PROTEOMIC ANALYSIS OF *E. COLI* USING 2D HPLC AND
MALDI-TOF MASS SPECTROMETRY

A Senior Thesis

By

CHRISTOPHER S. CAMPBELL

Submitted to the Office of Honors Programs
& Academic Scholarships
Texas A&M University

In partial fulfillment of the requirements of the

UNIVERSITY UNDERGRADUATE
RESEARCH FELLOWS

April 2002

Group:
Life Sciences 1
PROTEOMIC ANALYSIS OF *E. COLI* USING 2D HPLC AND MALDI-TOF MASS SPECTROMETRY

A Senior Thesis

By

CHRISTOPHER S. CAMPBELL

Submitted to the Office of Honors Programs
& Academic Scholarships
Texas A&M University

In partial fulfillment of the requirements of the

UNIVERSITY UNDERGRADUATE RESEARCH FELLOWS

Approved as to style and content by:

James C. Hu

Edward A. Funkhouser

April 2002

Group:

Life Sciences 1
ABSTRACT

Proteomic Analysis of *E. coli* Using 2D HPLC and MALDI-TOF Mass Spectrometry.

Christopher S. Campbell

Department of Biochemistry/Biophysics

Texas A&M University

Fellows Advisor: Dr. James C. Hu

Department of Biochemistry/Biophysics

In this post-genomic era, researchers are striving to find new ways to use the enormous amounts of data that have been collected. One obvious way is with proteomics, the large-scale identification of expressed proteins. We have developed a novel method for identifying proteins using two dimensions of non-denaturing high performance liquid chromatography (HPLC) and matrix assisted laser desorption ionization time of flight (MALDI-TOF) mass spectrometry. The first dimension of separation uses an anion exchange column and each of those fractions is run through the second dimension, a hydrophobic interaction column. The proteins were then dialyzed, denatured, and digested with trypsin before being subjected to mass spectrometry. Identifications were made based on the peptide masses. Using this method we have made 2012 protein identifications, 310 of which are unique. These numbers are comparable to other forms of proteomics such as 2-D gels.
This thesis is dedicated to Jimmy.

I'll always remember you Jimmy.
I would like to thank Dr. Hu and Matthew Champion for their guidance. In addition, I would like to thank Matt for help with the figures.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>v</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>INTRODUCTION</td>
</tr>
<tr>
<td>II</td>
<td>MATERIALS AND METHODS</td>
</tr>
<tr>
<td>III</td>
<td>RESULTS AND CONCLUSIONS</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>12</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Flow Chart</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Venn Diagram</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>Distributions of E(g) Numbers</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>Average E(g) Numbers</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>pI vs. MW Graphs</td>
<td>8</td>
</tr>
<tr>
<td>TABLE</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Functional Classifications of Proteins</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Pair-wise Interaction Candidates</td>
<td>10</td>
</tr>
</tbody>
</table>
INTRODUCTION

Proteomics is one of the most rapidly developing fields in the biological sciences. As such, a quick and efficient method for performing proteomic analyses is essential. The traditional method for proteomics involves the use of 2-D gels to visualize the proteins. 2-D gels first separate proteins based on their pI using isoelectric focusing. The second mode of separation is sodium dodecyl sulfate polyacrylamide gel electrophoresis. The spots are then excised and identified with mass spectrometry. However, there have been numerous complaints brought up against 2-D gels. Low abundance proteins are difficult to identify on 2-D gels. Many proteins have similar pIs, making separations difficult. 2-D gels also have an apparent bias towards proteins in the lower pI ranges. An alternate method that has been gaining popularity is the use of various forms of liquid chromatography for separating the proteins. Often affinity or reverse phase chromatography is used. These methods also use mass spectrometry to make identifications. There are problems with these methods as well. They are very costly and labor intensive. They also have an excessive false positive rate upwards of 30 percent. My thesis research involved developing an alternate method. Our method uses two forms of non-denaturing liquid chromatography in series; anion exchange and hydrophobic interaction. This separates the proteins into 380 fractions each containing 0-7 identifiable proteins. The fractions are each dialyzed, denatured and digested with trypsin before being subjected to analysis by MALDI-TOF mass spectrometry. Identifications are made using Protein Prospector.
MS-Fit software. Our method can be done with very low cost and only one or two people. We did the complete experiment four times.

Proteomics involves more than just identifications. Ideally, we would like to know what the proteins are interacting with. Since the mode of separation is non-denaturing, protein activity and complexes are preserved. We have proven β-galactosidase activity is maintained through both chromatography steps. Many, if not most proteins are believed to exist in complexes. Experiments such as those done by Ho et al. and Eisenberg et al. have been able to provide some evidence for the existence of specific complexes. However, multiple types of experiments are needed to get the whole picture. We hope to be able to identify protein complex candidates based on co-fractionation. The entire proteomic analysis has been done with 2 different pHs at which the anion exchange is run. Changing the pH changes which fractions the proteins elute into. By observing which proteins continue to co-fractionate after the shift, we have been be able to accumulate more circumstantial evidence for the existence of complexes.
MATERIALS AND METHODS

One liter of *E. coli* (MG1655) cells were grown in minimal glucose (M9) media. Cells were pelleted at 4000g for 20 minutes and resuspended in 200ml of 20mM Tris-HCl, 20mM NaCl, 1mM EDTA, pH 8.75. They were then centrifuged again and resuspended in 6ml of the same buffer. The cells were lysed via French press and half of the lysate was loaded onto a 1ml Waters column packed with SOURCE 15 Q resin. A gradient from 20mM to 1M NaCl was used. The pH set for the run was either 7.5 or 8.75. Five ml fractions were collected and each of them was loaded onto a 1ml Waters column packed with SOURCE 15 Phe resin. The gradient used went from 1.5M to 0M ammonium sulfate. Each 500μl fraction was collected directly into a Slide-A-Lyzer MINI Dialysis unit (Pierce 3,500 MWCO). The samples were dialyzed for 24 hours in 25mM ammonium bicarbonate. They were then denatured with heat (95°F for 20') and digested with 1μg of modified trypsin (Promega) each for 5 hours. The MALDI was done in a similar fashion to that previously described by Park et al.. Identifications from the peptide mass data were made with Protein Prospector MS-Fit software (prospector.ucsf.edu). Factors looked at for making the identifications include MOWSE score, sequence coverage, number of peptides matched, and trends in peptide error. All of the identified proteins were checked to make sure that they are present in the genomic DNA sequence of *E. coli* strain MG1655.
Figure 1. Flow chart describing the path of the proteins for identification by MALDI and for 2D gel analysis.
RESULTS AND CONCLUSIONS

A total of 2012 identifications have been made which include 310 different proteins identified. This is slightly more than the 271 different proteins identified by the Swiss 2-D project.\textsuperscript{10,11} Figure 2 shows the overlap between the two projects.

![Figure 2. Venn Diagram showing the number of proteins found in only our study, only the SWISS 2D-PAGE project, and those found in both.]

One thing that we wanted to determine was whether or not we had any biases in our identifications towards things with high abundance or high/low pl or molecular weight. To measure the abundance of the proteins we identified, we used E(g) numbers. E(g) numbers predict protein abundance based on the codon usage of their genes.\textsuperscript{12} The higher the number, the more abundant the protein is predicted to be. On average, our E(g) numbers were significantly higher than those of the entire predicted proteome of \textit{E. coli}, indicating that we do have a bias towards proteins of higher abundance. Figure 3 shows a comparison between the percentage of proteins in different E(g) ranges for our data and the entire proteome.
Our Data

Figure 3. Distribution of proteins for a range of E(g) numbers for A) the proteins that we have identified and B) the predicted proteome.

However, our numbers were still lower than those of other proteome projects. Figure 4 compares our E(g) numbers to those of other projects.
To check for pI and molecular weight biases, we simply compared the theoretical pIs and molecular weights for the proteins we identified and the entire genome. These were calculated using the prediction tools found on the SWISS-PROT website (www.expasy.org). There did not appear to be any significant difference between our pI and molecular weight distribution and that of the proteome. Figure 5 shows graphs of pI vs. molecular weight for our data as well as the proteome.

To identify candidates for protein complexes, we found all pairs of proteins that were found in the same fraction for both pH 7.5 and 8.75. Using this method, 125 candidate interactions were identified (Table 1). In addition, some known complexes were found to co-fractionate. For example, phenylalanine tRNA synthetase α and β subunits were found together. RNA polymerase subunits α, β, and β' were also seen in the same fractions. We believe that this data in conjunction with other experiments

**Figure 4.** Comparison of average E(g) number for a variety of protein sets, including the *E. coli* genome and three proteomics projects.
similar to those done by Ho et al. and Eisenberg et al. could result in fairly confident identification of novel complexes.

Figure 5. Graphs of pI versus molecular weight for A) the predicted proteome and B) the proteins that we identified.
### Table 1

Proteins that cofractionate at both pH7.5 and pH8.75. The 125 pairs are shown as 250 entries; each pair is listed with each partner first to aid finding proteins of interest.
To find out what kinds of proteins we identified, we looked at their functional classifications. Table 2 compares the percentages of functional classifications for our data as well as the entire genome. The lack of membrane proteins was expected because the membranes are pelleted after the cells are lysed. We identified a higher proportion of protein biosynthesis and nucleotide metabolism proteins most likely because of their high abundance.

<table>
<thead>
<tr>
<th>Category</th>
<th>% of Total</th>
<th>% of Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Biosynthesis/ Chaperonin</td>
<td>19%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Glycolysis TCA Carbon Utilization</td>
<td>15%</td>
<td>13.0</td>
</tr>
<tr>
<td>NT Metabolism</td>
<td>11%</td>
<td>1.4%</td>
</tr>
<tr>
<td>AA Synthesis</td>
<td>10%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Enzymatic Activities</td>
<td>10%</td>
<td>11.9</td>
</tr>
<tr>
<td>Biosynthetic Genes FA, DAP, LPS Cofactors</td>
<td>10%</td>
<td>8.7%</td>
</tr>
<tr>
<td>Transcription</td>
<td>4%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Replication</td>
<td>1%</td>
<td>2.7%</td>
</tr>
<tr>
<td>Membrane, Xport</td>
<td>0%</td>
<td>10.3</td>
</tr>
<tr>
<td>Hypothetical/ Unknown/ Putative</td>
<td>19%</td>
<td>43.2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

Table 2. Percentage of proteins in each functional classification for the proteins identified in this study and the *E. coli* genome. Classification come from Opitcek et al.11

To help confirm our identifications, 2D PAGE was performed on all of the first dimension fractions run at pH 7.5. Our identifications for each fraction were compared to the spots found on the gels. Both of the predicted molecular weight and pI as well as known migration patterns for previously identified proteins were looked at. The proteins identified by us had a much higher chance of correlating with a spot on these gels than proteins selected at random. Spots were assigned to 109 of the 219
proteins we identified at pH 7.5. Forty-one of these have not been identified by the SWISS-2D project.

At least one false positive was confirmed. UDP-glucose dehydrogenase from the K5 strain of *E. coli* was identified. To make sure that there was not something wrong with either our strain of *E. coli* or the MG1655 UDP-glucose dehydrogenase sequence, we sequenced the gene. The results showed that our strain does indeed have the MG1655 version of UDP-glucose dehydrogenase, and not that of K5. All other identified proteins came from MG1655. However, some IDs may still be false positives.

Future work on this project will likely involve proving that the system can be used to identify differences in protein expression under altered conditions. Repeating the experiment with cells that have been infected with lambda phage and looking for differences from the previous experiments is one possible way of doing this. Another would be to look at protein expression in outgrowth after starvation.
REFERENCES


