AN ABSTRACT OF THE DISSERTATION OF

Martha Degen Stapels for the degree of Doctor of Philosophy in Chemistry

presented on October 1, 2003.

Title: Proteomic Approach to the Analysis of DNA-Binding Proteins Using Mass

Spectrometry

Abstract Approved:

Redacted for privacy

Douglas F. Barofsky

In proteomic studies, separate experimental protocols have been necessary

to identify proteins, determine their function, and predict their three-dimensional

structure. In this study, a function-based separation of proteins was conceived to

fractionate proteins prior to enzymatic digestion. In the initial demonstration of

this technique, a DNA substrate was used to separate the DNA-binding proteins

from the rest of the proteins in a lysate in order to identify protein function and to

simplify the complex mixture of proteins. A total of 232 putative DNA-binding

proteins and over 540 proteins in all were identified from *E. coli*. Hypothetical or

unknown proteins were found, some of which bind to DNA. As a part of this

demonstration, changes in protein expression caused by different environmental

conditions (aerobic and anaerobic atmospheres) were observed. In a second

demonstration, aimed at determining the three-dimensional structure of the DNA-

binding proteins, binding sites were blocked with oligonucleotides, and the modified proteins were purified, enzymatically digested, and subjected to tandem mass spectrometry. The amino acids in the DNA-binding domains of three proteins were determined.

In a final application of function-based separation, DNA-binding proteins were digested with trypsin and the resulting peptides were separated using HPLC and subsequently analyzed using MALDI TOF/TOF and ESI Q-TOF instruments to study the complementary nature of the two ionization techniques, taking into account the differences between the mass analyzers. Based on the analysis of a large data set containing hundreds of peptides and thousands of individual amino acids, some of the currently held notions regarding the ionization processes were confirmed. ESI tends to favor the analysis of hydrophobic amino acids and peptides while MALDI is disposed toward mainly basic and aromatic species. These tendencies in ionization account in large part for the complementary nature of the peptides and proteins identified by the ESI and MALDI instruments and make it necessary to employ both types of instruments to gain the most information out of a given sample in a proteomics study.

Proteomic Approach to the Analysis of DNA-Binding Proteins

Using Mass Spectrometry

by

Martha Degen Stapels

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of

the requirements for the

degree of

Doctor of Philosophy

Presented October 1, 2003

Commencement June 2004

Doctor of Philosophy dissertation of <u>Martha Degen Stapels</u>
presented on <u>October 1, 2003</u>.

APPROVED:

Redacted for privacy

Major Professor, Representing Chemistry

Redacted for privacy

Chair of the Department of Chemistry

Redacted for privacy

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of
Oregon State University libraries. My signature below authorizes release of my
dissertation to any reader upon request.

Redacted for privacy

Martha Degen Stapels, Author

ACKNOWLEDGMENTS

I would first like to express my deepest gratitude to my research advisor, Professor Douglas Barofsky. His guidance, support, faith, and friendship have been essential to my success in graduate school. His mentoring has helped me to become an independent scientist, and for that I am most grateful.

I would also like to sincerely thank Brian Arbogast and Lilo Barofsky for their wisdom on all things mass spectrometry and for their friendship and support.

Thanks to Dr. Catalin Doneanu and the rest of the mass spectrometry group for teaching me invaluable biochemistry and mass spectrometry skills. Thanks to everyone in the mass spectrometry group for being part of the crazy family that we have formed down in the basement of ALS.

Many thanks to those individuals who served on my academic committee: Dr. Michael Schuyler, Dr. James Ingle, Dr. Kevin Gable, and Dr. George Pearson.

I would like to thank my family for their love and support.

Finally, I wish to acknowledge my husband Christopher Stapels for his humor, patience, and love that kept me going. Without him, this work would not have been possible.

CONTRIBUTION OF AUTHORS

The only authors of the manuscripts presented in this thesis are myself and my research advisor, Dr. Douglas Barofsky. The following individuals, although not co-authors, made technical contributions to the research presented in this thesis. Brian Arbogast performed the maintenance and set-up on all of the electrospray mass spectrometers. Elisabeth Lilo Barofsky performed MALDI mass spectrometry on many of the samples. Andrew Bauman and John Wilson assisted with the *E. coli* cell culture and Mac Wisdom performed some of the SDS-PAGE. Christie Hunter and Ewa Witkow at Applied Biosystems performed multidimensional chromatography experiments mentioned in Chapter 2. The Stratalinker used in Chapter 3 was courtesy of Dr. Dale Mosbaugh, Department of Environmental and Molecular Toxicology, Oregon State University. Dr. Nicholas Ambulos of the Department of Microbiology and Immunology at the University of Maryland at Baltimore generously supplied the oligonucleotides used in Chapter 3. Jeffrey Finch, Daniel Wall, and Michael Daly from Waters Corporation assisted in the installation and operation of the MALDIprep device described in Chapter 4.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

PROTEOMIC APPROACH TO THE ANALYSIS OF DNA-BINDING PROTEINS USING MASS SPECTROMETRY

1. INTRODUCTION

1.1. Proteomics

Coined after genomics, proteomics is the study of the proteome, the protein complement to the genome [1]. The recent sequencing of genomes, development of mass spectrometers capable of ionizing large biomolecules, and creation of browser-based bioinformatics tools have led to the current popularity of proteomic studies [2]. The first proteomic studies that involved mass spectrometers used gel electrophoresis to isolate proteins, but these studies were often time-consuming and led to the identification of only the most abundant proteins [3]. The advancement and application of mass spectrometers capable of tandem mass spectrometry (MS/MS) on biopolymers allowed for a new type of experiment, shotgun or mining proteomics [4]. In this type of analysis, MS/MS is performed on as many peptides as possible in a digested mixture of proteins and the spectra are then matched to proteins in a database. This approach has led to the identification of 1,910 proteins from the bacterium *Deinococcus radiodurans* [5] and over 1,500 proteins from yeast [4]. Besides mining experiments, proteomic studies are also used to compare proteins expressed under different conditions or

disease states, determine interactions between proteins, and characterize any posttranslational modifications present in a proteome [6].

The tools that are necessary for a mass spectrometric-based proteomic study include a database containing the proteins or genes from the organism to be studied, a technique to separate proteins or peptides, a mass spectrometer capable of ionizing proteins and peptides, and data analysis software to match the mass spectral data to the proteins in the database. If a database is not available for the organism being studied, *de novo* sequencing or searching based on homology to other organisms' proteins can be used to identify proteins. In order to quantify proteins observed with mass spectrometry, several different methods using isotopic labeling have been developed [7,8]. The most popular of these methods is the isotope-coded affinity tag (ICAT) approach, which involves labeling cysteines of proteins expressed under different conditions with heavy or light isotopes. The protein mixtures are combined and enzymatically digested; the resulting peptides are separated by high performance liquid chromatography (HPLC) and analyzed by MS/MS. The areas under the chromatographic peaks are used to determine the expression levels [7].

Proteomic studies have unique difficulties that are not present in the sequencing of a static genome. One of the main challenges is the temporal nature of the proteome, which reflects changes due to environmental conditions, presence of drugs, or simply over time. A tadpole shares the same genome as a mature frog, but their proteome is obviously different. The tendency of proteins to have splice

variants and posttranslational modifications also makes their analysis more challenging. All of these possibilities must be taken into account in a proteomic study. At present, mass spectrometry appears to be the only tool capable of handling all of these challenges.

1.2. Biological Mass Spectrometry

The 2002 Nobel prize in chemistry emphasized the profound effect the development of ionization techniques that make it possible to analyze large biomolecules by mass spectrometry has made on the study of proteins. The two ionization techniques used for proteins and peptides are electrospray ionization (ESI) [9] and matrix-assisted laser desorption ionization (MALDI) [10]. Since it ionizes analytes out of solution, ESI has been coupled with LC separations, oftentimes to analyze complex mixtures of peptides or proteins. As shown in Figure 1.1, a high voltage (1-5 kV) applied to a needle (usually coupled to a chromatographic column) results in a fine mist of charged analytes - the electrospray process. These droplets desolvate and undergo Coulomb explosions repeatedly until the charged analytes enter the mass spectrometer. Analytes can be multiply charged by ESI depending on their molecular weight, typically with one positive charge per 1,000 Da observed [11].
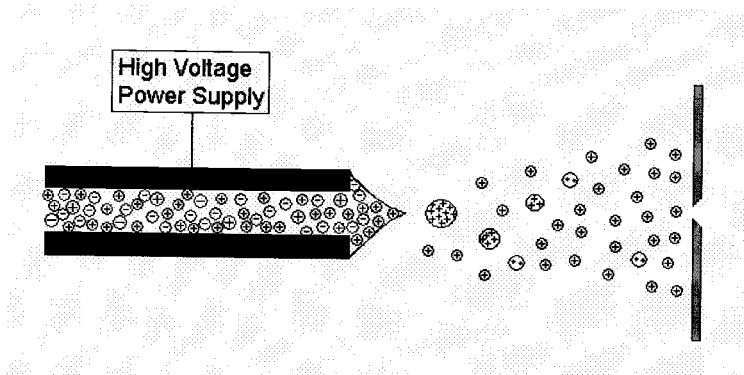
Figure 1.1. Electrospray Ionization. The end of the needle is usually held at positive potential with respect to the entrance to the mass spectrometer.

Unlike ESI, which analyzes molecules directly out of the liquid phase, MALDI involves mixing analytes in solution with matrix molecules and then depositing this mixture on a surface to produce a dry, crystalline mixture. The matrix molecules readily absorb ultraviolet light. When the crystal is exposed to pulses of photons (shots) from a UV laser, the matrix molecules become excited and assist in the transfer of the analyte molecules into the gas phase. An illustration of this process can be seen in Figure 1.2. MALDI usually produces ions that are singly charged, making the resulting spectra easy to interpret.
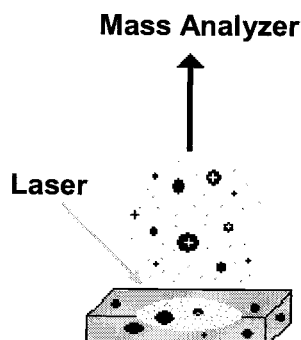
Figure 1.2. MALDI Ionization. The laser is used to ablate matrix and analyte molecules from the solid phase into the gas phase.

Until a short while ago, MALDI had been used almost exclusively for the direct analysis of mixtures of compounds, such as the peptide mixture resulting from the digest of a single protein. This has been so because the coupling of MALDI to liquid-phase separations has lagged behind that of ESI. Recently, commercial LC-MALDI devices have been developed that allow for the coupling of HPLC to MALDI instruments, although in an off-line manner [12]. Figure 1.3 shows the design of the MALDIprep sample collection module from Waters (Milford, MA). In the operation of this instrument, the eluate from an HPLC column is mixed with a matrix solution appropriate for the MALDI analysis. This mixture is then introduced into a capillary heated to 45-65°C, which, along with heated nitrogen gas, desolvates the analyte and matrix molecules as they are sprayed onto a MALDI sample plate. This instrument, and those like it, has made it possible to easily separate peptides prior to MALDI mass spectral analysis.
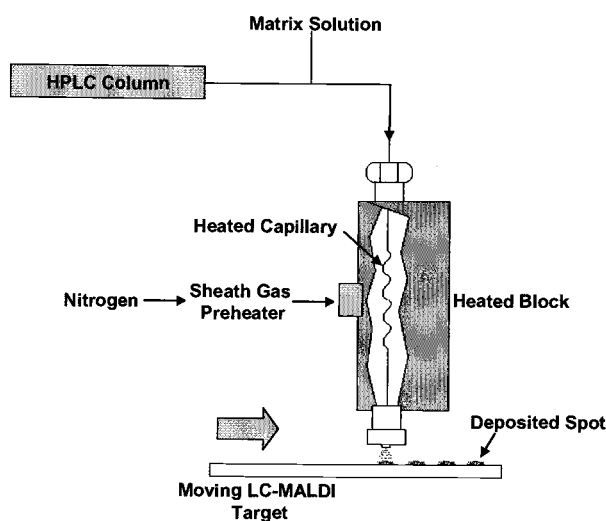
Figure 1.3. Design of the LC-MALDIprep device.

1.3. Photochemical Crosslinking

The photoreactivity of nucleic acids and their crosslinking to proteins were first shown in the 1960's [13], with the thymidine base subsequently being proven as the most photoreactive [14]. The photocrosslinking of proteins to DNA was first combined with mass spectrometry in 1994 in a study that determined the DNA-binding site of uracil DNA glycosylase (Ung) [15] prior to the publication of the crystal structure of Ung interacting with DNA. This study and those that followed it [16,17], all employed oligonucleotides made entirely of thymidine bases (dT_{20} and dT_{30}) in order to maximize the photoreactivity without perturbing the natural interactions between proteins and DNA. Unfortunately, the yield of crosslinking proteins to an oligonucleotide is less than 10%. Other studies using photochemical crosslinking and mass spectrometry have used photoreactive substrates, such as iodouracil [18], to increase the yield of crosslinked products, but it is not proven that these substrates mimic the natural systems accurately.

1.4. Tools for Data Analysis

The sequencing of genomes has led to a large number of DNA and protein databases that are publicly available. Peptide mass fingerprinting [19], where

peptide masses are generated from the mass spectral analysis of tryptic digestions of individual proteins, was the first application to use automated software to search the protein databases using masses of peptides. One such software, MOWSE (MOlecular Weight SEarch) computed the mass values directly from the databases taking the peptide sizes into account to match the experimental spectrum to the masses of a theoretical digest of a protein in the database [20]. Peptide mass fingerprinting is not useful for analyzing a mixture of digested proteins, although it is useful for determining the identity of proteins separated by gel electrophoresis.

Once mass spectrometers that could perform tandem mass spectrometry (MS/MS) became available for proteomics, new automated software was needed to search these spectra. One of the first commercially available software programs for MS/MS searching was Sequest [21]. This program uses an algorithm that creates a list of candidate peptides that have the same mass as the observed mass on which MS/MS was carried out. The program then calculates the masses of the fragment ions expected for each of the candidate peptides and compares the fragments matched to find the best match. Sequest quickly became the standard searching software tool for proteomic studies; unfortunately, it has a tendency to yield many false positive results. Other programs have been created to sort Sequest results so that there is less dependence on an analyst taking time to determine if a given result is correct [22,23]. Sequest is rather inflexible in terms of the ions it searches for (it cannot search for internal fragments, for example) and the number of modifications that can be present in one search. Unfortunately, this

software is not frequently updated, making it difficult to use with the newest instruments and applications.

Mascot is the other popular automated software tool that can be used to search MS/MS data [24]. It is based on the MOWSE algorithm, which uses probabilistic scoring to determine the significance of matches. The MS/MS fragment ions, which are selected by the user based on which instrument was used to collect the data, are used as a qualifier to the peptide mass value to determine the probability of a match being significant. Mascot is extremely flexible in terms of the number of modifications and choice of fragmentation patterns, and it is updated twice per year to ensure that it is capable of searching data from the latest mass spectrometers. There are auxiliary programs available for Mascot, such as the Mascot Daemon that automates searches and the Mascot Parser, which helps to sort the results. Unfortunately, both Mascot and Sequest try to match peptides to one protein even in mixtures that contain thousands of proteins. It would be preferable for the programs to simply list the best peptide match for each individual spectrum and then group any peptides from the same protein together. In the current programs, peptides matched with low scores to a certain protein will be considered significant, even if the individual peptides match well to other proteins. The results from both Sequest and Mascot must be visually inspected and manually sorted to ensure that there are minimal false positives.

Other MS/MS searching software programs have become available recently. ProbID [25], Sonar [26], and Salsa [27] are all examples of recent additions for the

analysis of MS/MS data. Salsa is especially useful for searching for unknown modifications to a known peptide or protein. This program uses a unique algorithm that searches for the difference between the ions that a known peptide would create, rather than searching for absolute mass values. Unfortunately, this software is currently only available for data acquired on instruments from ThermoFinnegan (San Jose, CA).

1.5. Research Objectives

All of the proteins in a proteome cannot be identified unless more than one dimension of separation is used prior to mass spectral analysis [4]. The first objective of this research was to create a method capable of fractionating proteins from a whole cell lysate and determine if this method could be used for observing changes in protein expression induced by culturing in different environmental conditions (Chapter 2). The second objective was to photochemically crosslink the mixtures of DNA-binding proteins to oligonucleotide substrates to determine the amino acids in the protein binding sites - this being a logical extension of the circumstance that the first objective was realized with the separation of proteins based on DNA-binding ability. The amino acids found to be crosslinked could then be compared to any published structural data for those proteins to demonstrate the accuracy of mass spectrometry in the investigation of protein-

DNA interactions (Chapter 3). Finally, the third objective of this thesis was to compare ESI and MALDI in order to determine the source of complementarity in these two techniques, given that both HPLC ESI and HPLC MALDI are now commercially available for proteomic studies (Chapter 4).

1.6. References

1. Pandey, A. and Mann, M. (2000) "Proteomics to Study Genes and Genomes," *Nature*, 405:837-846.

2. Liebler, D.C. (2002) "Introduction to Proteomics," *Humana Press*, Totowa, NJ, 1-12.

3. Tyers, M. and Mann, M. (2003) "From Genomics to Proteomics," *Nature*, 422:193-197.

4. Wolters, D.A.; Washburn, M.P.; Yates, J.R., III. (2001) "An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics," *Anal Chem*, 73: 5683-5690.

5. Lipton, M.S., Pasa-Tolic, L., Anderson, G.A., Anderson, D.J., Auberry, D.L., Battista, J.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., Masselon, C., Markillie, L.M., Moore, R.J., Romine, M.F., Shen, Y., Stritmatter, E., Tolic, N., Udseth, H.R., Venkateswaran, A., Wong, K-K., Zhao, R., and Smith, R.D. (2002) "Global Analysis of the *Deinococcus radiodurans* Proteome by Using Accurate Mass Tags," *Proc Natl Acad Sci USA*, 99:11049-11054.

6. Aebersold, R. and Mann, M. (2003) "Mass Spectrometry-Based Proteomics," *Nature*, 422:198-207.

7. 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.

8. Zhou, H., Ranish, J.A., Watts, J.D., Aebersold, R. (2002) "Quantitative Proteome Analysis by Solid-Phase Isotope Tagging and Mass Spectrometry," *Nature Biotechnol*, 20:512-515.

9. Yamashita, M. and Fenn, J.B. (1984) "Electrospray Ion Source. Another Variation on the Free-Jet Theme," *J Phys Chem*, 88:4451-4459.

10. Karas, M., Bachmann, D., Bahr, U., and Hillenkamp, F. (1987) "Matrix-Assisted Ultraviolet Laser Desorption Ionization of Non-Volatile Compounds," *Int J Mass Spectrom Ion Proc*, 78:53-68.

11. Chait, B.T. and Kent, B.H. (1992) "Weighing Naked Proteins: Practical, High-Accuracy Mass Measurement of Peptides and Proteins," *Science*, 257:1885-1894.

12. Wall, D.B., Berger, S.J., Finch, J.W., Cohen, S.A., Richardson, K., Chapman, R., Drabble, D., Brown, J., and Gostick, D. (2002) "Continuous Sample Desorption from Reversed-Phase Liquid Chromatography to Tracks on a Matrix-Assisted Laser Desorption/Ionization Precoated Target for the Analysis of Protein Digests," *Electrophoresis*, 23:3193-3204.

13. Alexander, P. and Morison, H. (1962) "Crosslinking of Deoxyribonucleic Acid to Protein Following Ultraviolet Irradiation of Different Cells," *Nature*, 194:882-883.

14. Shetlar, M.D. (1980) "Cross-linking of Proteins to Nucleic Acids by Ultraviolet Light," *Photochem Photobiol Rev*, 5:105-197.

15. Bennett, S.E., Jensen, O.N., Barofsky, D.F., and Mosbaugh, D.W. (1994) "UV-Catalyzed Cross-linking of *Escherichia coli* Uracil-DNA Glycosylase to DNA," *J Biol Chem*, 269:21870-21879.

16. Gafken, P.R. (2000) "Characterization of UV-Crosslinked Protein-Nucleic Acid Interfaces by MALDI-MS and ESI-MS/MS," *PhD thesis, Oregon State University*.

17. Doneanu, C.E. (2002) "Mass Spectrometric Analysis of UV-Crosslinked Protein-Nucleic Acid Complexes," *PhD thesis, Oregon State University*.

18. Steen, H., Petersen, J., Mann, M, and Jensen, O.N. (2001) "Mass Spectrometric Analysis of a UV-Crosslinked Protein-DNA Complex: Tryptophans 54 and 88 of *E. coli* SSB Cross-link to DNA," *Protein Science*, 10:1989-2001.

19. Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., and Watanabe, C. (1993) "Identifying Proteins from Two-Dimensional Gels by Molecular Mass Searching of Peptide Fragments in Protein Sequence Databases," *Proc Natl Acad Sci USA*, 90:5011-5015.

20. Pappin, D.J.C., Hojrup, P., and Bleasby, A.J. (1993) "Rapid Identification of Proteins by Peptide-Mass Fingerprinting," *Curr Biol*, 3:327-332.

21. Eng, J.K., McCormack, A.L., and Yates, J.R., III. (1994) "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database," *J Am Soc Mass Spec*, 5: 976-989.

22. 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.

23. MacCoss, M.J., Wu, C.C., and Yates, J.R., III. (2002) "Probability-Based Validation of Protein Identifications Using a Modified SEQUEST Algorithm," *Anal Chem*, 74:5593-5599.

24. Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999) "Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data," *Electrophoresis*, 20:3551-3567.

25. Zhang, N., Aebersold, R., and Schwikowski, B. (2002) "ProbID: A Probabilistic Algorithm to Identify Peptides Through Sequence Database Searching Using Tandem Mass Spectral Data," *Proteomics*, 2:1406-1412.

26. Field, H.I., Fenyo, D., and Beavis, R.C. (2002) "RADARS, a Bioinformatics Solution that Automates Proteome Mass Spectral Analysis, Optimizes Protein Identification, and Archives Data in a Relational Database," *Proteomics*, 2:36-47.

27. Liebler, D.C., Hansen, B.T., Davey, S.W., Tiscareno, L, and Mason, D.E. (2002) "Peptide Sequence Motif Analysis of Tandem MS Data with the SALSA Algorithm," *Anal Chem*, 74:203-210.

2. A PROTEOMIC STUDY OF *E. COLI* USING THE DNA-BINDING CAPABILITY OF PROTEINS

Martha D. Stapels and Douglas F. Barofsky

2.1. Introduction

Since the completion of the human genome project, the focus of biochemical research has shifted towards proteomics. A proteome is the protein complement to a genome, but due to posttranslational modifications and splice variants inherent in the expression of proteins, the "sequencing" of the human proteome will not be as straightforward as was the genome. Proteins exhibit many dimensions of information beyond the sequence of amino acids that make up their primary structure, so a proteomic study often includes the identification and quantification of all of the proteins in a cell, localization of the proteins within the cell, and determination of any post-translational modifications present on specific proteins [1,2]. In order to identify all of the proteins in a given cell, tissue, or organism, more than one dimension of separation has proven to be necessary. Two-dimensional gel electrophoresis has made it possible to visualize more than a thousand proteins from *E. coli* [3], and therefore most classical proteomic studies involve the visualization, comparison, and quantitation of protein spots on gels for cells grown under different conditions. Unfortunately, in-gel proteolytic digestions followed by mass spectrometry typically lead to the identification of only 300 or so of the most-abundant proteins [4], so other techniques must be employed in order to identify weakly-expressed proteins.

Multidimensional chromatography techniques have been used to identify over 1,500 proteins from the digestion of yeast cell lysate [5], but the

overwhelming complexity of whole cell lysate often limits the resolution and subsequent identification of proteins. Given that a cell typically contains thousands of proteins, a tryptic digest of all of the proteins in a cell might yield a million or more peptides with a wide range of concentrations; it would be extremely difficult to separate the peptides in such a mixture even with multidimensional chromatography. Separation of proteins prior to enzymatic or chemical digestion has been suggested and recently demonstrated as an effective means for reducing sample complexity. Anion exchange chromatography, for example, has been used to prefractionate proteins from *E. coli* and, when coupled with reversed-phase separation of proteins followed by enzymatic digestion, has led to the identification of 310 of its proteins using MALDI-MS [6]. Ion exchange chromatography, however, does not yield any information about the functions of the proteins identified.

The separation of proteins prior to digestion based on the protein function or activity has also been performed. Reagents that selectively bind to the active sites of tyrosine phosphatases [7], serine hydrolases [8] and cysteine proteases [9] have been used to separate proteins based on their activity. A protein's ability to bind to heparin was demonstrated as a means to prefractionate proteins in another investigation [10]. One company, Serenex, is using the separation of proteins by function with proprietary media in order to fractionate the human proteome and to assist with the high-throughput screening of chemical libraries against protein targets. Due to the proprietary nature of the Serenex ligands, it is not known to the

present authors if there are any specific to DNA-binding proteins [11]. Protein chips have been used in studies of DNA-binding proteins, but typically the chips used in these studies involve very specific sequences of DNA [12,13]. The premise behind the present study is that a separation of proteins based on their ability to bind to any sequence on a large DNA substrate might provide a useful prefractionation of proteins.

In this report, we describe the use of single-stranded calf thymus DNA (50 kb) as a large generic substrate for separating all of the DNA-binding proteins from the rest of the proteins in *E. coli*. We demonstrate that the proteins that bind to the DNA, as well as those that do not, can be digested using trypsin and then analyzed with LC-MS/MS. We evaluate the predicted expression values for the proteins found in the whole cell lysate versus those found using prefractionation. We compare the DNA-binding proteins expressed in *E. coli* grown under aerobic and anaerobic conditions. Finally, we relate our results with those obtained using multidimensional chromatography.

2.2. Experimental Section

2.2.1. Materials

Acetic acid (AA), trifluoroacetic acid (TFA), bromophenol blue, copper (II) chloride, ethylenediaminetetraacetic acid (EDTA), iodoacetamide, sodium

chloride, single-stranded calf thymus DNA cellulose, lysozyme, yeast extract, glycerol, urea, Tris-HCl, and Trizma (Tris-base) were purchased from Sigma Chemical Co. (St Louis, MO). Acrylamide (>99% pure), ammonium persulfate, 2-mercaptoethanol, bis N,N`- methylenebis acrylamide, and N,N,N`N`-tetramethylethylenediamine (TEMED) were from Bio-Rad (Hercules, CA). Tris(2-carboxyethyl)phosphine hydrochloride (TCEP·HCl) was purchased from Pierce Biotechnology (Rockford, IL). Ammonium sulfate, glycine, and sodium dodecyl sulfate (SDS) were from Life Technologies Inc. (Grand Island, NY). Sequencing-grade trypsin was purchased from Promega (Madison, WI). HPLC grade acetonitrile was supplied by Fisher Scientific (Pittsburgh, PA). Water was generated with a Milli-Q Ultrapure water purification system, and Ultrafree-0.5 centrifugal filter devices (NMWL 5,000) were purchased from Millipore Corp. (Bedford, MA). The TriChromRanger protein molecular weight marker mixture was a gift from Fisher Scientific (Hampton, NH).

2.2.2. Growth and Lysis of *E. coli*

Strain BL21 (*E. coli* B F⁻ *dcm ompT hsdS*(r$_8^-$ m$_8$)*gal*) from Stratagene (La Jolla, CA) was grown in LB media (1% bactone, 0.5% yeast extract, and 1% NaCl) at 37°C. *E. coli* was grown under aerobic and anaerobic conditions. The anaerobic environment was an atmosphere of 90% nitrogen, 5% hydrogen, and 5% carbon dioxide contained in a Bactron/Anaerobic Environmental Chamber (Sheldon Manufacturing, Cornelius, OR). Cells were frozen in liquid nitrogen,

resuspended in a buffer comprising 100 mM Tris (pH 7) and 1 mM EDTA, lysed ultrasonically after the addition of lysozyme, and centrifuged to remove insoluble material. Lysates were stored at -20°C until further use.

2.2.3. Separation of DNA-binding Proteins

Cell lysate (0.5 mL) was mixed with about 0.25 mL of single-stranded DNA cellulose at room temperature for 10 minutes. The mixture was centrifuged and the liquid was pipetted off in order to collect those proteins that did not bind to the DNA. The cellulose was then rinsed with 0.5 mL of DAB buffer (20 mM Tris pH 7.4 with 1 mM EDTA) to ensure that nonbinding proteins were not still associated with the DNA cellulose. The proteins that bound to the DNA were eluted off with increasing concentrations of sodium chloride (0.05 M, 0.2 M, 0.4 M, 0.6 M, 1.0 M, and 3.0 M). Each elution step consisted of mixing the cellulose with 0.5 mL of the salt solution, decanting, collecting the solution, repeating the first three steps two more times, and pooling the extracts. Each of the six fractions was concentrated and desalted using Ultrafree-0.5 centrifugal filter devices with a molecular weight cut-off of 5,000 Da. The desalting step (facilitated by adding 0.5 mL of fresh DAB buffer to the filter) was repeated three times to ensure that the sodium chloride was removed from each of the samples prior to tryptic digestion.

2.2.4. Slab Gel Electrophoresis

Small scale denaturing polyacrylamide gel electrophoresis was conducted using the Mini-Protean III mini-gel system (Bio-Rad) with 1 mm built-in spacers. Stacking gels (2 cm) comprised 3% acrylamide, 0.1% N,N'-methylenebis acrylamide, 125 mM Tris-HCl (pH 6.8), and 0.1% SDS. Resolving gels were made with acrylamide in 0.4% N,N'-methylenebis acrylamide, 375 mM Tris-HCl (pH 8.8), 0.1% SDS. Samples were mixed with 5× cracking buffer (0.125 M Tris-HCl (pH 6.8), 360 mM 2-mercaptoethanol, 25% (w/v) glycerol, 2.5% SDS, and 0.1% bromophenol blue), heated at 100°C for 15 min, and loaded onto the gel. Electrophoresis was conducted with 25 mM Tris-base, 200 mM glycine, and 0.1% SDS as the running buffer, at 200 V and room temperature until the tracking dye had migrated 0.5 cm from the bottom of the gel. Gels were then stained with 250 mM copper chloride for 10 min, rinsed with water, and scanned for recording purposes.

2.2.5. Digestion of the Soluble Fraction

Two different types of tryptic digestions were performed. Non-denatured proteins, in 10 µL fractions, were digested directly with the addition of 1.5 µL trypsin followed by an overnight incubation at 37°C. Fractions were also denatured in 8 M urea, reduced with 10 mM TCEP, and alkylated with 30 mM iodoacetamide prior to the addition of trypsin (1.5:10, v/v). All digests were stored at -20°C until a few minutes before their injection onto the LC.

2.2.6. Chromatography and Mass Spectrometry

Chromatography was performed on a 0.17 mm column packed with Jupiter C_{18} stationary phase (Phenomenex, Torrance, CA). Solvent A for liquid chromatography comprised 0.05% TFA and 0.1% acetic acid in 95:5 water:acetonitrile; solvent B contained the same acid modifiers in 5:95 water:acetonitrile. A linear gradient was performed from 10 to 60% B in 70 min followed by a steeper gradient up to 95% B over 20 min, for a total of 90 minutes of running time. Longer (180 min) and shorter (40 min) chromatography runs were used occasionally when it was known that the sample had many or few components, respectively. Specifically, the tryptic digest of the entire lysate and of the non-DNA-binding portion of the lysate each required 180 min of chromatographic separation due to the complexity of these samples. By contrast, the relatively few components contained in the fraction eluted off the DNA cellulose with 3.0 M NaCl required only 40 min of chromatographic separation. Mass spectra were obtained on a Finnegan LCQ ion trap mass spectrometer (ThermoFinnegan, San Jose, CA). Following a method similar to Wolters *et al.* [5], the LCQ was set using the Instrument Method file to acquire a full MS scan between *m/z* 500 and 1,500 and then to perform three consecutive MS/MS scans between *m/z* 200 and 2,000 of the three most intense ions recorded in the MS scan. Dynamic exclusion was used to repeat the MS/MS of any peptide three times over 0.5 min and then exclude that peptide for the next 10 min.

2.2.7. Data Analysis

SEQUEST (ThermoFinnegan) and Mascot (Matrix Science, London, UK) were used for protein identification. For both programs, only the *E. coli* database was searched in order to limit the data analysis time. Although trypsin was used for digestion, no enzyme was selected for database searching because the authors have frequently observed non-tryptic and partially-tryptic peptides in tryptic digests. The SEQUEST algorithm was used to analyze tandem mass spectra as previously described [14] with extremely conservative criteria invoked to interpret the SEQUEST results [5]. Specifically, all results had to have a ? Cn of 0.1 or greater to be considered. For a singly charged peptide, the cross-correlation score (Xcorr) had to be at least 1.9, and the peptide had to be tryptic. Doubly-charged tryptic peptides had to have an Xcorr greater than 2.2, and doubly charged nontryptic peptides had to have an Xcorr greater than 3.0. Triply charged tryptic or partially tryptic peptides had to have an Xcorr of 3.75 or greater. Mascot [15] was used to verify the protein identifications found using SEQUEST and also to search for any posttranslational modifications. In order for a Mascot result to be accepted, the individual ion scores had to indicate that the spectrum obtained experimentally was identical to that of the theoretical peptide matched to that spectrum. Generally, peptides with individual ion scores below 30 were not considered to be well matched from a Mascot search and were disregarded.

2.3. Results and Discussion

The purpose of this study was not to identify proteins that bind to specific sequences of DNA, but instead to use the DNA-binding properties of proteins to reduce the complexity of a mixture of proteins prior to proteolytic digestion in order to identify as many proteins as possible while gaining some insight into the function of a subset of the proteins. A scheme illustrating this function-based approach to identifying DNA-binding proteins in a proteome can be seen in Figure 2.1. Single-stranded DNA was chosen as the substrate since many DNA-binding proteins bind to both double- and single-stranded DNA. Also, double stranded DNA substrates anneal best in the presence of salts, which might affect the binding of some proteins. Certain proteins identified as DNA-binding in this study may have bound to our substrate due to an ion-exchange process (with the sugar-phosphate backbone of the DNA acting as a cation exchange material), but if this enables the fractionation of proteins, then it is beneficial to protein identification. Separation of proteins with a strong cation exchange material might prove to be a useful control experiment in the future and may aid in determining a protein's true DNA-binding character.
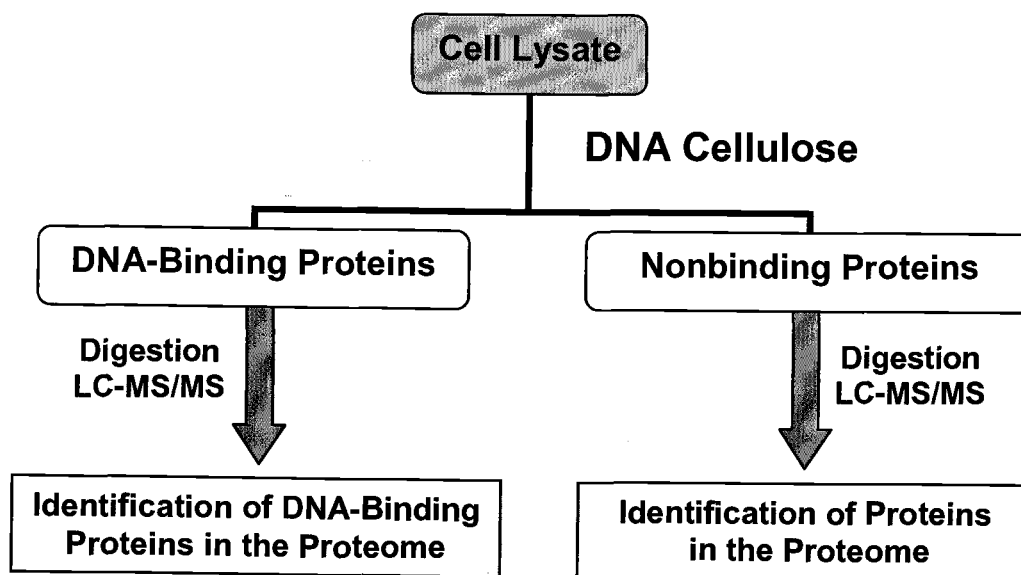
Figure 2.1. Scheme for the identification of DNA-binding proteins in a proteome.

Due to high variations in results obtained from the authors' previous experience with LC-MS/MS of tryptic digests, in-solution digestions were performed in two different ways in order to increase the number of proteins identified from *E. coli*. The first method consisted of mixing non-denatured fractions with trypsin and incubating at 37°C overnight. Even without denaturation, reduction, and alkylation, many excellent tandem mass spectra were obtained. The vast majority of peptides identified from the non-denaturing digestions did not contain cysteines. This fact is most likely due to the presence of nonreduced disulfide bonds in the tryptic peptide mixture. However, since the goal of this study was to simply identify proteins, not necessarily to obtain 100% coverage, this worked as an advantage in further reducing the complexity of the

peptide mixture prior to liquid chromatography. Since crosslinked peptides have significantly more mass than non-crosslinked peptides and, therefore, typically elute later in a chromatographic run, competition for ionization between cysteine-containing peptides and the rest of the peptides is decreased by the non-denaturing procedure. Just as the ICAT method only looks at cysteine-containing peptides [16], the digestion protocol used in this study yields predominantly non-cysteine-containing peptides.

The second digestion method comprised denaturation in urea, reduction with TCEP, and alkylation with iodoacetamide. Similar but not identical results were obtained for this more complicated and time-consuming technique. Cysteine-containing peptides were identified from the LC-MS/MS runs of the peptide mixtures obtained using these digestion conditions with most of the cysteines being alkylated. One disadvantage to this technique is that one must search for modified and unmodified cysteines in the data analysis when using either Mascot or SEQUEST; therefore, the searching inevitably takes longer due to the cysteine alkylation. Also, it has recently been reported that overalkylation can be a major problem in this procedure, especially when the original protein concentration is not specifically known [17], which is the case for mixtures of proteins. Overalkylation leads to additional sample complexity, complicates data analysis, and lowers confidence in protein identifications.

After the separation of the DNA-binding proteins from the rest of the lysate, a one-dimensional SDS-PAGE minigel was run to obtain a rough estimate

of the protein content in each of the fractions. Figure 2.2 shows the result of a minigel experiment where the DNA-binding portion was divided into four fractions corresponding respectively to salt elutions with 0.05, 0.15, 0.5, and 1.0 M NaCl. Both the whole lysate and the nonbinding portion of the lysate (the flow-through) contained very complex mixtures of proteins. The four fractions that were eluted off the column contain the various DNA-binding proteins that are the specific targets of this study. The fraction that eluted from the column with 0.5 M NaCl contained the majority of the DNA-binding proteins; hence in later studies, the number of elutions around that concentration was increased to further resolve the DNA-binding proteins. For the rest of the study, eight fractions were collected: lysate, flow-through, and eluates at strengths of 0.05 M, 0.2 M, 0.4 M, 0.6 M, 1.0 M, and 3.0 M NaCl.
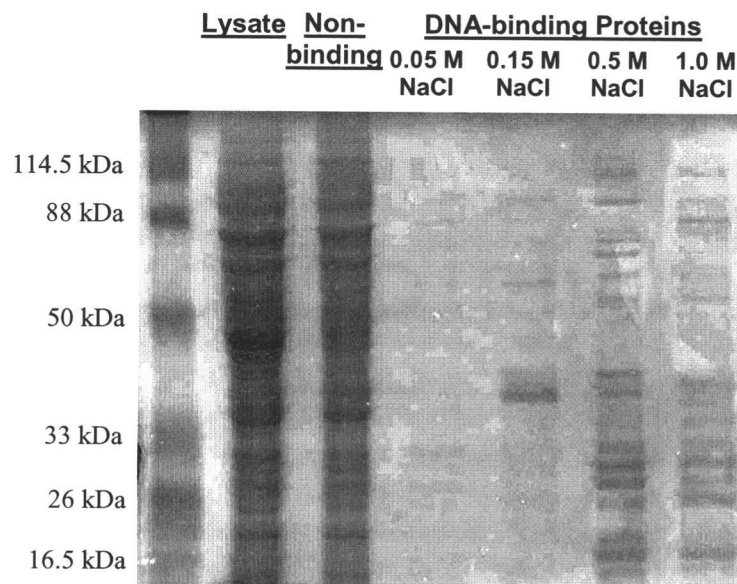
Figure 2.2. One-dimensional SDS-PAGE gel showing fractions off the DNA cellulose column for aerobically grown *E. coli*.

In order to determine whether the added function-based prefractionation step enabled identification of a greater number of proteins than obtained from the digest of the whole lysate, the numbers of proteins found in one set of experiments were compared. Digestion of the whole lysate followed by liquid chromatography and tandem mass spectrometry yielded 105 protein identifications. Digestions of the flow-through and DNA-binding fractions led to the identification of 200 proteins. It is important to look at the proteins found in the non-binding portion of the lysate since it is less complex than the whole lysate by virtue of most if not all of the DNA-binding proteins having been removed. The ribosomal proteins make up about 21% of the protein mass for *E. coli* [18], so if they are eliminated from a protein mixture by binding to DNA cellulose, it is easier to identify other proteins in the cell. Most of the ribosomal proteins (53 out of 56) were identified in this study, and the majority of them were observed in the DNA-binding fractions. A total of 541 proteins were identified from all of the analyses of the *E. coli* lysates, with 232 of those proteins binding to the DNA cellulose.

An alternative measure of the value of using a function-based separation in a proteomic study is to look at the predicted expression values of the different proteins observed. A separation should increase sensitivity enough to enable identification of proteins that are expressed at low levels, given that the majority of proteins are not highly expressed [5]. A listing of the expression values of some of the proteins observed in this study can be found in Table 2.1. The expression values listed are for predicted highly expressed (PHX) genes, which are calculated

by comparing codon usage frequencies between a gene of interest and genes that are known to be highly expressed. A protein with a PHX value of 1 or greater is predicted to be highly expressed. Digestion of the whole lysate leads primarily to the identification of proteins with PHX values greater than 1. By contrast, prefractionation based on DNA-binding ability leads to the identification of proteins with lower expression values, such as uracil DNA glycosylase, with a PHX value of 0.47 [19,20]. Of the proteins found from in-solution digestion of the whole cell lysate, 55% had PHX values greater than 1. Only 30% of the proteins found in the DNA-binding portion of the lysate had PHX values greater than 1, while 52% of those found in the flow-through had PHX values greater than 1.

Proteins	PHX Value	Lysate	Binding fractions
DNAK Protein	2.58	*	*
30S Ribosomal Protein S2	2.37		*
30S Ribosomal Protein S3	2.14	*	
30S Ribosomal Protein S1	2.12	*	*
Trigger Factor (TF)	2.05	*	*
50S Ribosomal Protein L3	1.90		*
50S Ribosomal Protein L9	1.88	*	*
Heat Shock Protein HTPG	1.85	*	
50S Ribosomal Protein L11	1.82	*	*
50S Ribosomal Protein L5	1.81	*	
50S Ribosomal Protein L21	1.71	*	
50S Ribosomal Protein L24	1.69		*
50S Ribosomal Protein L6	1.63	*	*
50S Ribosomal Protein L19	1.59		*
30S Ribosomal Protein S6	1.52	*	*
Single-Strand Binding Protein (SSB)	1.48		*
30S Ribosomal Protein S7	1.45	*	*
50S Ribosomal Protein L16	1.41		*
30S Ribosomal Protein S5	1.40	*	*
DNA Protection During Starvation Protein	1.13		*
Exodeoxyribonuclease III	0.64		*
Uracil DNA Glycosylase (UDG)	0.47		*

Table 2.1. Predicted expression values for some of the proteins found from *E. coli*. An asterisk indicates identification of a protein.

E. coli was grown under aerobic and anaerobic conditions in order to determine if any changes in protein expression or in protein function could be observed using the DNA-binding property as the basis for a prefractionation step. The first study of the effect of anaerbiosis on E. coli was performed in 1983 when 18 proteins were found to be overexpressed during anaerobic growth, as determined by audioradiographic measurements of two-dimensional SDS-PAGE [21,22]. The two one-dimensional SDS-PAGE minigels seen in Figure 2.3 show the DNA-binding proteins found in aerobically and anaerobically grown E. coli. Changes in protein concentration can be observed just by looking at the gels. Gel bands could have been extracted, digested, and analyzed by LC-MS/MS, but due to the poor yield of peptides and labor-intensiveness of in-gel digestions, in-solution digestions were employed. The fractions containing DNA-binding proteins, as well as the lysate and non-binding fractions, were digested with trypsin, and the resulting peptides were separated using 90 min LC runs with MS/MS. This procedure yielded the identification of 127 putative DNA-binding proteins in the aerobically and anaerobically grown E. coli.
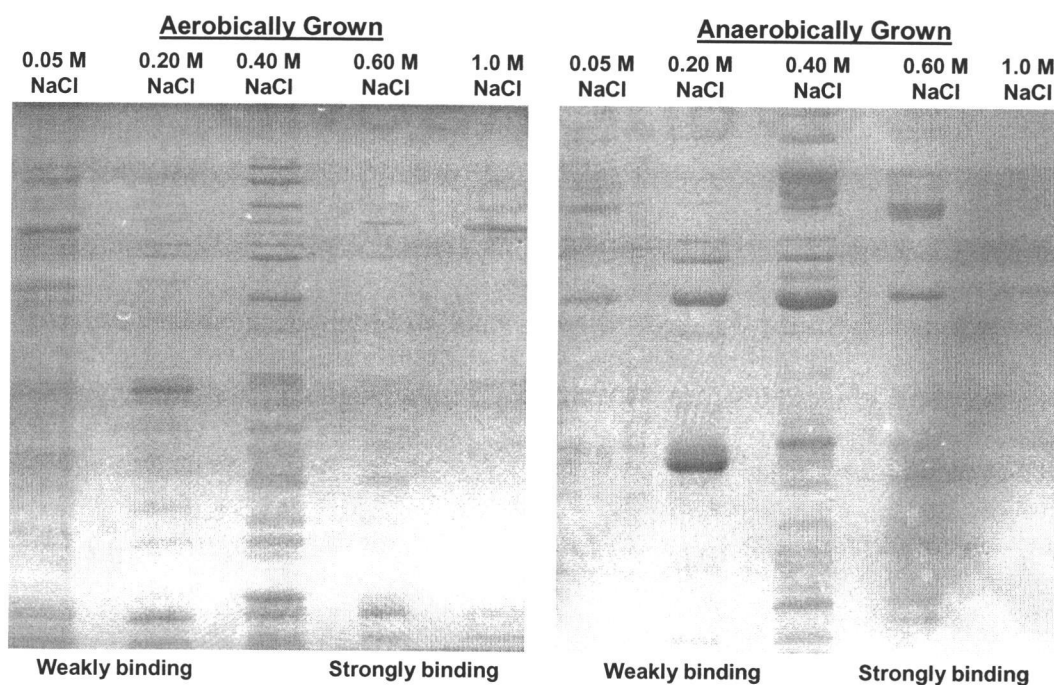
Figure 2.3. DNA-binding proteins in aerobically and anaerobically grown *E. coli.*

A "virtual gel" showing the DNA-binding fractions from *E. coli* grown under the two environmental conditions can be seen in Figure 2.4. The color of each cell corresponds to the number of peptides used to identify a given protein in a particular fraction. The number of peptides that are identified from a digest of a protein mixture is a rough indicator of protein concentration in that mixture. Although using relative isotopic abundance techniques, such as ICAT, is the most accepted way to determine expression levels in proteomic studies, the present authors chose to avoid the cost and labor involved with such techniques. Another group has explored the use of relative peak areas of peptides as they elute in LC-MS/MS to determine the relative concentration of proteins in a mixture [23], but this too can be labor-intensive. In order to see whether the changes in

expression level measured by counting the peptides identified from a given protein correspond to previous studies, pyruvate formate lyase (PFL) was examined as a model. This protein has been shown to be over-expressed 8-fold when *E. coli* is grown under anaerobic conditions [21]. In the present study, 16 peptides from PFL were found in the anaerobic lysate, whereas only 4 (one quarter fewer) peptides were found in the aerobic lysate. Thus, the number of peptides in this particular instance does indeed provide an estimation of expression that approximates the level to within an order of magnitude.

Intriguingly, 50 of the proteins that bound to the DNA-cellulose were heretofore unknown or hypothetical proteins based on genomic sequence. Some of the hypothetical proteins found to be DNA-binding are listed in Table 2.2. One hypothetical protein, a 14.9 kDa protein from the MINC-SHEA intergenic region, was found to bind strongly to the DNA cellulose when expressed under both aerobic and anaerobic conditions. This protein's function may not be affected by the environmental conditions under which the cells are grown. Another hypothetical protein, a 36.1 kDa protein from the LPP-AROD intergenic region, was found to bind strongly to the DNA cellulose only when expressed in aerobic cells. Obviously, the function of this protein changes when it is grown under different conditions. The functions of these hypothetical proteins are not yet known, but these results are interesting from the perspective that the expression of proteins that were unknown to date except by their genomic sequence has now been observed in conjunction with a DNA-binding function.
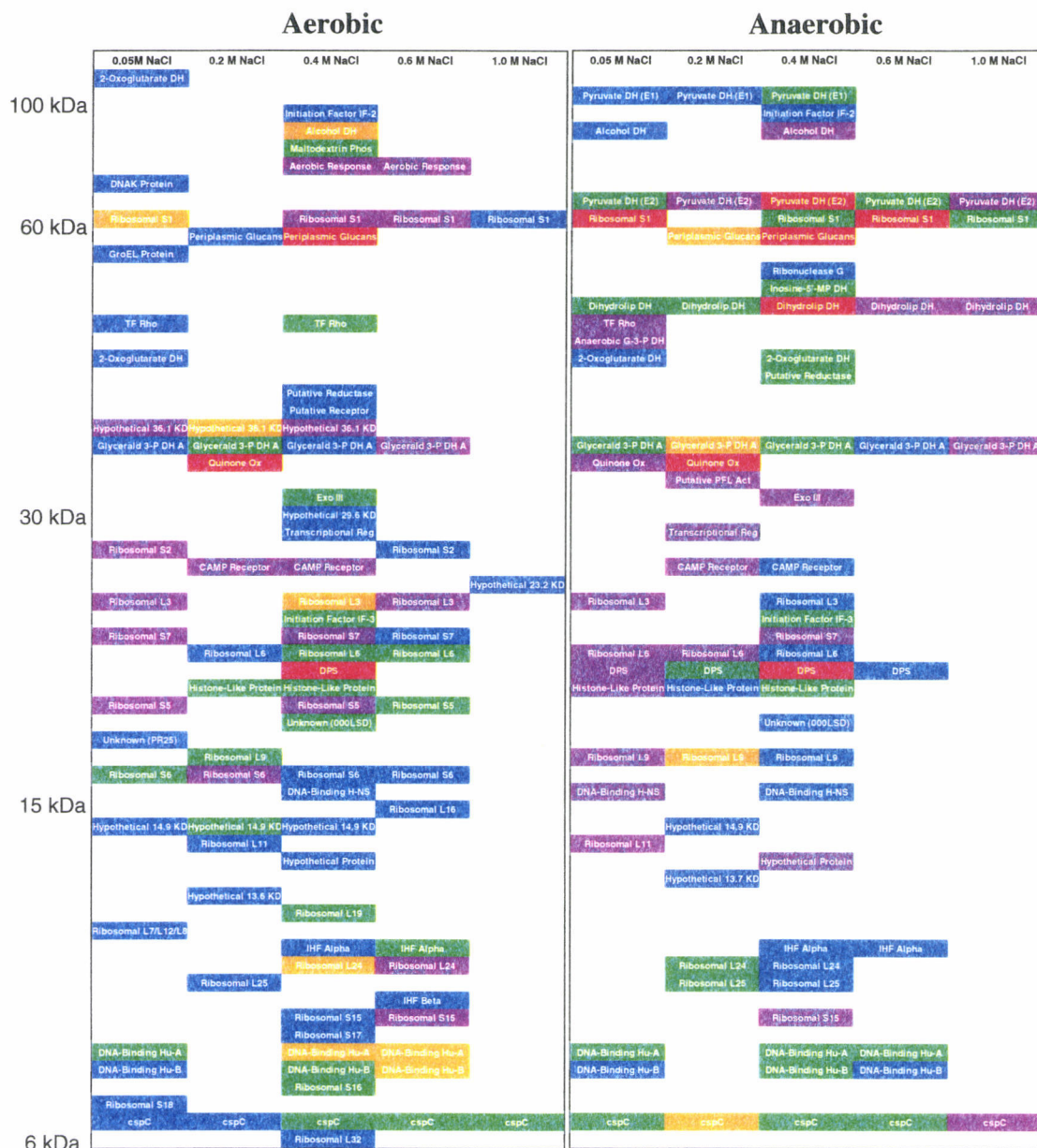
Figure 2.4. Virtual gels showing the DNA-binding proteins in aerobically (left) and anaerobically (right) grown *E. coli*. The proteins were found through in-solution digestion followed by tandem mass spectrometry. The red (■) boxes designate proteins that were identified from 12 or more peptides, orange (■) from 8 or more peptides, green (■) from greater than 4 peptides, blue (■) from 2 or 3 peptides, and purple (■) from only 1 peptide in that fraction. To conserve space,

Hypothetical/Unknown DNA-Binding Proteins	Peptides in Aerobically Grown *E. coli*	Peptides in Anaerobically Grown *E. coli*
Unknown Protein 2D_000LSD From 2D-PAGE Precursor		2
Hypothetical 36.1 kDa Protein in LPP-AROD Intergenic Region	10	
Hypothetical 29.6 kDa Protein in THRC-TALB Intergenic Region	3	
Hypothetical 14.9 kDa Protein in MINC-SHEA Intergenic Region	8	3
Hypothetical 13.7 kDa Protein in MTLR-LCTP Intergenic Region		2
Hypothetical 13.6 kDa Protein in ACRD-DAPE Intergenic Region	2	

Table 2.2. Hypothetical proteins found in DNA-binding fractions from *E. coli* grown under aerobic and anaerobic conditions.

Comparison of the DNA-binding proteins found in the aerobic and anaerobic lysates shows a few unexpected differences. The pyruvate dehydrogenase complex is a major component of the citric acid cycle, and a previous investigation showed that the proteins composing this complex are not highly expressed in anaerobically grown *E. coli* [24]. By contrast, the present authors found the complex to be highly expressed and DNA-binding in the anaerobic samples. Examining Figure 2.4, the pyruvate dehydrogenase complex in the anaerobic *E. coli* is clearly seen to be expressed in a way that causes it to bind to the DNA cellulose (Figure 2.4: 0.05 M, 0.2 M, and 0.4 M NaCl columns; 60-100 kDa), whereas this complex does not bind in the aerobic sample. Another surprising result was found with alcohol dehydrogenase (ADH), which is usually expressed under anaerobic conditions. It was not surprising to find a low level of ADH present in anaerobically grown *E. coli* (Figure 2.4: 0.05 M and 0.4 M columns; ~80 kDa). It was surprising, however, to find this protein strongly bound

to DNA primarily in the aerobic sample (Figure 2.4: 0.4 M column; ~80 kDa).

The sequence of ADH as well as the peptides found bound to DNA cellulose from

aerobic *E. coli*, can be seen in Figure 2.5. It has been shown through site-directed

mutagenesis that the aerobic activity of ADH is determined by the acidic residue

(E) at position 568 [25]. In the present experiment, 19% coverage was obtained

for aerobically expressed, DNA-binding ADH, including the peptide containing

the glutamic acid residue that is necessary for aerobic inactivation

($_{554}$IMWVMYEHPETHFEELALR$_{572}$). These results seem to suggest that ADH,

although expressed by aerobically grown *E. coli*, is not actually active and may

change its tertiary structure under aerobic conditions such that it binds to DNA.

MAVTNVAELNALVERVKKAQREYASFTQEQVDKIFRAAALAAADARIPLAKMAVAESGMGIVED$_{64}$
KVIKNHFASEYIYNAYKDEKTCGVLSEDDTFGTITIAEPIGIICGIVPTTNPTSTAIFKSLISL$_{128}$
KTRNAIIFSPHPRAKDATNKAADIVLQAAIAAGAPKDLIGWIDQPSVELSNALMHHPDINLILA$_{192}$
TGGPGMVKAAYSSGKPAIGVGAGNTPVVIDETADIKRAVASVLMSKTFDNGVICASEQSVVVVD$_{256}$
SVYDAVRERFATHGGYLLQGKELKAVQDVILKNGALNAAIVGQPAYKIAELAGFSVPENTKILI$_{320}$
GEVTVVDESEPFAHEKLSPTLAMYRAKDFEDAVEKAEKLVAMGGIGHTSCLYTDQDNQPARVSY$_{384}$
FGQKMKTARILINTPASQGGIGDLYNFKLAPSLTLGCGSWGGNSISENVGPKHLINKKTVAKRA$_{448}$
ENMLWHKLPKSIYFRRGSLPIALDEVITDGHKRALIVTDRFLFNNGYADQITSVLKAAGVETEV$_{512}$
FFEVEADPTLSIVRKGAELANSFKPDVIIALGGGSPMDAAKIMWVMYEHPETHFEELALRFMDI$_{576}$
RKRIYKFPKMGVKAKMIAVTTTSGTGSEVTPFAVVTDDATGQKYPLADYALTPDMAIVDANLVM$_{640}$
DMPKSLCAFGGLDAVTHAMEAYVSVLASEFSDGQALQALKLLKEYLPASYHEGSKNPVARERVH$_{704}$
SAATIAGIAFANAFLGVCHSMAHKLGSQFHIPHGLANALLICNVIRYNANDNPTKQTAFSQYDR$_{768}$
PQARRRYAEIADHLGLSAPGDRTAAKIEKLLAWLETLKAELGIPKSIREAGVQEADFLANVDKL$_{832}$
SEDAFDDQCTGANPRYPLISELKQILLDTYYGRDYVEGETAAKKEAAPAKAEKKAKKSA$_{891}$

Figure 2.5. Amino acid sequence of alcohol dehydrogenase from *E. coli*. Highlighted peptides are those that were found in DNA-binding fractions of lysate from aerobically grown cells. The underlined glutamic acid (E568) is one that was found to determine the activity of ADH under aerobic conditions. The peptides found had ion scores that ranged from 53 ($_{468}$PIALDEVITDGHK$_{480}$) to 79 ($_{489}$FLFNNGYADQITSVLK$_{504}$).

In order to test the robustness of our approach, the LCQ's performance, and the authors' skills at data analysis, samples were sent to Applied Biosystems (Foster City, CA) for analysis. Two complex samples, containing proteins respectively eluted off single-stranded DNA cellulose with 0.4 M NaCl (weakly binding) and 1.0 M NaCl (strongly binding) were digested with trypsin in the authors' laboratory. Using a multidimensional approach, the peptides were then applied to a strong cation exchange column, separated with eight increasing salt concentrations, subjected to reversed-phase liquid chromatography, and analyzed on a QSTAR quadrupole time-of-flight mass spectrometer (Data not shown). All but three of the proteins that were identified in the authors' laboratory using one-dimensional liquid chromatography were identified at Applied Biosystems. Over 430 proteins were identified using the multidimensional chromatographic approach; this result is comparable to the authors' results in which eight fractions off the DNA cellulose were analyzed.

2.4. Conclusions

This study demonstrates an effective use of a function-based separation. Many more proteins were identified when proteins were prefractionated based on DNA-binding ability prior to digestion (200 proteins) than when the whole cell lysate was digested and analyzed (105 proteins). Over 540 proteins from *E. coli*

were observed in all; 232 of those bound to DNA cellulose. Changes in binding affinities of proteins were also observed between cells grown in aerobic and anaerobic environments. These results were achieved without benefit of two-dimensional chromatography by using the DNA-binding property of the proteins to separate them from the rest of the lysate and thereby reduce the complexity of the digestion mixture prior to mass analysis. A comparison of the results obtained in this study with those obtained by scientists in another laboratory who used multidimensional chromatography shows that function-based separation yields comparable numbers of accurately identified proteins.

Function-based separation should be applicable well beyond the particular DNA-binding system that was used in this study to demonstrate its principles. Within the scope of DNA-binding, protein expression could be studied in the presence of other environmental stressors, such as agents that lead to oxidative stress or exposure to radiation. More generally, the ligand used to determine protein function could be varied to fractionate proteins based on their ability to bind other molecules, such as RNA, glucose, or specific pharmaceuticals. Finally, it should be possible to incorporate isotopic labeling into a function-based separation for quantitative purposes and into multidimensional chromatography experiments for increased resolution.

2.5. References

1. Griffin, T.J. and Aebersold, R. (2001) "Advances in Proteome Analysis by Mass Spectrometry," *J Biol Chem*, 276:45497-45500.

2. Tyers, M. and Mann, M. (2003) "From Genomics to Proteomics," *Nature*, 422:193-197.

3. O'Farrell, P.H. (1975) "High Resolution Two-Dimensional Electrophoresis of Proteins," *J Biol Chem*, 250:4007-4021.

4. VanBogelen, R.A., Abshire, K.Z., Moldover, B., Olson, E.R., and Neidhardt, F.C. (1997) "*Escherichia coli* Proteome Analysis Using the Gene-Protein Database," *Electrophoresis*, 18:1243-1251.

5. Wolters, D.A., Washburn, M.P., and Yates, J.R. III (2001) "An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics," *Anal Chem*, 73:5683-5690.

6. Champion, M.M., Campbell, C.S., Siegele, D.A., Russell, D.H., and Hu, J.C. (2002) "Proteome Analysis of *Escherichia coli* K-12 by Two-Dimensional Native-State Chromatography and MALDI-MS," Poster, *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics,* Orlando.

7. Lo, L.C., Pang, T.L., Kuo, C.H., Chiang, Y.L., Wang, H.Y., and Lin, J.J. (2001) "Design and Synthesis of Class-Selective Activity Probes for Protein Tyrosine Phosphatases" *J Proteome Research*, 1:35-40.

8. Liu, Y., Patricelli, M.P., and Cravatt, B.F. (1999) "Activity-Based Protein Profiling: The Serine Hydrolases," *Proc Natl Acad Sci USA*, 96:14694-14699.

9. Greenbaum, D., Medzihradszky, K.F., Burlingame, A., and Bogyo, M. (2000) "Epoxide Electrophiles as Activity-Dependent Cysteine Protease Profiling and Discovery Tools," *Chem Biol*, 7:569-581.

10. Shefcheck, K.J., Yao, X., and Fenselau, C. (2002) "Evaluation of Heparin for Fractionation of Cytosolic Proteins in Cancer Cells," Poster, *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics,* Orlando.

11. Dishman, R. (2002) "Rationalizing Proteomics: New Strategies for Streamlining Drug Discovery," *PharmaGenomics*, 2:58-62.

12. Bane, T.K., Leblanc, J.F., Lee, T.D., and Riggs, A.D. (2001) "Analysis of DNA Binding Proteins by Mass Spectrometry," Poster, *Proceedings of the 49th ASMS Conference on Mass Spectrometry and Allied Topics*, Chicago.

13. Bane, T.K., LeBon, J.M., Lee, T.D., and Riggs, A.D. (2002) "DNA-Protein and Protein-Protein Interactions by DNA Affinity Capture-LC/MS/MS," Poster, *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando.

14. Eng, J.K., McCormack, A.L., and Yates, J.R., III (1994) "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database," *J Am Soc Mass Spec*, 5:976-989.

15. Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999) "Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data," *Electrophoresis*, 20:3551-3567.

16. 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.

17. Boja, E.S. and Fales, H.M. (2001) "Overalkylation of a Protein Digest with Iodoacetamide," *Anal Chem,* 73:3576-3582.

18. Reilly, J.P. and Arnold, R.J. (1999) "Observation of *Escherichia coli* Ribosomal Proteins and their Posttranslational Modifications by Mass Spectrometry," *Anal Biochem*, 269:105-112.

19. Karlin, S. and Mrázek, J. (2000) "Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes," *J Bacter,* 182:5238-5250.

20. Karlin, S., Mrázek, J., Campbell, A., and Kaiser, D. (2001) "Characterizations of Highly Expressed Genes of Four Fast-Growing Bacteria," *J Bacter,* 183:5025-5040.

21. Smith, M.W. and Neidhardt, F.C. (1983) "Proteins Induced by Anaerobiosis," *J Bacter,* 154:336-343.

22. Smith, M.W. and Neidhardt, F.C. (1983) "Proteins Induced by Aerobiosis," *J Bacter,* 154:344-350.

23. Chelius, D. and Bondarenko, P.V. (2002) "Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry," *J Proteome Res*, 1:317-323.

24. Cassey, B., Guest, J.R., and Attwood, M.A. (1998) "Environmental Control of Pyruvate Dehydrogenase Complex Expression in *Escherichia coli*," *FEMS Micro Lett*, 159:325-329.

25. Holland-Staley, C.A., Lee, K., Clark, D.P., and Cunningham, P.R. (2000) "Aerobic Activity of *Escherichia coli* is Determined by a Single Amino Acid," *J Bacter*, 182:6049-6054.

3. MASS SPECTROMETRIC DETERMINATION OF CONTACT POINTS BETWEEN DNA AND PROTEINS CONTAINED IN A FUNCTIONAL SUBSET OF A PROTEOME

Martha D. Stapels and Douglas F. Barofsky

3.1. Introduction

The central goals in proteomic studies include the identification of all of the proteins in a proteome, the characterization of the function of the proteins, and the determination of their three-dimensional structure. In order to identify all of the proteins in a proteome, more than one dimension of separation has proven to be necessary. The first multi-dimensional separations of proteins involved gel electrophoresis, whereby proteins are initially separated on the basis of isoelectric point (pI) and then by molecular weight [1]. Recently, separation of proteins [2] or peptides [3] based on pI followed by separation based on hydrophobicity has led to the identification of greater numbers of proteins than can be identified through the earlier gel techniques. These protocols do not provide any clues to protein function, however. Affinity chromatography can be used to separate proteins to produce less complex samples while yielding information about the function of the proteins [4]. Separation of proteins based on function has been used to identify serine hydrolases [5], recognize DNA-binding proteins that bind to specific sequences [6], isolate cysteine proteases [7], and fractionate proteins based on their affinity to heparin [8].

The most common techniques used to determine protein structure are NMR [9] and x-ray crystallography [10,11]. Both of these methods require large quantities of purified protein. The progress of x-ray crystallographic studies is currently limited by the rate at which proteins can be crystallized [11], a step that

can be particularly difficult when the protein under investigation is interacting with a substrate. Hydrogen/deuterium exchange has been recently used with mass spectrometry (MS) to determine the structure of proteins [12], but this technique is limited mainly to purified proteins. Photochemical crosslinking is another technique that has been combined with mass spectrometry to predict the three-dimensional structure of proteins [13]. This technique has been used to determine the DNA-binding sites of many proteins, including uracil DNA glycosylase [14,15], replication protein A [16], and single-stranded binding protein [17]; however, all of these studies have involved proteins in purified form. In this study, the authors demonstrate how a protocol they recently introduced for separating proteins based on their ability to bind DNA [18] can be exploited in combination with photochemical crosslinking to both isolate DNA-binding proteins and determine the contact points between some of them and DNA.

3.2. Experimental Section

3.2.1. Materials

Trifluoroacetic acid (TFA), formic acid (FA), acetic acid (AA), a-cyano-4-hydroxycinnamic acid (HCCA), ethylenediaminetetraacetic acid (EDTA), sodium chloride (NaCl), ammonium phosphate, ammonium acetate, single-stranded calf thymus DNA cellulose, lysozyme, yeast extract, tryptone, and Trizma (Tris-base)

were purchased from Sigma Chemical Co. (St Louis, MO). Lyophilized sequencing-grade trypsin was purchased from Promega (Madison, WI). HPLC grade acetonitrile (ACN) was supplied by Fisher Scientific (Pittsburgh, PA). Nuclease P1 was purchased from Amersham Biosciences (Piscataway, NJ). For general use, water was generated with a Milli-Q Ultrapure water purification system (Millipore Corp. Bedford, MA); for HPLC, Burdick and Jackson water was purchased from Honeywell International, Inc. (Muskegon, MI). Ultrafree-0.5 centrifugal filter devices (molecular weight cut-off 5,000 and 30,000) were purchased from Millipore Corp. Applied Biosystems supplied the 4700 Proteomics Analyzer Calibration Mixture (Framingham, MA).

3.2.2. Growth and Lysis of *E. coli*

Escherichia coli Strain BL21 (*E. coli* B F⁻ *dcm ompT hsdS*(r_8^- m_8)*gal*) from Stratagene (La Jolla, CA) was grown in LB media (1% tryptone, 0.5% yeast extract, and 1% NaCl) at 37°C under aerobic conditions. Cells were frozen in liquid nitrogen, resuspended in a buffer comprising 100 mM Tris (pH 7) and 1 mM EDTA, lysed ultrasonically after the addition of 0.2 mg/mL lysozyme, and centrifuged to remove insoluble material. Lysates were stored at -20°C until further use.

3.2.3. Separation of DNA-Binding Proteins

Cell lysate (20 mL) was mixed at room temperature with 5 g of single-stranded DNA cellulose and allowed to equilibrate for 15 min. Centrifugation was used to separate the cellulose particles from the rest of the lysate, which contained non-binding proteins. The non-binding proteins were rinsed off the cellulose particles with the addition of 20 mL of DAB buffer (20 mM Tris pH 7.4 with 1 mM EDTA). This step was repeated three times, the non-binding proteins being decanted off and collected after each rinse. The proteins that bound to the DNA were released in fractions with increasing concentrations of sodium chloride (0.05 M, 0.2 M, 0.4 M, 0.6 M, 1.0 M). Each DNA-binding fraction was obtained by mixing the cellulose with 25 mL of the salt solution, decanting, collecting the solution, repeating the first three steps, and then pooling the two extracts. Each fraction was concentrated and desalted using an Ultrafree-0.5 centrifugal filter device with a molecular weight cut-off of 5,000 Da. Total protein concentration in each fraction after this point was estimated to be 0.2 pmol/μL based on the intensity of stained protein bands in a polyacrylamide gel (data not shown). The desalting step (facilitated by adding 0.5 mL of fresh 20 mM Tris pH 7.4 to the filter) was repeated three times to ensure that the sodium chloride and EDTA was removed from each of the samples.

3.2.4. Photochemical Crosslinking

Oligonucleotide dT_{20} was synthesized and purified by the Biopolymer Core Facility at the University of Maryland at Baltimore as previously described

[14,16]. The lyophilized dT_{20} was dissolved in DAB buffer to a concentration of 242 pmol/μL. In order to concentrate the oligonucleotide solutions to ~500 pmol/μL and to eliminate the EDTA prior to crosslinking, the solution was desalted and concentrated using Ultrafree-0.5 centrifugal filter devices with a molecular weight cut-off of 5,000 Da. To initiate the crosslinking, 100 μL of DNA-binding protein solution (0.4 M or 0.6 M NaCl eluted) was mixed with 50 μL of the dT_{20} solution and allowed to equilibrate on ice for 10 min. Of this protein/oligonucleotide mixture, 20 μL was saved on ice in the dark as a control, and 130 μL was irradiated with UV light (λ_{max} = 254 nm) for 30 min in a Stratalinker 1800 crosslinker (Stratagene, La Jolla, CA).

3.2.5. Purification of Crosslinked Proteins

To remove most of the uncrosslinked protein, the UV irradiated protein mixture was mixed with 150 μL of DNA cellulose and allowed to equilibrate at room temperature for 10 min. The crosslinked proteins were rinsed off the DNA cellulose with 300 μL of 20 mM Tris buffer (pH 7.4) four times. The crosslinked proteins were concentrated and washed with fresh Tris buffer using Ultrafree-0.5 centrifugal filter devices with a molecular weight cut-off of 30,000 Da. The final volume of each sample containing crosslinked proteins was 30 μL.

3.2.6. Enzymatic Digestions

All samples were digested in-solution with the addition of 2 μL of 0.5 μg/μL trypsin (in 0.1 M AA) to each sample. The solutions were vortexed, centrifuged, and placed on a 37°C water bath overnight to allow complete digestion to occur. Nuclease P1 was diluted with 50 mM ammonium acetate (pH 6.5) to yield 0.04 activity units per μL. The oligonucleotide tags were then digested with the addition of 10 μL of diluted nuclease P1 solution to each of the crosslinked samples, which were 32 μL in volume. These solutions were vortexed, centrifuged, and placed on a 37°C water bath for 4 hours.

3.2.7. Isolation of Crosslinked Peptides

An immobilized metal affinity chromatography (IMAC) kit from Pierce (Rockford, IL) was used on some of the crosslinked samples in order to isolate crosslinked peptides from the rest of the peptides in the digestion mixture. The kit contains a resin with chelated gallium (III) ions, which specifically bind to phosphopeptides [19] and to peptides crosslinked to DNA [17]. To adjust the pH of the peptide mixtures, 5% AA was added until the sample pH was approximately 2. The sample was then mixed with the resin for 10 minutes at room temperature. After washing non-specifically bound peptides off from the resin with 0.1 M AA, 50:50 ACN:0.1 M AA, and water, crosslinked peptides were eluted with 40 μL of 50 mM ammonium phosphate.

3.2.8. MALDI Mass Spectrometry

Samples for matrix assisted laser desorption ionization (MALDI) MS were mixed 1:3 (v/v) with a-cyano-4-hydroxycinnamic acid in 50% ACN and 0.1% TFA. MALDI-MS was performed on an Applied Biosystems 4700 Proteomics Analyzer with time-of-flight/time-of-flight (TOF/TOF) ion optics (Applied Biosystems, Inc., Framingham, MA). Data were acquired in the MALDI reflector mode using internal calibration standards in the ABI4700 CalMix supplied by Applied Biosystems. Data in the MS/MS mode were calibrated using fragment ions from Glu-fibrinopeptide B. Tandem mass spectra were acquired by accelerating the precursor ions to 8 keV, selecting them with a timed gate, and performing collision-induced dissociation (CID) at 1 keV. The timed gate was set to select precursors in a 3 Da-window; when two precursors were within 3 Da, the gate's window was reduced to 1 Da. Gas pressure (air) in the CID cell was set at 0.2 μTorr. Fragment ions were accelerated to 14 keV prior to entering the reflector.

3.2.9. Chromatography and Electrospray Mass Spectrometry

For LC-MS/MS experiments, each sample was mixed 1:1 (v/v) with solvent A (0.1% FA, 0.005% TFA, and 3% ACN in H_2O), and 6 μL of this solution was injected on column. Solvent B contained 0.1% FA and 0.005% TFA in 80% ACN. To desalt and concentrate each sample, a 5 mm by 0.32 mm C_{18} trap from LC Packing was used. A 15 cm long, 75 μm inner diameter PicoFrit

column from New Objective (Woburn, MA) packed in-house with Jupiter C_{18} from Phenomenex, Inc. (Torrance, CA) was used. The LC conditions started with 3% B for 5 min to wash the sample, followed by a gradient up to 40% B over 40 min, to 70% B at 50 min, to 90% B at 52 min, and held at 90% B until 60 min. A Waters CapLC system was used with a flow rate estimated to be 300 nL/min. The mass spectrometer used for ESI-MS/MS was a Quadrupole Time-of-Flight (Q-TOF) Global Ultima system from Micromass (Micromass, Ltd., Manchester, UK), operated with a spray voltage of 3.5 kV. Data-dependent MS/MS was used with a 0.5 sec survey scan and 2.5 sec MS/MS scans on the three most abundant peaks in the MS survey scan. In the neutral loss scanning mode [20], the collision energy was varied from 7 eV to 30 eV to look specifically for a loss of dT_2 (626.1 Da), and then MS/MS was performed (45 eV) on the precursors that yielded that loss.

3.2.10. Data Analysis

Mascot from Matrix Science Ltd. (London, UK) was used to search all of the tandem mass spectra. For the data obtained on the Q-TOF, files appropriate for Mascot (pkl files) were created using the Masslynx software from Waters, with a function that smoothes, calculates centroids, and assesses the quality of data. For data obtained with the MALDI TOF/TOF, spectra were converted into peak lists which were then searched using Mascot. The variable modifications searched for included oxidation of methionine, pyro-glu of N-terminal glutamic acid and

glutamine, and deamidation of asparagine and glutamine. Mascot was used to search for crosslinked oligonucleotide tags, dT_2 (626.1 Da) and dT_3 (930.1 Da) attached to any amino acid. Spectra were also searched manually looking for the product ion dT at m/z 305.06 or looking for a loss of dT_2 $(M-626.1)^+$.

3.3. Results

3.3.1. Experimental Strategy

The authors have recently described a protocol for prefractionating the members of a proteome based on protein function and demonstrated its application to DNA-binding proteins [18]. A scheme illustrating how this protocol can be extended to both identify protein function and characterize protein binding sites in a proteome can be seen in Figure 3.1. In the present study, DNA-binding proteins were first isolated from the rest of the proteins in an *E. coli* lysate based on their affinity for calf thymus DNA attached to cellulose particles. Single-stranded calf thymus DNA was chosen as a large, (50 kb) generic substrate so that as many DNA-binding proteins would be attracted as possible. The proteins captured by the calf thymus DNA compose a functional subset of the *E. coli* proteome. In this study, digesting these proteins with trypsin and performing tandem mass spectrometry (MS/MS) on the peptides led to the identification of 232 putative DNA-binding proteins from *E. coli*.
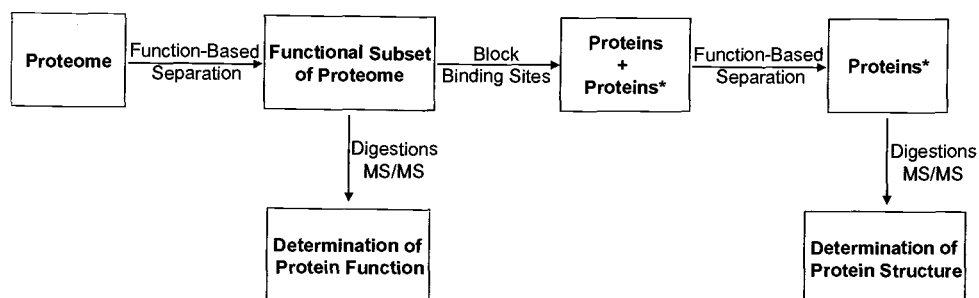
Figure 3.1. Scheme for the determination of the function and structure of proteins contained in a proteome. The chemical modification of the binding sites of proteins is denoted by an asterisk (*)

In order to covalently modify the DNA-binding sites in the proteins, the protein fractions were mixed with oligonucleotides and exposed to UV light. Only proteins from the 0.4 M and 0.6 M NaCl-eluted fractions were chosen for this part of the study because they are known to contain the majority of the DNA-binding proteins [18]. After irradiation, the samples contained a mixture of oligonucleotides, proteins, and proteins crosslinked to oligonucleotides. Based on previous studies [14], the yield of proteins crosslinked to the oligonucleotide dT_{20} is less than 5%. To enrich the amount of crosslinked protein relative to free protein, the solution was then mixed with DNA cellulose. In this instance, the noncrosslinked proteins bound to the DNA cellulose, while the crosslinked proteins rinsed off because their DNA-binding sites were already occupied by oligonucleotides. Proteins that originally bound to the DNA cellulose nonspecifically, due to ionic or other interactions, would again bind by the same interactions. In addition, DNA-binding proteins with more than one DNA-binding site could also bind to the DNA cellulose if only one of their binding sites was

occupied by a crosslinked oligonucleotide. Due to these possibilities, the second DNA cellulose binding step was omitted from some of the crosslinking experiments.

The samples containing crosslinked proteins were digested with trypsin to yield peptides, some of which were tagged with an oligonucleotide. This mixture was analyzed by MALDI-MS; however, no crosslinked peptides could be definitively identified from this analysis because true peptide-oligonucleotide crosslinks are difficult to distinguish from noncovalent adducts that form in a mixture containing free peptides and oligonucleotides [16]. Nuclease P1 was used to reduce the oligonucleotide tags to dimers or trimers so that the crosslinked peptides would have an appropriate size for MS/MS. Both LC-MS/MS and MALDI-MS/MS were used to analyze the crosslinked peptides at this stage.

3.3.2. Purification of Peptides Crosslinked to Oligonucleotides

Since noncrosslinked peptides were present in the solutions after digestion, IMAC, as described by Steen and Jensen [17] was used to enrich some of the samples in crosslinked peptides. The IMAC eluate was analyzed by both MALDI-MS/MS and ESI-MS/MS. This enrichment simplified the peptide mixture so that it contained only crosslinked and acidic peptides. This procedure proved to be especially useful in those cases where tryptic digestions were carried out on an entire crosslinked mixture that had not been subjected to a second function-based separation to bind the uncrosslinked proteins. In those cases where IMAC was used, no crosslinked peptides were found beyond those discovered when the entire

strategy shown in Figure 3.1 was employed. However, the IMAC purifications did yield the same crosslinked peptides as found from the analyses performed as diagramed in Figure 3.1. Higher confidence comes from these redundant results.

3.3.3. Tandem Mass Spectrometry of Peptides Crosslinked to Oligonucleotides

During MS/MS of a crosslinked peptide with either high- or low-energy collision induced dissociation, the peptide backbone fragments predominantly into y and b ions that are useful for identifying the peptide. Unfortunately, the nucleotide tag also fragments to create ions that complicate the spectrum. The fragmentation of the nucleotide tag dT_2 is illustrated in Figure 3.2. While these fragments are not useful for identification of the peptide, they are characteristic of a crosslinked species, and subsequent tandem mass spectra can be searched for this pattern. At this time, no automated searching algorithms are commercially available to search MS/MS data for both peptide and nucleotide fragments simultaneously.
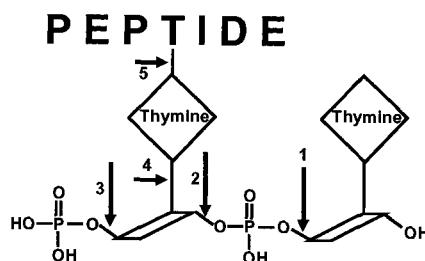
Figure 3.2. Fragmentation of the nucleotide tag dT_2 observed during MS/MS of crosslinked peptides. The fragmentation observed corresponds to well established pathways for nucleotides (a, c, and w ions) [21]. The numbering system shown in the diagram is used to clarify the nucleotide-peptide crosslink spectra presented in this paper.

Two representative high energy MS/MS mass spectra of crosslinked peptides can be seen in Figure 3.3. To the best of the present authors' knowledge, these spectra are the first examples of high energy MS/MS of peptides crosslinked to oligonucleotides ever reported. At 1 keV, the main fragment observed corresponds to cleavage of the entire dT_2 tag from the peptide. Three other nucleotide fragments can also be seen in both of the spectra. These fragments aid in identifying these spectra as those of crosslinked peptides, but they do not help to identify the peptide or protein from which they originate. The y ions seen in both of the spectra were used to identify the respective peptides; however, since there are so few peptide fragments, the Mascot ion scores were of low confidence. Also contributing to the low scores are the nucleotide fragments 2, 3, and 4, which Mascot attempts to fit as peptide fragments.

54
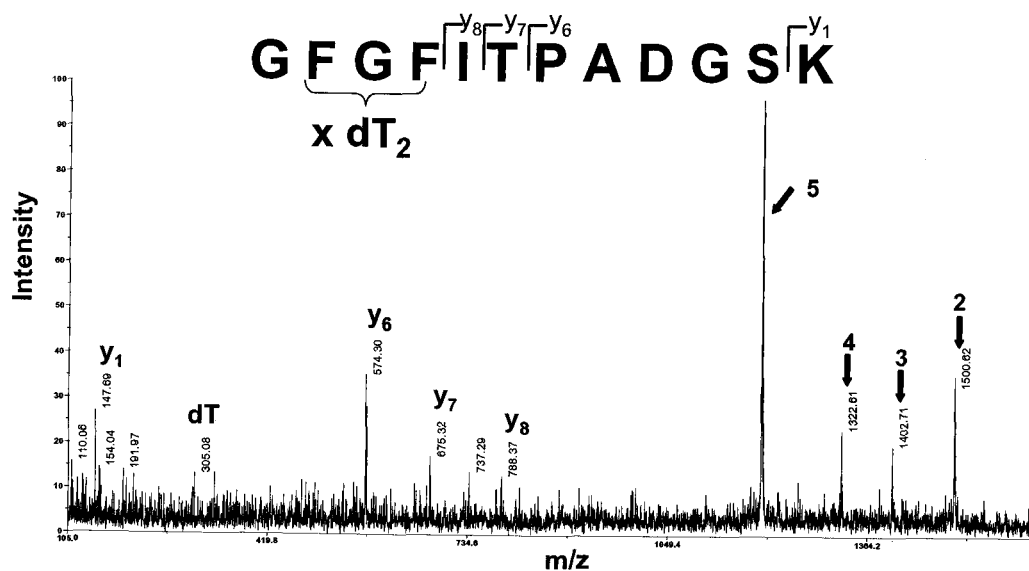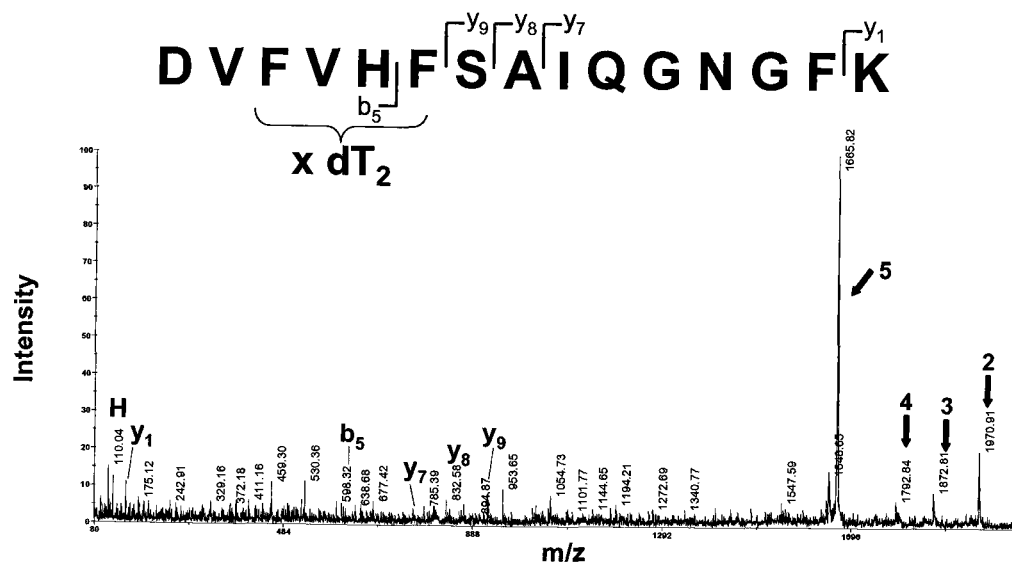
Figure 3.3. High energy MALDI tandem mass spectra corresponding to crosslinked peptides from CspC: (A) precursor with *m/z* 1,822.7; (B) precursor with *m/z* 2,292.8.

The low energy (40 eV) MS/MS spectra of the same crosslinked peptides, whose high energy spectra are shown in Figure 3.3, can be seen in Figure 3.4. The nucleotide fragments 1-5 are all present in the spectrum shown in Figure 3.4A, but they are not the largest peaks in this spectrum. The strong y-series (y_3 to y_{10}), along with some b ions (b_2 to b_4), enabled Mascot to assign this spectrum to a crosslinked peptide from CSP-C with an ion score of 27. As was so with the MALDI MS/MS spectrum (Figure 3.3A), this score is not high due to the presence of nucleotide fragments and ions with thymine attached. It can be readily inferred from the b- and y-ions that the dT_2 tag was attached to this peptide in the FGF region. It is not possible to narrow the assignment to one amino acid because isomers exist for each of the positions where the tag might be attached to the peptide. Specifically, the presence of y_9 and y_{10} ions in the spectrum with and without the thymine attached point to the existence of isomers that were unresolved by the HPLC separation. The dT_2 tag is probably attached to one or the other of the phenylalanine residues since aromatic amino acids tend to be in close proximity with nucleobases in base-stacking interactions in DNA-binding proteins [21]. The uncrosslinked version of this peptide was also observed in the chromatogram; it eluted 4 minutes after the crosslinked peptide and was identified with a Mascot ion score of 82.

Figure 3.4. Low energy ESI tandem mass spectra corresponding to doubly-charged crosslinked peptides from CspC: (A) precursor with *m/z* 911.87; (B) precursor with *m/z* 1,146.94. Fragments marked with a tilda (~) in these spectra correspond to those with a thymine attached. Fragments marked with an asterisk (*) are those that have the dT$_2$ tag attached.

The low energy MS/MS spectrum shown in Figure 3.4B corresponds to the high energy spectrum seen in Figure 3.3B. The presence of the extensive y-series in the low energy case, along with a short b-series (b_2 to b_5), led to the identification of this crosslinked peptide with a Mascot ion score of 34. Since this is a larger peptide than the one in Figure 3.4A, doubly charged fragments were observed, making the spectrum a bit more visually complex. Evidence for both peptide and nucleotide fragmentation are observed in this spectrum the same as they were in the previous spectra; however, the 2-5 fragments are mainly doubly charged in this case. As in the preceding example, this spectrum exhibits ion signals corresponding to isomers that prevent determining the position of the crosslink. The dT_2 tag's attachment can be localized to the FVHF region; in this case, the thymine is most likely crosslinked to either the phenylalanine or the histidine residues [21]. This peptide, in its uncrosslinked form, appeared in the same chromatogram 6 minutes after the crosslinked species, and it was identified with a Mascot ion score of 84.

Since the loss of dT_2 from both of these crosslinked peptides was observed with low energy MS/MS, it was decided that neutral loss scanning could be used to find more crosslinked peptides. Usually performed on phosphorylated peptides, this method uses an increase in collision energy to induce the neutral loss of a modification, and then MS/MS is performed only on those species displaying that signature loss [20]. Setting the neutral loss to 626.1 Da, the mass of the dT_2 tag, led to the identification of some of the same crosslinked peptides as were found

without scanning for the neutral loss, plus one additional crosslinked species. This is the first successful use of neutral loss scanning for the analysis of peptides crosslinked to oligonucleotides.

A summary of all the crosslinked peptides found using all of the methods employed in this study can be seen in Table 3.1. All of the crosslinks observed had a dT_2 tag attached, except for one peptide that carried a dT_3 tag. Larger tags were observed if nuclease P1 was not given enough time to completely digest the oligonucleotides (data not shown). Both tryptic (DVFVHFSAIQGNGFK) and partially tryptic (DVFVHFSAIQGN) versions of the same peptide from CspC were seen crosslinked to dT_2. Partially tryptic crosslinked peptides have been observed in previous studies that involved photochemical crosslinking and mass spectrometry [14,16].

Protein Observed	Crosslinked Peptides Observed
Cold Shock Protein C	GFGFITPADGSK x dT_2
	GFGFITPADGSK x dT_3
	DVFVHFSAIQGNGFK x dT_2
	DVFVHFSAIQGN x dT_2
Cold Shock Protein A	GFGFITPDDGSK x dT_2
50S Ribosomal Protein L24	VGFRFEDGK x dT_2

Table 3.1. Summary of crosslinked peptides observed from *E. coli* DNA-binding proteins. All of these crosslinked peptides were observed with both the ESI Q-TOF and MALDI TOF/TOF instruments.

3.4. Discussion

3.4.1. Cold Shock Proteins

Most of the crosslinked peptides observed in this study were from proteins in the cold shock protein (Csp) family. CspA, the major cold shock protein from *E. coli*, is induced by a decrease in temperature [23], but can be expressed at 37°C under exponential growth conditions [24]. The proteins in the Csp family, which share the nucleic acid binding sequence motifs RNP1 and RNP2 with the Y-box proteins, are known to bind single-stranded (ss) RNA and DNA with micromolar affinity [25]. CspA has been shown to bind polypyrimidine ssDNA with especially high affinity [26], so the observation of this protein interacting with dT_{20} is not surprising. Based on experiments with a polythymidine template, CspA has been shown to require 6-7 bases per molecule of protein for binding [25]; therefore, dT_{20} is definitely large enough to serve as a substrate for this protein.

There is a great deal of homology between CspA and CspC, which is expressed at 37°C and is not induced by a decrease in temperature [27]. As illustrated in Figure 3.5, the two proteins are 68% homologous, with the two crosslinked peptides from CspC being 85% homologous to the corresponding peptides from CspA. All of the amino acids that were found to be crosslinked in CspC are conserved in CspA. There are currently no published structures available for CspC, and there are no published structures for either protein

interacting with DNA. Three-dimensional structures of CspA have, however, been solved through NMR [9] and crystallographic studies [10]; these show that CspA is a Greek-key ß-barrel protein. Several basic and aromatic amino acid residues in CspA, which are conserved in all cold shock proteins, were found to be arranged on one side of the protein forming a potential nucleic acid-binding surface. Looking at the crystal structure from CspA with the crosslinked amino acids highlighted (Figure 3.6), the putative DNA-binding site is apparent. Site-directed mutagenesis of F_{18}, F_{20}, and F_{31} from CspA to serines has been shown to affect the binding of ssDNA [28]. The mutation of these same phenylalanine residues to alanines, along with substitution of H_{33} with glutamine, eliminated the retardation of ssDNA in a gel-shift assay involving CspB [29]. The residues found to be essential for DNA binding in the site-directed mutagenesis studies are the same residues that were found to be photochemically crosslinked in this study.

Cold Shock Protein A

Cold Shock Protein C
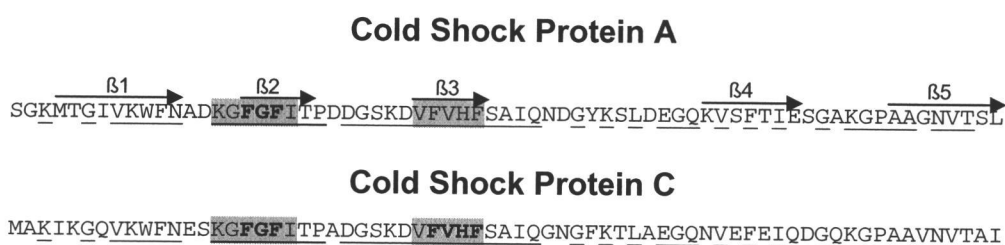
Figure 3.5. Sequences for *E. coli* CspA and CspC. The homologous sequences are underlined, and the crosslinked amino acids are in bold. The highlighted sequences correspond to the RNP1 and RNP2 sequence motifs commonly found in ssRNA-binding proteins. The five ß-strands observed in CspA are denoted by arrows [9].
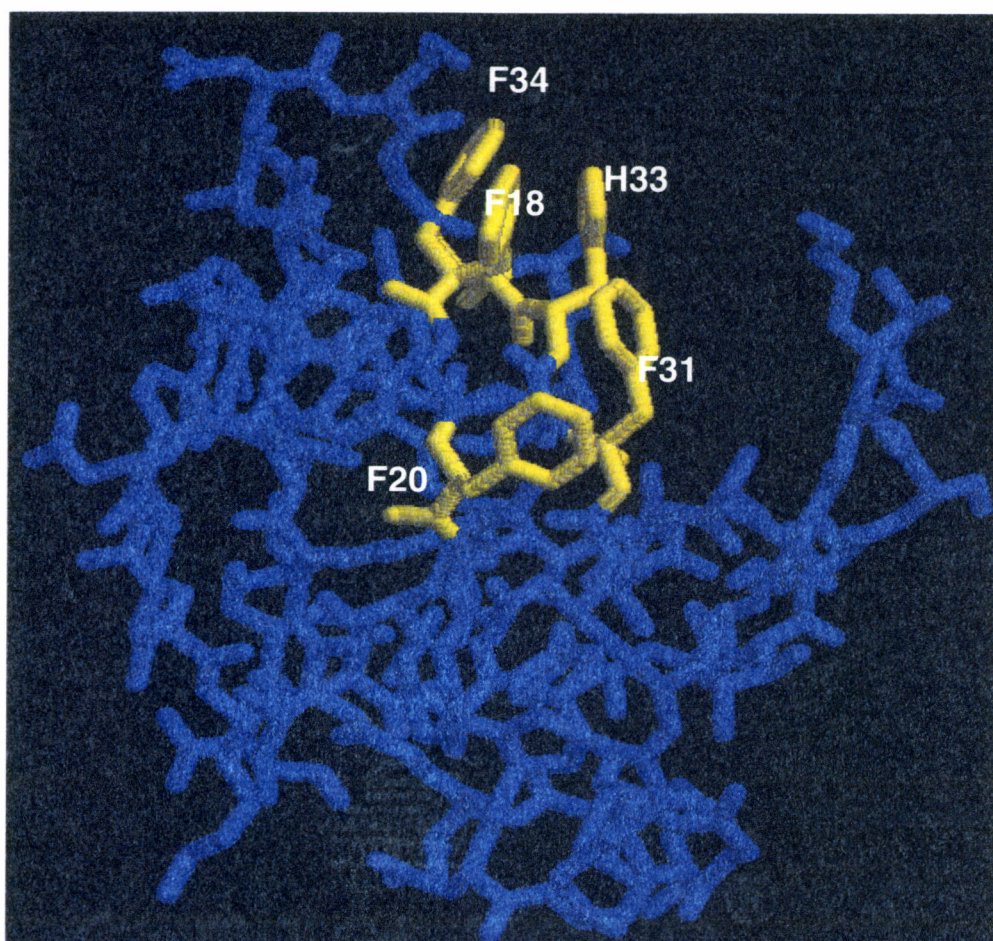
Figure 3.6. Crystal structure of CspA [10], with the aromatic amino acids found to be crosslinked in peptides from CspC highlighted in yellow. This drawing was created using the Rasmol program [32].

3.4.2. 50S Ribosomal Protein L24

One peptide from 50S ribosomal protein L24 was found to be crosslinked in this study (Table 3.1). The crosslinked peptide (VGFRFEDGK) contains basic and aromatic amino acids, so it is logical for this peptide to be involved in binding to nucleotides. Unfortunately, no high resolution structures of this protein from *E. coli* are available to compare with the crosslinked peptide data produced in this study. A low resolution electron microscopy experiment, however, shows that the FRF region of the crosslinked peptide from this protein is in close proximity to the ssRNA in the *E. coli* ribosome [30], but the low resolution of the microscopy prevented the positions of the amino acid side chains from being determined. Previous studies have shown that L24 binds to ssRNA [31], so it is possible that this protein was interacting with the ssDNA in the crosslinking experiment. The results from the present study may be used to predict the amino acids that are in contact with ssDNA in the binding site of this protein.

3.5. Conclusions

Both high and low energy CID were performed on crosslinked peptides in this study. For peptides attached to oligonucleotide tags, it appears that the high energy CID results mainly in the fragmentation of the oligonucleotide tag, which is useful for identifying the oligonucleotide but not for identifying the peptide. To sequence the peptides that are crosslinked to oligonucleotides, it appears that low-energy CID is most effective. Other studies comparing low and high energy CID on a glutathione-modified peptide found that the high energy CID in the TOF/TOF was most effective at breaking apart the peptide backbone, while low energy CID primarily yielded a loss of glutamate from the modification [33]. It appears that both types of CID are necessary to obtain as much information as possible from modified peptides in proteomic studies.

The results of this study demonstrate that photochemical crosslinking and mass spectrometry can be used to predict the three-dimensional structures of proteins with known structures (CspA) and somewhat unknown structures (CspC and L24) without isolation of the individual proteins. With the mass spectrometric tools available today, modified and unmodified peptides from mixtures of proteins can be easily identified. Adding covalent modifications to proteins to predict their structure does not severely complicate the analysis of mass spectral data, given that the mass of the modification is known. The procedure used in this study allows for the analysis of more than one protein at a time in order to determine the

function and structure of proteins in a proteome. This capability is necessary if all of the proteins in a proteome are to be eventually studied.

Separation of proteins based on function and covalent modification of binding sites could be used to determine the function and structure of many different classes of proteins other than DNA-binding proteins. For example, a recent study involved the reaction of recombinantly expressed and purified rat protein disulfide isomerase (PDI) with a chemically reactive version of glutathione to determine which cysteine in the active site was modified by a toxic environmental contaminant [33]. The same results could be obtained using a protocol like that used in this study. Glutathione-binding proteins, like PDI, could be separated from the rest of the rat proteome using an affinity column with glutathione attached. The glutathione-binding proteins could be eluted and then modified with the reactive moiety to block the active sites on the proteins. The affinity column could then be used to separate the unreacted proteins from the modified proteins, which would then be ready for proteolytic digestion and tandem mass spectrometry. This same protocol could be used for identifying and determining the structures of drug targets, AMP-binding proteins, or heme-binding proteins, as long as an affinity column and a reactive substrate that could covalently block the binding site were available.

3.6. References

1. O'Farrell, P.H. (1975) "High Resolution Two-Dimensional Electrophoresis of Proteins," *J Biol Chem*, 250: 4007-4021.

2. Champion, M.M., Campbell, C.S., Siegele, D.A., Russell, D.H., and Hu, J.C. (2002) "Proteome Analysis of *Escherichia coli* K-12 by Two-Dimensional Native-State Chromatography and MALDI-MS," Poster, *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics,* Orlando.

3. Wolters, D.A., Washburn, M.P., and Yates, J.R., III (2001) "An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics," *Anal Chem,* 73: 5683-5690.

4. Griffin, T.J. and Aebersold, R. (2001) "Advances in Proteome Analysis by Mass Spectrometry," *J Biol Chem*, 276: 45497-45500.

5. Liu, Y., Patricelli, M.P., and Cravatt, B.F. (1999) "Activity-Based Protein Profiling: The Serine Hydrolases," *Proc Natl Acad Sci USA*, 96: 14694-14699.

6. Bane, T.K., Leblanc, J.F., Lee, T.D., and Riggs, A.D. (2001) "Analysis of DNA Binding Proteins by Mass Spectrometry," Poster, *Proceedings of the 49th ASMS Conference on Mass Spectrometry and Allied Topics*, Chicago.

7. Greenbaum, D., Medzihradszky, K.F., Burlingame, A., and Bogyo, M. (2000) "Epoxide Electrophiles as Activity-Dependent Cysteine Protease Profiling and Discovery Tools," *Chem Biol*, 7: 569-581.

8. Shefcheck, K.J., Yao, X., and Fenselau, C. (2002) "Evaluation of Heparin for Fractionation of Cytosolic Proteins in Cancer Cells," Poster, *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics,* Orlando.

9. Feng, W., Tejero, R., Zimmerman, D.E., Inouye, M., and Montelione, G.T. (1998) "Solution NMR Structure and Backbone Dynamics of the Major Cold-Shock Protein (CspA) from *Escherichia coli*: Evidence for Conformational Dynamics in the Single-Stranded RNA-Binding Site," *Biochem*, 37:10881-10896.

10. Schindelin, H., Jiang, W., Inouye, M., and Heinemann, U. (1994) "Crystal Structure of CspA, the Major Cold Shock Protein of *Escherichia coli*," *Proc Natl Acad Sci USA*, 91:5119-5123.

11. Janssen, D. (2003) "Making an Enzyme Structure Crystal Clear," *Proteomics & Genomics*, 3:16-24.

12. Yan, X., Zhang, H., Watson, J., Schimerlik, M.I., and Deinzer, M. (2002) "Hydrogen/Deuterium Exchange and Mass Spectrometric Analysis of a Protein Containing Multiple Disulfide Bonds: Solution Structure of Recombinant Macrophage Colony Stimulating Factor-Beta (rhM-CSFß)," *Protein Science*, 11:2113-2124.

13. Steen, H. and Jensen, O.N. (2002) "Analysis of Protein-Nucleic Acid Interactions by Photochemical Cross-linking and Mass Spectrometry," *Mass Spectrom Rev*, 21:163-182.

14. Gafken, P.R. (2000) "Characterization of UV-Crosslinked Protein-Nucleic Acid Interfaces by MALDI-MS and ESI-MS/MS," *PhD thesis, Oregon State University*.

15. Bennett, S.E., Jensen, O.N., Barofsky, D.F., and Mosbaugh, D.W. (1994) "UV-Catalyzed Cross-linking of *Escherichia coli* Uracil-DNA Glycosylase to DNA," *J Biol Chem*, 269:21870-21879.

16. Doneanu, C.E. (2002) "Mass Spectrometric Analysis of UV-Crosslinked Protein-Nucleic Acid Complexes," *PhD thesis, Oregon State University*.

17. Steen, H., Petersen, J., Mann, M., and Jensen, O.N. (2001) "Mass Spectrometric Analysis of a UV-Crosslinked Protein-DNA Complex: Tryptophans 54 and 88 of *E. coli* SSB Cross-link to DNA," *Protein Science*, 10:1989-2001.

18. Stapels, M.D. (2003) *this thesis*, Chapter 2.

19. Poseqitz, M.C. and Tempst, P. (1999) "Immobilized Gallium (III) Affinity Chromatography of Phosphopeptides," *Anal Chem*, 71:2883-2892.

20. Bateman, R.H., Carruthers, R., Hoyes, J.B., Jones, C., Langridge, J.I., Millar, A., and Vissers, J.P.C. (2002) "A Novel Precursor Ion Discovery Method on a Hybrid Quadrupole Orthogonal Acceleration Time-of-Flight (Q-TOF) Mass Spectrometer for Studying Protein Phosphorylation," *J Am Soc Mass Spectrom*, 13:792-803.

21. McLuckey, S., Van Berkel, G.J., and Glish, G.L. (1992) "Tandem Mass Spectrometry of Small Multiply Charged Oligonucleotides," *J Am Soc Mass Spectrom*, 3:60-70.

22. Ollis, D.L. and White, S.W. (1987) "Structural Basis of Protein-Nucleic Acid Interactions," *Chem Rev*, 87:981-995.

23. Goldstein, J., Pollit, N.S., and Inouye, M. (1990) "Major Cold Shock Protein of *Escherichia coli*," *Proc Natl Acad Sci USA*, 87:283-287.

24. Brandi, A., Spurio, R., Gualerzi, C.O., and Pon, C.L. (1999) "Massive Presence of the *Escherichia coli* 'Major Cold-Shock Protein' CspA under Non-Stress Conditions," *EMBO J*, 18:1653-1659.

25. Ermolenko, D.N. and Makhatadze, G.I. (2002) "Bacterial Cold-Shock Proteins," *Cell Mol Life Sci*, 59:1902-1913.

26. Lopez, M.M. and Makhatadze, G.I. (2000) "Major Cold Shock Proteins, CspA from *Escherichia coli* and CspB from *Bacillus subtilis*, Interact Differently with Single-Stranded DNA Templates," *Biochim Biophys Acta*, 1479:196-202.

27. Graumann, P.L. and Marahiel, M.A. (1999) "Cold Shock Proteins CspB and CspC are Major Stationary-Phase-Induced Proteins in *Bacillus subtilis*," *Arch Microbiol*, 171:135-138.

28. Hillier, B.J., Rodriguez, H.M., and Gregoret, L.M. (1998) "Coupling Protein Stability and Protein Function in *Escherichia coli* CspA," *Fold Des*, 3:87-93.

29. Schroder, K., Graumann, P., Schnuchel, A., Holak, T.A., and Marahiel, M.A. (1995) "Mutational Analysis of the Putative Nucleic Acid-Binding Surface of the Cold-Shock Domain, CspB, Revealed an Essential Role of Aromatic and Basic Residues in Binding of Single-Stranded DNA Containing the Y-box Motif," *Mol Microbiol*, 16:699-708.

30. Gao, H.X., Sengupta, J., Valle, M.; Korostelev, A., Eswar, N., Stagg, S.M., Van Roey, P., Agrawal, R.K., Harvey, S.C., Sali, A., Chapman, M.S., and Frank, J. (2003) "Study of the Structural Dynamics of the *E. coli* 70S Ribosome Using Real-Space Refinement," *Cell*, 113:789-801.

31. Stelzl, U. and Nierhaus, K.H. (2001) "A Short Fragment of 23S rRNA Containing the Binding Sites for Two Ribosomal Proteins, L24 and L4, is a Key Element for rRNA Folding During Early Assembly," *RNA*, 7:598-609.

32. Sayle, R.A. and Milner-White, E.J. (1995) "RASMOL: Biomolecular Graphics for All," *Trends Biochem Sci*, 20:374-377.

33. Kaetzel, R.S. (2003) "Proteomic Approach to 1,2-Dichloroethane Bioactivation and Reaction with Redox-Active Protein Disulfide Isomerase," *PhD thesis, Oregon State University*.

4. COMPLEMENTARY USE OF MALDI AND ESI FOR THE HPLC-MS/MS ANALYSIS OF DNA-BINDING PROTEINS

Martha D. Stapels and Douglas F. Barofsky

4.1. Introduction

Since the advent of mass spectrometric proteomic studies, electrospray ionization (ESI) has mainly been used to sequence peptides separated by HPLC [1], while matrix-assisted laser desorption/ionization (MALDI) has been used to mass fingerprint mixtures of peptides [2,3]. This division of effort came about because ESI can be easily coupled on-line to receive the eluate from an HPLC column, and MALDI can readily be used to analyze mixtures of peptides without resorting to separation. Separation of peptides prior to MALDI, however, reduces ion suppression [4] and leads to higher signal to noise ratios and greater sensitivity [5]. On-line coupling of MALDI to HPLC has been attempted with continuous flow [6], aerosol [7,8], or rotating ball inlet [9] interfaces, but none of these approaches have been widely used due to carryover effects, high sample consumption rates, or low mass resolution [5]. HPLC separations have also been captured onto MALDI targets for off-line analysis using various instruments that deposit a series of spots or a continuous streak of eluate [5,10]. Off-line analysis of an HPLC separation is advantageous since the sampling of peaks from the eluate is not dependent upon the cycling time of the mass spectrometer. In the present study, a robotic device (MALDIprep sample collection module) from Waters (Milford, MA) was used to fractionate an HPLC separation into discrete spots on a MALDI sample plate. This sample collection module uses a heated capillary nebulizer to desolvate the column eluate and then spray it onto a MALDI

plate [11,12]. Since the eluate is desolvated during deposition onto the plate, column flow rates from 200 nL/min to 50 μL/min can be used with this instrument [10].

It has been previously noted that ESI and MALDI are complementary ionization techniques that, when used together, lead to the identification of more peptides than can be identified by either technique alone [13-18]. A recent study by Bodnar *et al.* exploited the complementary nature of LC MALDI and LC ESI to yield improved coverage in the identification of ribosomal proteins. In that study, 51 proteins were found, with 8 proteins unique to ESI and 11 proteins unique to MALDI [18]. Other studies have demonstrated the tendency for MALDI to ionize basic residues preferentially [19-24] and for ESI to favor hydrophobic amino acids [1]. Medzihradszky *et al.* described this complementary nature in a study that analyzed an HPLC fraction containing peptides created from an in-gel digestion of an SDS-PAGE gel band. In that study, nine peptides were identified with MALDI, whereas fifteen peptides were identified with ESI. The six peptides unique to the ESI experiment either had ended with a lysine residue or contained no basic amino acid residues at all [16]. No large-scale proteomic study that examines the source of the complementarity of the two ionization methods has been published to date.

In the present study, the complementary nature of ESI and MALDI is evaluated for a large set of peptides acquired on ESI and MALDI instruments. For both instruments, HPLC was used to separate the peptides prior to mass analysis. The ESI system used an on-line separation, and the MALDI system was coupled to

HPLC off-line. This work illustrates the benefit of using different ionization techniques as well as different mass analyzers for proteomic studies.

4.2. Experimental Section

4.2.1. Materials

Trifluoroacetic acid (TFA), formic acid (FA), a-cyano-4-hydroxycinnamic acid (HCCA), ethylenediaminetetraacetic acid (EDTA), sodium chloride (NaCl), ammonium citrate, single-stranded calf thymus DNA cellulose, lysozyme, yeast extract, tryptone, and Trizma (Tris-base) were purchased from Sigma Chemical Co. (St Louis, MO). Lyophilized sequencing-grade trypsin was purchased from Promega (Madison, WI). HPLC grade acetonitrile (ACN) was supplied by Fisher Scientific (Pittsburgh, PA). For general use, water was generated with a Milli-Q Ultrapure water purification system (Millipore Corp. Bedford, MA); for HPLC, Burdick and Jackson water was purchased from Honeywell International, Inc. (Muskegon, MI). Ultrafree-0.5 centrifugal filter devices (molecular weight cut-off 5,000) were purchased from Millipore Corp. Applied Biosystems supplied the 4700 Proteomics Analyzer Calibration Mixture (Framingham, MA).

4.2.2. Growth and Lysis of *E. coli*

Escherichia coli Strain BL21 (*E. coli* B F⁻ *dcm ompT hsdS*(r₈⁻ m₈)*gal*) from Stratagene (La Jolla, CA) was grown in LB media (1% tryptone, 0.5% yeast extract, and 1% NaCl) at 37°C under aerobic conditions. Cells were frozen in liquid nitrogen, resuspended in a buffer comprising 100 mM Tris (pH 7) and 1 mM EDTA, lysed ultrasonically after the addition of lysozyme, and centrifuged to remove insoluble material. Lysates were stored at -20°C until further use.

4.2.3. Sample Preparation

Cell lysate (20 mL) was mixed at room temperature with 5 g of single-stranded DNA cellulose and allowed to equilibrate for 15 min. Centrifugation was used to separate the cellulose particles from the rest of the lysate, which contained non-binding proteins. The non-binding proteins were rinsed off the cellulose particles with the addition of 20 mL of DAB buffer (20 mM Tris pH 7.4 with 1 mM EDTA). This step was repeated three times, the non-binding proteins being decanted off and collected after each rinse. The proteins that bound to the DNA were released in fractions with increasing concentrations of sodium chloride (0.05 M, 0.2 M, 0.4 M, 0.6 M, 1.0 M). Each DNA-binding fraction was obtained by mixing the cellulose with 25 mL of the salt solution, decanting, collecting the solution, repeating the first three steps, and then pooling the two extracts. Each fraction was concentrated and desalted using an Ultrafree-0.5 centrifugal filter device with a molecular weight cut-off of 5,000 Da. Total protein concentration in

each fraction after this point was estimated to be 0.2 pmol/μL based on the intensity of stained protein bands in a polyacrylamide gel (data not shown). The desalting step (facilitated by adding 0.5 mL of fresh DAB buffer to the filter) was repeated three times to ensure that the sodium chloride was removed from each of the samples prior to tryptic digestion. Each sample was digested in-solution with the addition of 2 μL of 0.5 μg/μL trypsin (in 0.1 M AA) to 20 μL of sample. The samples were vortexed, centrifuged, and placed on a 37°C water bath overnight to allow complete digestion to occur.

4.2.4. Chromatography and Electrospray Mass Spectrometry

For each chromatography experiment, 15 μL of sample was mixed with 15 μL of solvent A (0.1% FA, 0.005% TFA, and 3% ACN in H_2O), and 5 μL of this solution was injected. Every sample was run three times on each instrument. Solvent B contained 0.1% FA and 0.005% TFA in 80% ACN. A 5 mm by 0.32 mm C_{18} trap from LC Packing (Sunnyvale, CA) and a 15 cm long, 75 μm inner diameter PicoFrit column from New Objective (Woburn, MA) packed in-house with Jupiter C_{18} from Phenomenex (Torrance, CA) were used for the ESI experiments. The LC conditions started with 3% B for 5 min to wash the sample, followed by a gradient up to 30% B over 50 min, to 50% B at 60 min, to 70% B at 65 min, and held at 90% B from 72 to 78 min. A Waters CapLC system with a flow rate estimated to be 300 nL/min was used to deliver solvent. The mass spectrometer used for ESI-MS/MS was a quadrupole time-of-flight (Q-TOF)

Global Ultima system from Micromass (Micromass, Ltd., Manchester, UK) operated with a spray voltage of 3.5 kV. Data-dependent MS/MS was generate using a 0.5 sec MS survey scan and 2.5 sec MS/MS scans on the three most abundant peaks found in the survey scan. The CID energy was between 25 and 65 eV depending on the mass and charge state of the precursor ion. Mass spectra were calibrated using fragment ions generated from MS/MS of Glu-fibrinopeptide B (MW 1570.68).

4.2.5. Chromatography and MALDI Mass Spectrometry

A Symmetry300 5 μm C_{18} trap and a 150 mm long x 0.32 mm inner diameter Symmetry column packed with 5 μm C_{18} particles, both from Waters, were used for the MALDI experiments. Solvent A and solvent B both contained 0.1% TFA, with A including 1% ACN and B containing 99% ACN. The gradient used on this system was exactly the same as the one used in the ESI experiments. A controlled flow rate of 3 μL/min was delivered by a separate Waters CapLC system. The eluate from the column was mixed with 1 μL/min of 0.6 mg/mL HCCA and 0.08 mg/mL ammonium citrate in 50:50 ACN:H_2O containing 0.1% TFA. Ammonium citrate was added to the matrix solution to reduce the intensity of matrix peaks in MALDI mass spectra. This solution flowed via a 75 μm capillary at a combined rate of 4 μL/min into a Waters MALDIprep sprayer installed in the authors' laboratory courtesy of the Life Sciences R & D Laboratory, Waters Corporation. A spotting time of 0.75 min per spot, a nitrogen

flow rate of 7 psi, and a temperature gradient of -2°C/7 min (65°C to 45°C over 20 to 90 min) was used for each run. MALDI-MS and -MS/MS were performed in an automated (batch-mode) fashion on an Applied Biosystems 4700 Proteomics Analyzer with time-of-flight/time-of-flight (TOF/TOF) ion optics (Applied Biosystems, Inc., Framingham, MA). Data were acquired in the MALDI reflector mode using five spots of the internal calibration standard (ABI4700 Calibration Mixture). The plate file was updated to calibrate each run from each plate separately. Mass spectra were obtained from each sample spot using 500 shots per spectrum. Nine tandem mass spectra were then obtained from each spot; peaks with a signal to noise (S/N) ratio greater than 20 were chosen by the instrument. If more than nine peaks were found with a S/N ratio greater than 20, the instrument performed MS/MS on the nine most intense peaks starting with the least intense of these nine and progressing in order of increasing signal strength to the most intense; each MS/MS comprised 1,500 shots. Tandem mass spectra were acquired by accelerating the precursor ions to 8 keV, selecting them with the timed gate set to a window of 3 Da, and performing collision induced dissociation (CID) at 1 keV. Gas pressure (air) in the CID cell was set at 0.2 μTorr. Fragment ions were accelerated to 14 keV before entering the reflector.

4.2.6. Data Analysis

Mascot [25] from Matrix Science Ltd. (London, UK) was used to search all of the tandem mass spectra. For the data obtained on the Q-TOF, files appropriate

for Mascot (pkl files) were created using the Masslynx software from Waters with a processing macro (JustPkl) that smoothes, centroids, and assesses the quality of data. The MaxEnt3 function, which also converts the isotopic envelope to a singly charged ^{12}C-peak, was used in a few instances for comparison. GPS Explorer software from Applied Biosystems was used to create and search files with Mascot for the data obtained on the TOF/TOF. Most of the Mascot search parameters were held constant for all of the data generated in this study. These included the variable modifications (protein N-acetylation, oxidation of methionine, and pyro-glu of N-terminal glutamine and glutamic acid), cleavage by trypsin, up to three missed cleavages, and taxonomic specification (*E. coli*). Some search parameters unique to each instrument were used. In the case of the Q-TOF, the peptide and fragment mass tolerances were set to 0.5 Da and ESI Q-TOF was chosen as the instrument type. With the TOF/TOF data, the chosen instrument type was MALDI TOF/TOF, the peptide mass tolerance was set to 0.5 Da, and the fragment mass tolerance was set to 0.8 Da. In order for a Mascot result to be accepted with either the ESI or MALDI data, the individual ion scores had to indicate that the spectrum obtained experimentally was identical to that of the theoretical peptide matched to that spectrum. If a Mascot ion score indicated homology rather than identity, the spectrum was visually inspected to determine if it was to be accepted. To calculate theoretical isoelectric point (pI), aliphatic index, and hydrophobicity, the ProtParam tool from the Expasy molecular biology server (http://us.expasy.org/tools/protparam.html) was used online.

4.3. Results and Discussion

In order to compare the two LC systems, chromatograms from each were studied. Figure 4.1 shows a base peak chromatogram of one tryptic digest mixture of one fraction of DNA-binding proteins run on the ESI Q-TOF. Peaks that yielded quality tandem mass spectra appear in the chromatogram between 16 and 55 min. A base peak chromatogram from the same sample analyzed on the MALDI TOF/TOF can be seen in Figure 4.2. A time delay was used on this system, so that the MALDIprep did not collect samples until 20 minutes into the LC run. This was done so that sample would not be spotted onto the sample plate until a few minutes before peptides would start eluting and, therefore, no space on the plate would be wasted. In this particular chromatogram, peaks with high-quality tandem mass spectra were observed between 38 and 82 minutes. Both LC systems, therefore, eluted peptides over a span of about 40 minutes, with the Q-TOF collecting mass spectral data in real time and the TOF/TOF collecting data after the LC eluate was collected onto a MALDI plate. Disregarding the absolute time scales, the visual appearance of the two chromatograms is similar.

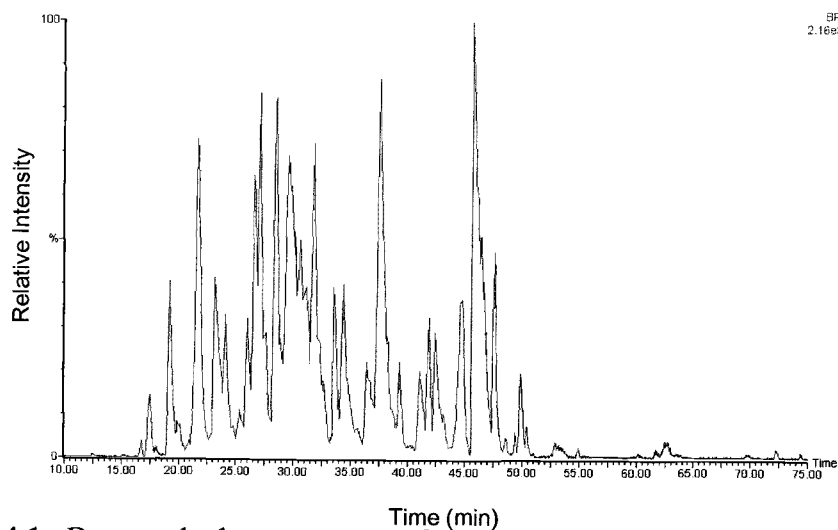**Figure 4.1.** Base peak chromatogram of a tryptic digest of the 0.6 M NaCl fraction of DNA-binding proteins recorded on-line with the Q-TOF.
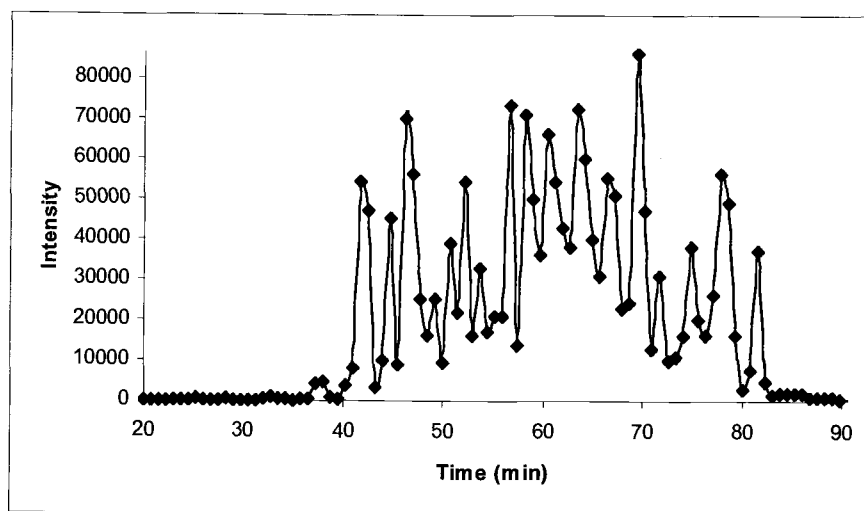

Figure 4.2. Base peak chromatogram of a tryptic digest of the 0.6 M NaCl fraction of DNA-binding proteins. This sample was sprayed onto a MALDI plate with the MALDIprep device and then analyzed on the MALDI TOF/TOF. Each data point corresponds to a separate spot on the plate.

There were a couple of additional dissimilarities between the two chromatographic systems that require comment. The first and most obvious was the vastly different column sizes and, therefore, flow rates used. The column used for the Q-TOF has an inner diameter of 75 μm, while the column used for the TOF/TOF has an inner diameter of 320 μm. If both columns were tested on an ESI instrument, the 75 μm column would show higher sensitivity, greater chromatographic resolution, and lead to more peptide identifications. However, with the MALDIprep, the sensitivity is increased by the concentrating effect of depositing a large volume (3 μL) of sample onto each spot. This benefit is unique to the MALDIprep, since other commercially available LC-MALDI instruments do not concentrate the eluate prior to spotting [14]. In terms of chromatographic resolution, compounds that elute within the 45-second spotting time whether they are resolved or not will all be deposited onto the same spot. Since the MALDI TOF/TOF can perform tandem mass spectrometry on more than 10 peptides contained in a given spot, lack of chromatographic resolution is not a problem unless the sample is extremely complex. The second dissimilarity was the higher concentration of acid (0.1% TFA) used in the HPLC solvent in the MALDI experiment than used in the ESI experiment (0.005% TFA). Since the MALDIprep decouples the chromatography from the mass spectrometry, additives that improve the chromatography, but are not allowed on-line because they suppress ionization, can be used. There was also an operational difference between the two mass spectrometric experiments that is worth comment. The Q-

TOF was run with dynamic exclusion, wherein a peptide would only be selected for tandem mass spectrometry a limited number of times. The software used in this study for TOF/TOF analysis did not have this capability so peptides were repeatedly analyzed if they were found in more than one spot.
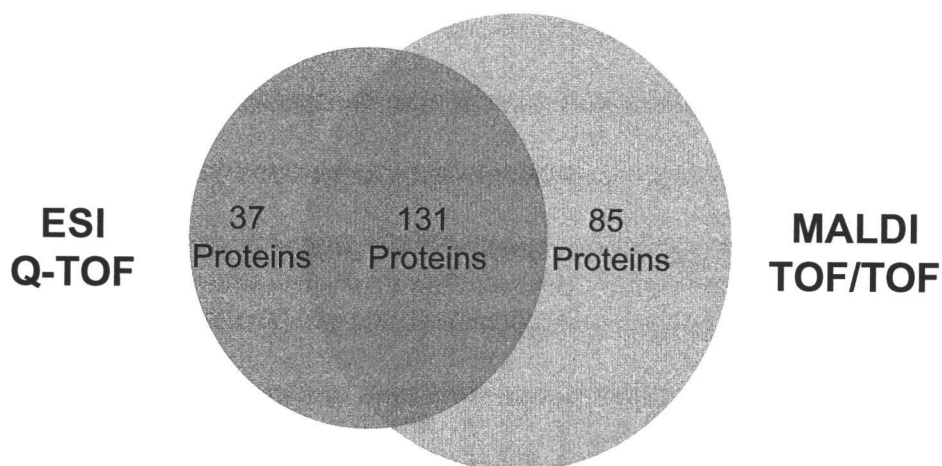
Figure 4.3. Venn diagram showing the number of proteins observed with the ESI Q-TOF and with the MALDI TOF/TOF. A total of 253 proteins were observed in this study.

After the Mascot searches were completed, the peptides and proteins with significant scores were summarized to compare the results of the two techniques. A Venn diagram comparing the proteins observed by ESI Q-TOF and MALDI TOF/TOF can be seen in Figure 4.3. A majority of the proteins observed (131 out of 253) were found by both techniques; consequently, confidence in those findings is very high. The MALDI TOF/TOF identified more proteins, probably due to the fact that, in this system, the chromatography is decoupled from the mass

spectrometry. In the ESI experiment, the Q-TOF was limited to the number of peptides it could process via MS/MS in real time, and therefore identified fewer proteins. The fact that the MALDI instrument enabled more peptide and protein identifications in our experiments should not be regarded as an indication of superiority since the ESI instrument would probably perform similarly with variable-flow chromatography in a "peak parking" experiment [26].

In Figure 4.4, a Venn diagram showing the number of peptides observed from each technique can be seen. For these identifications, a majority of peptides (402 out of 667), 130 by ESI and 272 by MALDI, were only identified by one technique or the other but not both. The peptide per protein average for the set of 253 proteins identified by both techniques is 2.64 (=667/253), higher than that of either the ESI experiment (2.35=395/168) or the MALDI experiment (2.49=537/216) alone. Thus, by combining the two techniques the peptide per protein average is increased by approximately 10%. This increase, which is a quantitative sign of the complementary nature of the two techniques, is beneficial since it leads to higher confidence in the identification of proteins.
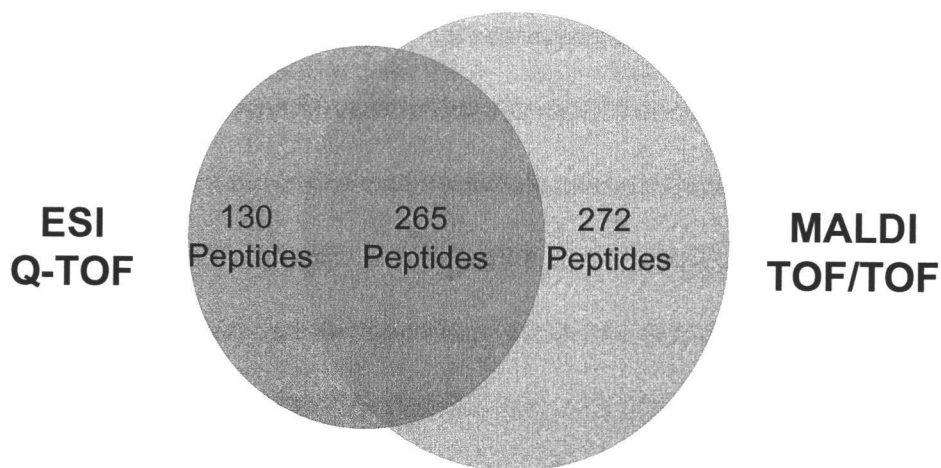
Figure 4.4. Venn diagram showing the number of peptides observed with the ESI Q-TOF and with the MALDI TOF/TOF. In this study, 667 peptides were identified with high confidence using tandem mass spectrometry.

The parameters listed in Table 4.1 were used to further compare the two techniques. The peptide mass-error was lower for the Q-TOF data; consequently, confidence in peptide identifications is inherently higher, even though Mascot does not take this into account with its scoring algorithm. The peptide mass-error in the TOF/TOF data was relatively high, but it could be significantly decreased by teeing in a mass standard with the matrix solution to provide an internal calibrant in each sample spot. The mass-error observed for the peptide fragments was also higher in the TOF/TOF data. For the Q-TOF, calibration was performed on the fragments of a standard peptide, Glu-fibrinopeptide B. The higher error in the MS/MS spectra produced by the TOF/TOF is undoubtedly due to the fact that this instrument was only calibrated in the MS mode. The Mascot search parameters were set to reflect this higher error in the TOF/TOF data. Setting the fragment

tolerance to 0.8 Da (the same tolerance used for the TOF/TOF peptide fragments) for the Mascot search of the Q-TOF data did not lead to higher ion scores or to more peptide identifications.

	Ave. Peptide Mass error	Low Peptide MW	High Peptide MW	Ave. Peptide MW	Ave. Ion Score	High Ion Score
ESI Q-TOF	0.049	1036.54	3569.89	1583.99	60.05	139
MALDI TOF/TOF	0.120	900.61	3546.62	1541.52	64.77	252

Table 4.1. Comparison of peptides observed with the ESI Q-TOF and MALDI TOF/TOF. Peptide mass error and peptide molecular weights are all given in Da.

The average molecular weight of peptides observed with the Q-TOF was slightly higher than that observed with the TOF/TOF. The mass to charge range for peptides subjected to MS/MS analysis in the 4700 TOF/TOF was set from 900 to 4000, and peptides were identified over almost that entire range. The Q-TOF collected tandem mass spectra for doubly and triply charged peptides between m/z 300 and 1680. The results of the present study are in contrast with those of two previous studies in which it was suggested that the TOF/TOF identifies higher molecular weight peptides [14, 17].

The average Mascot ion score for peptides obtained from the TOF/TOF data was almost 5 points higher than for that obtained from the Q-TOF data. Looking at the highest ion score in Table 4.1 for each instrument helps explain the difference between the results of the two techniques. In general, the TOF/TOF yields tandem mass spectra with more complete series of product ions, especially

for precursor peptides in the higher mass range. An illustration of this can be seen in Figure 4.5. Figure 4.5A shows the tandem mass spectrum obtained on the TOF/TOF for a peptide (MW 2396.3) from ribosomal protein L18. The entire b series plus most of the y series (y_4 to y_{21}) was identified in this spectrum; the ion score for this spectrum is an exceptional 252. Also present in this spectrum, though unlabelled and partially unidentified by Mascot, were many internal fragment ions. Smaller peptides generally yield more high energy fragments in the TOF/TOF, which can make it possible to differentiate isobaric residues such as leucine and isoleucine [27]. The tandem mass spectrum of the same peptide (triply charged) obtained on the ESI Q-TOF can be seen in Figure 4.5B. A great deal of the b-series (b_2 to b_{15}, except b_9) and the y series (y_1 to y_{16}, excluding y_3) can be seen in this spectrum. The Mascot ion score of 105 for this spectrum is excellent by most standards, but due to the lack of b_{16} to b_{22} and y_{17} to y_{21}, the score for this spectrum is notably lower than the one obtained on the TOF/TOF. Changing the data conversion file from JustPkl to the MaxEnt3 function for the Q-TOF data did not yield higher Mascot scores or more peptide identifications. Excluding the spectrum with the score of 252 from the TOF/TOF data did not change the average score significantly due to the presence of other large peptides with remarkably high scores.
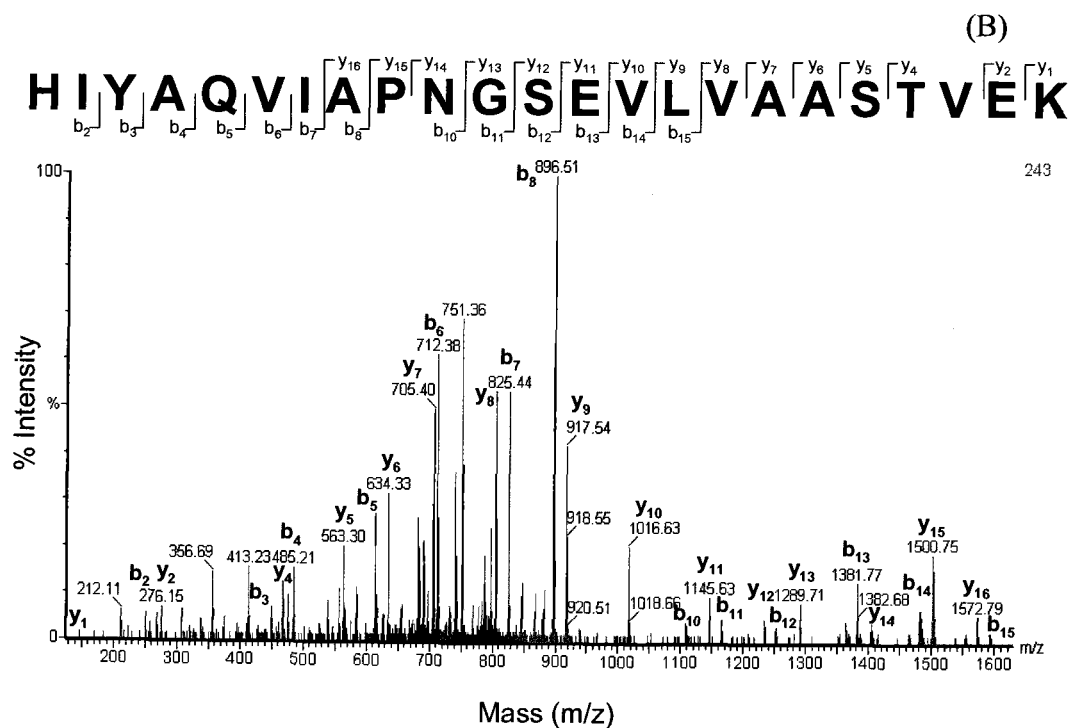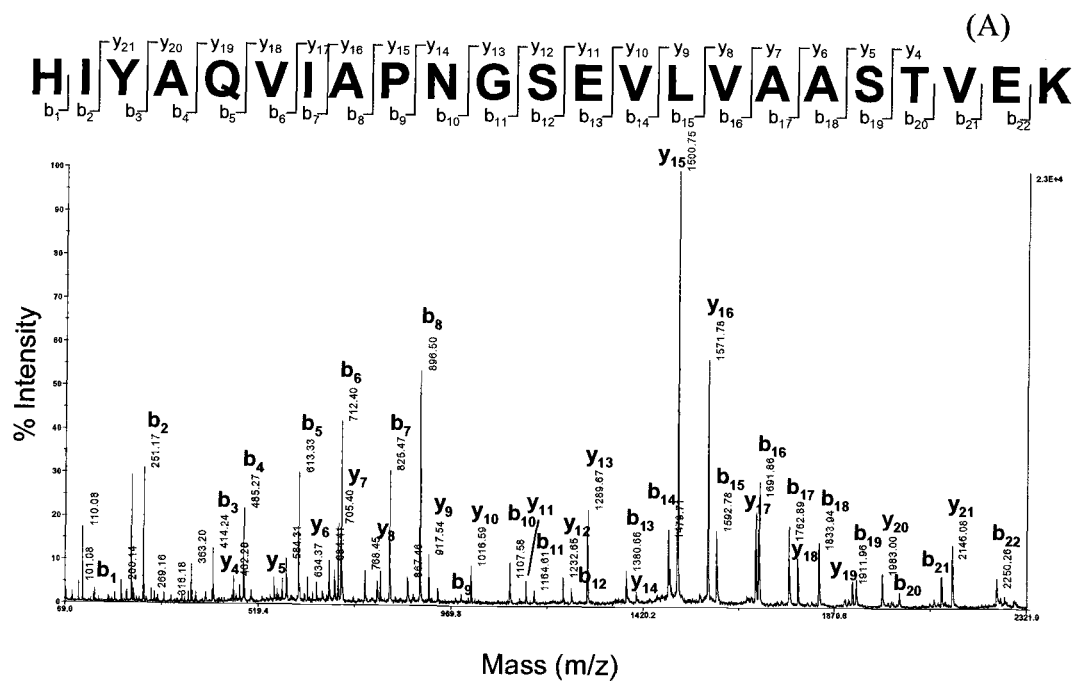
Figure 4.5. Tandem mass spectra of a peptide from 50S ribosomal protein L18.
(A) Spectrum produced from singly charged peptide by the MALDI TOF/TOF;
(B) spectrum produced from triply-charged peptide by the ESI Q-TOF.

Some other parameters used for comparing the peptides observed with ESI and MALDI in this study are listed in Table 4.2. As in previous studies [13,15,16], the ESI method identified more peptides that end in lysine than in arginine; this is reflected in the relatively high K/R ratio of 1.38. The MALDI technique identified more peptides with an arginine at the C-terminus; this is accounted for by the relatively low K/R ratio of 0.89. When only peptides that are unique to each ionization technique are considered (see Figure 4.4), the difference in the ratios is even more striking; the ESI method yielding a K/R ratio of 1.84 and the MALDI technique one of 0.61. This observation further emphasizes the complementary nature of ESI and MALDI in the analysis of tryptic peptides. Little distinction between the two ionization processes is observed in the isoelectric point (pI) of the peptides analyzed in this study; the ESI-observed peptides have a slightly more acidic isoelectric point. Others have noted that MALDI tends to preferentially ionize basic peptides [19-24], and the findings in this study strongly support this observation. Both the aliphatic index and hydrophobicity measure indicate a tendency for the ESI experiment to identify more hydrophobic peptides than the MALDI experiment. ESI is purported to ionize hydrophobic peptides efficiently because both the charges and the hydrophobic amino acids tend to occupy the outer surface of the droplets in the electrospray process [1]. Ironically, it has also been reported that MALDI preferentially ionizes hydrophobic amino acids due to their interaction with the matrix molecules [4,22]. This latter observation with MALDI was, however, made

with the dried-droplet approach to crystallization, which results in large crystals quite unlike the much smaller, finer crystals produced by the MALDIprep sprayer used in the present study.

	K/R Ratio	Theoretical pI	Aliphatic Index	Hydrophobicity
ESI Q-TOF	1.38	4.45	96.18	-0.121
ESI Q-TOF only	1.84	4.26	100.3	-0.13
MALDI TOF/TOF	0.89	4.67	87.99	-0.248
MALDI TOF/TOF only	0.61	4.78	81.27	-0.392

Table 4.2. Comparison of properties of peptides observed with the ESI Q-TOF and the MALDI TOF/TOF. The properties of peptides that were unique to the ESI and MALDI instruments are also listed. Aliphatic index refers to the fraction of a peptide that is occupied by residues with aliphatic side chains (A,V,I,L). The hydrophobicity was calculated according to the GRAVY (Grand Average of Hydropathicity) index by Kyte and Doolittle [28], where 4.5 is defined as most hydrophobic (I) and -4.5 is defined as most hydrophilic (R).
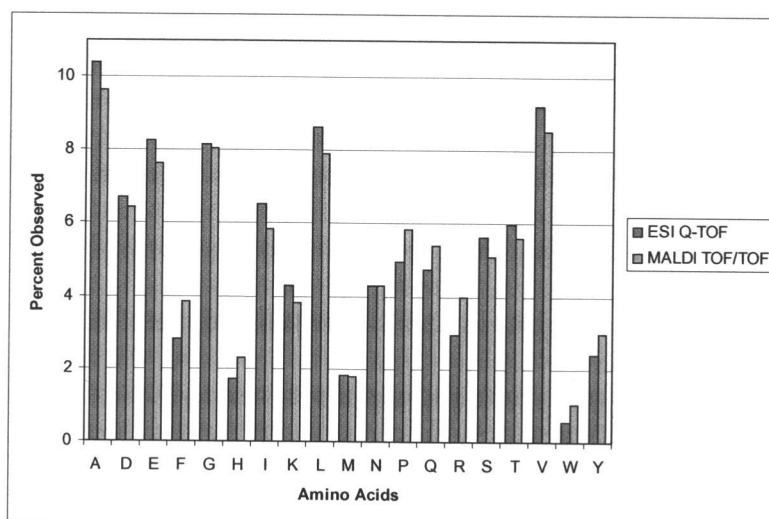

Figure 4.6. Amino acids observed by MALDI and ESI. Cysteine was not included in this chart; since the samples were not reduced or alkylated, only one cysteine was observed in the entire study.

To examine the types of peptides observed with each technique a bit further, the distribution of amino acids observed was also examined. The peptides that were identified in this study were made up of more than 9,000 individual amino acids. In Figure 4.6, the frequency (expressed as a percentage) with which each amino acid appeared respectively in ESI and MALDI mass spectra is shown. For example, less than 2% of all the amino acids observed by either ESI or MALDI were methionine. The known preference for lysine to be observed by ESI and for arginine to be observed by MALDI can be seen in this graph. Similar biases in favor of one ionization mode or the other can be seen for several of the other amino acids. The patterns among the amino acids become more apparent when the amino acids are grouped into classes as shown in Figure 4.7. As reported earlier, MALDI favors the ionization of basic and aromatic amino acids [19-24].
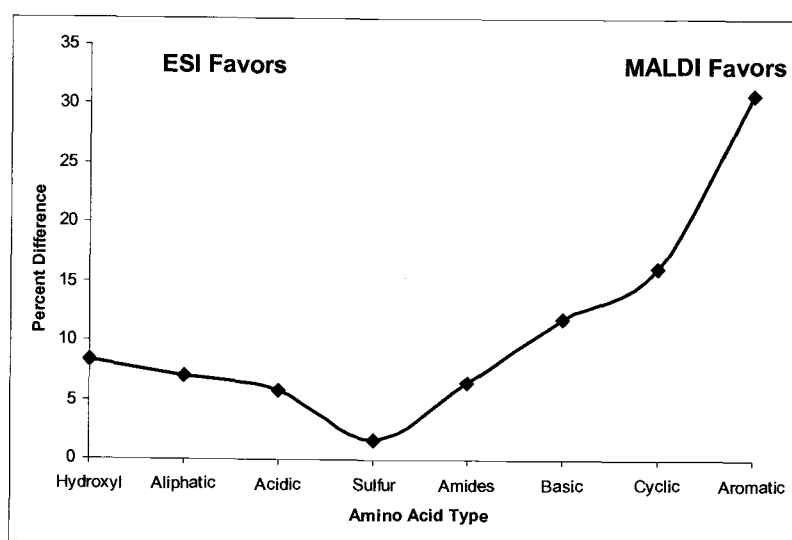

Figure 4.7. Amino acid types observed by ESI and MALDI. The amino acid type is plotted against the absolute value of the normalized percent difference in order to measure the preference of the ionization techniques for certain classes of amino acids.

It has been proposed that the tendency for aromatic amino acids to preferentially ionize via MALDI might be due to photoexcitation of those moieties during ionization [23,24]. In our study, MALDI also tended to favor the ionization of amide-containing and cyclic (proline) amino acids. The tendency for MALDI to ionize these amino acid types may be due to their relatively higher gas-phase basicities [29]. ESI led to the identification of greater numbers of peptides richer in residues with aliphatic, acidic, and hydroxyl groups. At least one previous report indicates the preference of ESI for hydrophobic amino acids [1], but there is no known explanation for its bias toward amino acids containing hydroxyl or acidic moieties. These differences in amino acid biases, which emerge from a large dataset, again emphasize the complementary nature of ESI and MALDI in identifying peptides.

4.4. Conclusions

Five fractions of DNA-binding proteins were digested and analyzed in triplicate by LC-ESI Q-TOF and LC-MALDI TOF/TOF in a study of the complementary nature of these techniques. This is one of the first direct comparisons of LC-MALDI and LC-ESI using the MALDIprep device. Running one sample twice on the ESI instrument or on the MALDI instrument increases the number of peptides identified on the average by 25%, whereas running the same

sample once on an ESI instrument and once on a MALDI instrument increases the number of peptides identified by 49% on the average. This study shows that the complementary nature of ESI and MALDI is due to the biases of the two ionization techniques for certain classes of amino acids. At present, most researchers use ESI in large-scale proteomic studies, because of the ease with which it couples to LC. Now that instruments are commercially available for automated LC-MALDI, MALDI may become more popular for use in proteomics. It is clear from this as well as other studies that using both ionization techniques can significantly increase the number of peptides observed in a large proteomic study; however, the present study indicates how the character of the sample can be used as a guide for choosing between ESI and MALDI when only one technique can be used. For example, MALDI seems to be more applicable to basic peptides such as those found in histones or to aromatic peptides found in RNA-binding proteins while ESI might be better employed for hydrophobic peptides like those found in membrane proteins or acidic peptides found in neurofilaments.

4.6. References

1. Cech, N.B. and Enke, C.G. (2000) "Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides," *Anal Chem*, 72:2717:2723.

2. Aebersold, R. (2003) "A Mass Spectrometric Journey into Protein and Proteome Research," *J Am Soc Mass Spectrom*, 14:685-695.

3. Mann, M., Schevchenko, A., Wilm, M., and Vorm, O. (1996) "Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels," *Anal Chem*, 68:850-858.

4. Kratzer, R., Eckerskorn, C., Karas, M., and Lottspeich, F. (1998) "Suppression Effects in Enzymatic Peptide Ladder Sequencing Using Ultraviolet-Matrix Assisted Laser Desorption/Ionization-Mass Spectrometry," *Electrophoresis*, 19:1910-1919.

5. Karger, B.L., Preisler, J., Hu, P., and Rejtar, T. (2000) "Capillary Electrophoresis-Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry Using a Vacuum Deposition Interface," *Anal Chem*, 72:4785-4795.

6. He, L., Li, L., and Lubman, D.M. (1995) "Continuous-Flow MALDI Mass Spectrometry Using an Ion Trap/Reflectron Time-of-Flight Detector," *Anal Chem*, 67:4127-4132.

7. Zhan, O., Gusev, A., and Hercules, D.M. (1999) "A Novel Interface for On-Line Coupling of Liquid Capillary Chromatography with Matrix-Assisted Laser Desorption/Ionization Detection," *Rapid Comm Mass Spectrom*, 13:2278-2283.

8. Murray, K.K. (1997) "Coupling Matrix-Assisted Laser Desorption/Ionization to Liquid Separation," *Mass Spectrom Rev*, 16:283-299.

9. Orsnes, H., Graf, T., Degn, H., and Murray, K.K. (2000) "A Rotating Ball Inlet for On-Line MALDI Mass Spectrometry," *Anal Chem*, 72:251-254.

10. Wall, D.B., Berger, S.J., Finch, J.W., Cohen, S.A., Richardson, K., Chapman, R., Drabble, D., Brown, J., and Gostick, D. (2002) "Continuous Sample Desorption from Reversed-Phase Liquid Chromatography to Tracks on a Matrix-Assisted Laser Desorption/Ionization Precoated Target for the Analysis of Protein Digests," *Electrophoresis*, 23:3193-3204.

11. Biemann, K. and Gagel, J.J., U.S.A. Patent 4,843,243, June 27, 1989.

12. Prevast, T.H. and Dwyer, J.L., U.S.A. Patent 5,772,964, June 30, 1998.

13. Yao, X., Heller, M., Mattou, H., and Menzel, C. (2003) "Trypsin Catalyzed ^{16}O-to-^{18}O Exchange for Comparative Proteomics: Tandem Mass Spectrometry Comparison Using MALDI-TOF, ESI-QTOF, and ESI-Ion Trap Mass Spectrometers," *J Am Soc Mass Spectrom*, 14:704-718.

14. Burlingame, A.L., Hansen, K.C., Schmitt-Ulms, G., Chalkley, R.J., Hirsch, J., and Baldwin, M.A. (2003) "Mass Spectrometric Analysis of Protein Mixtures at Low Levels Using Cleavable ^{13}C-ICAT and Multi-Dimensional Chromatography," *Mol Cell Proteomics*, 2:299-314.

15. Baldwin, M.A., Medzihradszky, K.F., Lock, C.M., Fisher, B., Settineri, T.A., and Burlingame, A.L. (2001) "Matrix-Assisted Laser Desorption/Ionization Coupled with Quadrupole/Orthogonal Acceleration Time-of-Flight Mass Spectrometry for Protein Discovery, Identification, and Structural Analysis," *Anal Chem*, 73:1707-1720.

16. Medzihradszky, K.F., Leffler, H., Balwin, M.A., and Burlingame, A.L. (2001) "Protein Identification by In-Gel Digestion, High-Performance Liquid Chromatography, and Mass Spectrometry: Peptide Analysis by Complementary Ionization Techniques," *J Am Soc Mass Spectrom*, 12:215-221.

17. Jonscher, K. (2003) "Proteomics: To Spray or Not to Spray," *Genomics & Proteomics*, 3:31-33.

18. Bodnar, W.M., Blackburn, R.K., Krise, J.M., and Moseley, M.A. (2003) "Exploiting the Complementary Nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for Increased Proteome Coverage," *J Am Soc Mass Spectrom*, 14:971-979.

19. Chen, C.H., Zhu, Y.F., Lee, K.L., Tang, K., Allman, S.L., and Taranenko, N.I. (1995) "Revisit of MALDI for Small Proteins," *Rapid Commun in Mass Spectrom*, 9:1315:1320.

20. Vertes, A., Olumee, Z., Sadeghi, M., and Tang, X. (1995) "Amino Acid Composition and Wavelength Effects in Matrix-assisted Laser Desorption/Ionization," *Rapid Commun in Mass Spectrom*, 9:744-752.

21. Jungblut, P.R., Krause, E., and Wenschuh, H. (1999) "The Dominance of Arginine-Containing Peptides in MALDI-Derived Tryptic Mass Fingerprints of Proteins," *Anal Chem.* 71:4160-4165.

22. Chait, B.T. and Cohen, S.C. (1996) "Influence of Matrix Solution Conditions on the MALDI-MS Analysis of Peptides and Proteins," *Anal Chem*, 68:31-37.

23. Burkitt, W.I., Giannakopulos, A.E., Sideridou, F., Bashir, S., and Derrick, P.J. (2003) "Discrimination Effects in MALDI-MS of Mixtures of Peptides-Analysis of the Proteome," *Aust J Chem*, 56:369-377.

24. Valero, M., Giralt, E., and Andreu, D. (1998) "An Evaluation of Some Structural Determinants for Peptide Desorption in MALDI-TOF Mass Spectrometry," in *Peptides 1996* (Eds Ramage, R.; Epton, R.), Mayflower Scientific, Kingswinford, 855-856.

25. Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999) "Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data," *Electrophoresis*, 20: 3551-3567.

26. Davis, T.D. and Lee, M.T. (1998) "Rapid Protein Identification Using a Microscale Electrospray LC/MS System on an Ion Trap Mass Spectrometer," *J Am Soc Mass Spectrom,* 9:194-201.

27. Medzihradszky, K.F., Campbell, J.M., Baldwin, M.A., Falick, A.M., Juhasz, P., Vestal, M.L., and Burlingame, A.L. (2000) "The Characteristics of Peptide Collision-Induced Dissociation Using a High-Performance MALDI-TOF/TOF Tandem Mass Spectrometer," *Anal Chem*, 72:552-558.

28. Kyte, J. and Doolittle, R.F. (1982) "A Simple Method for Displaying the Hydropathic Character of a Protein," *J Mol Biol*, 157:105-132.

29. Harrison, A.G. (1997) "The Gas-Phase Basicities and Proton Affinities of Amino Acids and Peptides," *Mass Spectrom Rev*, 16:201-217.

5. SUMMARY AND CONCLUSION

Mass spectrometry is quickly becoming the tool of choice in proteomic studies. In order to study protein expression, function, and structure, new ways to use this tool must be developed. The most innovative proteomic studies often employ established biochemical tools and make them applicable to mass spectrometric characterization. The following is a summary of the outcomes of the three studies involving mass spectrometric characterization of DNA-binding proteins presented in this thesis.

In the first study, it was found that the use of a function-based separation enables the identification of more proteins than can be identified from the digest of a whole cell lysate while providing insight into the biological roles of the identified proteins. Other multidimensional separation methods, such as strong cation exchange followed by reversed-phase chromatography of peptides, allow for the identification of many proteins, but do not yield protein function information. Changes in protein expression due to environmental stress can be observed with two-dimensional gel electrophoresis, but a change in protein function cannot be observed as easily. The separation of proteins based on DNA-binding function led to the identification of over 500 proteins, with 232 being classified as DNA-binding. Changes in protein function were observed when the proteins were expressed under aerobic and anaerobic conditions. Alcohol dehydrogenase, which should not be expressed in aerobically grown *E. coli*, was

found to be strongly expressed and DNA-binding under these conditions. The pyruvate dehydrogenase complex of proteins also exhibited unexpected behavior. This complex, usually involved in aerobic metabolism, was expressed and DNA-binding under anaerobic conditions. This is the first proteomic study using a function-based separation to analyze proteins expressed under different conditions. These findings demonstrate the usefulness of function-based separations for proteomic studies.

In the second study, mass spectrometry was used to study DNA-protein interactions. In all of the studies prior to the present one, recombinantly overexpressed, purified proteins were involved. Since mixtures DNA-binding proteins were easily isolated from *E. coli* and purified proteins are expensive and not readily available, photochemical crosslinking experiments in this study were undertaken with the mixtures of DNA-binding proteins. From these experiments, three proteins were found crosslinked, with the majority of the crosslinked peptides observed being from CspC. This protein was found to be abundant in most of the DNA-binding protein fractions. As it turned out, an oligonucleotide containing all thymidine is optimal for its binding site; however, this was not known until after the crosslinking experiment. Another oligonucleotide with a different DNA sequence could be used to crosslink to other proteins in the DNA-binding protein mixture. The crosslinked amino acids match with the NMR, x-ray crystallographic, and site-directed mutagenesis studies on cspA, which show that the data obtained by mass spectrometry is biologically relevant. The amino acids

from the third protein found to be crosslinked, 50S ribosomal protein L24, are probably involved in base-stacking interactions with DNA, but there are no published high resolution structures available for comparison.

The combination of function-based separation with covalent modification of binding sites could be used with many other proteins and substrates to determine protein function and structure within proteomic studies. The most obvious use for this combination of techniques is with drug targets within the pharmaceutical industry. An active drug could be attached to a column and all of the proteins that interact with that compound could be isolated. This would enable the identification of drug targets as well as proteins that cause unwanted side effects. The active sites of the proteins could then be blocked with a reactive version of the drug, and the modified proteins could then be isolated, digested, and analyzed with tandem mass spectrometry. This would identify the amino acids involved in the binding sites, which in turn, would lead to accurate three-dimensional models of the in-solution structure. These models could then be used to create a compound that binds more effectively into the target's binding site and less so into the active sites of proteins responsible for side effects.

In the third study, which was conducted with a large data set of peptides from DNA-binding proteins, it was found that MALDI and ESI prefer to ionize different amino acids, which may account for the complementary nature of the peptides identified with each technique. MALDI has not been used in large-scale proteomic studies, except for peptide mass fingerprinting of proteins digested in-

gel, due to its previous difficulty in coupling with liquid-phase separations. Now that HPLC MALDI is commercially available, large scale comparisons of ESI and MALDI are possible. From the results of this study, it was found that to gain the most information out of a sample in a proteomic study, both ESI and MALDI should be used. While running a sample twice with the same ionization source causes a 25% increase in the number of peptides observed, running the same sample once with ESI and once with MALDI leads to a 49% increase in the number of peptides observed. If only one ionization source is available, the character of the protein of interest should be the guide to choosing which ionization technique to use. ESI favors hydrophobic and acidic peptides, while MALDI tends to ionize basic and aromatic species. It is clear from this research that these preferences need to be taken into account before a sample is analyzed.

BIBLIOGRAPHY

Aebersold, R. (2003) "A Mass Spectrometric Journey into Protein and Proteome Research," *J Am Soc Mass Spectrom*, 14:685-695.

Aebersold, R. and Mann, M. (2003) "Mass Spectrometry-Based Proteomics," *Nature*, 422:198-207.

Alexander, P. and Morison, H. (1962) "Crosslinking of Deoxyribonucleic Acid to Protein Following Ultraviolet Irradiation of Different Cells," *Nature*, 194:882-883.

Baldwin, M.A., Medzihradszky, K.F., Lock, C.M., Fisher, B., Settineri, T.A., and Burlingame, A.L. (2001) "Matrix-Assisted Laser Desorption/Ionization Coupled with Quadrupole/Orthogonal Acceleration Time-of-Flight Mass Spectrometry for Protein Discovery, Identification, and Structural Analysis," *Anal Chem*, 73:1707-1720.

Bane, T.K., Leblanc, J.F., Lee, T.D., and Riggs, A.D. (2001) "Analysis of DNA Binding Proteins by Mass Spectrometry," Poster, *Proceedings of the 49th ASMS Conference on Mass Spectrometry and Allied Topics*, Chicago.

Bane, T.K., LeBon, J.M., Lee, T.D., and Riggs, A.D. (2002) "DNA-Protein and Protein-Protein Interactions by DNA Affinity Capture-LC/MS/MS," Poster, *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando.

Bateman, R.H., Carruthers, R., Hoyes, J.B., Jones, C., Langridge, J.I., Millar, A., and Vissers, J.P.C. (2002) "A Novel Precursor Ion Discovery Method on a Hybrid Quadrupole Orthogonal Acceleration Time-of-Flight (Q-TOF) Mass Spectrometer for Studying Protein Phosphorylation," *J Am Soc Mass Spectrom*, 13:792-803.

Bennett, S.E., Jensen, O.N., Barofsky, D.F., and Mosbaugh, D.W. (1994) "UV-Catalyzed Cross-linking of *Escherichia coli* Uracil-DNA Glycosylase to DNA," *J Biol Chem*, 269:21870-21879.

Biemann, K. and Gagel, J.J., U.S.A. Patent 4843243, June 27, 1989.

Boja, E.S. and Fales, H.M. (2001) "Overalkylation of a Protein Digest with Iodoacetamide," *Anal Chem*, 73:3576-3582.

Bodnar, W.M., Blackburn, R.K., Krise, J.M., and Moseley, M.A. (2003) "Exploiting the Complementary Nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for Increased Proteome Coverage," *J Am Soc Mass Spectrom*, 14:971-979.

Brandi, A., Spurio, R., Gualerzi, C.O., and Pon, C.L. (1999) "Massive Presence of the *Escherichia coli* 'Major Cold-Shock Protein' CspA under Non-Stress Conditions," *EMBO J*, 18:1653-1659.

Burkitt, W.I., Giannakopulos, A.E., Sideridou, F., Bashir, S., and Derrick, P.J. (2003) "Discrimination Effects in MALDI-MS of Mixtures of Peptides-Analysis of the Proteome," *Aust J Chem*, 56:369-377.

Burlingame, A.L., Hansen, K.C., Schmitt-Ulms, G., Chalkley, R.J., Hirsch, J., and Baldwin, M.A. (2003) "Mass Spectrometric Analysis of Protein Mixtures at Low Levels Using Cleavable ^{13}C-ICAT and Multi-Dimensional Chromatography," *Mol Cell Proteomics*, 2:299-314.

Cassey, B., Guest, J.R., and Attwood, M.A. (1998) "Environmental Control of Pyruvate Dehydrogenase Complex Expression in *Escherichia coli*," *FEMS Micro Lett*, 159:325-329.

Cech, N.B. and Enke, C.G. (2000) "Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides," *Anal Chem*, 72:2717:2723.

Chait, B.T. and Cohen, S.C. (1996) "Influence of Matrix Solution Conditions on the MALDI-MS Analysis of Peptides and Proteins," *Anal Chem*, 68:31-37.

Chait, B.T. and Kent, B.H. (1992) "Weighing Naked Proteins: Practical, High-Accuracy Mass Measurement of Peptides and Proteins," *Science*, 257:1885-1894.

Champion, M.M., Campbell, C.S., Siegele, D.A., Russell, D.H., and Hu, J.C. (2002) "Proteome Analysis of *Escherichia coli* K-12 by Two-Dimensional Native-State Chromatography and MALDI-MS," Poster, *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando.

Chelius, D. and Bondarenko, P.V. (2002) "Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry," *J Proteome Res*, 1:317-323.

Chen, C.H., Zhu, Y.F., Lee, K.L., Tang, K., Allman, S.L., and Taranenko, N.I. (1995) "Revisit of MALDI for Small Proteins," *Rapid Commun in Mass Spectrom*, 9:1315:1320.

Davis, T.D. and Lee, M.T. (1998) "Rapid Protein Identification Using a Microscale Electrospray LC/MS System on an Ion Trap Mass Spectrometer," *J Am Soc Mass Spectrom*, 9:194-201.

Dishman, R. (2002) "Rationalizing Proteomics: New Strategies for Streamlining Drug Discovery," *PharmaGenomics*, 2:58-62.

Doneanu, C.E. (2002) "Mass Spectrometric Analysis of UV-Crosslinked Protein-Nucleic Acid Complexes," *PhD thesis, Oregon State University*.

Eng, J.K., McCormack, A.L., and Yates, J.R., III. (1994) "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database," *J Am Soc Mass Spec*, 5: 976-989.

Ermolenko, D.N. and Makhatadze, G.I. (2002) "Bacterial Cold-Shock Proteins," *Cell Mol Life Sci*, 59:1902-1913.

Feng, W., Tejero, R., Zimmerman, D.E., Inouye, M., and Montelione, G.T. (1998) "Solution NMR Structure and Backbone Dynamics of the Major Cold-Shock Protein (CspA) from *Escherichia coli*: Evidence for Conformational Dynamics in the Single-Stranded RNA-Binding Site," *Biochem*, 37:10881-10896.

Field, H.I., Fenyo, D., and Beavis, R.C. (2002) "RADARS, a Bioinformatics Solution that Automates Proteome Mass Spectral Analysis, Optimizes Protein Identification, and Archives Data in a Relational Database," *Proteomics*, 2:36-47.

Gafken, P.R. (2000) "Characterization of UV-Crosslinked Protein-Nucleic Acid Interfaces by MALDI-MS and ESI-MS/MS," *PhD thesis, Oregon State University*.

Gao, H.X., Sengupta, J., Valle, M.; Korostelev, A., Eswar, N., Stagg, S.M., Van Roey, P., Agrawal, R.K., Harvey, S.C., Sali, A., Chapman, M.S., and Frank, J. (2003) "Study of the Structural Dynamics of the *E. coli* 70S Ribosome Using Real-Space Refinement," *Cell*, 113:789-801.

Goldstein, J., Pollit, N.S., and Inouye, M. (1990) "Major Cold Shock Protein of *Escherichia coli*," *Proc Natl Acad Sci USA*, 87:283-287.

Graumann, P.L. and Marahiel, M.A. (1999) "Cold Shock Proteins CspB and CspC are Major Stationary-Phase-Induced Proteins in *Bacillus subtilis*," *Arch Microbiol*, 171:135-138.

Greenbaum, D., Medzihradszky, K.F., Burlingame, A., and Bogyo, M. (2000) "Epoxide Electrophiles as Activity-Dependent Cysteine Protease Profiling and Discovery Tools," *Chem Biol*, 7:569-581.

Griffin, T.J. and Aebersold, R. (2001) "Advances in Proteome Analysis by Mass Spectrometry," *J Biol Chem*, 276: 45497-45500.

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.

Harrison, A.G. (1997) "The Gas-Phase Basicities and Proton Affinities of Amino Acids and Peptides," *Mass Spectrom Rev*, 16:201-217.

He, L., Li, L., and Lubman, D.M. (1995) "Continuous-Flow MALDI Mass Spectrometry Using an Ion Trap/Reflectron Time-of-Flight Detector," *Anal Chem*, 67:4127-4132.

Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., and Watanabe, C. (1993) "Identifying Proteins from Two-Dimensional Gels by Molecular Mass Searching of Peptide Fragments in Protein Sequence Databases," *Proc Natl Acad Sci USA*, 90:5011-5015.

Hillier, B.J., Rodriguez, H.M., and Gregoret, L.M. (1998) "Coupling Protein Stability and Protein Function in *E. coli* CspA," *Fold Des*, 3:87-93.

Holland-Staley, C.A., Lee, K., Clark, D.P., and Cunningham, P.R. (2000) "Aerobic Activity of *Escherichia coli* is Determined by a Single Amino Acid," *J Bacter*, 182:6049-6054.

Janssen, D. (2003) "Making an Enzyme Structure Crystal Clear," *Proteomics & Genomics*, 3:16-24.

Jonscher, K. (2003) "Proteomics: To Spray or Not to Spray," *Genomics & Proteomics*, 3:31-33.

Jungblut, P.R., Krause, E., and Wenschuh, H. (1999) "The Dominance of Arginine-Containing Peptides in MALDI-Derived Tryptic Mass Fingerprints of Proteins," *Anal Chem.* 71:4160-4165.

Kaetzel, R.S. (2003) "Proteomic Approach to 1,2-Dichloroethane Bioactivation and Reaction with Redox-Active Protein Disulfide Isomerase," *PhD thesis, Oregon State University.*

Karas, M., Bachmann, D., Bahr, U., and Hillenkamp, F. (1987) "Matrix-Assisted Ultraviolet Laser Desorption Ionization of Non-Volatile Compounds," *Int J Mass Spectrom Ion Proc*, 78:53-68.

Karger, B.L., Preisler, J., Hu, P., and Rejtar, T. (2000) "Capillary Electrophoresis-Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry Using a Vacuum Deposition Interface," *Anal Chem*, 72:4785-4795.

Karlin, S. and Mrázek, J. (2000) "Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes," *J Bacter,* 182:5238-5250.

Karlin, S., Mrázek, J., Campbell, A., and Kaiser, D. (2001) "Characterizations of Highly Expressed Genes of Four Fast-Growing Bacteria," *J Bacter,* 183:5025-5040.

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.

Kratzer, R., Eckerskorn, C., Karas, M., and Lottspeich, F. (1998) "Suppression Effects in Enzymatic Peptide Ladder Sequencing Using Ultraviolet-Matrix Assisted Laser Desorption/Ionization-Mass Spectrometry," *Electrophoresis*, 19:1910-1919.

Kyte, J. and Doolittle, R.F. (1982) "A Simple Method for Displaying the Hydropathic Character of a Protein," *J Mol Biol*, 157:105-132.

Liebler, D.C. (2002) "Introduction to Proteomics," *Humana Press*, Totowa, NJ, 1-12.

Liebler, D.C., Hansen, B.T., Davey, S.W., Tiscareno, L, and Mason, D.E. (2002) "Peptide Sequence Motif Analysis of Tandem MS Data with the SALSA Algorithm," *Anal Chem*, 74:203-210.

Lipton, M.S., Pasa-Tolic, L., Anderson, G.A., Anderson, D.J., Auberry, D.L., Battista, J.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., Masselon, C., Markillie, L.M., Moore, R.J., Romine, M.F., Shen, Y., Stritmatter, E., Tolic, N., Udseth, H.R., Venkateswaran, A., Wong, K-K., Zhao, R., and Smith, R.D. (2002) "Global Analysis of the *Deinococcus radiodurans* Proteome by Using Accurate Mass Tags," *Proc Natl Acad Sci USA*, 99:11049-11054.

Liu, Y., Patricelli, M.P., and Cravatt, B.F. (1999) "Activity-Based Protein Profiling: The Serine Hydrolases," *Proc Natl Acad Sci USA*, 96:14694-14699.

Lo, L.C., Pang, T.L., Kuo, C.H., Chiang, Y.L., Wang, H.Y., and Lin, J.J. (2001) "Design and Synthesis of Class-Selective Activity Probes for Protein Tyrosine Phosphatases" *J Proteome Research*, 1:35-40.

Lopez, M.M. and Makhatadze, G.I. (2000) "Major Cold Shock Proteins, CspA from *Escherichia coli* and CspB from *Bacillus subtilis*, Interact Differently with Single-Stranded DNA Templates," *Biochim Biophys Acta*, 1479:196-202.

MacCoss, M.J., Wu, C.C., and Yates, J.R., III. (2002) "Probability-Based Validation of Protein Identifications Using a Modified SEQUEST Algorithm," *Anal Chem*, 74:5593-5599.

Mann, M., Schevchenko, A., Wilm, M., and Vorm, O. (1996) "Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels," *Anal Chem*, 68:850-858.

McLuckey, S., Van Berkel, G.J., and Glish, G.L. (1992) "Tandem Mass Spectrometry of Small Multiply Charged Oligonucleotides," *J Am Soc Mass Spectrom*, 3:60-70.

Medzihradszky, K.F., Campbell, J.M., Baldwin, M.A., Falick, A.M., Juhasz, P., Vestal, M.L., and Burlingame, A.L. (2000) "The Characteristics of Peptide Collision-Induced Dissociation Using a High-Performance MALDI-TOF/TOF Tandem Mass Spectrometer," *Anal Chem*, 72:552-558.

Medzihradszky, K.F., Leffler, H., Balwin, M.A., and Burlingame, A.L. (2001) "Protein Identification by In-Gel Digestion, High-Performance Liquid Chromatography, and Mass Spectrometry: Peptide Analysis by Complementary Ionization Techniques," *J Am Soc Mass Spectrom*, 12:215-221.

Murray, K.K. (1997) "Coupling Matrix-Assisted Laser Desorption/Ionization to Liquid Separation," *Mass Spectrom Rev*, 16:283-299.

O'Farrell, P.H. (1975) "High Resolution Two-Dimensional Electrophoresis of Proteins," *J Biol Chem*, 250: 4007-4021.

Ollis, D.L. and White, S.W. (1987) "Structural Basis of Protein-Nucleic Acid Interactions," *Chem Rev*, 87:981-995.

Orsnes, H., Graf, T., Degn, H., and Murray, K.K. (2000) "A Rotating Ball Inlet for On-Line MALDI Mass Spectrometry," *Anal Chem*, 72:251-254.

Pandey, A. and Mann, M. (2000) "Proteomics to Study Genes and Genomes," *Nature*, 405:837-846.

Pappin, D.J.C., Hojrup, P., and Bleasby, A.J. (1993) "Rapid Identification of Proteins by Peptide-Mass Fingerprinting," *Curr Biol*, 3:327-332.

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999) "Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data," *Electrophoresis*, 20:3551-3567.

Poseqitz, M.C. and Tempst, P. (1999) "Immobilized Gallium (III) Affinity Chromatography of Phosphopeptides," *Anal Chem*, 71:2883-2892.

Prevast, T.H. and Dwyer, J.L., U.S.A. Patent 5,772,964, June 30, 1998.

Reilly, J.P. and Arnold, R.J. (1999) "Observation of *Escherichia coli* Ribosomal Proteins and their Posttranslational Modifications by Mass Spectrometry," *Anal Biochem*, 269:105-112.

Sayle, R.A. and Milner-White, E.J. (1995) "RASMOL: Biomolecular Graphics for All," *Trends Biochem Sci*, 20:374-377.

Schindelin, H., Jiang, W., Inouye, M., and Heinemann, U. (1994) "Crystal Structure of CspA, the Major Cold Shock Protein of *Escherichia coli*," *Proc Natl Acad Sci USA*, 91:5119-5123.

Schroder, K., Graumann, P., Schnuchel, A., Holak, T.A., and Marahiel, M.A. (1995) "Mutational Analysis of the Putative Nucleic Acid-Binding Surface of the Cold-Shock Domain, CspB, Revealed an Essential Role of Aromatic and Basic Residues in Binding of Single-Stranded DNA Containing the Y-box Motif," *Mol Microbiol*, 16:699-708.

Shefcheck, K.J., Yao, X., and Fenselau, C. (2002) "Evaluation of Heparin for Fractionation of Cytosolic Proteins in Cancer Cells," Poster, *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics,* Orlando.

Shetlar, M.D. (1980) "Cross-linking of Proteins to Nucleic Acids by Ultraviolet Light," *Photochem Photobiol Rev,* 5:105-197.

Smith, M.W. and Neidhardt, F.C. (1983) "Proteins Induced by Aerobiosis," *J Bacter,* 154:344-350.

Smith, M.W. and Neidhardt, F.C. (1983) "Proteins Induced by Anaerobiosis," *J Bacter,* 154:336-343.

Steen, H. and Jensen, O.N. (2002) "Analysis of Protein-Nucleic Acid Interactions by Photochemical Cross-linking and Mass Spectrometry," *Mass Spectrom Rev,* 21:163-182.

Steen, H., Petersen, J., Mann, M, and Jensen, O.N. (2001) "Mass Spectrometric Analysis of a UV-Crosslinked Protein-DNA Complex: Tryptophans 54 and 88 of *E. coli* SSB Cross-link to DNA," *Protein Science,* 10:1989-2001.

Stelzl, U. and Nierhaus, K.H. (2001) "A Short Fragment of 23S rRNA Containing the Binding Sites for Two Ribosomal Proteins, L24 and L4, is a Key Element for rRNA Folding During Early Assembly," *RNA,* 7:598-609.

Tyers, M. and Mann, M. (2003) "From Genomics to Proteomics," *Nature,* 422:193-197.

Valero, M., Giralt, E., and Andreu, D. (1998) "An Evaluation of Some Structural Determinants for Peptide Desorption in MALDI-TOF Mass Spectrometry," in *Peptides 1996* (Eds Ramage, R.; Epton, R.), Mayflower Scientific, Kingswinford, 855-856.

VanBogelen, R.A., Abshire, K.Z., Moldover, B., Olson, E.R., and Neidhardt, F.C. (1997) "*Escherichia coli* Proteome Analysis Using the Gene-Protein Database," *Electrophoresis,* 18:1243-1251.

Vertes, A., Olumee, Z., Sadeghi, M., and Tang, X. (1995) "Amino Acid Composition and Wavelength Effects in Matrix-assisted Laser Desorption/Ionization," *Rapid Commun in Mass Spectrom,* 9:744-752.

Wall, D.B., Berger, S.J., Finch, J.W., Cohen, S.A., Richardson, K., Chapman, R., Drabble, D., Brown, J., and Gostick, D. (2002) "Continuous Sample Desorption from Reversed-Phase Liquid Chromatography to Tracks on a Matrix-Assisted Laser Desorption/Ionization Precoated Target for the Analysis of Protein Digests," *Electrophoresis*, 23:3193-3204.

Wolters, D.A.; Washburn, M.P.; Yates, J.R., III. (2001) "An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics," *Anal Chem*, 73: 5683-5690.

Yamashita, M. and Fenn, J.B. (1984) "Electrospray Ion Source. Another Variation on the Free-Jet Theme," *J Phys Chem*, 88:4451-4459.

Yan, X., Zhang, H., Watson, J., Schimerlik, M.I., and Deinzer, M. (2002) "Hydrogen/Deuterium Exchange and Mass Spectrometric Analysis of a Protein Containing Multiple Disulfide Bonds: Solution Structure of Recombinant Macrophage Colony Stimulating Factor-Beta (rhM-CSFß)," *Protein Science*, 11:2113-2124.

Yao, X., Heller, M., Mattou, H., and Menzel, C. (2003) "Trypsin Catalyzed ^{16}O-to-^{18}O Exchange for Comparative Proteomics: Tandem Mass Spectrometry Comparison Using MALDI-TOF, ESI-QTOF, and ESI-Ion Trap Mass Spectrometers," *J Am Soc Mass Spectrom*, 14:704-718.

Zhan, O., Gusev, A., and Hercules, D.M. (1999) "A Novel Interface for On-Line Coupling of Liquid Capillary Chromatography with Matrix-Assisted Laser Desorption/Ionization Detection," *Rapid Comm Mass Spectrom*, 13:2278-2283.

Zhang, N., Aebersold, R., and Schwikowski, B. (2002) "ProbID: A Probabilistic Algorithm to Identify Peptides Through Sequence Database Searching Using Tandem Mass Spectral Data," *Proteomics*, 2:1406-1412.

Zhou, H., Ranish, J.A., Watts, J.D., Aebersold, R. (2002) "Quantitative Proteome Analysis by Solid-Phase Isotope Tagging and Mass Spectrometry," *Nature Biotechnol*, 20:512-515.