-fdx*-ORF3 gene cluster involved in the assembly of Fe-S clusters in *Escherichia coli*." J Biochem **126**(5): 917-926.

Taniguchi, M., K. Miura, H. Iwao, and S. Yamanaka (2001). "Quantitative assessment of DNA microarrays-Comparison with Northern blot analyses." Genomics **71**(1): 34-39.

Tao, H., C. Bausch, C. Richmond, F. R. Blattner, and T. Conway (1999). "Functional genomics: Expression analysis of *Escherichia coli* growing on minimal and rich media." J Bacteriol **181**(20): 6425-6440.

Tatsuta, T., T. Tomoyasu, B. Bukau, M. Kitagawa, H. Mori, K. Karata, and T. Ogura (1998). "Heat shock regulation in the ftsH null mutant of *Escherichia coli*: Dissection of stability and activity control mechanisms of σ^{32} *in vivo*." Mol Microbiol **30**(3): 583-593.

Tseng, C.-P., J. Albrecht, and R. P. Gunsalus (1996). "Effect of microaerophilic cell growth conditions on expression of the aerobic (*cyoABCDE* and *cydAB*) and anaerobic (*narGHJI*, *frdABCD*, and *dmsABC*) respiratory pathway genes in *Escherichia coli*." J Bacteriol **178**(4): 1094-1098.

Tseng, G. C., M.-K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong (2001). "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects." Nucleic Acids Res **29**(12): 2549-2557.

Tu, Y., G. Stolovitzky, and U. Klein (2002). "Quantitative noise analysis for gene expression microarray experiments." P Natl Acad Sci USA **99**(22): 14031-14036.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). "Significance analysis of microarrays applied to the ionizing radiation response." P Natl Acad Sci USA **98**(9): 5116-5121.

Ugulava, N. B., B. R. Gibney, and J. T. Jarrett (2001). "Biotin synthase contains two distinct iron-sulfur cluster binding sites: Chemical and spectroelectrochemical analysis of iron-sulfur cluster interconversions." Biochemistry-US **40**(28): 8343-8351.

Ugulava, N. B., K. K. Surerus, and J. T. Jarrett (2002). "Evidence from Mössbauer spectroscopy for distinct $[2Fe\text{-}2S]^{2+}$ and $[4Fe\text{-}4S]^{2+}$ cluster binding sites in biotin synthase from *Escherichia coli*." J Am Chem Soc **124**(31): 9050-9051.

VanBogelen, R. A., P. M. Kelley, and F. C. Neidhardt (1987). "Differential induction of heat shock, SOS, and oxidation stress regulons and accumulation of nucleotides in *Escherichia coli*." J Bacteriol **169**(1): 26-32.

Vardar, F. and M. D. Lilly (1982). "Effect of cycling dissolved oxygen concentrations on product formation in penicillin fermentations." Eur J Appl Microbiol **14**(4): 203-211.

Varghese, S., Y. Tang, and J. A. Imlay (2003). "Contrasting sensitivities of *Escherichia coli* aconitases A and B to oxidation and iron depletion." J Bacteriol **185**(1): 221-230.

Voet, D. and J. G. Voet (1995). Biochemistry. New York, NY, John Wiley & Sons, Inc.

Wei, Y., J.-M. Lee, C. Richmond, F. R. Blattner, J. A. Rafalski, and R. A. LaRossa (2000). "High-density microarray-mediated gene expression profiling of *Escherichia coli*." J Bacteriol **183**(2): 545-556.

Wendisch, V. F., D. P. Zimmer, A. Khodursky, B. Peter, N. Cozzarelli, and S. Kustu (2001). "Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays." Anal Biochem **290**(2): 205-213.

Wild, J., W. A. Walter, C. A. Gross, and E. Altman (1993). "Accumulation of secretory protein precursors in *Escherichia coli* induces the heat shcok response." J Bacteriol **175**(13): 3992-3997.

Winkler, S. A. (1995). Development of a Fermentation Process for the Production of Recombinant Heparinase I in *Eschericia coli*. Master's Thesis, Department of Chemical Engineering. Cambridge, MA, Massachusetts Institute of Technology: 108.

Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules (2001). "Assessing gene significance from cDNA microarray expression data via mixed models." J Comput Biol **8**(6): 625-637.

Xu, B., M. Jahic, G. Blomsten, and S.-O. Enfors (1999). "Glucose overflow metabolism and mixed-acid fermentation in aerobic large-scale fed-batch processes with *Escherichia coli*." Appl Microbiol Biot **51**(5): 564-571.

Yang, L., R. T. Lin, and E. B. Newman (2002). "Structure of the Lrp-regulated *serA* promoter of *Escherichia coli* K-12." Mol Microbiol **43**(2): 323-333.

Yang, M. C. K., Q. G. Ruan, J. J. Yang, S. Eckenrode, S. Wu, R. A. McIndoe, and J. X. She (2001). "A statistical procedure for flagging weak spots greatly improves normalization and ratio estimates in microarray experiments." Physiological Genomics **7**(1): 45-53.

Yang, Y. H., S. Dudoit, P. Luu, and T. P. Speed (2001). Normalization for cDNA Microarray Data. SPIE BiOS, San Jose, California.

Yegneswaran, P. K., M. R. Gray, and B. G. Thompson (1991). "Experimental simulation of dissolved oxygen fluctuations in large fermentors: Effect on *Streptomyces clavuligerus*." Biotechnol Bioeng **38**(10): 1203-1209.

Yura, T. and K. Nakahigashi (1999). "Regulation of the heat-shock response." Curr Opin Microbiol **2**(2): 153-158.

Zheng, M., F. Åslund, and G. Storz (1998). "Activation of the OxyR transcription factor by reversible disulfide bond formation." Science **279**(5357): 1718-1721.

Zheng, M., B. Doan, T. D. Schneider, and G. Storz (1999). "OxyR and SoxRS regulation of *fur*." J Bacteriol **181**(15): 4639-4643.

Zheng, M., X. Wang, B. Doan, K. A. Lewis, T. D. Schneider, and G. Storz (2001). "Computation-directed identification of OxyR DNA binding sites in *Escherichia coli*." J Bacteriol **183**(15): 4571-4579.

Zheng, M., X. Wang, L. J. Templeton, D. R. Smulski, R. A. LaRossa, and G. Storz (2001). "DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide." J Bacteriol **183**(15): 4562-4570.