# FUNCTIONAL PROTEOMICS IN *Escherichia coli*

A Dissertation

by

MATTHEW MAURICE CHAMPION

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2005

Major Subject Biochemistry

# FUNCTIONAL PROTEOMICS IN *Escherichia coli*

A Dissertation

by

MATTHEW MAURICE CHAMPION

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

| | |
|---|---|
| Chair of Committee, | James C. Hu |
| Committee Members, | Deborah A. Siegele |
| | Ryland Young |
| | David H. Russell |
| | Donald A. Pettigrew |
| Head of Department | Gregory D. Reinhart |

December 2005

Major Subject Biochemistry

# ABSTRACT

Functional Proteomics in *Escherichia coli.* (December 2005)

Matthew Maurice Champion, B.S., The University of Iowa

Chair of Advisory Committee: Dr. James C. Hu

Cells respond to their environment with programmed changes in gene expression. Cataloging these changes at the protein level is key towards understanding the physiology of an organism. Multi-subunit and multi-protein complexes are also important and pathogenic and physiologic processes. In order to identify expressed proteins and potential protein complexes, we utilized a combination of non-denaturing chromatography and peptide mass fingerprinting. This approach allows us to identify the components of protein mixtures, as well as information lost in traditional proteomics, such as subunit associations. Applying this methodology to cells at both mid-exponential and stationary phase growth conditions, we identified several thousand proteins from each cell-state of *E. coli* corresponding to hundreds of unique gene products. The co-purification of proteins when fractionated at varying pHs could suggest the components of higher order complexes. This non-denaturing proteomic approach should provide physiological data unavailable by other means. The components of several known cellular complexes were also evident in this analysis. To characterize proteins associated with nucleic acid binding, we also performed proteome analysis on log and stationary phase cells grown in LB separated over heparin chromatography at neutral pH, which enriches for these proteins. The complete analysis of these identifications is discussed.

# DEDICATION

To my Wife, Patty

# ACKNOWLEDGMENTS

First I would like to extend a debt of gratitude to my advisor, Jim Hu. You have been an outstanding and challenging mentor and friend throughout my graduate career. I have come to appreciate the wonder that is the unique combination of science you practice and teach. Thank you for the instruction in a mixture of classical approaches with modern technologies. Thank you for continuing to challenge myself and others to push the idea, without forgetting the experiments that led us to this path. Thank you for entertaining my typically crazy hypotheses, and for arguing with me to make sense and digest our results. The best thing I can say to express my thanks is to describe that over my training, I have realized that an advisor is neither supervisor, nor professor, but something different and synergistic. People often look back upon the number of instructors they have had in their years that changed the very nature of how they saw a problem. This number is typically very small. It is a distinct honor to have had several professors here who make this a possibility for me and none more so than Jim. Above all, thank you for this opportunity and for education.

I also thank the many people who made this work possible: All of the members of the Hu Lab, in particular Chris Campbell for his work on many of the early experiments, and Gwen Knapp for being a friend and intellectual ear. I thank my close friends, Peter McCormick, Pablo Sobrado and Sam Perkins for many discussions and arguments about science and life. I would like to thank my committee members, Dr's Ry Young and Don Pettigrew for their time and assistance with various matters of my education and Debby Siegele, for her near-constant core-dump of *E. coli* into my brain and overall advice on the life of the germ. Dr. David Russell, (referred to often as just DR) has been absolutely

amazing for his willingness to let me into his lab essentially unannounced, and provided virtually unlimited support, both technical and material. No portion of these experiments would have been possible without the incredible access afforded to me by Dave and the members of the lab. I felt a part of the lab at every stage, and I hoped I was able to make a small contribution to the academic environment in his research group. In particular, several members of Dr. Russell's research group provided me most of the training and knowledge I have about biological mass spectroscopy. Zee-Yong Park provided most of my initial training on the MALDI MS systems, and introduced me to the basics of peak extraction and database searching. Bill Russell, in particular, was always available with a different perspective, troubleshooting, or general scientific discussion. His insight was invaluable in generating and improving my data. Many people and classmates provided assistance throughout this process, including Tony Reeves, Stephen Kopytek, and Leonardo Mariño.

The experiments in Chapter II were performed with assistance. Chris Campbell provided protein ID and fractional analysis, Dr. Siegele provided strains and microarray data, and Dr. Russell provided MS training, facilities and consumables. Experimental design, LC separations, MALDI-MS, peak picking, and bioinformatics were performed by me. Annotation of the 2D gels was performed by me with assistance from James C. Hu and Dr. Siegele. In Chapter III, the work was supported by the laboratories of Drs. Siegele, Russell and Hu, the bulk of the work was performed by me. Extensive analysis and bioinformatics resources were provided by Lili Niu, who was instrumental in designing and setting up the database and bioinformatics systems by which most of the queries in this data were performed. Most of the analysis in this chapter would not be possible without her contribution. The results of these and continuing experiments are available on the internet at http://eep.tamu.edu. The initial work in Chapter IV, analysis, pilot experiments and separations were developed by me. I also performed the non-tandem MS identifications, one-dimensional separations and most of the analysis. Additional separation and cell-growth work in this chapter was performed by Lili Niu, in the laboratory of Dr. J. Hu. MS and MS/MS analysis was performed by Dr. S. Perkins, and the analysis of the 2D gels were greatly facilitated by the patience and careful eye of Dr. Siegele. All relevant bioinformatics and queries were performed by L. Niu, and feedback was provided by me.

I thank my family and 'non-science' friends for their love and support, and my wife, Patty for her patience and for pushing me. My work, life and nature of the understanding of the scientific method are forever altered because of all of you. Thank you for giving me free range over our ideas and trying to guide me to meaningful questions.

# TABLE OF CONTENTS

# LIST OF TABLES

Table Page

# LIST OF FIGURES

FIGURE Page

FIGURE Page

# CHAPTER I

# INTRODUCTION

## Overall Problem

Bacteria exist in essentially in two states, vegetative growth and stationary phase. Gram negative bacteria such as *Escherichia coli* enter stationary phase when they encounter conditions not suitable for rapid growth. *E. coli* must prepare for extended periods of starvation in the absence of sporulation. This dramatic change in physiology results in morphological and biochemical changes to the cells and an array of programmed changes in gene expression.

In stationary phase, the cells undergo alterations to the inner and outer membrane lipid composition, thickening and increases in the crosslinking of the peptidoglycan cell wall length, and changes in cell size and buoyant density (Huisman *et al.*, 1996). Some of the macromolecular changes include dimerization of the ribosomes and accumulation of storage compounds, like trehalose (Hengge-Aronis, 1996; Huisman *et al.*, 1996). The genome of the cell is condensed into a damage-resistant complex with the DNA binding protein Dps.

Morphologically, the cells are smaller, more rounded and more resistant than rapidly growing cells to environmental stresses like acid and detergent

This dissertation follows the style and format of *Molecular Microbiology.*

(Hengge-Aronis, 1996; Huisman *et al.*, 1996). Changes in size of the bacteria are explained by two phenomena: reductive division and dwarfing. Reductive division is the process by which bacteria entering stationary phase must complete division of genomes for which they have already begun replication (Nystrom, 2004). Dwarfing by contrast is the general loss of cell mass due to degradation of cell material including components of the inner membrane, cell wall and cytosol. Most of these alterations require extensive changes in protein synthesis mediated by a complex program of gene expression.

The major questions pertaining to stationary phase at the protein level are: What is the complement of protein components in stationary phase relative to exponential growth? What are the changes in the localization and interactions of these proteins? And, what insights into regulation of stationary phase can be gleaned from a global understanding of changes in protein content?

**Regulation of Stress Responses**

The accumulation of stationary phase specific gene products reveals a time-dependent progression through stationary phase (Groat *et al.*, 1986; Matin, 1991; Stephani *et al.*, 2003). Subsets of these genes are expressed in waves. Many genes involved in stationary phase are controlled by alternative sigma factors, primarily RpoS. During glucose starvation, levels of RpoH increase, inducing synthesis of the heat shock proteins DnaK, GroEL, and HtpG (Jenkins *et al.*, 1988). RpoE levels also increase in stationary phase (Nitta *et al.*, 2000). Catabolite repression, small-molecule metabolites such as the alarmone ppGpp, and proteolysis all play roles in the programmed entry into stationary phase.

Stationary phase is the common response for cells experiencing different forms of nutrient limitation. Studies of cells subjected to nutrient starvation or changes in media composition have resulted in the identification of many pathways and regulators of stress-specific gene expression (Jenkins *et al.*, 1988; Loewen *et al.*, 1998; Schultz *et al.*, 1988; Stephani *et al.*, 2003; Weichart *et al.*, 2003). These include nitrogen, phosphorus, and carbon starvation (Corbin *et al.*, 2003; Groat *et al.*, 1986; Klose *et al.*, 1994; Zimmer *et al.*, 2000). Although RpoS is involved in all of these responses, there is also a significant RpoS-independent component to the shared stationary phase program.

Gene expression associated with stationary phase is generally studied with reference to RpoS (Chatterji and Ojha, 2001; Hirsch and Elliott, 2002; Matin, 1991; Nystrom, 2003, 2004). Through a combination of increased expression and changes in proteolysis, $\sigma^S$ levels increase in response to a variety of environmental stresses, resulting in increases in expression of more than 100 RpoS-dependent gene products (Chatterji and Ojha, 2001; Corbin *et al.*, 2003; Hengge-Aronis, 1996; Lacour and Landini, 2004; Patten *et al.*, 2004; Tani *et al.*, 2002; Vijayakumar *et al.*, 2004). RpoS regulates other stress responses including acid growth and osmotic shock (Arnold *et al.*, 2001; Jishage and Ishihama, 1995; Weber *et al.*, 2005). Many of the genes associated with stationary phase are not necessarily specific to that response.

The focus on RpoS may overemphasize the role of transcriptional activation in stationary phase. Genes involved in growth are repressed, and proteins synthesized prior to entry into stationary phase are degraded. Understanding of the stationary phase response would benefit from a more complete characterization of the associated changes in the intracellular biochemistry of *E. coli*. This includes changes in protein levels, protein-protein interactions, and protein modifications.

The complexity of regulation in stationary phase response is typical of general regulatory responses and lends itself to global analysis as many different pathways are activated and repressed, creating a network of dependent expression. Additional interactions in these networks reveal further complexity in tuning the bacterial stress response. Involvement of Lrp and CRP/cAMP in the stationary response for example add additional levels of complexity which have been difficult to fully understand with current models (Tani *et al.*, 2002; Weber *et al.*, 2005; Zinser and Kolter, 2000). An example of the degree to which even a small number of gene products are interconnected in stationary phase is shown in Figure 1.1. These 7 proteins have more than 30 defined pleiotropic dependencies within stationary phase alone (Hengge-Aronis, 1996). Integration of the effects resulting from regulation due to multiple stimuli has been problematic and is best addressed with global approaches.

## Global Stationary Phase Gene Identification by DNA Microarrays

Global approaches strive to identify components of a particular organism or pathway. They are not typically hypothesis-driven and the goal is to generate an unbiased collection of changes or identifications. The most widely employed global analysis of cell change is the DNA microarray. In these experiments, mRNA or cDNA from cells is hybridized to arrays containing DNA from most or all of the open reading frames in the genome. Quantitative measurements can be made on the changes in mRNA levels between cells or cell states. Changes in potential protein content are inferred from this providing an organism-level examination of which genes are activated or repressed in response to a change in growth condition, or a mutation.

Many global studies of stationary phase-dependent gene transcription have been performed utilizing DNA microarrays. (Arnold *et al.*, 2001; Tani *et al.*, 2002; Weber *et al.*, 2005; Weichart *et al.*, 2003; Zhu *et al.*, 2001). These studies have identified upwards of 100 genes with increases in mRNA levels upon entering stationary phase in either rich or minimal media (Patten *et al.*, 2004; Tani *et al.*, 2002; Tao *et al.*, 1999; Weber *et al.*, 2005). Gene arrays have also been employed to examine changes in expression under different conditions including nitrogen starvation, acid tolerance, and osmotic shock (Arnold *et al.*, 2001; Patten *et al.*, 2004; Weber and Jung, 2002; Zhu *et al.*, 2001; Zimmer *et al.*, 2000). All of these experiments revealed overlapping and distinct gene products from each stress response.

Weber *et al.* in particular tested three of these conditions [glucose exhaustion, acid growth at (pH 5.0), and osmotic shock (NaCl)] and established a set of 140 'core' RpoS dependent stress response genes and several hundred other RpoS-dependent proteins specific to a particular growth stress (Weber *et al.*, 2005). This and other studies show that RpoS dependence alone is not sufficient to explain many observations and changes in expression during stationary phase (Chatterji and Ojha, 2001; Hirsch and Elliott, 2002; Loewen *et al.*, 1998; Nystrom, 2003). These studies have provided useful insight into the regulatory networks and the beginnings of how the expression of gene products is fine-tuned for specific conditions.

The major results from transcriptional studies are that the number of proteins which are induced upon entry into stationary phase is large and functionally diverse. Of

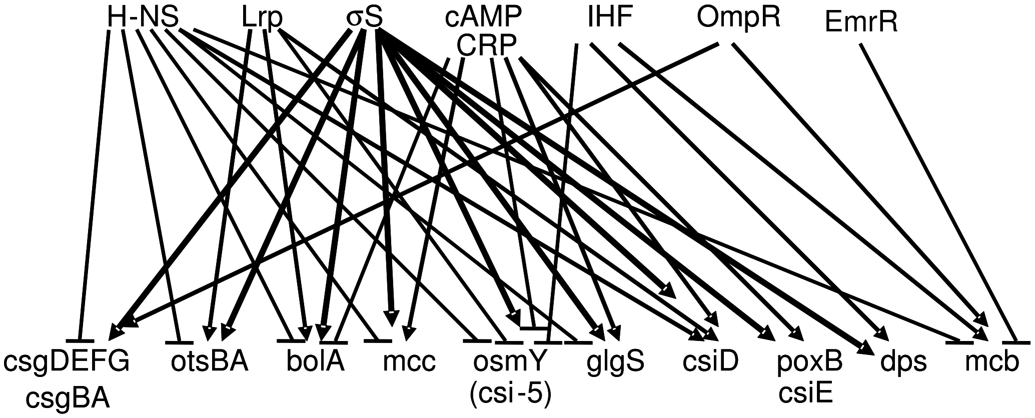


Figure 1.1.) Illustration of regulatory dependencies within a subset of stationary-phase induced genes from *E. coli*. (Adapted from (Hengge-Aronis, 1996). On the top are known regulatory proteins, and shown in lines are positive and/or negative regulatory interactions with genes and gene systems across the bottom. The arrows do not necessarily imply a direct interaction, but rather illustrate a portion of the complexity present in stationary phase-specific gene expression.

the 140 shared general stress response genes in Weber *et al*., 62% of them have unknown or putative functions. Arrays identify increases and decreases in levels of mRNA in response to specific starvation conditions. Overall, these experiments describe more than 500 genes with increased expression upon entry into stationary phase. Other methods should reveal additional changes in protein content. Proteomics, which involves directidentification of the proteins present, has the capacity to identify many aspects of global physiology that arrays cannot determine.

## Aspects of Global Regulation Are Not Observable by mRNA Examination

A fundamental question is, "how do the proteins from *E. coli* change when it starves?" Cataloging the changes in gene transcription is necessary but not sufficient to understand the components of the stationary phase response. mRNA levels do elucidate likely changes in protein levels in balanced growth, but this correlation is likely to not hold during stationary phase. For example, RpoS levels are regulated post-transcriptionally, by changes in translation and protein stability (Hengge-Aronis, 1996). In addition, mRNA cannot reveal localization, interactions, or modification of polypeptides.

Proteases are responsible for a large amount of regulation in stationary phase; changes in the interaction with ClpXP are a major source of RpoS accumulation in stationary phase *E. coli* (Loewen *et al.*, 1998; Mogk *et al.*, 2003; Stephani *et al.*, 2003; Weichart *et al.*, 2003). Surprisingly, the global signals for overall changes in protease specificity are not known although oxidation of proteins in late stationary phase has been suggested as a possible signal for regulation of proteolytic targeting (Nystrom, 2004). Specific protease systems such as ClpXP and *trans*-translation have preferred substrates, and the extent to which these preferences are regulated is not understood (Flynn *et al.*, 2003; Neher *et al.*, 2003; Sauer *et al.*, 2004). Weichart *et al.* identified 18 proteins as potential substrates of ClpXP or ClpAP that were altered in Clp$^-$ mutants during entry into stationary phase (Weichart *et al.*, 2003). Flynn *et al.* identified preferred substrates of ClpXP by utilizing protease mutants to trap substrates inside the proteosome; trapped proteins were later identified by mass spectrometry. Adopting this strategy to identify changes in protease substrates at different time points would greatly add to our understanding of the role of proteolysis in stationary phase (Flynn *et al.*, 2003; Stephani *et al.*, 2003).

Thus, global approaches to addressing the protein content of *E. coli* directly under specific growth conditions and the development of methods to identify them were the focus of this research. Our research addressed these major issues: (1) the separation of bacterial proteins for identification. (2) the development of these methods to interrogate the protein content of *E. coli* in response to carbon starvation, and (3) integration of this biological knowledge into generating improvements to enhance data collection and validation. The incomplete picture of and our lack of knowledge of the function of most of the genes expressed in stationary phase underscores the fact that our understanding of this response is incomplete.

## Proteomics

Genomics is the study of the organization and programmed implementation of the genetic blueprint. In contrast, the proteome of a cell is the complement of its expressed proteins at points in time (Wilkins *et al.*, 1996). The complement of proteins in an organism changes in absolute and relative abundance, modifications, and protein-protein interactions. Proteins thus represent the output of the genetic blueprint and the sum result of the genetic evolution in organisms. Indeed, the extent to which organisms are able to regulate the number of protein interactions is a major source evolutionary diversity in higher organisms (Baltimore, 2001; Graveley, 2001; Levine and Tjian, 2003).

The idea of cataloging protein content from cells certainly existed prior to 1995. Labs were performing experiments that would be called "Proteomics" today. Subsequent technological developments created a vocabulary for these experiments, which led to the coining of "Proteome" in 1995 by Wilkins *et al*. (Wasinger *et al.*, 1995; Wilkins *et al.*, 1996).

Overall, progress in the field has progressed from being technology oriented to biology oriented. Early experiments focused on developing the capacity to identify macromolecules from biological samples (Jensen *et al.*, 1997; Shevchenko *et al.*, 1996a; Shevchenko *et al.*, 1996b). These experiments were necessary to develop the hardware and informatic tools essential for handling and interpretation of the data (Pappin *et al.*, 1993; Washburn and Yates, 2000; Wolters *et al.*, 2001). Relatively quickly, technology-driven experiments as probes of systems biology were performed including interaction studies and whole cell identifications (Gavin *et al.*, 2002; Washburn and Yates, 2000).

Proteomics is currently an emerging analytical field, and the development of methods can supersede initial biological significance. The goal of this is that an understanding of function and regulation should follow identification. In other words, improvements in identification of proteins will lead to understanding of function and physiology. The growth of this field is tremendous. Figure 1.2 illustrates this by comparing the publication rates from PubMed search terms "proteome" and "proteomics" by year. It is shown plotted with total PubMed hits and a search term, “glycolysis,” for comparison. PubMed has an average growth rate of about 3.6% per annum since 1995. ‘Proteome’ or ‘Proteomics’ has an average growth rate in publications of about 132% per year since 1995, or 976% in total. ‘Glycolysis,’ by contrast has an annual growth rate of about 1.8%, which does not even keep abreast with the core PubMed ‘inflation’ rate.

## Technological Challenges to Large-Scale Protein Identification

Overall identification has been limited by the capacity to identify more than a few component proteins in a complex mixture without additional fractionation. The problems of identification using proteomics can be broken into two categories: Complexity and dynamic range. Complexity is the number of constituent compounds present within a sample, and dynamic range is the difference in concentration from the least abundant component to the most abundant. Both of these properties are crucial when considering sample analysis.

Even bacteria with relatively small-to-average numbers of predicted genes such as *E. coli* or *H. influenzae* encode for several thousand gene products (Blattner *et al.*, 1997; Fleischmann *et al.*, 1995). Unlike the genome, protein content cannot be amplified, thus identifying any component from a mixture with a dynamic range of six orders of magnitude is virtually impossible without enrichment (Corthals *et al.*, 2000). More complex organisms can possess dynamic ranges in protein concentrations in excess of $10^{12}$, increasing the difficulty in selecting specific components (Huber, 2003; Stasyk and Huber, 2004). Fractionation is necessary because no single technique can resolve components at biological levels of complexity. Most analytical techniques have an ability to resolve three to four orders of magnitude in complexity (number of resolvable

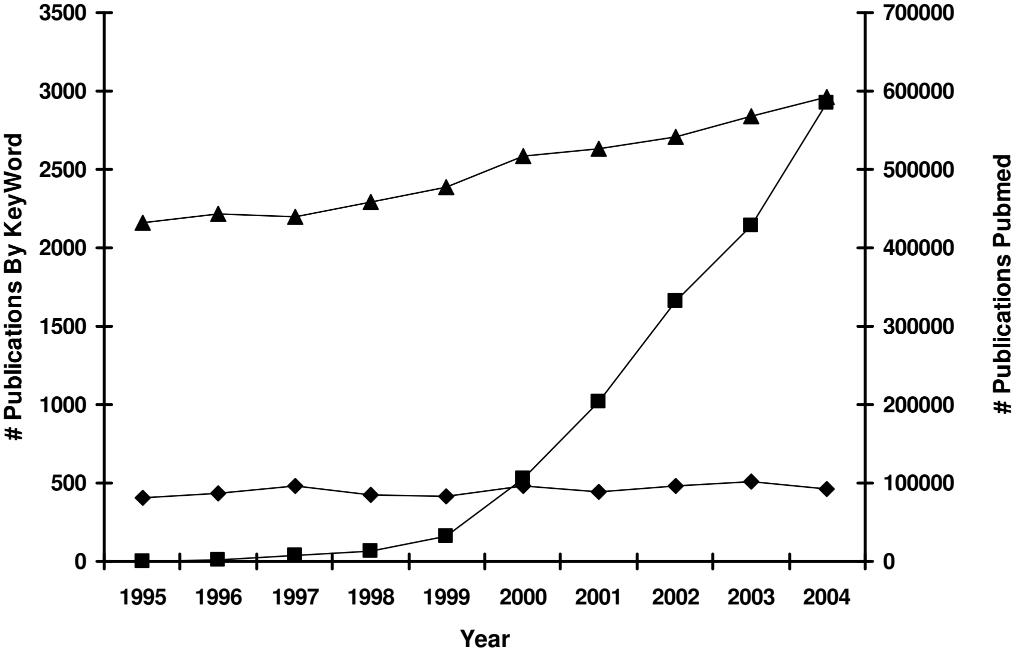


Figure 1.2.) Citation hits by year for key search terms. Pubmed citations were determined by performing year limited searches for proteome or proteomics [Squares], glycolysis [Diamonds] and all Pubmed hits [Triangles]. Pubmed is accessed through the web at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi. Meta analysis was performed using the standard advanced search features for the keywords listed or no keyword at all for total publication rates.

components) and concentration sensitivity (dynamic range). In order to examine the contents of a bacterium, fractionation of two or three orders of magnitude is necessary to potentially identify most components.

Fractionation of an organism is centered on two goals: Enriching components of interest up to sufficient levels to be amenable for analysis, and separating the total mixture until each component 'part' is within the resolving power of the analytical technique. At present, studies utilizing direct analysis of total cell protein are limited to identifying tens of high abundance components. This may be sufficient for microorganism identification in a clinical setting, but it is not applicable to studying physiology (Loo *et al.*, 2001; Park and Russell, 2001).

**Fractionation**

Fractionation involves separation of analytes via specific chemical properties. Two major approaches are utilized: gel electrophoresis, and chromatography. Stasyk *et al*. reviews these two main approaches to fractionation and their application to proteome identifications (Stasyk and Huber, 2004). All strategies to perform biochemical separations must be done at a scale to provide sufficient material for further analysis. Virtually all of these separation techniques were developed prior to proteomics, and only recently have instrumentation and interfaces been developed that allow their output to be analyzed by mass spectrometry.

Fractionation strategies are limited in resolving power. Modern techniques cannot resolve more than hundreds to thousands of individual components, which necessitates a tradeoff since virtually all organisms possess more than that number of proteins. An experiment cannot increase its depth of coverage without sacrificing particular portions of the information (proteins). Although the ultimate solution may come from improved instrumentation, current solutions focus on the separation of molecules into simpler subsets. Typically, increasing the dimensions of separation multiplies the resolving power of each technique. In practice, no set of techniques is completely orthogonal and the net number of resolved components will be less than the theoretical maximum.

**Electrophoretic Separations**

The most common methods used to separate proteins using electrophoresis are 1 and 2D gel electrophoresis. Discontinuous 1D gel electrophoresis introduced by Laemmli and the subsequent introduction of 2D gel electrophoresis allowed for the simultaneous examination of multiple gene products (Klose, 1975; Laemmli, 1970; O'Farrell, 1975). This allowed visualization of changes in intensity and expression of hundreds to thousands of bands/proteins. It was not practical to identify large numbers of proteins separated until biological mass spectrometry was more developed. Identification of the components of these bands has been made routine, although success rates vary (Jensen *et al.*, 1997; Jensen *et al.*, 1998; Shevchenko *et al.*, 1996a; Shevchenko *et al.*, 1996b). Rabbilloud  and Görg highlight that 2D electrophoresis has enabled the visualization of thousands of gene products, and the dynamic range and resolution of 2D gels is still superior to other methods (Gorg *et al.*, 2000; Gorg *et al.*, 2004, 2005).

Gel electrophoresis has many distinct advantages. The equipment necessary to perform these separations is present in nearly every laboratory. For simple separations, 1D SDS PAGE makes highly focused bands, which are easily removed for identification.

Sensitivity of detection can be increased by two to three orders of magnitude either by changing the staining method (e.g. Coomassie to silver, Sypro etc.) or adding in a second dimension of separation by pI (2D GE). Gel electrophoresis is also compatible with histochemical and immunological detection, as well as autoradiography. Gel electrophoresis can be used to measure the pI and molecular weight of proteins. 2D gels in particular provide unit resolution of proteins, which is difficult to achieve with other analytical techniques. These many advantages are why to see gel electrophoresis continues to be utilized in most proteome experiments, despite some intrinsic disadvantages to the technology.

Topographical presentation of protein spots is a popular means by which proteomes are represented. This is not surprising; there is a familiarity in the presentation format. As a means to visualize the proteome of *E. coli*, virtual 2D gels of the predicted pI and MW of the *E. coli* genome have been independently generated by several groups (Cavalcoli *et al.*, 1997; Corthals *et al.*, 2000; Link *et al.*, 1997). Our representation of the data is presented later in this work. Figure 1.3 illustrates some of these representations of a virtual 'gel' genome. Of interest is the distinct gap in predicted pI from the *E. coli* proteome around pH 7.1-7.4, which corresponds with the pH inside of the bacteria. This makes sense as proteins are generally least soluble at their pI. This gap is not apparent from actual 2D PAGE of cell lysates themselves. This could be due to sample artifacts and modification of proteins or due to differences between predicted pI and where spots eventually migrate. Görg *et al.* outlines the acquisition of data from 2D gels in a protocol-descriptive manner, including the steps necessary to hydrate, load, run and

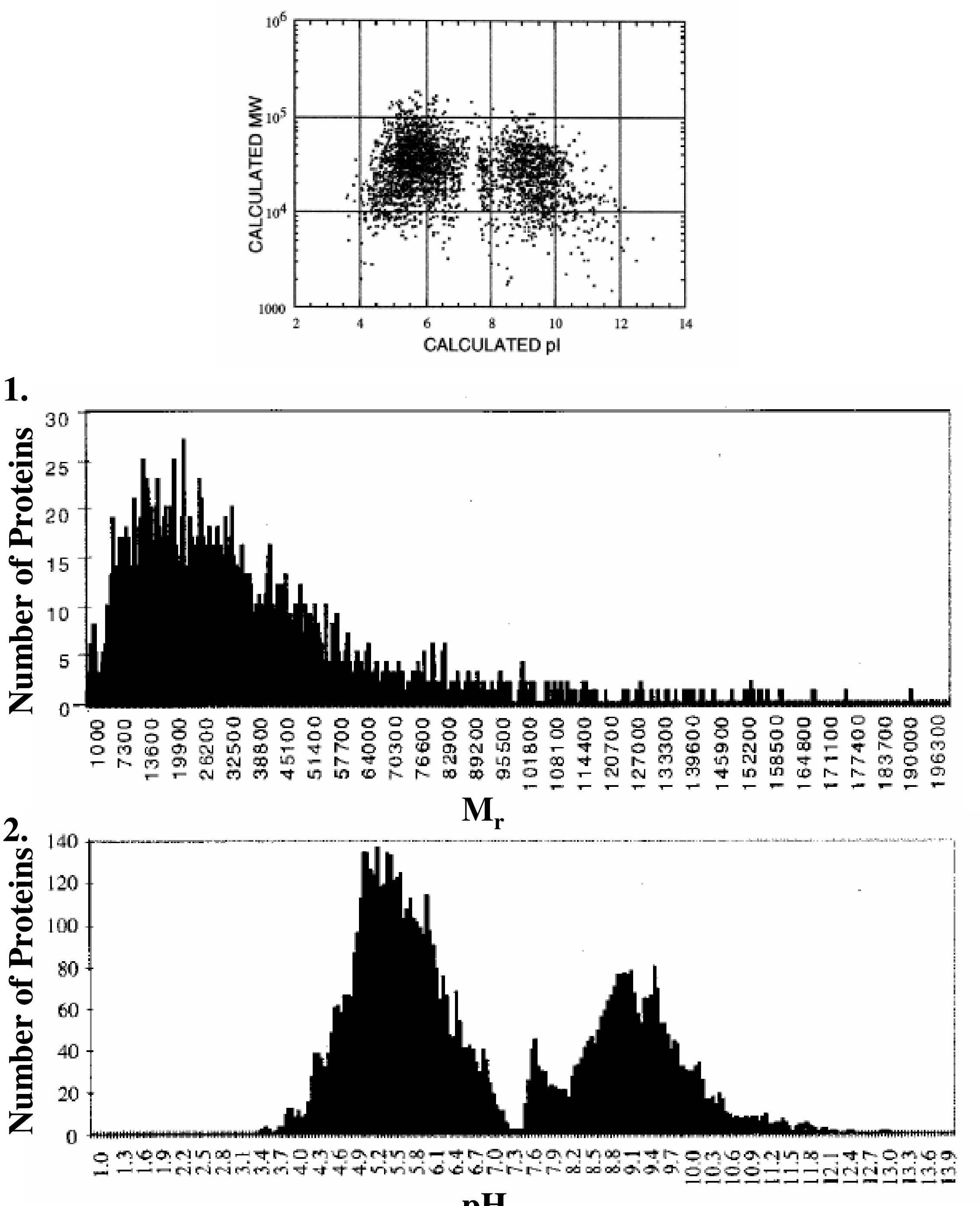


Figure 1.3.) Graphical representation of *E. coli* proteome in 2D gel-like formats. These are graphical representations of the *E. coli* proteome in ‘virtual 2D gel syle from two sources, note the slight differences between presentations depending on the equation utilized to determine pI/M$_r$, From VanBogelen *et al.*, 1999. pI and molecular weight were determined from the Compute pI/M$_r$ tool on ExPAsy. Bottom 1 and 2. pI and Molecular weight were determined by internal calculations and published pI reference points (Bjellqvist *et al.*, 1982; Cavalcoli *et al.*, 1997).

visualize modern IPG (immobilized pH) 1$^{st}$ dimension gel strips (Gorg *et al.*, 2005; Rabilloud, 2002).

The main disadvantages of 2D gels are the lack of analytical level reproducibility in sample sets and the difficulties of automation (Rabilloud, 2002). Despite the high resolving power, 2D gels lose some of this capacity due to the fact that many proteins exist as multiple spots on a gel. This is due to biological modification or artifacts during the separation process. The solid-state nature of 1D and 2D gels does not lend itself well to high throughput due to the many mechanical operations necessary to excise and process samples. Most groups however, have only limited gel running needs, in which case massive parallel separations and identification are unnecessary. For traditional protein identification, 1D gel electrophoresis is still the main source from which bands are obtained.

Gel electrophoresis also suffers from a limited loading capacity. The chemical properties of the separation also create problems. The first dimension of 2D GE is not effective at separating strongly acidic and basic compounds, and large and small biomolecules present a problem in SDS PAGE. pI as a fractionation dimension is underresolved relative to LC (liquid chromatography) methods as most proteins tend to separate within a the same narrow isoelectric range. Isoelectric fractionation is performed under strongly denaturing and reducing conditions, which modifies polypeptides and creates artifacts. In practice this limits the useful pI range to a narrow window. Multiple pH ranges and Zoom-gel pH ranges exist, which have helped in mitigating some of the difficulties in isoelectric separations (Gorg *et al.*, 2004, 2005).

## Liquid Chromatography Fractionation

LC separations of biomolecules are a mainstay of biochemical characterization. Molecules can be separated by chemical or physical properties. Cation and anion exchange rely upon differences in charge of either proteins or peptides. Hydrophobic interaction, size exclusion, and various affinity separations are also common (Alberts *et al.*, 1968; Champion *et al.*, 2003; Opiteck *et al.*, 1997; Opiteck *et al.*, 1998).

Many additional fractionations have been utilized to select for certain sub-classes of proteins, including heparin (cation exchange), hydroxyapatite, and anion exchange (Champion *et al.*, 2003; Fountoulakis and Takacs, 1998; Langen *et al.*, 2000; Lee and Lee, 2004). And as we describe in this work, combinations of these fractionations can be employed together to generate very specific subsets of intact cellular proteins prior to identification.

Today, the most common approach for high dynamic-range analysis is to digest the components of complicated mixtures of proteins (whole cell lysates) and then separate the peptides over multiple dimensions of HPLC. Generating most of the biomass as peptides eliminates the need for biologically compatible separation techniques. The in-solution properties of peptides are significantly more robust and generic than those of intact proteins, and allow high resolution separation methods which are compatible for direct MS and MS/MS ionization is (Ducret *et al.*, 1998; Opiteck *et al.*, 1998; Yates, 1998).

These fractionations are typically strong cation exchange, followed by reverse phase ($C_{18}$) separation directly into mass spectrometers. Two of the first examples of this approach were performed by Woburn *et al.* where samples were fractionated by 2D gel electrophoresis prior to reverse phase separation of the peptide digest, which was then developed into an in-line 2D LC separation of total digest directly by cation exchange and RPC (reversed phase chromatography) (Washburn and Yates, 2000; Washburn *et al.*, 2001; Wolters *et al.*, 2001). This approach of determining protein identification from mixtures of peptides is only possible because computational power exists to reconstruct the proteins from constituent sequences, and tandem MS/MS data enables identification without requiring all peptides to be present together.

Multidimensional separation of peptides is not the only LC fractionation strategy employed. In Fountoulakis *et al*. an LC separation from *E. coli* lysates fractionated over hydroxyapatite resin, then separated via one and two-dimensional gels prior to MS analysis eliminated the need for specific sample preparation to that type of LC (Fountoulakis *et al.*, 1999a; Fountoulakis *et al.*, 1999b).

### Electrophoretic/LC Combinations

Initially, hybrid-style separations were common. One of the first LC/MS approaches was performed by Opiteck *et al*. (Opiteck *et al.*, 1997), where originally 2D gels were utilized, but eventually abandoned for an all LC fractionation approach. Today, many different types of sample preparation are utilized, but reversed phase chromatography of the peptide fragments, often directly into a mass spectrometer, is the preferred method for analysis. In addition to being performed under solvent systems compatible with mass spectrometric analysis, extremely high sensitivity can be achieved with low (nl/min) flow systems and hyper fine column diameters (typically 75µm ID). Because of this common entry-point with RPC, the conditions can be standardized, improving the reliability of an otherwise difficult technique. (Zhen *et al.*, 2004). This allows for samples generated from gels, for example, to be relatively easily analyzed with LC/MS/MS technology.

However, SDS-PAGE and 2D gel electrophoresis remain the most common means of primary biological separation and MS sample preparation (Rabilloud, 2002). There have been efforts to assign complete proteomes in gel-free systems (Chong *et al.*, 2001; Ducret *et al.*, 1998; Link *et al.*, 1999; Loo *et al.*, 2001; Washburn *et al.*, 2001; Yates, 1998). For bacteria in particular, they are quite viable. In cases where *E. coli* was examined, 2D gels were capable of resolving approximately 70 to 80% of the predicted proteome, a number which diminishes substantially in more complex species (Hoogland *et al.*, 2000; Link *et al.*, 1997; Tonella *et al.*, 2001).

Several groups designed 2D GE-like separations, which would have higher capacity and be more robust. Loo *et al*. illustrated this by visualizing LC separation techniques, and creating virtual 2D gel profiles from separations where MS replaced SDS PAGE as the primary means by which molecular weight was determined (Loo *et al.*, 2001). More complex organisms and samples require additional fractionation and depletion of abundant components. The most obvious conclusion from the data available is that the overall limitation of 2D gel proteomics is total dynamic range relative to multi-dimensional LC separations. Multi dimensional HPLC approaches are clearly more capable of identifying greater raw numbers of protein components from complex mixtures (Ducret *et al.*, 1998; Washburn and Yates, 2000; Washburn *et al.*, 2001; Wolters *et al.*, 2001).

The multi-dimensional separations by Yates *et al*. were excellent illustrations of the evolution of the fractionation from off-line LC coupled to 2D electrophoresis to

MS/MS analysis, to an on-line peptide fractionation followed by automated MS/MS analysis, for a review see (Yates, 2004). K. Resing *et al*. performed extensive refinement of MuDPIT (Multi Dimensional Protein Identification Technology), by employing gas phase fractionation in the mass spectrometer prior to precursor selection. Substantial improvements were made in software identification as well as the optimized fractionations (Resing and Ahn, 2005; Wysocki *et al.*, 2005). Vollmer *et al*. performed optimization of variables in off-line 2D LC of peptides for MS/MS analysis. Fractionation of *E. coli* lysates and standard protein mixtures utilized strong cation exchange with fraction collection followed by reversed-phase LC/MS/MS analysis. Efficiency is typically monitored as a function of unique peptides identified, not necessarily proteins found, which in itself is descriptive of the goals of sample preparation and fractionation in general, that is they are not necessarily oriented towards specific biological questions (Vollmer *et al.*, 2004).

**Identification**

Identification of peptides and proteins is made by several means, including antibodies, chemical assay, and reporter compounds. Analytically, identification is now achieved predominantly by mass spectrometry (Brancia *et al.*, 2001; Gevaert and Vandekerckhove, 2000; Jensen *et al.*, 1996; Link *et al.*, 1997; Wasinger and Humphery-Smith, 1998; Wilkins *et al.*, 1998). Identification by mass spectrometry became possible after the discovery of the soft ionization techniques of electrospray by John Fenn and Matrix Assisted Laser Desorption-Ionization (MALDI) by Kraus, Hillenkamp, and Tanaka for which John Fenn and Koichi Tanaka were awarded Nobel Prizes in 2002 (http://www.nobel.se). The addition of genome sequences and databases and the development of statistical models for interpreting protein and peptide spectra further enable the routine identification of proteins from separated biological samples.

MS and MS/MS based approaches for identification would not be routine without the large number of complete genome sequences. Annotated genomes allow identification of potential proteins by comparing peptide fragments to virtual proteins predicted by the genome. In most cases, the genome(s) of interest are translated *in silico*, and a database of all predicted peptides and sequence products is generated. Fragment lists or MS/MS fragment ions, determined using biological mass spectrometry are then compared to these predicted databases, from which protein identification is inferred.

Systematic identification of proteins by mathing peptide masses and fragment masses to a peptide of origin is difficult. Many peptides generate similar fragments and changes in sample complexity, mass accuracy, and resolution dictate the degree of ambiguity in identification. Peptides generated by a single protease also tend to occur within a narrow mass range. Protein assignment by MS or MS/MS analysis is essentially a hypothesis that it is present in a sample, and validation of these results is increasingly complex.

Peptide-mass fingerprinting (PMF) is the process by which proteins are identified on the basis of masses of proteolytic fragments (Cottrell, 1994). Many studies have established criteria by which PMF identifications are made (Jensen *et al.*, 1997; Jungblut *et al.*, 1999). Although the general process of comparing observed fragment masses to calculated ones was described in 1984, the systematic identification of proteins from organisms required genome sequences to be available (Cottrell, 1994; Gevaert and Vandekerckhove, 2000; Gras *et al.*, 1999; Lee *et al.*, 2002a). Wise *et al*. described an *in* *silico* dissection of the minimum number of peptides and potential specific cleavage sites needed for unambiguous identification of proteins from PMF. Overall, they concluded that no single enzyme was sufficient to provide uniqueness, but this work relied upon the several assumptions (Wise *et al.*, 1997a; Wise *et al.*, 1997b) which now can be largely overcome using current instrumentation. High resolution (High mass accuracy below 10-20 ppm) MALDI-TOF MS, for example can resolve mixtures of several proteins digested together in a single spectra (Jensen *et al.*, 1997; Park and Russell, 2000, 2001; Russell *et al.*, 2001) Additional assumptions made by Wise were not consistent with results from empirical experiments. These studies however, did establish potential limits of digest agents to identify all proteins, which spans from 8.4 to 18.1% of a total proteome. Limitations of fractionation and sensitivity dominate the depth of a proteome. This is consistent with the identification rate reported from papers that utilized 1 and 2D gel electrophoresis followed by PMF (MALDI) identification (Hoogland *et al.*, 2000; Jungblut *et al.*, 1999; Shevchenko *et al.*, 1996a; Tonella *et al.*, 1998; Tonella *et al.*, 2001).

For a given set of peptides to identify a protein, a unique peptide need not exist; the combination of many peptides within the same fraction can be sufficient to determine an ID. This is a common approach among software packages and is often an Occam's razor approach, where a peptide matching multiple entries from a database is more likely to have originated from the protein for which a number of the peptides were already matched.

Several automated approaches exist to compare empirical MS data to theoretical digests and report scores or confidence metrics for protein ID. Originally, MS and

MS/MS data were compared to predicted ions manually, an impractical approach for complex mixtures. One of the earliest such algorithms to automate PMF was designed by Bleasby *et al*. in 1993 called MOWSE or (MOlecular Weight SEarch) which described a scoring matrix based on the coincident detection of many peptides from the same protein in a single experiment (Pappin *et al.*, 1993; Perkins *et al.*, 1999). The database against which data are to be searched is generated *in silico* by computing all of the peptides that would be generated from digesting each protein with a particular enzyme. The caclulated peptides are separated into bins corresponding to their molecular weight, and each peptide is assigned a score based on its fractional abundance in a particular bin. For example, if a peptide of MW 1000.000 Da is in a bin of 3000 Da width with 9,999 other peptides its relative 'uniqueness' would be 0.0001. Therefore, the number of submitted peptides matching one of these virtual peptides +/- mass error tolerances determines the score any protein receives, and their respective uniqueness scores are multiplied together. There are additional correction factors for protein size, but these are small relative to the product score.

This works extremely well under certain circumstances: First, the sequences generated are from a sequenced and annotated genome. Second, the number of peptides is from relatively non-complex samples. Non-complex in this case refers to the analysis of a single to about 10 or fewer proteins worth of peptides. Ambiguity in identification arises because the algorithms are generally not sensitive to mass error sensitive with regard to scoring. At some point, given enough submitted peptides a database will generate high-scoring matches to almost anything, limiting the utility of this approach.

Today, detection and identification of mixtures of proteins with automated search software is routine. The MOWSE algorithm is still the most common means by which peptides are scored. Variations of this system underlie peptide-level scoring for two of the most popular search routines, Protein Prospector (MS-FIT) from Karl Clauser at the University of CA, San Francisco, and Mascot, from Matrix Science. More recently, error-sensitive searches and Bayesian scoring based on empirical data have further refined peptide mass fingerprinting to the point where it is extremely robust and discriminating for simpler mixtures of peptides (Zhang and Chait, 2000).

In the absence of tandem MS data, several criteria have been applied to determine whether peptides observed in PMF are sufficient for calling a protein ID. Detecting small numbers of peptides for a protein is generally not considered sufficient for reliable identification. The most common threshold applied to identification is sequence coverage. This is accomplished by determining how much of the polypeptide sequence as a % of the total is represented in the data. A generally accepted number for PMF sequence coverage is ≥20-25% in simple mixtures, defined above (Baldwin, 2004; Jungblut *et al.*, 1999).

The major reason more complicated analysis is difficult using PMF is that as the number of peptides increases the false positive identifications due to random peak matching begin to approach the number of true positive identifications (Cargile *et al.*, 2004; Keller *et al.*, 2002). A portion of this is due to large peak lists and ambiguity in assigning masses, but it is also due in part to random hits on numerous orphan peptide masses that frequent spectra (Keller *et al.*, 2002). In Karty *et al.* a protein identification study in *C. crescentus*, one-third of the observed peptides could not be assigned to proteins. They were able to assign identities to 75% of these unexplained peptides with expanded assumptions about trypsin, peptide modifications, contaminants, and handling artifacts like deamidation and oxidation (Karty *et al.*, 2002). Even taking into account multiple proteins in a spot, there is no readily available explanation for the remaining 25% of the unannotated masses. This trend is consistent with data taken from tandem (MS/MS) experiments as well (Gavin *et al.*, 2002; Washburn and Yates, 2000).

An increasingly popular approach involves the use of tandem mass spectrometry to select precursor masses of peptides or proteins and subject them to ion activation followed by fragmentation and subsequent detection of the daughter ions for sequence comparison (Wolters *et al.*, 2001). In practice, these fragment ions are rarely used to directly sequence the polypeptide. Instead, like fingerprinting, empirical daughter MS/MS ions are compared to virtual daughter ions for each predicted peptide in a database. An example of a MALDI spectrum typically utilized in peptide mass fingerprinting is show as Figure 1.4A, and a typical precursor ion-selection and fragmentation by Collisionally Induced Dissociation (CID) is shown as 1.4B. In MALDI MS, the spectrum shows hundreds of individual precursor peptides from a mixture of several proteins digested with trypsin. The tandem MS/MS spectrum in (B) shows the fragment ions from one single tryptic precursor subjected to CID. Each peptide from a digest theoretically provides all of the degenerate information within its fragments. Thus tandem MS/MS data has greater information content, increasing the likelihood of an unambiguous identification. Tandem MS/MS data also has the advantage of potentially determining on which amino acid post-translational modifications such as phosphorylation and glycosylation occur (Yates, 2004).

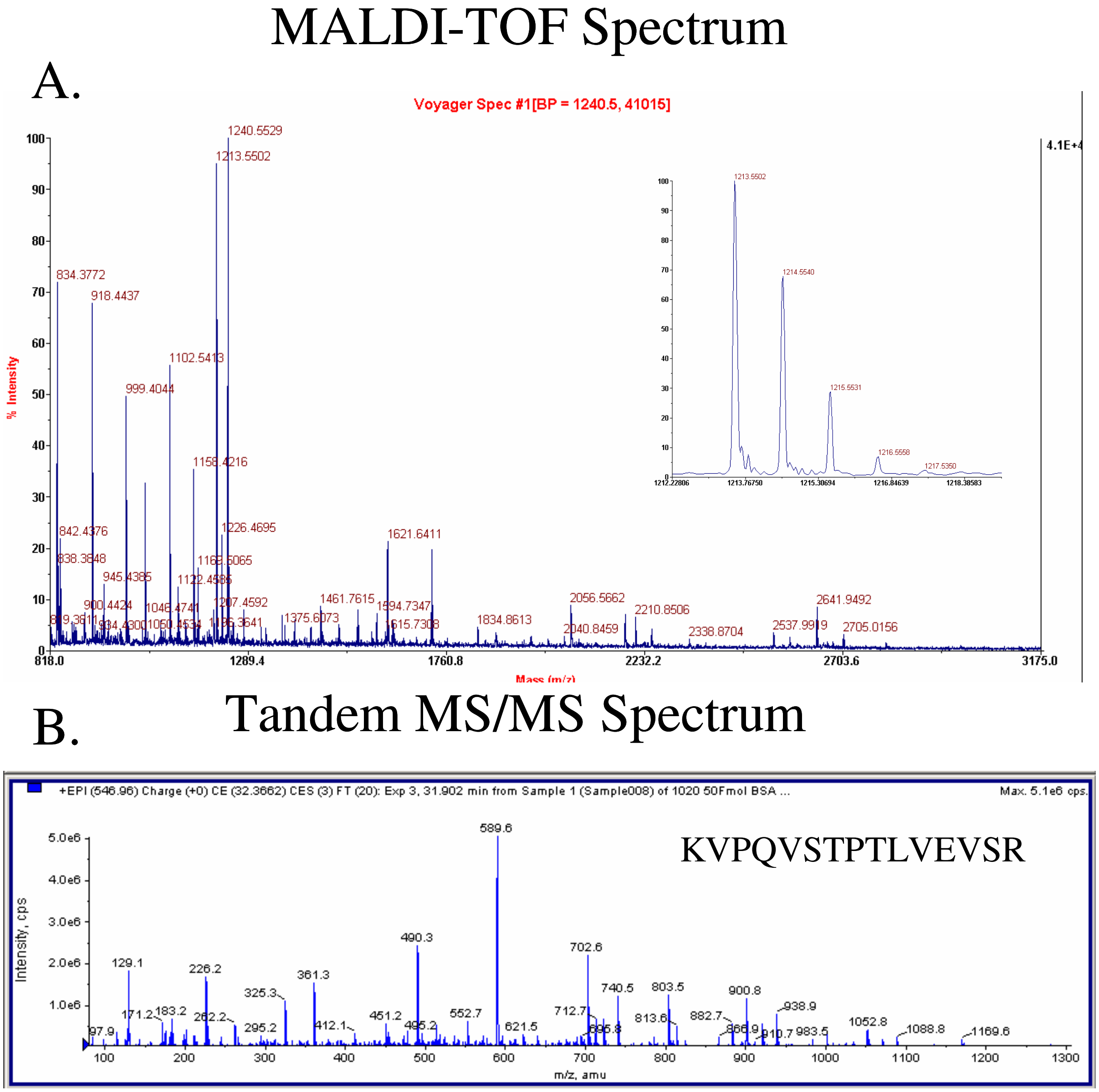


Figure 1.4.) Example spectra of peptides from MS instrumentation. A. Example spectrum from a MALDI-TOF instrument of a protein digest. Individual peaks are singly charged peptides from a protein digest from *E. coli*. Comparing the observed masses of these peptides with theoretical digestions of the predicted proteins can yield a match to particular open reading frames in a process known as peptide mass fingerprinting (PMF). The inset shows a zoom-in of a single ion, illustrating the high-resolving power of the instrumentation. B. Example spectrum from anESI-QqQ linear ion-trap instrument of a peptide from a protein tryptic digest. Fragmentation is provided by voltage activation and CID with $N_2$ gas. In this case, an m/z $3^+$ ion at 546.96 Da has fragmented to produce the characteristic y and b-ion series. Comparison of the observed ions with those predicted from fragmentations of peptides from theoretical digestions in the databases can yield a match to specific proteins with fewer peptides and greater confidence than for PMF.

Searching and identification of MS/MS data typically is performed by comparison of the data to theoretical fragmentations of genomic *in silico* digests. Matching of fragment spectra uses what is essentially a more complicated version of a peptide mass fingerprint. In some cases, a PMF is performed first to eliminate easy to assign peptides and the tandem data is collected on peptides that do not readily present a PMF match (Zhen *et al.*, 2004).

In most cases, assignment of tandem data is based upon two criteria: error tolerance relative to the precursor or peptide mass, and the % of predicted fragment ions accounted for by the MS/MS spectra. In typical CID spectra, particular ions dominate the fragmentation pattern, typically y and b series fragments (Burlingame *et al.*, 1998; Clauser *et al.*, 1999). b and y ions are the ions that form on either side of the peptide bond. b-ions are to the N-terminus, and y-ions are from the C-terminusThe dominant fragmentation ions are weighted in the scoring algorithm and the aggregate score is presented as a traditional probability, or confidence interval. Probability-based scores describe ID's as if the same data were searched many times, this answer would match the correct 'hit X% of the time. For confidence scores, the data are described as a sufficiently large population of n confidence, n% of those are correct answers.

Two major conclusions can be drawn from several studies utilizing theoretical and empirical data to estimate average false-positive rates.. First, the overall false positive rate from all scoring matrices is high without additional checking and validation, in some cases exceeds 40%. Second, protein identification on the basis of a single tandem peptide is weak and generally results in extremely high rates of false positive identification (Cargile *et al.*, 2004; Mrowka *et al.*, 2001). Have also been described (Resing and Ahn, 2005).

More recently, Steven Carr *et al.* proposed more rigorous guidelines for the presentation of protein identifications from MS and MS/MS data (Carr *et al.*, 2004). The need for these guidelines was two fold: First, a significant number of protein 'hits' being reported in the literature were likely false positives. Second generally accepted criteria are needed for calling an ID based on a combination of high quality data across multiple peptides and searched with database matching software under the correct conditions/parameters. Their guidelines for reporting were as follows:

First, report sufficient information on the nature of the data collection and processing to generate peak lists, including the parameters and software used in searching. Second, the sequence coverage obtained for each protein and number of identified peptides for each protein, including modifications, common or otherwise should be reported. Third, single hit protein assignments must include significant supporting evidence, including charge, m/z, ion-matching score and others. Since many thousands of spectra exist for a proteome, context of which spectra to report in detail is important. For, MS/MS spectra, they should be relevant for identifications generating specific biological conclusions. PMF data should include 'orphan' peptides in addition to peptides that match the described protein. This provides a means to evaluate the false positive rate and a description of the mass accuracy, resolution, calibration means, and contaminant exclusion. Substantial effort should be made to eliminate multiple entries in databases due to overlapping peptide sequences, or isoforms for which no distinguishing evidence exists.

Efforts such as these, developed from a large body of published research and empirical experience should increase our ability to validate database dependent results for protein identification. Studies in bacterial systems typically have a significant advantage here, as elimination and estimation of false positives is easier. Sequences of random proteins can be included to generate false-positive threshold levels for validation of results (Steve Gygi, personal communication,). Despite this, identification of false-positive results is not trivial even in microorganisms.

In systems with excellent biological and biochemical characterization, like *E. coli* assessment of false positives is still difficult. In Corbin *et al.*, a whole cell lysate proteome from *E. coli,* protein hits made on the basis of single peptides were excluded from the general ID pool on the basis of lack of multiple hits, and no additional validation was performed on the data.

Computational approaches to reducing false positive assignments have been described in several places, Peng and Cargile performed what is likely the most straightforward approach and using 'Medusa' databases of reversed protein sequences to co-search with empirical data (Cargile *et al.*, 2004; Peng and Gygi, 2001). By graphing the true negative hits from garbage datasets against the forward (correct) databases, a threshold can easily be established, limiting false positive identifications to an extremely low level, perhaps as low as 1%. This technique comes with considerable cost in the form of false negatives. A hit of low confidence or from small numbers of peptides as compared to comparably scoring matches from the reverse database, are thrown out. But real peptides are likely to match artificial databases to some extent by chance and in cases where an ID had comparable matches to a reverse data set, it would likely be described incorrectly as a false-positive.

## Validation

In a collection of data from a high complexity proteome sample, hits from search programs ostensibly fall into three pools; first, are high confidence hits, which are likely true-positives. These generally require little, if any additional validation. Second, are likely true negatives, which warrant no further attention. Third, are a group which contains both true positives and false positives (true negatives). The focus of validation efforts should be centered on approaches for the confirmation of these data, and determining whether general approaches can be designed from specific instances of known results.

The most reliable means of evaluating identifications is through independent observation from different experimental approaches. There are, however, populations of peptides and potential ID's for which these cutoff approaches are insufficient to establish identity and by inference, presence in the sample. Validation of proteomics results can be thought of as a two-step process: First, are data examination techniques including computational and literature driven approaches. Second, are the orthogonal validations of protein identifications including deference to the underlying biology.

### Data Examination Techniques

Criticism of data collection, reporting procedures and standards vary widely by experiment and group (Baldwin, 2004). In general, these approaches require manual interpretation and designed computer programs to assess identification. As is the case in the work presented here, validation was performed almost exclusively by hand, in both the determination of MS peak lists and database matching from peptide mass fingerprinting. In its entirety, validation was assessed in part by taking advantage of the fact that bacteria are well-suited for generation of data sets which can be reasonably annotated for generation of large-scale computational approaches.

Challenges in even well characterized model systems are still present however. The difficulty in comparing multiple proteome studies can be made by comparing the data obtained from several large-scale proteome studies performed on *E. coli*: The SWISS 2D protein database of *E. coli*, which identified proteins by gel-spot analysis, the 2D non denaturing LC /MS described in this dissertation and a MuDPIT analysis of peptides from whole cell lysates of *E. coli* each identified several hundred proteins (Wolters *et al.*, 2001). Despite the degree under which nearly identical cells were studied, the overlap, and therefore the reproducibility in protein identifications were small, about 30% (Chapter III & Chapter IV). The MIPS database (http://mips.gsf.de/) is one such database that integrates literature for protein-protein interactions as true-positive validation of additional studies. In another genetic approach, Weber *et al*. performed gene-array analysis in *E. coli* to determine a core set of stress genes by comparing expression profiles under several different stress conditions, and arrived at a set of 180 genes which were common to all stress/growth conditions (Weber *et al.*, 2005). In our recent published works (Champion *et al.*, 2003; Marino-Ramirez *et al.*, 2004), we provide comparisons of data generated from multiple sources to serve as both validation and contrast of experimental approaches. A large literature-based approach to mapping multiple identifications in *E. coli* is also housed in EchoBase, from The University of

Table 1.1.) Instances of GlyA by Protein ID in EchoBase.
This table lists all of the entries in Echobase describing protein identification of the glyA gene product from various lysates/preparations of *E. coli*. Listed by year of publication, it also describes the major relevant technique utilized for the resolution/separation and identification.

| Author | Technique |
|---|---|
| VanBogelen et al. 1996 | 2D PAGE & N-term sequencing |
| Link et al. 1997 | 2D PAGE & N-term sequencing |
| Wasinger et al. 1998 | 2D PAGE & N-term sequencing |
| Tonella et al. 1998, 2000 | 2D PAGE & N-term sequencing |
| | MALDI MS & MS/MS sequencing |
| Blankenhorn et al. 1999 | N-term sequencing |
| Fountoulakis et al. 1999 | Chromatography and MALDI MS |
| Champion et al. 2002 | 2D LC, MALDI MS |
| Yan et al. 2002 | 2D DiGE, MALDI MS, MS/MS |
| Birch et al. 2003 | Chromatography and MALDI MS |
| Corbin et al. 2003 | 2D LC, MS/MS sequencing |

York and Glaxo-Smith Kline (http://www.biolws1.york.ac.uk/echobase/). This group examined virtually every ORF from *E. coli* and has generated a list of every experiment that reported its detection under certain conditions. The protein GlyA for example, is annotated in Echobase several dozen times, but was identified by at least nine independent proteome experiments in *E. coli* likely assuring its existence in the cell under the conditions tested. Additionally, several different strains of *E. coli* were used in the studies and differential identification technologies as well. Table 1.1 summarizes the myriad of proteome studies detecting this specific protein, and the methods utilized for identification (Corbin *et al.*, 2003; Fountoulakis *et al.*, 1999a; Lambert *et al.*, 1997; Link *et al.*, 1997; Tonella *et al.*, 2001; VanBogelen *et al.*, 1996; Wasinger and Humphery-Smith, 1998; Yano *et al.*, 2002). The most important detail of these results were the number of fundamentally different approaches Used in the determination. They included Edman sequencing, PMF, tandem MS/MS data, and correlation of the 2D gel migration position to predicted $M_w$ and pI.

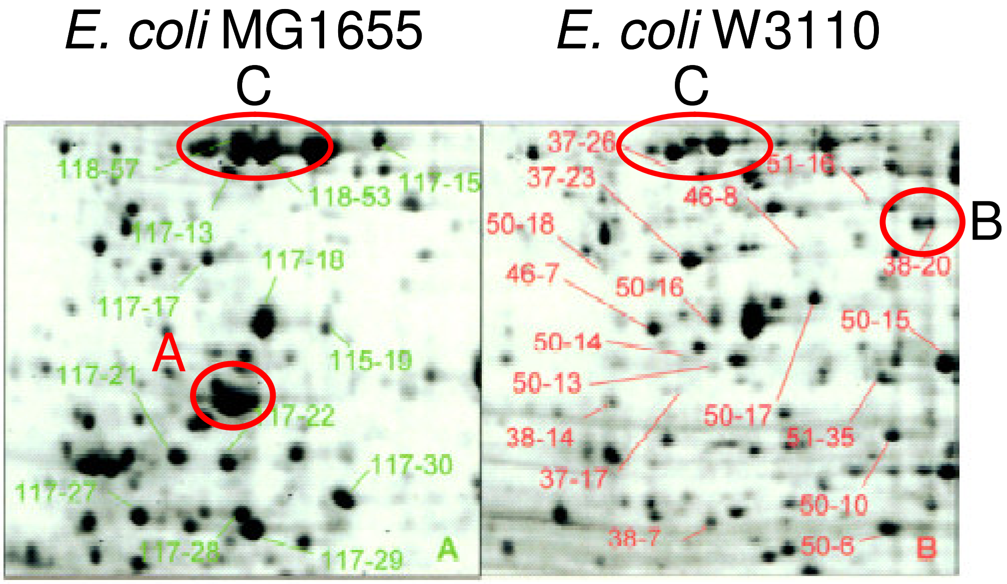
### Orthogonal Validation

Orthogonal validation is important because it takes advantage of experiments performed using multiple analyses and incorporates biological knowledge about specific systems and organisms. The process by which integration of these different sources of validation are incorporated into a 'result' a source of considerable investment. Adequate metrics to describe the individual contributions of individual components are lacking and are a source of considerable investment.

In Chapter II for example, the correlation of the identified proteins to spots observed on 2D gels was performed by hand and helped validate the approaches we used for identification, and enabled visualization by spot intensity of specific changes in protein content independent of MS driven identification. Some computational analysis was available for the alternative LC analysis performed in Chapter IV, but a large amount of post-proteome MS/MS data was acquired on observed 2D gel spots.

Figure 1.5 shows a section of a 2D gel from two different strains of *E. coli* K12 (MG1655) and (W3110) which differ only slightly in genome. Although in general gels between similar strains of bacteria are similar, many spots are different. The spot marked 'A' in Figure 1.5 is essentially absent in the W3110 preparation. The spot cluster marked B indicates a set of spots not present in the MG1655 lysate, and the spot circled C in both gels appears to exhibit differential expression. Even across small sections of the 2D gel these obvious differences in expression and presence of protein spots is obvious, which should give pause to studies in higher organisms where samples and controls are often pooled or aggregated. In particular to this figure, the bacteria were grown using different media. The MG1655 cells were grown in M9 minimal glucose media and *E. coli* strain W3110 was grown in LB (Luria–Bertaini). This also validates observations made about the need for standards in bacterial preparations for physiologic analysis. A more complete understanding of the differences in extremely closely related species of bacteria should provide information about the extent to which variance is a problem overall, improving our recognition of true differences.

The knowledge gained from these experiments is designed to allow us to interpret differences in these organisms when they are exposed to the environment. Biologically, bacteria are ideally suited to further studies in these areas, as extensive knowledge of basic physiology is present in a well-annotated and relatively unmodified proteome.



Reprinted from Lee et al., Biotech. And Bioneng, 2003 84:(7), 12-30, 801-814.

Figure 1.5.) Identical sections from two 2D SDS-PAGE from MG1655 and W3110 *E. coli*. Presented here are two insets of a 2DGE PAGE from *E. coli* MG1655 (Left) and W3110 (Right). Shown in A, B, are spot patterns that differ in apparent presence between the two strains, and C spots that appear up-regulated relative to one another. Reprinted from Lee *et al.,* 2003, (Lee and Lee, 2003).

Many fundamental biological questions have been answered in bacterial systems, and many biological pathways and chemistries are well-conserved even in mammals. In many cases, powerful genetic tools exist to complement and direct the research performed. Repeat experiments and comparisons between data sets are facilitated by these facts and separations can be perfected and incorporated as they develop. Amplifying clonal populations is trivial, enabling validation from repetition and reproducibility. This also generates sample quantities amenable to the detection limits of instrumentation.

Bacterial proteomes also present the possibility of generating corroborating evidence for many biochemical identifications via enzymatic assay, analysis of knockouts, or constitutive expression for validation/confirmation. The ease at which these independent manipulations can be performed enhances the degree to which proteomics approaches and technologies will be performed on microorganisms, generating testable data. For bacteria, a goal of validation should be to generate large and descriptive enough data sets that are consistent enough with the underlying biology such that novel findings can be believed, and not dismissed as false positives or 'noise.'

My efforts in the following dissertation were to expand our understanding of the protein content in *E. coli* under multiple conditions by combining several biochemical techniques, with protein identification by biological mass spectrometry, (MALDI TOF and MALDI TOF-TOF). Additionally, I wanted to apply these novel techniques to gain unique insight into the growth transition, not merely repeat identification of proteins already known to be expressed or repressed during stationary phase. We were able to implement several instrumental and processing improvements to the generation of the data, eventually describing one of the largest whole-cell proteomes in *E. coli* to date.

One of the most important pieces of information not currently available from gene arrays, or even large-scale protein interaction mapping studies (Gavin *et al.*, 2002; Marino-Ramirez *et al.*, 2004) are the differences in protein association and localization as a function of cell state, an observation reported here with implications for our continued experiments and understanding of the global rearrangement and changes in expression when organisms experience different environments.

# CHAPTER II

# PROTEOME ANALYSIS OF *Escherichia coli* BY NATIVE STATE CHROMATOGRAPHY AND MALDI MS UNDER EXPONENTIAL GROWTH CONDITIONS[*]

## Summary

To identify proteins expressed in *E. coli* K-12 MG1655 during exponential growth in defined medium, we separated soluble proteins of *E. coli* over two dimensions of native state high performance liquid chromatography, and examined the components of the protein mixtures in each of 380 fractions by peptide mass fingerprinting. To date, we have identified the products of 310 genes covering a wide range of cellular functions. Validation of protein assignments was made by comparing the assignments of proteins to specific first-dimension fractions to proteins visualized by 2-D gel electrophoresis. Co-fractionation of proteins suggests the possible identities of components of multiprotein complexes. This approach which can yieldyields high-throughput gel-independent identification of proteins or canproteins. It can also be used to assign identities to spots visualized by 2-D gels, and should be useful to evaluate differences in expressed proteome content and protein complexes among strains or between different physiological states.

[*] Portions of this chapter have been reprinted with permission from Proteome analysis of *Escherichia coli* K-12 by two-dimensional native state chromatography and MALDI-MS by Champion MM, Campbell CS, Siegele DA, Russell DH and Hu JC, 2003. *Molecular Microbiology,*47(2), 383-396. ©2003 Blackwell Publishing Ltd.

## Introduction

Partial proteomes have now been mapped for several microorganisms, and to date, they represent a large body of biological data for the proteins present within an organism (Washburn and Yates, 2000). Other efforts are underway to characterize proteomes under different growth or stress conditions (Liu *et al.*, 2005; VanBogelen, 1999; VanBogelen *et al.*, 1999a; VanBogelen *et al.*, 1999b). Our efforts were to map the proteome of *E. coli* in as much detail as possible, validate the assignments made after separation and identification with biological mass spectrometry, and analyze the content of these data. In order to map and identify the expressed proteins of *E. coli* grown to exponential phase we separated whole-cell lysates over two dimensions of non-denaturing chromatography. Each of the fractions from this separation were subjected to tryptic digestion and analysis by high resolution MALDI-DE-R-TOF mass spectrometry. The individual peptides were assigned to proteins by peptide mass fingerprinting and each proteome was performed twice at two different pH's. In this analysis we identified 2012 proteins from *E. coli* corresponding to a non redundant proteome of 310 unique assignments. Proteins are recovered from every predicted functional annotation class and cover a wide range of predicted abundances, but are biased towards moderate or highly expressed proteins. Co-fractionation of proteins at multiple pH's via the non-denaturing chromatography can also suggest the potential partners in multi-protein complexes. Pair-wise analysis of proteins that cofractionate at multiple pH's indicates at least 125 such protein pairs in these proteomes.

## Results

Our general approach is shown in Figure 2.1A. Whole-cell lysates are fractionated over two dimensions of native-state HPLC, a strong anion exchange column (AIX), followed by a second separation on a hydrophobic interaction resin (HIC). Proteins in each fraction are then digested with trypsin and identified from the masses of tryptic fragments, which are determined by MALDI-TOF mass spectrometry.

Figures 2.1B and 2.1C show the separation of the clarified crude lysate of *E. coli*. For the first dimension anion exchange, we utilized a shallow, segmented salt gradient to distribute the proteins roughly equally over 19 protein-containing fractions plus flow-through (Figure 2.1B). About 20% of the total protein by weight is in the flow-through, which was processed separately (see below). The large peak of UV-absorbing material in fractions 22-25 contains primarily nucleic acids (data not shown). Figure 2.1C shows the elution profile for a typical second dimension separation of one of the 19 anion-exchange fractions after HIC. The two dimensions of chromatography separate the soluble proteins into 380 fractions. Many of these individual fractions contain 5-20 proteins visible by silver staining of 1-D SDS gels (data not shown). The separation was performed four times, using lysates from independent cultures. Two different pH conditions (pH 7.50 and pH 8.75) were used for the anion exchange step and two lysates were processed at each pH.

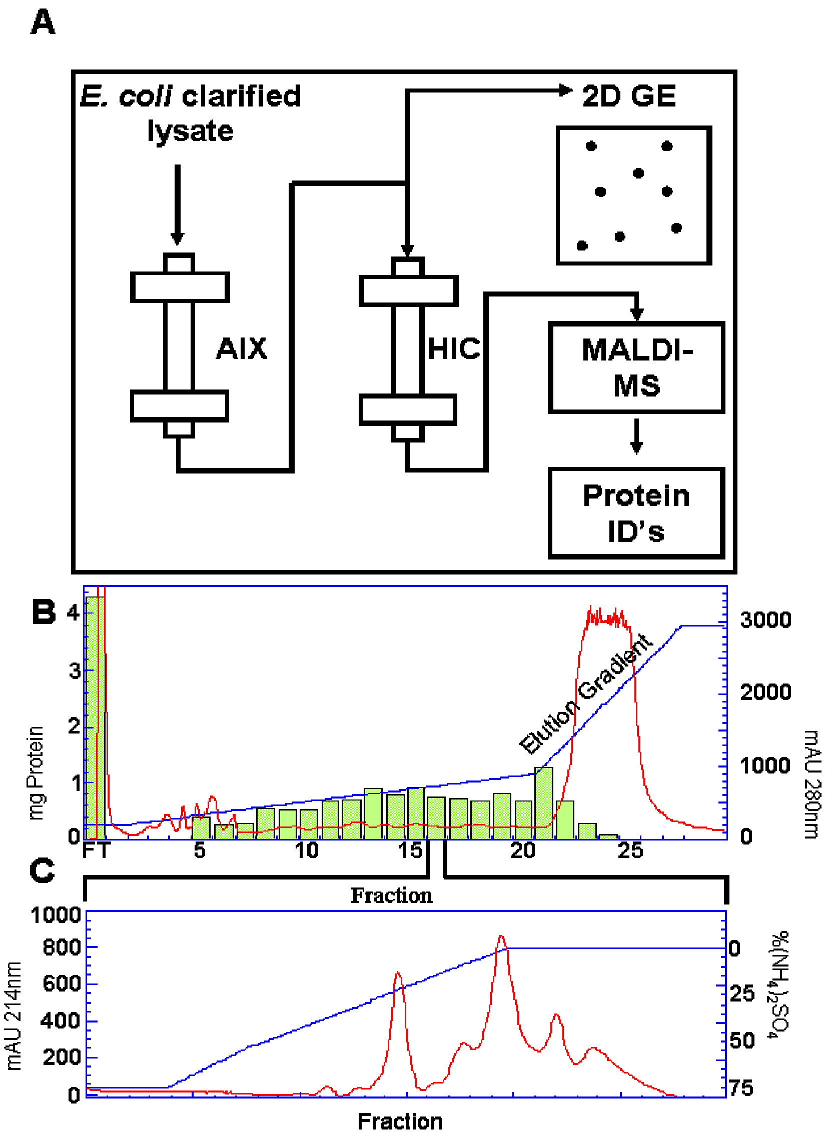


Figure 2.1.) Proteomics by native-state LC/MS. A.) Flowchart. Clarified crude lysates of *E. coli* MG 1655 first separated over an anion exchange (AIX) column, collected into 20 fractions, either run on a 2-D PAGE or separated over a hydrophobic interaction resin (HIC). These fractions are digested and identified using MALDI-DE-R-TOF MS and peptide mass fingerprinting. B) Chromatogram of 1$^{st}$ dimension separations. Typical chromatogram trace for cell lysates separated on SOURCE 15Q anion-exchange column (pH 7.50). Traces show UV 280nm absorbing material, gradient (NaCl) and the bars quantitate the protein in each fraction by Bradford assay. C) Typical chromatogram for 2nd dimension fractionation, in this case, of AIX fraction #16.

## Proteins from the Ion-exchange Flow-through

Approximately 20% of the protein by mass flowed through the ion-exchange column. As this is the only fraction that contains such a high amount of protein, we suspected that 30S and 50S ribosomal subunits were in the ion-exchange flow through. Consistent with this possibility, fractionation of the flow-though via a Superose 12 size-exclusion column revealed RNA and abundant proteins in the void volume consistent with the presence of a ribonucleoprotein complex > 300,000 MW (data not shown). MALDI-MS and peptide mass fingerprinting identified several ribosomal proteins in theion-exchange flow-through refractionated by cation-exchange chromatography or SDS-PAGE (data not shown). Tandem mass-spectrometry performed on trypsinized ion-exchange flow-through using a Thermo Finnigan LCQ Deca identified an additional 18 ribosomal proteins from peptide sequences (data not shown) and no attempts were made to identify the remaining proteins in the ion-exchange flow-through in this chapter.

## Protein Identification

The masses of tryptic peptides from digestion of each fraction were determined by Matrix-Assisted Laser Desorption Ionization-Delayed Extraction Reflectron-Time-of-Flight mass spectrometry (MALDI-DE-R-TOF) as described in the Experimental Procedures. Figure 2.2 shows a MALDI-DE-R-TOF spectrum from one of the 380 fractions. In total, nearly 2,000 spectra were collected and annotated for their peptide masses. The average mass error (m/z) for each fraction was 20 ppm, with a standard deviation of 20 ppm.

Proteins in each fraction were identified by recursive matching of observed peptide masses from tryptic digests to peptides predicted in the *E. coli* genome as described in the Experimental Procedures. Table 2.1 summarizes the number of proteins found in each lysate, and the overlap between experiments. Overall, 2,012 proteins were identified, corresponding to a nonredundant set of 310 gene products. A full list of the protein identities and the fractions where they are found is available as supplementary data.

## Comparison with Proteins Observed by 2-D PAGE

To test the validity of our protein assignments we compared the proteins identified in each AIX fraction to the proteins observed by 2-D PAGE. Figure 2.3 shows 2-D gels for the 19 fractions from the first dimension of chromatography. To generate a list of proteins we expect to see on each gel, we merged assignments made from MS data of second dimension (HIC) fractions for each of the AIX-fractions. We then examined the appropriate 2-D gel for a spot migrating at the expected MW and pI. Since many proteins seen on 2-D gels migrate at positions that differ significantly from their predicted positions (Link *et al.*, 1997) we used published and indexed 2-D gel maps of *E. coli* to identify spots wherever possible (Hoogland *et al.*, 2000; Tonella *et al.*, 1998; Tonella *et al.*, 2001). By combining predicted spot positions with known spot migrations, we can examine the correlation between our protein assignments and spots that can be identified on the 2-D gels. Figure 2.4 shows one example of an annotated 2-D gel. In this case, we can correlate 16 of the 17 proteins we identify with spots on the 2-D gel. For this gel, only one protein identified from the MS data, GreA, did not match a corresponding spot. Figure 2.4 indicates the predicted and actual migration of GreA as

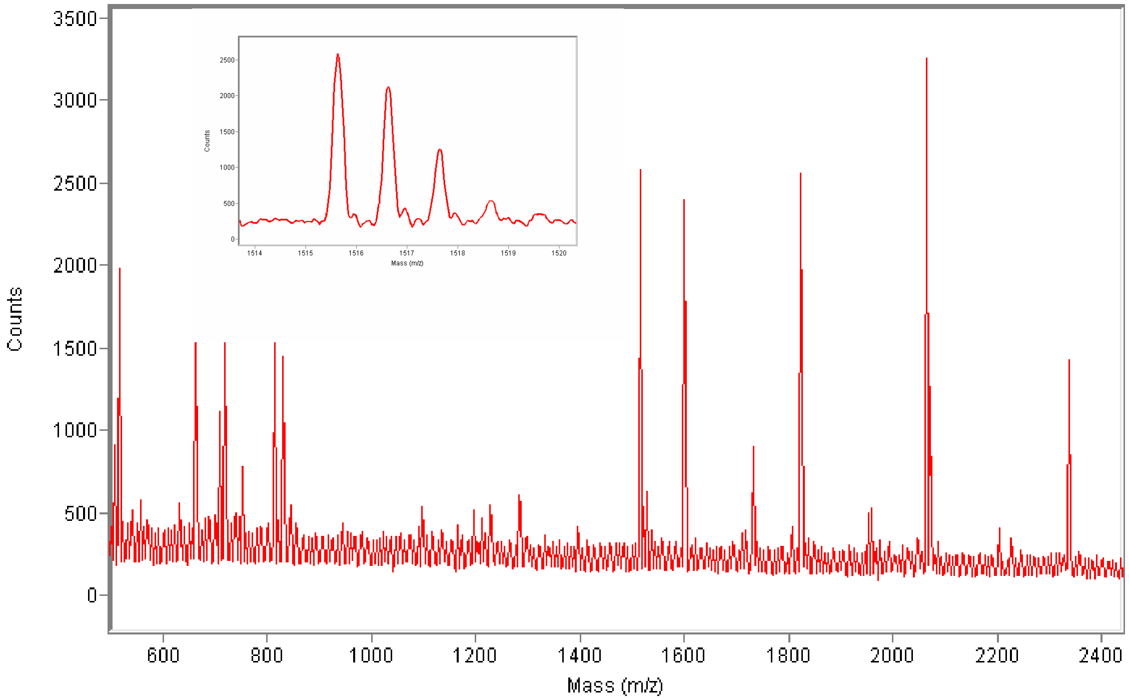


Figure 2.2.) Typical MALDI-DE-R-TOF spectra for protein digest from a 2nd dimension column fraction. MALDI-MS was performed on column fractions, peak annotation was done using GRAMS 32 software, and data exported for peptide mass fingerprinting with MS-FIT. The inset shows the well-resolved mono-isotopic distribution of an individual tryptic fragment in this spectra.

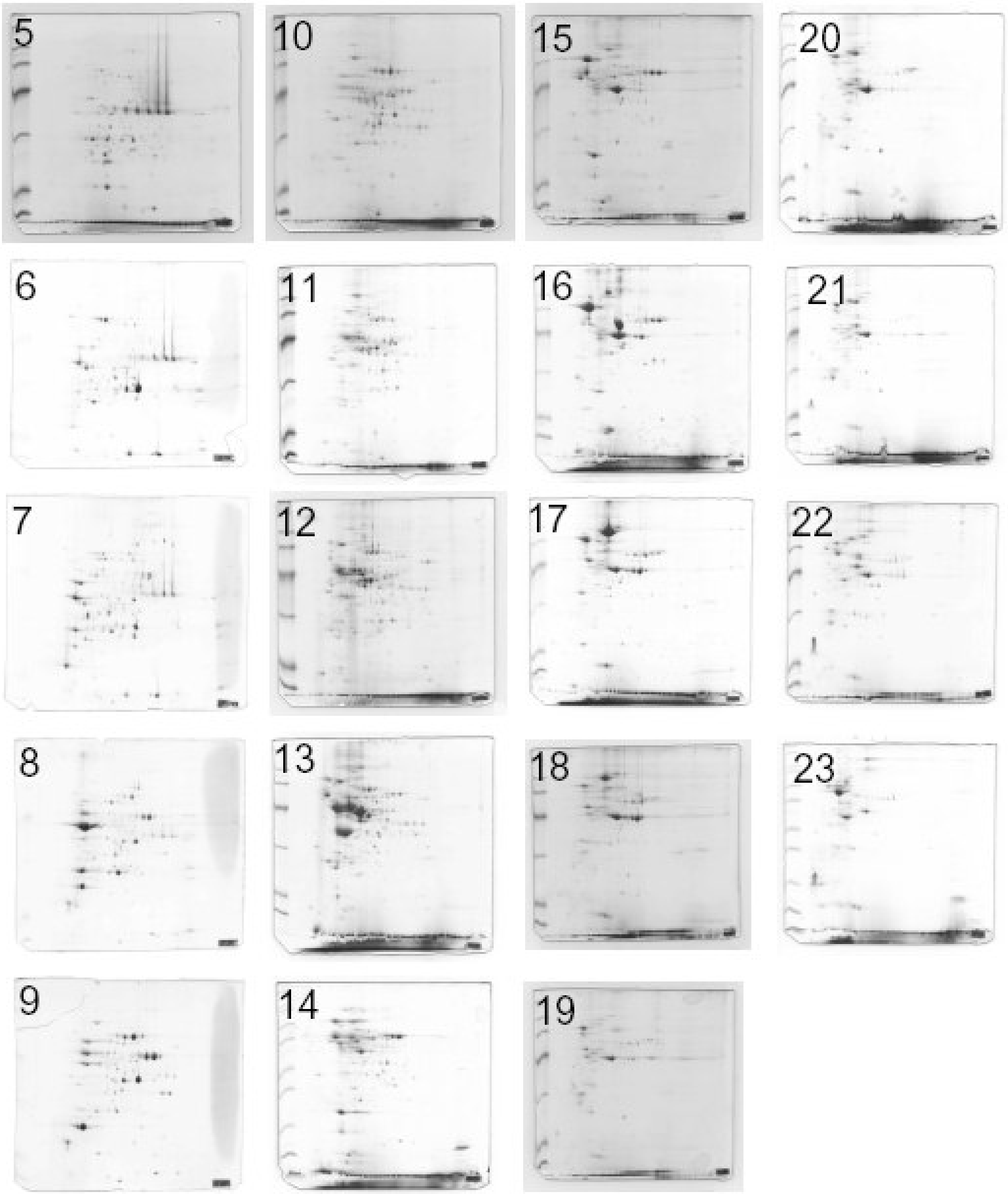


Figure 2.3.) Two dimensional gels of anion exchange (AIX) (SOURCE 15Q) fractions. Approximately 300µg pf protein from each AIX fraction was subjected to 2D PAGE analysis and stained with Coomassie Blue. Isoelectric focusing was done using IPG strips with a non-linear pH 3-11 gradient. Gels are numbered by the AIX fractions that were run.

Table 2.1.) Protein identification totals from *E. coli* lysates.
Each pH was performed twice and totals are listed above. Total number of proteins identified includes the same proteins found in multiple fractions, unique ID's are non-redundant totals for each experiment. The total of 310 identifications is the total of all unique ID's from all 4 proteomes, with redundant entries removed.

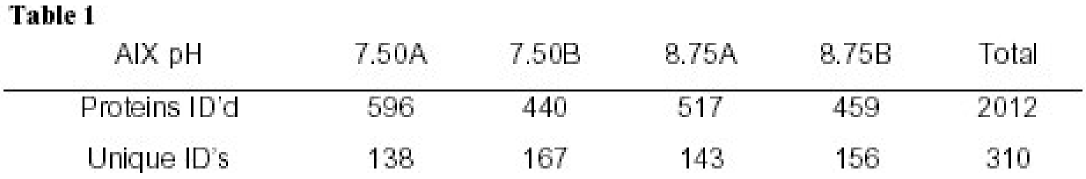
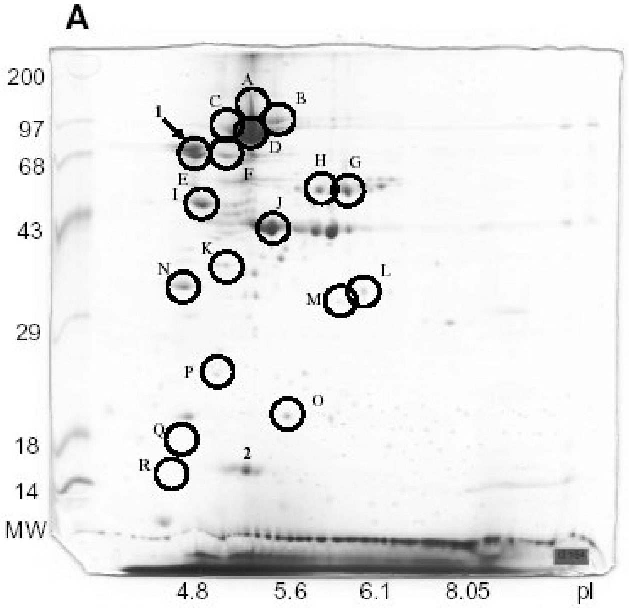
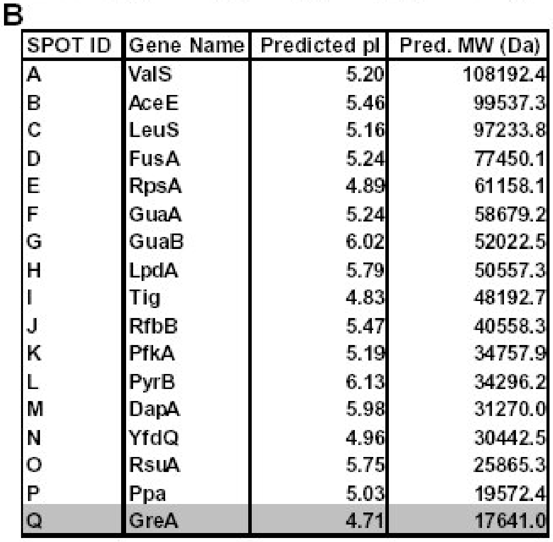


**Table 1**

| AIX pH | 7.50A | 7.50B | 8.75A | 8.75B | Total |
|---|---|---|---|---|---|
| Proteins ID'd | 596 | 440 | 517 | 459 | 2012 |
| Unique ID's | 138 | 167 | 143 | 156 | 310 |





B

| SPOT ID | Gene Name | Predicted pI | Pred. MW (Da) |
|---|---|---|---|
| A | ValS | 5.20 | 108192.4 |
| B | AceE | 5.46 | 99537.3 |
| C | LeuS | 5.16 | 97233.8 |
| D | FusA | 5.24 | 77450.1 |
| E | RpsA | 4.89 | 61158.1 |
| F | GuaA | 5.24 | 58679.2 |
| G | GuaB | 6.02 | 52022.5 |
| H | LpdA | 5.79 | 50557.3 |
| I | Tig | 4.83 | 48192.7 |
| J | RfbB | 5.47 | 40558.3 |
| K | PfkA | 5.19 | 34757.9 |
| L | PyrB | 6.13 | 34296.2 |
| M | DapA | 5.98 | 31270.0 |
| N | YfdQ | 4.96 | 30442.5 |
| O | RsuA | 5.75 | 25865.3 |
| P | Ppa | 5.03 | 19572.4 |
| Q | GreA | 4.71 | 17641.0 |

Figure 2.4.) Comparison of proteins identified by LC/LC MS and by 2-D gels. A) Annotated 2D gel of AIX (SOURCE15Q) fraction 17. B) Identities and predicted pI and MW for proteins expected to be in this fraction based on peptide mass fingerprinting of HIC fractions from AIX fraction 17. Spot ID Q, highlighted in grey, shows the predicted migration of GreA, which is not visible on this gel.

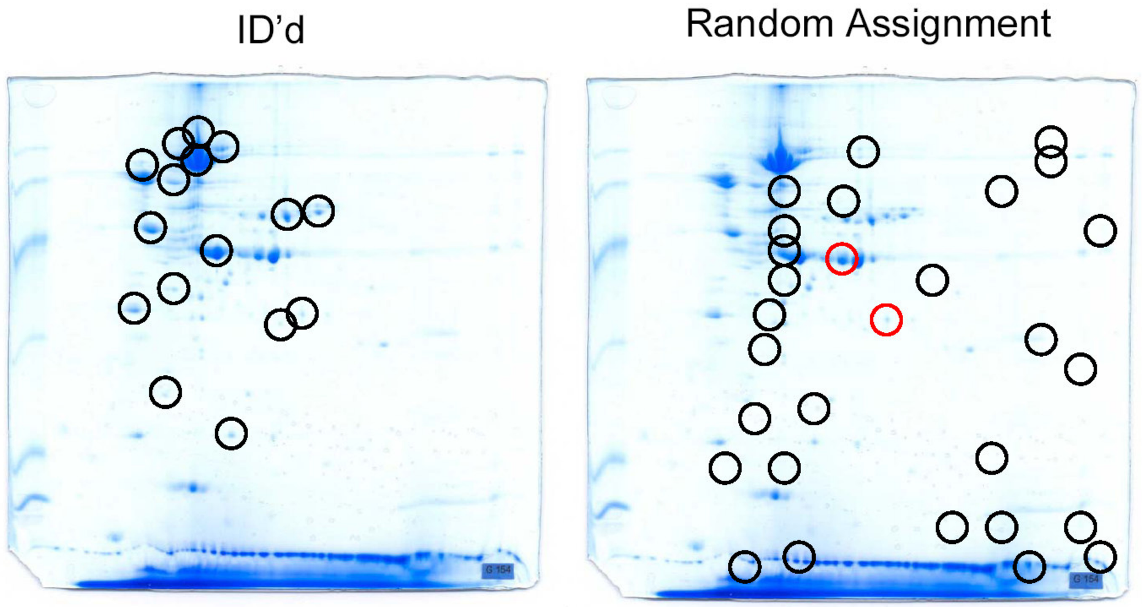


Figure 2.5.) Comparison of proteins assigned to gel vs. random assignment. This is an identical 2D gel from the exponential phase cell growth annotated with the pI and MW ofproteins assigned to its fraction from 2D LC MALDI, vs. a random assignment of the pI and MWof proteins from the entire list of identified proteins. Random spots which matched actual spots on the gel are highlighted in red.

<Q> and <R> respectively. As a control, we randomly selected bins of 30 proteins from our complete list of 310 identified proteins, and attempted to match them to the positions of the observed spots for several gels. An example of this matching is shown as Figure 2.5. Only two of these randomly selected proteins matched a spot at the appropriate MW and pI; one of them was in the list of expected proteins identified in the corresponding fraction. Thus, the correlation between the proteins identified by MS and those observed on the gel is much better than would be expected by chance.

In addition to the spots identified with MS correlation, all of the gels resolved spots with identities that could not be assigned from MS data for that fraction. Some of these are clearly multiple spots produced from the same protein; indeed, some of these are annotated in databases of *E. coli* proteins identified by 2-D gels. In other cases, we can make assignments when the same spot is seen in gels from a series of contiguous fractions, reflecting the changing abundance of each protein as it elutes from the ion exchange column. Although the protein might not be identified by MS in one fraction, it could be identified in one or more of the contiguous fractions. For example, we can see the spot marked 1 in Figure 2.4 on gels from AIX fractions 9-13. In AIX fraction 17, this spot is unidentified, but in AIX fractions 18, 20, 21 and 22 it is identified as DnaK, the major Hsp70 homolog in *E. coli*. We also observe several spots such as spot 2, which could not be identified unambiguously by either MALDI-MS or comparison with published gel annotations. Of the 219 unique proteins we identify at pH 7.50, we can assign spots on the 2D gels for 109 of them (57%). Of these, 41 (38%) were not previously annotated in the SWISS-2D database.

## Classification of Proteins Identified by Function, pI/MW, and Abundance

To determine whether our method is biased toward or against particular kinds of proteins, we compared the proteins found in *E. coli* fractions to the different gene classes defined by Blattner et al. and the Riley lab web page (A perl script designed by a summer student, Fouad Kahn enabled scripting of proteome function utilized in Figure 2.6). We find proteins predicted to be in all of the functional classes, but fewer proteins annotated as membrane proteins (transport, cell structure) are seen then would be expected in a random sampling. This is likely a consequence of how we prepared our samples, which requires that our proteins remain soluble. We recover a greater fraction of proteins involved in metabolism, which probably reflects abundance more than gene function.

We also examined the predicted pI and molecular weights of the proteins we identified and compared them to the distribution of pI and molecular weights of all of the annotated ORFs in the *E. coli* genome (Figure 2.7). For comparison, we examined the distribution of proteins seen in the SWISS-2DPAGE database. SWISS-2DPAGE identified very few proteins for pI ranges above 7.0. The proteins we observe cover the whole pI range observed for the genome, with a slight bias toward proteins with pIs between 4 and 6. This may reflect the pIs of proteins in the optimal separation range for the anion exchange step, and/or it could be a consequence of a bias against very basic proteins, which would tend to be in the flowthrough of the anion exchange column. Only 18.3% of the proteins we identified have a predicted pI above 7.0, while the expected frequency for the genome is 35.5%. 2-D gel data has an expected bias toward proteins that resolve well by isoelectric focusing. Only 7.7% of the SWISS-2D identifications have a predicted pI above 7.0. Both methods mirror the genomic distribution in

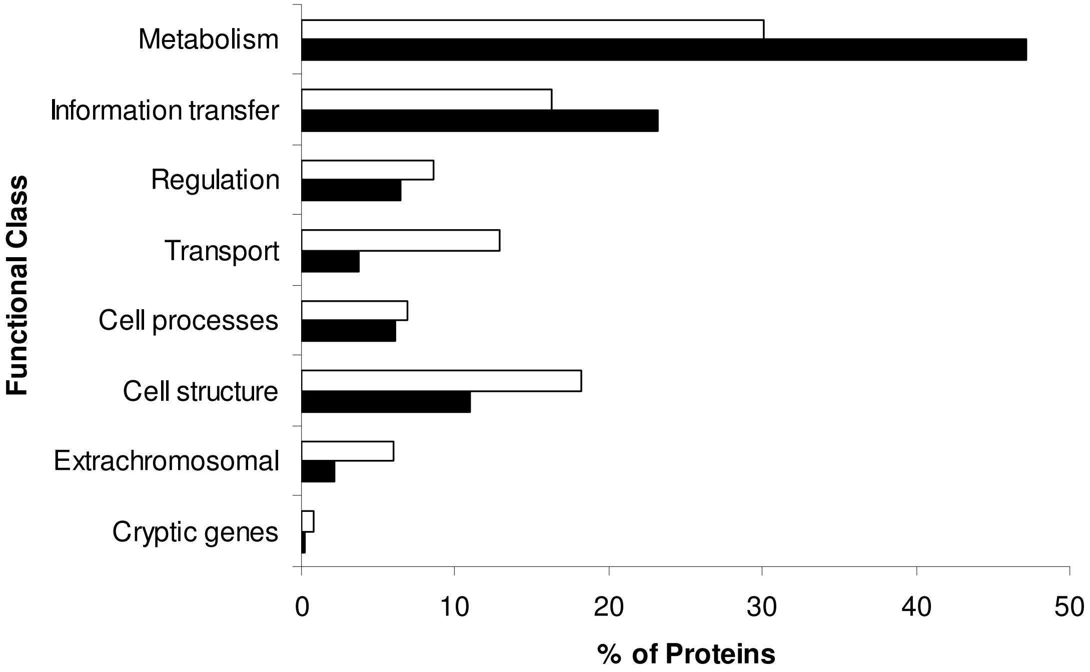


Figure 2.6.) Classification of identified proteins by function. Functional classification categories are from Blattner et al. (1997), and Riley et al. (http://genprotec.mbl.edu/start). Open and filled bars show the % of the genome (Open) and the proteins identified in this study, (Filled) respectively, assigned to each functional class. Although MG1655 does not contain any plasmids, extrachromosomal genes include prophage genes. The Cryptic category includes 43 genes not predicted to be expressed. Our single 'hit' of a translated cryptic gene was hofB, which is a putative transport protein.

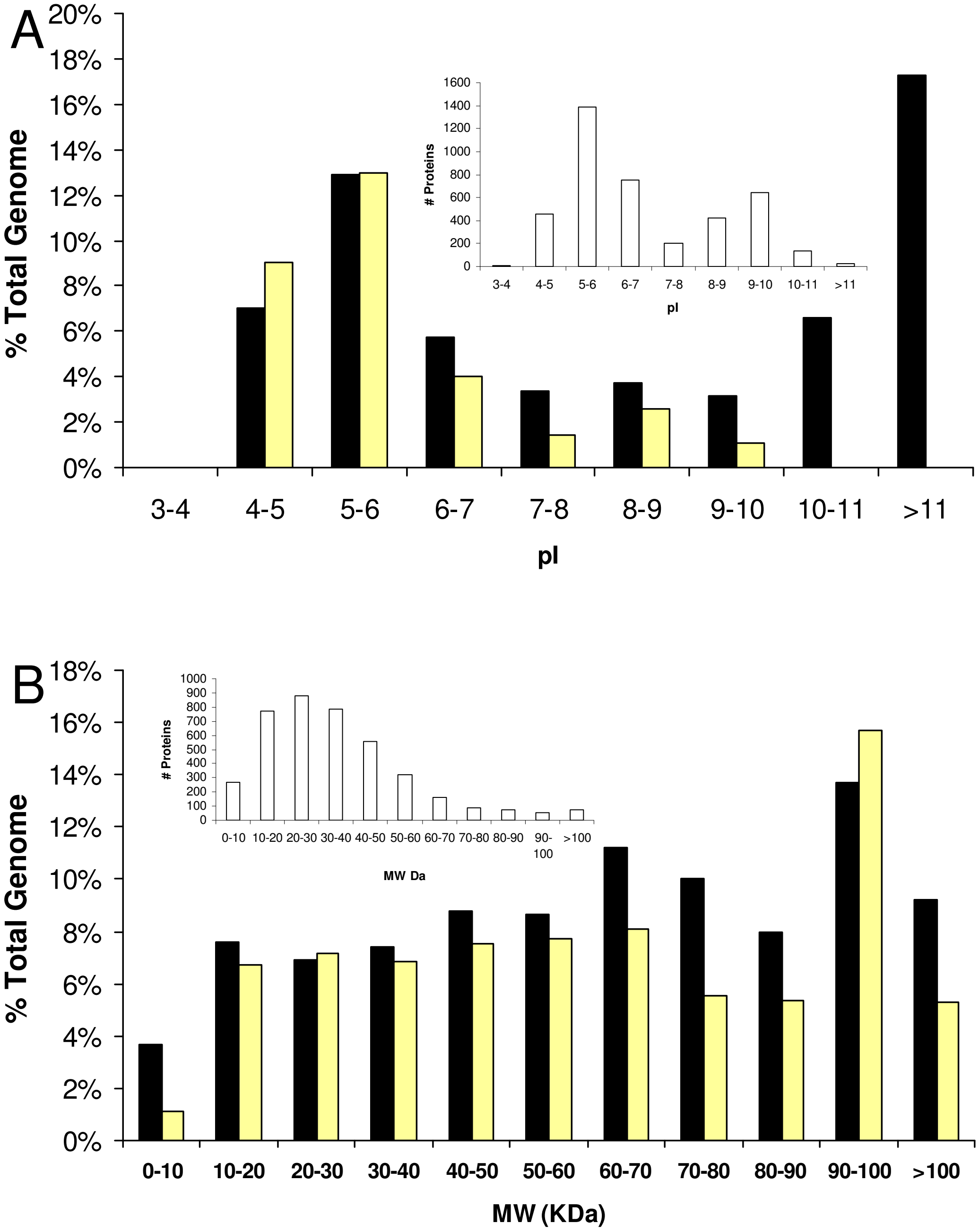


Figure 2.7.) pI and molecular weight distributions for expressed proteomes. A) distribution of the predicted pI's, and B) distribution of the predicted molecular weights of the identified proteins for annotated ORF's from the genome sequence of MG1655 (Inset), the proteins identified in this study (Black ), and the annotated SWISS-2D database for *E. coli* (Grey). Each bar shows the number of proteins identified as a fraction of the number of annotated proteins from the complete genome in that pI or MW range.

molecular weight, and recover very few small peptides/proteins. This partially reflects the fact that smaller proteins, on average, have fewer diagnostic tryptic peptides than larger proteins. As expected from the gel-independence of our methods, we identify more low-molecular weight proteins that do not resolve well on conventional SDS-PAGE.

All proteome methods to identify expressed proteins are biased toward those that are abundant. Since direct measurements of abundance are not available for most *E. coli* proteins, we used two criteria to evaluate the correlation between our protein identifications and the actual abundance of the proteins. First, we examined the overlap between our identifications and the proteins for which synthetic rates have been measured by pulse-labelling and 2-D electrophoresis. In *E. coli* strain W3110 grown in minimal MOPS + glucose, under similar growth phase conditions, 51 proteins were indexed by Neidhardt et al. (VanBogelen *et al.*, 1996). All 51 of these were identified in our study. Assuming that the quantitation made from pulse labeled cells approximates the steady-state levels of proteins, from their data we detect proteins with the lowest stated abundance of ≥0.2% of total protein, or about 500 copies per cell. This agrees well with reconstruction experiments we performed using known amounts of beta-galactosidase spiked into column fractions, where we estimate sensitivity of about 250 copies/cell (data not shown). However, our absolute detection sensitivity is more limited by ion-supression of more abundant peptides and low-abundance proteins diluted over multiple fractions than lack of ability to recover low-abundance proteins.

We also examined the expression levels of each protein predicted from sequence analysis. Mrázek and Karlin (Karlin and Mrazek, 2000; Karlin *et al.*, 2001) described an

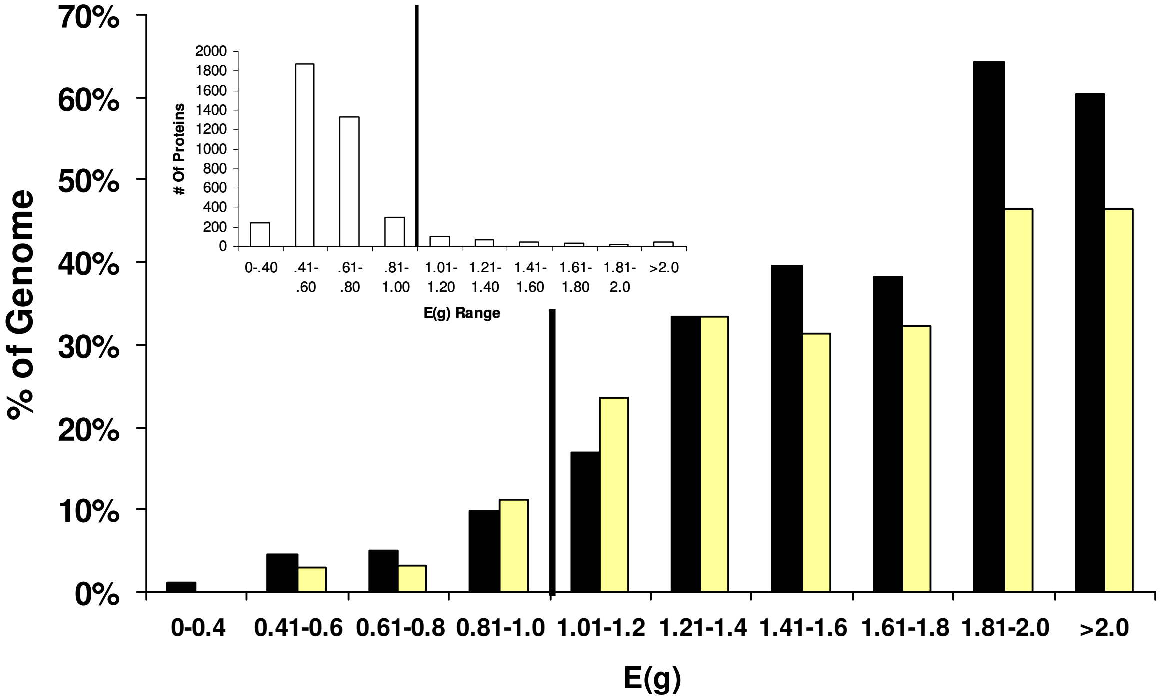


Figure 2.8.) Predicted expression levels for identified proteins. E(g) value distributions were determined for the proteins identified in this study (Black), and the entire annotated SWISS 2DPAGE for *E. coli* (Grey) and the MG1655 genome (Inset). Vertical Black lines indicate the cutoff for genes that are in the PHX class (**P**redicted **H**ighly **E**xpressed).

algorithm to predict abundance based on comparing codon usage of a gene of interest to that of several abundant protein classes, including ribosomal proteins, chaperonins, and translation factors. The predicted expression level is expressed as an E(g) ratio, where values greater than 1 are considered to be 'Predicted Highly Expressed' or (PHX). Figure 2.8 compares the distribution of E(g) values for the proteins we identified to the values for all of the annotated genes from E. coli K-12. 39% of our identifications fall within the PHX class, while only 8% of the proteins in the genome are predicted to have an E(g) value >1.0. Proteins identified in the SWISS-2DPAGE gel database have a similar distribution as the proteins we identify by our method - 45% of the SWISS 2D identifications have E(g) values greater than 1.0. We seem to find more proteins from the lower expression classes.

Correlation with Gene Expression Assayed by DNA Microarrays

We examined the correlation between the 310 proteins we identified and gene expression by performing microarray experiments to identify expressed mRNAs. In total mRNA prepared from cells 3 independent cultures grown under identical conditions to those used for protein samples, we observed 3860 genes were expressed at 2-fold >4 SD above over the background in all three cultures. In published array experiments, expression of similar comparable numbers of genes is observed in exponential phase cultures of the same *E. coli* strain also growing in minimal glucose medium (Courcelle *et al.*, 2001; Tao *et al.*, 1999; Wei *et al.*, 2001). Hybridization was observed for 94% (290 of 310) of the genes encoding the proteins we observed. This is consistent with the large number of high-abundance genes we identify, the array obviously also has a few thousand transcripts for genes not seen in the proteome. This is expected, as the array samples the entire transcriptome space, and our proteome study is non-saturating, even when combined with the entire SWISS 2D. This is a trend, not reversed in even more advanced and through studies (Liu *et al.*, 2005; Tani *et al.*, 2002).

## Co-fractionation of Native Complexes

Because multiprotein complexes should remain intact through both chromatographic dimensions, it may be possible to identify protein complexes by analyzing chromatographic co-fractionation of subunits. The general idea is to apply a ‘guilt by association’ analysis to our entire proteome snapshot. In ten cases, gene names suggest that two or more cofractionating proteins share a common function. However, simply examining the cofractionation over two columns is likely to generate a very high background of false positives. Proteins in the same fraction could cofractionate because they are physically associated or because they just happen to fractionate similarly. In traditional purification protocols, coincidental cofractionation is reduced by either increasing the specificity of purification steps (e.g. affinity chromatography) or adding more steps to the purification (additional chromatographic steps).

Instead of adding additional purifications steps to the separation, we performed parallel separations in which the pH of the buffers used in the anion exchange step was changed. "pH scouting" is often used to optimize ion exchange separations and is based on how titration of ionizable surface groups on the protein alters their elution positions. At either pH, stable complexes will co-elute from the anion exchange column, while proteins in the same fraction by coincidence are free to migrate elsewhere, depending on their individual chromatographic properties.

This is illustrated by PheS and PheT, the α and β subunits, respectively, of an $\alpha_2\beta_2$ heterotetrameric tRNA charging enzyme. PheS and PheT cofractionate at both pH 7.50 and pH 8.75. At pH 7.50, eight other proteins are found in the same fractions as PheS and PheT: AccA, AceE, AsnS, GltB, GroES, RfbB, RpsA, and Tig. At pH of 8.75 RplJ and Tig are found cofractionating with PheS and PheT. Since Tig is an abundant chaperonin, it is likely that its interaction with PheS and PheT is nonspecific or coincidental.

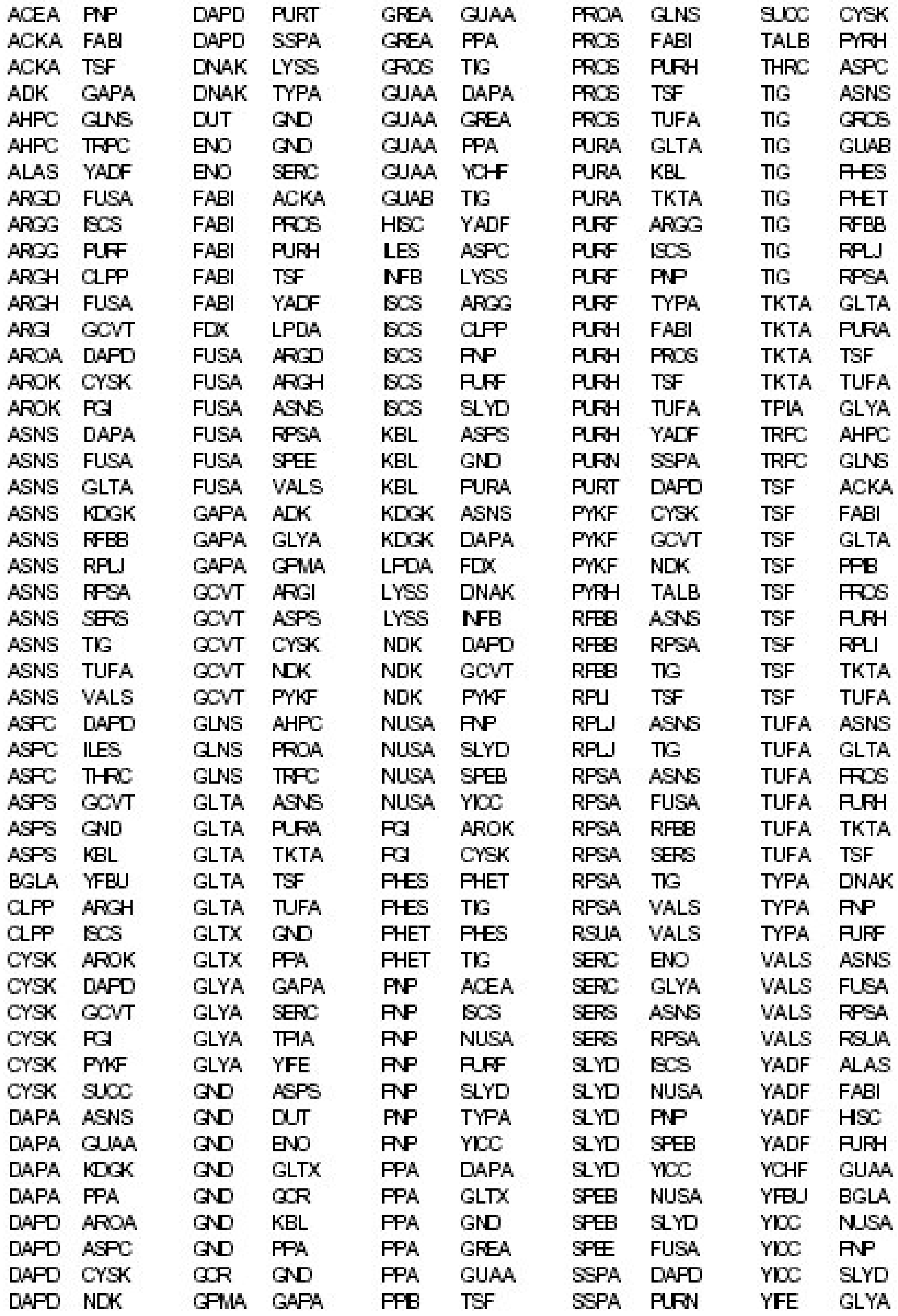
By applying this analysis to all of the proteins we identified 125 pairs of proteins that cofractionated at both pH 7.50 and pH 8.75 (Table 2.2). These potential interactions include several, like PheS and PheT that have been previously described or that seem plausible from functional annotations. This is clearly an underestimate of the stable complexes; knowncomplexes: Known complexes such as RNA core polymerase $\alpha_2\beta\beta'$ and the ClpX, ClpP *E. coli* proteosome were identified as co-fractionating in only one sample, or only at one pH.

## Materials and Methods

### *E. coli* Lysates

*E. coli* K-12 strain MG1655 (Blattner *et al.*, 1997) was grown overnight in M9 minimal medium (Miller, 1972) containing glucose (0.4%), uridine (50μg/mL), $CaCl_2$ (100μM), $MgSO_4$ (2mM). 1L cultures of the same medium + (0.1% w/v) Casamino Acids (Difco) were inoculated with 10ml of the overnight and grown to mid-log (OD 600 =0.5). Cells were harvested by centrifugation at 4,000xg for 20' in a JA-10 rotor (Beckman) and washed by resuspension in 20mM Tris Cl, 20mM NaCl, 1mM EDTA, pH

Table 2.2.) Proteins that cofractionate at both pH 7.50 and pH 8.75.
The 125 pairs are shown as 250 entries; each pair is listed with each partner first to aid finding proteins of interest.



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ACEA | PNP | DAPD | PURT | GREA | GUAA | PROA | GLNS | SUCC | CYSK |
| ACKA | FABI | DAPD | SSPA | GREA | PPA | PROS | FABI | TALB | PYRH |
| ACKA | TSF | DNAK | LYSS | GROS | TIG | PROS | PURH | THRC | ASPC |
| ADK | GAPA | DNAK | TYPA | GUAA | DAPA | PROS | TSF | TIG | ASNS |
| AHPC | GLNS | DUT | GND | GUAA | GREA | PROS | TUFA | TIG | GROS |
| AHPC | TRPC | ENO | GND | GUAA | PPA | PURA | GLTA | TIG | GUAB |
| ALAS | YADF | ENO | SERC | GUAA | YCHF | PURA | KBL | TIG | PHES |
| ARGD | FUSA | FABI | ACKA | GUAB | TIG | PURA | TKTA | TIG | PHET |
| ARGG | ISCS | FABI | PROS | HISC | YADF | PURF | ARGG | TIG | RFBB |
| ARGG | PURF | FABI | PURH | ILES | ASPC | PURF | ISCS | TIG | RPLJ |
| ARGH | CLPP | FABI | TSF | INFB | LYSS | PURF | PNP | TIG | RPSA |
| ARGH | FUSA | FABI | YADF | ISCS | ARGG | PURF | TYPA | TKTA | GLTA |
| ARGI | GCVT | FDX | LPDA | ISCS | CLPP | PURH | FABI | TKTA | PURA |
| AROA | DAPD | FUSA | ARGD | ISCS | PNP | PURH | PROS | TKTA | TSF |
| AROK | CYSK | FUSA | ARGH | ISCS | PURF | PURH | TSF | TKTA | TUFA |
| AROK | PGI | FUSA | ASNS | ISCS | SLYD | PURH | TUFA | TPIA | GLYA |
| ASNS | DAPA | FUSA | RPSA | KBL | ASPS | PURH | YADF | TRPC | AHPC |
| ASNS | FUSA | FUSA | SPEE | KBL | GND | PURN | SSPA | TRPC | GLNS |
| ASNS | GLTA | FUSA | VALS | KBL | PURA | PURT | DAPD | TSF | ACKA |
| ASNS | KDGK | GAPA | ADK | KDGK | ASNS | PYKF | CYSK | TSF | FABI |
| ASNS | RFBB | GAPA | GLYA | KDGK | DAPA | PYKF | GCVT | TSF | GLTA |
| ASNS | RPLJ | GAPA | GPMA | LPDA | FDX | PYKF | NDK | TSF | PPIB |
| ASNS | RPSA | GCVT | ARGI | LYSS | DNAK | PYRH | TALB | TSF | PROS |
| ASNS | SERS | GCVT | ASPS | LYSS | INFB | RFBB | ASNS | TSF | PURH |
| ASNS | TIG | GCVT | CYSK | NDK | DAPD | RFBB | RPSA | TSF | RPLI |
| ASNS | TUFA | GCVT | NDK | NDK | GCVT | RFBB | TIG | TSF | TKTA |
| ASNS | VALS | GCVT | PYKF | NDK | PYKF | RPLI | TSF | TSF | TUFA |
| ASPC | DAPD | GLNS | AHPC | NUSA | PNP | RPLJ | ASNS | TUFA | ASNS |
| ASPC | ILES | GLNS | PROA | NUSA | SLYD | RPLJ | TIG | TUFA | GLTA |
| ASPC | THRC | GLNS | TRPC | NUSA | SPEB | RPSA | ASNS | TUFA | PROS |
| ASPS | GCVT | GLTA | ASNS | NUSA | YICC | RPSA | FUSA | TUFA | PURH |
| ASPS | GND | GLTA | PURA | PGI | AROK | RPSA | RFBB | TUFA | TKTA |
| ASPS | KBL | GLTA | TKTA | PGI | CYSK | RPSA | SERS | TUFA | TSF |
| BGLA | YFBU | GLTA | TSF | PHES | PHET | RPSA | TIG | TYPA | DNAK |
| CLPP | ARGH | GLTA | TUFA | PHES | TIG | RPSA | VALS | TYPA | PNP |
| CLPP | ISCS | GLTX | GND | PHET | PHES | RSUA | VALS | TYPA | PURF |
| CYSK | AROK | GLTX | PPA | PHET | TIG | SERC | ENO | VALS | ASNS |
| CYSK | DAPD | GLYA | GAPA | PNP | ACEA | SERC | GLYA | VALS | FUSA |
| CYSK | GCVT | GLYA | SERC | PNP | ISCS | SERS | ASNS | VALS | RPSA |
| CYSK | PGI | GLYA | TPIA | PNP | NUSA | SERS | RPSA | VALS | RSUA |
| CYSK | PYKF | GLYA | YIFE | PNP | PURF | SLYD | ISCS | YADF | ALAS |
| CYSK | SUCC | GND | ASPS | PNP | SLYD | SLYD | NUSA | YADF | FABI |
| DAPA | ASNS | GND | DUT | PNP | TYPA | SLYD | PNP | YADF | HISC |
| DAPA | GUAA | GND | ENO | PNP | YICC | SLYD | SPEB | YADF | PURH |
| DAPA | KDGK | GND | GLTX | PPA | DAPA | SLYD | YICC | YCHF | GUAA |
| DAPA | PPA | GND | GOR | PPA | GLTX | SPEB | NUSA | YFBU | BGLA |
| DAPD | AROA | GND | KBL | PPA | GND | SPEB | SLYD | YICC | NUSA |
| DAPD | ASPC | GND | PPA | PPA | GREA | SPEE | FUSA | YICC | PNP |
| DAPD | CYSK | GOR | GND | PPA | GUAA | SSPA | DAPD | YICC | SLYD |
| DAPD | NDK | GPMA | GAPA | PPIB | TSF | SSPA | PURN | YIFE | GLYA |

8.75 and centrifuged again. The pellet was resuspended in 6 ml of the same buffer and lysed by three passes through a chilled small French-pressure cell at 1,000 PSI. The lysate was centrifuged at 15,000xg for 25' in a JA-20 rotor. The supernatent was filtered through a non-binding 0.45µm syringe filter prior to chromatography.

### Two-dimensional Electrophoresis

300µg of material from each anion exchange fraction was diverted for analysis by 2D PAGE. 2D PAGE was performed at the Protein Chemistry Laboratory at Texas A&M University (http://www.calabreso.com/pcl/users.html). Briefly, acetone precipitated anion-exchange samples were reswelled into Igphor immobilized gradient gels (14cm pH 3-10 NL) (Pharmacia) and focused for approximately 60-80,000 Volt-hours. After reduction and exchange in SDS and DTT, 12% SDS gels (13x16cm) were run in the second dimension and stained with Gel Code Blue. pI's were determined by fitting a nonlinear standard curve from Pharmacia as a function of gel length and adjusted to proteins with known migration (e.g. DnaK, GroEL). Apparent molecular weight was determined by a standard ladder applied to the leftmost portion of the gel after loading of the iso-electric gel strip.

### Chromatography

For the liquid separation of clarified lysates, the following procedure was used. Approximately half of the cell-lysate was applied to a 1ml SOURCE 15Q (Pharmacia) resin packed into a Waters AP-1 glass column preequilibrated in 20mM NaCl, 30mM Bis-tris, 15mM Tris-Cl at a pH of 7.50 or 8.75. A segmented gradient from 20mM to 1M NaCl was run over approximately 150 column volumes at a flow rate of 3ml/min on an

ÄKTA Explorer HPLC. 5ml fractions were collected from the anion-exchange separation. For the second dimension, each anion exchange fraction was brought up to 1.5M $(NH_4)_2SO_4$ 100mM $KPO_4$ buffer at pH 7.0. This was applied to a 1ml SOURCE 15Phe (Pharmacia) resin packed into a Waters AP-1 glass column preequilibrated in the same buffer. A segmented gradient of 1.5M to 0M $(NH_4)_2SO_4$ was used over approximately 15 column volumes. 0.5mL fractions were collected directly into microdialysis cassettes (Pierce), and arrayed into foam racks and exhaustively dialyzed against 25mM ammomium bicarbonate. Denaturation, digestion and MALDI were performed essentially as described elsewhere (Park and Russell, 2000)(Park et al., 2000). Spectra were acquired on a Perseptive Biosystems Voyager Elite XL TOF with a pulsed nitrogen laser at 337nm. The dried samples were resuspended in 100-270μl of water mixed with MALDI matrix (35mM α-cyano 4 hydroxy cinnaminic acid/MeOH) to a final matrix concentration of ≈10mM and <0.5μl was spotted in duplicate onto 35mM overlayers of matrix in MeOH (Edmondson and russell, 1996). The samples were analyzed in reflectron mode with 25kV accelerating voltage, a grid voltage of 17.5kV and a delayed extraction time of 150ns. Signals from 100 laser shots were averaged per spectrum. Two-point calibration was performed using Angiotensin I and Neurotensin ($[M+H]^+$ =1296.6853, $[M+H]^+$= 1672.9175) and a low mass gate of 500 Da was used. MALDI DE-R-TOF spectra were taken from digests of each of the 380 fractions from the HPLC separations. Four separate lysates prepared on different days were used to generate the proteome separations. Two different pH's were utilized in the anion-exchange separation, each performed twice. Peak picking was done by the operator using Grams 386 software, and peptide-mass fingerprinting was performed as described below.

In total, nearly 2,000 spectra were annotated and analyzed for protein content by peptide mass fingerprinting.

## Peptide Mass Fingerprinting

Proteins were identified from the resulting peptides using MS-FIT and Protein Prospector. The algorithm that generates the MOWSE score, and the ranking for MS-FIT outputs is detailed in (Pappain et al, 1993(Pappin *et al.*, 1993) and (http://prospector.ucsf.edu). The utility of peptide-mass fingerprinting of single proteins and more recently protein mixtures, has been shown and applied to multiple experimental systems. (Schevencho et al, 1997 Park, et al, 2000).(Jensen *et al.*, 1997; Mann *et al.*, 1993; Shevchenko *et al.*, 1996b; Yates *et al.*, 1993). Peptide masses were searched against the most current SWISS-PROT database with no constraints on pI or MW. No post-translational modifications were allowed and species was limited to *E. coli*. A mass error of 300 ppm was applied and 1 missed cleavage was allowed. For our case, the following database matching criteria in MS-FIT were applied. 1.) The identified protein must come from the correct strain of *E. coli.* Since multiple strains of *E. coli* have been sequenced, false-positive protein matches often occur with different strains. For example, a common false positive protein from bacterial searches of *E. col*i is TraI, a gene located on the F' plasmid and not present in the *E. coli* strain we used, K-12 MG1655. 2.) The sequence coverage of the putative protein identified must be greater than 25%. 3.) The assigned peptides for any given identification must have mass accuracy error consistent in magnitude and trend with other peptides assigned to the same protein. Our average mass error was 20 ppm, with a standard deviation of 20 ppm, and our error rarely exceeded 50 ppm. The MOWSE algorithm is insensitive to error regardless which is why a high (300 ppm) tolerance was allowed. Multiple proteins could be identified in the same fraction, by removing the peptides assigned to the first protein and resubmitting the remaining peaks in a recursive process. Identification of the same protein in adjacent fractions in both separation dimensions, allows many orphan peptides to be assigned when they fell below threshold criteria.

## DNA Microarrays

Total RNA was isolated from 3 independent cultures grown under identical conditions as those used for protein samples. Cells were grown to $OD_{600}$ 0.51, 0.58 and 0.55 respectively. RNA isolation, synthesis of $^{33}$P-labeled cDNA probes using *E. coli* gene-specific primers (Sigma-Genosys, The Woodlands, TX), and hybridization to Panorama *E. coli* gene arrays (Sigma-Genosys), wasere performed as described previously (Arnold *et al.*, 2001) with the following modifications. Before cultures were harvested by centrifugation, 1/8 volume of ice-cold Ethanol/Phenol stop solution (5% water-saturated phenol (pH<7.0) in ethanol) was added to stop RNA degradation (Lee *et al.*, 2002b; Lin-Chao and Cohen, 1991). Prior to cDNA synthesis, RNA samples were treated with RQ1 RNAase-free DNAase (Promega Biotech, Madison, WI), followed by two extractions with phenol and phenol:CHCl3, ethanol precipitation and resuspension in DEPC-treated deionized water. For quantitation, filters were exposed to a phosphorimager screen, which was scanned at 100 micron resolution using a Fujix BAS2000 phosphorimager. The Fujix BAS image files were analyzed using Visage HDG Analyzer software (R.M. Lupton, Inc., Jackson, MI) running on a Sun Microsystems ULTRA10 workstation. The integrated intensity (I.I.) of each spot is the sum of the value of each pixel within the boundaries of the spot minus the local background. The I.I. values, which are expressed in arbitrary units, were exported to Microsoft Excel for further analysis. The 294 blank spots on the arrays were used to define a background expression level (0.45 ± 0.34 arbitrary units). Based on the visual examination of individual spots, we concluded that I.I. values >1.8 (background plus 4 SD) represented real signals. This cut-off was used in identifying transcripts for the genes encoding the proteins we observed.

### Database Generation

Output proteins from MS-FIT were indexed by SWISS-PROT ID as the unique key, and treated as text tables. All manipulation of identified proteins was done using scripts written in Perl or Microsoft Excel. Functional annotation was performed with the indexed list from the Riley Lab, at the following web address. Molecular weight and pI predictions were based on the 'pI Tool' located on the Expasy web site (http://expasy.org/tools/pi_tool.html) at the Swiss Institute of Bioinformatics. Lists of the proteins identified, their frequency and expression data (Eg), and all other data manipulations, and SWISS-2DPAGE comparisons are available as supplementary material on-line.

## Discussion

### Identities of Expressed Proteins

Understanding the physiology of a cell involves knowing what proteins are expressed under a given set of circumstances. Although powerful methods for genome-wide expression profiling based on examining mRNA are widely available, the correlation between mRNA and protein levels is imperfect, and direct examination of cellular protein content is needed. While 2-D gels have been applied to cataloging of catalog the expressed proteins in *E. coli* for many years, a variety of technical issues prevent efficient identification of the genes that encode the proteins seen as thousands of spots on 2-D gels. For example, protein recovery from gels is often low, samples are sometimes difficult to digest *in situ*, and the loading capacity of gels limits the amount of material that can be recovered from spots. Nevertheless, combining multiple 2-D gels with microsequencing or mass spectrometry has allowed the identification of 273 proteins from *E. coli* in mid-exponential phase growth in minimal glucose medium (Tonella *et al.*, 1998; Tonella *et al.*, 2001).

Above, we describe a complementary gel-independent approach based on multidimensional liquid chromatography. Although the resolution of chromatrography is much lower than gel electrophoresis, the samples that are obtained are much more efficiently processed for protein identification using the power of mass spectrometry to deconvolute complex mixtures of proteins found in chromatographic fractions. Using this method, we identified 310 proteins expressed in exponential-phase *E. coli* growing in M9 glucose media supplemented with amino acids (casein hydrolysate).

Figure 2.9 compares our results with the SWISS-2DPAGE proteome from *E. coli* *(Hoogland et al., 2000; Tonella et al., 1998)*. Taken together, our studies and the SWISS-2DPAGE identify 467 proteins. Of these, 116 were identified by both studies, while the native-state LC/LC MS approach described here identified 194 proteins that were not previously annotated in the SWISS-2DPAGE database. The SWISS-2DPAGE database identified 157 proteins that were not seen our experiments. Clearly, the two methods complement each other to provide a more complete understanding of the protein content of *E. coli* than either would alone.

Although our primary reason for performing a parallel analysis of column fractions by both peptide mass fingerprinting and 2-D gels was to validate the identifications made by the former, the concordance between a predicted pI/MW for a protein identified by mass fingerprinting and a spot on a gel also can be used to assign an identity to the spot without having to recover protein or peptides from the gels. In this way, we assigned identities to 41 spots that were previously unidentified in the SWISS-2D database.

Neither our method nor the combination of our method with 2-D gels is detecting all of the proteins we expect to be present in the cell. Although determining the number of expressed proteins from 2-D gels requires making assumptions about the number of spots per protein, on the order 1,500-2,500 genes about 1,000-3,000 expressed proteins are in reasonable agreement with the number of expressed *E. coli* genes seen by microarray experiments. (See above and A. Khodursky, personal communication). Why aren't we identifying 70-90% of the proteins we expect to see? While some are likely to be in the fractions we were not able to process, e.g. membrane proteins in the insoluble pellet and proteins in the flowthrough from the ion-exchange column, these are unlikely to account for the bulk of the proteins we are missing.

Because MALDI-ToF is capable of exquisite sensitivity with pure peptides, the amount of material in our samples is not limiting. However, peptides in mixtures compete for ionization, leading to suppression of the weaker signals. The two chromatographic separations used here help to alleviate that problem relative to

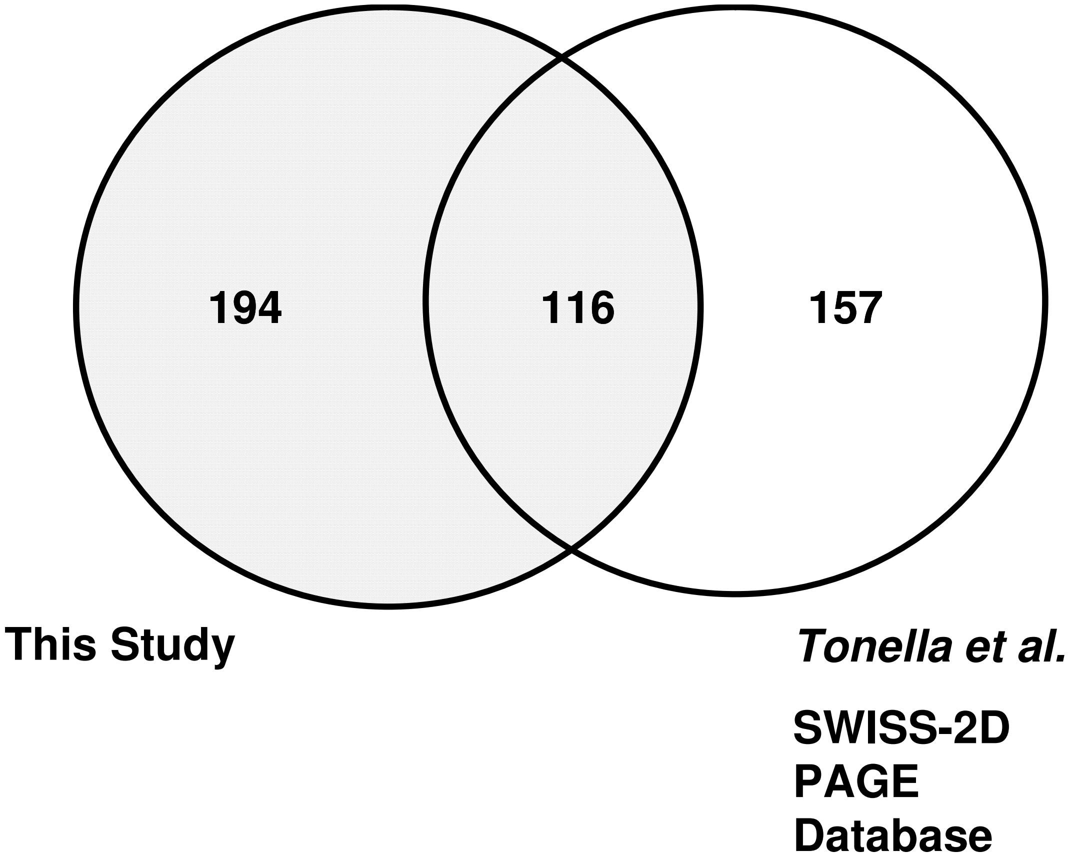


Figure 2.9.) Venn diagram of overlap between *E. coli* proteome projects/data sets. This diagram illustrates the distribution of the non-redundant protein assignments from each proteome, and the overlap between the two. The data sets together identify 467 non-redundant proteins from *E. coli.*

unfractionated extracts, where extremely abundant ribosomal proteins dominate the spectra. However, the complexity of our mixtures even after chromatography limits the numbers of proteins we can identify, possibly due to incomplete digestion, and the ability to identify enough tryptic peptides from each polypeptide for an unambiguous assignment for some proteins.

Recently, high resolution separation of tryptic peptides by multidimensional HPLC and mass spectrometric analysis of peptides and peptide fragments produced by collision induced dissociation have been used to provide a large-scale analysis of the yeast proteome (Washburn and Yates, 2000; Washburn *et al.*, 2001; Wolters *et al.*, 2001). Based on the published work on yeast, this approach would be expected to identify many more proteins in either our column fractions or in tryptic digests of the unfractionated *E. coli* lysates. Although tandem MS approaches yield greater numbers of identifications, the large amounts of instrument time and computational power required to search the data make it impractical for rapid identification of proteins and interrogating multiple physiologic states. Performing our type of analysis on all 380 fractions generated by the two chromatographic dimensions used here would involve prohibitive amounts of instrument time (weeks) with current technology.

## Protein Complexes

Identification of expressed polypeptides as the products of specific genes provides information about gene expression, but it is important to remember Benzer's modification of the "one gene-one enzyme" hypothesis of Beadle and Tatum (Beadle, 1945; Dronamraju, 1991) to "one cistron-one polypeptide." Individual polypeptides are not equivalent to proteins because proteins have quaternary structure and are often composed of multiple subunits. Indeed, it is becoming clear that many intracellular processes are carried out by larger multisubunit complexes than had been previously imagined (Alberts, 1998).

For these reasons, major efforts are ongoing to map the "interactomes" in several organisms by a variety of techniques including large-scale two-hybrid studies (Ito *et al.*, 2001; Schwikowski *et al.*, 2000; Uetz *et al.*, 2000; Uetz and Hughes, 2000) protein chips (Nelson *et al.*, 2000; Rabilloud, 2002; Zhu *et al.*, 2001) and identification of proteins that coimmunoprecipitate or copurify with specific baits for which there are antibodies (Tong *et al.*, 2002) or expressed versions with affinity tags (Butland *et al.*, 2005; Gavin *et al.*, 2002; Ho *et al.*, 2002). Although each of these methods is able to identify a subset of the interactions known to occur in a proteome, all of them identify only subsets of the previously known interactions and, presumably, only subsets of the unknown interactions they seek to find. In addition to high fractions of false negatives, each approach has its own kinds of false positives.

In our approach, cofractionation through partial purification provides suggestive evidence for interactions. Copurification is the classical method of biochemical identification of the subunits in a multisubunit protein; whatever remains at a reasonable stoichiometry after an activity is purified to homogeneity or near homogeneity is considered to be a subunit. Purification to homogeneity is impractical on a genome-wide scale for two reasons: first, because of the exponential increase in the scale of the experiment with the addition of each fractionation and second, because there is no universal assay by which to follow the activities of all possible complexes.

Instead, we examined the concordance of cofractionating proteins through parallel partial purifications. In each individual preparation any given protein complex is contaminated by many other proteins that copurify coincidentally. If we can use conditions that differentially change the elution of proteins, then we should find a different subset of the proteome contaminating the same complex. Subunits of stable complexes should stay together through all of our purification steps.

As with other proteomics methods, our method will include both false positives and false negatives. False positives will occur simply because some proteins will copurify coincidentally over both of our fractionation schemes. Here, we used two different pHs in the ion exchange dimension as our different schemes. The changes in elution position that we need to alter the spectrum of contaminants seem to be larger for the weakly charged proteins that elute from the column first than for the more strongly anionic proteins that elute later at higher salt concentrations. This is as expected; these proteins are probably enriched for acidic residues that do not titrate significantly over the pH range we used. In addition, the titration of ionizable groups in these proteins will have a smaller effect as a fraction of the total charge compared to the more weakly anionic proteins. Despite these limitations, the use of two pHs has already significantly reduced the background of coincidental copurification. Other fractionation schemes that are based on larger differences in the physical basis for separation may reduce the false positives further.

False negatives have two major sources. First, we know from the long history of protein biochemistry that some complexes will not survive the purification steps. Cell lysis and fractionation involve significant dilution from intracellular conditions, and the changes in salt and pH that accompany each fractionation step, as well as binding to the chromatographic matrices *per se*, will disrupt some complexes. Second, the low density of protein identications in each experiment will lead to missed identifications of proteins that are successfully copurified. This is clearly happening in our experiments for some known complexes, including RNA polymerase core. Although we identify the α subunit of RNA polymerase in each of the four expressed proteomes, and β and β' are seen on gels in the expected fractions, diagnostic peptides for β and β' were only found in one out of four experiments.

Despite these caveats, 125 pairs of putative interactions are detected in our experiments. Several are either known or plausible from the operon structure of the genes involved. Further study will be needed to determine which of the interactions represent real complexes.

### Future Applications for *E. coli* and Other Bacteria

The prospect of efficiently assigning gene identities to expressed proteins provides renewed impetus to the analysis of the expressed proteome as a function of bacterial physiology. It is important to note that the approach described here is more accessible to small labs than most other large-scale proteomics methods; although the process is clearly amenable to automation at many steps; this study was done entirely without the benefit of robotics. Moreover, unlike approaches based on predigesting samples before separation, the ability to correlate identities made by native-state LC/LC MS with quantitation from 2-D gels, which are even more accessible to small-scale projects, makes the use of both methods much more powerful than the sum of the two.

Similarly, most large-scale interactome projects require either the construction of large numbers of strains expressing affinity-tagged proteins or the generation of large numbers of high-quality antibodies (note that polyclonal antibodies often cross-react with many bacterial proteins). By contrast, cofractionation can be done with any wild-type or mutant strain without further strain construction.

The native-state LC/LC MS approach should be broadly applicable beyond *E. coli*. Proteomics based on peptide mass fingerprinting is strongly dependent on the availablity of complete genome sequences; proteins can only be identified if they are in the database. With the rapid advances in the availability of complete geneome sequences, this is no longer a problem for many microorganisms. The small size of bacterial genomes is a significant factor in the success of the approach described here, which used peptide mass fingerprinting alone to identify several components in complex mixtures. (Eriksson *et al.*, 2000) calculated the theoretical information content intrinsic to a peptide mass as a function of the genome size of the subject organism. This study validates their theoretical calculations and shows that for a genome of the size and complexity of *E. coli*, peptide mass fingerprinting is able to successfully deconvolute mixtures of proteins generated by two dimensions of chromatography.

Within the specter of *E. coli*, we can also utilize this method to begin to globally map the potential differences in interactome(s) and protein content between two separate cell-states. Chapter III illustrates the application of the overall proteome methodology to just such a question. Since we have improved the ability to resolve complex mixtures of peptides, and are able to detect peptides at lower abundance levels, the basic Mid-exponential phase proteomes at two different pH's were repeated in order to more effectively analyze the stationary phase content of *E. coli*.  Additionally, further characterization of the proteome has revealed significant changes in intracellular composition and arrangement upon the shift to stationary phase.

# CHAPTER III

# PROTEOME ANALYSIS OF *Escherichia coli* BY NATIVE STATE CHROMATOGRAPHY AND MALDI MS UNDER EXPONENTIAL AND STARVATION GROWTH CONDITIONS REVEALS SIGNIFICANT CHANGES IN INTRACELLULAR PROTEIN BEHAVIOR

## Summary

Proteins from exponentially growing and stationary phase E. coli strain MG1655 were separated and identified via non-denaturing HPLC coupled to off-line mass spectrometry. Two chromatographic dimensions were used: anion exchange and hydrophobic interaction chromatography; in parallel separations the anion exchange step was performed at pH 7.50 and pH 8.75. Analysis of exponential phase E. coli grown in MOPS glucose minimal medium identified 2,533 proteins corresponding to 389 unique gene products. A parallel analysis of proteins from cells in stationary phase for 3 hours yielded 2,308 corresponding to 362 genes. In total, 520 unique proteins were identified between the two cell states. Changes in the number of times a protein was identified were correlated with likely changes in the abundance of proteins in exponential vs. stationary phase, enabling patterning and profiling of changes in the proteomes. Proteins were identified across all cell functional and abundance categories. Ultimately, differences in the chromatographic elution profiles of proteins identified in these samples suggest large-scale changes in the biochemical properties of the two proteomes, possibly

due to changes in protein-protein interactions. Ninety-four proteins were identified that exhibited significant changes in elution between exponential and stationary phase.

## Introduction

Outside the laboratory, bacteria spend most of their time in nongrowing states. Although it does not form form spores, *E. coli* responds to nutrient starvation by going through dramatic global physiologic changes, resulting in cells that are smaller, rounder, biochemically altered, and more resistant to a variety of environmental challenges (Huisman *et al.*, 1996). These changes require protein synthesis and a complex program of gene expression. At least 115 genes have been identified and annotated that are either expressed in or specific to stationary-phase growth conditions (Hengge-Aronis, 2002a, b; Matin, 1991; Tani *et al.*, 2002) and upwards of 200 have been suggested to be stationary-phase specific (Chatterji and Ojha, 2001; Tani *et al.*, 2002), by a combination of reporter fusions, microarrays and studies using 2-dimensional protein gel electrophoresis.

The expression of various stationary phase specific genes involves combinations of the alternative sigma factors RpoS, RpoH, and RpoN, catabolite repression, and the stringent response (Chatterji and Ojha, 2001; Hengge-Aronis, 1996; Hirsch and Elliott, 2002; Matin, 1991). A wide variety of other transcription factors and regulatory RNAs are also involved in regulating gene expression in stationary phase (Lease *et al.*, 2004; Rao and Kornberg, 1999). In addition, protein degradation and peptidase activities are involved in entry into and exit from stationary phase (Becker *et al.*, 2000; Weichart *et al.*, 2003). Different subsets of proteins are induced during limitation for carbon, nitrogen, or phosphate. Some proteins are induced by more than one kind of nutrient limitation, and some, called Pex genes, are induced regardless of how the cells enter stationary phase

(Lomovskaya *et al.*, 1994; Schultz *et al.*, 1988). Induction of genes in stationary phase does not occur all at once; several waves of genes are induced as cells enter a starved state (Groat *et al.*, 1986; Matin, 1991). The importance of stationary-phase regulated genes is seen from the fact that mutations in some of these genes affect the survival of *E. coli* in stationary phase (Zinser and Kolter, 1999).

The pleotropic natures of stationary phase, from different sources of stimulation, are ideal to study from a post-genomics frame to put into context many of these global cell changes. Not all stationary phases are equal, different subsets of genes are expressed depending on the nature of the stress. This makes it an excellent model cell-state because despite extensive studies on stationary phase gene expression in *E. coli*, our understanding of how changes in gene expression are reflected in the proteome is very incomplete. 2-D gel studies have identified changes in the synthesis of about 30 proteins after carbon starvation (Matin, 1991; McCann *et al.*, 1991); these studies were done long before current proteomic methods became available for identification of the proteins. Many of the proteins whose levels change in stationary phase are either identified only as spots or are inferred to change from changes in mRNA. Changes in mRNA will also miss changes in protein levels due to turnover, modification, or assembly into different multiprotein complexes. This suggests that current knowledge of the global changes in the *E. coli* proteome during stationary phase is still incomplete, but utilizing the extensive genetic and molecular work in the field provides a means to an ‘omics positive control unavailable to unannotated organisms. One such advantage here is the ability to use experimental evidence to validate and compare our data. This also directed determinations of changes between cell-states. Globally then, the evidence available suggests larger dynamic forces occurring within the cell upon cessation of exponential growth, which we are beginning to explore with these methods.

Although no single method detects all proteins in the proteome, a variety of methods provide large-scale identifications (Ducret *et al.*, 1998; Jensen *et al.*, 1997; Washburn *et al.*, 2001; Yates *et al.*, 1993). We have described an approach to characterization of microbial proteomes based on combining multidimensional nondenaturing liquid chromatography with protein identification by mass spectrometry (Champion *et al.*, 2003). Proteins are separated in their native states, providing information not only about what is present, but also about the chromatographic behavior of the proteins, which reflect basic biochemical properties.

Here, we describe the application of this approach, which provides information that is complementary to other proteomic methods, to examine one snapshot of stationary phase physiology at a specific time and from a specific way of entering stationary phase: carbon limitation due to glucose exhaustion in defined medium. Below we confirm the stationary phase regulation of many genes, and suggest stationary phase regulation of expression of others, and identify widespread changes in the chromatographic properties of many proteins.

## Materials and Methods

### *E. coli* Cell Culture and Lysate Preparation

MOPS glucose minimal media (1L) contained: 0.1% Glucose, 19mM $NH_4Cl$, 1.32 mM $K_2HPO_4$, 0.2 μg/ml thiamine, 10 μg/ml uridine, 0.52 mM $MgCl_2$, 0.25 μM $CaCl_2$ (1x MOPS 8.37 g MOPS, 0.72 g tricine, 48 mg $K_2SO_4$, 2.92 g NaCl, 3 mg $FeSO_4$-$7H_2O$,

≅1.4 g KOH Δto pH 7.4.). One Liter cultures of *E. coli* MG1655 were grown in minimal MOPS glucose media in a non-baffled Fernbach flask in a New Brunswick G76 rotary bath shaker at 37°C and 250 RPM. Exponential phase cultures were grown at an $OD_{600}$ of 0.5. Stationary phase cells were grown under identical conditions and harvested 3 hours after inflection from exponential-phase growth to an $OD_{600}$ of 3.2. Cells were rapidly chilled on ice and harvested by centrifugation. Cell lysates were prepared as described previously (Champion *et al.*, 2003).

2D Liquid Chromatography

Chromatography was performed as described earlier (Champion *et al.*, 2003). The amount of lysate loaded on the first column was adjusted so that protein from equivalent cell ODs was used. Half of the original exponential lysate volume was loaded onto each anion exchange column, approximately 2.5 ml, and 500-700 µl of the stationary phase lysate was utilized.

Sample Preparation and Mass Spectrometry

0.5 ml samples were dialyzed as described against 25 mM ammonium bicarbonate, denatured for 20 min at 90°C and digested for four hours at 37°C by the addition of 1µg of sequencing grade trypsin (Promega, Madison, WI) to each sample. Samples contained between 0-20 µg of protein, estimated by overall Coomassie intensities on 2D gels. After trypsin digestion, the samples were lyophilized and resuspended in an equal volume 1:1 of ($H_2O$ + 0.1%TFA) and MALDI matrix, [alpha-cyano 10 mg/ml in 66%MeCN, 0.1% TFA]. (Typically 100 µl of each was added.)(Park and Russell, 2000). <0.5µl spots were deposited in duplicate onto a stainless steel

MALDI target and spectra were acquired manually (Spot to Spot) on an Applied Biosystems Voyager DE-STR MALDI in reflectron mode with 22K accelerating voltage, 60% grid voltage and a delay time of 150 ns. 120 spectra were averaged for each sample and peak processing was performed using Data Explorer software (Applied Biosystems). The following parameters were applied to extracted peaks: A S/N filter of >5:1, two baseline corrections, the built-in scripts for noise filtering, (correlation factor 0.7), isotope deconvolution (Adduct H C6H5NO) and monoisotoic mass filtering to eliminate $^{13}$C isotope peaks.

Sample Identification by Peptide Mass Fingerprinting

Peak lists from each of the fractions were analyzed using the ProFound peptide mass fingerprinting engine located at Rockefeller University http://129.85.19.192/profound_bin/WebProFound.exe (Zhang and Chait, 2000). Manual identification was done using the *E coli* database in SWISS_PROT utilizing a mass error of 300 ppm. Additional settings per software were: *monoisotopic data*, *30 Proteins Listed*, and *single protein mixture* settings (See below).

Proteins were identified as positive hits if they fulfilled three sets of criteria: 1.) If the systematic error on the matched peptides was linear (e.g. in the identification metric, the graph of ProFound mass error of the matched peptides was linear as matched mass increased), and, if after correction, high mass accuracies of <20ppm were obtained (<20ppm standard error). 2.) Sequence coverage of the protein exceeded 25% and 3.) Proteins that satisfied all criteria except for sequence coverage were considered hits if they were identified multiple times in adjacent fractions meeting all previous criteria. This third criteria of identification identified the ‘tails’ of peaks in the elution profiles of proteins in the separations. This does not increase the number of unique protein identifications but recursive mass matching using these ‘likely’ entries reduces false positives by removing that subset of peptides from the matching lists, reducing random peak matches (Jensen *et al.*, 1997; Lee *et al.*, 2002a; Pappin *et al.*, 1993).

Multiple proteins were identified in each fraction by recursive mass matching where the identified peptides from the top ‘hit’ were removed and the remaining peak list was resubmitted to ProFound with identical parameters. For the top proteins identified in most fractions, the scoring system utilized in ProFound was sufficient to annotate high confidence hits as true positives. Proteins that fell below the Z score threshold were often in high complexity fractions that contained multiple unassigned peaks or contained multiple proteins where one protein was significantly more represented by greater numbers of tryptic fragments. Additionally, recursive mass matching resolves these differences by eliminating dominant positive peptides from the mass list allowing more accurate scoring of remaining peaks. Recursive mass matching abrogated the need to use the ‘potential protein mix’ selections in ProFound.

## Bioinformatics

All protein sorting, indexing and annotation was performed using Microsoft Excel, and conversion of the database numbers to Blattner # and MW/pI was performed using the pI/MW tool on the Swiss Prot web site (http://www.expasy.ch) (Gasteiger *et al.*, 2003). Assignment of Proteins identified by Blattner #’s (b#’s) to functional categories was done using a web-based script and the data were taken from the M. Riley group as compiled from the MG1655 (http://tofu.tamu.edu, http://oligomers.tamu.edu)

(Marino-Ramirez *et al.*, 2004). Metabolic mapping was done via BioCyc (http://biocyc.org) using the relative gene expression metabolic mapping utility, where relative 'intensities' were defined as the ratio of the number of times a protein was identified in stationary phase samples over the number of identifications in exponential growth (Karp *et al.*, 2000; Karp *et al.*, 2002). Contents of the SWISS2D were obtained from http://expasy.ch and included all pH gel ranges for *E. coli* K12 except for data obtained from DiGE (Amersham Biosciences) experiments. The contents of *E. coli* proteomes compared in this work were obtained from the supplementrary data of Tani et al., and Corbin, Paliy et al. (Corbin *et al.*, 2003; Tani *et al.*, 2002) via text capture software in Adobe Acrobat (Adobe Inc.). All of the identifications for each fraction and from published datasets were entered into a MySQL database to facilitate data mining (L. Niu and J.C. Hu, unpublished). This is maintained as the EEP 'Experiments in *E. coli* Proteomics' website (http://eep.tamu.edu). Additional supplementary data not specifically mentioned are also available at the following internet address: http://eep.tamu.edu/nondelc/index.php?page=results.html.

## Results

### Identification of Proteins from Exponential and Stationary Phase Cells

To understand how the proteome of *E. coli* changes in stationary phase, we compared proteins from MG1655 cultures that were exponentially growing in MOPS glucose minimal medium to proteins from the same strain grown in the same medium, but incubated in stationary phase at 37°C for 3 hours before harvesting. Proteome analysis was done as described previously (Champion *et al.*, 2003). Proteins were fractionated by nondenaturing 2-D chromatography in order to preserve, as much as possible, native protein structure and protein-protein interactions. The two dimensions of chromatography, anion exchange and hydrophobic interaction, resolved each proteome into 380 fractions. Separations were repeated using two different pH's (7.50 and pH 8.75) for the anion exchange step, yielding 760 fractions for each proteome. Proteins from each of the 1,520 fractions, each of which contains many proteins, were identified by peptide mass fingerprint analysis of tryptic fragments.

The number of identifications made in each fraction were strikingly different for the exponential and stationary phase samples (Figure 3.1A). More identifications were made in stationary phase from the middle numbered anion-exchange fractions, while more identifications were made from late fractions in exponential phase. On average, 3.3 ± 2.1 and 3.4 ± 2.3 proteins were identified per fraction for log phase proteins fractionated at pH 7.50 and 8.75, respectively. For stationary phase proteins, at pH 7.50 and 8.75 the average number of proteins identified per fraction were 3.3 ± 2.2 and 2.7 ± 1.9. About 15% of the fractions yielded no MS data whatsoever. We were able to identify many proteins from some fractions including a single fraction with 11 identified proteins. A histogram of the distribution of the number of proteins identified per fraction is displayed as Figure 3.1B.

Overall, 4,841 proteins were identified from the 1,520 fractions (Appendix NR table). A large fraction of these represent the same gene product being identified in more than one fraction, at both pHs, and in both exponential and stationary phase cells. Table 3.1 shows a variety of ways in which this redundancy is distilled, by eliminating duplicate identifications within a fractionation run, within a cell state, or for the complete

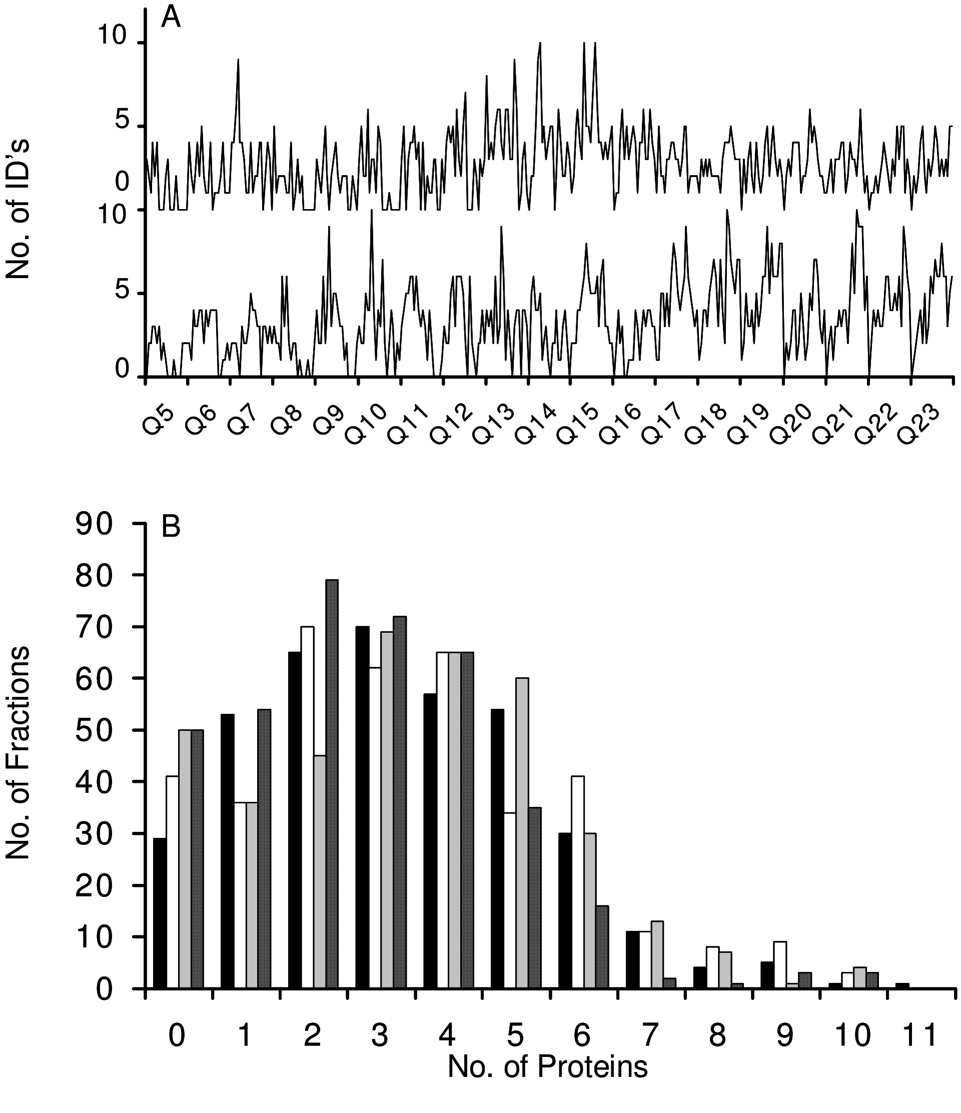


Figure 3.1.) Distribution of numbers of protein identifications. A) Number of identified proteins per fraction from separation of log phase (lower line) and stationary phase (upper line) proteins. Labels indicate the ion exchange first-dimension fraction used to generate each second-dimension elution profile. The periodic dips in the number of proteins identified correspond to early Phe (HIC) fractions, which contained little protein, and thus yielded no identifications. Anion exchange was done at pH8.75. B.) Distribution of the numbers of proteins identified per HIC (Phe) fraction from exponential samples at pH 7.50 (Black), exponential at pH 8.75 (White), stationary phase samples at pH 7.50 (Hatched), and stationary at pH 8.75 (Shaded).

Table 3.1.) Identified protein totals from MOPS glucose *E. coli* proteomes.
Each cell state was examined at pH 7.50 and 8.75. IDs count the number of fractions where a particular protein was identified. Genes counts the number of gene products found. Nonredundant counts each gene product only once from the combined subtotals.

| | pH 7.50 | pH8.75 | Combined |
|---|---|---|---|
| **Log Phase** | | | |
| **ID's** | 1,243 | 1,290 | 2,533 |
| **Genes** | 278 | 290 | 568 |
| **Nonredundant Log** | | | 389 |
| | | | |
| **Stationary Phase** | | | |
| **ID's** | 1,270 | 1,038 | 2,308 |
| **Genes** | 269 | 241 | 510 |
| **Nonredundant Log** | | | 362 |
| | | | |
| **Totals** | | | |
| **ID's** | 2,513 | 2,328 | 4,841 |
| **Genes** | 547 | 531 | 1,078 |
| **Nonredundant Total** | 375 | 387 | 520 |

dataset. This yields 389 gene products from the exponential phase sample, 362 from the stationary phase sample and 520 overall. Table 3.2 lists the proteins assigned to eachcell-state and found in common. 231 proteins were found in both proteomes, 158 proteins were found only in exponential phase cells and 131 from stationary phase only.

### Correlation of Numbers of Identifications with Protein Abundances

As noted above, many gene products were identified in multiple fractions within the same separation experiment. Although the peak intensity in MALDI-MS is not a quantitative measure of peptide abundance, there is an empirical correlation between protein abundance and the number of times a protein was found across the fractionation. Using number of identifications as a proxy for abundance, we can generate a virtual elution profile for any protein in each of the chromatography dimensions by counting the number of Phe fractions where a protein is found from each Q fraction, and vice versa. Figure 3.2 shows the virtual elution profiles for six proteins from the mid-exponential phase cell lysate. In those cases where the same protein can be visualized by 2-D gels of the Q fractions, peak positions inferred from the virtual elution profile match the peaks inferred from spot intensities (virtual and actual 2D gel profiles of fractionated extracts are available on the EEP website
http://eep.tamu.edu/nondelc/index.php?page=results.html)

Overall, the the number of overall identifications of a protein should be correlated with its abundance. Figure 3.3A shows the distribution of redundant identifications from all four complete proteome profiles. About 75% of the gene products were identified between 1 and 10 times. 170 proteins were identified in just a single fraction, which

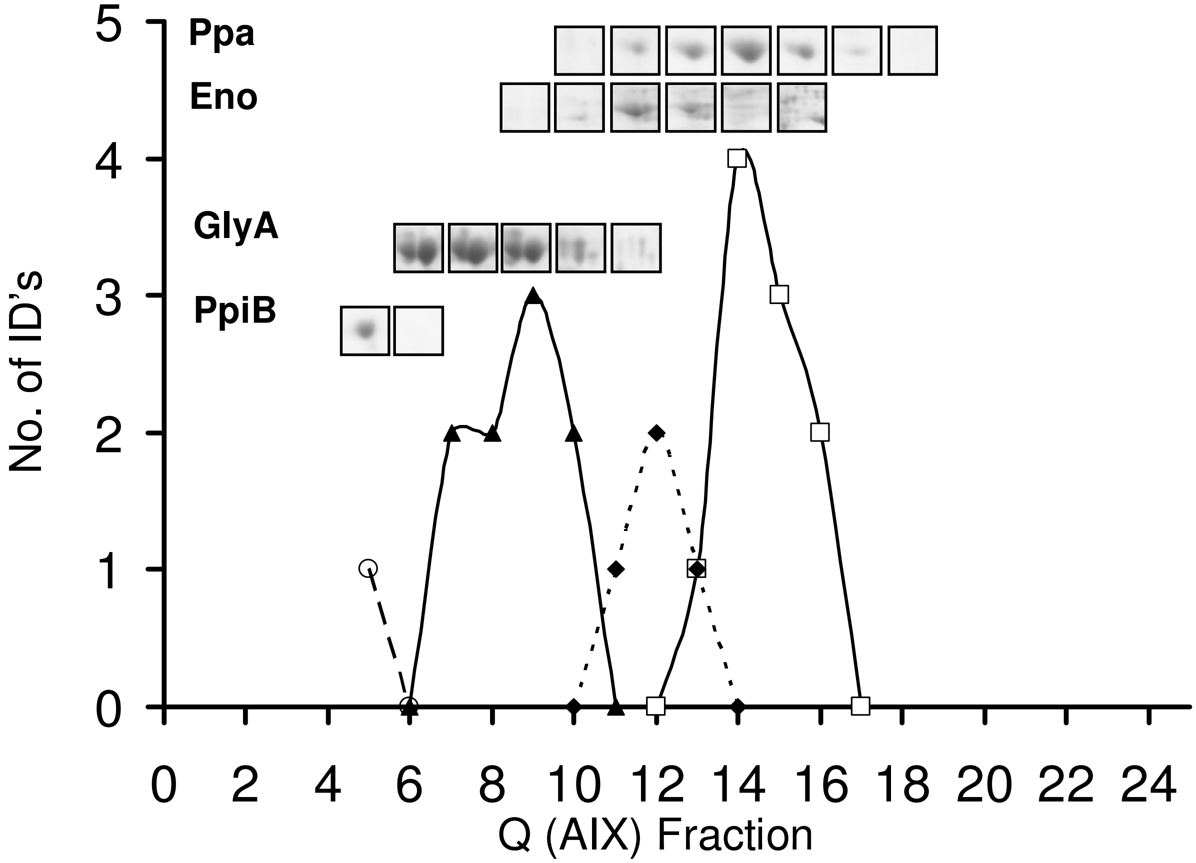


Figure 3.2.) Virtual elution profiling of proteins. Virtual elution profiles generated for these proteins illustrate elution patterns and allow correlation of MS generated ID frequencies with spot density observed on 2D PAGE. The elution positions for four selected proteins are shown as follows; □Ppa, ΔGlyA, ♦Eno and ○PpiB. Spot densities from the 2D PAGE of the anion-exchange gradient fractions for the region associated with each protein are given above the trace for comparison. PpiB for example, is only observed as an intense staining spot in anion-exchange fraction 5, which is identical to its identification with MS. Identification frequency appears to trend with spot density.

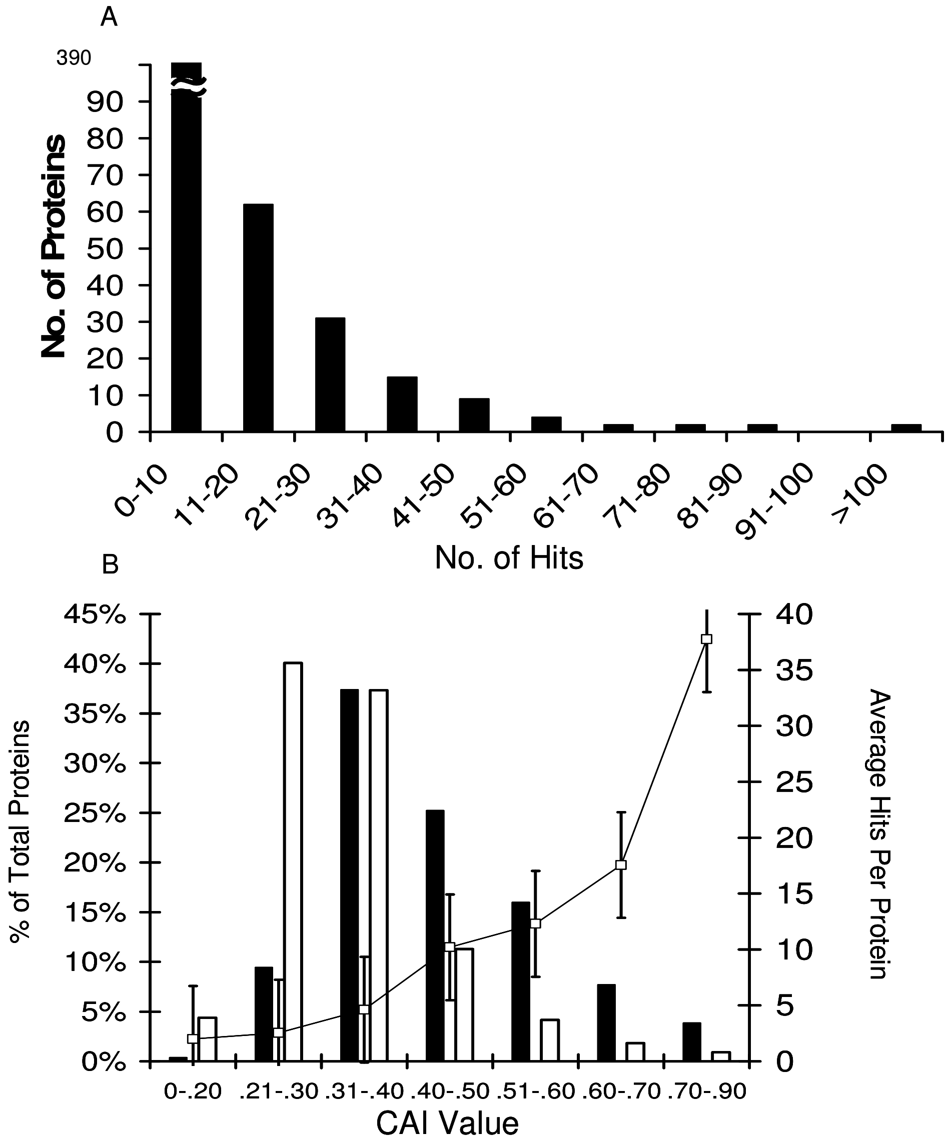


Figure 3.3.) Identification frequency and predicted expression. A.) Histogram of the number of times each protein was identified. B.) Codon Adaptation Index (CAI) bins as a fraction of identified proteins (Black) and the *E.coli* genome (White). The line shows the average number of times proteins in each CAI category were identified, illustrating that the average CAI is higher for those proteins that are identified more often A higher CAI value is indicative of higher predicted expression levels.

Table 3.2.) Non redundant protein assignments by cell state.
Non redundant list of proteins uniquely assigned to each cell state by gene name, and those found in both cell states. A.) 158 proteins identified only in exponential cells, B.)131 proteins identified only in stationary phase cells and C.) 231 proteins identified in both cell states. Reference numbers indicate proteins that were identified by other reference proteomes in either cell state. *Ratio* in C indicates the fold greater number of ID's in exponential vs. stationary phase proteomes.

1. SWISS 2D Page http://au.expasy.org/ch2d/ W3110 OD$_{600}$ = 1.0 MOPS Glucose (Hoogland *et al.*, 2000)
2. Cyber Cell http://redpoll.pharmacy.ualberta.ca/CCDB/index.html K12 exponential growth (Sundararaj *et al.*, 2004)
3. Corbin & Paily et al., MG1655 OD$_{600}$ = 0.4 Minimal 0.2% glycerol (Corbin *et al.*, 2003)
4. Champion et al., MG1655 OD$_{600}$ = 0.5 M9 glucose (Champion *et al.*, 2003)

**A.** **Mid Exponential Cells Only**

| SWISS ID | Name | Ref. | SWISS ID | Name | Ref. | SWISS ID | Name | Ref. |
|---|---|---|---|---|---|---|---|---|
| P00452 | nrdA | 3 | P12758 | udp | 1, 2, 4 | P33940 | mqo | |
| P00547 | thrB | 1, 2 | P14375 | zraR | | P35340 | ahpF | 1, 2, 3 |
| P00582 | polA | 1, 3 | P15039 | purR | 2, 4 | P36649 | cueO | |
| P00584 | glgC | 3 | P15042 | ligA | | P37013 | norR | |
| P00859 | atpF | 2 | P16244 | cpxR | | P37313 | dppF | |
| P00935 | metB | 3 | P16431 | hycE | | P37666 | ghrB | 3, 4 |
| P02351 | rpsB | 2, 3 | P17117 | nfsA | 4 | P37744 | rfbA | 3, 4 |
| P02359 | rpsG | 3 | P17315 | cirA | 2 | P37751 | wbbK | 4 |
| P02366 | rpsK | 4 | P17579 | kdsA | 1, 2, 3 | P38134 | etk | |
| P02384 | rplA | 1, 2, 3 | P17854 | cysH | 3, 4 | P38489 | nfsB | 1, 3, 4 |
| P02413 | rplO | 3, 4 | P18843 | nadE | 1, 3 | P39174 | fliY | 1, 2, 3 |
| P02418 | rplI | 1, 2, 3, 4 | P19494 | lrp | 3 | P39272 | dcuS | |
| P02420 | rplS | 4 | P19797 | metR | 4 | P39290 | rlmB | |
| P02917 | livJ | 1, 2, 3 | P21177 | fadB | | P39323 | ytfP | |
| P02927 | mglB | 1, 2, 3 | P21774 | fabZ | 2, 4 | P40874 | solA | 3 |
| P02933 | envZ | | P23486 | fre | | P41407 | azoR | |
| P03002 | rho | 2, 3 | P23843 | oppA | 1, 2, 3, 4 | P42593 | fadH | |
| P04384 | metK | 1, 2, 3 | P23847 | dppA | 1, 2, 3, 4 | P42608 | exuR | |
| P04391 | argI | 1, 3, 4 | P23863 | cmk | 2, 4 | P45473 | yhbS | 2 |
| P04422 | aspA | 2, 3, 4 | P23869 | ppiB | 1, 2, 4 | P45563 | xapA | |
| P04951 | kdsB | 1, 3, 4 | P23908 | argE | | P45770 | yrdA | 2 |
| P04968 | ilvA | | P24233 | ndk | 1, 2, 3, 4 | P46853 | yhhX | 3, 4 |
| P04983 | rbsA | | P24249 | fabH | 2 | P46880 | glk | 3 |
| P05380 | groS | 1, 3, 4 | P24555 | ptrB | | P52065 | yggX | 1, 4 |
| P05826 | glnB | 2 | P24991 | dsbA | 1, 2, 4 | P52073 | glcE | |
| P05838 | sspA | 1, 2, 4 | P25537 | rng | | P52647 | ydbK | 3 |
| P06960 | argF | 1, 3, 4 | P25716 | fabG | 2, 3 | P53635 | sodC | 2 |
| P06961 | cca | | P25740 | rfaG | | P75678 | ykfA | |
| P06968 | dut | 1, 2, 4 | P25741 | rfaP | | P75805 | yliJ | 3, 4 |
| P06980 | gshA | 3 | P25748 | galS | | P75876 | yccW | |
| P07001 | pntA | 3 | P26266 | fepE | | P76316 | dcyD | 3 |
| P07016 | sucB | 1, 2 | P26282 | folP | 4 | P76373 | ugd | |
| P07459 | sucD | 1, 2, 3 | P27126 | rfaS | | P77258 | nemA | 2 |
| P07638 | aroA | 1, 4 | P27252 | rpiA | 1, 2, 4 | P77391 | yeaG | 3 |
| P07649 | truA | | P27300 | lpxK | | P77690 | arnB | |

Table 3.2 Continued…

| SWISS ID | Name | Ref. | SWISS ID | Name | Ref. | SWISS ID | Name | Ref. |
|---|---|---|---|---|---|---|---|---|
| P07672 | apt | 2,4 | P27511 | folE | 4 | P77718 | thiI | |
| P07762 | glgB | | P27827 | yifE | 2, 4 | P77804 | ydgA | 2, 3, 4 |
| P07862 | ddlB | | P27828 | rffE | 4 | P80449 | folX | 4 |
| P08179 | purN | 4 | P28860 | glpX | 2 | Q46829 | bglA | 3, 4 |
| P08193 | accD | 2 | P29015 | ribC | 1 | Q46933 | tas | |
| P08201 | nirB | | P29464 | pyrH | 2, 4 | | | |
| P08244 | pyrF | 1, 2, 4 | P30136 | thiC | 3, 4 | | | |
| P08400 | phoR | | P30177 | ybiB | 3 | | | |
| P08506 | dacC | | P30747 | moaC | 4 | | | |
| P08837 | crr | 1, 2, 3, 4 | P30854 | evgA | 2 | | | |
| P09030 | xthA | 1, 2, 4 | P30867 | accA | 2, 3, 4 | | | |
| P09151 | leuA | 3 | P31216 | ychF | 1, 2, 3, 4 | | | |
| P09158 | speE | 2, 4 | P31456 | yidS | 4 | | | |
| P09170 | rbfA | 1, 4 | P31473 | yieN | | | | |
| P09371 | fadR | | P32130 | yihI | | | | |
| P09374 | pflA | 2 | P32665 | gldA | 2 | | | |
| P09550 | ubiX | | P33137 | mdoH | | | | |
| P09625 | trxB | 1, 2, 3, 4 | P33138 | clpX | 2, 3, 4 | | | |
| P09743 | deoD | 4 | P33221 | purT | 3, 4 | | | |
| P10177 | eda | 1, 2 | P33225 | torA | | | | |
| P10371 | hisA | 2 | P33232 | lldD | 2 | | | |
| P10423 | iap | | P33234 | adiY | | | | |
| P11445 | argB | 3 | P33643 | rluD | | | | |
| P12281 | moeA | 3 | P33937 | napA | | | | |

Table 3.2. Continued...

**B.** **Stationary Phase Only**

| SWISS ID | Name | Ref. | SWISS ID | Name | Ref. | SWISS ID | Name | Ref. |
|---|---|---|---|---|---|---|---|---|
| P00470 | thyA | | P24182 | accC | 2, 3, 4 | P77154 | ycjT | |
| P00478 | pyrI | 1, 2, 3, 4 | P24186 | folD | | P77202 | dsbG | |
| P00562 | metL | 3 | P24192 | hypD | 1 | P77212 | ykgC | |
| P00888 | aroF | | P24230 | recG | | P77239 | cusB | |
| P00895 | trpE | 3 | P24231 | pmbA | 2, 3 | P77381 | djlB | |
| P00907 | carA | 1, 2, 3, 4 | P25520 | galU | 2, 3 | P77432 | ydeV | |
| P00959 | metG | 1, 2, 3 | P25522 | mnmE | | P77433 | ykgG | 2 |
| P00961 | glyS | 3 | P25526 | gabD | 3 | P77581 | argM | |
| P02432 | rpmE | | P25906 | ydbC | 2 | P77645 | miaB | 3 |
| P03004 | dnaA | | P26607 | barA | | P77674 | ydcW | 3 |
| P03024 | galR | | P27127 | rfaB | | P77713 | yagH | |
| P03026 | arcA | | P27246 | marA | | Q46812 | ssnA | |
| P04286 | ftsI | | P27430 | dps | 1, 2, 3 | Q46857 | dkgA | 3 |
| P06710 | dnaX | | P27550 | acs | 3 | | | |
| P06971 | fhuA | 2 | P28302 | gadB | 3, 4 | | | |
| P07003 | poxB | 1, 3 | P28904 | treC | | | | |
| P07004 | proA | 1, 2, 3, 4 | P30850 | rnb | 3 | | | |
| P07024 | ushA | 2, 3 | P30958 | mfd | 3 | | | |
| P07651 | deoB | 2, 3 | P31660 | prpC | | | | |
| P08328 | serA | 3 | P31806 | yjeF | | | | |
| P08331 | cpdB | 2 | P32176 | fdoG | | | | |
| P08394 | recB | | P32664 | nudC | | | | |
| P08531 | araG | | P32719 | alsE | | | | |
| P08956 | hsdR | | P33013 | dacD | | | | |
| P09126 | hemD | | P33345 | yehH | | | | |
| P09152 | narG | | P33602 | nuoG | 2, 3 | | | |
| P09157 | sodB | 1, 2, 4 | P33920 | yejK | 3 | | | |
| P09546 | putA | 2 | P36683 | acnB | 1, 2, 3 | | | |
| P10121 | ftsY | 4 | P36767 | rdgC | | | | |
| P10413 | htpG | 1, 2, 3 | P36938 | pgm | 2, 3 | | | |
| P10443 | dnaE | | P37095 | pepB | 2, 3, 4 | | | |
| P11056 | bfr | 3 | P37177 | ptsP | 3 | | | |
| P11585 | relA | | P37192 | gatY | 2, 3 | | | |
| P13009 | metH | 1, 3 | P37196 | treF | | | | |
| P13031 | glgP | 3 | P37689 | gpmM | 1, 4 | | | |
| P13035 | glpD | 1, 2, 3 | P39168 | mgtA | | | | |
| P13482 | treA | 2, 3 | P39285 | yjeP | | | | |
| P14081 | selB | | P39321 | ytfN | | | | |
| P15038 | helD | | P39336 | yjgL | | | | |
| P15254 | purL | 3 | P39453 | torS | | | | |
| P15288 | pepD | 2 | P42620 | yqjG | | | | |
| P15723 | dgt | 3 | P43329 | hrpA | | | | |
| P16916 | rhsA | | P43672 | uup | | | | |
| P16918 | rhsC | | P43675 | gssA | | | | |
| P16926 | mreC | | P45545 | yhfS | | | | |
| P17109 | menD | | P45766 | yhdW | | | | |

Table 3.2 Continued…

| SWISS ID | Name | Ref. | SWISS ID | Name | Ref. |
|---|---|---|---|---|---|
| P17112 | mnmG | | P46837 | yhgF | 3 |
| P17115 | gutQ | | P52054 | yggS | 2, 4 |
| P17580 | spoT | | P52645 | ydbH | 2 |
| P18775 | dmsA | | P52648 | znuC | |
| P18840 | ansA | 2 | P55798 | pphA | |
| P19319 | narZ | | P75780 | fiu | |
| P19636 | eutC | | P75793 | ybiW | |
| P21169 | speC | | P75870 | yccS | |
| P21599 | pykA | 3 | P75914 | ycdX | 4 |
| P23538 | pps | 1, 3 | P76015 | dhaK | 3 |
| P23852 | hepA | | P76143 | lsrF | |
| P23892 | cadA | | P76328 | yodD | |

Table 3.2 Continued…

### C. Found In Both Cell States

| SWISS | | | | SWISS | | | | SWISS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Name | Ratio | Ref. | ID | Name | Ratio | Ref. | ID | Name | Ratio | Ref. |
| P00350 | gnd | 0 | 2, 3, 4 | P06998 | pfkA | 1 | 1, 2, 4 | P22767 | argG | -1 | 1, 2, 3, 4 |
| P00353 | asd | 1 | 1, 2, 3 | P07011 | prfA | 2 | | P22992 | zwf | -2 | 1, 3 |
| P00363 | frdA | 0 | | P07012 | prfB | 0 | 3 | P23480 | bcp | -1 | 1, 2, 3 |
| P00370 | gdhA | 0 | 1, 2, 3 | P07118 | valS | 0 | 3, 4 | P23721 | serC | 1 | 1, 2, 3, 4 |
| P00391 | lpd | 0 | 1, 2, 3, 4 | P07395 | pheT | -1 | 1, 2, 3, 4 | P23839 | yicC | 0 | 3, 4 |
| P00453 | nrdB | 0 | 4 | P07460 | sucC | 0 | 1, 2, 3, 4 | P23851 | rluC | 1 | 4 |
| P00477 | glyA | 1 | 1, 2, 3, 4 | P07682 | lepA | 0 | 3 | P23882 | fmt | 0 | 1, 2, 3 |
| P00479 | pyrB | -1 | 1, 2, 3, 4 | P07813 | leuS | 0 | 1, 2, 3, 4 | P23893 | hemL | 1 | 3 |
| P00490 | malP | 0 | 3 | P08177 | lon | -1 | 3 | P24167 | aroK | 0 | 1, 2, 4 |
| P00496 | purF | 1 | 3, 4 | P08186 | manX | 0 | 1, 2 | P24171 | dcp | 0 | 3, 4 |
| P00509 | aspC | 0 | 1, 2, 3, 4 | P08200 | icd | 0 | 1, 2, 3, 4 | P24183 | fdnG | -1 | |
| P00510 | ilvE | 2 | 1,2 | P08312 | pheS | 0 | 1, 2, 3, 4 | P25516 | acnA | -3 | 3 |
| P00561 | thrA | -2 | 3 | P08324 | eno | -2 | 1, 2, 3, 4 | P25524 | codA | 1 | 3, 4 |
| P00574 | rpoA | 0 | 1, 2, 3, 4 | P08398 | pyrG | -1 | 3, 4 | P25528 | fdx | 1 | 4 |
| P00575 | rpoB | 0 | 1, 2, 3, 4 | P08660 | lysC | 0 | 3 | P25532 | upp | 0 | 1, 2, 3 |
| P00577 | rpoC | 0 | 3, 4 | P08839 | ptsI | -2 | 1, 2, 3, 4 | P25540 | ribE | 0 | 2 |
| P00822 | atpA | 0 | 1, 2, 3 | P08859 | glpK | -2 | 1, 2, 3 | P25553 | aldA | -1 | 1, 2, 3 |
| P00824 | atpD | 0 | 1, 2, 3 | P08936 | hns | 3 | 1, 2, 3, 4 | P25665 | metE | 1 | 3, 4 |
| P00837 | atpG | 0 | 3, 4 | P08997 | aceB | 1 | 3, 4 | P25715 | fabD | 2 | 1, 2 |
| P00864 | ppc | -1 | 3 | P09028 | purE | 3 | 4 | P25739 | purB | -1 | 2, 3 |
| P00882 | deoC | 1 | 1, 2, 4 | P09029 | purK | 2 | 1, 2, 3, 4 | P26427 | ahpC | 3 | 1, 2, 3, 4 |
| P00886 | aroG | 1 | 1, 3 | P09097 | gyrA | 1 | 1, 2, 3, 4 | P26612 | amyA | -1 | 4 |
| P00891 | gltA | 0 | 3, 4 | P09156 | serS | 0 | 1, 2, 3, 4 | P27248 | gcvT | 0 | 2, 4 |
| P00909 | trpC | -1 | 3, 4 | P09159 | speD | 0 | | P27249 | glnD | 1 | |
| P00923 | fumA | -1 | 3 | P09372 | grpE | -1 | 1, 2 | P27298 | prlC | 1 | 3 |
| P00928 | trpA | 1 | 1, 2, 3, 4 | P09373 | pflB | -1 | 1, 2, 3, 4 | P27302 | tktA | 2 | 2, 3, 4 |
| P00934 | thrC | 0 | 1, 2, 3, 4 | P09831 | gltB | -1 | 3, 4 | P27854 | ubiB | 3 | |
| P00936 | cyaA | -2 | | P09832 | gltD | -1 | 1, 2, 3 | P28242 | uspA | 1 | 1, 2 |
| P00955 | thrS | 2 | 3 | P10373 | hisF | 2 | 2 | P28688 | ppk | -2 | 4 |
| P00956 | ileS | 1 | 3, 4 | P11071 | aceK | 1 | | P28694 | mog | 1 | 1, 4 |
| P00957 | alaS | -1 | 1, 2, 3, 4 | P11096 | cysK | 1 | 1, 2, 3, 4 | P29132 | fabI | 3 | 1, 2, 3, 4 |
| P00962 | glnS | 0 | 1, 2, 3, 4 | P11446 | argC | 1 | 4 | P29217 | yceH | 0 | 4 |
| P00968 | carB | -3 | 3, 4 | P11447 | argH | 1 | 3, 4 | P29680 | hemE | 1 | 2 |
| P02339 | ssb | 1 | 1, 2, 4 | P11537 | pgi | 0 | 3, 4 | P30125 | leuB | 1 | 3, 4 |
| P02349 | rpsA | 0 | 1, 2, 3, 4 | P11604 | fbaA | 1 | 1, 3, 4 | P30148 | talB | 2 | 1, 2, 3, 4 |
| P02354 | rpsD | 2 | 3 | P11648 | pepA | -1 | 3 | P30746 | moaB | 0 | 1, 2 |
| P02358 | rpsF | 0 | 1, 2, 4 | P11665 | pgk | 1 | 1, 2, 3, 4 | P30856 | slyD | 1 | 2, 4 |
| P02386 | rplC | 1 | 3 | P11668 | yggE | 0 | 4 | P31119 | aas | 0 | |
| P02387 | rplB | -1 | 3 | P11875 | argS | 0 | 1, 3 | P31120 | glmM | 0 | 3, 4 |
| P02408 | rplJ | 0 | 3, 4 | P12283 | purA | 1 | 2, 3, 4 | P31142 | sseA | 0 | 2, 3, 4 |
| P02416 | rplQ | 2 | 4 | P13029 | katG | 0 | 1, 3 | P31217 | gpmA | 0 | 1, 2, 3, 4 |
| P02990 | tufA/B | 1 | 1, 2, 3, 4 | P13030 | lysS | 0 | 1, 3, 4 | P31465 | yieF | -3 | |
| P02995 | infB | 0 | 3, 4 | P13034 | glpC | 1 | 2, 4 | P31663 | panC | 1 | 1, 2, 3 |
| P02996 | fusA | 1 | 1, 2, 3, 4 | P14178 | pykF | 1 | 2, 3, 4 | P32132 | typA | 0 | 3, 4 |
| P02997 | tsf | 1 | 1, 2, 3, 4 | P14825 | lysU | 0 | | P32661 | rpe | 2 | 2 |
| P03003 | nusA | 0 | 1, 2, 3, 4 | P14926 | fabB | 0 | 2, 3, 4 | P33136 | mdoG | 2 | 1, 2, 3 |

Table 3.2 Continued…

| SWISS | | | | SWISS | | | | SWISS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ID** | **Name** | **Ratio** | **Ref.** | **ID** | **Name** | **Ratio** | **Ref.** | **ID** | **Name** | **Ratio** | **Ref.** |
| P03815 | clpB | -2 | 1, 2, 3, 4 | P15002 | hemB | 2 | 4 | P33195 | gcvP | 0 | 3, 4 |
| P03948 | dapD | 2 | 1, 2, 3, 4 | P15034 | pepP | 0 | 3, 4 | P33363 | bglX | 2 | 2 |
| P04036 | dapB | 0 | 1, 2, 4 | P15046 | ackA | 2 | 1, 2, 3, 4 | P33570 | tktB | -2 | 3, 4 |
| P04079 | guaA | 1 | 3, 4 | P15639 | purH | 0 | 2, 3, 4 | P33918 | rsuA | 1 | 2, 4 |
| P04425 | gshB | 2 | 4 | P15640 | purD | 1 | 2, 3, 4 | P36541 | hscA | 2 | |
| P04475 | dnaK | 0 | 1, 2, 3, 4 | P15716 | clpA | 0 | 3 | P36766 | hpt | 0 | 1, 2, 4 |
| P04790 | tpiA | 2 | 1, 2, 4 | P15877 | gcd | 0 | 2 | P37330 | glcB | -1 | 3 |
| P04804 | hisS | -1 | 3, 4 | P16659 | proS | -1 | 1, 2, 3, 4 | P37350 | ygdH | -2 | |
| P04805 | gltX | 0 | 3, 4 | P16700 | cysP | 2 | 1, 3 | P37595 | iaaA | 0 | |
| P04816 | livK | 3 | 1, 2, 3 | P16703 | cysM | 3 | 1 | P37647 | kdgK | 2 | 2, 4 |
| P04825 | pepN | -2 | 3, 4 | P16936 | speB | 3 | 4 | P37747 | glf | 0 | 3, 4 |
| P05020 | pyrC | -1 | 3, 4 | P17169 | glmS | 0 | 3, 4 | P37759 | rfbB | 1 | 2, 3, 4 |
| P05021 | pyrD | 1 | 1, 2, 4 | P17242 | asnS | 0 | 1, 3, 4 | P37901 | tpx | 4 | 1, 2, 3, 4 |
| P05055 | pnp | -1 | 1, 2, 3, 4 | P17288 | ppa | 1 | 1, 2, 4 | P39171 | iscS | 1 | 3, 4 |
| P05082 | adk | 2 | 1, 2, 3, 4 | P17547 | adhE | -1 | 2, 3 | P39172 | znuA | 1 | 1 |
| P05194 | aroD | 2 | 1, 2, 4 | P17846 | cysI | 2 | 3, 4 | P39184 | pta | -1 | 1, 2, 3 |
| P05313 | aceA | -1 | 1, 2, 3, 4 | P18274 | dksA | -1 | 1, 2 | P39377 | iadA | 1 | 2 |
| P05640 | dapA | 1 | 1, 2, 3, 4 | P18335 | argD | 1 | 1, 3, 4 | P39435 | fabF | 1 | 4 |
| P05793 | ilvC | 1 | 1, 2, 3, 4 | P19245 | clpP | 0 | 1, 2, 4 | P40120 | ydcG | 1 | 2, 4 |
| P06139 | groL | -1 | 1, 2, 3 | P19641 | ispB | -1 | 4 | P40681 | galM | 2 | 1, 2, 3, 4 |
| P06149 | dld | 1 | 3 | P21155 | purC | 0 | 2, 3, 4 | P42607 | uxaC | 0 | |
| P06711 | glnA | 0 | 1, 2, 3, 4 | P21165 | pepQ | 0 | 2, 3, 4 | P42632 | tdcE | -1 | |
| P06715 | gor | 0 | 1, 2, 3, 4 | P21170 | speA | 0 | 3 | P52643 | ldhA | -1 | |
| P06721 | metC | 1 | 3 | P21179 | katE | -1 | 3, 4 | P52697 | ybhE | 1 | 3, 4 |
| P06958 | aceE | -1 | 1, 2, 3, 4 | P21346 | greA | 0 | 1, 2, 3, 4 | P71295 | fbaB | 0 | 3, 4 |
| P06959 | aceF | 2 | 1, 2, 3, 4 | P21499 | rnr | 0 | 3 | P76492 | yfbU | 3 | 2, 3, 4 |
| P06977 | gapA | 1 | 1, 2, 3, 4 | P21889 | aspS | 0 | 1, 3, 4 | P76558 | maeB | -3 | 3 |
| P06981 | guaB | 0 | 1, 2, 3, 4 | P22106 | asnB | -3 | 2, 3, 4 | P77241 | ppiD | -1 | |
| P06982 | gyrB | 1 | 3 | P22256 | gabT | 1 | 3 | P77254 | der | -1 | |
| P06986 | hisC | 2 | 3, 4 | P22257 | tig | 0 | 1, 2, 3, 4 | P78258 | talA | 0 | 2, 3 |
| P06994 | mdh | 3 | 3, 4 | P22259 | pck | 1 | 3, 4 | P80063 | gadA | -6 | |

likely reflects their relatively low abundance (see below). The remainder of the proteins were identified at least 11 times. These 350 proteins account for a majority (68%) of the 4,845 total non-redundant identifications. Twelve proteins were identified in at least 50 separate fractions. These proteins, which account for 23% of all non-redundant identifications, include known abundant proteins such as Ef-G, DnaK, Ribosomal Subunit S1, and ClpP. The remaining 8 proteins are; PyrB, GyrB, GadA, AceA, IscS, LpdA, MetE, and GltB were also identified in $\geq$ 50 fractions.

To further evaluate this correlation between incidence of identification and protein levels, we examined the correlation between frequency of identification and codon usage. Codon adaptation indices (CAI) (Sharp and Li, 1987) predict potential expression level based upon codon usage relative to the codons used in a reference set of highly expressed genes. The bars in Figure 3.3B show the distribution of CAI values (Higher CAI = higher predicted expression levels) for the 520 proteins we identified as compared to the genome. Although we identify proteins from all predicted expression classes, identifications are biased towards proteins predicted to have higher expression. The line in Figure 3.3B shows the average number of times each protein in each bin of CAI values was identified. The number of identifications is clearly positively correlated with predicted expression levels.

### Functional Analysis Reveals a Similar Distribution between Cell-States

Based on functional classifications, (http://genprotec.mbl.edu/) (Serres *et al.*, 2004) both stationary and exponential phase proteomes show significant overrepresentation in proteins assigned to metabolism and underrepresentation in the transport and cell structure categories compared to the representation of these categories in the complete genome (Figure 3.4). The functional distribution of identified proteins is not significantly different between exponential and stationary phase. Of note however, is the higher than average representation of proteins from the category of cell processes in stationary-unique genes relative to log and stationary cells, even compared to the genome. There is also a higher representation of genes in the 'Regulation' category. Since microarray based studies, and other proteome efforts (Grunenfelder *et al.*, 2001; Tani *et al.*, 2002) support the idea that most genes products ≈87-90% are present or change expression subtly at different cell-states, overall functional assignment is expected to be non-remarkable. VanBogelen et al. describe methods to differentiate cell states of microbes using integrated data from a multiple array of biochemical information including expression and protein ratios (VanBogelen *et al.*, 1999b). Surprisingly, functional classification of just those proteins identified uniquely in exponential or stationary phase from this methodology reveals more dramatic differences between cell-states. The content of one of these altered categories, Cell Processes, is shown in Table 3.3. It has been filtered for those proteins that were found in both cell states in 'Cell Processes.' Comparing the known and predicted gene functions of these proteins, this category appears over-represented for proteins involved in stress and starvation responses. This was determined by examining the gene annotations in Genbank, Genprotec, and the SWISS PROT for keywords including, "stationary," "stress," etc., similar to Tani or additional references illustrating increases in expression during stationary phase (Hengge-Aronis, 1996, 2002b; Tani *et al.*, 2002). This is distinct from the comparative

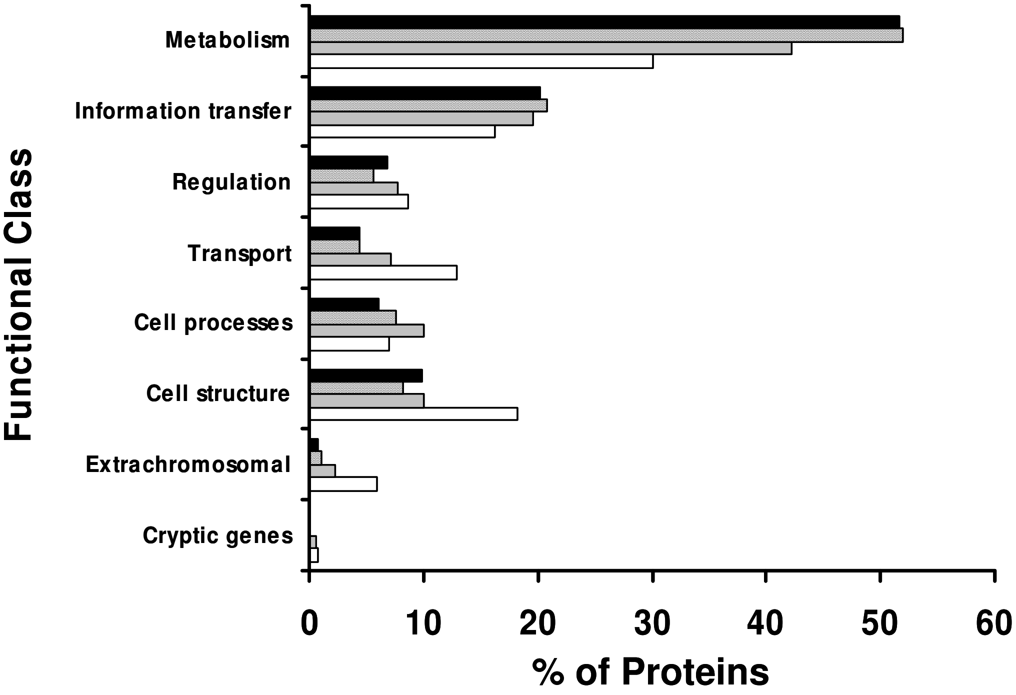


Figure 3.4.) Classification of proteins from exponential and stationary phase. Classification of the proteins identified in each cell state and the genome. Functional classification from Blattner et al., and Riley et al., (was used to classify the non-redundant list of identified proteins (MOPS Exponen. phase ID's, Black; Stationary Phase ID's, Hatch Up, Unique to Stationary Phase Grey, Genome, White). The distribution of functions in the MG1655 Genome is shown for comparison (White).

Table 3.3.) Protein identifications from a single functional category.
Annotated description of specific protein functions from a set of proteins that were uniquely assigned to either mid-exponential or late-stationary phase. The category ‘Cell Processes’ is illustrated because numerical inspection of the differences shows the largest trend in this category. Footnotes provide references for proteins annotated as stationary specific outside of gentprotec.

| Cell Process Proteins From Log Phase | | |
|---|---|---|
| SWISS PROT ID | Gene Name | Function |
| P35340 | acpC | Detoxification |
| P25537 | cafA | Cell Division |
| P16244 | cpxR | Osmo Regulation |
| P08506 | dacC | Peptidoglycan Syn. |
| P07862 | ddlB | Peptidoglycan Syn. |
| P24991 | dsbA | S-S Bond Formation |
| P26266 | fepE | Fe Uptake/Storage |
| P05380 | GroES | Chaperonin |
| P33137 | mdoH | Osmo Regulation |
| P05838 | sspA | AA Starvation |
| P38489 | nfnB | Metabolism |
| P25746 | rfaG | LPS Synthesis |
| P09030 | xthA | DNA Repair |

| Cell Process Proteins From Stationary Phase | | |
|---|---|---|
| SWISS PROT ID | Gene Name | Function |
| P03026 | acrA | Decreases Aerobic Genes |
| P11056 | bfr | Iron Storage, Inc. in Stat. |
| P77239 | cusB | Metal/Cellular Resistance |
| P27430 | dps | DNA Protection Starvation |
| P04286 | FtsI | Septum Formation |
| P10121 | FtsY | Protein Transport |
| P13482 | freA | Osmo Protection/Reg. |
| P25520 | galU | Stationary Phase Survival[1] |
| P17112 | gidA | Unknown |
| P27246 | marA | Induces Stress Genes[2] |
| P24231 | pmbA | Possib inc. C Storage |
| P55798 | pphA | Misfolded Protein Recog. |
| P09157 | sodB | Oxidative Stress |
| P17580 | spoT | Increase in pppGpp |
| Q46812 | ssnA | Expressed In Stat. Phase[3] |
| P25522 | thdF | Cell Detoxification |
| P75780 | YbiL | Prob. Fe Transport |

analysis provided in Table 3.2, which describes the proteins found in both cell states relative to occurences in other proteomes.

## Differences Between Exponential and Stationary Phase Proteomes – Abundance Changes

We evaluated potential changes in protein abundance between cell states by comparing the frequency with which a protein was identified from exponential and stationary phase samples. In addition to the 131 proteins that were not found at all in the exponential phase samples, there were 25 proteins that were identified in >2 fold more fractions in stationary phase than in exponential phase. There were 48 proteins that had >2 fold more occurrences in exponential growth than in stationary phase in addition to the 158 that were not found in stationary phase. ClpB, a heat-shock protein involved in proteolysis was identified in a significant number of fractions in both cell states, but its expression is known to increase due to stress, stationary growth etc (Weichart *et al.*, 2003). ClpB was identified in 32 more fractions (3.9x) in stationary phase cells than in mid-exponential cells. Likewise, Tkt2 (transketolase) was seen in 4 fold more fractions in stationary phase than exponential cells. By contrast, DapD, which is involved in cell membrane synthesis was identified in 4 fold more fractions (20 vs 5) in exponential than stationary cells. A list of the significant differential identifications are available in the supplementary data on the website, (http://eep.tamu.edu), the Appendix and metabolic changes are highlighted in figure 3.6 (discussed later).

### Differences Between Exponential and Stationary Phase Proteomes – Changes in Elution Profiles

Based on the number of identifications in each chromatographic dimension, we calculated an average peak position for the elution of each protein. Surprisingly, although many proteins are found in both cell states, many of them have dramatically altered chromatographic behavior in exponential and stationary phase samples.   The peak positions of 42 proteins shifted 3 or more fractions between exponential and stationary phase anion-exchange separations performed at pH 7.50, while 61 proteins had shifted peak positions at pH 8.75.  This likely underestimates the chromatographic differences in elution positions; although the elution of proteins into specific fractions is quite reproducible when comparing different exponential phase proteome samples or different stationary phase samples by 2D gels (data not shown), the pattern of proteins seen in comparable fractions from exponential and stationary phase are dramatically different (Niu et al, in preparation).  Limitations in the resolution of the chromatography mean that many real changes in elution were not scored due to insufficient separation.

## Discussion

Comparing the exponential and stationary phase proteomes, we see differences in the proteins detected as changes in the levels of proteins assayed by the identification 'hit' frequency from nondenaturing LC/MS spot intensities on 2D gels and changes in the elution positions of proteins in both chromatographic dimensions.  Our interpretation of these results is then focused on three questions:  First, do the changes we see accurately reflect what is happening in the cells?  Second, what is the molecular basis for the differences we detect between exponential and stationary phase proteomes? Third, what is the biological significance of these changes between cell states?

## Changes in Protein Content Between Exponential and Stationary Cells

Figure 3.5 compares the number of identified proteins from the exponential phase growth in MOPS medium to the total results from the cells grown to stationary phase. Together, both cell states identified a total of 520 proteins which is about double the number we identified previously (Champion *et al.*, 2003). Many proteins that are present in one cell state, but absent in another, correspond to proteins known to be induced or repressed in stationary phase. Changes in expression level that drop a protein below the detection threshold in either proteome would score as all-or-none differences. For some cases we can also tell that a protein has dropped below the detection limits of staining on the 2D gels. Failure to detect a protein can also occur for a variety of less interesting reasons such as ion suppression in the MALDI due to changes in the dynamic range of cofractionating proteins.

Our observation that many proteins change their chromatographic properties between exponential and stationary phase raises the possibility that some proteins missing in only one proteome have changed so that they no longer bind to the first dimension of separation. However, the complexity of the ion-exchange flowthrough makes it difficult to handle using our methods. It is important to remember that the failure to identify a protein by MALDI-MS in one state does not rule out its presence in the other. Indeed, some proteins identified as ‘unique’ to one cell state are clearly false negatives in the other, based on our prior biological knowledge. For example, several

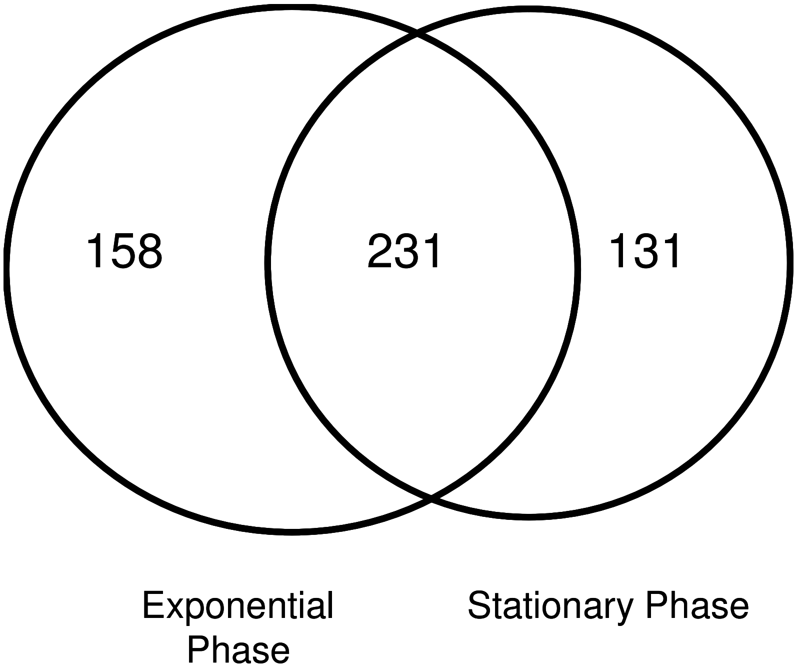


Figure 3.5.) Venn diagram of protein identifications between cell phases. Distribution and overalp of the total unique protein ID's from the exponential and stationary proteomes. Both proteins identify comparable numbers of proteins, and the overlap between the two cell states is substantial, (avg. = 62%) and each cell state also describes a large unique set of identifications. This is the sum of each proteome at two anion-exchange pH's per cell state.

components of the DNA replisome were only found in stationary phase, although the DNA replication machinery is clearly present in growing cells.

Overall, we found 131 distinct proteins that were unique to stationary phase cells. Comparisons with other studies can be used to filter out likely false negatives that should be present in exponential phase as well. We compared our stationary-specific identifications to our previous work (Champion *et al.*, 2003), the 404 high confidence ID's from Corbin (Corbin *et al.*, 2003), the mid-exponential data from Cybercell and the late log data from SWISS2D. 56 of our stationary phase-specific gene assignments were found in one of these other exponential phase datasets. Many of these are likely to be false negatives that we missed in our exponential phase samples, which yields a net 75 stationary-phase specific proteins identified unique in this work.These cross-hits are also listed in Table 3.2.

Some of them, such as Dps and TktB, may represent false positives from the other published exponential phase samples. In particular, Dps is known to be strongly induced in stationary phase (Almiron *et al.*, 1992; Altuvia *et al.*, 1994), has never been identified in any exponential-only cell preparation in our laboratory, and microarrays (Tani *et al.*, 2002), but was assigned as a exponential gene product in Corbin et al. (Corbin *et al.*, 2003). The combined results of this comparison, as well as the published on-going SWISS 2D *E. coli* proteome project are presented as Figure 3.7 (Hoogland *et al.*, 2000; Tonella *et al.*, 2001).

Although it is not possible to quantify proteins based on the intensities of peptide peaks in MALDI spectra, we show here that the frequency of identification of a protein by multidimensional LC fractions correlates with protein abundance. Proteins will generally be present above their detection thresholds in fractions consistent with their peak profiles. MS is a concentration dependent measurement, and peptides at higher concentration typically generate better signal-to-noise averages. The 10 most commonly identified proteins in our samples were identified in 1001 instances, or about 20% of the total identifications. These proteins include DnaK (229 times), GltB (200 times), FusA [EF-G] (83 times), and Tig (82 Times) as the top 4. These are all known to be highly abundant proteins. The least identified proteins were identified in only one fraction and include proteins like ZraR, (Regulatory Protein) and GlgB (Alpha glycan branching enzyme) which represent low abundance proteins and those we were unable to identifiy our criteria. These identification frequencies are consistent overall with changes in relative abundance.

It is important to view this as a qualitative correlation and not a quantitative one. First, the dynamic range of identification frequency is sharply limited to less than two orders of magnitude, while proteins in *E. coli* are known to vary in abundance by about five orders of magnitude (Corthals *et al.*, 2000). Second, while there is a correlation, specific proteins may give disparate results due to the same variables that affect whether a protein is identified at all. Nevertheless, changes in identification frequency can provide preliminary evidence that the abundance of a protein has increased or decreased. For example, GadA was identified in 66 separate Q/Phe fractions in stationary phase, but only once in exponential cells. Similarly, DapD was identified 20 times in exponential phase and only 5 times in stationary phase cells.

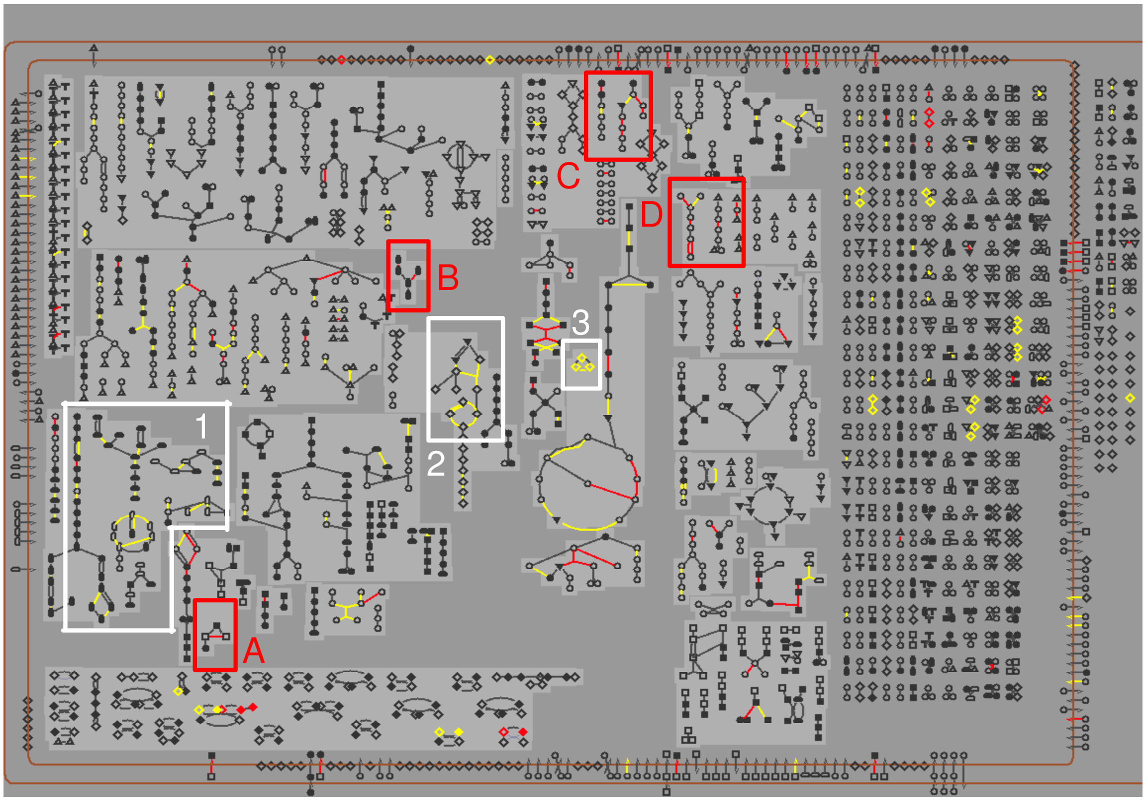


Figure 3.6.) Distribution of cell state regulated metabolic proteins in *E. coli*. This metabolic map generated with BioCyc (2002) illustrates pathway specific differences between the proteins identified in each cell state. The expression ratios are red (Expressed in stationary phase preferentially) to yellow (Expressed in log phase preferentially). Highlighted in red and white respectively are several pathways that illustrate cell-state distinctions: Stationary-phase Pathways: A.) Trehalose degradation B.) ppGpp biosynthesis, C.) Anaerobic respiration, and D.) Amino acid degradation. In exponential cells, 1.) *De novo* nucleotide synthesis and salvage, purines and pyrimidines. 2.) Fatty acid biosynthesis and elongation, and 3.) Pyruvate dehydrogenase complex (glycolysis), are highlighted.

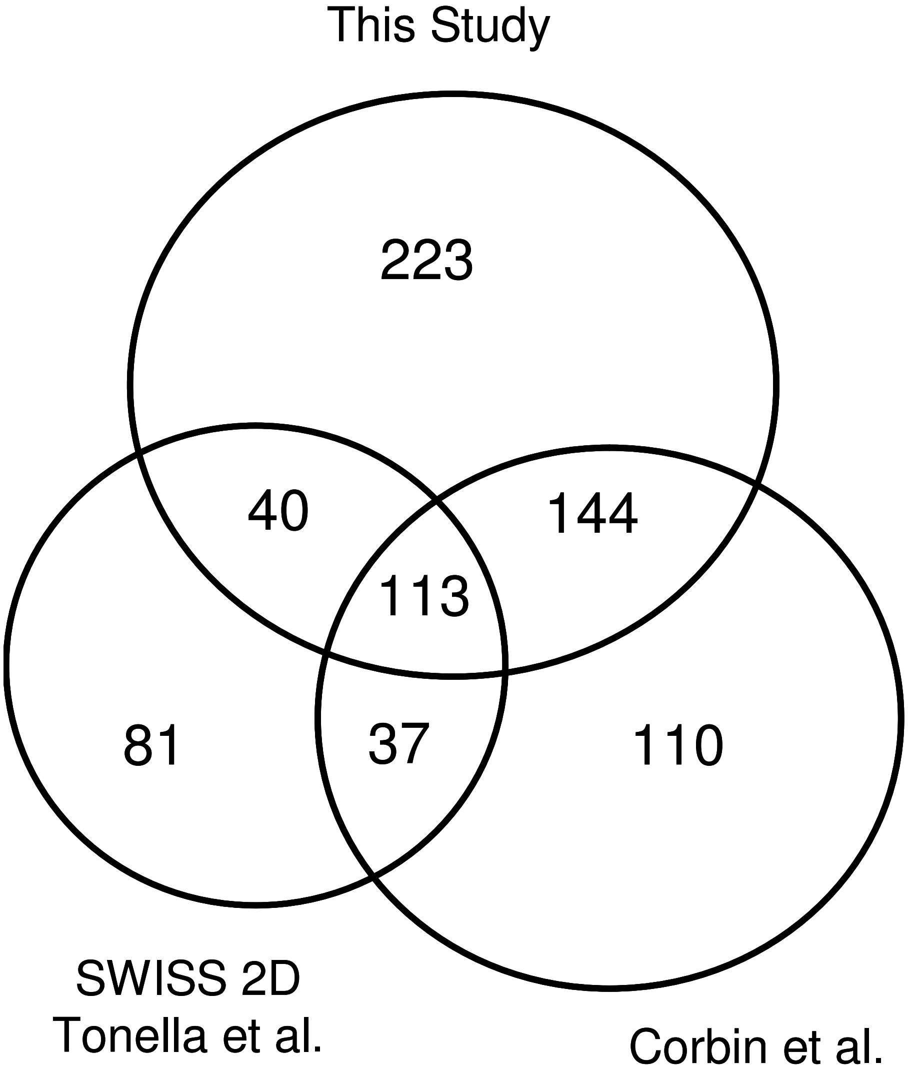


Figure 3.7.) Venn diagram comparing comprehensive *E. coli* proteomes. This Venn diagram shows the distribution of the unique protein identifications made several whole -cell proteomes in *E. coli*. The SWISS 2D utilized 2D gel electrophoresis spanning multiple pH range IEF gels coupled to peptide mass fingerprinting for identification. Corbin et al., utilized MDLC LC/MS/MS separation of whole-cell digests. This study utilized pre-digestion fractionation, and peptide mass fingerprinting for ID.

We used the metabolic mapping tools provided by the EcoCyc database to map changes in expression on a metabolic diagram of *E. coli* (Karp *et al.*, 2000; Karp *et al.*, 2002). Figure 3.6 highlights many pathways which appear to have changes in protein levels between cell states. Some of these, such as trehalose synthesis are known to be induced in stationary phase growth (Cayley *et al.*, 1991; Strom and Kaasen, 1993; Welsh *et al.*, 1991). Proteins involved in menaquinone and THF (tetrahydrofolate) synthesis are significantly overrepresented in stationary cells, and their role in alternate anaerobic respiration has been discussed by Unden et al., (Unden, 1988; Unden and Bongaerts, 1997). However, their specific role in stationary phase survival has not been described. The methylcitrate cycle, responsible for propionate utilization, appears to be present preferentially in stationary phase but its mRNA levels have been reported as constant under numerous conditions tested (Brock *et al.*, 2002). Additionally, several components of LPS and KDO synthesis are found to be present or bias towards exponential cells, consistent with their requirement for rapid cell growth and division. Although they are not highlighted in the figure, enzymes involved in acetate utilization and the BarA sensor kinase pathway are also overrepresented in stationary phase cells. Glycolysis is unchanged within our rough limits of detection which is consistent with other measurements made by array s data.

We also compared our list of proteins induced in stationary phase to changes in determined by other studies. Tani et al. (Tani *et al.*, 2002) investigated changes in mRNA expression of stationary specific genes in Lrp$^+$ and Lrp$^-$ background strains, and identified 53 genes with known stationary phase expression, and an additional 102 stationary-specific genes thus implicated by their research (Tani *et al.*, 2002). We identified proteomically 24 proteins from this list of from stationary phase *E. coli* cells via MS identification.  Of these, 13 were identified from their class of novel, predominantly uncharacterized stationary phase specific genes, and 14 of the 24 genes were not identified in exponentially grown bacteria. Of the unique uncharacterized hits there were 5 gene products which were identified at least 4 times.  Further examination of other proteins identified in both cell states, but annotated as 'stationary phase' reveals cases where identification frequency analysis is illuminating.  KatE, GadA, and TktB were found at both cell states and both KatE and GadA are known to be induced in stationary phase (De Biase *et al.*, 1999; Mulvey *et al.*, 1988; Sak *et al.*, 1989).  In our study, KatE was identified 23 separate times in stationary phase cells and 12 times in mid-exponential cells.  GadA, also, was identified 66 times in stationary phase, and only once in mid exponential cells, illustrating the ability of frequency to differentiate between a binary identification and a change in abundance.  This technique enables a graduated identification of a stationary phase response instead of a + or – result from a traditional proteome.

### Elution Changes Occur Between Cell States

The most surprising finding from these experiments was the degree to which many proteins altered their chromatographic properties between exponential and stationary phase. Physiological change on this scale seems to involve more than just expressing different polypeptides – the changes in the chromatographic properties of proteins found in both exponential and stationary phase were extensive and dramatic.

Comparing exponential and stationary phase cells at pH 7.50 and 8.75 revealed 94 proteins that shifted by more than 3 anion-exchange fractions from their average elution position (Table 3.5). Of these, 9 proteins were found to shift at both pH 7.50 and pH 8.75, which means the changes in elution were not an artifact of the pH of separation. Because the correlation with levels is qualitative, smaller shifts in elution position cannot be reliably determined from numbers of identifications. However, in the 2-D gel analysis of ion exchange fractions from the two physiological states, similar fractions are almost unrecognizable as pairs. Since most of the proteins identified in this study do not significantly change elution position between cell-states, these changes do not reflect a systematic shift such as is seen when we change the pH of the anion exchange step. Graphically, elution changes can also be presented as a virtual elution profile in figure 3.8. In this case, several proteins are shown which have substantial elution shifts between exponential and stationary growth (SerC) and those that do not move much, RpoA).

Many proteins that shift their elution, such as include GapA (glyceraldehyde 3-P dehydrogenase), SerC (phosphoserine amino transferase), and FrdA (fumerate reductase) as examples, have not been previously thought of as changing during the shift from log to stationary phase. Of course, these kinds of changes would be missed by virtually every other approach to cell physiology.

Although this is the first large-scale study cataloging changes in the chromatographic properties of proteins as a function of physiology of which we know, there is ample precedent for changes in the post-translational modification or subunit

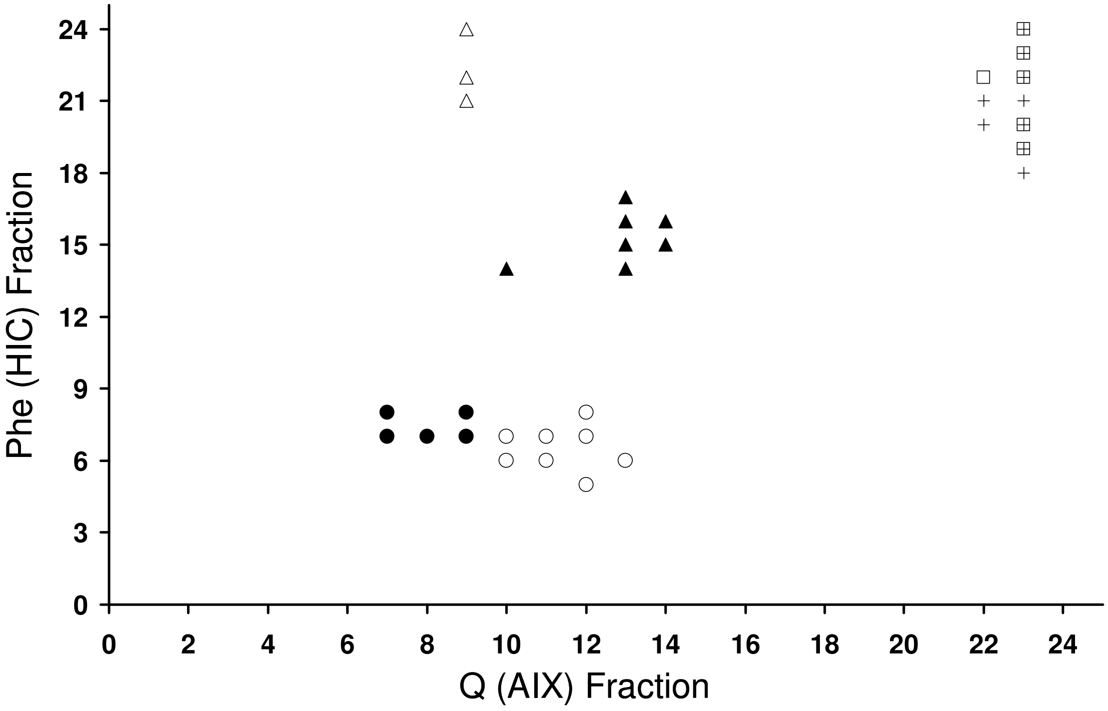


Figure 3.8.) Elution profiles of changes in protein elution at different cell states. Changes in average elution position of ≥ 3 fractions were observed for 117 polypeptides between mid-exponential and stationary phase cells. Shown are elution profiles for three example proteins, which shift in either or both chromatographic dimensions. PpC (P00864) exponential Δ and stationary ▲.SerC (P23721) exponential ○ and stationary ●. RpoA (P00574) exponential □ and stationary + does not alter its elution as a function of cell state like most of the proteins identified in this study.

Table 3.4.) Potential interacting protein pairs.
Each entry is the list of protein pairs that were found to co-fractionate at both pH 7.50 and 8.75 in their average elution position in Mid-exponential cells, stationary phase cells, and those identified in both growth conditions. The lists are sorted in alphabetic order, non redundant for all interactions. E.g. PheS- PheT is not re-listed as PheT – PheS. There are 81 pairs in log-phase, 52 in stationary, and 18 shared by both cell states.

**Mid Exponential**

AceA-MetK
aceA-rpe
aceA-sbt
aceA-slyD
alaS-galM
aldA-rpsA
argC-katE
argD-lysS
argG-iscS
argH-atpD
argH-dnaK
argH-iscS
argH-rpsF
aroA-nfnB
aroA-talB
asnS-leuB
asnS-metE
asnS-valS
aspC-sucD
aspC-tktA
bglA-yfbU
clpB-rpoA
clpP-dnaK
codA-ilvE
cysK-pykF
cysK-sucC
cysK-sucD
cysM-moeA
cysM-proS
dapA-yrdA
dapB-fdx
dapD-sucC
dapD-sucD
deoC-mdoG
dnaK-ilvE
dnaK-ldhA
dnaK-rpsF
dut-gnd
folE-tig
fusA-katG
fusA-trxB
gabT-guaA
gabT-pyrB
gapA-gpmA
ilvE-rpsF
gdhA-iscS
glnS-yfbU
glyA-tpiA
gpmA-livJ
greA-ybhE
gshB-ilvE
guaA-pfkA
guaA-pyrB
guaB-ssb
hisA-yifE
ilvE-rpsF
katE-pepP
leuB-metE
leuB-serS
moaB-yceH
ndk-pykF
ndk-sucD
nusA-pnp
nusA-yicC
pck-talB
pepP-speD
pepP-sseA
pnp-yicC
ppa-ybhE
ppiB-rplI
purK-znuA
pykF-sucC
pyrB-speD
rplQ-sseA
rpoA-rpoC
sbt-slyD

**Stationary Phase**

acnA-aspS
acnA-pgi
aldA-serS
argG-rpsF
asnS-leuS
aspS-cysK
aspS-gcvT
aspS-zwf
bfr-katE
bfr-moaB
bfr-mog
bfr-sseA
clpP-ldhA
clpP-rpsF
dapD-pck
dcp-purC
eno-gltX
fabB-gltX
fabB-ppa
fbaB-talB
gadA-leuS
gadA-pyrB
gadB-kdgK
gadB-pyrB
gadB-upp
gadB-uxaC
gltA-tktA
gltA-upp
gltX-ppa
glyA-serC
gnd-gor
gnd-tktB
gpmA-pgk
greA-lpd
greA-yieF
guaA-yieF
hisS-purA
katE-sseA
leuB-pyrI
maeB-pnp
maeB-purF
moaB-yieF
pepN-serS
pnp-slyD
purA-tktA
pyrI-serS
thrC-treA
upp-uxaC
yceH-yieF

**Both Cell States**

argG-gdhA
aspC-sucC
aspS-gnd
clpB-ldhA
cysK-gcvT
dapB-yceH
eno-gnd
gltA-kdgK
gltA-purA
gltA-uxaC
gltB-tig
katE-speD
kdgK-uxaC
moaB-mog
moaB-sseA
pheS-pheT
pnp-purF
proS-purH

Table 3.4 Continued…

| **Mid Exponential** | | **Stationary Phase** | **Both Cell States** |
|---|---|---|---|
| eno-serC | sucC-sucD | kdgK-upp | |
| fabI-purH | trxB-valS | ldhA-rpsF | |
| fabZ-tsf | | leuB-purH | |

Table 3.5.) Proteins with altered elution between exponential and stationary phase. This table lists the 94 proteins that had alterted anion-exchange elution properties between exponential and stationary phase. Proteins for which two shifts are given were found shifting between cell states at both pH's (e.g. Log pH 7.50 to Stat 7.50; & Log 8.75 to Stat 8.75).

| SWISS ID | Avg..Shift Log - Stat | SWISS ID | Avg..Shift Log - Stat |
|---|---|---|---|
| P00350 | 3.7 | P10373 | 6.6 |
| P00353 | 3 | P11096 | 4.2 |
| P00363 | 3.0, 8.0 | P11537 | 3.3 |
| P00477 | 5.3 | P11604 | 3.3 |
| P00479 | 3.7 | P11665 | 3.5 |
| P00509 | 4.7 | P12283 | 4.6, 5.1 |
| P00510 | 3.3 | P13029 | 5.6 |
| P00561 | 3 | P14178 | 4.4 |
| P00837 | 5 | P14926 | 4 |
| P00864 | 3.9, 3.0 | P15716 | 5 |
| P00886 | 6.5 | P15877 | 15 |
| P00891 | 3.1, 3.9 | P16659 | 4.2 |
| P00923 | 12.5 | P16700 | 3 |
| P00936 | 13.7 | P16703 | 3.8 |
| P00957 | 4.6 | P17846 | 7.2 |
| P00962 | 4.1, 3.8 | P19245 | 3 |
| P00968 | 8 | P21499 | 10.2 |
| P02354 | 5.3 | P21889 | 3.9 |
| P02387 | 10.3, 3.5 | P22106 | 3 |
| P02995 | 8.7 | P22992 | 3 |
| P02997 | 3.5 | P23721 | 3.4, 3.9 |
| P03003 | 6 | P23839 | 3.3 |
| P03948 | 4.5 | P23882 | 3 |
| P05640 | 3.3 | P24171 | 3.2 |
| P06139 | 4.8 | P25524 | 5.8 |
| P06149 | 12 | P25540 | 9 |
| P06715 | 3.1 | P25553 | 3.3 |
| P06959 | 5.8 | P25739 | 3 |
| P06977 | 5.5, 4.2 | P26427 | 4.2 |
| P06981 | 3 | P27249 | 3 |
| P06998 | 3.4 | P27302 | 3.2 |
| P07012 | 3 | P27854 | 3 |
| P07118 | 5.6 | P28688 | 5.3 |
| P07395 | 3 | P29132 | 4.4, 4.4 |
| P07460 | 3.6 | P30125 | 5.2 |

Table 3.5 Continued…

| SWISS ID | Avg..Shift Log - Stat |
|---|---|
| P07682 | 10 |
| P07813 | 4 |
| P08177 | 8 |
| P08200 | 3.8 |
| P08312 | 3 |
| P08398 | 3 |
| P08859 | 7.3 |
| P08936 | 4.7 |
| P09029 | 3 |
| P09097 | 14 |
| P09372 | 9 |
| P09373 | 4.7 |
| P33136 | 7.3 |
| P33195 | 3.5 |
| P33363 | 8 |
| P36766 | 4.5 |
| P37350 | 3.8 |
| P37647 | 3.3 |
| P37759 | 3 |
| P37901 | 3 |
| P39171 | 4.2 |
| P39435 | 8.2 |
| P42607 | 3.7 |
| P80063 | 11.7 |

structure of proteins that could account for changes in elution profiles. Some covalent modifications are not detectable on 2D gels, and changes in subunit structure can only be examined in the native state. Ishihama et al. (Ozaki *et al.*, 1991) described the changes in chromatographic behavior of core and holoenzyme of RNA polymerase due to association of various sigma factors, which associate as a response to cell and nutritional conditions. In their work, core bound to $\sigma^{70}$ was chromatographically distinct from forms of holoenzymnes recovered from stationary-phase grown cells. These forms of the enzyme were found to contain $\sigma^{70}$ but changes in these peaks of were observed over a 72 hour growth curve. Groat et al., on stationary phase growth in *E. coli* also suggests large ”molecular realignments” in response to starvation, which is consistent with data presented here(Groat *et al.*, 1986). Transcriptional studies were also performed on the different ‘forms’ of holoenzyme, and promoter recognition differences were seen between these chromatographically distinct RNAP’s. The large dynamic change in metabolic behavior we observe associated with stationary phase is consistent with the idea that these pathways could alter their multi subunit composition as well.

## Protein Complexes

The experiments described above may dramatically change our picture of the differences in *E. coli* between growing and nongrowing cells. While it has long been known that significant changes in gene expression are seen in stationary phase cells (Kusano and Ishihama, 1997; Ozaki *et al.*, 1991; Tani *et al.*, 2002; Weichart *et al.*, 2003), our results suggest that these changes in expression of mRNAs and gene products represent only a small fraction of the differences in the intracellular biochemistry of these two physiological states.

As noted above, one possible molecular explanation for changes in the elution patterns of a protein is a change in the composition of a multisubunit complex. Previously, we have argued that large-scale cofractionation studies provide suggestive evidence for the existence of complexes (Champion *et al.*, 2003) and recently complexes such as the degradosome have been characterized in *E. coli* and other bacteria (Carpousis, 2002). Several studies have also demonstrated the utility of examing differences in protein content as an indicator of cell-state/condition(McAtee *et al.*, 1998; Perrot *et al.*, 2000; VanBogelen *et al.*, 1999b). Additionally, global changes in cell composition, which likely corresponds to differences in gene expression and protein content as well were described by Makinoshima et al. (Makinoshima *et al.*, 2003) In their case in particular, differences in cell buoyancy were RpoS dependant. Table 3.4 shows proteins that cofractionated over two dimensions of chromatography both at pH 7.50 and 8.75. Complexes were observed in both or either cell-states and in total, 90 distinct non-redundant pair wise interactions were observed between the two pH's of exponentially grown cells, and 70 unique potential interactions were seen in stationary phase cells. As noted previously (Champion *et al.*, 2003), these lists should be viewed as strongly enriched for potential interactions rather than as claims for the existence of specific complexes..

Further study of individual proteins is needed to examine whether proteins are in a complex or merely cofractionating by coincidence. Some of these cofractionations are certainly significant. In addition to changes in cofractionation between cell states, we examined which proteins both cofractionate and have elution shifts between cell states. DnaK and IscS (NifS homolog) co-fractionated at both pH's in both cell-states. In both cases, however, the protein pair is found in a lower anion-exchange fraction at pH 8.75 than it is at pH 7.50. There are additional protein pairs that that demonstrate the 'super shift' of cell state and co-elution. GltA and PyrB for example co elute between log and stationary cells and shift between cell states as well (5 Q Fractions). The extent to which co-elution and changes in fractionation are significant is not entirely clear, but changes in complexes in which we identify more than one component would be expected to show up by this analysis. Several dozen proteins that coeluted between log and stationary phase also had changes in absolute elution position.

## Conclusions

In sum, we demonstrate here, one of the largest sets of validated proteins assignments made to date in *E. coli*, performed on a scale compatible with physiologic experimentation, and were able to assess the differences between identical cells grown under multiple conditions. Additionally, the unique aspects of the sample preparation yield a large amount of additional information about the nature of the macromolecular components that exist within the cells and the potential changes they undergo as their environment becomes less hospitable.

Orthogonal approaches also provide additional validation of the protein assignments. Validation of protein assignments made from a combination of multiple growth conditions, the overlap between independent projects, and cross-correlation with parallel 2D gels, generates a significantly better vetted dataset for these hundreds of proteins in *E. coli*. The physiologic utility of these profiles also provides a snapshot into system-wide and specific changes that occur when organisms change growth state. These data are consistent with the hypothesis that changes in cellstate are associated with changes in the larger intramolecular architecture in the cell. We observe unique changes in protein content due to cell state, protein abundance, and perhaps more importantly, changes in the elution characteristics of intact proteins, which likely reflects changes in protein association. We would also expect that the gains in protein identification, differential expression and elution profiling would benefit from the addition of tandem MS/MS data, providing a different confidence in assigned proteins and likely greater numbers of validated protein hits. As a demonstration of the utility of these approaches, however, the rapidity and ease of use of MALDI TOF MS, and the lessened data complexity of functional proteomics in bacterial systems, makes this an ideal approach for those labs not readily equipped to dedicate massive amounts of MS and MS/MS time to process and analyze samples.

As our ability to identify more proteins and protein complexes becomes available, we can slice up the growth changes of *E. coli* into increasingly smaller temporal slices, and hope to appreciate the vast changes cells undergo as nutritional availability changes. Our conclusion is that there are significantly larger global & structural changes occurring within the cell under transitions from steady-state to limited growth conditions than previously captured.

# CHAPTER IV

# ‘RATIONAL PROTEOMICS’: DIRECTED APPROACHES TOWARDS TARGETING SPECIFIC PROTEIN SUBSETS

## Summary

Chromatography can decrease the complexity of proteome samples and alter the types of proteins identified. Pilot experiments separating the cellular proteins of *E. coli* via heparin HPLC and analyzing the fractions resulted in the identification of 118 proteins from *E. coli* lysates from liquid cultures grown to both exponential and stationary phase. Extension of the method to utilize tandem MS/MS techniques on both SDS PAGE gel bands, and the reversed phase separation of digested peptides resulted in the addition of 1358 protein ID’s. This corresponded to a non-redundant set of 318 identified proteins, including 240 from exponential phase cells and 234 from stationary phase cells. 156 proteins were shared between the two cell states.

## Introduction

Identifying all or most of the proteins in a cell under various conditions has proven difficult, and experiments performed to identify many of the proteins present within a cell typically identify only highly abundant proteins with high confidence. Large numbers of single spectrum ‘hits’ are reported for proteins predicted or known to be in low abundance, but these assignments rarely carry the same weight as the identifications of more highly abundant proteins (Baldwin, 2004; Carr *et al.*, 2004). Therefore, there is considerable interest in refining the techniques to increase the numbers of identified proteins and, more importantly, to find particular sets of proteins for further study. For this study, we utilized alternative separation as a means to both fractionate and enrich for proteins produced in *E. coli*. Our rationale is that modifying the sample preparation prior to mass spectrometry is capable of generating novel data of potentially more focused physiologic interest. This can be envisioned as a hybrid of two extremes; a pull-down experiment which is focused on the fewest number of components from a large cell extract and complete analysis of whole cell lysate, using MDLC and MUDPIT experiments designed to detect everything (Washburn and Yates, 2000; Washburn *et al.*, 2001; Wolters *et al.*, 2001). The addition of tandem MS/MS data provided additional validation through sequence fragment information and larger numbers of ID's due to the greater information content available per peptide.

Unlike genome sequencing, the 'status' at which a proteome is complete is limited by our ability to understand biologically when each particular protein species is actually expressed and present. By this standard, a complete proteome for any organism is not likely to be obtained soon. One of the major obstacles to these approaches is the dynamic range and complexity of proteins present within a cell. In blood plasma for example, the concentration of high abundance proteins like immunoglobins and albumin can be >12 orders of magnitude greater than that of low abundance cytokines, in addition to a complexity of many thousands of different protein products. This is likely an understatement of the actual complexity, as protein-protein interactions and post-translational modifications are not considered (Corthals *et al.*, 2000; Huber, 2003; Stasyk and Huber, 2004). Difficulties are also present even in bacteria and lower eukaryotes, as a large fraction of proteome ID's from these organisms are also highly abundant proteins (Corbin *et al.*, 2003; Opiteck *et al.*, 1997; Opiteck *et al.*, 1998). In many laboratories, the desire for data from a particular system or pathway is outweighed by a need to generate the largest number of identifications. Obtaining a high number of protein identifications does not necessarily advance the goal of the experiment, once that process is no longer a means on itself. The experimental goal is not the proteome, it is the underlying biological information that is sought.

One trend is to perform enriching separations for the identification of classes of proteins. Birch et al. (Birch *et al.*, 2003) recently demonstrated the use of reactive dye columns to enrich proteins in an effort to characterize perturbations on cell physiology in a more specific manner than shotgun proteomics.. This approach works for enriching proteins for some analysis, it does not specifically target proteins associated with a physiologic function. Fountoulakis and colleagues utilized chromatographic approaches to enrich for low abundance proteins via separation by hydroxyapatite LC (Fountoulakis and Takacs, 1998; Fountoulakis *et al.*, 1999a; Fountoulakis *et al.*, 1999b). In this case the chromatography resin was not selective enough as an even higher proportion of the enriched proteins were also highly abundant. These studies demonstrated that the diversity in chromatography chemistry is useful for generating subsets of the total protein content. Low-abundance proteins could be identified because of chromatographic enrichment, where typically maximal fractionation is desired. The rationale for our following experiments was to merge the separation and the selectiveness of chromatography with a ‘rational’ approach to enrich for specific subsets of the proteome.

Heparin chromatography has long been utilized as a means to purify basic proteins and in particular overexpressed DNA binding proteins. Here we tested whether its ability to act as a nucleic acid mimetic would enrich for nucleic acid binding proteins and allow identification of a significant number of these low abundance proteins. When applied to samples from two physiological states, the proteins we identified included a significant number of proteins known to bind to nucleic acids and anionic small molecules. These results suggest an effective means to rationally sub-fractionate more complex cellular mixtures and identify specific and relevant proteins in the process.

## Results

### One-dimensional Separation and Identification of Bacterial Proteins

Figure 4.1 shows the general strategy used in these studies. In general, Lysates from *E. coli* K-12 MG1655 (Blattner *et al.*, 1997) grown in LB or MOPS glucose and harvested during exponential or after entry into stationary phase were separated via heparin cation-exchange chromatography at pH 7.2. Heparin is a 'weak' cation-exchange resin, which should bind a limited subset of soluble *E. coli* proteins under these conditions. Individual fractions were collected and analyzed by SDS-PAGE, MALDI TOF-MS, and MALDI MS/MS to identify proteins.

Figure 4.2 shows 1D SDS-PAGE of the first 16 eluted heparin column fractions from lysates of the two physiological states. Unlike the flowthrough from anion exchange columns used previously (Chapter II) about 90% of the total cell protein as measured examining the gels relative to cell-lysates is in the flow-through fraction (Data not shown). Moreover, the pattern of proteins in the flow through seen by 1D SDS-PAGE is indistinguishable by eye from the proteins in the starting material, consistent with the column only removing a small subset of specific proteins from the extract. This

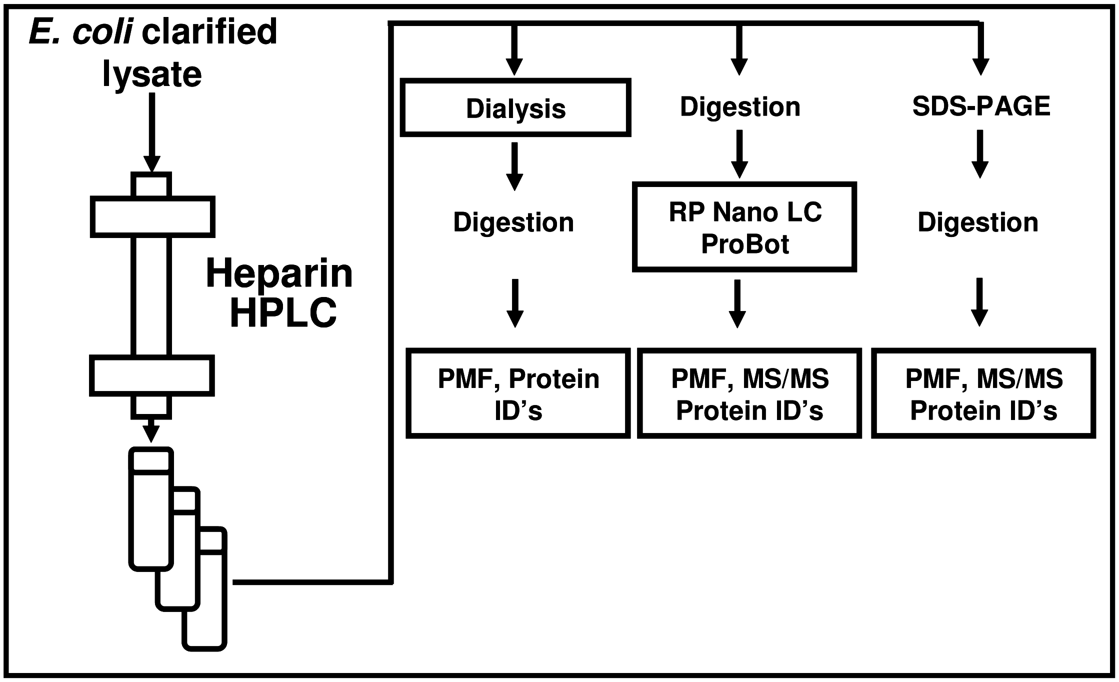


Figure 4.1.) Flowchart of heparome identification. Clarified cell lysates were loaded on heparin-HPLC columns. Heparin-binding proteins were collected and portions of fractions were separated by SDS-PAGE. Proteins were excised from gels and digested by trypsin and identified by MS and MS/MS analysis. Additional portions of the fraction were dialyzed, digested with trypsin, and separated by LC for identification by MS/MS (TOF-TOF).

is also possible with an overloaded column system, but the amount of material injected onto the column is not consistent with that observation. Overloaded columns also tend to lose resolving power, and are slow to return to baseline, none of which are observed on the gel or from the chromatogram (Data not shown). The large peak in fraction 6 as well as the identified proteins contain a significant portion of the ribosomal proteins identified in this study.

Direct analysis of heparin column fractions initially used methods similar to those previously described for analysis of 2-D native-state chromatography fractions (Chapters II & III). This method differes from these because it relies less on the ability of chromatography to reduce the complexity of the sample to a level that can be handled by MALDI and by PMF (Jensen *et al.*, 1997; Park and Russell, 2000, 2001). These experiments benefited from an additional RPLC separation of the peptides after digestion. This preliminary analysis was needed examine the complexity of a heparin separated proteome, as a typical anion exchange fraction is significantly too complex to analyze without additional fractionation (Data not shown). The massive amount of material in a typical separation creates ion-suppression effects that generally make all but the most intense peptides invisible. Figure 4.3 illustrates a typical MALDI spectrum from a fraction analyzed directly from the heparin column. The number of peaks observed in these spectra indicates significant spectral complexity, but it is not beyond the resolving and ionization capacity of the MALDI, even at high mass (Inset). The exponential phase fraction 7 (Fig. 4.3 Top) contained 197 annotated peptides, and 184 were seen in the stationary phase heparin fraction 7 (Fig. 4.3 Bottom).

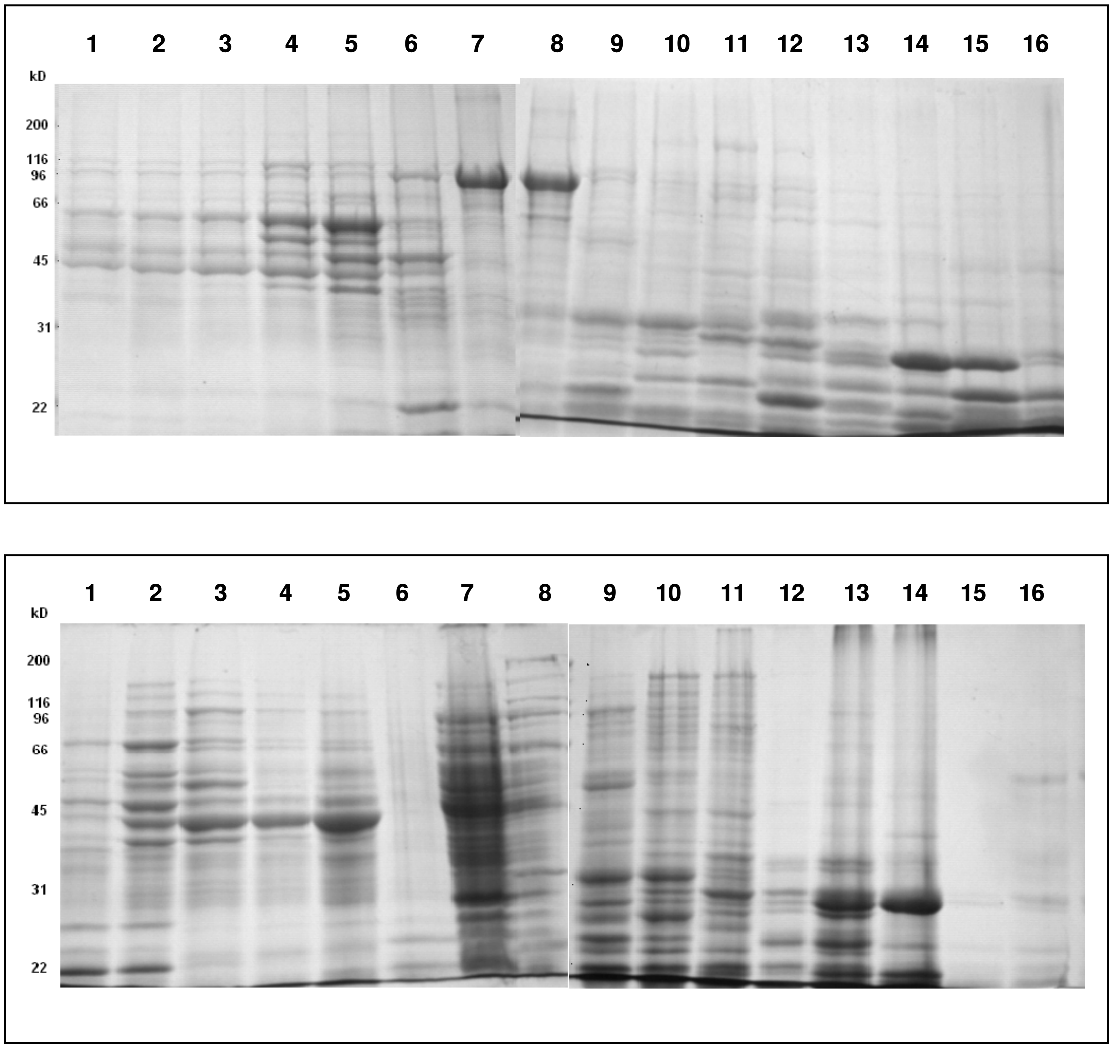


Figure 4.2.) SDS PAGE separation of heparin column fractions of exponential (top) and stationary (bottom) phase cells. The first 16 fractions from the heparin elution were collected, precipitated and separated over a 10% SDS-PAGE (Materials and Methods).

Having established that the heparin fractions could be analyzed directly, 459 bands were excised from gels similar to those shown in Figure 4.2. The gel slices were washed, digested in-gel with trypsin, and the resultant peptides were analyzed using MALDI-TOF MS and MS/MS analysis. For MS data, additional portions were analyzed without 1D PAGE separation as well.

The distribution of the identified proteins by fraction from the combined methods (SDS-PAGE) and (LC MALDI) is presented as Figure 4.4. Of note is the relatively even number of protein identifications across the fractions, illustrating that our LC conditions generate a reasonably flat elution profile. This is well-matched to our goals. Interestingly, a single fraction (6) contained approximately 3 fold as many protein identifications as the other fractions. This was observed between cell states as well and is not likely an artifact of a single heparin separation or empirically 'lucky' fraction. This fraction also contained a large number of the ribosomal proteins, which could explain part of the difference. Approximately half of the total ribomsomal proteins from *E. coli* were identified in this fraction in exponential phase and stationary phase. In total 1358 proteins were identified from heparin separated lysates of *E. coli* grown to either exponential or late stationary phase in minimal MOPS media. In total, 318 unique proteins were identified as heparin binding (Table 4.1A,) 240 unique proteins were identified from exponential cells, and 234 proteins from late stationary phase (Table 4.1B). There were 156 proteins identified in both cell states (Table 4.1C).

Table 4.1 lists the total proteins identified from both cell states. The list contains all proteins identified from both phases sorted by heparin fraction. Whether a protein

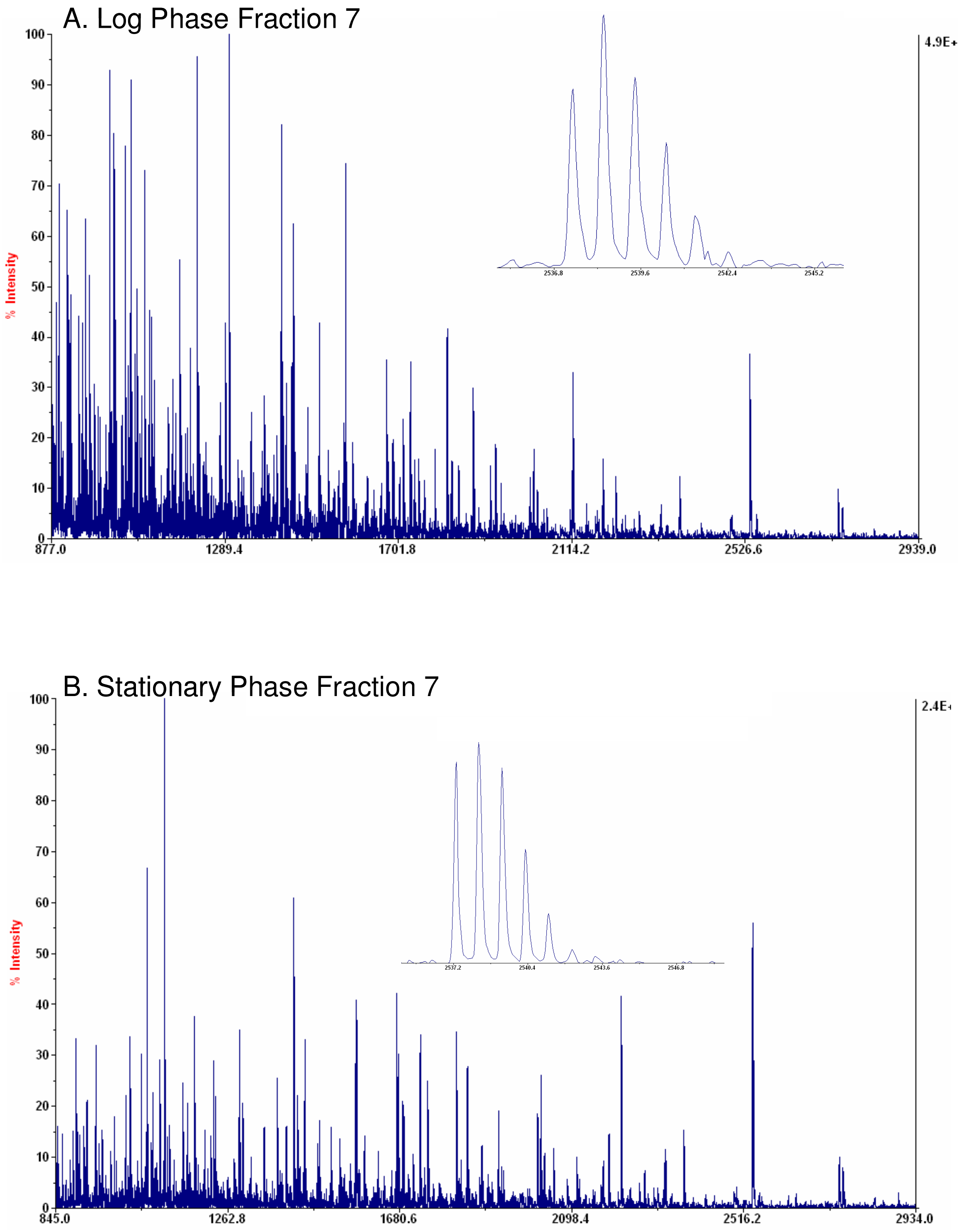


Figure 4.3.) MALDI spectra from heparin proteome. These are typical MALDI-DE RTOF spectra from the analysis of the heparin proteome. The top panel is fraction 7 from the exponential phase cells and the bottom panel is fraction 7 from stationary phase cells. The inset shows a blow up of a high mass tryptic peak to illustrate the large mass range in which high resolution data is obtainable, even with tryptic digestion and a high ion population.

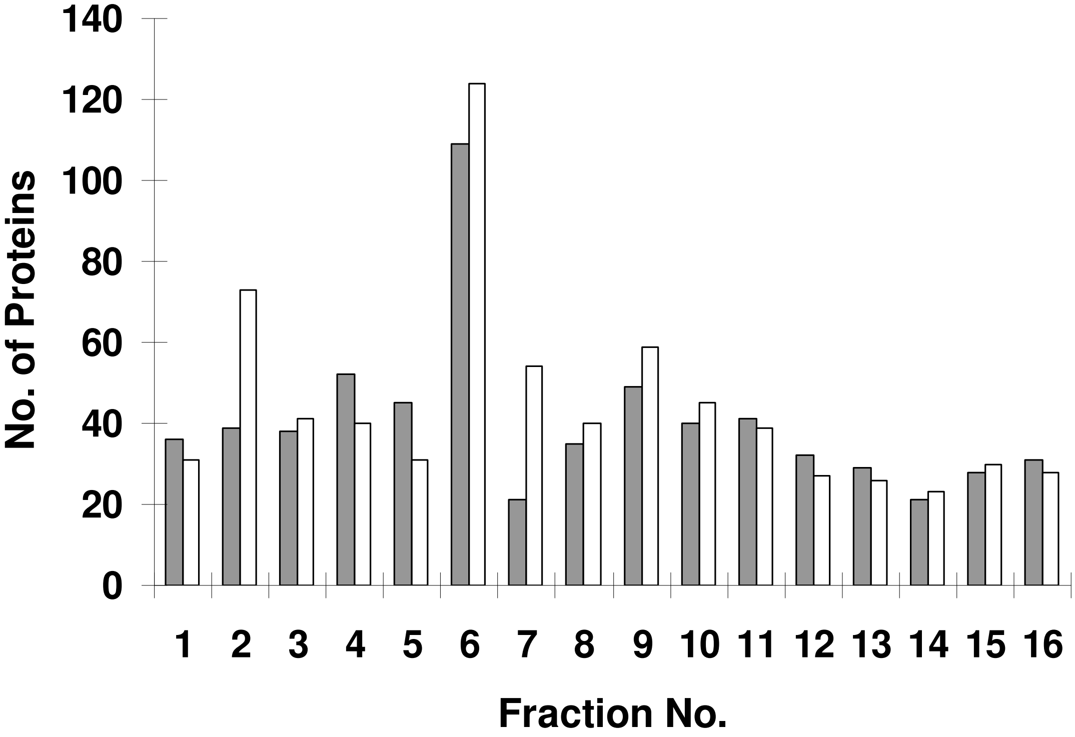


Figure 4.4.) Distribution of protein identifications. The number of unique proteins in every fraction was plotted for fractions from the exponential phase sample (gray) and the stationary phase sample (white). The overall pattern is similar between exponential and stationary phase, and in addition, no biases in peptide separation were observed. This refers to a bias from one cell phase containing a disproportionate number of proteins in a fraction.

was identified uniquely in a particular cell-state is indicated by the bold type. Several proteins known to be expressed in stationary phase were identified, the results of which are shown in Table 4.1. Dps, which is expressed in stationary phase cells, was only identified by MS/MS data from the stationary phase cells. Dps is also a DNA binding protein, so its presence on this LC substrate in a cell state specific manner is revealing and provides general validation of the method as a whole.

### Global Protein Information

Figure 4.5 highlights the overlap between the two heparin proteomes, the identifications of which are listed in Table 4.1B and C. Proteins were assigned to functional classification as compared to the *E. coli* genome. The results of this analysis are presented in Figure 4.5B. The distribution of the identifications is plotted against the genome for comparison. Although there are not many striking differences in the distribution of functional assignments, there is a large overrepresentation in proteins annotated as information transfer, which represents proteins involved in the central dogma, namely binding and processing the genome, tRNA, and mRNA molecules. There is also a slight increase in proteins in the regulation class of regulation relative to the genome, and a reasonably even representation in metabolism. This is significant because our previous studies had large biases towards proteins involved in metabolism (high abundance; Chapter II Figure (2.6). We interpret this as evidence for the selective enrichment of proteins in the heparin proteome.

Table 4.1.) Heparin binding proteins identified from exponential and stationary phase proteomes. (A) Total proteins identified from each cell sate by a combination of in-gel digestion/MS/MS and Nano-LC MALDI MS/MS analysis. Unique proteins are given as the number of specific gene products. Each protein is only counted once in the combined experiments, and total numbers. (B) List of proteins identified in exponential phase cells binding to the heparin separation. 240 unique proteins were identified from heparin separations of whole-cell lysates of *E. coli.* Gene name and SWISS Pro ID are given for each entry. Entries in bold (84) were found only in exponential phase cells. (C) List of proteins identified in stationary phase cells. In bold are the 78 proteins that were uniquely identified in stationary phase cells.

A.

| **Nr Protein ID's** | **In-Gel MS/MS** | **LC MALDI MS/MS** | **Combined** |
|---|---|---|---|
| **Exponential Phase** | 151 | 178 | 240 |
| **Stationary Phase** | 117 | 196 | 234 |
| **Combined** | 195 | 241 | 318 |

B. Table 4.1 Continued…

accA (P30867), accC (P24182), accD (P08193), aceE (P06958), adhE (P17547), **alaS (P00957)**, arcA (P03026), **argA (P08205)**, aroC (P12008), **aroH (P00887)**, atpA (P00822), atpD (P00824), **bglX (P33363)**, **ccmH (P33925)**, **creA (P08367)**, **crl (P24251)**, crp (P03020), **cysH (P17854)**, **dapE (P24176)**, **deaD (P23304)**, def (P27251), **dinG (P27296)**, dksA (P18274), dnaK (P04475), **era (P06616)**, erfK (P39176), fabA (P18391), fabB (P14926), **fabF (P39435)**, fabG (P25716), fabH (P24249), fabI (P29132), **fabR (P27307)**, **fadR (P09371)**, ftnA (P23887), **fur (P06975)**, **gapA (P06977)**, ghrA (P75913), **gidB (P17113)**, **glf (P37747)**, glgB (P07762), **glgC (P00584)**, **glmU (P17114)**, **glnE (P30870)**, glnS (P00962), gltB (P09831), gltX (P04805), **glyA (P00477)**, gnd (P00350), guaB (P06981), gyrA (P09097), **gyrB (P06982)**, hflX (P25519), **hisG (P10366)**, hisS (P04804), hns (P08936), **hscC (P77319)**, hupA (P02342), hupB (P02341), **hycE (P16431)**, ihfA (P06984), ihfB (P08756), infA (P02998), infB (P02995), infC (P02999), iscS (P39171), kdgR (P76268), **lexA (P03033)**, ligA (P15042), lon (P08177), lpd (P00391), **lpxB (P10441)**, lrp (P19494), malP (P00490), map (P07906), mdoG (P33136), **metE (P25665)**, metF (P00394), **metG (P00959)**, metJ (P08338), mfd (P30958), **miaA (P16384)**, **minE (P18198)**, **mnmG (P17112)**, moaB (P30746), **moaC (P30747)**, mprA (P24201), mraW (P18595), **mug (P43342)**, **nuoG (P33602)**, **nusB (P04381)**, **ompA (P02934)**, ompC (P06996), **ompF (P02931)**, ompR (P03025), **ompT (P09169)**, oppA (P23843), oxyR (P11721), **panC (P31663)**, parC (P20082), **parE (P20083)**, **pbpC (P76577)**, pcnB (P13685), pdxH (P28225), pepA (P11648), pgk (P11665), **pheS (P08312)**, **pheT (P07395)**, **pnp (P05055)**, **pntA (P07001)**, ppc (P00864), ppk (P28688), prc (P23865), **proC (P00373)**, purU (P37051), rbfA (P09170), **rcsB (P14374)**, relA (P11585), **rhlB (P24229)**, rhlE (P25888), rho (P03002), **rlpB (P10101)**, rluC (P23851), **rluE (P75966)**, **rne (P21513)**, rnr (P21499), rob (P27292), rplA (P02384), rplB (P02387), rplC (P02386), rplD (P02388), rplE (P02389), rplF (P02390), rplI (P02418), rplJ (P02408), rplK (P02409), rplL (P02392), rplM (P02410), rplN (P02411), rplO (P02413), rplP (P02414), rplQ (P02416), rplR (P02419), rplS (P02420), rplT (P02421), rplU (P02422), rplV (P02423), rplW (P02424), rplX (P02425), rplY (P02426), rpmA (P02427), rpmB (P02428), rpmC (P02429), rpoA (P00574), rpoB (P00575), rpoC (P00577), **rpoD (P00579)**, rpsA (P02349), rpsB (P02351), rpsC (P02352), rpsD (P02354), rpsE (P02356), rpsF (P02358), rpsG (P02359), rpsH (P02361), rpsI (P02363), rpsJ (P02364), rpsK (P02366), rpsM (P02369), rpsN (P02370), rpsO (P02371), rpsP (P02372), rpsQ (P02373), rpsR (P02374), rpsS (P02375), rpsT (P02378), rpsU (P02379), **rsd (P31690)**, rsmC (P39406), **rsuA (P33918),** secA (P10408), **selB (P14081)**, skp (P11457), slyA (P55740), **spoT (P17580)**, stpA (P30017), sucA (P07015), sucB (P07016), suhB (P22783), talB (P30148), **tgt (P19675)**, thrS (P00955), **tktA (P27302)**, **trpR (P03032)**, trpS (P00954), **ttk (P06969)**,

tufA/tufB (P02990), **tyrR (P07604)**, tyrS (P00951), **ugpQ (P10908)**, **uidR (Q59431)**, **ung (P12295)**, uvrB (P07025), **uvrY (P07027)**, xthA (P09030), yaaA (P11288), yaeH (P37048), **yafC (P30864)**, **yafJ (Q47147)**, **yafL (Q47151)**, ybaK (P37175), ybeZ (P77349), **ybgK (P75745)**, ybiB (P30177), **ycbB (P22525)**, ycgK (P76002), **yciK (P31808)**, ycjX (P76046), **ydcC (P28917)**, **ydfH (P77577)**, ydgH (P76177), ydjA (P24250), yegQ (P76403), **yejK (P33920)**, **yfaA (P17994)**, **yfhD (P30135)**, **yfiF (P33635)**, **yfjH (P52123)**, yggH (P32049), yggS (P52054), yhaJ (P42623), yhhX (P46853), **yibK (P33899)**, **yibQ (P37691)**, yihA (P24253), **yjeQ (P39286)**, yjgA (P26650), ykgM (P71302), ynhG (P76193), zwf (P22992),

C. Table 4.1 Continued…

accA (P30867), accC (P24182), accD (P08193), **aceA (P05313)**, aceE (P06958), **aceF (P06959)**, adhE (P17547), **ahpC (P26427)**, **aidB (P33224)**, **allR (P77734)**, arcA (P03026), aroC (P12008), atpA (P00822), atpD (P00824), **atpF (P00859)**, **bfr (P11056)**, **carB (P00968)**, **cbpA (P36659)**, **clpA (P15716)**, **cpsB (P24174)**, crp (P03020), **dacB (P24228)**, def (P27251), **dinB (Q47155)**, dksA (P18274), dnaK (P04475), **dnaX (P06710)**, **dps (P27430)**, **eda (P10177)**, **envC (P37690)**, erfK (P39176), fabA (P18391), fabB (P14926), fabG (P25716), fabH (P24249), fabI (P29132), **fmt (P23882)**, ftnA (P23887), **gadB (P28302)**, ghrA (P75913), glgB (P07762), **glnK (P38504)**, glnS (P00962), gltB (P09831), gltX (P04805), gnd (P00350), **gpmA (P31217)**, **groL (P06139)**, **grxB (P39811)**, guaB (P06981), **gutQ (P17115)**, gyrA (P09097), hflX (P25519), **hisB (P06987)**, hisS (P04804), hns (P08936), **htpG (P10413)**, hupA (P02342), hupB (P02341), **icd (P08200)**, ihfA (P06984), ihfB (P08756), **ilvE (P00510)**, infA (P02998), infB (P02995), infC (P02999), iscS (P39171), **ivy (P45502)**, kdgR (P76268), **lacZ (P00722)**, ligA (P15042), lon (P08177), lpd (P00391), **lpxD (P21645)**, lrp (P19494), malP (P00490), **malQ (P15977)**, **manX (P08186)**, map (P07906), **mdh (P06994)**, mdoG (P33136), metE (P25665), metJ (P08338), **metQ (P28635)**, mfd (P30958), **mglA (P23199)**, moaB (P30746), mprA (P24201), mraW (P18595), **mukB (P22523)**, **nadR (P27278)**, **nagA (P15300)**, **nikE (P33594)**, **nusG (P16921)**, ompC (P06996), ompR (P03025), oppA (P23843), oxyR (P11721), parC (P20082), pcnB (P13685), pdxH (P28225), pepA (P11648), pgk (P11665), **phoB (P08402)**, **polA (P00582)**, ppc (P00864), ppk (P28688), **ppx (P29014)**, prc (P23865), **purC (P21155)**, **purR (P15039)**, purU (P37051), rbfA (P09170), **rbsA (P04983)**, relA (P11585), **relE (P07008)**, rhlE (P25888), rho (P03002), **rlmB (P39290),** rluC (P23851), **rmf (P22986)**, **rnc (P05797)**, rnr (P21499), rob (P27292), rplA (P02384), rplB (P02387), rplC (P02386), rplD (P02388), rplE (P02389), rplF (P02390), rplI (P02418), rplJ (P02408), rplK (P02409), rplL (P02392), rplM (P02410), rplN (P02411), rplO (P02413), rplP (P02414), rplQ (P02416), rplR (P02419), rplS (P02420), rplT (P02421), rplU (P02422), rplV (P02423), rplW (P02424), rplX (P02425), rplY (P02426), rpmA (P02427), rpmB (P02428), rpmC (P02429), **rpmD (P02430)**, rpoA (P00574), rpoB (P00575), rpoC (P00577), rpsA (P02349), rpsB (P02351), rpsC (P02352), rpsD (P02354), rpsE (P02356), rpsF (P02358), rpsG (P02359), rpsH (P02361), rpsI (P02363), rpsJ (P02364), rpsK (P02366), rpsM (P02369), rpsN (P02370), rpsO (P02371), rpsP (P02372), rpsQ (P02373), rpsR (P02374), rpsS (P02375), rpsT (P02378), rpsU (P02379), rsmC (P39406), secA (P10408), **selA (P23328)**, **serA (P08328)**, skp (P11457), slyA (P55740), **sodB (P09157)**, **speG (P37354)**, **spy (P77754)**, **sthA (P27306)**, stpA (P30017), sucA (P07015), sucB (P07016), suhB (P22783), talB (P30148), thrS (P00955), **tig (P22257)**, **tpx (P37901)**, trpS (P00954), **truA (P07649)**, **truD (Q57261)**, **tsf (P02997)**, tufA/tufB (P02990), tyrS (P00951), uvrB (P07025), **wcaI (P32057)**, **wrbA (P30849)**, xthA (P09030), yaaA (P11288), **yaeB (P28634)**, yaeH (P37048), **yaiN (P55756)**, ybaK (P37175), **ybeX (P77392)**, ybeZ (P77349), ybiB (P30177), ycgK (P76002), **yciH (P08245)**, yciK (P31808), **ydcP (P76104)**, ydgH (P76177), ydjA (P24250), **yeaO (P76243)**, yegQ (P76403), **yffB**

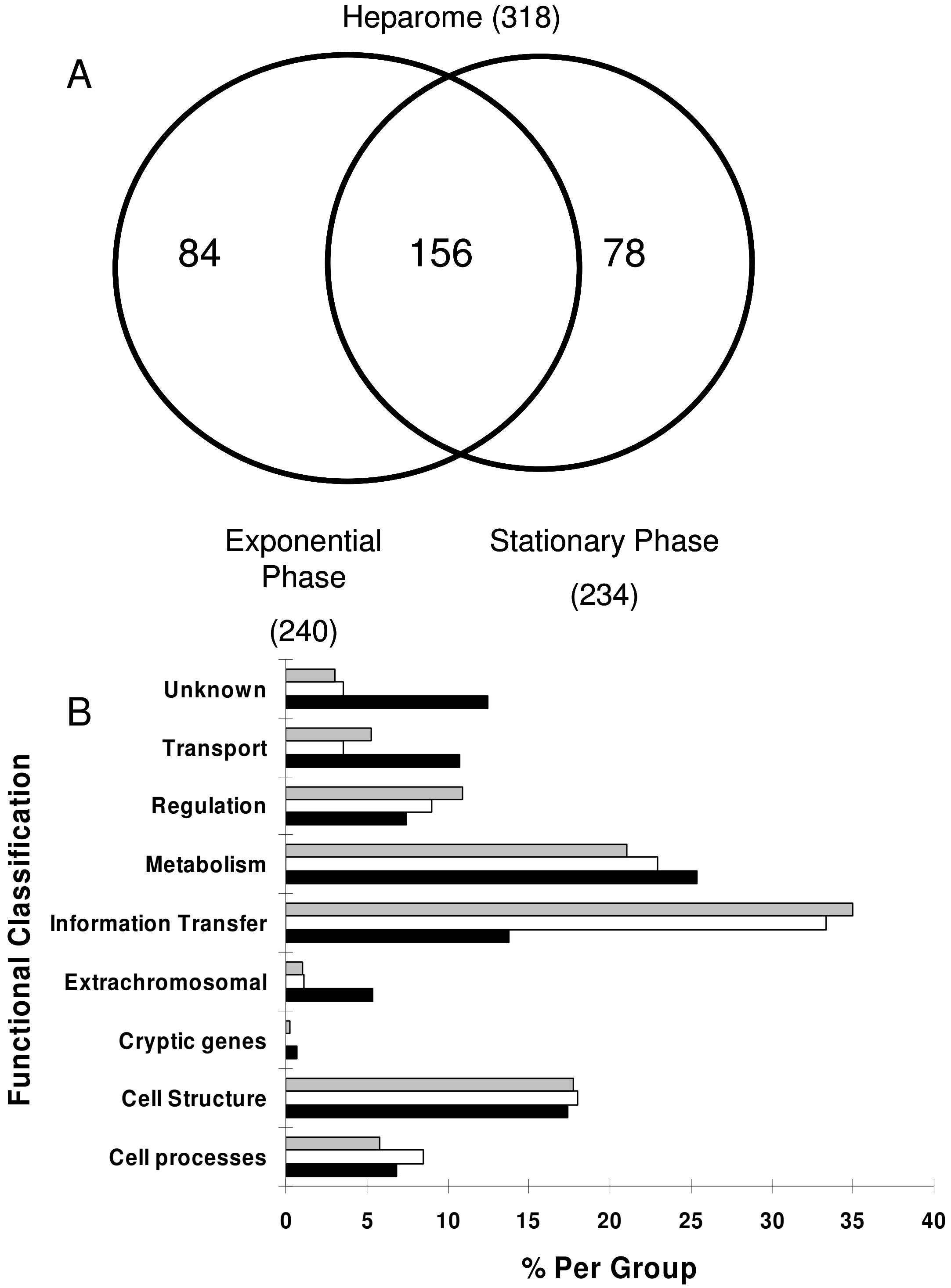


Figure 4.5.) Comparison of heparin-binding proteins from exponential and stationary phase cells. A) Venn diagram. 240 identifications were made in exponential phase cells, 234 in stationary phase cells and in total, 318 proteins are identified as heparin-binding proteins. B) Functional classification. Genome: Black bars, exponential heparome: Grey bars, stationary heparome: White bars. This shows the distribution of the gene assignments of the heparin identified proteins by cell-state as compared to the genome. Functional categories are from the Riley lab and http://eep.tamu.edu.

Comparing functional assignments for the heparin proteome relative to those obtained from other whole cell proteomes of *E. coli* is a different way in which to assess the qualitative differences in the types of proteins retained on heparin HPLC. The results of this are illustrated as Figure 4.6A (Corbin *et al.*, 2003; Sundararaj *et al.*, 2004; Tonella *et al.*, 2001). In white, is the portion of those proteomes that overlaps with the data identified by this heparin study alone. Since we do not expect to identify as many proteins in the heparin proteome, the relative contribution of the unique proteins identified from this study should be addressed. There are more heparin protein ID's in the categories of Information Transfer and Regulation. This is in contrast to Metabolism and Cell Processes, which can already be detected by traditional proteome methods. 51% (59 of the 116) unique proteins identified by heparin are found within the category of Information Transfer, and 47 of these are non-ribosomal proteins. The unique proteins found in this study are biased towards a functional category that is underrepresented in previous work.

Codon adaptation indices are useful as a means to predict potential protein abundances (Sharp and Li, 1987). In order to assess whether the heparin proteome increases the identification of lower-abundance proteins, we plotted these CAI values in Figure 4.6B as a distribution of the proteins identified from the heparin proteome compared to the distribution from other proteome projects. The heparome is enriched for proteins predicted to be in the lowest 5 classes of abundance. In particular, the number of unique identifications in the lowest three categories is substantial. More than 70% of the proteins in the genome are predicted to have CAI values between 0.2 to 0.4 Of the 116 proteins uniquely identified in the heparome data compared to other whole cell

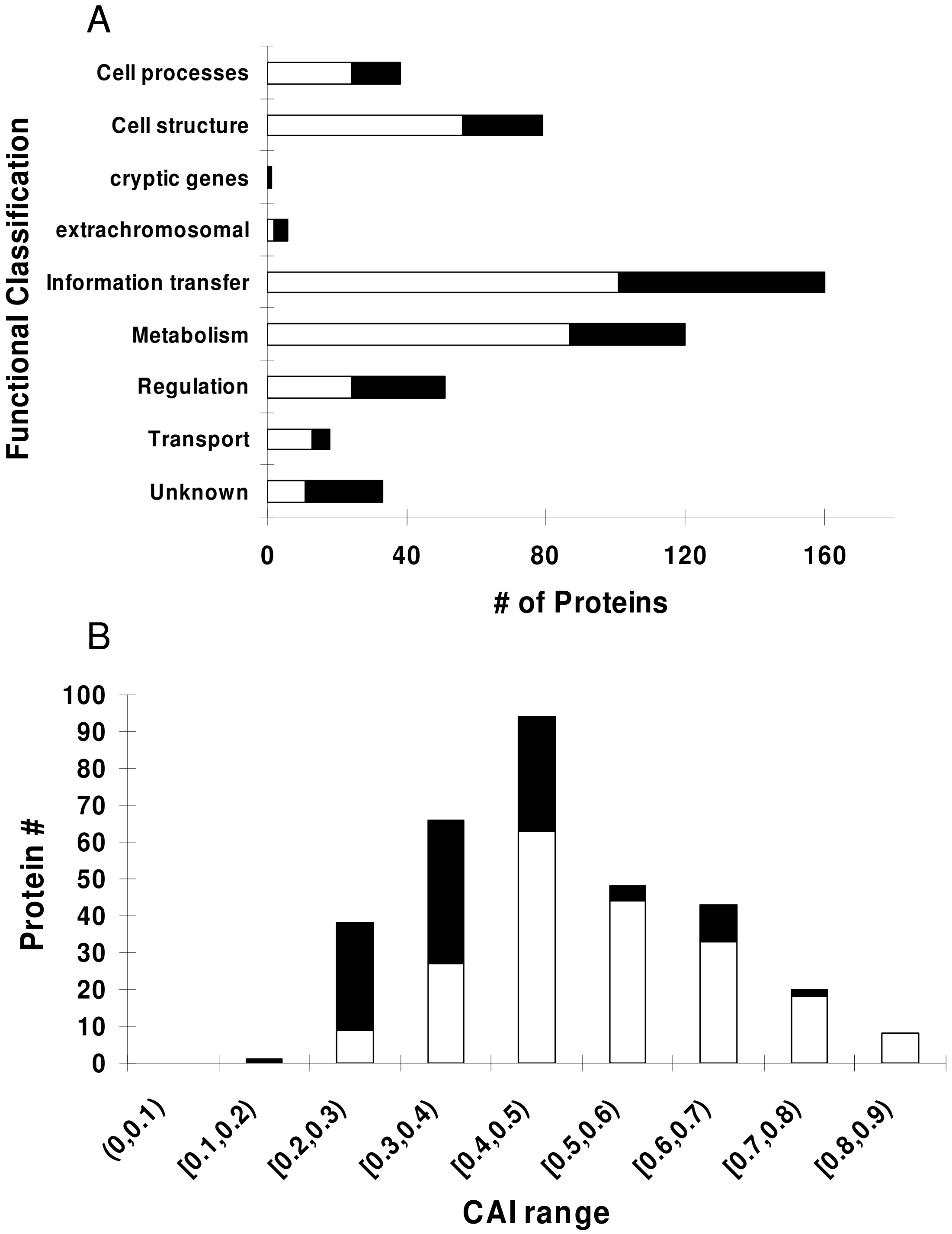


Figure 4.6.) Proteins identified as heparin-binding proteins. A) Functional classification of heparin-binding proteins (Riley M., 1998). Proteins found only by Heparome (black) are mainly in cell structure, information transfer, metabolism, regulation and unknown. B) Distribution of CAI (Eyre-Walker A., 1996) range of heparin-binding proteins. Proteins found only by Heparome (black) are mainly in CAI range from 0.2 to 0.5. In white are the fraction of the heparin identified proteins seen in other major *E. coli* proteomes. The entire bar represents the complete functional distribution of proteins identified in this work combined with several other whole cell proteomes.

proteomes, 100 of them have CAI values less than 0.5. This represents more than 86% of the total unique proteins identified by this method.

Abundance/Elution Changes

In order to identify proteins with potential changes in abundance between cell states, heparin-binding proteins were examined by 2D IEF PAGE (details in Materials & Methods). Proteins from at least 32 spots were observed to be up regulated in stationary phase and 19 were down regulated. These spots were excised, in-gel digested and identified by MS/MS analysis. Several spots contained no identifiable polypeptides, and several spots matched the same protein. Dps, for example was identified in 6 of the up-regulated spots from stationary phase. All of these spots had the same parent MW (isobaric), but had different pI's, indicating potential post-translational modification. Overall, 15 proteins were identified as being up regulated in stationary phase and 12 were identified as down regulated (Table 4.2).

Among the 156 proteins found in both stationary phase and exponential cells, 13 had distinct changes in elution position between the two growth states. These proteins are listed in Table 4.3. The elution position was the average elution fraction of all hits for that protein from the heparin dimension of separation. For example, peptide deformylase had an average elution position of fraction 5.6 in the exponential phase fractionation, which shifted to fraction 2 (lower NaCl) in the stationary phase samples. These proteins are likely to be interesting because shifting could be indicative of changes in protein-protein interactions, or post-translational modification. Many known complexes are retained on and coelute from the heparin column. PheS, and PheT, a heterotetrameric tRNA binding complex co-elute together on heparin as well as anion exchange

(Champion *et al.*, 2003). Other protein complexes that are stable through the heparin separation are sometimes identified together, and in some cases at both cell phases. Like our previous experiments, these identifications could be expanded on, and potentially utilized to identify new complexes, or changes to existing ones.

## Discussion

The identification of classes of proteins will play an increasing role in proteomics as specific questions are addressed by an increasingly biological user base in mass spectrometry. The proteins in this study compared heparin binding proteins from exponential and stationary phase from *E. coli*. These results can be thought of as two fold: The results show that this method is capable of identifying enough interesting protein candidates in each cell state to ascertain physiologic differences. Second, we were able to identify proteins of lower abundance overall. Acceptance of the ability to identify proteins by these means is widely discussed (Baldwin, 2004; Cargile *et al.*, 2004; Carr *et al.*, 2004; Eriksson *et al.*, 2000; Keller *et al.*, 2002; Wise *et al.*, 1997a; Wise *et al.*, 1997b), but generation of useful biological insight is limited.

The use of heparin as a first dimension of separation has been applied several times, notably to the proteome of *H. influenzae*, although its use as a means to enrich for a subset of the proteome has only recently been examined (Fountoulakis and Takacs, 1998; Langen *et al.*, 2000). Alternate separations appear to work in application toaddressing some of the questions raised here. Although the number of identified proteins from the heparome is lower than seen from multidimensional experiments, the degree to which the identified proteins are enriched for a particular type is significant.

Table 4.2.) Heparin binding proteins with changes in abundance observed in exponential and stationary phase cells.
Shown are the gene products identified by MALDI MS/MS analysis from spots on 2D PAGE that were observed to change intensity between stationary and log cells.

**Up-regulated in stationary phase**

| sp_id | gn | b_number | product |
|---|---|---|---|
| P12008 | aroC | b2329 | Chorismate synthase |
| b0812 | dps | P27430 | Stress response DNA-binding protein; starvation-induced resistance to H2O2; Fe-binding and storage protein; forms biocrystals with DNA |
| P52084 | elaB | b2266 | hypothetical protein |
| P45502 | ivy | b0220 | Inhibitor of C-lysozyme |
| P03033 | lexA | b4043 | Global regulator (repressor) for SOS regulon |
| P00391 | lpd | b0116 | Lipoamide dehydrogenase, NADH-dependent; E3 component of pyruvate and 2-oxoglutarate dehydrogenase complexes; also functions as glycine cleavage system L protein; binds Zn(II) |
| P08186 | manX | b1817 | Mannose phosphotransferase system, EIIAB component |
| P33136 | mdoG | b1048 | Periplasmic oligosaccharide synthesis |
| P08338 | metJ | b3938 | Methionine sulfoximine plus methylmethionine sensitivity; repressor |
| P18595 | mraW | b0082 | SAM-dependent protein methyltransferase, membrane-associated; cellular function unknown, expressed gene in dcw gene cluster; non-essential |
| P02934 | ompA | b0957 | Outer membrane protein A (II*); alkali-inducible |
| P02358 | rpsF | b4200 | 30S ribosomal subunit protein S6; suppressor of dnaG-Ts |
| P39406 | rsmC | b4371 | 16S rRNA m2G1207 methyltransferase, SAM-dependent |
| P76550 | yffS | b2450 | CPZ-55 prophage; putative transcriptional regulator |
| P46853 | yhhX | b3440 | Putative oxidoreductase, expressed protein; ydgJ paralog |

**Down-regulated in stationary phase**

| sp_id | gn | b_number | product |
|---|---|---|---|
| P17854 | cysH | b2762 | Phosphoadenylyl sulfate (PAPS) reductase |
| P24253 | engB | b3865 | GTPase essential for cell cycle |
| P75913 | ghrA | b1033 | Glyoxylate/hydroxypyruvate reductase; activity higher on glyoxylate than hydroxypyruvate |
| P00350 | gnd | b2029 | Gluconate-6-phosphate dehydrogenase, decarboxylating |
| P06987 | hisB | b2022 | Bifunctional enzyme imidazoleglycerolphosphate (IGP) dehydratase, histidinol phosphatase |
| P25665 | metE | b3829 | Methionine synthase, cobalamin-independent; 5-methyltetrahydropteroyltriglutamate-homocysteine methyltransferase; binds Zn(II) |
| P75787 | mntR | b0817 | conserved protein, Winged helix domain |
| P02388 | rplD | b3319 | 50S ribosomal subunit protein L4; erythromycin sensitivity |
| P02418 | rplI | b4203 | 50S ribosomal subunit protein L9 |
| P02351 | rpsB | b0169 | 30S ribosomal subunit protein S2; binds Zn(II) |
| P02352 | rpsC | b3314 | 30S ribosomal subunit protein S3 |
| P33918 | rsuA | b2183 | 16S RNA pseudouridine 516 synthase |

Table 4.3.) Proteins with elution shifts between exponential and stationary phase. Proteins are shown here that were found to have substantial (>2 fraction) shifts when comparing exponential average elution position and stationary phase average elution position.

| | | | | Fractions | |
|---|---|---|---|---|---|
| SP_ID | Gene Name | B # | Product | Exp. | Stat. |
| P27251 | def | b3287 | peptide deformylase | 5, 6 | 2 |
| P04475 | dnaK | b0014 | chaperone Hsp70; DNA biosynthesis; autoregulated heat shock proteins | 1 | 6, 7 |
| P23887 | ftnA | b1905 | cytoplasmic ferritin (an iron storage protein) | 1 | 6 |
| P33136 | mdoG | b1048 | periplasmic glucans biosynthesis protein | 8, 9 | 6 |
| P18595 | mraW | b0082 | putative apolipoprotein | 12 | 10 |
| P06996 | ompC | b2215 | outer membrane protein 1b (Ib;c) | 1, 2, 4 | 6 |
| P20082 | parC | b3019 | DNA topoisomerase IV subunit A | 11 | 9 |
| P00864 | ppc | b3956 | phosphoenolpyruvate carboxylase | 1 | 6 |
| P28688 | ppk | b2501 | polyphosphate kinase | 1 | 10 |
| P11585 | relA | b2784 | (p)ppGpp synthetase I (GTP pyrophosphokinase); regulation of RNA synthesis, Stringent Factor | 6 | 12, 14 |
| P23851 | rluC | b1086 | orf hypothetical protein | 15 | 13 |
| P37048 | yaeH | b0163 | putative structural protein | 10 | 6 |
| P42623 | yhaJ | b3105 | putative transcriptional regulator LYSR-type | 4 | 2 |

A recent study by Shefcheck et al., (Shefcheck *et al.*, 2003) explored the utility of applying heparin separations to cytosolic proteins from a cancer cell line. In their case, heparin separations were followed by 2D electrophoresis/MS. Their study demonstrated the utility of the chromatography as a means to enrich cationic proteins, but made little effort to identify the proteins thusly separated. In that case, 14 proteins were identified, and the extent to which these bound nucleic acid or the level of enrichment was not examined. One of the major additions of our work to this process were technical development in databases, separations, and the general idea of enhanced protein information from changes to chromatography. Biologically, understanding the differences observed in the cells requires extensive testing of the observations and validation of the changes in cellular association witnessed through protein identification.

## Validation of Confidence in Protein ID

In order to describe differences between cell states accurately, extensive validation and checking of the protein ID's is essential. Fortunately, there is some guidance in the field now. Recently, Carr et al. established a set of parameters for the display and dissemination of LC/MS/MS data, which is designed to exclude a significant fraction of false positive identification reported in the literature (Carr *et al.*, 2004). Reexamination of some manuscripts, for example, has identified false positive rates greater than 50%, and most of these can be attributed to single peptide hits reported. Identifications based on only a single peptide MS/MS match were eliminated from the list of nonredundant entries. And MS/MS confidence scores were kept at >0.005 (Materials and Methods). For purposes of elution positioning, lower scoring peptides would be considered an ID if adjacent fractions contained enough high stringency information to make an identification. In this manner, peak tails and elution position can be more accurately determined without falsely increasing the total number of unique protein identifications. Our two dimensional gel identifications were also utilized as controls for protein ID relative to specific heparin fractions. Another source of validatiaon was that proteins were often identified from both LC MALDI MS/MS and MS/MS on in-gel digests from the same fraction.

### Unique and Significant Protein Sets by Heparin Separation

The extent to which the heparin separation results in unique identifications was examined further. Table 4.1A-C also illustrates the proteins which were identified in both cell states. 45% (156/318) of the heparin identifications were seen in both log and stationary phase cells and each cell-state identified 84 and 78 unique proteins respectively (Figure 4.5A). It is also comparable to the amount of overlap observed in general between different proteomes, regardless of the fractionations utilized. This illustrates that the development of the method provides unique, descriptive protein information about different cell states.

In practice, this demonstrates an interesting observation: Parallel rounds of chromatography and/or utilizing cells grown under different conditions have a similar effect on the number of non redundant protein identifications as large-scale MDLC approaches (Washburn and Yates, 2000; Yates, 1998, 2004). Multiple injections of the same sample often incorporating inclusion/exclusion of precursor ions has been effective in increasing the identification of proteins from human samples (Resing and Ahn, 2005; Wysocki *et al.*, 2005). In order to be of physiologic use, a proteome must generate specific identifications relative to each cell state, in this case stationary phase growth. An approach is to identify as many polypeptides as possible, which paradoxically results in comparable identifications to this focused and more information rich approach.

These data also represent a contribution to the overall pool of confirmed protein identifications in *E. coli.* It does this by utilizing front end chromatography to select types of proteins not readily observed by other means. Performing the separation under two growth conditions further enhanced the number protein identifications. Figure 4.5B supports this assertation. In our original work one of the categories that was significantly over represented was metabolism. Here, the data in metabolism is comparable to the distribution in the genome, and information transfer (containing a higher proportion of nucleic acid binding proteins) is disproportionately higher. Information transfer only accounted for about 15% of the identifications in our original study (Chapter II Figure 2.6). The implication of this is that the heparin separation enriches relative to proteins observed in global studies and identifies a subset of proteins uniquely in the absence of any functional bias. It is clear from these data that a ‘proteome’ catalog of of every expressed ORF from *E. coli* wil require both differential separations and changes in cell growth.

The proportion of the unique heparin proteins falling within underrepresented identification categories is very high. It has been difficult in the past to identify many proteins in the lower predicted abundance categories, or functional categories dominated by regulatory proteins and low abundance polypeptides. In Figure 4.6B the enhancement of the lower abundance classes is striking compared to the previous proteomes.

Heparome, Parts of a Whole

The addition of MS/MS data enhanced our ability to identify proteins. It can be seen in the fact that the heparin proteome, which discards 90% of the starting material, identified essentially as many non-redundant proteins as the 2D LC separation performed earlier (Chapter II). In total, current proteomic studies have identified upwards of 1163 gene products in *E. coli* composing about 27% of the predicted genome (Champion *et al.*, 2003; Corbin *et al.*, 2003; Sundararaj *et al.*, 2004; Tonella *et al.*, 2001). These approaches included 2D gel electrophoresis, 2D MudPIT analysis, and 2D non denaturing LC separations, which comprise most of the available technology available to separate and identify proteins. The contribution of the heparome to the total identification of an *E. coli* proteome is presented as Figure 4.7A. In this case, heparomics adds less than 10% to the total of unique entries, but 37% of its identifications were unique. This is significant considering the relatively limited set of proteins retained on the column(s). Using this separation technique then, provides an orthogonal means by which we can expand our coverage of the model organism as well as isolate and characterize specific subsets of proteins. As a percentage of the available genome, these data compare well with those observed in yeast and higher eukaryotes (Bantscheff *et al.*, 2004). To best facilitate examination of differences between cell-states a selective method must also generate unique results and in this case, as shown in Table 4.1 (B,C) approximately 25% of the identifications made were unique to each cell state.

To look at the contribution of proteomes another way, Figure 4.7B graphs the codon adaptation index of all identified proteins from the proteome projects against the number of proteins present in each category from the genome. The most obvious lesson

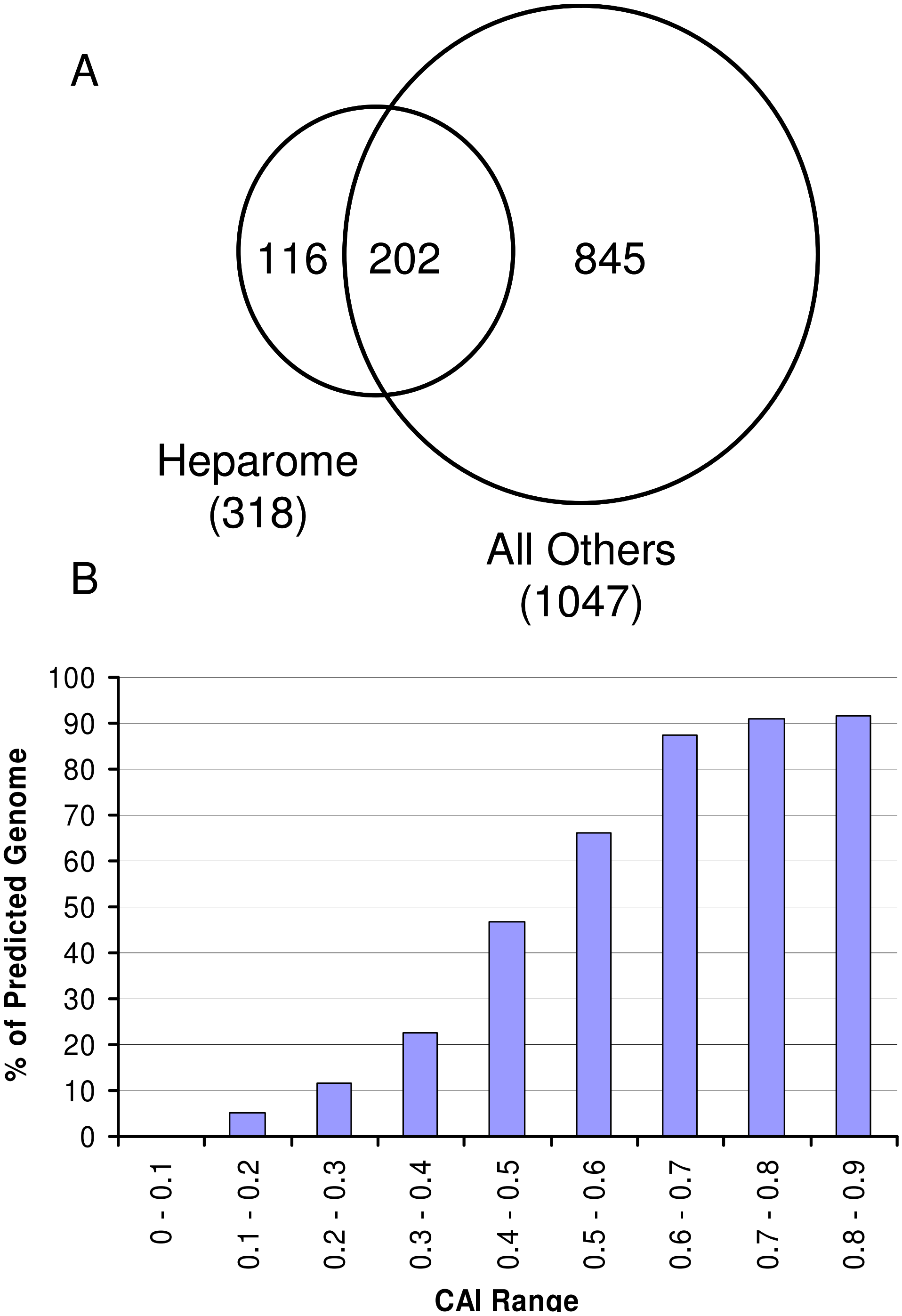


Figure 4.7.) Summary of other *E. coli* proteome efforts. A) Venn diagram of summary of Heparome and additional proteome efforts. 318 proteins were found in the heparome binding proteins and 1047 proteins were found by 5 other proteomic studies (Tonella L, 2001; Corbin RW., 2003; Cybercell project; (Chapter II); and in Chapter III). 116 proteins were only found in Heparome. In total 1163 proteins were identified for *E. coli* K12. B) Distribution of CAI range of proteins identified by current proteomic studies. 1163 proteins identified by current proteomic studies (Tonella L, 2001; Corbin RW., 2003; Cybercell project; Chapter II; and Chapter III; and these data). These data indicate that overall, coverage of predicted abundant proteins is excellent, and coverage falls off as CAI values are lower.

is that these projects do an extremely good job of identifying proteins that are predicted to be highly abundant. In fact, the 4 other proteome studies represented here identify upwards of 90% of the predicted proteins from the three highest CAI categories. Unfortunately, all proteomes suffer to some extent from abundance biases, and this figure illustrates that in order to identify more low abundance proteins, significant technological developments must still occur. Hopefully, the observations made from a comparison of exponential and stationary phase cells can be utilized to identify proteins for further characterization. A discussion of how observed differences in proteomes might be expanded on genetically is presented in Chapter V.

Another interesting aspect is the extent to which the large numbers of identifications of highly abundant bacterial proteins, like the elongation factors, heat shock proteins, or highly othr metabolic proteins do not seem to dominate the ID landscape as much. Proteins that were identified virtually everywhere in our previous studies, were found here in just a single cell state (Tig, PnP,) and the relative number of redundant identifications generated by these few proteins is less (data not shown).

### Orthogonal Separations as a Complement to Standard Approaches, Future Applications

There has been a long-standing interest in studying proteins that bind to nucleic acids, in particular mRNA and genomic DNA. Major cell processes in cells revolve around interactions with these two major classes of molecules, most notably in DNA-DNA replication, DNA-mRNA transcription, and mRNA-Protein via translation. The majority of the transcription factors and repressors in cells are typically DNA binding molecules present in extremely low abundance within cells, and thus are difficult to even detect even by enrichment based approaches.

The degree to which this study yielded significant identifications is slightly surprising, since a typical 1-D separation of cellular lysates is far too complex to resolve well on MALDI for PMF, and typically requires additional separation such as 2D gels or 2D chromatography to separate out the tryptic peaks and identify the proteins unambiguously. This is probably because of poor retention of heparin for most cellular proteins, one reason it is often chosen for *in vitro* purification of known DNA binding proteins. An examination of Figure 4.2 A&B also illustrates the comparative simplicity of proteins separated in this manner relative to a 2D gel of a whole cell lysates.

## Materials and Methods

### *E. coli* Lysates

*E. coli* strain MG1655 (Blattner *et al.*, 1997) was grown essentially as described in Chapter III (Champion et al., 2003). Two 1L cultures were inoculated with a 1:400 dilution from an overnight 5ml culture in minimal MOPS-glucose media. The MOPS media contained (0.4% glucose, 19mM $NH_4Cl$, 1.32mM $K_2HPO_4$, 2µg/ml thiamine, 10µg/ml uridine, 0.52 mM $MgCl_2$, 0.25µM $CaCl_2$, 8.37g/L MOPS, 0.72g/L tricine, 48mg/L $K_2SO_4$, 2.92g/L NaCl, 3mg/L Fe-$SO_4$-$7H_2O$ and additional micronutrients). Cells were grown with aeration in a non baffled Fernbach flask in a water bath shaker at 250RPM and 37$^o$C. One liter of the culture was harvested at mid-exponential phase ($OD_{600}$ = 0.5) and the other liter was harvested at late stationary phase ($OD_{600}$ = 2.0); approximately 17 hours after cells reached early exponential phase ($OD_{600}$ = 0.2). Cells were pelleted in a JA10 rotor (Beckman) at 4500 x g for 20 minutes and rinsed twice in chilled buffer containing 50mM $NaPO_4$, 50mM NaCl, pH 7.2 prior to lysis by French pressure cell. 5mM PMSF, 5mM DTT and 5mM *p*-aminobenzamidine were added to the lysis buffer, but not utilized in the washing buffer.

## Chromatography

Chromatography was performed on the system and hardware described elsewhere (Chapter II, Champion et al., 2003). Approximately 5 to 7ml of the lysates containing ≅50mg of protein was loaded onto a 5ml Hi-Trap Heparin column (Amersham Biosciences), washed with 8 column volumes of buffer (50mM $NaPO_4$, 50mM NaCl pH 7.2) and a linear gradient of 50 mM – 1 M NaCl over 10 column volumes was applied. 3ml fractions were collected and 20 fractions from the linear portion of the gradient were collected for analysis.

Nano LC for LC-MALDI analysis was carried out by loading 10 µl onto a Ultimate HPLC using a Famos autosampler (LCPackings). A $C_{18}$ Pepmap column (0.75mm x150mm) was used (LCPackings). The wash phase (A) was 2% ACN/0.1% TFA and organic phase (B) was 85% ACN/5% IPA/0.1% TFA. The gradient was 5% B to 90% B over a 60 minute period, followed by a 15 min wash. Eluted peptides were mixed with a 7.5 mg/ml α-cyano matrix solution in 60% MeCN. Spots were deposited every 6 seconds by a Probot (LC Packings) and spotted 24 x 26 on two stainless steel MALDI plates.

## One and 2D Gel Electrophoresis

For each of the first 16 elution fractions, 1 ml out of every 3 ml fraction (approx 80 µg per 1ml) was precipitated with 14% TCA and washed with ice cold acetone 2 times and dried. Proteins were separated at 20V/cm on 10% (37.5:1) format SDS PAGE (5x8cm) Laemmli gels (Klose, 1975; Laemmli, 1970; O'Farrell, 1975). Gels were stained with G250 Coomassie Brilliant Blue. Two dimensional PAGE was performed in the Protein Chemistry Lab at Texas A&M University essentially as described elsewhere (Champion *et al.*, 2003). For the first 16 elution fractions 400 µl of each 3ml fraction was precipitated by TCA and run on Igphor immobilized pH gradient gels (14cm pH 3-10NL) (Amersham Pharmacia) and focused for 60,000 V/h. After reduction and alkylation, SDS PAGE was perfomed with gels containing 12% polyacrylamide 37.5:1 bis-acrylamide.

## Dialysis and Digestion

In-gel digestions were carried out upon 459 protein bands excised from the SDS PAGE (Laemmli) using the Montage In-Gel Digest Kit (Millipore). Standard Montage digestion procedures were followed with the following additions. Four additional washing steps were added prior to gel slice drying, alternating treatments of 25 mM ammonium bicarbonate/ 5% MeCN and 25 mM ammonium bicarbonate/50% MeCN. The gel slices were also rehydrated in trypsin (0.02 µg/µl trypsin, 25 mM ammonium bicarbonate) and digested for 4 hours at 37C. Digestion wells were washed an additional 3 times with 100 µl wash solution (0.2% TFA) and step eluted with standard elution buffer followed by a 80% MeCN elution. Eluants were dried down on a speed vacuum and resuspended in matrix (see below). In solution digestion was performed by using 200 µl aliquots of each heparin fraction, which were then dialyzed three times for 4 hours each against 25 mM ammonium bicarbonate and subsequently dried on a speed vacuum at medium temperature. These were then reconstituted in 30 µl 50 mM ammonium bicarbonate, mixed with 2 µl of 20 µg/ml trypsin, and incubated overnight at 37° C. The samples were then frozen at -20° C prior to analysis.

MALDI-MS PMF and MS/MS

MALDI TOF/TOF spectra of in-gel digests were acquired by resuspension of the speed vacuum dried digest 3.0 µl CHCA (alpha cyano) at 10mg/ml in 75% MeCN, 0.05% TFA. Data were acquired in batch mode on the Applied Biosystems 4700. MS spectra were collected at 50 shots/spectra, 40 sub-spectra per spot for 2000 shots. Spectra were internally calibrated on angiotensin and fibrinogen peptide B (Sigma). The mass range was 800-5000 m/z with a focus mass of 1900 m/z. MS/MS spectra were collected at 50 shots/spectra, 60 sub-spectra per spot for a total of 3000 shots. The mass window for MS/MS spectra was selected as 50 relative resolution of the parent peak (FWMH). The metastable suppressor and the Colission Induced Dissociation were on with atmospheric gas present.

Additional portions of the heparin LC fractions were desalted by microdialysis, digested with trypsin as described above, and the resulting peptides were identified by MS and PMF. This MS was obtained on a Voyager DE-STR MALDI-TOF instrument (Applied Biosystems) using parameters identical to those described elsewhere (Park and Russell, 2000, 2001; Russell *et al.*, 2001) and Chapter II & III.

Analysis of spectral data was done using the Applied Biosystems GPS Explorer Software Version 2.0 with the Mascot search engine. The parameters were as follows: Taxonomy, *Escherichia coli*; Database, Swiss Prot; Enzyme, Trypsin; Max. Missed Cleavages, 1; Variable Modifications, Oxidation (M); Peptide Tolerance, 100 ppm; and

Significance Threshold of ≥90%. Peak filtering for MS was from 800-5000 Da with a 15 S/N (signal-to-noise) ratio. Peak filtering for MS/MS was from 60 Da to 20 Da below each precursor mass with a minimum S/N filter of 8 and a mass tolerance of 100 ppm. Positive identifications had a minimum protein confidence interval of 99.5% and systematic error.

### MW Validation from Gels

In order to validate the identifications obtained from MS and MS/MS with the in-gel digestions, we compared the observed molecular weights from the gel slices digested to the predicted molecular weight of the open reading frame. Among 833 redundant protein identifications made by in-gel digestion MS MS/MS analysis 77% of them were isolated within 30% of their predicted molecular weights (Data not shown). Although database searches can be constrained to specific MW or pI ranges, not all proteins resolve on SDS PAGE proportional to their molecular weights. Ribosomal proteins, in particular, are basic, and migrate differently than their predicted MW. It was therefore, more beneficial to search with loose tolerances on MW with respect to the observed migration position on the gel.

# CHAPTER V

# CONCLUSIONS

This work focused on the development of methods to identify the soluble protein content of E. coli. The data represent both contributions toward a complete bacterial proteome and novel insights into the scope of changes in protein content and interactions as cells encounter different environments. This chapter will focus on the history of how these approaches were developed and our current summary for the rearrangement in intracellular components as cells experience altered environments.

## State of Proteomics at Onset

This project began after a confluence of three events: The introduction of Dr. Russell's group to the Biochemistry Department, a seminar given by Dr. Don Hunt from the University of Virginia, and a realization that the techniques available and being described might be applicable to entire cell systems (Ficarro *et al.*, 2003). When this work began, identification of proteins on the basis of mass spectral information was non-trivial and labor intensive. There was a lack of adequate analytical software, and automation was essentially non-existent. The tools necessary to perform the peptide-mass fingerprinting experiments described here were becoming easier, but tandem MS/MS data for protein identification was relatively new and poorly validated for large-scale approaches.

Jensen et al. (Jensen *et al.*, 1997), established the major principle of recursive mass matching, whereby complex mixtures of peptides from several proteins could be deconvolved by selectively eliminating sequences from matching proteins and researching the remaining peak lists. This basic process is compatible with virtually any Peptide Mass Fingerprinting (PMF) approach and is utilized extensively in this work and others (Binz *et al.*, 1999; Kaji *et al.*, 2000; Langen *et al.*, 1997; Lee *et al.*, 2002a; Wasinger *et al.*, 1995; Westbrook *et al.*, 2001; Wilkins *et al.*, 1999). Reconstruction of mixtures containing at least ten separate protein digests was reported by Park et al. (Park and Russell, 2001), and Washburn et al., and Gygi et al., (Gygi *et al.*, 1999; Washburn and Yates, 2000), illustrated the feasibility of identifying proteins from whole-cell lysates using a combination of multidimensional chromatography and ion-trap tandem mass spectrometers (Washburn *et al.*, 2001). These studies and others illustrated the utility of biological mass spectrometry in protein identification, identification of proteins from mixtures, and large-scale approaches for the identification of many proteins from a cell.

## Know Your *E. coli*

*E. coli* is a Gram-negative bacterium. It is a member of a large family of bacteria commonly know as enterics. *E. coli* was a pioneering organism in the development of modern genetics, phage biology, motility, protein translocation, and a large portion of our understanding of the central dogma. The 4 compartments of the cytoplasm, periplasm, and the inner and outer cell membranes of *E. coli* contain many different types of molecules and we were interested in the approximately 2,100,000 copies of protein present within a typical cell. These two million proteins are encoded by approximately 4200 genes, of which 2000-3000 might be expressed and translated at a given point in time. For *E. coli,* a general description of its characteristics are given as Table 5.1. (Gorg *et al.*, 2004; Link *et al.*, 1997; Link *et al.*, 1999; Sundararaj *et al.*, 2004; Tani *et al.*, 2002). Most proteins are present as oligomers or within oligomeric complexes within the cell. This is significant, because it the majority of the proteins within a cell are present as a part of complexes and these complexes are large in size (Alberts, 1998; Sundararaj *et al.*, 2004). The presence of other molecules such as nucleotides and lipids would provide severe interference for MS analysis, relative to those conditions suited for peptide analysis, and the dynamic range and complexity of proteins in the organism are too large to analyze without fractionation. Initial experiments centered on improvements in fractionation of whole-cell lysates. *A. priori* this might seem a trivial problem. 4000 gene products separated into approximately 20 equal protein fractions would leave about 200 proteins per fraction. A second separation into 20 additional fractions would yield a net average of 10 proteins per fraction, which is generally within the limits of complexity of peptide analysis by MALDI-TOF-MS (Single MS). In fact, multi-dimensional separations of bacteria and small eukaryotes, such as yeast, routinely miss most of the expressed proteome regardless of the complexity of the front-end separation (Anderson and Anderson, 1998; Butt *et al.*, 2001; Corbin *et al.*, 2003; Corthals *et al.*, 2000; Figeys *et al.*, 1998; Fountoulakis and Takacs, 1998; Loo *et al.*, 2001; Lopez, 2000). This is known, because virtually all proteomes to date identify fewer numbers of proteins than we observe being expressed by other experiments.

The genome and proteome provide a relatively incomplete representation of the contents of a cell, although studies today provide more temporal insight into changes in protein and mRNA content at different states or time points (Corbin *et al.*, 2003; Hengge-Aronis, 2002b; Weber *et al.*, 2005). Unfortunately, information collected from gene-array experiments, or cell proteomes often cannot take into account changes in association or localization of proteins as a function of time, cell state, or environment, Table 5.1.) Table of micro and macromolecular contents of a typical *E. coli* cell. Reprinted from Project Cyber Cell, (http://redpoll.pharmacy.ualberta.ca/CCDB/) (Sundararaj *et al.*, 2004).

| General Statistics | | Large Molecule Statistics | |
|---|---|---|---|
| Cell length | 2 um or 2x10-6 m | Number of cell walls/cell | 1 |
| Cell diameter | 0.8 um or 0.8x10-6 m | Number of membranes/cell | 2 |
| Cell total volume | 1x10-15 L or 1x10-18 m3 | Number of chromosomes/cell | 2.3 (at mid log phase) |
| Cell aqueous volume | 7 x 10-16 L | Number of mRNA/cell | 4000 |
| Cell surface area | 6x10-12 m2 | Number of rRNA/cell | 18,000 |
| Cell wet weight | 1x10-15 kg or 1x10-12 g | Number of tRNA/cell | 200,000 |
| Cell dry weight | 3.0x10-16 kg or 3.0x10-13 g | Number of all RNA/cell | 222,000 |
| Periplasm volume | 6.5x10-17 L | Number of polysaccharides/cell | 39,000 |
| Cytoplasm volume | 6.7x10-16 L | Number of murein molecules/cell | 240,000-700,000 |
| Envelope volume | 1.6x10-16 L | Number of lipopolysaccharide/cell | 600,000 |
| Nuclear (DNA+protein) volume | 1.6x10-16 L | Number of lipids/cell | 25,000,000 |
| Inner Membrane thickness | 8x10-9 m | Number of all lipids/cell | 25,000,000 |
| Outer Membrane thickness | 8x10-9 - 15x10-9 m | Number of phosphatidylethanolamine | 18,500,000 |
| Periplasm thickness | 1x10-8 m | Number of phosphatidylglycerol | 5,000,000 |
| Average size of protein | 360 residues | Number of cardiolipin | 1,200,000 |
| Average diameter of ave. protein | 5 nm | Number of phosphatidylserine | 500,000 |
| Average MW of protein | 40 kD | Number of LPS (MW = 10kD) | 600,000 |
| Average prot. oligomerization state | 4 proteins/complex | Average SA of lipid molecule | 25 Ang2 |
| Average MW of protein entity | 160 kD | Fraction of lipid bilayer=lipid | 40% |
| Average size of mRNA | 1100 bases | Fraction of lipid bilayer=protein | 60% |
| Average length of mRNA | 370 nm | Number of outer membrane proteins | 300,000 |
| Average MW of all RNAs | 400 kD | Number of porins (subset of OM) | 60,000 |
| Average MW of single DNA | 3.0x109 D or 3.0x106 kD | Number of lipoproteins (OM) | 240,000 |
| Average MW of all DNA | 7 x 106 kD | Number of inner membrane proteins | 200,000 |
| Average length of DNA (chrom.) | 1.55 mm | Number of nuclear proteins | 100,000 |
| Diameter of chromosome | 490 um | Number of cytoplasmic proteins | 1,000,000 (excluding ribo proteins) |
| Diameter of condensed chromosome | 17 um | Number of ribosomal proteins | 900,000 |
| Spacing between small organics | 3.6 nm/molecule | Number of periplasmic proteins | 80,000 |
| Spacing between ions | 2.1 nm/molecule | Number of all proteins in cell | 2,600,000 |
| Ave. spacing between proteins | 7 nm/molecule | Number of external proteins (flag/pili) | 1,000,000 |
| Spacing between protein entities | 9 nm/molecule | Number of all proteins | 3,600,000 |
| Mean Velocity of 70 kD protein (cytoplasm) | 3 nm/ms = 3x10-6 m/s | | |
| Mean Velocity of 40 kD protein (cytoplasm) | 5 nm/ms = 5x10-6 m/s | | |
| Mean Velocity of 30 kD protein (cytoplasm) | 7 nm/ms = 7x10-6 m/s | | |
| Mean Velocity of 14 kD protein (cytoplasm) | 10 nm/ms = 10x10-6 m/s | | |
| Mean Velocity of small molecules (cytoplasm) | 50 nm/ms = 5x10-5 m/s | | |
| Mean Velocity of protein in H2O | 27 nm/ms = 2.7x10-5 m/s | | |
| Mean Velocity of small molecules in H2O | 87 nm/ms = 8.7x10-5 m/s | | |
| Concentration of protein in cell | 200-320 mg/mL (5-8 mM) | | |
| Concentration of RNA in cell | 75-120 mg/mL (0.5-0.8 mM) | | |
| Concentration of DNA in cell | 11-18 mg/mL (5 nM) | | |
| Volume occupied by water | 70% | | |
| Volume occupied by protein | 17% | | |
| Volume occupied by all RNA | 6% | | |
| Volume occupied by rRNA | 5% | | |
| Volume occupied by tRNA | 0.8% | | |
| Volume occupied by mRNA | 0.2% | | |
| Volume occupied by DNA | 1% | | |
| Volume occupied by ribosomes | 8% | | |
| Volume occupied by lipid | 3% | | |
| Volume occupied by LPS | 1% | | |
| Volume occupied by murein | 1% | | |
| Volume occupied by glycogen | 1% | | |
| Volume occupied by ions | 0.3% | | |
| Volume occupied by small organics | 1% | | |
| Translation rate | 40 aa/sec | | |
| RNA polymerase transcription rate | 70 nt/sec | | |

and in particular, post translational modifications. One of the obvious interpretations of the differences we observed in protein elution and <association> are changes to the polypeptide sequence. This method did not develop into a good discovery based-tool for identifying and cataloging interactions.

Several research groups have recently described a set of interacting protein partners from *E. coli* based upon methods similar to Gavin et al., and VonMering et al. (Butland *et al.*, 2005; Gavin *et al.*, 2002; von Mering *et al.*, 2002). In Butland et al., about 1,000 proteins were selectively tagged, pulled from cells and separated by 1D SDS PAGE. The resulting protein bands were identified by ESI mass spectrometry, and interaction maps were generated from the resulting protein ID's. Although the number of non-redundant interactions from this data set is relatively small, at 716 it represents a significant step towards understanding this portion of the proteome not visible by most experiments. Importantly 85% of these interactions were *novel* that suggests that these and other screens are far from saturation. Orthogonal approaches, and additional tagging coupled with IP's (immunoprecipitation) are likely to yield a significant amount of additional data. These approaches to collect data across specific cell states or environmental conditions are also labor intensive, which somewhat precludes their use in categorically differentiating cell-states or performing time-point physiologic studies. Since each tag is in a different strain of bacteria, elucidating changes due to physiology on many complexes at once would entail several thousand separate cultures, and interpretation would be difficult. Understanding more about the changes in protein content of a cell, in addition to potential changes in localization and oligomerization would provide an understanding of physiology unobtainable by traditional means.

## The Native State Approach

Initially, our experimental approach was designed to obtain broad separation of the proteins in order to identify as many protein components as possible. However, most separations designed to do this only fractionate proteins after digestion, which involves chromatography and MS resources that were unavailable to us at the time. We then applied intact separation procedures for whole-cell soluble lysates, and these fell into two major categories: Ion exchange, or size-exclusion chromatography. Ultimately, the primary separation utilized for most of the work presented was anion exchange, as it offered a few distinct advantages: First, modern ion-exchange resins have extremely high capacities (>50mg protein/ml resin) and can tolerate strong buffer and pH conditions necessary for effective cleaning, since we wanted to minimize sample handling prior to HPLC separation. Minimal sample handling, or loading samples with small numbers of discrete manipulation/loss steps, should decrease losses associated with multi-step methods. Second, anion-exchange is non-denaturing for many proteins near physiologic pH, and is in a range at which most of them bind and resolve. Third, one of the most common ‘contaminants’ in bacterial proteomics are the ribosomal proteins which are abundant, and virtually overwhelming relative to monitoring changes in protein content due to cell state. Anion-exchange offered a distinct advantage of having poor retention for a large portion of the ribosome and ribosomal proteins, which likely eluted in our HPLC flow-through. A second dimension of non-denaturing HPLC was needed, and hydrophobic interaction chromatography (HIC) was utilized for most of the separations performed in this work SOURCE 15Phe (Amersham). This was chosen because interactions on HIC columns are thought to be substantially different from those observed on ion-exchange columns.

### pH Profiling for Changes

One advantage of this separation procedure was the ability to alter the anion-exchange conditions of chromatography without redesigning the overall method. This was primarily accomplished by altering the pH at which the proteins were separated in the first dimension anion-exchange step. This accomplished two things: First, it allowed us to separate and identify proteins that were either not retained, or stable at one particular pH, without increasing the complexity of the separation/identification. Second, it enabled the generation of several hundred protein-protein interaction hypotheses, or potential interacting protein pairs.

The utilization of multiple pH's in the anion-exchange dimension increased the total number of protein ID's by approximately two-fold, and was responsible for the observation that proteins would shift not only at different pH, but between cell states as well. Growing cells under different conditions also increased identifications by approximately two-fold. This was a consequence of many factors including differential protein expression, changes in the cellular microenvironment, and different cell chemistry. One would likely expect different elution profiles of the cellular proteins, so in a way, the results were not surprising.

These data suggest that proteomes in general should explore more linear/parallel separations prior to using multi-dimensional separations in order to increase total protein identification. An example of this was the use of a different cation-exchange resin, heparin, which selected for significantly more basic proteins than the quaternary amine utilized in the SOURCE 15Q anion-exchange separations of the earlier work in Chapters II and III. Consequently, we were able to significantly add to the total numbers of proteins ID's in all *E. coli* proteomes, selecting for particular subsets of cellular proteins, and include the orthogonal data provided by repeating these results in cells grown under different conditions. All of these changes in chromatography increased the final samples arithmetically, unlike the geometric increases associated with additional separation dimensions.

## Changes in Cell Are Widespread

It is widely accepted that as cells change state, there are large morphologic changes that can be explained by changes in protein content. Phenotypic analysis is also largely described on the basis of changes in protein content (e.g. mutant analysis). Some proteins decrease in expression or presence, others increase etc., but this is not connected in parallel to overall changes in expression, modification, and oligomeric association.

Specific localization of many proteins is now understood as common in bacteria, including MinCD, FtsZ (ZipA) and many other proteins. Lai et al., sampled the protein content in mini-cells, which are polar derived, to assess differences in protein content relative to the general cell cytosol (Lai *et al.*, 2004). They observed several such proteins, YaiF and OmpW in particular, that had a strong preference for the mini cell. Recently, Roseman, et al., observed that portions of the pyruvate dehydrogenase complex localize to different locations in exponential *vs*. stationary cells (Patel *et al.*, 2004). Liu et al., probed changes in gene expression from *E. coli* utilizing different carbon sources, also observed changes in the distribution of RNAP between a high and low quality carbon source (Liu *et al.*, 2005).

As function following form, we can hypothesize that the observation that that protein localization is cell state dependent, and that a large change in protein expression occurs between cell states, that the massive changes in chromatography between cell-states we observed immediately suggests changes in protein-protein interactions or changes in post-translational modification because of growth conditions. The major summary of these data was presented in chapters II through IV. The other observation from the separate pH fractionations of each proteome is the ability to compare changes in co-elution between cell-state as well as between pH. Because of this, we were able to infer concerted changes in protein association due to changes in elution position where the only dependent variable is culture age.

These studies were instrumental in defining the dynamic nature of the bacterial content during the cell cycle, growth under different conditions, or the observation that protein localization plays specific roles in bacterial physiology. One such conclusion from the data presented here is how common these behaviors are likely to be. It appears that the overall rearrangements of the macromolecular contents of the bacterium are significantly more extensive than previously expected. In fact, it would now appear somewhat surprising to find examples of polypeptide chains which are true monomers, or do not appear to alter in a significant fashion due to changes in association or gene expression as a result of changes in the local environment.

Within the limits of this work, determination of protein complexes was not highly effective. The major obstacle to this was a lack of independent validation of complexes for which previous evidence was not available. In general, the methods seemed robust to the preservation of many complexes, but an examination of the data illustrates a false-positive rate that is likely high. Potentially a different dimension of separation specific for a set of complexes or a more traditional approach to complex purification might have yielded a better set of testable interactions, but from the work of Butland et al., it seems likely that parallel analysis of traditional complex purification is more effective (Butland *et al.*, 2005).

Significant morphologic changes within cells during starvation are difficult to rationalize with changes in protein levels alone. In, 1986 Groat et al., proposed that changes beyond simple protein expression are occurring when bacteria enter stationary phase (Groat *et al.*, 1986). It is continually surprising how complicated bacteria make regulation for scientists determined to generate simple models to explain their behavior.

## Future Work

Chapter IV gives some insight into the directions the proteome studies are likely to take next. Improvements in our ability to identify proteins, with the goal of increasing the ability to resolve differences between cell-states, are the primary motivators for innovation in this area. This is likely to be accomplished in two areas; changes in separation, and changes in mass spectrometric acquisition.

## Separation

One of the key developments made in this research was the large amount of insight gained by performing most of the separation under non-denaturing conditions. Insights from these data are still being interpreted now, and a more complete picture of the cellular polypeptide behavior will be difficult to obtain. In the absence of performing separations in this manner, however, most of this information would have been lost, or with the goal of simply identifying additional proteins, would have been ignored. In fact, several studies have performed primary separations under non-denaturing conditions, but these were typically designed as orthogonal forms of chromatography such as SEC or designed to capture specific types of biomolecules, such as dye-reactive columns, hydroxyapatite, or heparin (Fountoulakis and Takacs, 1998; Fountoulakis *et al.*, 1999a; Langen *et al.*, 2000; Shefcheck *et al.*, 2003). These efforts were centered on the types of proteins identified, or to gain larger numbers of ID'd proteins and efforts were not made to dissect the data for patterns of elution, co-elution or changes in chromatographic behavior.

Changes in chromatography do not need be multi-dimensional in order to improve the type and amount of data collected. In fact, as we illustrated in Chapter IV, the use of the glycosaminoglycan heparin facilitated two things: First, the different ion-exchange separation enabled the identification of many different proteins not observed in the anion-exchange driven data including 116 proteins not observed in the major published *E. coli* proteomes. Second, the polyanionic properties of heparin tend to enrich for nucleotide binding proteins, which includes regulatory proteins and replication-transcription-translation factors. These are typically not observed in traditional proteomes likely due to their low abundance (Corbin *et al.*, 2003; Tonella *et al.*, 2001).

One direction that was not adequately explored was the use of single-dimension separations, coupled to LC/MS/MS (or LC-MALDI) analysis. The availability of these instruments and nano-flow HPLC systems makes higher-throughput analysis of whole cell lysates more practical. Primary separations on various FPLC resins, such as NADH, ADP agarose etc., coupled to high resolving reversed-phase separations of peptides would enable significant numbers of unique proteins to be identified, and would preserve biological information about the types and classes of proteins separated. These results would likely generate readily interpretable data, due to its relative simplicity and a combination of several such resins might generate more protein identifications than the larger-scale multi-dimensional separations currently utilized. As a biological tool for the bacterial community a complement of LC separation conditions designed to catalog the complement of flavin, or ADP binding proteins etc., would have substantial utility.

### Mass Spectrometry

The major advances in this work in mass spectrometry were facilitated by the acquisition of new instrumentation. Originally, we utilized a Voyager Elite XL-TOF MALDI instrument (Applied Biosystems) to collect and analyze all of our data. Functionally, this instrument had extremely high resolution >15,000 FWMH, but the laser pulse frequency was slow (1Hz eventually upgraded to 3Hz), which decreased our ability to process thousands of samples. The next instrument we used was essentially an improved version of the Elite XL TOF, a Voyager DE-STR (Applied Biosystems). In addition to greater sensitivity, the laser pulse frequency was 20Hz, which significantly increased our ability to acquire data.

Ultimately, the trend in the field is to add in tandem MS data capability. This is accomplished using ESI tandem instruments, like Q-TOF style instrumentation, or ion-traps, or in this case, the acquisition of a tandem TOF instrument, a 4700 TOF-TOF. The increase in data quality on a TOF-TOF instrument increases the confidence of the proteins we identify, and allows identification utilizing fewer numbers of peptides, since the peptide fragmentation pattern provides additional data not present in precursor masses and fingerprints.

Although instrument developments will continue to improve the quality and sensitivity of the collected data, these are likely to diminish in importance relative to data analysis and database systems capable of organizing the results. Bacterial proteomics is not generally a material limited field and gains in sensitivity are more important as a means to increase the dynamic range of the instrument, not simply the detection of the lowest abundance precursor molecules. Full utilization of tandem acquisition methods and advances in automated data processing will have a greater impact on the data than advances in instrumentation. However, instrument advances are typically coupled to advances in storage and processing, so it is often difficult to separate the two.

The efforts of my lab to develop and maintain on-line databases of the proteome results are on-going. The question of to to efficiently archive and access the data is being addressed as additional projects in the laboratory. At the time of this writing, it can be found at http://eep.tamu.edu (**EEP** = **E**xperiments in ***E**. coli* **P**roteomics) which contains an interface for searching and visualizing the data from these experiments. The underlying data structure also makes possible queries that greatly facilitate investigating empirical questions based on the results. Initially, all of these *post-hoc* experiments were performed on spreadsheet programs or paper, which was laborious and difficult to maintain. As an example the observation of elution changes being present in the chromatography based on cell state was only made after extensive manual data mining, something that would have been easier with a database format comparable to **EEP**.

## Are the Changes Real?

Additional biochemical evidence for the observations we made still needs to be collected or obtained.. A focused effort on several of the putative targets from the initial studies is underway, but at the onset, it was not within the scope of the original project. The most direct evidence that these changes are real is that are observed as shifting spots on 2D gels of fractionated whole-cell lysates from exponential and stationary phase cells. This does not eliminate the separation itself as the source of the difference, however. Additional evidence for the changes being real is two-fold: First, the changes are consistent and reproducible. 2D gels of the ion-exchange fractions collected and/or run years apart on different columns, with different cell-preparations and by different students show extremely consistent trends in spot position and cell-state shifts. The fact that shifting patterns are present when different types of chromatography are utilized indicates it is not an artifact of the particular anion-exchange setup.

The method we developed was effective at identifying changes and additional validation on specific complexes/ID's are underway. Specific identification of a migrating protein having a known or determined role in stationary phase survival is indirect evidence that the shift is related to the physiology. Using MS and MS/MS analysis we are investigating additional shifting proteins for potential post-translational modifications.

Thus, if the changes I observed reflect changes in the cell environment, three questions are raised: First, how do the proteins change in multimeric partners? Second, how do the proteins change in localization? Third, is how are the proteins modified during entry into stationary-phase? Global approaches to identifying the interactome are best answered with traditional immuno- (tag) precipitations coupled to MS/MS analysis, as has been done in yeast and more recently, *E. coli* (Butland *et al.*, 2005; Gavin *et al.*, 2002).

The second question could be assessed using confocal fluorescence microscopy, where several putative partners are labeled and examined under different growth conditions (Liu *et al.*, 2005; Weber *et al.*, 2005). The third question is interesting because post-translational modification in bacteria is not well studied and is thought to play a much smaller role in physiology than transcriptional regulation Overall, this represents a confluence of two fields, classical biochemistry coupled to traditional physiology where the former is interested in separation and purification to the exclusion of examining everything else, and the latter is concerned with the examination of everything with the exclusion of separation.

It is interesting to consider the broader physiology when examining these results. The anthropomorphic question arises, “Why might *E. coli* do this?” A likely answer is that regulatory networks in bacterium are more complex than expected. One of the major observations from the genome projects of higher organisms was the strikingly minimal increase in the number of predicted genes (Levine and Tjian, 2003). The increase in physiologic complexity therefore may arise from increasing the number of connections and signals between existing components, rather than creating components *de novo* to perform additional functions. Thus, in bacteria, broad changes in gene regulation are achieved via traditional models, but adaptation and fine-tuning are more efficiently handled by changes in interactions. Figure 5.1 summarizes a general model of the types of rearrangements occurring within cells. This highlights changes in expression

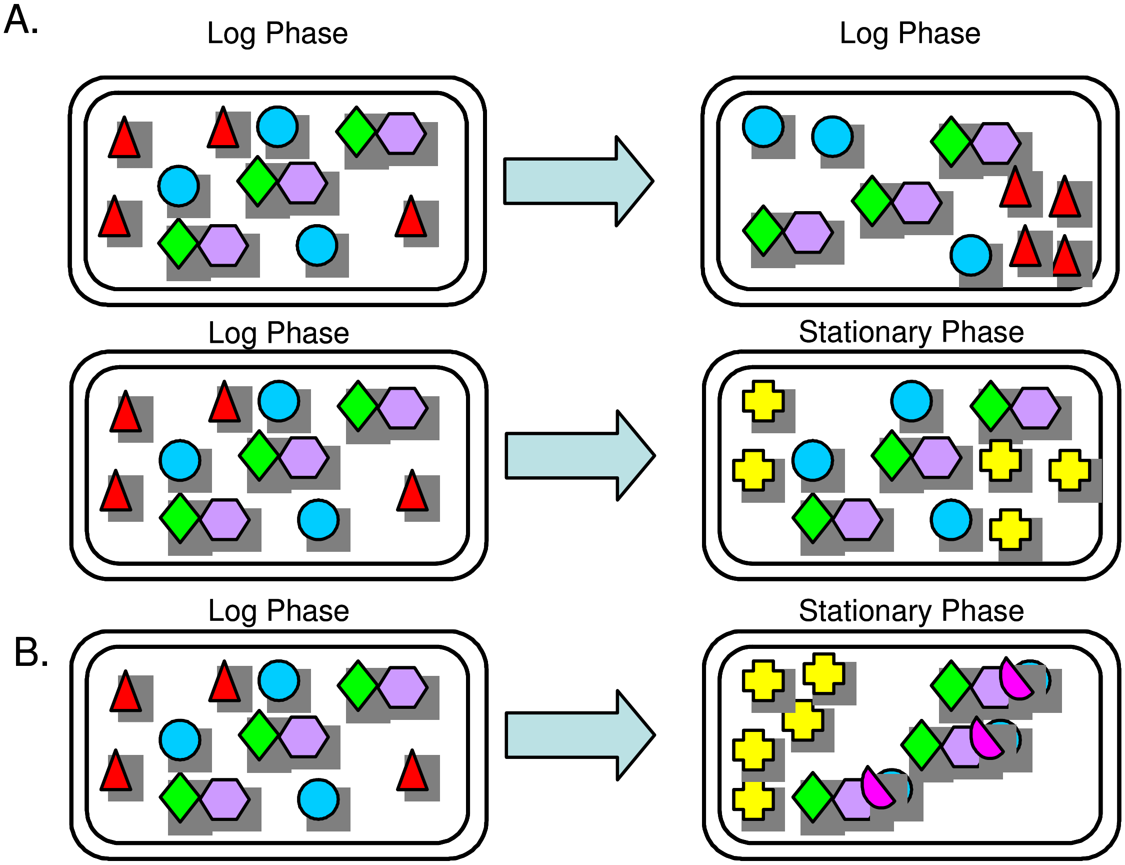


Figure 5.1.) Rearrangements in *E. coli* cells. This representation, based on other studies and observations here, illustrates how we perceive changes in cell-state relating to changes in protein content. A. represents changes in both expression and localization that occur during entry into stationary phase. B. represents the combination of these events, where a change in functional association is accompanied with an alteration in localization or modification.

and changes in localization. Overall, we think a much larger complement of these changes are occurring than is thought, and many of these events could be occurring together.

## Concluding Remarks

During these studies I was constantly challenged and surprised at the complexity of *E. coli* physiology under relatively mundane conditions. The work was extremely challenging and for a moderate period in the beginning, I was uncertain of its potential success. So much of the data collected was novel to my eyes; I found I was continually elated to see a spectrum or a band on a gel.

Something that should have been pursued in parallel to the MS studies were orthogonal experiments on the generated results. It was a loss to generate these putative results in arguably the best model-organism, then not take advantage of the tools and awesome power of genetics to further characterize these results. It is important not to understate the importance of the informatic tools in the analysis of these results, as I acknowledged, I am deeply grateful for the assistance of my advisor, Jim Hu and Lili Niu for maintaining and designing the database and query analysis, without which few of these results would be available from which to ask, "What if…?"

# REFERENCES

Alberts, B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**: 291-294.

Alberts, B.M., Amodio, F.J., Jenkins, M., Gutmann, E.D., and Ferris, F.L. (1968) Studies with DNA-cellulose chromatography. I. DNA-binding proteins from *Escherichia coli. Cold Spring Harb Symp Quant Biol* **33**: 289-305.

Almiron, M., Link, A.J., Furlong, D., and Kolter, R. (1992) A novel DNA-binding protein with regulatory and protective roles in starved *Escherichia coli*. *Genes Dev* **6**: 2646-2654.

Altuvia, S., Almiron, M., Huisman, G., Kolter, R., and Storz, G. (1994) The *dps* promoter is activated by OxyR during growth and by IHF and sigmaS in stationary phase. *Mol Microbiol* **13**: 265-272.

Anderson, N.L., and Anderson, N.G. (1998) Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis* **19**: 1853-1861.

Arnold, C.N., McElhanon, J., Lee, A., Leonhart, R., and Siegele, D.A. (2001) Global analysis of *Escherichia coli* gene expression during the acetate- induced acid tolerance response. *J Bacteriol* **183**: 2178-2186.

Baldwin, M.A. (2004) Protein identification by mass spectrometry: issues to be considered. *Mol Cell Proteomics* **3**: 1-9.

Baltimore, D. (2001) Our genome unveiled. *Nature* **409**: 814-816.

Bantscheff, M., Ringel, B., Madi, A., Schnabel, R., Glocker, M.O., and Thiesen, H.J. (2004) Differential proteome analysis and mass spectrometric characterization of germ line development-related proteins of *Caenorhabditis elegans*. *Proteomics* **4**: 2283-2295.

Beadle, G.W. (1945) Biochemical genetics. *Chemical Reviews* **37**: 15-96.

Becker, G., Klauck, E., and Hengge-Aronis, R. (2000) The response regulator RssB, a recognition factor for sigmaS proteolysis in *Escherichia coli*, can act like an anti-sigmaS factor. *Mol Microbiol* **35**: 657-666.

Binz, P.A., Muller, M., Walther, D., Bienvenut, W.V., Gras, R., Hoogland, C., Bouchet, G., Gasteiger, E., Fabbretti, R., Gay, S., Palagi, P., Wilkins, M.R., Rouge, V., Tonella, L., Paesano, S., Rossellat, G., Karmime, A., Bairoch, A., Sanchez, J.C., Appel, R.D., and Hochstrasser, D.F. (1999) A molecular scanner to automate proteomic research and to display proteome images. *Anal Chem* **71**: 4981-4988.

Birch, R.M., O'Byrne, C., Booth, I.R., and Cash, P. (2003) Enrichment of *Escherichia coli* proteins by column chromatography on reactive dye columns. *Proteomics* **3**: 764-776.

Bjellqvist, B., Ek, K., Righetti, P.G., Gianazza, E., Gorg, A., Westermeier, R., and Postel, W. (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J Biochem Biophys Methods* **6**: 317-339.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**: 1453-1462.

Brancia, F.L., Butt, A., Beynon, R.J., Hubbard, S.J., Gaskell, S.J., and Oliver, S.G. (2001) A combination of chemical derivatisation and improved bioinformatic tools optimises protein identification for proteomics. *Electrophoresis* **22**: 552-559.

Brock, M., Maerker, C., Schutz, A., Volker, U., and Buckel, W. (2002) Oxidation of propionate to pyruvate in *Escherichia coli*. Involvement of methylcitrate dehydratase and aconitase. *Eur J Biochem* **269**: 6184-6194.

Burlingame, A.L., Boyd, R.K., and Gaskell, S.J. (1998) Mass spectrometry. *Anal Chem* **70**: 647R-716R.

Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433**: 531-537.

Butt, A., Davison, M.D., Smith, G.J., Young, J.A., Gaskell, S.J., Oliver, S.G., and Beynon, R.J. (2001) Chromatographic separations as a prelude to two-dimensional electrophoresis in proteomics analysis. *Proteomics* **1**: 42-53.

Cargile, B.J., Bundy, J.L., and Stephenson, J.L., Jr. (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. *J Proteome Res* **3**: 1082-1085.

Carpousis, A.J. (2002) The *Escherichia coli* RNA degradosome: structure, function and relationship in other ribonucleolytic multienzyme complexes. *Biochem Soc Trans* **30**: 150-155.

Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics* **3**: 531-533.

Cavalcoli, J.D., VanBogelen, R.A., Andrews, P.C., and Moldover, B. (1997) Unique identification of proteins from small genome organisms: theoretical feasibility of high throughput proteome analysis. *Electrophoresis* **18**: 2703-2708.

Cayley, S., Lewis, B.A., Guttman, H.J., and Record, M.T., Jr. (1991) Characterization of the cytoplasm of *Escherichia coli* K-12 as a function of external osmolarity. Implications for protein-DNA interactions in vivo. *J Mol Biol* **222**: 281-300.

Champion, M.M., Campbell, C.S., Siegele, D.A., Russell, D.H., and Hu, J.C. (2003) Proteome analysis of *Escherichia coli* K-12 by two-dimensional native-state chromatography and MALDI-MS. *Mol Microbiol* **47**: 383-396.

Chatterji, D., and Ojha, A.K. (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Curr Opin Microbiol* **4**: 160-165.

Chong, B.E., Hamler, R.L., Lubman, D.M., Ethier, S.P., Rosenspire, A.J., and Miller, F.R. (2001) Differential screening and mass mapping of proteins from premalignant and cancer cell lines using nonporous reversed-phase HPLC coupled with mass spectrometric analysis. *Anal Chem* **73**: 1219-1227.

Clauser, K.R., Baker, P., and Burlingame, A.L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* **71**: 2871-2882.

Corbin, R.W., Paliy, O., Yang, F., Shabanowitz, J., Platt, M., Lyons, C.E., Jr., Root, K., McAuliffe, J., Jordan, M.I., Kustu, S., Soupene, E., and Hunt, D.F. (2003) Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc Natl Acad Sci U S A* **100**: 9232-9237.

Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F., and Sanchez, J.C. (2000) The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis* **21**: 1104-1115.

Cottrell, J.S. (1994) Protein identification by peptide mass fingerprinting. *Pept Res* **7**: 115-124.

Courcelle, J., Khodursky, A., Peter, B., Brown, P.O., and Hanawalt, P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics* **158**: 41-64.

De Biase, D., Tramonti, A., Bossa, F., and Visca, P. (1999) The response to stationary-phase stress conditions in *Escherichia coli*: role and regulation of the glutamic acid decarboxylase system. *Mol Microbiol* **32**: 1198-1211.

Dronamraju, K.R. (1991) Profiles in genetics: George Wells Beadle and the origins of the gene-enzyme concept. *J Hered* **82**: 443-446.

Ducret, A., Van Oostveen, I., Eng, J.K., Yates, J.R., 3rd, and Aebersold, R. (1998) High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. *Protein Sci* **7**: 706-719.

Edmondson, R.D., and Russell, D.H. (1996) Evaluation of Matrix-Assisted Laser Desorption Ionization-Time-of-Flight Mass Measurement accuracy by using delayed extraction. *J Am Soc Mass Spectrom* **7**: 995-1001.

Eriksson, J., Chait, B.T., and Fenyo, D. (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal Chem* **72**: 999-1005.

Ficarro, S., Chertihin, O., Westbrook, V.A., White, F., Jayes, F., Kalab, P., Marto, J.A., Shabanowitz, J., Herr, J.C., Hunt, D.F., and Visconti, P.E. (2003) Phosphoproteome analysis of capacitated human sperm. Evidence of tyrosine phosphorylation of a kinase-anchoring protein 3 and valosin-containing protein/p97 during capacitation. *J Biol Chem* **278**: 11579-11589.

Figeys, D., Gygi, S.P., Zhang, Y., Watts, J., Gu, M., and Aebersold, R. (1998) Electrophoresis combined with novel mass spectrometry techniques: powerful tools for the analysis of proteins and proteomes. *Electrophoresis* **19**: 1811-1818.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496-512.

Flynn, J.M., Neher, S.B., Kim, Y.I., Sauer, R.T., and Baker, T.A. (2003) Proteomic discovery of cellular substrates of the ClpXP protease reveals five classes of ClpX-recognition signals. *Mol Cell* **11**: 671-683.

Fountoulakis, M., and Takacs, B. (1998) Design of protein purification pathways: application to the proteome of *Haemophilus influenzae* using heparin chromatography. *Protein Expr Purif* **14**: 113-119.

Fountoulakis, M., Takacs, M.F., Berndt, P., Langen, H., and Takacs, B. (1999a) Enrichment of low abundance proteins of *Escherichia coli* by hydroxyapatite chromatography. *Electrophoresis* **20**: 2181-2195.

Fountoulakis, M., Takacs, M.F., and Takacs, B. (1999b) Enrichment of low-copy-number gene products by hydrophobic interaction chromatography. *J Chromatogr A* **833**: 157-168.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**: 3784-3788.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**: 141-147.

Gevaert, K., and Vandekerckhove, J. (2000) Protein identification methods in proteomics. *Electrophoresis* **21**: 1145-1154.

Gorg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., and Weiss, W. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21**: 1037-1053.

Gorg, A., Weiss, W., and Dunn, M.J. (2004) Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **4**: 3665-3685.

Gorg, A., Weiss, W., and Dunn, M.J. (2005) Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **5**: 826-827.

Gras, R., Muller, M., Gasteiger, E., Gay, S., Binz, P.A., Bienvenut, W., Hoogland, C., Sanchez, J.C., Bairoch, A., Hochstrasser, D.F., and Appel, R.D. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* **20**: 3535-3550.

Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* **17**: 100-107.

Groat, R.G., Schultz, J.E., Zychlinsky, E., Bockman, A., and Matin, A. (1986) Starvation proteins in *Escherichia coli*: kinetics of synthesis and role in starvation survival. *J Bacteriol* **168**: 486-493.

Grunenfelder, B., Rummel, G., Vohradsky, J., Roder, D., Langen, H., and Jenal, U. (2001) Proteomic analysis of the bacterial cell cycle. *Proc Natl Acad Sci U S A* **98**: 4681-4686.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**: 994-999.

Hengge-Aronis, R. (1996) Regulation of Gene Expression during Entry into Stationary Phase. In Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology*. Vol. 2. Neidhardt, F.C., Curtiss III, R., Ingraham, J., Lin, E., Low, K.B., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M. and Umbarger, H. (eds). Washington DC: ASM Press, pp. 1497-1508.

Hengge-Aronis, R. (2002a) Stationary phase gene regulation: what makes an *Escherichia coli* promoter sigmaS-selective? *Curr Opin Microbiol* **5**: 591-595.

Hengge-Aronis, R. (2002b) Recent insights into the general stress response regulatory network in *Escherichia coli*. *J Mol Microbiol Biotechnol* **4**: 341-346.

Hirsch, M., and Elliott, T. (2002) Role of ppGpp in *rpoS* stationary-phase regulation in *Escherichia coli*. *J Bacteriol* **184**: 5077-5087.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., and Tyers, M. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisia*e by mass spectrometry. *Nature* **415**: 180-183.

Hoogland, C., Sanchez, J.C., Tonella, L., Binz, P.A., Bairoch, A., Hochstrasser, D.F., and Appel, R.D. (2000) The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res* **28**: 286-288.

Huber, L.A. (2003) Is proteomics heading in the wrong direction? *Nat Rev Mol Cell Biol* **4**: 74-80.

Huisman, G., Siegele, D.A., Zambrano, M.M., and Kolter, R. (1996) Morphological and Physiological Changes during Stationary Phase. In Escherichia col*i and* Salmonella typhimurium: *Cellular and Molecular Biology*. Vol. 2. Neidhardt, F.C., Curtiss III, R., Ingraham, J., Lin, E., Low, K.B., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M. and Umbarger, H. (eds). Washington, DC: ASM Press, pp. 1672-1679.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U S A* **98**: 4569-4574.

Jenkins, D.E., Schultz, J.E., and Matin, A. (1988) Starvation-induced cross protection against heat or $H_2O_2$ challenge in *Escherichia coli*. *J Bacteriol* **170**: 3910-3914.

Jensen, O.N., Vorm, O., and Mann, M. (1996) Sequence patterns produced by incomplete enzymatic digestion or one- step Edman degradation of peptide mixtures as probes for protein database searches. *Electrophoresis* **17**: 938-944.

Jensen, O.N., Podtelejnikov, A.V., and Mann, M. (1997) Identification of the components of simple protein mixtures by high- accuracy peptide mass mapping and database searching. *Anal Chem* **69**: 4741-4750.

Jensen, O.N., Larsen, M.R., and Roepstorff, P. (1998) Mass spectrometric identification and microcharacterization of proteins from electrophoretic gels: strategies and applications. *Proteins* **Suppl**: 74-89.

Jishage, M., and Ishihama, A. (1995) Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of sigma70 and sigma38. *J Bacteriol* **177**: 6832-6835.

Jungblut, P.R., Schaible, U.E., Mollenkopf, H.J., Zimny-Arndt, U., Raupach, B., Mattow, J., Halada, P., Lamer, S., Hagens, K., and Kaufmann, S.H. (1999) Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Mol Microbiol* **33**: 1103-1117.

Kaji, H., Tsuji, T., Mawuenyega, K.G., Wakamiya, A., Taoka, M., and Isobe, T. (2000) Profiling of *Caenorhabditis elegans* proteins using two-dimensional gel electrophoresis and matrix assisted laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis* **21**: 1755-1765.

Karlin, S., and Mrazek, J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol* **182**: 5238-5250.

Karlin, S., Mrazek, J., Campbell, A., and Kaiser, D. (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J Bacteriol* **183**: 5025-5040.

Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S.M., and Pellegrini-Toole, A. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res* **28**: 56-59.

Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., and Gama-Castro, S. (2002) The EcoCyc database. *Nucleic Acids Res* **30**: 56-58.

Karty, J.A., Ireland, M.M., Brun, Y.V., and Reilly, J.P. (2002) Artifacts and unassigned masses encountered in peptide mass mapping. *J Chromatogr B Analyt Technol Biomed Life Sci* **782**: 363-383.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**: 5383-5392.

Klose, J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**: 231-243.

Klose, K.E., North, A.K., Stedman, K.M., and Kustu, S. (1994) The major dimerization determinants of the nitrogen regulatory protein NTRC from enteric bacteria lie in its carboxy-terminal domain. *J Mol Biol* **241**: 233-245.

Kusano, S., and Ishihama, A. (1997) Stimulatory effect of trehalose on formation and activity of *Escherichia coli* RNA polymerase E sigma38 holoenzyme. *J Bacteriol* **179**: 3649-3654.

Lacour, S., and Landini, P. (2004) SigmaS-dependent gene expression at the onset of stationary phase in *Escherichia coli*: function of sigmaS-dependent genes and identification of their promoter sequences. *J Bacteriol* **186**: 7186-7195.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**: 680–685.

Lai, E.M., Nair, U., Phadke, N.D., and Maddock, J.R. (2004) Proteomic screening and identification of differentially distributed membrane proteins in *Escherichia coli. Mol Microbiol* **52**: 1029-1044.

Lambert, L.A., Abshire, K., Blankenhorn, D., and Slonczewski, J.L. (1997) Proteins induced in *Escherichia coli* by benzoic acid. *J Bacteriol* **179**: 7595-7599.

Langen, H., Gray, C., Roder, D., Juranville, J.F., Takacs, B., and Fountoulakis, M. (1997) From genome to proteome: protein map of *Haemophilus influenzae*. *Electrophoresis* **18**: 1184-1192.

Langen, H., Takacs, B., Evers, S., Berndt, P., Lahm, H.W., Wipf, B., Gray, C., and Fountoulakis, M. (2000) Two-dimensional map of the proteome of *Haemophilus influenzae*. *Electrophoresis* **21**: 411-429.

Lease, R.A., Smith, D., McDonough, K., and Belfort, M. (2004) The small noncoding DsrA RNA is an acid resistance regulator in *Escherichia coli*. *J Bacteriol* **186**: 6179-6185.

Lee, K., Bae, D., and Lim, D. (2002a) Evaluation of parameters in peptide mass fingerprinting for protein identification by MALDI-TOF mass spectrometry. *Mol Cells* **13**: 175-184.

Lee, K., Bernstein, J.A., and Cohen, S.N. (2002b) RNase G complementation of rne null mutation identifies functional interrelationships with RNase E in *Escherichia coli. Mol Microbiol* **43**: 1445-1456.

Lee, P.S., and Lee, K.H. (2003) *Escherichia coli*--a model system that benefits from and contributes to the evolution of proteomics. *Biotechnol Bioeng* **84**: 801-814.

Lee, W.C., and Lee, K.H. (2004) Applications of affinity chromatography in proteomics. *Anal Biochem* **324**: 1-10.

Levine, M., and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature* **424**: 147-151.

Lin-Chao, S., and Cohen, S.N. (1991) The rate of processing and degradation of antisense RNAI regulates the replication of ColE1-type plasmids in vivo. *Cell* **65**: 1233-1242.

Link, A.J., Robison, K., and Church, G.M. (1997) Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **18**: 1259-1313.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., and Yates, J.R., 3rd (1999) Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17**: 676-682.

Liu, M., Durfee, T., Cabrera, J.E., Zhao, K., Jin, D.J., and Blattner, F.R. (2005) Global transcriptional programs reveal a carbon source foraging strategy by *Escherichia coli. J Biol Chem* **280**: 15921-15927.

Loewen, P.C., Hu, B., Strutinsky, J., and Sparling, R. (1998) Regulation in the *rpoS* regulon of *Escherichia coli. Can J Microbiol* **44**: 707-717.

Lomovskaya, O.L., Kidwell, J.P., and Matin, A. (1994) Characterization of the sigma38-dependent expression of a core *Escherichia coli* starvation gene, *pexB. J Bacteriol* **176**: 3928-3935.

Loo, R.R., Cavalcoli, J.D., VanBogelen, R.A., Mitchell, C., Loo, J.A., Moldover, B., and Andrews, P.C. (2001) Virtual 2-D gel electrophoresis: visualization and analysis of the *E. coli* proteome by mass spectrometry. *Anal Chem* **73**: 4063-4070.

Lopez, M.F. (2000) Better approaches to finding the needle in a haystack: optimizing proteome analysis through automation. *Electrophoresis* **21**: 1082-1093.

Makinoshima, H., Aizawa, S., Hayashi, H., Miki, T., Nishimura, A., and Ishihama, A. (2003) Growth phase-coupled alterations in cell structure and function of *Escherichia coli*. *J Bacteriol* **185**: 1338-1345.

Mann, M., Hojrup, P., and Roepstorff, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* **22**: 338-345.

Marino-Ramirez, L., Minor, J.L., Reading, N., and Hu, J.C. (2004) Identification and mapping of self-assembling protein domains encoded by the *Escherichia coli* K-12 genome by use of lambda repressor fusions. *J Bacteriol* **186**: 1311-1319.

Matin, A. (1991) The molecular basis of carbon-starvation-induced general resistance in *Escherichia coli*. *Mol Microbiol* **5**: 3-10.

McAtee, C.P., Fry, K.E., and Berg, D.E. (1998) Identification of potential diagnostic and vaccine candidates of *Helicobacter pylori* by "proteome" technologies. *Helicobacter* **3**: 163-169.

McCann, M.P., Kidwell, J.P., and Matin, A. (1991) The putative sigma factor *KatF* has a central role in development of starvation-mediated general resistance in *Escherichia coli*. *J Bacteriol* **173**: 4188-4194.

Miller, J. (1972) *Experiments in Molecular Genetics.* Cold Spring Harbor Press, New York: Cold Spring Harbor Laboratory.

Mogk, A., Schlieker, C., Strub, C., Rist, W., Weibezahn, J., and Bukau, B. (2003) Roles of individual domains and conserved motifs of the AAA+ chaperone ClpB in oligomerization, ATP hydrolysis, and chaperone activity. *J Biol Chem* **278**: 17615-17624.

Mrowka, R., Patzak, A., and Herzel, H. (2001) Is there a bias in proteome research? *Genome Res* **11**: 1971-1973.

Mulvey, M.R., Sorby, P.A., Triggs-Raine, B.L., and Loewen, P.C. (1988) Cloning and physical characterization of katE and katF required for catalase HPII expression in *Escherichia coli*. *Gene* **73**: 337-345.

Neher, S.B., Flynn, J.M., Sauer, R.T., and Baker, T.A. (2003) Latent ClpX-recognition signals ensure LexA destruction after DNA damage. *Genes Dev* **17**: 1084-1089.

Nelson, R.W., Nedelkov, D., and Tubbs, K.A. (2000) Biosensor chip mass spectrometry: a chip-based proteomics approach. *Electrophoresis* **21**: 1155-1163.

Nitta, T., Nagamitsu, H., Murata, M., Izu, H., and Yamada, M. (2000) Function of the sigma(E) regulon in dead-cell lysis in stationary-phase *Escherichia coli*. *J Bacteriol* **182**: 5231-5237.

Nystrom, T. (2003) Conditional senescence in bacteria: death of the immortals. *Mol Microbiol* **48**: 17-23.

Nystrom, T. (2004) Stationary-phase physiology. *Annu Rev Microbiol* **58**: 161-181.

O'Farrell, P.H. (1975) High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**: 4007-4021.

Opiteck, G.J., Lewis, K.C., Jorgenson, J.W., and Anderegg, R.J. (1997) Comprehensive on-line LC/LC/MS of proteins. *Anal Chem* **69**: 1518-1524.

Opiteck, G.J., Ramirez, S.M., Jorgenson, J.W., and Moseley, M.A., 3rd (1998) Comprehensive two-dimensional high-performance liquid chromatography for the isolation of overexpressed proteins and proteome mapping. *Anal Biochem* **258**: 349-361.

Ozaki, M., Wada, A., Fujita, N., and Ishihama, A. (1991) Growth phase-dependent modification of RNA polymerase in *Escherichia coli*. *Mol Gen Genet* **230**: 17-23.

Pappin, D.J.C., Hojrup, P., and Bleasby, A.J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology* **3**: 327-332.

Park, Z.Y., and Russell, D.H. (2000) Thermal denaturation: a useful technique in peptide mass mapping. *Anal Chem* **72**: 2667-2670.

Park, Z.Y., and Russell, D.H. (2001) Identification of individual proteins in complex protein mixtures by high-resolution, high-mass-accuracy MALDI TOF-mass spectrometry analysis of in-solution thermal denaturation/enzymatic digestion. *Anal Chem* **73**: 2558-2564.

Patel, H.V., Vyas, K.A., Li, X., Savtchenko, R., and Roseman, S. (2004) Subcellular distribution of enzyme I of the *Escherichia coli* phosphoenolpyruvate: glucose phosphotransferase system depends on growth conditions. *Proc Natl Acad Sci U S A* **101**: 17486-17491.

Patten, C.L., Kirchhof, M.G., Schertzberg, M.R., Morton, R.A., and Schellhorn, H.E. (2004) Microarray analysis of RpoS-mediated gene expression in *Escherichia coli* K-12. *Mol Genet Genomics* **272**: 580-591.

Peng, J., and Gygi, S.P. (2001) Proteomics: the move to mixtures. *J Mass Spectrom* **36**: 1083-1091.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**: 3551-3567.

Perrot, F., Hebraud, M., Charlionet, R., Junter, G.A., and Jouenne, T. (2000) Protein patterns of gel-entrapped *Escherichia coli* cells differ from those of free-floating organisms. *Electrophoresis* **21**: 645-653.

Rabilloud, T. (2002) Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* **2**: 3-10.

Rao, N.N., and Kornberg, A. (1999) Inorganic polyphosphate regulates responses of *Escherichia coli* to nutritional stringencies, environmental stresses and survival in the stationary phase. *Prog Mol Subcell Biol* **23**: 183-195.

Resing, K.A., and Ahn, N.G. (2005) Proteomics strategies for protein identification. *FEBS Lett* **579**: 885-889.

Russell, W.K., Park, Z.Y., and Russell, D.H. (2001) Proteolysis in mixed organic-aqueous solvent systems: applications for peptide mass mapping using mass spectrometry. *Anal Chem* **73**: 2682-2685.

Sak, B.D., Eisenstark, A., and Touati, D. (1989) Exonuclease III and the catalase hydroperoxidase II in *Escherichia coli* are both regulated by the *katF* gene product. *Proc Natl Acad Sci U S A* **86**: 3271-3275.

Sauer, R.T., Bolon, D.N., Burton, B.M., Burton, R.E., Flynn, J.M., Grant, R.A., Hersch, G.L., Joshi, S.A., Kenniston, J.A., Levchenko, I., Neher, S.B., Oakes, E.S., Siddiqui, S.M., Wah, D.A., and Baker, T.A. (2004) Sculpting the proteome with AAA(+) proteases and disassembly machines. *Cell* **119**: 9-18.

Schultz, J.E., Latter, G.I., and Matin, A. (1988) Differential regulation by cyclic AMP of starvation protein synthesis in *Escherichia coli*. *J Bacteriol* **170**: 3903-3909.

Schwikowski, B., Uetz, P., and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18**: 1257-1261.

Serres, M.H., Goswami, S., and Riley, M. (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res* **32 Database issue**: D300-302.

Sharp, P.M., and Li, W.H. (1987) The Codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**: 1281-1295.

Shefcheck, K., Yao, X., and Fenselau, C. (2003) Fractionation of cytosolic proteins on an immobilized heparin column. *Anal Chem* **75**: 1691-1698.

Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Boucherie, H., and Mann, M. (1996a) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci U S A* **93**: 14440-14445.

Shevchenko, A., Wilm, M., Vorm, O., Jensen, O.N., Podtelejnikov, A.V., Neubauer, G., Mortensen, P., and Mann, M. (1996b) A strategy for identifying gel-separated proteins in sequence databases by MS alone. *Biochem Soc Trans* **24**: 893-896.

Stasyk, T., and Huber, L.A. (2004) Zooming in: fractionation strategies in proteomics. *Proteomics* **4**: 3704-3716.

Stephani, K., Weichart, D., and Hengge, R. (2003) Dynamic control of Dps protein levels by ClpXP and ClpAP proteases in *Escherichia coli*. *Mol Microbiol* **49**: 1605-1614.

Strom, A.R., and Kaasen, I. (1993) Trehalose metabolism in *Escherichia coli*: stress protection and stress regulation of gene expression. *Mol Microbiol* **8**: 205-210.

Sundararaj, S., Guo, A., Habibi-Nazhad, B., Rouani, M., Stothard, P., Ellison, M., and Wishart, D.S. (2004) The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of *Escherichia coli*. *Nucleic Acids Res* **32 Database issue**: D293-295.

Tani, T.H., Khodursky, A., Blumenthal, R.M., Brown, P.O., and Matthews, R.G. (2002) Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis. *Proc Natl Acad Sci U S A* **99**: 13471-13476.

Tao, H., Bausch, C., Richmond, C., Blattner, F.R., and Conway, T. (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J Bacteriol* **181**: 6425-6440.

Tonella, L., Walsh, B.J., Sanchez, J.C., Ou, K., Wilkins, M.R., Tyler, M., Frutiger, S., Gooley, A.A., Pescaru, I., Appel, R.D., Yan, J.X., Bairoch, A., Hoogland, C., Morch, F.S., Hughes, G.J., Williams, K.L., and Hochstrasser, D.F. (1998) '98 *Escherichia coli* SWISS-2DPAGE database update. *Electrophoresis* **19**: 1960-1971.

Tonella, L., Hoogland, C., Binz, P.A., Appel, R.D., Hochstrasser, D.F., and Sanchez, J.C. (2001) New perspectives in the *Escherichia coli* proteome investigation. *Proteomics* **1**: 409-423.

Tong, A.H., Drees, B., Nardelli, G., Bader, G.D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C.W., Fields, S., Boone, C., and Cesareni, G. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**: 321-324.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J.M. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**: 623-627.

Uetz, P., and Hughes, R.E. (2000) Systematic and large-scale two-hybrid screens. *Curr Opin Microbiol* **3**: 303-308.

Unden, G. (1988) Differential roles for menaquinone and demethylmenaquinone in anaerobic electron transport of *E. coli* and their *fnr*-independent expression. *Arch Microbiol* **150**: 499-503.

Unden, G., and Bongaerts, J. (1997) Alternative respiratory pathways of *Escherichia coli:* energetics and transcriptional regulation in response to electron acceptors. *Biochim Biophys Acta* **1320**: 217-234.

VanBogelen, R.A., Abshire, K.Z., Pertsemlidis, A., Clark, R.L., and Neidhardt, F.C. (1996) Gene-Protein Database of *Escherichia coli* K-12: Edition 6. In Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology*. Vol. 2. Neidhardt, F.C., Curtiss, R.C.I., Ingraham, J.L., Lin, E.C.C., Low, K.B., Magasanik, B., Reznikoff, W.S., Riley, M., Schaechter, M. and Umbarger, H.E. (eds). Washington, DC: ASM Press, pp. 2067-2117.

VanBogelen, R.A. (1999) Generating a bacterial genome inventory. Identifying 2-D spots by comigrating products of the genome on 2-D gels. *Methods Mol Biol* **112**: 423-429.

VanBogelen, R.A., Greis, K.D., Blumenthal, R.M., Tani, T.H., and Matthews, R.G. (1999a) Mapping regulatory networks in microbial cells. *Trends Microbiol* **7**: 320-328.

VanBogelen, R.A., Schiller, E.E., Thomas, J.D., and Neidhardt, F.C. (1999b) Diagnosis of cellular states of microbial organisms using proteomics. *Electrophoresis* **20**: 2149-2159.

Vijayakumar, S.R., Kirchhof, M.G., Patten, C.L., and Schellhorn, H.E. (2004) RpoS-regulated genes of *Escherichia coli* identified by random *lacZ* fusion mutagenesis. *J Bacteriol* **186**: 8499-8507.

Vollmer, M., Horth, P., and Nagele, E. (2004) Optimization of two-dimensional off-line LC/MS separations to improve resolution of complex proteomic samples. *Anal Chem* **76**: 5180-5185.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**: 399-403.

Washburn, M.P., and Yates, J.R., 3rd (2000) Analysis of the microbial proteome. *Curr Opin Microbiol* **3**: 292-297.

Washburn, M.P., Wolters, D., and Yates, J.R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**: 242-247.

Wasinger, V.C., Cordwell, S.J., Cerpa-Poljak, A., Yan, J.X., Gooley, A.A., Wilkins, M.R., Duncan, M.W., Harris, R., Williams, K.L., and Humphery-Smith, I. (1995) Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium. Electrophoresis* **16**: 1090-1094.

Wasinger, V.C., and Humphery-Smith, I. (1998) Small genes/gene-products in *Escherichia coli* K-12. *FEMS Microbiol Lett* **169**: 375-382.

Weber, A., and Jung, K. (2002) Profiling early osmostress-dependent gene expression in *Escherichia coli* using DNA macroarrays. *J Bacteriol* **184**: 5502-5507.

Weber, H., Polen, T., Heuveling, J., Wendisch, V.F., and Hengge, R. (2005) Genome-wide analysis of the general stress response network in *Escherichia coli*: sigmaS-dependent genes, promoters, and sigma factor selectivity. *J Bacteriol* **187**: 1591-1603.

Wei, Y., Lee, J.M., Richmond, C., Blattner, F.R., Rafalski, J.A., and LaRossa, R.A. (2001) High-density microarray-mediated gene expression profiling of *Escherichia coli. J Bacteriol* **183**: 545-556.

Weichart, D., Querfurth, N., Dreger, M., and Hengge-Aronis, R. (2003) Global role for ClpP-containing proteases in stationary-phase adaptation of *Escherichia coli*. *J Bacteriol* **185**: 115-125.

Welsh, D.T., Reed, R.H., and Herbert, R.A. (1991) The role of trehalose in the osmoadaptation of *Escherichia coli* NCIB 9484: interaction of trehalose, $K^+$ and glutamate during osmoadaptation in continuous culture. *J Gen Microbiol* **137 ( Pt 4)**: 745-750.

Westbrook, J.A., Yan, J.X., Wait, R., and Dunn, M.J. (2001) A combined radiolabelling and silver staining technique for improved visualisation, localisation, and identification of proteins separated by two-dimensional gel electrophoresis. *Proteomics* **1**: 370-376.

Wilkins, M.R., Pasquali, C., Appel, R.D., Ou, K., Golaz, O., Sanchez, J.C., Yan, J.X., Gooley, A.A., Hughes, G., Humphery-Smith, I., Williams, K.L., and Hochstrasser, D.F. (1996) From proteins to proteomes: large scale protein identification by two- dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)* **14**: 61-65.

Wilkins, M.R., Gasteiger, E., Tonella, L., Ou, K., Tyler, M., Sanchez, J.C., Gooley, A.A., Walsh, B.J., Bairoch, A., Appel, R.D., Williams, K.L., and Hochstrasser, D.F. (1998) Protein identification with N and C-terminal sequence tags in proteome projects. *J Mol Biol* **278**: 599-608.

Wilkins, M.R., Gasteiger, E., Gooley, A.A., Herbert, B.R., Molloy, M.P., Binz, P.A., Ou, K., Sanchez, J.C., Bairoch, A., Williams, K.L., and Hochstrasser, D.F. (1999) High-throughput mass spectrometric discovery of protein post- translational modifications. *J Mol Biol* **289**: 645-657.

Wise, M.J., Littlejohn, T., and Humphery-Smith, I. (1997a) Better cutters for protein mass fingerprinting: preliminary findings. *ISMB* **5**: 340-343.

Wise, M.J., Littlejohn, T.G., and Humphery-Smith, I. (1997b) Peptide-mass fingerprinting and the ideal covering set for protein characterisation. *Electrophoresis* **18**: 1399-1409.

Wolters, D.A., Washburn, M.P., and Yates, J.R., 3rd (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* **73**: 5683-5690.

Wysocki, V.H., Resing, K.A., Zhang, Q., and Cheng, G. (2005) Mass spectrometry of peptides and proteins. *Methods* **35**: 211-222.

Yano, H., Kuroda, S., and Buchanan, B.B. (2002) Disulfide proteome in the analysis of protein function and structure. *Proteomics* **2**: 1090-1096.

Yates, J.R., 3rd, Speicher, S., Griffin, P.R., and Hunkapiller, T. (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal Biochem* **214**: 397-408.

Yates, J.R., 3rd (1998) Mass spectrometry and the age of the proteome. *J Mass Spectrom* **33**: 1-19.

Yates, J.R., 3rd (2004) Mass spectral analysis in proteomics. *Annu Rev Biophys Biomol Struct* **33**: 297-316.

Zhang, W., and Chait, B.T. (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem* **72**: 2482-2489.

Zhen, Y., Xu, N., Richardson, B., Becklin, R., Savage, J.R., Blake, K., and Peltier, J.M. (2004) Development of an LC-MALDI method for the analysis of protein complexes. *J Am Soc Mass Spectrom* **15**: 803-822.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R.A., Gerstein, M., and Snyder, M. (2001) Global analysis of protein activities using proteome chips. *Science* **293**: 2101-2105.

Zimmer, D.P., Soupene, E., Lee, H.L., Wendisch, V.F., Khodursky, A.B., Peter, B.J., Bender, R.A., and Kustu, S. (2000) Nitrogen regulatory protein C-controlled genes of *Escherichia coli*: scavenging as a defense against nitrogen limitation. *Proc Natl Acad Sci U S A* **97**: 14674-14679.

Zinser, E.R., and Kolter, R. (1999) Mutations enhancing amino acid catabolism confer a growth advantage in stationary phase. *J Bacteriol* **181**: 5800-5807.

Zinser, E.R., and Kolter, R. (2000) Prolonged stationary-phase incubation selects for *lrp* mutations in *Escherichia coli* K-12. *J Bacteriol* **182**: 4361-4365.

# APPENDIX

Supplementary Data for Chapter II

Additional Data is also available at http://eep.tamu.edu

Appendix A-1. Non redundant protein ID list from original proteome. It also includes the frequency of protein ID.

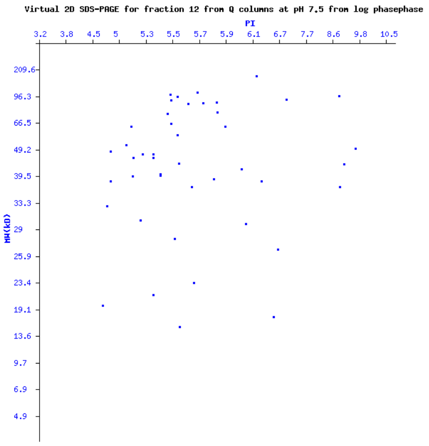
| SWISS ID | # Times ID'd | SWISS ID | # Times ID'd | SWISS ID | # Times ID'd |
|---|---|---|---|---|---|
| P02997 | 67 | P00479 | 11 | P21165 | 5 |
| P22257 | 58 | P02408 | 11 | P21774 | 5 |
| P02996 | 57 | P02995 | 11 | P30746 | 5 |
| P04079 | 50 | P11665 | 11 | P30867 | 5 |
| P02349 | 45 | P24233 | 11 | P31120 | 5 |
| P00350 | 44 | P30148 | 11 | P32132 | 5 |
| P02418 | 39 | P33221 | 11 | P37647 | 5 |
| P39171 | 39 | Q46829 | 11 | P40711 | 5 |
| P06958 | 38 | P00909 | 10 | P00453 | 4 |
| P17242 | 36 | P04790 | 10 | P02371 | 4 |
| P06977 | 35 | P04805 | 10 | P02422 | 4 |
| P15639 | 33 | P06986 | 10 | P06715 | 4 |
| P00391 | 31 | P07395 | 10 | P06959 | 4 |
| P00477 | 27 | P27248 | 10 | P06960 | 4 |
| P11096 | 25 | P30856 | 10 | P07004 | 4 |
| P36857 | 25 | P04036 | 9 | P07638 | 4 |
| P08936 | 24 | P05082 | 9 | P08179 | 4 |
| P00928 | 23 | P06994 | 9 | P09831 | 4 |
| P05640 | 23 | P08312 | 9 | P14926 | 4 |
| P26427 | 23 | P25524 | 9 | P15002 | 4 |
| P29132 | 23 | P29464 | 9 | P22106 | 4 |
| P12283 | 22 | P00478 | 8 | P23843 | 4 |
| P23721 | 22 | P02339 | 8 | P30136 | 4 |
| P27302 | 22 | P07813 | 8 | P31142 | 4 |
| P02990 | 21 | P08200 | 8 | P31803 | 4 |
| P17288 | 21 | P08324 | 8 | P37901 | 4 |
| P16659 | 20 | P31216 | 8 | P80449 | 4 |
| P07118 | 19 | P37759 | 8 | P00575 | 3 |
| P05055 | 18 | P76492 | 8 | P00577 | 3 |
| P31217 | 18 | P04475 | 7 | P00907 | 3 |
| P00574 | 17 | P07912 | 7 | P00934 | 3 |
| P00891 | 17 | P11537 | 7 | P00956 | 3 |
| P06981 | 17 | P18335 | 7 | P02420 | 3 |
| P13030 | 17 | P22259 | 7 | P05020 | 3 |
| P21889 | 16 | P22767 | 7 | P07672 | 3 |
| P00509 | 15 | P27827 | 7 | P09029 | 3 |
| P03003 | 15 | P02372 | 6 | P09030 | 3 |
| P03948 | 15 | P04804 | 6 | P09158 | 3 |
| P00962 | 14 | P06711 | 6 | P09373 | 3 |
| P14178 | 14 | P06998 | 6 | P09625 | 3 |
| P15046 | 14 | P07460 | 6 | P15034 | 3 |
| P16936 | 14 | P11604 | 6 | P17169 | 3 |
| P00496 | 13 | P23839 | 6 | P25528 | 3 |
| P05313 | 13 | P33918 | 6 | P25665 | 3 |
| P05838 | 13 | P02419 | 5 | P27828 | 3 |
| P11447 | 13 | P04391 | 5 | P32164 | 3 |
| P19245 | 13 | P05380 | 5 | P37747 | 3 |
| P09156 | 12 | P06968 | 5 | P45578 | 3 |
| P21155 | 12 | P08374 | 5 | P75805 | 3 |
| P21346 | 12 | P09170 | 5 | P00957 | 2 |

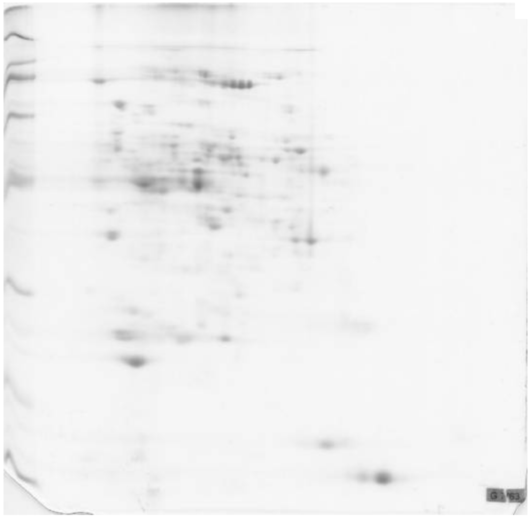
Appendix A-1 Appendix Continued…

| SWISS ID | # Times ID'd | SWISS ID | # Times ID'd | SWISS ID | # Times ID'd |
|---|---|---|---|---|---|
| P00968 | 2 | P04951 | 1 | P31130 | 1 |
| P02392 | 2 | P05021 | 1 | P31220 | 1 |
| P02416 | 2 | P05194 | 1 | P31451 | 1 |
| P04825 | 2 | P05850 | 1 | P31453 | 1 |
| P08244 | 2 | P06139 | 1 | P31456 | 1 |
| P09028 | 2 | P06987 | 1 | P32147 | 1 |
| P09097 | 2 | P07010 | 1 | P33138 | 1 |
| P09743 | 2 | P07102 | 1 | P33195 | 1 |
| P10378 | 2 | P08398 | 1 | P33570 | 1 |
| P11446 | 2 | P08837 | 1 | P33633 | 1 |
| P11668 | 2 | P08839 | 1 | P36645 | 1 |
| P12758 | 2 | P08997 | 1 | P36663 | 1 |
| P15039 | 2 | P09157 | 1 | P36766 | 1 |
| P21179 | 2 | P09200 | 1 | P37028 | 1 |
| P22783 | 2 | P09378 | 1 | P37048 | 1 |
| P22885 | 2 | P09454 | 1 | P37095 | 1 |
| P23851 | 2 | P10101 | 1 | P37197 | 1 |
| P23869 | 2 | P10121 | 1 | P37651 | 1 |
| P24167 | 2 | P10366 | 1 | P37666 | 1 |
| P24991 | 2 | P13034 | 1 | P37744 | 1 |
| P25888 | 2 | P15640 | 1 | P37751 | 1 |
| P27252 | 2 | P16528 | 1 | P38489 | 1 |
| P28694 | 2 | P16688 | 1 | P39265 | 1 |
| P29217 | 2 | P16921 | 1 | P39320 | 1 |
| P36950 | 2 | P17117 | 1 | P39343 | 1 |
| P37689 | 2 | P17846 | 1 | P39356 | 1 |
| P39330 | 2 | P17854 | 1 | P39435 | 1 |
| P40681 | 2 | P19641 | 1 | P40120 | 1 |
| P46853 | 2 | P19797 | 1 | P43781 | 1 |
| P52054 | 2 | P23836 | 1 | P45392 | 1 |
| P75864 | 2 | P23847 | 1 | P45465 | 1 |
| P00501 | 1 | P23863 | 1 | P45467 | 1 |
| P00837 | 1 | P23932 | 1 | P45535 | 1 |
| P00882 | 1 | P24171 | 1 | P46132 | 1 |
| P00961 | 1 | P24182 | 1 | P52065 | 1 |
| P00963 | 1 | P24234 | 1 | P52697 | 1 |
| P02356 | 1 | P24238 | 1 | P56604 | 1 |
| P02358 | 1 | P24253 | 1 | P71295 | 1 |
| P02363 | 1 | P25538 | 1 | P75844 | 1 |
| P02364 | 1 | P25895 | 1 | P75914 | 1 |
| P02366 | 1 | P26282 | 1 | P75915 | 1 |
| P02370 | 1 | P26428 | 1 | P75969 | 1 |
| P02409 | 1 | P26612 | 1 | P76008 | 1 |
| P02413 | 1 | P27511 | 1 | P76052 | 1 |
| P02428 | 1 | P27836 | 1 | P76056 | 1 |
| P02999 | 1 | P28302 | 1 | P76069 | 1 |
| P03815 | 1 | P28688 | 1 | P76250 | 1 |
| P04422 | 1 | P30125 | 1 | P76259 | 1 |
| P04425 | 1 | P30745 | 1 | P76513 | 1 |
| P77565 | 1 | P30747 | 1 | P76577 | 1 |
| P77601 | 1 | P30979 | 1 | P76641 | 1 |
| P77754 | 1 | P31057 | 1 | P77493 | 1 |
| P77770 | 1 | P77804 | 1 | Q47140 | 1 |
|  |  |  |  | Q47269 | 1 |

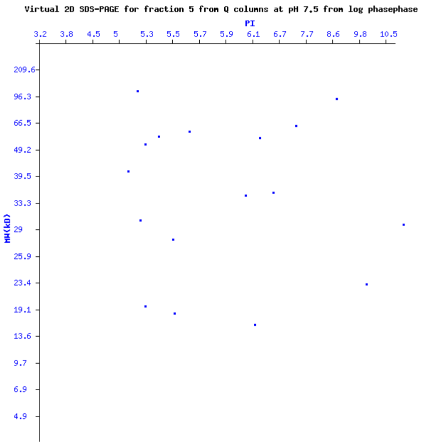
Supplemental Data For Chapter III
A-2. Example Virtual and Actual 2D gel from Log Fraction 12 &5 pH 7.50
http://eep.tamu.edu/nondelc

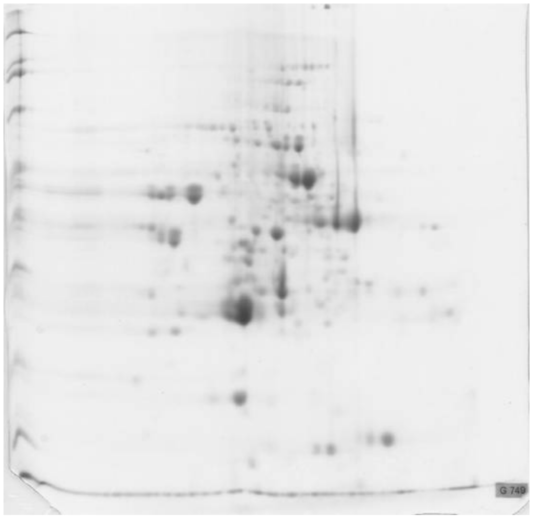
## pH 7.50 Frac 12





## pH 7.50 Frac 5





## A-3 Non Redundant List of Identified Proteins From Log and Stationary Phase Cells.

**Non-redundant list of proteins identified by 2D-LC**

hits_exp: hits in exponential phase at pH7.5 and pH8.75

hits_stat: hits in stationary phase at pH7.5 and pH8.75

ratio: Round(log(hits_exp/hits_stat,2),0)

notes: proteins is also found in one or more other proteomic study

Some other studies of log phase E.coli K12

1. Swiss_2D PAGE : late log (OD600=1), E. coli K12 W3110, MOPS
2. Cyber Cell Project 2D PAGE: mid log E. coli K12
3. Corbin's 2D LC: mid log (OD600=0.4), E.coli K12 MG1655, minimal medium, 0.2% glycerol
4. Champion's: mid log (OD600=0.5), E.coli K12 MG1655, M9, glucose

Elution shift: protein has 3 or more elution shift from log to stationary phase in either or both of Q and Phe columns

Stat. Δ: "+" means up-regulated in stationary phase, "-" means down-regulated in stationary phase, blank means no significant cha

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P31119 | aas | 2 | 2 | 0 | | + | |
| P05313 | aceA | 17 | 39 | -1.2 | 1, 2, 3, 4 | | + |
| P08997 | aceB | 3 | 2 | 0.58 | 3, 4 | | |
| P06958 | aceE | 17 | 26 | -0.61 | 1, 2, 3, 4 | + | |
| P06959 | aceF | 14 | 3 | 2.22 | 1, 2, 3, 4 | + | - |
| P11071 | aceK | 2 | 1 | 1 | | | |
| P15046 | ackA | 3 | 1 | 1.58 | 1, 2, 3, 4 | | - |
| P25516 | acnA | 2 | 15 | -2.91 | 3 | | + |
| P17547 | adhE | 3 | 5 | -0.74 | 2, 3 | | |
| P05082 | adk | 11 | 3 | 1.87 | 1, 2, 3, 4 | | - |
| P26427 | ahpC | 27 | 4 | 2.75 | 1, 2, 3, 4 | + | - |
| P00957 | alaS | 11 | 22 | -1 | 1, 2, 3, 4 | + | |
| P25553 | aldA | 12 | 25 | -1.06 | 1, 2, 3 | + | + |
| P26612 | amyA | 2 | 3 | -0.58 | 4 | | |
| P11446 | argC | 4 | 2 | 1 | 4 | | |
| P18335 | argD | 17 | 8 | 1.09 | 1, 3, 4 | | - |
| P22767 | argG | 9 | 19 | -1.08 | 1, 2, 3, 4 | | + |
| P11447 | argH | 24 | 12 | 1 | 3, 4 | | |
| P11875 | argS | 1 | 1 | 0 | 1, 3 | | |
| P05194 | aroD | 3 | 1 | 1.58 | 1, 2, 4 | | - |
| P00886 | aroG | 2 | 1 | 1 | 1, 3 | + | |
| P24167 | aroK | 1 | 1 | 0 | 1, 2, 4 | | |
| P00353 | asd | 6 | 3 | 1 | 1, 2, 3 | + | |
| P22106 | asnB | 1 | 9 | -3.17 | 2, 3, 4 | + | + |
| P17242 | asnS | 21 | 18 | 0.22 | 1, 3, 4 | | |
| P00509 | aspC | 14 | 11 | 0.35 | 1, 2, 3, 4 | + | |
| P21889 | aspS | 13 | 14 | -0.11 | 1, 3, 4 | + | |
| P00822 | atpA | 11 | 9 | 0.29 | 1, 2, 3 | | |
| P00824 | atpD | 12 | 9 | 0.42 | 1, 2, 3 | + | |
| P00837 | atpG | 1 | 1 | 0 | 3, 4 | + | |

A-3 Continued

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P23480 | bcp | 1 | 2 | -1 | 1, 2, 3 | | |
| P33363 | bglX | 4 | 1 | 2 | 2 | + | - |
| P00968 | carB | 4 | 26 | -2.7 | 3, 4 | + | + |
| P15716 | clpA | 3 | 4 | -0.42 | 3 | + | |
| P03815 | clpB | 11 | 43 | -1.97 | 1, 2, 3, 4 | | + |
| P19245 | clpP | 14 | 11 | 0.35 | 1, 2, 4 | + | |
| P25524 | codA | 11 | 7 | 0.65 | 3, 4 | + | |
| P00936 | cyaA | 1 | 4 | -2 | | + | + |
| P17846 | cysI | 11 | 3 | 1.87 | 3, 4 | + | - |
| P11096 | cysK | 28 | 17 | 0.72 | 1, 2, 3, 4 | + | |
| P16703 | cysM | 7 | 1 | 2.81 | 1 | + | - |
| P16700 | cysP | 3 | 1 | 1.58 | 1, 3 | + | - |
| P05640 | dapA | 9 | 4 | 1.17 | 1, 2, 3, 4 | + | - |
| P04036 | dapB | 16 | 13 | 0.3 | 1, 2, 4 | | |
| P03948 | dapD | 20 | 5 | 2 | 1, 2, 3, 4 | + | - |
| P24171 | dcp | 6 | 5 | 0.26 | 3, 4 | + | |
| P00882 | deoC | 6 | 3 | 1 | 1, 2, 4 | | |
| P77254 | der | 2 | 3 | -0.58 | | | |
| P18274 | dksA | 1 | 2 | -1 | 1, 2 | | |
| P06149 | dld | 2 | 1 | 1 | 3 | + | |
| P04475 | dnaK | 101 | 128 | -0.34 | 1, 2, 3, 4 | | |
| P08324 | eno | 5 | 23 | -2.2 | 1, 2, 3, 4 | | + |
| P14926 | fabB | 6 | 5 | 0.26 | 2, 3, 4 | + | |
| P25715 | fabD | 5 | 1 | 2.32 | 1, 2 | | - |
| P39435 | fabF | 8 | 3 | 1.42 | 4 | + | - |
| P29132 | fabI | 18 | 3 | 2.58 | 1, 2, 3, 4 | + | - |
| P11604 | fbaA | 5 | 3 | 0.74 | 1, 3, 4 | + | |
| P71295 | fbaB | 8 | 8 | 0 | 3, 4 | | |
| P24183 | fdnG | 2 | 5 | -1.32 | | + | + |
| P25528 | fdx | 9 | 5 | 0.85 | 4 | | |
| P23882 | fmt | 1 | 1 | 0 | 1, 2, 3 | + | |
| P00363 | frdA | 3 | 3 | 0 | | + | |
| P00923 | fumA | 1 | 2 | -1 | 3 | + | |
| P02996 | fusA | 51 | 32 | 0.67 | 1, 2, 3, 4 | + | |
| P22256 | gabT | 3 | 2 | 0.58 | 3 | | |
| P80063 | gadA | 1 | 66 | -6.04 | | + | + |
| P40681 | galM | 5 | 1 | 2.32 | 1, 2, 3, 4 | | - |
| P06977 | gapA | 28 | 15 | 0.9 | 1, 2, 3, 4 | + | |
| P15877 | gcd | 1 | 1 | 0 | 2 | + | |
| P33195 | gcvP | 3 | 3 | 0 | 3, 4 | + | |
| P27248 | gcvT | 4 | 3 | 0.42 | 2, 4 | | |
| P00370 | gdhA | 7 | 6 | 0.22 | 1, 2, 3 | | |
| P37330 | glcB | 2 | 4 | -1 | 3 | + | |
| P37747 | glf | 1 | 1 | 0 | 3, 4 | | |

A-3 Continued

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P31120 | glmM | 1 | 1 | 0 | 3, 4 | | |
| P17169 | glmS | 8 | 10 | -0.32 | 3, 4 | | |
| P06711 | glnA | 12 | 13 | -0.12 | 1, 2, 3, 4 | | |
| P27249 | glnD | 2 | 1 | 1 | | + | |
| P00962 | glnS | 10 | 9 | 0.15 | 1, 2, 3, 4 | + | |
| P13034 | glpC | 2 | 1 | 1 | 2, 4 | + | |
| P08859 | glpK | 1 | 3 | -1.58 | 1, 2, 3 | + | + |
| P00891 | gltA | 17 | 16 | 0.09 | 3, 4 | + | |
| P09831 | gltB | 80 | 120 | -0.58 | 3, 4 | | |
| P09832 | gltD | 6 | 11 | -0.87 | 1, 2, 3 | | |
| P04805 | gltX | 6 | 6 | 0 | 3, 4 | | |
| P00477 | glyA | 29 | 16 | 0.86 | 1, 2, 3, 4 | + | |
| P00350 | gnd | 19 | 21 | -0.14 | 2, 3, 4 | + | |
| P06715 | gor | 5 | 6 | -0.26 | 1, 2, 3, 4 | + | |
| P31217 | gpmA | 13 | 10 | 0.38 | 1, 2, 3, 4 | | |
| P21346 | greA | 18 | 15 | 0.26 | 1, 2, 3, 4 | | |
| P06139 | groL | 8 | 12 | -0.58 | 1, 2, 3 | + | |
| P09372 | grpE | 1 | 2 | -1 | 1, 2 | + | |
| P04425 | gshB | 4 | 1 | 2 | 4 | | - |
| P04079 | guaA | 29 | 18 | 0.69 | 3, 4 | + | |
| P06981 | guaB | 15 | 20 | -0.42 | 1, 2, 3, 4 | + | |
| P09097 | gyrA | 9 | 5 | 0.85 | 1, 2, 3, 4 | + | |
| P06982 | gyrB | 2 | 1 | 1 | 3 | + | |
| P15002 | hemB | 3 | 1 | 1.58 | 4 | | - |
| P29680 | hemE | 2 | 1 | 1 | 2 | | |
| P23893 | hemL | 2 | 1 | 1 | 3 | | |
| P06986 | hisC | 10 | 2 | 2.32 | 3, 4 | | - |
| P10373 | hisF | 10 | 3 | 1.74 | 2 | + | - |
| P04804 | hisS | 3 | 7 | -1.22 | 3, 4 | + | + |
| P08936 | hns | 7 | 1 | 2.81 | 1, 2, 3, 4 | + | - |
| P36766 | hpt | 2 | 2 | 0 | 1, 2, 4 | | |
| P36541 | hscA | 3 | 1 | 1.58 | | | - |
| P37595 | iaaA | 2 | 2 | 0 | | | |
| P39377 | iadA | 2 | 1 | 1 | 2 | | |
| P08200 | icd | 25 | 22 | 0.18 | 1, 2, 3, 4 | + | |
| P00956 | ileS | 10 | 4 | 1.32 | 3, 4 | | - |
| P05793 | ilvC | 13 | 9 | 0.53 | 1, 2, 3, 4 | | |
| P00510 | ilvE | 10 | 2 | 2.32 | 1,2 | + | - |
| P02995 | infB | 12 | 12 | 0 | 3, 4 | + | |
| P39171 | iscS | 43 | 17 | 1.34 | 1, 3, 4 | + | - |
| P19641 | ispB | 2 | 5 | -1.32 | 4 | | + |

A-3 Continued…

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| sp_id | eco_gn | Exp. | Stat. | ratio | notes | Elution Shift | Stat. Δ |
| P21179 | katE | 12 | 23 | -0.94 | 3, 4 | | |
| P13029 | katG | 8 | 11 | -0.46 | 1, 3 | + | |
| P37647 | kdgK | 6 | 2 | 1.58 | 2, 4 | + | - |
| P52643 | ldhA | 5 | 12 | -1.26 | | | + |
| P07682 | lepA | 2 | 2 | 0 | 3 | + | |
| P30125 | leuB | 12 | 7 | 0.78 | 3, 4 | + | |
| P07813 | leuS | 7 | 6 | 0.22 | 1, 2, 3, 4 | + | |
| P04816 | livK | 20 | 3 | 2.74 | 1, 2, 3 | + | - |
| P08177 | lon | 2 | 3 | -0.58 | 3 | + | |
| P00391 | lpd | 39 | 38 | 0.04 | 1, 2, 3, 4 | | |
| P08660 | lysC | 5 | 4 | 0.32 | 3 | | |
| P13030 | lysS | 14 | 10 | 0.49 | 1, 3, 4 | | |
| P14825 | lysU | 1 | 1 | 0 | | | |
| P76558 | maeB | 1 | 8 | -3 | 3 | | + |
| P00490 | malP | 11 | 8 | 0.46 | 3 | | |
| P08186 | manX | 1 | 1 | 0 | 1, 2 | | |
| P06994 | mdh | 12 | 2 | 2.58 | 3, 4 | | - |
| P33136 | mdoG | 4 | 1 | 2 | 1, 2, 3 | + | - |
| P06721 | metC | 3 | 2 | 0.58 | 3 | | |
| P25665 | metE | 38 | 15 | 1.34 | 3, 4 | + | - |
| P30746 | moaB | 18 | 18 | 0 | 1, 2 | | |
| P28694 | mog | 4 | 2 | 1 | 1, 4 | | |
| P00453 | nrdB | 4 | 5 | -0.32 | 4 | | |
| P03003 | nusA | 9 | 7 | 0.36 | 1, 2, 3, 4 | + | |
| P31663 | panC | 2 | 1 | 1 | 1, 2, 3 | | |
| P22259 | pck | 3 | 2 | 0.58 | 3, 4 | | |
| P11648 | pepA | 1 | 2 | -1 | 3 | | |
| P04825 | pepN | 1 | 4 | -2 | 3, 4 | + | + |
| P15034 | pepP | 12 | 10 | 0.26 | 3, 4 | | |
| P21165 | pepQ | 5 | 6 | -0.26 | 2, 3, 4 | | |
| P06998 | pfkA | 11 | 4 | 1.46 | 1, 2, 4 | + | - |
| P09373 | pflB | 12 | 31 | -1.37 | 1, 2, 3, 4 | + | + |
| P11537 | pgi | 4 | 3 | 0.42 | 3, 4 | + | |
| P11665 | pgk | 10 | 6 | 0.74 | 1, 2, 3, 4 | + | |
| P08312 | pheS | 11 | 13 | -0.24 | 1, 2, 3, 4 | + | |
| P07395 | pheT | 11 | 16 | -0.54 | 1, 2, 3, 4 | + | |
| P05055 | pnp | 15 | 22 | -0.55 | 1, 2, 3, 4 | | |
| P17288 | ppa | 17 | 12 | 0.5 | 1, 2, 4 | | |
| P00864 | ppc | 5 | 10 | -1 | 3 | + | |
| P77241 | ppiD | 2 | 3 | -0.58 | | | |
| P28688 | ppk | 3 | 14 | -2.22 | 4 | + | + |
| P07011 | prfA | 5 | 1 | 2.32 | | | - |

A-3 Continued…

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P07012 | prfB | 2 | 2 | 0 | 3 | + | |
| P27298 | prlC | 2 | 1 | 1 | 3 | | |
| P16659 | proS | 5 | 10 | -1 | 1, 2, 3, 4 | + | |
| P39184 | pta | 1 | 2 | -1 | 1, 2, 3 | | |
| P08839 | ptsI | 3 | 10 | -1.74 | 1, 2, 3, 4 | | + |
| P12283 | purA | 17 | 12 | 0.5 | 2, 3, 4 | + | |
| P25739 | purB | 2 | 4 | -1 | 2, 3 | + | |
| P21155 | purC | 3 | 3 | 0 | 2, 3, 4 | | |
| P15640 | purD | 2 | 1 | 1 | 2, 3, 4 | | |
| P09028 | purE | 8 | 1 | 3 | 4 | | - |
| P00496 | purF | 4 | 2 | 1 | 3, 4 | | |
| P15639 | purH | 8 | 10 | -0.32 | 2, 3, 4 | + | |
| P09029 | purK | 3 | 1 | 1.58 | 1, 2, 3, 4 | + | - |
| P14178 | pykF | 12 | 7 | 0.78 | 2, 3, 4 | + | |
| P00479 | pyrB | 31 | 49 | -0.66 | 1, 2, 3, 4 | + | |
| P05020 | pyrC | 8 | 12 | -0.58 | 3, 4 | + | |
| P05021 | pyrD | 2 | 1 | 1 | 1, 2, 4 | | |
| P08398 | pyrG | 1 | 2 | -1 | 3, 4 | + | |
| P37759 | rfbB | 6 | 3 | 1 | 2, 3, 4 | + | |
| P25540 | ribE | 1 | 1 | 0 | 2 | + | |
| P23851 | rluC | 7 | 3 | 1.22 | 4 | | - |
| P21499 | rnr | 3 | 3 | 0 | 3 | + | |
| P32661 | rpe | 6 | 2 | 1.58 | 2 | | - |
| P02387 | rplB | 4 | 6 | -0.58 | 3 | + | |
| P02386 | rplC | 2 | 1 | 1 | 3 | + | |
| P02408 | rplJ | 1 | 1 | 0 | 3, 4 | + | |
| P02416 | rplQ | 20 | 5 | 2 | 4 | | - |
| P00574 | rpoA | 13 | 10 | 0.38 | 1, 2, 3, 4 | | |
| P00575 | rpoB | 2 | 2 | 0 | 1, 2, 3, 4 | | |
| P00577 | rpoC | 5 | 4 | 0.32 | 3, 4 | | |
| P02349 | rpsA | 39 | 28 | 0.48 | 1, 2, 3, 4 | + | |
| P02354 | rpsD | 18 | 4 | 2.17 | 3 | + | - |
| P02358 | rpsF | 25 | 23 | 0.12 | 1, 2, 4 | | |
| P33918 | rsuA | 4 | 2 | 1 | 2, 4 | | |
| P23721 | serC | 15 | 10 | 0.58 | 1, 2, 3, 4 | + | |
| P09156 | serS | 6 | 7 | -0.22 | 1, 2, 3, 4 | | |
| P30856 | slyD | 8 | 4 | 1 | 2, 4 | + | |
| P21170 | speA | 2 | 2 | 0 | 3 | | |
| P16936 | speB | 7 | 1 | 2.81 | 4 | | - |
| P09159 | speD | 7 | 5 | 0.49 | | | |
| P02339 | ssb | 15 | 6 | 1.32 | 1, 2, 4 | | - |
| P31142 | sseA | 11 | 15 | -0.45 | 2, 3, 4 | | |
| P07460 | sucC | 7 | 5 | 0.49 | 1, 2, 3, 4 | + | |
| P78258 | talA | 4 | 3 | 0.42 | 2, 3 | | |
| P30148 | talB | 16 | 5 | 1.68 | 1, 2, 3, 4 | + | - |
| P42632 | tdcE | 1 | 2 | -1 | | | |
| P00561 | thrA | 1 | 3 | -1.58 | 3 | + | + |
| P00934 | thrC | 6 | 6 | 0 | 1, 2, 3, 4 | | |

A-3 Continued…

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P00955 | thrS | 10 | 2 | 2.32 | 3 | + | - |
| P22257 | tig | 45 | 37 | 0.28 | 1, 2, 3, 4 | | |
| P27302 | tig | 20 | 6 | 1.74 | 2, 3, 4 | + | - |
| P33570 | tig | 4 | 16 | -2 | 3, 4 | | + |
| P04790 | tig | 10 | 2 | 2.32 | 1, 2, 4 | | - |
| P37901 | tig | 12 | 1 | 3.58 | 1, 2, 3, 4 | + | - |
| P00928 | tig | 8 | 5 | 0.68 | 1, 2, 3, 4 | | |
| P00909 | tig | 1 | 2 | -1 | 3, 4 | + | |
| P02997 | tig | 26 | 11 | 1.24 | 1, 2, 3, 4 | + | - |
| P02990 | tig | 2 | 1 | 1 | 1, 2, 3, 4 | | |
| P32132 | tig | 7 | 6 | 0.22 | 3, 4 | | |
| P27854 | tig | 7 | 1 | 2.81 | | + | - |
| P25532 | tig | 5 | 7 | -0.49 | 1, 2, 3 | + | |
| P28242 | tig | 6 | 3 | 1 | 1, 2 | | |
| P42607 | tig | 5 | 4 | 0.32 | | + | |
| P07118 | tig | 25 | 19 | 0.4 | 3, 4 | + | |
| P52697 | tig | 8 | 4 | 1 | 3, 4 | | |
| P29217 | tig | 6 | 8 | -0.42 | 4 | | |
| P40120 | tig | 4 | 2 | 1 | 2, 4 | | |
| P76492 | tig | 6 | 1 | 2.58 | 2, 3, 4 | | - |
| P37350 | tig | 1 | 5 | -2.32 | | + | + |
| P11668 | tig | 1 | 1 | 0 | 4 | | |
| P23839 | tig | 4 | 4 | 0 | 3, 4 | + | |
| P31465 | tig | 2 | 12 | -2.58 | | | + |
| P39172 | tig | 16 | 7 | 1.19 | | + | - |
| P22992 | tig | 2 | 7 | -1.81 | 1, 3 | + | + |
| P30867 | tig | 1 | 0 | log only | 2, 3, 4 | | |
| P08193 | tig | 1 | 0 | log only | 2 | | |
| P33234 | tig | 1 | 0 | log only | | | |
| P35340 | tig | 1 | 0 | log only | 1, 2, 3 | | |
| P07672 | tig | 4 | 0 | log only | 2, 4 | | - |
| P11445 | tig | 1 | 0 | log only | 3 | | |
| P23908 | tig | 3 | 0 | log only | | | - |
| P06960 | tig | 3 | 0 | log only | 1, 3, 4 | | - |
| P04391 | tig | 3 | 0 | log only | 1, 3, 4 | | - |
| P77690 | tig | 1 | 0 | log only | | | |
| P07638 | tig | 6 | 0 | log only | 1, 4 | | - |
| P04422 | tig | 2 | 0 | log only | 2, 3, 4 | | - |
| P00859 | tig | 1 | 0 | log only | 2 | | |
| P41407 | tig | 1 | 0 | log only | | | |
| Q46829 | tig | 2 | 0 | log only | 3, 4 | | - |
| P06961 | tig | 1 | 0 | log only | | | |
| P17315 | tig | 1 | 0 | log only | 2 | | |
| P33138 | tig | 2 | 0 | log only | 2, 3, 4 | | - |
| P23863 | tig | 5 | 0 | log only | 2, 4 | | - |
| P16244 | tig | 2 | 0 | log only | | | - |
| P08837 | tig | 2 | 0 | log only | 1, 2, 3, 4 | | - |
| P36649 | tig | 2 | 0 | log only | | | - |

A-3 Continued…

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P17854 | tig | 7 | 0 | log only | 3, 4 | | - |
| P08506 | tig | 3 | 0 | log only | | | - |
| P39272 | tig | 1 | 0 | log only | | | |
| P76316 | tig | 2 | 0 | log only | 3 | | - |
| P07862 | tig | 2 | 0 | log only | | | - |
| P09743 | tig | 2 | 0 | log only | 4 | | - |
| P23847 | tig | 3 | 0 | log only | 1, 2, 3, 4 | | - |
| P37313 | tig | 1 | 0 | log only | | | |
| P24991 | tig | 1 | 0 | log only | 1, 2, 4 | | |
| P06968 | tig | 6 | 0 | log only | 1, 2, 4 | | - |
| P10177 | tig | 1 | 0 | log only | 1, 2 | | |
| P02933 | tig | 2 | 0 | log only | | | - |
| P38134 | tig | 2 | 0 | log only | | | - |
| P30854 | tig | 4 | 0 | log only | 2 | | - |
| P42608 | tig | 1 | 0 | log only | | | |
| P25716 | tig | 2 | 0 | log only | 2, 3 | | - |
| P24249 | tig | 1 | 0 | log only | 2 | | |
| P21774 | tig | 7 | 0 | log only | 2, 4 | | - |
| P21177 | tig | 1 | 0 | log only | | | |
| P42593 | tig | 5 | 0 | log only | | | - |
| P09371 | tig | 1 | 0 | log only | | | |
| P26266 | tig | 2 | 0 | log only | | | - |
| P39174 | tig | 1 | 0 | log only | 1, 2, 3 | | |
| P27511 | tig | 3 | 0 | log only | 4 | | - |
| P26282 | tig | 1 | 0 | log only | 4 | | |
| P80449 | tig | 4 | 0 | log only | 4 | | - |
| P23486 | tig | 3 | 0 | log only | | | - |
| P25748 | tig | 1 | 0 | log only | | | |
| P37666 | tig | 2 | 0 | log only | 3, 4 | | - |
| P52073 | tig | 1 | 0 | log only | | | |
| P32665 | tig | 1 | 0 | log only | 2 | | |
| P07762 | tig | 1 | 0 | log only | | | |
| P00584 | tig | 1 | 0 | log only | 3 | | |
| P46880 | tig | 1 | 0 | log only | 3 | | |
| P05826 | tig | 1 | 0 | log only | 2 | | |
| P28860 | tig | 1 | 0 | log only | 2 | | |
| P05380 | tig | 6 | 0 | log only | 1, 3, 4 | | - |
| P06980 | tig | 2 | 0 | log only | 3 | | - |
| P10371 | tig | 6 | 0 | log only | 2 | | - |
| P16431 | tig | 1 | 0 | log only | | | |
| P10423 | tig | 1 | 0 | log only | | | |
| P04968 | tig | 1 | 0 | log only | | | |
| P17579 | tig | 2 | 0 | log only | 1, 2, 3 | | - |
| P04951 | tig | 2 | 0 | log only | 1, 3, 4 | | - |
| P09151 | tig | 2 | 0 | log only | 3 | | - |
| P15042 | tig | 3 | 0 | log only | | | - |
| P02917 | tig | 2 | 0 | log only | 1, 2, 3 | | - |
| P33232 | tig | 4 | 0 | log only | 2 | | - |

A-3 Continued…

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P27300 | tig | 1 | 0 | log only | | | |
| P19494 | tig | 5 | 0 | log only | 3 | | - |
| P33137 | tig | 2 | 0 | log only | | | - |
| P00935 | tig | 1 | 0 | log only | 3 | | |
| P04384 | tig | 7 | 0 | log only | 1, 2, 3 | | - |
| P19797 | tig | 1 | 0 | log only | 4 | | |
| P02927 | tig | 3 | 0 | log only | 1, 2, 3 | | - |
| P30747 | tig | 3 | 0 | log only | 4 | | - |
| P12281 | tig | 2 | 0 | log only | 3 | | - |
| P33940 | tig | 1 | 0 | log only | | | |
| P18843 | tig | 1 | 0 | log only | 1, 3 | | |
| P33937 | tig | 1 | 0 | log only | | | |
| P24233 | tig | 12 | 0 | log only | 1, 2, 3, 4 | | - |
| P77258 | tig | 1 | 0 | log only | 2 | | |
| P17117 | tig | 1 | 0 | log only | 4 | | |
| P38489 | tig | 6 | 0 | log only | 1, 3, 4 | | - |
| P08201 | tig | 1 | 0 | log only | | | |
| P37013 | tig | 1 | 0 | log only | | | |
| P00452 | tig | 1 | 0 | log only | 3 | | |
| P23843 | tig | 9 | 0 | log only | 1, 2, 3, 4 | | - |
| P09374 | tig | 1 | 0 | log only | 2 | | |
| P08400 | tig | 1 | 0 | log only | | | |
| P07001 | tig | 1 | 0 | log only | 3 | | |
| P00582 | tig | 1 | 0 | log only | 1, 3 | | |
| P23869 | tig | 5 | 0 | log only | 1, 2, 4 | | - |
| P24555 | tig | 1 | 0 | log only | | | |
| P08179 | tig | 1 | 0 | log only | 4 | | |
| P15039 | tig | 3 | 0 | log only | 2, 4 | | - |
| P33221 | tig | 3 | 0 | log only | 3, 4 | | - |
| P08244 | tig | 1 | 0 | log only | 1, 2, 4 | | |
| P29464 | tig | 1 | 0 | log only | 2, 4 | | |
| P09170 | tig | 4 | 0 | log only | 1, 4 | | - |
| P04983 | tig | 1 | 0 | log only | | | |
| P25740 | tig | 1 | 0 | log only | | | |
| P25741 | tig | 1 | 0 | log only | | | |
| P27126 | tig | 3 | 0 | log only | | | - |
| P37744 | tig | 2 | 0 | log only | 3, 4 | | - |
| P27828 | tig | 2 | 0 | log only | 4 | | - |
| P03002 | tig | 2 | 0 | log only | 2, 3 | | - |
| P29015 | tig | 2 | 0 | log only | 1 | | - |
| P39290 | tig | 1 | 0 | log only | | | |
| P33643 | tig | 1 | 0 | log only | | | |
| P25537 | tig | 1 | 0 | log only | | | |
| P27252 | tig | 2 | 0 | log only | 1, 2, 4 | | - |
| P02384 | tig | 1 | 0 | log only | 1, 2, 3 | | |
| P02418 | tig | 38 | 0 | log only | 1, 2, 3, 4 | | - |
| P02413 | tig | 1 | 0 | log only | 3, 4 | | |
| P02420 | tig | 4 | 0 | log only | 4 | | - |

A-3 Continued…

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P02351 | tig | 1 | 0 | log only | 2, 3 | | |
| P02359 | tig | 1 | 0 | log only | 3 | | |
| P02366 | tig | 1 | 0 | log only | 4 | | |
| P53635 | tig | 1 | 0 | log only | 2 | | |
| P40874 | tig | 2 | 0 | log only | 3 | | - |
| P09158 | tig | 2 | 0 | log only | 2, 4 | | - |
| P05838 | tig | 2 | 0 | log only | 1, 2, 4 | | - |
| P07016 | tig | 2 | 0 | log only | 1, 2 | | - |
| P07459 | tig | 6 | 0 | log only | 1, 2, 3 | | - |
| Q46933 | tig | 1 | 0 | log only | | | |
| P30136 | tig | 1 | 0 | log only | 3, 4 | | |
| P77718 | tig | 1 | 0 | log only | | | |
| P00547 | tig | 1 | 0 | log only | 1, 2 | | |
| P33225 | tig | 1 | 0 | log only | | | |
| P07649 | tig | 1 | 0 | log only | | | |
| P09625 | tig | 7 | 0 | log only | 1, 2, 3, 4 | | - |
| P09550 | tig | 1 | 0 | log only | | | |
| P12758 | tig | 2 | 0 | log only | 1, 2, 4 | | - |
| P76373 | tig | 1 | 0 | log only | | | |
| P37751 | tig | 2 | 0 | log only | 4 | | - |
| P45563 | tig | 3 | 0 | log only | | | - |
| P09030 | tig | 1 | 0 | log only | 1, 2, 4 | | |
| P30177 | tig | 2 | 0 | log only | 3 | | - |
| P75876 | tig | 1 | 0 | log only | | | |
| P31216 | tig | 3 | 0 | log only | 1, 2, 3, 4 | | - |
| P52647 | tig | 2 | 0 | log only | 3 | | - |
| P77804 | tig | 1 | 0 | log only | 2, 3, 4 | | |
| P77391 | tig | 1 | 0 | log only | 3 | | |
| P52065 | tig | 1 | 0 | log only | 1, 4 | | |
| P45473 | tig | 4 | 0 | log only | 2 | | - |
| P46853 | tig | 4 | 0 | log only | 3, 4 | | - |
| P31456 | tig | 1 | 0 | log only | 4 | | |
| P31473 | tig | 1 | 0 | log only | | | |
| P27827 | tig | 6 | 0 | log only | 2, 4 | | - |
| P32130 | tig | 3 | 0 | log only | | | - |
| P75678 | tig | 1 | 0 | log only | | | |
| P75805 | tig | 3 | 0 | log only | 3, 4 | | - |
| P45770 | tig | 3 | 0 | log only | 2 | | - |
| P39323 | tig | 2 | 0 | log only | | | - |
| P14375 | tig | 1 | 0 | log only | | | |
| P24182 | tig | 0 | 1 | stationary only | 2, 3, 4 | | |
| P36683 | tig | 0 | 11 | stationary only | 1, 2, 3 | | + |
| P27550 | tig | 0 | 10 | stationary only | 3 | | + |
| P32719 | tig | 0 | 1 | stationary only | | | |
| P18840 | tig | 0 | 1 | stationary only | 2 | | |
| P08531 | tig | 0 | 1 | stationary only | | | |
| P03026 | tig | 0 | 1 | stationary only | | | |
| P77581 | tig | 0 | 3 | stationary only | | | + |

A-3 Continued…

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P00888 | tig | 0 | 1 | stationary only | | | |
| P26607 | tig | 0 | 2 | stationary only | | | + |
| P11056 | tig | 0 | 36 | stationary only | 3 | | + |
| P23892 | tig | 0 | 1 | stationary only | | | |
| P00907 | tig | 0 | 4 | stationary only | 1, 2, 3, 4 | | + |
| P08331 | tig | 0 | 1 | stationary only | 2 | | |
| P77239 | tig | 0 | 2 | stationary only | | | + |
| P33013 | tig | 0 | 1 | stationary only | | | |
| P07651 | tig | 0 | 11 | stationary only | 2, 3 | | + |
| P15723 | tig | 0 | 1 | stationary only | | | |
| P76015 | tig | 0 | 2 | stationary only | 3 | | + |
| P77381 | tig | 0 | 1 | stationary only | | | |
| Q46857 | tig | 0 | 1 | stationary only | 3 | | |
| P18775 | tig | 0 | 2 | stationary only | | | + |
| P03004 | tig | 0 | 1 | stationary only | | | |
| P10443 | tig | 0 | 4 | stationary only | | | + |
| P06710 | tig | 0 | 1 | stationary only | | | |
| P27430 | tig | 0 | 11 | stationary only | 1, 2, 3 | | + |
| P77202 | tig | 0 | 1 | stationary only | | | |
| P19636 | tig | 0 | 1 | stationary only | | | |
| P32176 | tig | 0 | 4 | stationary only | | | + |
| P06971 | tig | 0 | 1 | stationary only | 2 | | |
| P75780 | tig | 0 | 1 | stationary only | | | |
| P24186 | tig | 0 | 1 | stationary only | | | |
| P04286 | tig | 0 | 1 | stationary only | | | |
| P10121 | tig | 0 | 1 | stationary only | 4 | | |
| P25526 | tig | 0 | 4 | stationary only | 3 | | + |
| P28302 | tig | 0 | 20 | stationary only | 3, 4 | | + |
| P03024 | tig | 0 | 1 | stationary only | | | |
| P25520 | tig | 0 | 1 | stationary only | 2, 3 | | |
| P37192 | tig | 0 | 1 | stationary only | 2, 3 | | |
| P13031 | tig | 0 | 1 | stationary only | 3 | | |
| P13035 | tig | 0 | 1 | stationary only | 1, 2, 3 | | |
| P00961 | tig | 0 | 1 | stationary only | 3 | | |
| P37689 | tig | 0 | 1 | stationary only | 1, 4 | | |
| P43675 | tig | 0 | 1 | stationary only | | | |
| P17115 | tig | 0 | 1 | stationary only | | | |
| P15038 | tig | 0 | 1 | stationary only | | | |
| P09126 | tig | 0 | 1 | stationary only | | | |
| P23852 | tig | 0 | 2 | stationary only | | | + |
| P43329 | tig | 0 | 4 | stationary only | | | + |
| P08956 | tig | 0 | 1 | stationary only | | | |
| P10413 | tig | 0 | 1 | stationary only | 1, 2, 3 | | |
| P24192 | tig | 0 | 1 | stationary only | 1 | | |
| P76143 | tig | 0 | 4 | stationary only | | | + |
| P27246 | tig | 0 | 1 | stationary only | | | |
| P17109 | tig | 0 | 1 | stationary only | | | |
| P00959 | tig | 0 | 2 | stationary only | 1, 2, 3 | | + |

A-3 Continued…

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P13009 | tig | 0 | 1 | stationary only | 1, 3 | | |
| P00562 | tig | 0 | 4 | stationary only | 3 | | + |
| P30958 | tig | 0 | 2 | stationary only | 3 | | + |
| P39168 | tig | 0 | 1 | stationary only | | | |
| P77645 | tig | 0 | 1 | stationary only | 3 | | |
| P25522 | tig | 0 | 2 | stationary only | | | + |
| P17112 | tig | 0 | 1 | stationary only | | | |
| P16926 | tig | 0 | 1 | stationary only | | | |
| P09152 | tig | 0 | 2 | stationary only | | | + |
| P19319 | tig | 0 | 1 | stationary only | | | |
| P32664 | tig | 0 | 1 | stationary only | | | |
| P33602 | tig | 0 | 1 | stationary only | 2, 3 | | |
| P37095 | tig | 0 | 3 | stationary only | 2, 3, 4 | | + |
| P15288 | tig | 0 | 1 | stationary only | 2, 3 | | |
| P36938 | tig | 0 | 1 | stationary only | 2, 3 | | |
| P24231 | tig | 0 | 1 | stationary only | 2, 3 | | |
| P07003 | tig | 0 | 1 | stationary only | 1, 3 | | |
| P55798 | tig | 0 | 1 | stationary only | | | |
| P23538 | tig | 0 | 3 | stationary only | 1, 3 | | + |
| P07004 | tig | 0 | 2 | stationary only | 1, 2, 3, 4 | | + |
| P31660 | tig | 0 | 1 | stationary only | | | |
| P37177 | tig | 0 | 2 | stationary only | 3 | | + |
| P15254 | tig | 0 | 8 | stationary only | 3 | | + |
| P09546 | tig | 0 | 13 | stationary only | 2 | | + |
| P21599 | tig | 0 | 1 | stationary only | 3 | | |
| P00478 | tig | 0 | 8 | stationary only | 1, 2, 3, 4 | | + |
| P36767 | tig | 0 | 1 | stationary only | | | |
| P08394 | tig | 0 | 1 | stationary only | | | |
| P24230 | tig | 0 | 1 | stationary only | | | |
| P11585 | tig | 0 | 5 | stationary only | | | + |
| P27127 | tig | 0 | 1 | stationary only | | | |
| P16916 | tig | 0 | 1 | stationary only | | | |
| P16918 | tig | 0 | 1 | stationary only | | | |
| P30850 | tig | 0 | 1 | stationary only | 3 | | |
| P02432 | tig | 0 | 1 | stationary only | | | |
| P14081 | tig | 0 | 1 | stationary only | | | |
| P08328 | tig | 0 | 8 | stationary only | 3 | | + |
| P09157 | tig | 0 | 1 | stationary only | 1, 2, 4 | | |
| P21169 | tig | 0 | 2 | stationary only | | | + |
| P24169 | tig | 0 | 1 | stationary only | | | |
| P17580 | tig | 0 | 1 | stationary only | | | |
| Q46812 | tig | 0 | 1 | stationary only | | | |
| P00470 | tig | 0 | 1 | stationary only | | | |
| P39453 | tig | 0 | 1 | stationary only | | | |
| P13482 | tig | 0 | 7 | stationary only | 2, 3 | | + |
| P28904 | tig | 0 | 1 | stationary only | | | |

A-3 Continued…

| | | hits | | | | | |
|---|---|---|---|---|---|---|---|
| **sp_id** | **eco_gn** | **Exp.** | **Stat.** | **ratio** | **notes** | **Elution Shift** | **Stat. Δ** |
| P37196 | tig | 0 | 1 | stationary only | | | |
| P00895 | tig | 0 | 1 | stationary only | 3 | | |
| P07024 | tig | 0 | 1 | stationary only | 2, 3 | | |
| P43672 | tig | 0 | 1 | stationary only | | | |
| P77713 | tig | 0 | 1 | stationary only | | | |
| P75793 | tig | 0 | 1 | stationary only | | | |
| P75870 | tig | 0 | 1 | stationary only | | | |
| P75914 | tig | 0 | 2 | stationary only | 4 | | + |
| P77154 | tig | 0 | 2 | stationary only | | | + |
| P25906 | tig | 0 | 1 | stationary only | 2 | | |
| P52645 | tig | 0 | 1 | stationary only | 2 | | |
| P77674 | tig | 0 | 6 | stationary only | 3 | | + |
| P77432 | tig | 0 | 1 | stationary only | | | |
| P33345 | tig | 0 | 2 | stationary only | | | + |
| P33920 | tig | 0 | 1 | stationary only | 3 | | |
| P76633 | tig | 0 | 1 | stationary only | | | |
| P52054 | tig | 0 | 1 | stationary only | 2, 4 | | |
| P45766 | tig | 0 | 1 | stationary only | | | |
| P45545 | tig | 0 | 1 | stationary only | | | |
| P46837 | tig | 0 | 1 | stationary only | 3 | | |
| P31806 | tig | 0 | 1 | stationary only | | | |
| P39285 | tig | 0 | 1 | stationary only | | | |
| P39336 | tig | 0 | 1 | stationary only | | | |
| P77212 | tig | 0 | 1 | stationary only | | | |
| P77433 | tig | 0 | 1 | stationary only | 2 | | |
| P76328 | tig | 0 | 2 | stationary only | | | + |
| P42620 | tig | 0 | 5 | stationary only | | | + |
| P39321 | tig | 0 | 1 | stationary only | | | |
| P52648 | tig | 0 | 1 | stationary only | | | |

# VITA

| | |
|---|---|
| Name: | Matthew Maurice Champion |
| Address: | Applied Biosystems 353 Hatch Dr. Foster City, CA 94404 |
| Email Address: | champimm@comcast.net |
| Education: | B.S., Microbiology, The University of Iowa, 1997, Iowa City, IA<br>Ph.D., Biochemistry, Texas A&M University, 2005, College Station, TX |
| Professional Publications: | Champion, M.M., Campbell, C.S., Siegele, D.A., Russell, D.H., and Hu, J.C. (2003) Proteome analysis of Escherichia coli K-12 by two-dimensional native-state chromatography and MALDI-MS. *Mol Microbiol* **47**: 383-396.<br>Miller, M.A., McGowan, S.E., Gantt, K.R., Champion, M., Novick, S.L., Andersen, K.A., Bacchi, C.J., Yarlett, N. Britigan, B.E., and Wilson, M.E. (2000) Inducible Resistance to oxidant stress in the protozoan *Leishmania chagasi*. *J Biol Chem* **275**: 33883-33889.<br>Schopfer L.M., Champion M.M., Tamblyn N., Thompson, C.M., and Lockridge, O. (2005) Characteristic mass spectral fragments of the organophosphorus agent FP-biotin and FP-Biotinylated peptides from trypsin and bovine albumin (Tyr410). *Analytical Biochemistry* **345**: 122-132. |