

Copyright 2011 Dana E. Robinson

AUTOMATED DATA ACQUISITION AND ANALYSIS FOR HIGH-RESOLUTION
FOURIER TRANSFORM MASS SPECTROMETRY OF PROTEINS AND PEPTIDES

BY

DANA E. ROBINSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemistry
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Robert B. Gennis, Chair
Professor Alexander Scheeline
Professor Jonathan V. Sweedler
Professor Peter M. Yau

This thesis details the progress made in automating the data acquisition and data processing of MS and MS/MS data from Fourier transform ion cyclotron resonance mass spectrometers. These instruments have made great strides in the past few years, moving from labor-intensive, largely manual systems to the current flagship of the Kelleher group – a 12 tesla hybrid LTQ FT capable of analyzing hundreds of proteins per week with minimal user intervention.

In early automation work centered around an 8.5 tesla hybrid quadrupole FT-ICR mass spectrometer, three strategies for automated operation of the instrument were developed. In one method, a charge state deconvolution algorithm was used to identify species of interest for isolation using a SWIFT waveform, which were then fragmented in the ICR cell using IRMPD. In another method, broad 20-60 m/z wide sections of the mass spectrum of a complex mixture were fragmented in parallel and this multiplexed fragment data was analyzed using an iterative algorithm. In the third method, samples were analyzed using a THRASH-based "quad march" method where data were acquired using consecutive, wide (20-60 m/z) isolation windows, analyzed using the THRASH algorithm, and then selected species were selected for MS/MS fragmentation. Results from the application of this platform to a survey of the *Methanosarcina acetivorans* proteome are presented.

Later work focused on a commercial 12 tesla hybrid linear ion trap FT-ICR mass spectrometer. This automation scheme used hybrid online/offline data acquisition to take advantages of the features of both online LC-MS (efficient separation and rapid data collection) and offline direct infusion MS/MS (project-wide target selection, better MS/MS data). The workflow for the online portion of this scheme is a modified form of the THRASH-based "quad march" data acquisition scheme from the earlier 8.5 T automation work. The centerpiece of this

platform is a database known as the Automation Warehouse, which acts as a repository for the intact mass data observed in a proteome project and stores the overall state of the project. Custom software binds together the raw data, the data stored in the warehouse and ProSight. Results from the application of this platform to a survey of proteins from HeLa cell nuclei are presented.

In an attempt to begin applying the above platforms to membrane proteins, MS analysis of a putative cross-link in the active site of the C-type heme-copper oxygen reductase from *Vibrio cholera* was performed. Though the sequences in the region differed, other members of the HCO superfamily contained a similar cross-link (confirmed by mass spectrometry and crystal structures) that is important in the catalytic cycle of the enzyme. Computer modeling of the C-type oxidase suggested that the cross-link would be present, though the key amino acid would be located on a different helix. MS/MS analysis of a tryptic digest confirmed the presence of the cross-link and the evolutionary migration of the key amino acid.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.....	1
CHAPTER 2: AUTOMATION OF AN 8.5 TESLA HYBRID QUADRUPOLE FOURIER TRANSFORM MASS SPECTROMETER	11
CHAPTER 3: AUTOMATION OF A 12 TESLA HYBRID LINEAR ION TRAP FOURIER TRANSFORM MASS SPECTROMETER	54
CHAPTER 4: EVOLUTIONARY MIGRATION OF THE HISTIDINE-TYROSINE CROSS- LINK IN THE HEME-COPPER OXYGEN REDUCTASES.....	86
CHAPTER 5: PROTON CHANNEL AND ELECTRON DELIVERY MUTANTS PROVIDE INSIGHT INTO THE FORMATION OF THE HIS-TYR COVALENT CROSSLINK IN CYTOCHROME C OXIDASE	107

CHAPTER 1: INTRODUCTION

1.1 Functional Genomics

Functional genomics is defined by its focus on the dynamic qualities of gene expression in a cell or organism, including transcription, translation, protein and RNA modification and protein-protein interactions. The various parts of this changing "state" of a cell's gene products can be measured using a variety of techniques such as microarrays, serial analysis of gene expression (SAGE) (Velculescu, et al., 1995), mass spectrometry (Aebersold and Mann, 2003) and high-throughput DNA sequencing-based methods such as ChIP-seq (Robertson, et al., 2007). The common feature of these techniques is that they are designed to interrogate the state of the cell as a whole, at the system level, and produce copious amounts of data which must then be filtered, processed and converted into biological information about the system. As biology moves further into less reductionist and more holistic, systems-based approaches to investigating organisms and communities, this increased data load and analysis burden quickly becomes a limiting factor in the biological sciences.

1.2 Protein Mass Spectrometry

A common goal in functional genomics is to understand the proteomic state of a cell. This includes not only the catalog of expressed proteins but their modification state, including post-translational processing. Mass spectrometry is one of the most common methods used to

determine this information. The methods used to interrogate proteins using mass spectrometry fall into two broad camps. The most common method involves digesting the protein(s) of interest using a proteolytic enzyme such as trypsin. The masses of these peptides and their fragments from MS/MS experiments are then determined using the mass spectrometer. Software algorithms such as Mascot (Perkins, et al., 1999) and SEQUEST (Eng, et al., 1994) are used to determine which proteins were present in the original sample. This is known as the bottom-up method of protein mass spectrometry. The other technique does not rely on digesting the protein(s) of interest and instead the intact mass is determined followed by fragmentation in the mass spectrometer. This technique is known as the top-down method of protein mass spectrometry (Kelleher, 2004). Although both top-down and bottom-up mass spectrometry involve determining the mass of a polypeptide and its fragments, the information that is obtained, both in theory and practice can be quite different. Figure 1.1 shows the main differences in the information that can be obtained from the two methods.

Although not as commonly used as the bottom-up methodology, top-down offers several important advantages in the information obtained. Most importantly, top-down analysis obtains the mass of the intact protein, vastly increasing the specificity of the identification. This intact mass, when combined with robust fragmentation data, can also help with the determination of the total modification state of the protein, since the mass difference from the gene sequence can be determined. This "overall state" of the protein under analysis is typically missing from a bottom-up experiment. In particular, the combinatorial knowledge of a protein's modifications is lost. As an example, in figure 1.1, the bottom-up data is missing the most C-terminal modification and the combinatorial status of the two modifications toward the N-terminus cannot be determined.

In a top-down experiment, however, all three modifications could potentially be observed in parallel and some information could be gleaned about the occupancy of the modifications.

Bottom-up analysis does have its advantages, however. The small size of tryptic peptides means that they are more easily analyzed by lower-resolution mass spectrometers and that they are unlikely to have extensive gas-phase secondary structure which can interfere with fragmentation. Identification of a particular protein is often more likely in a bottom-up experiment as well since only a small fraction of the many peptides produced by the digestion need to be amenable to analysis in the mass spectrometer (size, hydrophobicity, residue content, etc.) in order to generate an identification.

1.3 The Measurement Challenge

The analytical challenges that a mass-spectrometry-based proteomic study faces are significant. The most daunting is the dynamic range of protein expression in a cell, which can range over nine orders of magnitude. This enormous difference in expression levels can mean that abundant housekeeping proteins will often hide important proteins with low copy number, such as transcription factors. Another factor contributing to the challenge is the mapping of a single gene to multiple gene products via post-translational processing such as variable intron/exon splicing and covalent modification of amino acids (*e.g.* phosphorylation). Post-translational processing also introduces combinatorial complexity, where several modifications can appear on a single gene product in various combinations. This is especially problematic with heavily-modified eukaryotic proteins such as histones. Also, unlike a genome sequencing project, a proteome project has a temporal aspect, as the state of the proteome can vary

enormously depending on environmental conditions. Proteins are also a heterogeneous class of biomolecules and can vary greatly in their properties such as size, pI and solubility, making any single separation and analysis platform unlikely to give a comprehensive picture of a cell's proteome. Clearly, in order to get a reasonable overview of the true state of a cell's proteome, data acquisition and analysis systems will have to be employed which can interrogate the proteome quickly and completely, returning full protein characterizations (knowledge of the full state of the protein, as opposed to simple identifications which are knowledge that some form of the gene was present) to the user.

1.4 The Heme-Copper Oxygen Reductases

The heme-copper oxygen reductases (Garcia-Horsman, et al., 1994) couple the reduction of O₂ to proton pumping in aerobic organisms, generating an electrochemical gradient which is then used to produce ATP for the cell. These multi-subunit integral membrane proteins share a common structure, differing mainly in the types of heme used in the enzyme and the source of the input electrons, which can be either a cytochrome *c* or quinol. A key feature of all HCO reductases is a binuclear active site consisting of a heme group and a copper atom. This active site also contains highly conserved histidine and tyrosine residues which are experimentally found to be cross-linked in the A- and B-type reductases and believed to play a role in catalytic oxygen reduction. This work investigates the presence of this histidine-tyrosine crosslink in the C-type HCO reductases from *Vibrio cholerae*.

In addition to helping to understand the role of the crosslink, the enzymes of the HCO reductase superfamily provide excellent test proteins for membrane protein mass spectrometry. Many of the proteins are easily expressed in common prokaryotic species and represent an upper

bound on the difficulty of working with alpha helix bundle integral membrane proteins due to their large size (typically 60-75 kDa), number of trans-membrane helices (usually 12 in subunit I) and high hydrophobicity. They are large, multi-helix, multi-subunit, highly hydrophobic proteins with known post-translational modifications, both the HY crosslink on subunit I and covalently bound hemes on subunit II of some members, among other modifications. Sample handling and data acquisition techniques which work for the HCOR enzymes would be good candidates for application to a membrane protein survey using mass spectrometry.

1.5 Thesis Overview

The intention of this work is to demonstrate automated data acquisition and processing of proteomics data obtained from Fourier transform ion cyclotron resonance mass spectrometers. Chapter 2 contains early automation work based on a custom 8.5 tesla hybrid quadrupole FT-ICR MS using the MIDAS data acquisition system and embedded tool command language (Tcl) scripting libraries. The tools and workflow are described, platform validation results are demonstrated, and the results from a survey of *M. acetivorans* soluble proteins are presented. Chapter 3 discusses later work based on a 12 tesla hybrid linear ion trap mass spectrometer. Again, the platform, tools and data warehouse are discussed and some early results are presented. Chapter 4 presents a study into the nature of a histidine-tyrosine crosslink in the active site of the heme-copper oxidase subfamily of the oxygen reductases. The techniques used to analyze these proteins could be used by the automation platform in a middle-down survey of a cell's membrane proteome.

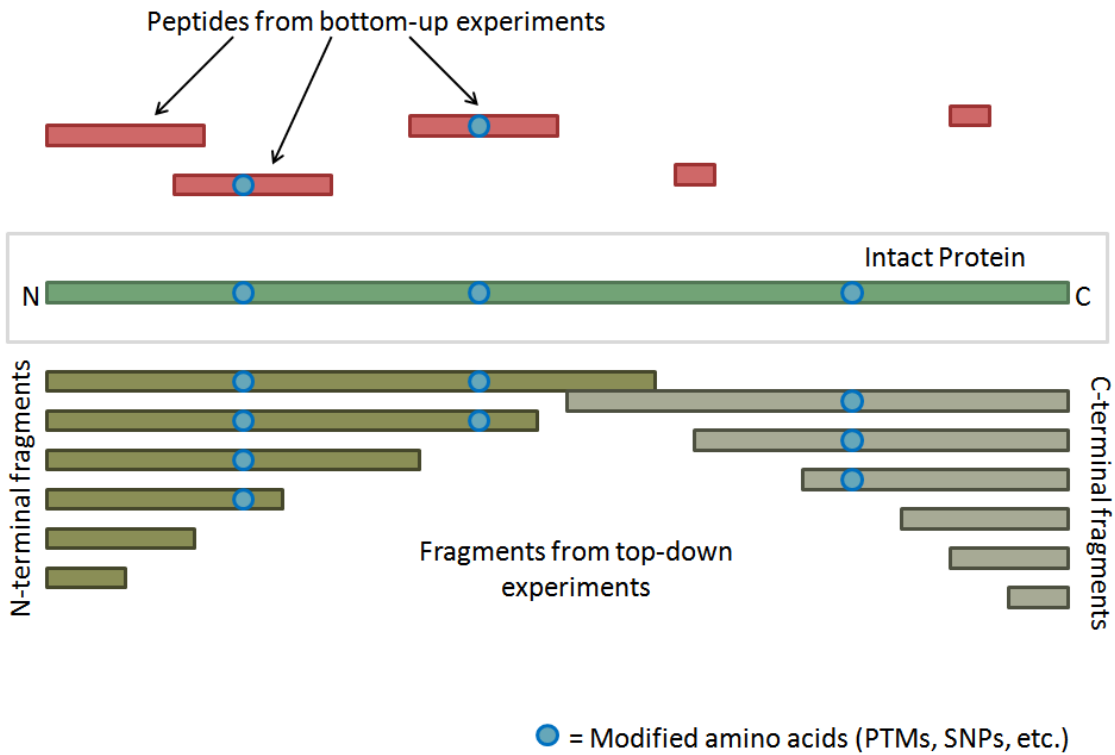


Figure 1.1 Bottom-up and top-down data compared. Bottom-up data is shown above the representation of the intact protein and top-down is shown below. The advantages of top-down analysis are the knowledge of the intact mass and potentially better localization of protein processing events such as post-translational modifications.

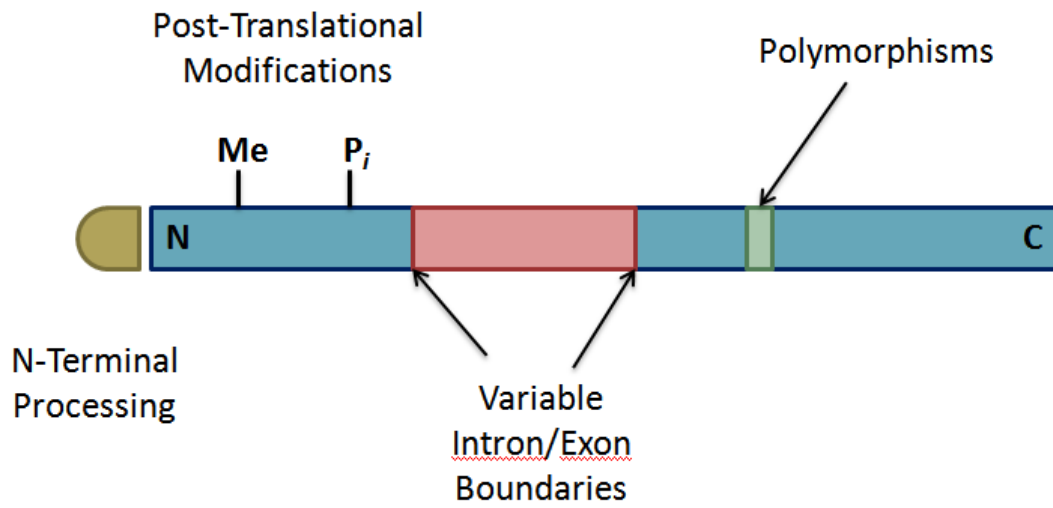


Figure 1.2 Protein processing events which can contribute the number of gene products produced by a particular DNA sequence.

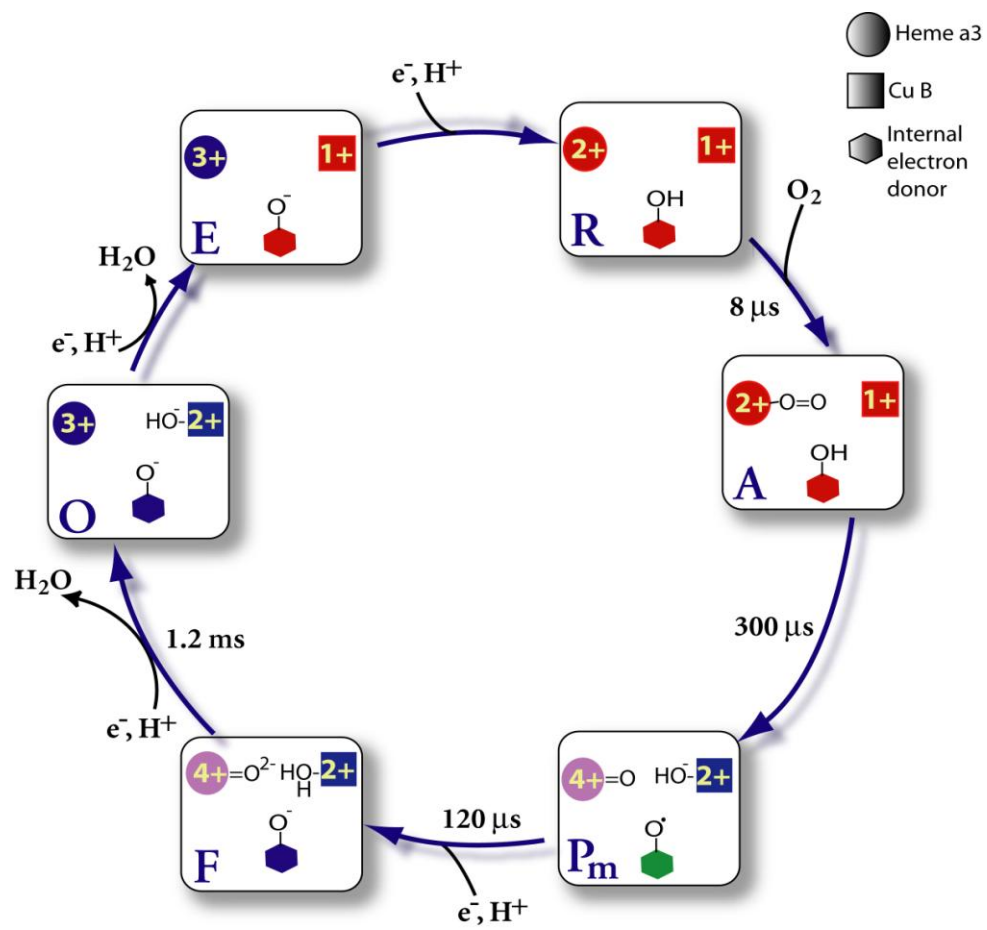


Figure 1.3. The catalytic cycle of heme-copper oxygen reductase. The active site tyrosine is redox active during the catalytic cycle, donating a hydrogen atom (a proton and an electron) to break the O-O bond.

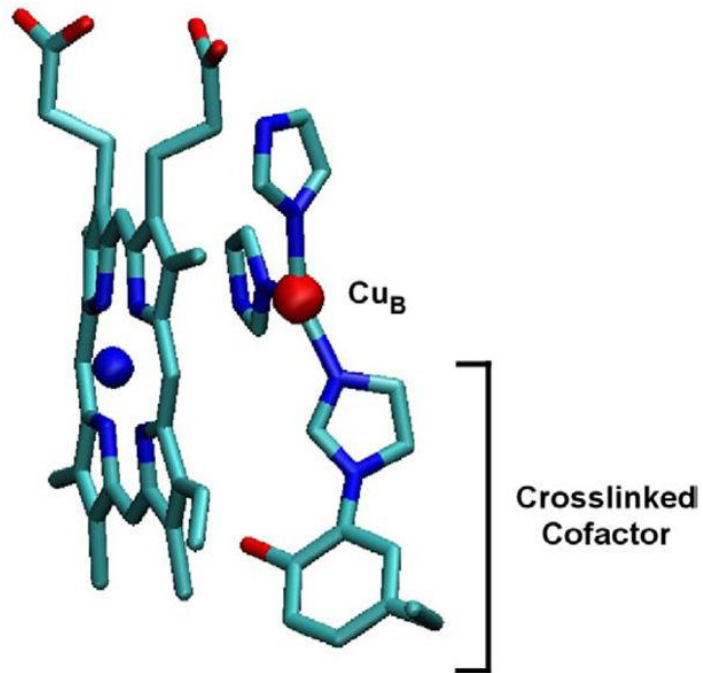


Figure 1.4. Structure of the active-site cofactor from *Rhodobacter sphaeroides* A-type heme-copper oxygen reductase. The cofactor is formed by a covalent cross-link between the Ne of a Cu_B histidine ligand and the Cε of the active site tyrosine and is present throughout the catalytic cycle. The farnesyl tail has been removed from the heme for clarity. This figure was generated using VMD software (Humphrey, et al., 1996).

1.6 Literature Cited

Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics, *Nature*, **422**, 198-207.

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

Garcia-Horsman, J.A., Barquera, B., Rumbley, J., Ma, J. and Gennis, R.B. (1994) The superfamily of heme-copper respiratory oxidases, *J Bacteriol*, **176**, 5587-5600.

Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: Visual molecular dynamics, *J Mol Graphics*, **14**, 33-&.

Kelleher, N.L. (2004) Top-down proteomics, *Anal Chem*, **76**, 196a-203a.

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.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y.J., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O.L., He, A., Marra, M., Snyder, M. and Jones, S. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, *Nature Methods*, **4**, 651-657.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial Analysis of Gene-Expression, *Science*, **270**, 484-487.

CHAPTER 2: AUTOMATION OF AN 8.5 TESLA HYBRID QUADRUPOLE FOURIER TRANSFORM MASS SPECTROMETER

The contents of this chapter were adapted from the following articles: Steven M. Patrie, Jonathan T. Ferguson, Dana E. Robinson, Dave Whipple, Michael Rother, William W. Metcalf, Neil L. Kelleher (2006) "Top Down Mass Spectrometry of <60-kDa Proteins from *Methanosarcina acetivorans* Using Quadrupole FTMS with Automated Octopole Collisionally Activated Dissociation." Molecular & Cellular Proteomics 5(1): 14-25. Cell culture was performed by Michael Rother. Steve Patrie developed OCAD and co-developed the strategies and platform and performed data analysis. Jon Ferguson and Dave Whipple performed data analysis. This work was supported by the Packard Foundation and NSF Foundation Grant CHE-0134953. Steven M. Patrie, Dana E. Robinson, Fanyu Meng, Yi Du, Neil L. Kelleher (2004) "Strategies for Automating Top-Down Protein Analysis with Q-FTICR MS." International Journal of Mass Spectrometry 234: 175-184. Cell culture and some data analysis were performed by Fanyu Meng and Yi Du. Steve Patrie co-developed the strategies and platform and performed some data analysis. This work was supported by the Packard Foundation, the Sloan Foundation, NSF grant CHE-0134953 and NIH grant GM 067193. The custom 8.5 T mass spectrometer on which this platform is based was constructed with assistance from Alan Marshall's group at the National High Magnetic Field Laboratory in Tallahassee, Florida (NSF CHE 99-09502).

2.1 Introduction

With the staggering complexity of biological systems, measurement approaches with high resolution, dynamic range and throughput are required. Particularly, “functional proteomics” methodologies are needed for studying the evolution of cellular phenotype from the genome through the proteome and complex networks that develop within a cell (Godovac-Zimmermann and Brown, 2001). One aspect of functional proteomics emphasizes the identification of cellular proteins, in addition to the characterization of their post-translational state (*i.e.*, protein expression ratios (Gygi, et al., 1999; Gygi, et al., 2002; Haab, et al., 2001; Tonge, et al., 2001; Unlu, et al., 1997), modifications (Mann and Jensen, 2003; Meri and Baumann, 2001) and protein interactions (Fields and Song, 1989; Figeys, et al., 2001; Templin, et al., 2003)). The prevailing methodology for identification and characterization relies upon separation of protein mixtures (often with 2D gel electrophoresis (Fey and Larsen, 2001; Gorg, et al., 2000; Rabilloud, 2002)), enzymatic digestion followed by chromatographic fractionation and analysis with mass spectrometry (Aebersold and Mann, 2003; Henzel, et al., 1993; Lin, et al., 2003). The peptide mass and fragmentation data, generated by tandem MS (MS/MS), are used to search the organism’s protein sequence database (Clauser, et al., 1999; Eng, et al., 1994; Lin, et al., 2003; Nesvizhskii, et al., 2003) to identify the original proteins present. Although reliable for separation of thousands of species (Klose and Kobalz, 1995), protein identification is limited by dynamic range, high sample consumption, reproducibility, and the ability to separate proteins of extreme acidity, basicity, or mass (Gygi, et al., 2000; Lin, et al., 2003; Ong and Pandey, 2001).

“Shotgun-proteomics” eliminates the gel separation by digesting whole cell lysates without prior fractionation (Aebersold and Mann, 2003; McDonald and Yates, 2002; Wu and MacCoss, 2002). Several studies combining multi-dimensional separations with mass spectrometry have extended the dynamic range for protein identification (Peng, et al., 2003; Wolters, et al., 2001). Conrads, *et al.* created a version of the “shotgun” approach, where protein identification is based upon accurate mass analysis of proteolytic digests (<1 ppm mass error) (Conrads, et al., 2000; Smith, et al., 2002). This “accurate mass tag” (AMT) approach was used to obtain 6997 AMTs identifying >61% of the predicted proteins in the genome of *Deinococcus radiodurans* (Lipton, et al., 2002).

Even though high-throughput analysis of the proteome is possible with bottom-up-based approaches, validity of protein identification is often limited by low peptide match redundancy, a high false positive rate and the complexity of data analysis as the proteome size increases (Keller, et al., 2002; Peng, et al., 2003). A contemporary study by Peng *et al.* (Peng, et al., 2003) estimated that 13% of all yeast proteins identified were false positives before manual interpretation of the dataset. Of the remaining identified proteins, 34% were identified with only one peptide match. This lack of sequence coverage leads to incomplete characterization of proteins identified.

A complementary technique to peptide-based proteomics, called the “top-down” approach, is based on MS/MS analysis of intact protein ions without prior digestion (Kelleher, 2000; Kelleher, et al., 1999; VerBerkmoes, et al., 2002). Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) (Comisar.Mb and Marshall, 1974; Marshall, et al., 1998) is one of a few techniques capable of top-down protein analysis from mixtures (ionized by electrospray, ESI) due to its high resolution and mass accuracy (Kelleher, et al., 1997) in MS/MS

mode (Kelleher, et al., 1998). This approach is advantageous because all fragment ions generated are formed within the instrument itself, via MS/MS, and correlate to the specific protein being analyzed. Additionally, with the intact molecular weight information, 100% sequence coverage can be obtained, improving detection of post-translational processing events (Forbes, et al., 2001). The ability to identify and characterize intact proteins with ESI/FT-ICR MS makes top-down proteomics a desirable target for technology development as the evolution of large molecule MS continues to unfold.

Classical and shotgun proteomics both utilize multiple separation techniques to enhance dynamic range for MS/MS of peptides (Link, et al., 1999). Incorporating this kind of front end methodology has the same effect for the top-down approach (Meng, et al., 2002). In the results reported here, *Saccharomyces cerevisiae* samples created by a relatively new two-dimensional approach for proteome fractionation (Meng, et al., 2002) and samples from *Methanosarcina acetivorans* created using an isoelectric focusing strategy contain many proteins above 10 kDa. Even with the increased dynamic range of FTMS at high magnetic field, limitations associated with sample complexity, ESI signal suppression, and chemical noise create a significant challenge above 10 kDa.

This work shows the progress that was made towards acquiring automated, high-resolution top-down MS/MS data using a hybrid quadrupole FT-ICR mass spectrometer (Patrie, et al., 2004). Three different data acquisition modes are presented with a focus on the THRASH-based (Horn, et al., 2000) "quad march" strategy. Results from proteomic surveys of *Saccharomyces cerevisiae* and *Methanosarcina acetivorans* that were obtained using the platform are presented.

2.2 The Data Acquisition Platform

The automation platform is implemented on a custom-built 8.5 T quadrupole FT-ICR mass spectrometer (Q-FTMS) which is very similar to the 9.4 T passively-shielded system maintained at the National High Magnetic Field Laboratory by Marshall and co-workers (Patrie, et al., 2004; Senko, et al., 1997). The instrument is controlled by the MIDAS data station (Senko, et al., 1996) which utilizes an embedded tool command language (Tcl) interpreter to facilitate automated control of the instrument hardware (Freitas, et al., 2003). A library of Tcl modules were created in-house to control various functions used in automation, such as file naming, fragmentation method (IRMPD, CAD, SORI and/or ECD), isolation mode (quadrupole and/or SWIFT), spectrum processing (deconvolution (Zhang and Marshall, 1998) or THRASH (Horn, et al., 2000)), and transition from sample to sample, figure 2.1A. The modular design of the library facilitates the rapid generation of different experimental event sequences used in automation. Figure 2.1B outlines the platform used for the automatic detection and fragmentation of intact protein ions. From the modular event sequence (figure 2.1A, right) the user defines specific parameters unique to each Tcl module (figure 2.1B, left). This format for automation development releases the responsibility of the user to understand the details of the underlying Tcl code/hardware interface, while creating a dynamic environment for an automation platform.

Current methodologies used in the top-down platform have general formats which include steps for molecular ion determination, isolation of targeted species, and finally fragmentation. With current capabilities each of these steps may be linked under a fully-automated platform or separated by some manual intervention. Outlined in figure 2.2 are three examples of modular event sequences implemented to date. A fully-automated deconvolution-

based method (Johnson, et al., 2002) is used (figure 2.2A) to automatically acquire and process the intact spectrum (25 scans) into protein M_r values with calculated charge states. The five most abundant charge states for an individual protein are isolated using a SWIFT waveform generated on-the-fly and subsequently fragmented. In a multiplexing strategy, (figure 2.2B) several protein candidates are isolated simultaneously using the quadrupole filter and fragmented in parallel. In this fully-automated script, after the intact spectrum (25 scans) is acquired, MS/MS of intact proteins is done by arbitrarily fragmenting all protein ions in predefined 20–60 m/z sections (1–5 consecutive sections, 25–50 scans each) (Patrie, et al., 2004). With this strategy, proteins are identified by iterative database searching, starting with the most abundant precursor ions.

Lastly, a method based on isotopically-resolved peaks, automatic processing, filtering, and a combination of quadrupole and/or SWIFT isolation of precursor ions is used (figure 2.2C). In this format, the quadrupole mass filter serves a dual function. First, the intact spectrum is sampled with enhanced sensitivity by selective accumulation of ions using consecutive quadrupole isolation windows of ~ 60 m/z width, defined as a “quad-march.” The data generated are automatically sent to a remote computer (2.8 GHz Pentium IV, 512 MB DDR RAM) for reduction by a modified version of THRASH. Generated lists of pseudo-molecular ions are combined and filtered (off-line) to obtain protein charge state distributions in a custom data filter. The filter enables features such as dynamic exclusion of previously identified proteins and exclusion of adduct peaks (*i.e.*, phosphate, sodium, and potassium). The filter also integrates user-definable searches of molecular ion spectra to facilitate detection of common post-translational modifications of known mass shift such as phosphorylation. After manual inspection of the four to six species to be targeted, new Tcl scripts are generated to isolate

precursor ions with the quadrupole mass filter and are finally fragmented. All masses reported correspond to monoisotopic values unless otherwise specified.

Samples were presented to the instrument with the NanoMate 100 nanospray robot (Advion BioSciences, Ithaca, NY). Ten microliters of each sample was loaded into a 96-well plate and covered with an aluminum seal. Larger sample loads (15–20 μL) reliably gave over 1 h of running time with ~ 200 nL/min flow rate at 0.2 psi back pressure and a chip ESI voltage of 1.5 kV. Ions were accumulated in an 18 cm octopole (1.5 MHz at 500 volts peak-to-peak, -10 V DC offset) after traversing through the quadrupole (ABB Extrel trifilter (Pittsburgh, PA), 20 cm long 9.5 mm diameter rods, controlled by a 1.2 MHz 300 watt QC150 RF power supply, -25 V offset) in either RF-only or mass selection mode. Nitrogen gas (1 millitorr) facilitated accumulation and dissociation in this region. For improved transmission through the quadrupole, ions were collisionally axialized in a 20 cm octopole (1.5 MHz at 500 volts peak-to-peak, -10 V DC offset at 1 millitorr, located behind the skimmer) for 100 ms prior to transfer through the quadrupole. The transfer was then repeated as necessary to accumulate the desired ion population. Direct pressure readings were not available for the octopole accumulation region (low millitorr), so relative adjustments in pressure were monitored via an ionization gauge from Helix Technologies (Longmont, CO) in an adjacent region in the vacuum chamber. Octopole RF and amplitude were controlled via a 33120A 15 MHz frequency generator from Hewlett Packard (Agilent).

In most MS/MS experiments, infrared multiphoton dissociation (IRMPD) or collisional dissociation in the accumulation octopole were used. IRMPD is performed using a 75W CO_2 laser (10.6 μm) with a user-determined irradiation period (~ 0.25 ms) and power level (37–75 W). In this work, only the irradiation period was set by the automation software. Collisional

dissociation was induced in the front octopole (OCAD) by lowering the axial offset of the octopole to a voltage automatically calculated and set by the automation software. To optimize fragmentation, a secondary manually-controlled power supply was coupled to the octopole DC offset supplied by MIDAS to allow on-the-fly adjustments of the fragmentation power.

Fragment mass values generated by the THRASH algorithm were analyzed for *b*-, and *y*-ions and sequence tag (Mann and Wilm, 1994; Mortz, et al., 1996) information using the ProSight PTM (Taylor, et al., 2003) website, either interactively or using a command-line batch processing version developed for analyzing multiple MS/MS spectra without user intervention.

2.3 Validation of the Methodology Using ALS-PAGE/RPLC of Proteins From *Saccharomyces cerevisiae*

Preparation of Samples From Saccharomyces cerevisiae. Protein samples from *S. cerevisiae* were produced by two-dimensional fractionation of cell lysates using continuous-elution gel electrophoresis with an acid-labile surfactant (ALS, Waters Corporation, Milford, MA) to facilitate subsequent reversed-phase liquid chromatography (RPLC). SDS-PAGE was used to determine molecular weight ranges of the fractions from the first dimension of separation. These ALS-PAGE/RPLC fractions, generated as previously described (Johnson, et al., 2002; Meng, et al., 2002), were then lyophilized and resuspended in electrospray solution (78:20:2 acetonitrile:water:acetic acid). Solvents and other reagents were obtained from Sigma Chemical Co. (St. Louis, MO, USA).

Database Searching. The probability scores reported are with 50 ppm fragment ion tolerance and a ± 1000 Da search window (unless stated otherwise) around the candidate protein to accommodate mass shifts associated with post-translational modifications. All data was

externally calibrated on a bovine ubiquitin spectrum. For identification of proteolytic products, the 0–60 kDa region of the database was searched in 5 kDa windows at 50 ppm fragment mass tolerance to localize protein candidates based on number of fragment ion matches. In all cases presented, only one protein is identified at one time with >99% confidence unless multiple species were fragmented in parallel or the identified protein is part of a gene family with nearly identical sequences. If proteins were identified as products of duplicate non-identical genes and the fragmentation data could not discern between the two, they were still considered unique identifications only if each intact protein form was observed.

Decon-Directed Automation. Automatic deconvolution of the spectrum obtained from intact yeast proteins allowed the detection (figure 2.3A), isolation (figure 2.3B), and fragmentation (figure 2.3C) of a species observed in one 25–30 kDa ALS–PAGE/RPLC fraction. The five most abundant charge states for this 29.3 kDa protein (and its phosphate adduct, +98 Da) were SWIFT isolated and fragmented by IRMPD. Ten fragment ions (six *b* and four *y*) matched the 40s ribosomal protein s4 with <25 ppm mass accuracy and a probability score of 5×10^{-8} . This protein was found to lack the N-terminal Met but otherwise harbors no other post-translational modifications. There is no commercial instrument that executes data acquisition in this manner. The future of this approach lies in processing mixtures of ever-larger proteins, while those of smaller M_r values can be measured with isotopic resolution even on lower B_0 FTICR systems.

Multiplexed Fragmentation and the Informatics Challenge. The quadrupole enhancement to FTMS applied to one ALS–PAGE/RPLC fraction containing 10–15 kDa proteins yielded the figure 2.4B versus figure 2.4A improvement over standard FTMS. The broadband spectrum revealed three proteins in the 860–870 m/z region (figure 2.4A) whereas

targeted accumulation yielded nine proteins with a $\sim 5\times$ increase in S/N using about $\sim 45\%$ less sample (figure 2.4B). Collisional fragmentation of all proteins in the 860–870 m/z region as they exited the quadrupole yielded 75 fragment ions from 1 to 13 kDa (figure 2.4C). Figure 2.5 contains a screenshot of the ProSight PTM output for a search of the 12–16 kDa region of the yeast database (50 ppm mass tolerance), showing the six most probable candidates for identification. Two proteins were identified with $>99\%$ confidence ($P_{\text{score}} = 0.0002$ and 0.01) resulting from 13 to 11 *b/y* fragment matches, respectively. Both proteins were observed with start methionine removed and are products of the duplicate 40s ribosomal protein s17 genes which differ by one amino acid on the C-terminal end (Val versus Asn). Neither of the identified proteins were present in the quadrupole isolated region targeted for fragmentation (Fig. 4B), however, both of these values matched molecular ions in the broadband spectrum outside the quadrupole window (figure 2.4A, inset) (15713.6-9 Da theoretical versus 15713.9-9 Da experimental, and 15728.7-9 Da theoretical versus 15728.8-9 Da experimental). Upon further inspection of the intact masses within the quad-window (figure 2.4B), one of the dominant species, 15615.0-9 Da, is consistent within 20 ppm with the C-terminally truncated version of both s17 ribosomal proteins (proteolytic loss of either a C-terminal valine or asparagine). This explains why both proteins were primarily identified from N-terminal fragment ions, with the two C-terminal *y* ions due to random fragment matches. This type of proteolysis could either be artifactual or biologically-relevant, but many such cases have been observed to date with yeast (Johnson, et al., 2002; Meng, et al., 2002), an organism notorious for its proteolytic capacity.

As the data from figures 2.4 and 2.5 illustrate, >11 of 75 fragment masses must match to obtain 99% confidence in identification ($P_{\text{score}} = 0.01$), when searching ~ 2500 protein forms in the 4000 Da window (Meng, et al., 2001). Figure 2.5 also shows four other candidate proteins

were returned in the search. The intact mass for two of them corresponds to species observed in figure 2.4B (13835.0-8 Da theoretical versus 13835.2-8 Da experimental and 14639.2-9 Da theoretical versus 14639.3-9 Da experimental), with 10 and 6 *b/y* fragment ions matching, respectively. However, with $P_{\text{score}} > 0.01$ these were not considered unambiguously identified. Exclusion of fragment masses (and their ammonia and water losses) related to the already identified 15.6 kDa protein lowered the P_{score} of the 13.8 kDa protein to 0.0008. A similar exclusion of these fragment products from the peak list and searching again resulted in a $P_{\text{score}} = 0.01$ for the 14.6 kDa protein present. In both cases no fragment ions for the identified proteins were lost during the filtering event. Improved confidence in identification may also be achieved by internal calibration of the spectrum with fragment ions identified during the “first pass.” Thus, narrowing the fragment mass search tolerance in Prosight PTM from ± 50 to ± 10 ppm makes reported P_{scores} drop by ~ 4 orders of magnitude.

The above data convey that as the fragmentation spectra increase in complexity (due to a combination of secondary fragmentation, water and ammonia loss, as well as fragmenting multiple proteins) the probability of spurious matches also increases. This is compounded by the increased chance for a false hit when searching the whole database or even a large portion of it. On the other hand, lowering the tolerance and restricting searches by the knowledge of observed intact mass values (*i.e.*, top-down) all serve to improve overall identification confidence. However, the latter presumes the database contains the correct protein form (or nearly so). As protein PTM complexity increases (e.g., in higher eukaryotes), alternative identification techniques such as the sequence tag approach will be needed. Application of ECD (Zubarev, et al., 1998) to the quad-enhanced multiplexing approach is one avenue to explore, with the MS/MS spectral complexity anticipated to be exceptionally high.

THRASH-Directed Automation. Two-dimensional fractionation of yeast lysates followed by ESI/Q-FTMS of the resulting samples typically yields 1–13 molecular ion masses per fraction, visible after 25 scans (no prior mass selection). However, spectral quality of samples with low protein concentration is reduced further by incomplete desolvation, chemical noise, and ESI signal suppression. This is typified by the results in figure 2.6A where two yeast proteins were discernable from the high level of background noise in the broadband spectrum. Selective accumulation of 60 m/z sections enhanced the dynamic range of the localized areas ~20-fold (figure 2.6B). Using this procedure across the 900–1100 m/z region increased the number of observed charge state distributions from 2 to 13 with observed masses from 11 to 16 kDa.

After filtering, five of the observed intact ions were subsequently targeted for fragmentation by IRMPD. To maximize transmission of targeted ions, the mass selection window is kept at ~30 m/z and a SWIFT isolation is used to remove residual molecular ions that contaminate the spectrum. Fig. 6 also contains the isolation (figures 2.6C and D) and MS/MS spectra (figures 2.6E and F) for two of the five proteins targeted from the “quad-march” spectra (figure 2.6B). Four of five proteins targeted for fragmentation were identified with an average of nine fragment ions, yielding $P_{\text{scores}} < 0.002$. To obtain yet more backbone cleavages, each of the proteins were manually targeted by collision induced fragmentation in the accumulation octopole, yielding a sequence tag of at least four amino acids for each of the proteins (data not shown).

For complex mixtures, such as in figure 2.6, increased experimental duration due to sampling the spectrum in 60 m/z sections is offset by the >20-fold increase in S/N for the observed species. This is explained by improved accumulation efficiency for the mass-selected species per unit time in the second accumulation octopole as well as reduced dephasing of ion

packets in the ICR cell at high ion populations (Patrie, et al., 2004). Also, the size-dependent fractionation leads to mixtures of proteins within a ~5–7 kDa window. Therefore, with limited mass ranges at least two charge states for all proteins present are usually detectable within three adjacent 60 m/z sections eliminating the need for quad-based sampling of the entire spectral region.

Throughput for Top-Down Strategies Based on The Platform. Nine ALS–PAGE/RPLC fractions in the 10–20 kDa molecular weight range were processed with the “quadmarch”/THRASH-directed approach (figure 2.2C), with intact proteins detected in three 60 m/z Q-FTMS spectra. About 690 individual isotopic clusters were detected yielding 34 charge state distributions (after filtering), 10 of which were believed to be oxidation or phosphate adducts. The remaining 24 were targeted with the quadrupole (no SWIFT clean up) using collisional dissociation in the accumulation octopole for fragmentation. Table 2.1 contains the 20 proteins identified with an average of ~16 b/y fragment ions matching with all $P_{\text{score}} < 0.003$. These were predominantly ribosomal proteins with 18 of 20 detected without their start met and four of these also N-terminally acetylated. An oxidized form (+16 Da) of a 12 kDa heat shock protein was observed in one ALS–PAGE/RPLC fraction. The modification was localized to a 47 amino acid stretch on the backbone (Pro45–Ala91). Three other proteins detected were truncated products due to proteolysis reactions that can occur during cell lysis or perhaps within the yeast cells.

Using this method of operation we can typically perform the broadband (10 scans), three quad-march spectra (10 scans each), process the data, and fragment six proteins (25–50 scans depending on initial abundance) in 45 min. Identification of the proteins is performed manually,

facilitated by ProSight PTM, with general processing times of <5 min per protein except where multiple proteins have been isolated and fragmented in parallel.

2.4 Application of the Platform to a Proteome-Wide Survey of the Soluble Proteins From *Methanosarcina acetivorans*

Preparation of Samples From M. acetivorans. Protein samples from *M. acetivorans* (C2A), grown on MeOH (Galagan, et al., 2002), were produced by two-dimensional fractionation of cell lysates. Two different separation platforms were used in this study. The first was continuous elution PAGE facilitated by an acid-labile surfactant (ALS) from Waters Corp. (Milford, MA) followed by reversed-phase liquid chromatography (RPLC) (Meng, et al., 2002). The second was the ProteomeLab PF 2D protein fractionation system from Beckman Coulter (Fullerton, CA) with chromatofocusing as the first dimension and non-porous silica RPLC in the second (Chong, et al., 2001). Cell lysates were produced by suspension of cells in either lysis buffer (Meng, et al., 2002) or according to protocols outlined in the PF 2D operator manual. Suspended cells were lysed by 15–30 s of microsonication. For the ALS-PAGE/RPLC run 10–15 mg of protein (determined by Bradford assay) was used. For the PF 2D run 1–5 mg of protein (determined by Bradford assay) was used. Molecular weight ranges of the ALS-PAGE fractions were determined by SDS-PAGE.

PF 2D (or ALS-PAGE/RPLC) fractions were lyophilized, resuspended in electrospray solution (78:20:2 acetonitrile:water:acetic acid), and analyzed on a custom hybrid quadrupole FT-ICR mass spectrometer (Patrie, et al., 2004). Samples were presented to the instrument with the NanoMate 100 from Advion BioSciences (Ithaca, NY). Ten to 20 μ L of sample was loaded into a 96-well plate and covered with an aluminum seal. This sample load reliably gave over 1 h

of running time with a 200 nL/min typical flow rate at 0.2 p.s.i. back pressure and a chip voltage of 1500 V. Solvents and other reagents were obtained from Sigma.

Mass Spectrometry and Database Searching. A broadband spectrum (10 scans) of intact pseudomolecular ions was obtained followed by sampling of a 200–300 m/z -wide region of the spectrum in consecutive 60 m/z windows (10–16 scans each) using the quadrupole mass filter, as described in chapters 2.2 and 2.3. Each mass spectrum (for intact proteins) was processed with a remote version of the THRASH algorithm with three processing protocols: zero truncations, one zero fill, Hamming apodization from 450–2000 m/z ; one truncation, one zero fill, Hamming apodization from 450–2000 m/z ; and zero truncations, one zero fill, Hamming apodization from 550–1800 m/z . Only intact masses observed with at least three different charge states at $S/N > 3$ were considered for MS/MS. All masses are monoisotopic unless presented with an italicized “-X” where “X” represents the most abundant isotope in the isotopic envelope (a “0” corresponds to the monoisotopic mass).

Experimental MS/MS mass peak lists were analyzed for *b* and *y* ions and also processed on-the-fly with a remote version of the THRASH algorithm. MS/MS spectra were processed with three different levels of truncation: 0, 1, and 2, which corresponds to 512K, 256K, and 128K processing of the original 512K dataset. Protein identification was performed using a command line version of ProSight PTM (LeDuc, et al., 2004; Taylor, et al., 2003) used for batch processing of multiple MS/MS spectra without user intervention. Identification of the protein was based on any of the three individual datasets. The probability scores < 0.01 with at least seven matching ions (the estimated minimum number of fragment ions needed to achieve 99% confidence in protein assignment in the most gene-dense region of the *M. acetivorans* database) were retained from the search with fragment ion tolerance set at 30 ppm to accommodate

externally calibrated ions. To accommodate mass shifts associated with post-translational modifications, proteolysis and parallel fragmentation of multiple species, the entire database was searched in 1000-Da increments.

In some cases, manual scrutiny of Δm values of an identified protein was performed in the Single Protein Mode of ProSight PTM to characterize selected proteins. Combinations of THRASH outputs for differently processed MS/MS datasets can increase the degree of localization for Δm values in the protein by inclusion of all fragment ions that are optimized for either resolution (512K dataset) or S/N (128K dataset).

CAD Fragmentation in the Accumulation Octopole (OCAD). Collisionally-activated dissociation in the accumulation octopole (OCAD) was used extensively to fragment protein species using the automated platform. In this technique, fragmentation was induced by increasing the potential difference between the DC offsets of the focusing octopole and the accumulation octopole as depicted in figure 2.7. This procedure resulted in a pronounced decrease in parent ion and an increase in fragment ion populations in the FT broadband spectrum as shown with carbonic anhydrase (29.6 kDa) in figure 2.7c.

Figure 2.8 shows an example of OCAD fragmentation on a protein from *M. acetivorans*. An important advantage of this technique is that the abundances of the fragment ions increase with accumulation time while the fragment identities do not change appreciably. This suggests that the fragmentation occurs upon entry into the ~10 millitorr environment of the octopole, avoiding further fragmentation which can produce difficult-to-assign internal fragments. The improved S/N from extended accumulation of fragment ions provides increased identification confidence and protein characterization power. For the identified protein in figure 2.8, 12, 36, and 47 fragment ions matched when using 5, 20, and 50 s accumulation periods, respectively

(figure 2.9, a–c). The number of fragment ion matches for a single 50 s scan was slightly better than that observed for 10 co-added 10-s scans (figure 2.9, c versus d). A total of 140 s was required for the 10 co-added scans; thus a 2.5-fold improvement in experiment throughput was achieved by extended accumulation of the fragment ions. This is largely attributed to improved duty cycle with more of the experimental event sequence devoted to accumulation of the fragment ions and less associated with the transfer, excitation and detection events.

With increased confidence based on the number of fragment matches for the assigned protein, overall P_{scores} dropped with an observed plateau of $\sim 1 \times 10^{-35}$. This plateau was attributed to the observed n/f ratio, which did not change dramatically from the 20 to 50 s accumulation times (figure 2.9e). The plateau of n/f at $\sim 70\%$ can be explained by fragmentation produced secondary fragmentation from increased ion space charge resulting in unassigned fragment ions in the observed spectrum (internal ions not searched). However, upon comparison of a single 10 s scan with an experiment of 10 co-added 10 s scans, a similar decrease in fragment matches to total fragment numbers was observed. This indicates that at low accumulation times there are low abundance fragment ions not matching the predicted b/y ions for the protein, which can only be observed by increased accumulation time or co-adding scans. The lack of space charge-induced fragmentation is typical for standard operation of this instrument with accumulations times < 30 s and protein concentrations $< 10 \mu\text{M}$.

Automated Fragmentation with OCAD. With OCAD, the axial DC offset can be adjusted to optimize fragmentation for the mass and charge of selected species (Equation 1) (Patrie, et al., 2006). This response to the OCAD acceleration voltage was determined empirically to be approximately linear over the 700-1200 m/z range for a variety of proteins and was used by the automation software for automated MS/MS of selected species.

$$V_{\text{OCAD}} = (-0.06 \times m/z) + (0.0004 \times \text{mass}) + 8 \quad (\text{Equation 1})$$

Application of OCAD to THRASH-directed automation is shown in figure 2.10. A single protein dominated one particular broadband spectrum (figure 2.10a). A quad march of the 800–1000 m/z region of the spectrum yielded 397 pseudomolecular ions of which 223 were eliminated from the intact mass candidate list by a filtering program because they were either 1) singly charged contaminant species or 2) adducts due to oxidation, inorganic phosphate, or sodium, or 3) the intact mass only occurred once in all quad-march spectra. The remaining 174 ions were part of 15 distinct charge state distributions. The nine most abundant species were targeted for mass selection and fragmentation with MS/MS spectra for three of the four identified proteins is shown in Figure 2.10, b and c. A total of 42 min was required for the entire automated experiment with 9 min for the broadband/quad march experiments and 33 min to isolate and fragment the nine intact protein charge states. For the 15.6 and 7.9 kDa proteins of figure 2.10b, a 100x and 200x increase in S/N was achieved by extended accumulation after mass selection, respectively. Such improvements in S/N led to detection and localization of Δm values in three of four proteins identified (table 2.2). Characterization of Δm values was highly dependent upon the number of fragment ions matching with site-specific localization of a deamidation and disulfides achieved.

Medium Throughput Top Down Mass Spectrometry with Automated OCAD. An ALS-PAGE/RPLC run (28 ALS-PAGE fractions) from *M. acetivorans* was analyzed using the THRASH-directed automation OCAD. Sixty-five proteins were identified with P_{scores} below 0.01 (table 2.3). Improved dynamic range associated with OCAD yielded an average of 14 b/y

fragment ions per protein compared with seven *b/y* fragment ions from previous studies (Meng, et al., 2001; Taylor, et al., 2003), greatly increasing the specificity of intact protein identification. Of all the proteins identified, 20% were from 0–10 kDa, 60% were from 10–20 kDa, 10% were from 20–30 kDa, and the rest were from above 30 kDa with several of these attributed to proteolysis events in the cell or during sample preparation (Meng, et al., 2002). The majority of the proteins identified were conserved, hypothetical, ribosomal and/or predicted proteins with a small percentage associated with ATP synthesis or involved in methanogenesis (Galagan, et al., 2002).

For the proteins identified, 66% of those observed matched theoretical masses to within 2 Da. Three of these had disulfide bonds indicating incomplete cysteine reduction during sample preparation. Of the remaining 34%, four had mis-predicted start sites (Table I), and four were truncated versions of the predicted protein with all fragment ions associated with one terminal end of the protein. Finally 12 proteins were identified with Δm values > 10 Da, which could not be explained with the fragmentation dataset. For six of the 12, the identification came in a multiplexed format; manual validation of intact masses could not be performed due to low S/N levels.

In a recent set of studies published in 2004, the first detailed investigation into the *M. acetivorans* proteome was presented yielding identification for 10% of the predicted ORFs. Of the 412 identified proteins, 122 proteins were unique to acetate-grown versus 102 unique proteins for methanol-grown cells providing insight into differential protein expression (Li, et al., 2005; Li, et al., 2005). A comparison of the identified proteins obtained with the ALSPAGE/RPLC method with those identified by the complementary bottom up-based method showed a significant bias toward the identification of proteins with higher pI. Of the 37 proteins

identified with the top-down approach (that were not identified by the complementary method), 23 of them had a pI > 9. The identification of proteins with high pI is not surprising because the separation method used does not discriminate based upon this characteristic. However, the ionization efficiency in the mass spectrometer should be greater for the more basic proteins. Also, a direct comparison of the two complementary methods is difficult due to the limited coverage of this pI region in the bottom up method.

Enhanced Throughput and Mass Range with Chromatofocusing and RPLC. For ALS-PAGE/RPLC, discrete molecular mass bands (7 kDa) were reported for *S. cerevisiae* (Meng, et al., 2004); however, the application of this separation platform to thermophiles from the archaea has proven difficult with poor resolution observed from the ALS separation (Forbes, et al., 2004). Similarly, mass spectrometric analysis of the *M. acetivorans* ALS-PAGE fractions was significantly biased toward lower molecular mass (< 25 kDa) species because of their presence in the higher molecular weight fractions as well as reduced concentration of the large proteins in their respective fractions. As an alternative, chromatofocusing (Chong, et al., 2001) was used for the first dimension to separate proteins based on pI. Figure 2.11a contains a chromatogram for a whole cell lysate from *M. acetivorans*. From pH 8.5 to pH 4.0, increasingly acidic proteins eluted from the column. Figure 2.11, b and c, contains representative RPLC traces for fractions from figure 2.11a. Extended gradients were implemented to lower the number of species per fraction (such as figure 2.11c) for first dimension fractions with high absorbance values. Automated MS/MS using OCAD of the 60 RPLC fractions, contained in figure 2.11c, identified 56 intact proteins from 5–59 kDa with an average of 14 *b/y* ions and $P_{\text{scores}} < 0.001$. Twenty-one of the 56 proteins identified are >20 kDa. The increased mass range relative to the analogous ALS-PAGE runs is largely attributed to the high resolution observed for elution of proteins off the

chromatofocusing column and reduced oxidation of proteins during 2D fractionation. An extended mass range for top-down protein identification of whole cell lysates is shown in figure 2.12. MA4159, H⁺-transporting ATP synthase, subunit B (50.3 kDa), was identified with 20 *b/y* fragment matches yielding a seven-amino acid sequence tag. One-hundred percent characterization of the primary amino acid sequence was not achieved, and an apparent mass shift of +20 Da exists; although heavy adduction from sodium (+22) and phosphate (+98) limited clear assignment of the intact mass. This extent of salt adduction was not observed at lower mass, implying this large protein was not fully denatured during sample preparation. In the future, improving denaturing conditions during sample preparation should help to minimize adduction and further extend the mass range for top down proteomics. The optimization of sample pretreatment to minimize potentially artifactual modifications (*e.g.* proteolytic cleavage, oxidation, and deamidation) remains a challenge for the future.

As conveyed in Table 2.2, four proteins had mispredicted translational start sites, implying an error rate of 2% in automated annotation of ORFs in microbial genomes. Further four disulfides were detected (table 2.3) with the one on MA1775 thought to be relevant *in vivo* due to the CXXC motifs present in these proteins. Lastly six Δm values were unexplained. With a this automated hybrid quadrupole FT-ICR mass spectrometer and improved fragmentation coupled with chromatofocusing and RPLC, an approximate analysis rate of 100 proteins/week can be attained in a sustained manner.

2.5 Conclusions

This early automation work shows three strategies for acquiring robust, high-resolution MS/MS data on a hybrid quadrupole FT-ICR mass spectrometer. The three strategies were

validated using ALS-PAGE/RPLC fractions from *S. cerevisiae* and then applied to a proteome-wide survey of *M. acetivorans*.

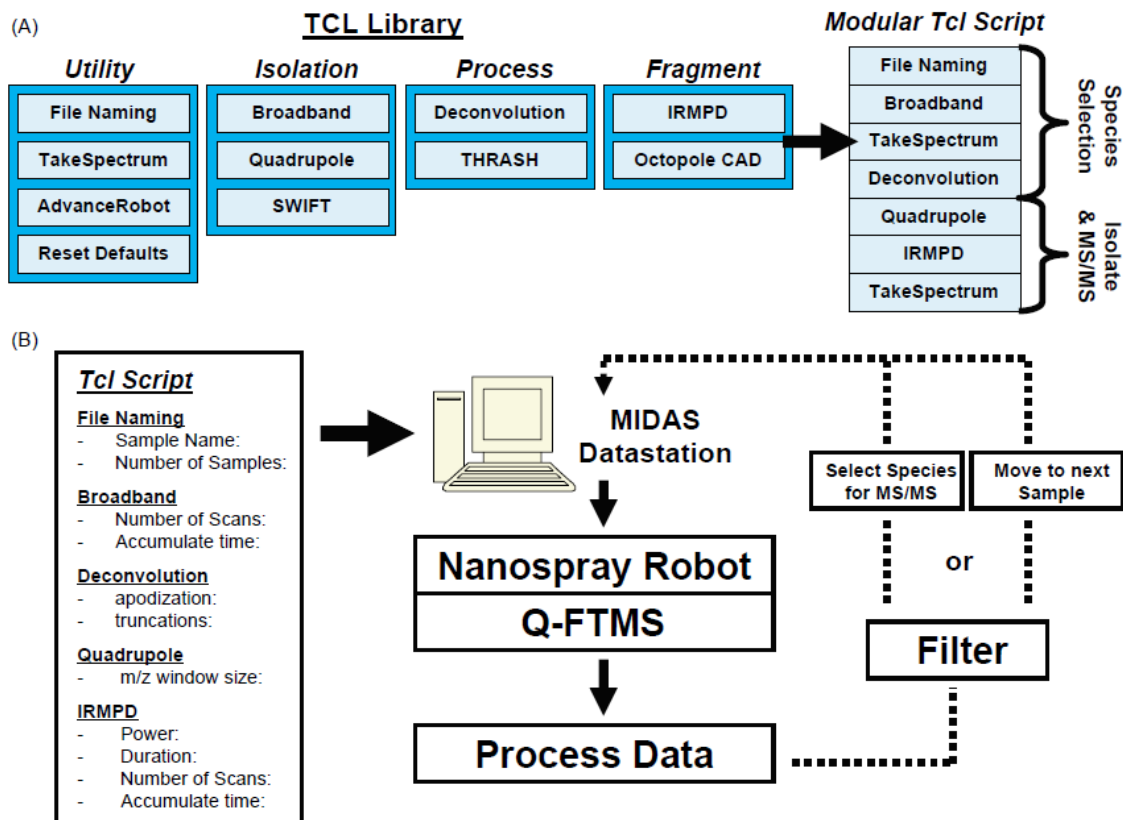


Figure 2.1. (A) Modular Tcl library used for automatic sample processing on a custom quadrupole-FTMS with a nanospray robot and the MIDAS datastation. The modular format of the library facilitates generation of diverse event sequences (at right). (B) General platform of the highly automated top-down platform starting with selection of user-definable properties for the modular experiment sequence.

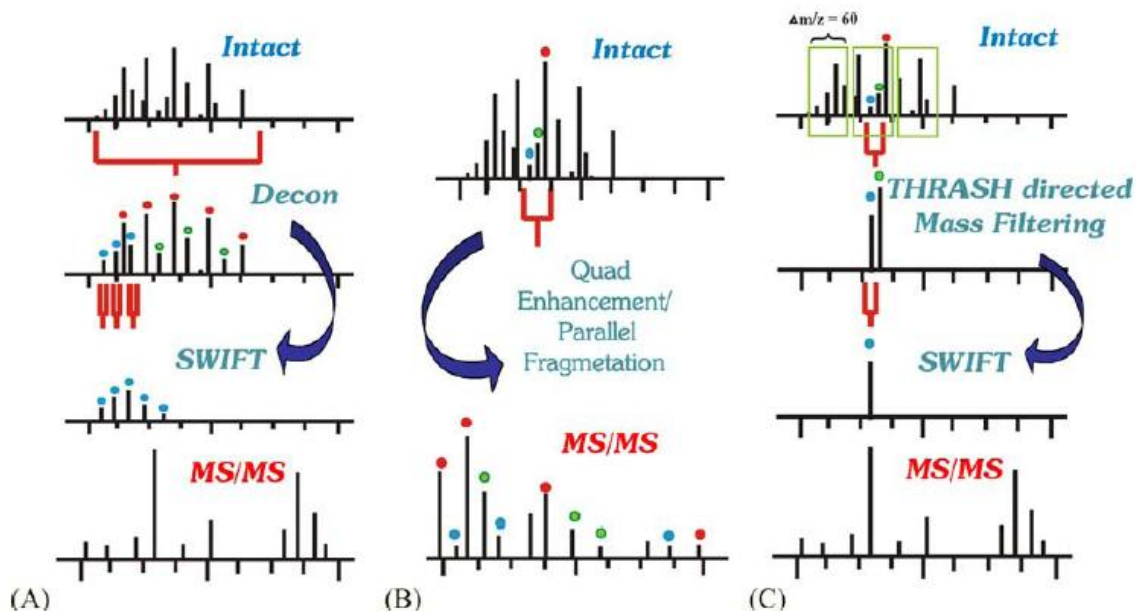


Figure 2.2. (A) Deconvolution-directed automation where the intact spectrum is deconvoluted, processed and either the quadrupole or SWIFT is used for subsequent isolation of intact ions prior to fragmentation. (B) In a multiplexing experiment, multiple species are isolated simultaneously for fragmentation in parallel. (C) A quadrupole targeting experiment, including sampling of the intact spectrum in 60 m/z sections with THRASH-directed detection of intact ions for fragmentation.

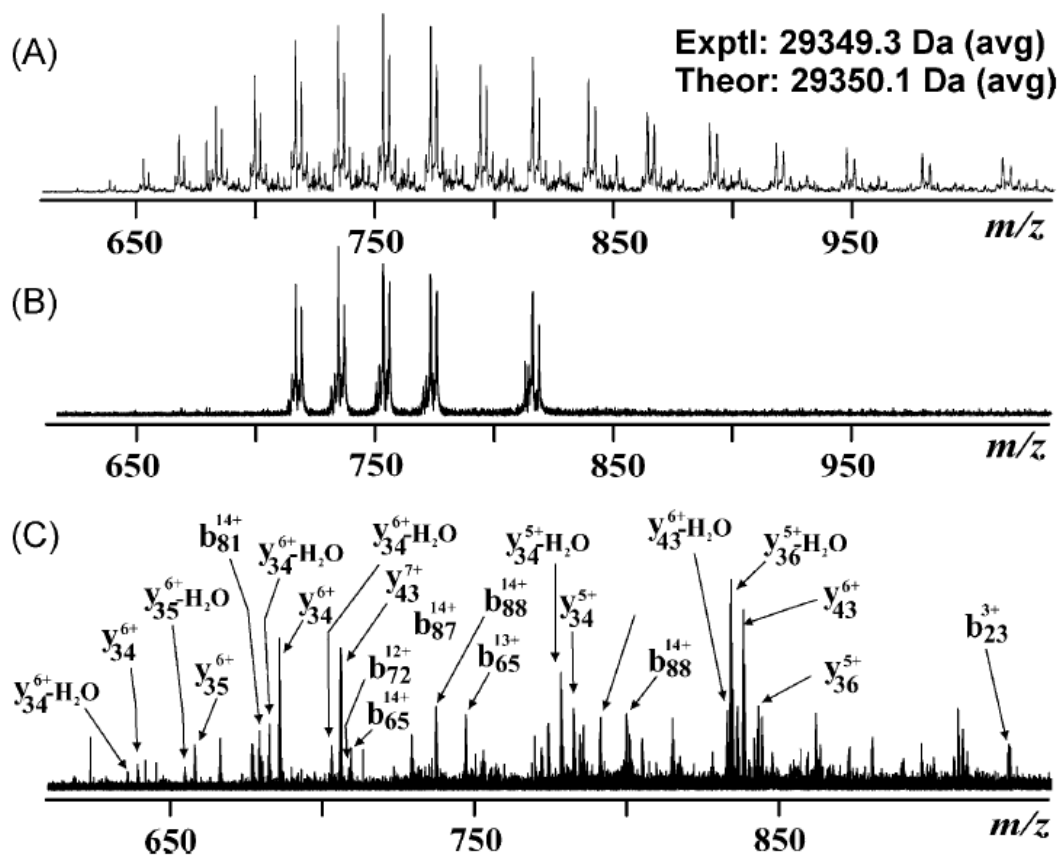


Figure 2.3. Implementation of the deconvolution-based method for identification of a 29.3 kDa yeast protein. The method included automatic reduction of the intact spectrum (A), with on-the-fly SWIFT isolation (B) and fragmentation (C). The contaminating species in (B) is a +98 Da phosphate adduct.

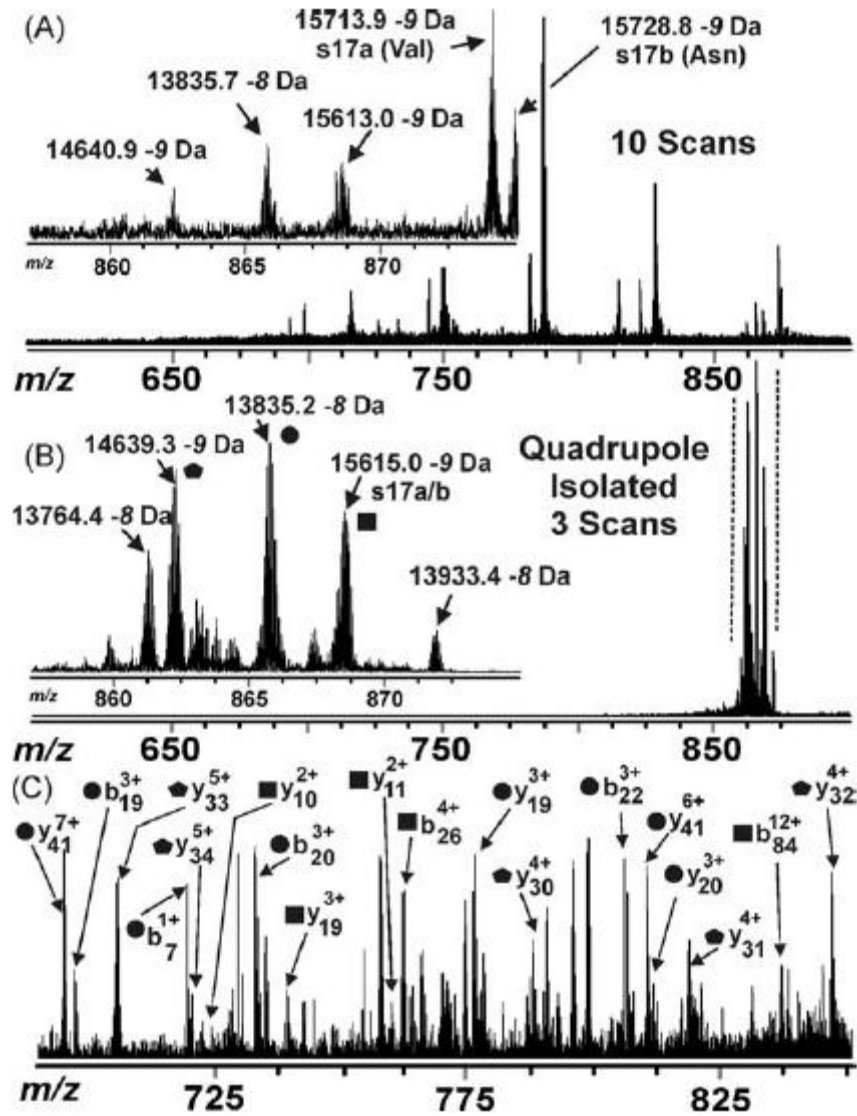


Figure 2.4. (A) Broadband ESI/FT mass spectrum (no mass selection) for a yeast ALS-PAGE/RPLC fraction (4.1s experiment length, 10 scans). (B) Selective accumulation (with mass selection) of all ions in the 860-870 m/z region (7.8s experiment length, 3 scans), compare inset from (A) and (B). (C) Fragment ions identified in the 690-850 m/z region from the spectrum after parallel fragmentation of all species in (B).








Absolute Mass View		ProSight PTM				
Absolute Mass File: 12000-16000.absmass.1.csv		Intact Mass: 14000 Da				
Click on  for Fragmentation Details		Intact Tolerance: 2000 Da				
		Search: B/Y ions				
		Minimum Number of Matches: 2				
		Delta M Mode: NO				
		Frag Mass Type: Monoisotopic				
		Fragment Ion List: /tmp/pklAbMabS				
		Fragment Tolerance: 50 ppm				
Show <input type="button" value="ALL Results"/> Sort By: <input type="button" value="Total Ions"/> <input type="button" value="Descending"/> <input type="button" value="Filter"/>						
<input type="checkbox"/> Group by Gene ID						
ID	Description	Sequence	Mass	B-ions	Y-ions	Pscore
 7043	>NR_SC:SW-R17B_YEAST SW:R17B_YEAST P1412.7 saccharomyces cerevisiae (baker's yeast t). 40s ribosomal pr otein s17-b (rp51b). 12/1998; PIR:R4B7B ribosomal protein S 17.e.B, cytosolic - yeast (Saccharomyces cerev	GVVRTKIVKRSKALIERVYPKLTLDFTM KRLCDELATIQSKRLRNKIAGYTHLMKRI QKGPVVGISFKLQEEERERKDYVPEVSAL DLSRSGVNLVNDHQTSDLVKSLGLKLPISV INVSAQDRRYRERM	15729.2	11	2	0.00021577
 7046	>NR_SC:SW-R17A_YEAST SW:R17A_YEAST P0240.7 saccharomyces cerevisiae (baker's yeast t). 40s ribosomal pr otein s17-a (rp51a). 12/1998; PIR:R5B51 ribosomal protein S 17.e.A, cytosolic - yeast (Saccharomyces cerev	GVVRTKIVKRSKALIERVYPKLTLDFTM KRLCDELATIQSKRLRNKIAGYTHLMKRI QKGPVVGISFKLQEEERERKDYVPEVSAL DLSRSGVNLVNDHQTSDLVKSLGLKLPISV INVSAQDRRYRERM	15714.2	11	0	0.00965659
 1526	>NR_SC:SW-RL35_YEAST SW:RL35_YEAST P3974.1 saccharomyces cerevisiae (baker's yeast t). 60s ribosomal pr otein l35. 12/1998; PIR:S30770 ribosomal protein L35.e, cyto solic - yeast (Sacch aromyces cerevisiae) ; gj1	AGVKAYELRTKSKEQLASQLVDLKKELAEI KVQKLSRPSLPKIKTVRKSIACVLTVIMEQ QREAVRQLYKGGKYQPKDLRAKTRALRRR LTKFEASQVTEKQRKKQLAAPPORRYAIA	13835.4	6	4	0.0568935
 2511	>NR_SC:SW-RL32_YEAST SW:RL32_YEAST P3806.1 saccharomyces cerevisiae (baker's yeast t). 60s ribosomal pr otein l32. 12/1998; PIR:S45410 ribosomal protein L32.e, cyto solic - yeast (Sacch aromyces cerevisiae) ; gj5	(41) ASLPHPKIVKGHYKFKFRHSDRYHR VAENWRKQKGDIVVRRFRGNISQPKIGY GSKKTKFLSPSGHKTFLVANVKDLETLTM HTKTYAAEIAHNISAKNRVVILARAKALGI KVTNPKGRALALEA	14682.2	1	6	6.29413
 2510	>NR_SC:SW-RL32_YEAST SW:RL32_YEAST P3806.1 saccharomyces cerevisiae (baker's yeast t). 60s ribosomal pr otein l32. 12/1998; PIR:S45410 ribosomal protein L32.e, cyto solic - yeast (Sacch aromyces cerevisiae) ; gj5	ASLPHPKIVKGHYKFKFRHSDRYHRVAEN WRKQKGDIVVRRFRGNISQPKIGYGSNK KTKFLSPSGHKTFLVANVKDLETLTMHTK YAAEIAHNISAKNRVVILARAKALGIKVTN PKGRALALEA	14640.1	0	6	23.5983
 2509	>NR_SC:SW-RL32_YEAST SW:RL32_YEAST P3806.1 saccharomyces cerevisiae (baker's yeast t). 60s ribosomal pr otein l32. 12/1998; PIR:S45410 ribosomal protein L32.e, cyto solic - yeast (Sacch aromyces cerevisiae) ; gj5	MASLPHPKIVKGHYKFKFRHSDRYHRVAE NWRKQKGDIVVRRFRGNISQPKIGYGSN KTKFLSPSGHKTFLVANVKDLETLTMHTK TYAAEIAHNISAKNRVVILARAKALGIKVT NPKGRALALEA	14771.3	0	6	23.5983

Figure 2.5. A representation of the ProSight PTM output for the multiplexed fragmentation data shown in figure 4C. Circled numbers indicate either the intact theoretical average mass values for the protein candidates (middle column) or their respective P_{score} (far right column). P_{score} values below 0.01 indicate >99% confidence in the identification. Search was performed on the $14,000 \pm 2000$ Da region of the yeast database with a 50 ppm fragment mass tolerance on all monoisotopic fragment mass values. The two uncircled mass values correspond to different protein forms of identified 60s ribosomal protein l32.

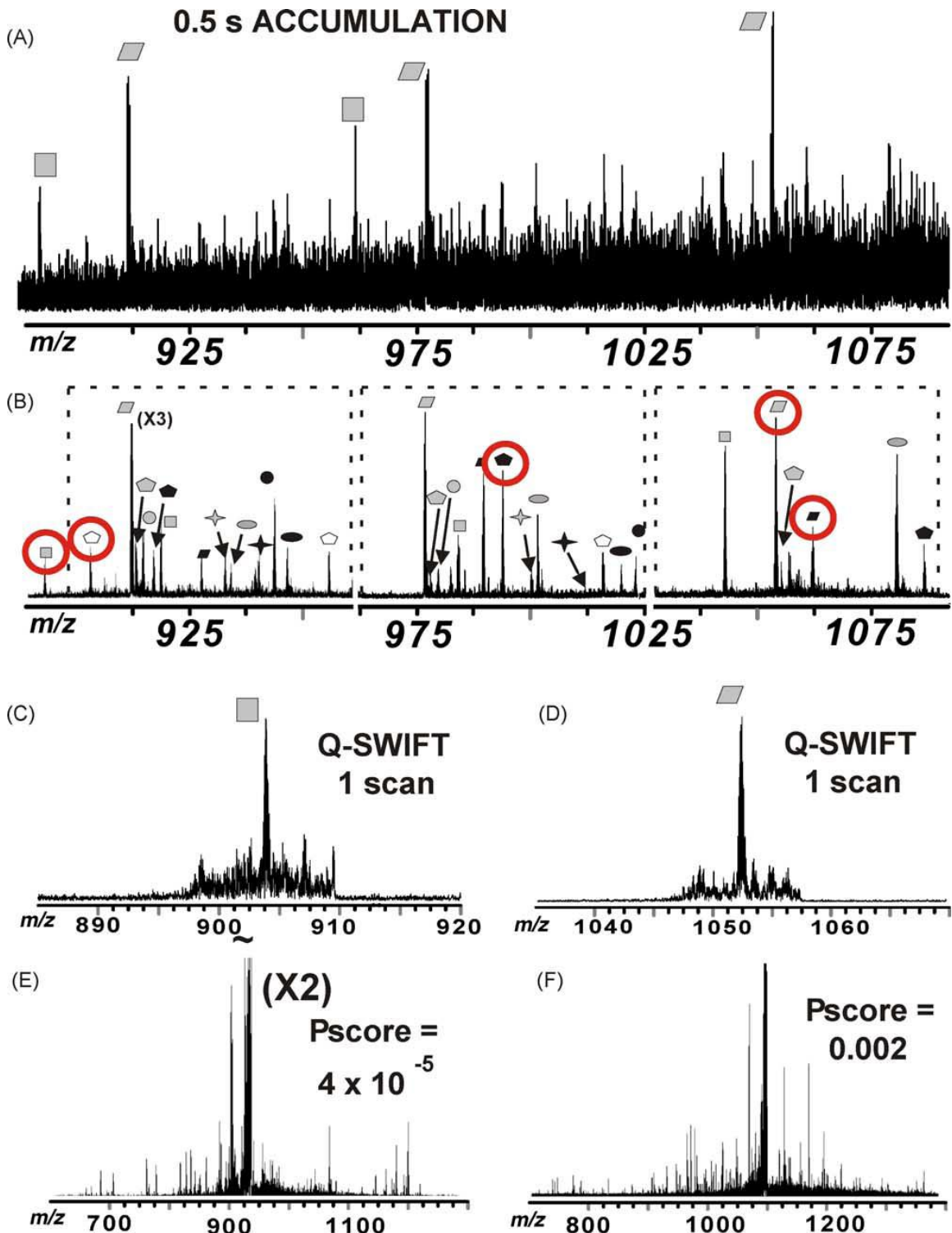


Figure 2.6. (A) The 900-1100 m/z region of a broadband ESI/FT mass spectrum for a yeast ALS-PAGE/RPLC fraction (10 scans). (C and D) Selective accumulation followed by SWIFT isolation (one scan) of two (of five targeted) proteins with more than two charge states in (B). (E and F) IRMPD fragmentation data (25 scans) for isolated ions shown in C and D, respectively. Proteins identified were *hypothetical 12.0 kDa*, *glyceraldehydes 3-phosphate dehydrogenase 3*, *acidic ribosomal protein P0.e* and *2-phosphoglycerate dehydratase* with the latter three observed as 16.2, 13.7 and 12.5 kDa proteolysis products, respectively.

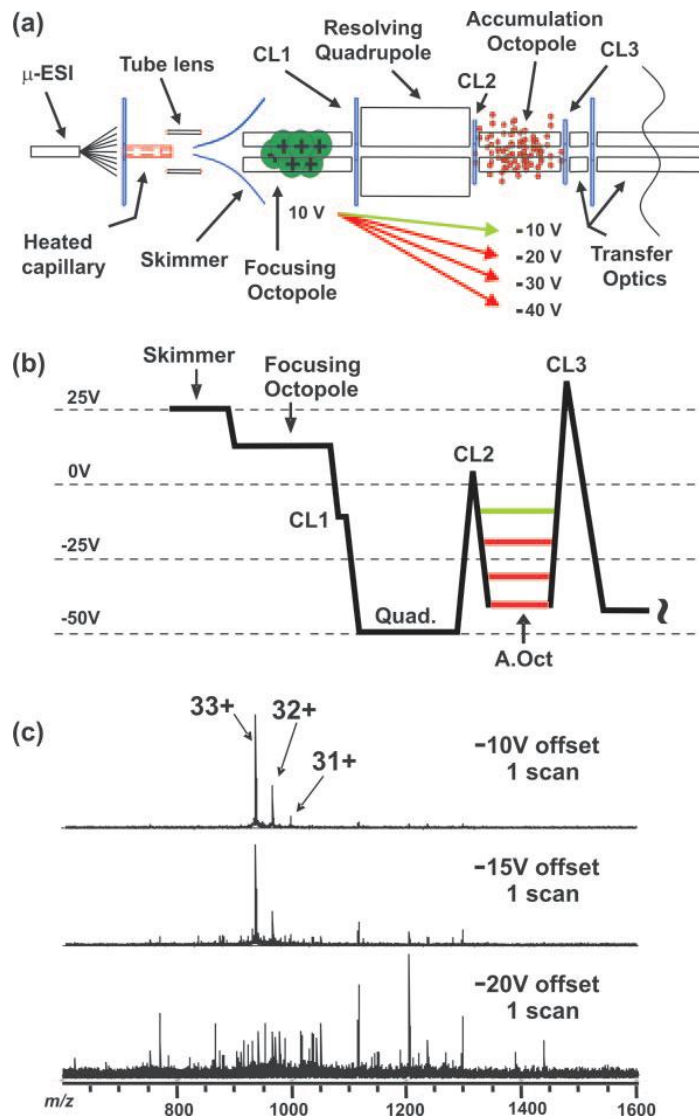


Figure 2.7. External ion optics for the Q-FT mass spectrometer highlighting components used for OCAD. (a) Schematic highlighting the voltages applied between the focusing octopole and the accumulation octopole, which dictates the degree of fragmentation by OCAD. (b) Representative voltages placed on different lenses during the transfer of ions from the focusing octopole, through the filtering quadrupole, to the accumulation octopole (A.Oct) with (red)/without (green) OCAD (c) OCAD-induced fragmentation for the mass-selected 33+ charge state for carbonic anhydrase at various accumulation octopole voltages. 31+ and 32+ charge states are present due to charge transfer of the 33+ with neutral ions in the accumulation octopole after mass selection.

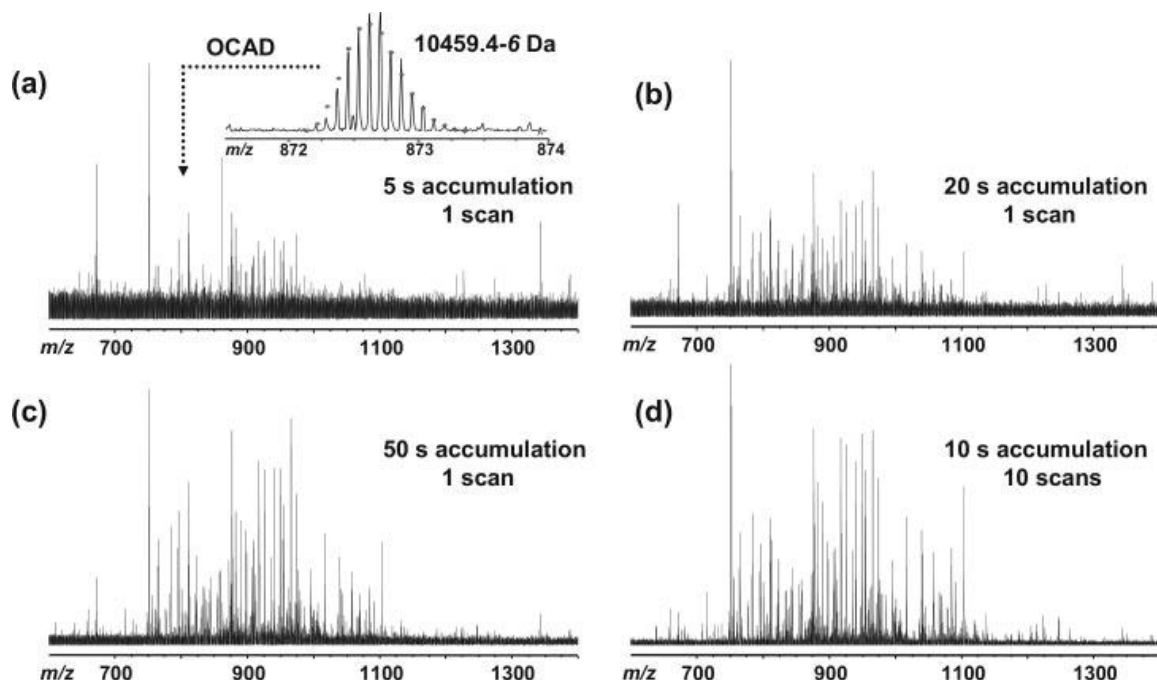


Figure 2.8. OCAD of an *M. acetivorans* protein. Shown is OCAD at -35 V of a mass-selected *M. acetivorans* protein ($[M + 12H]^{12+}$) at (a) 5 s (b) 20 s (c) 50 s and (d) 10 s, 10 co-added scans accumulation.

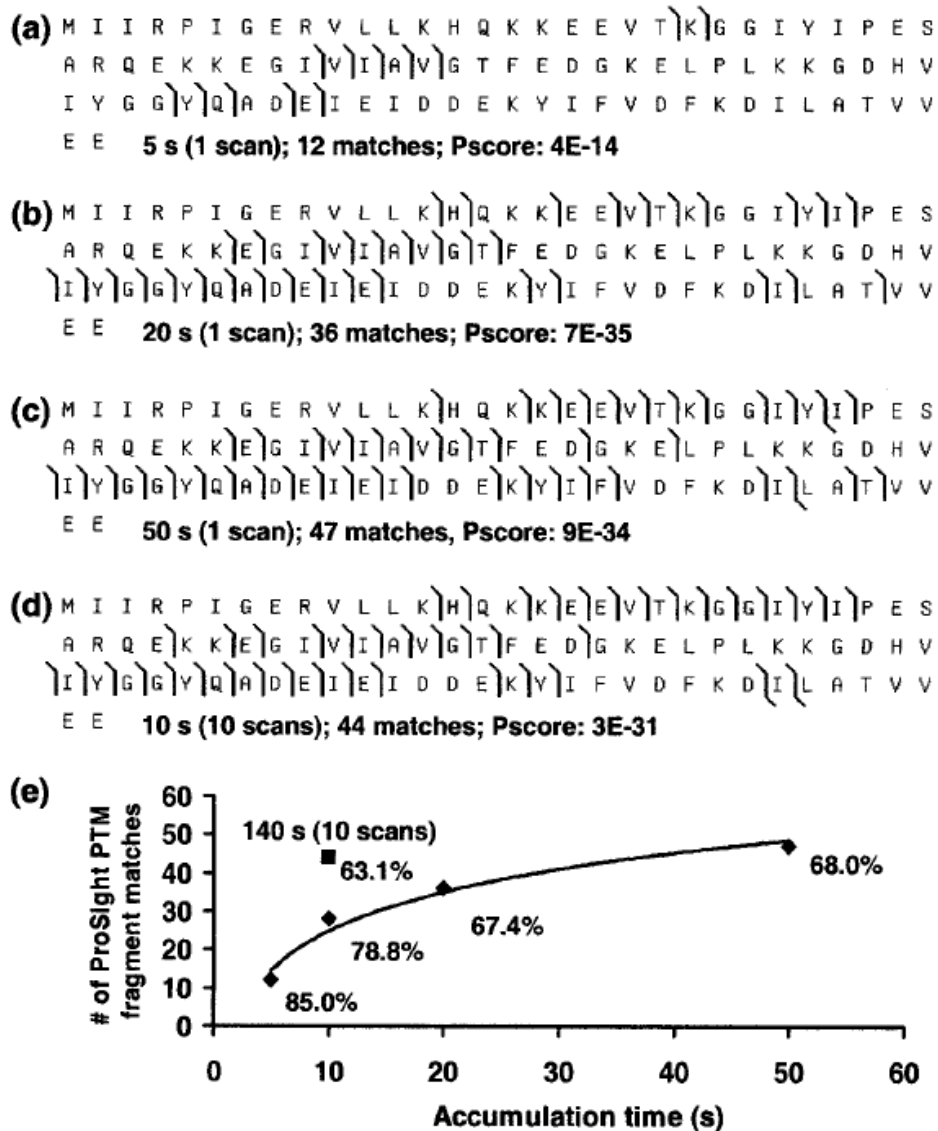


Figure 2.9. Comparison of OCAD data presented in Fig. 2. Shown is the primary sequence for identified *M. acetivorans* protein groES (MA0630) from OCAD data presented in Fig. 2 at 5 s (a), 20 s (b), 50 s (c), and 10 s (d) 10 co-added scans. Dividers indicate points of N-terminal (*b* ions) and C-terminal (*y* ions) cleavage. (e) Plot indicating number of fragment matches versus total accumulation time. Percent values indicate the relative ratio of number of fragment ions matched to MA0630 relative to the total number of ions detected multiplied by 100%.

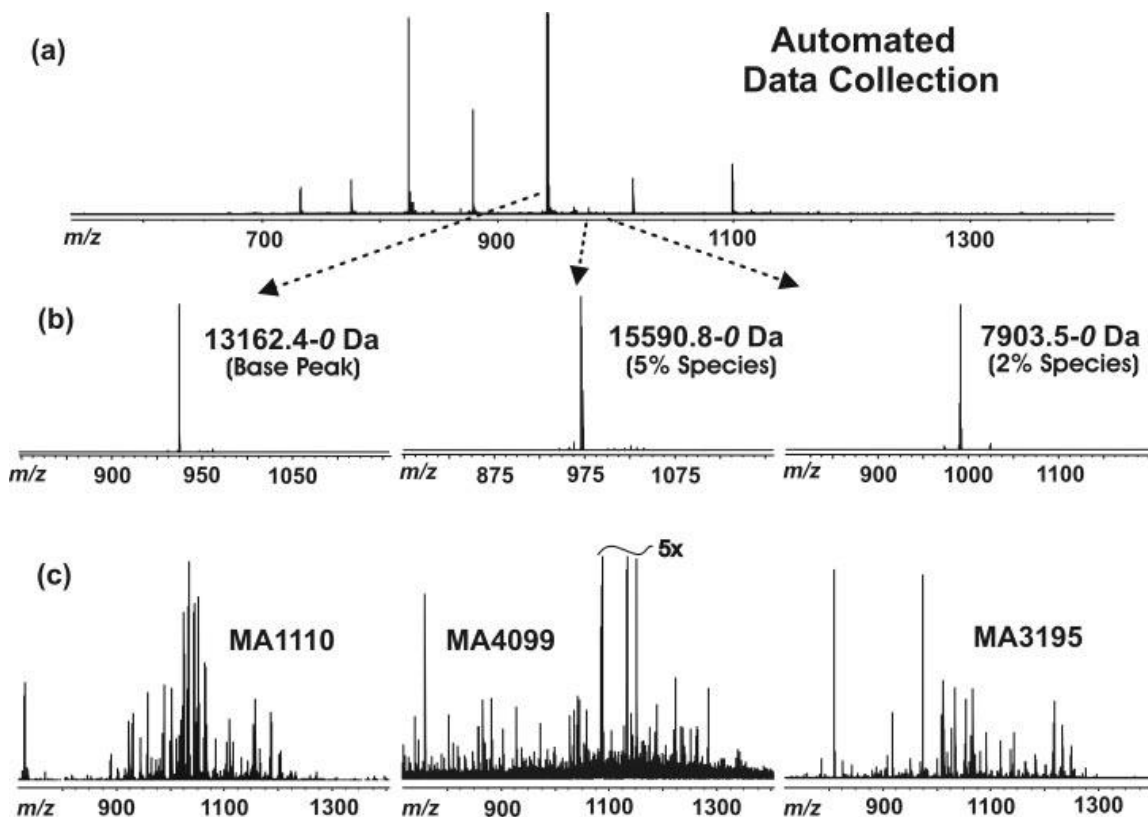


Figure 2.10. Automated processing with OCAD. (a) Broadband ESI-Q-FTMS mass spectrum for one 2D *M. acetivorans* fraction (1s accumulation, 1 scan). Shown are the mass selection (6 s accumulation, 3 scans) (b) and fragmentation (6 s accumulation, 5 scans) (c) of molecular ions observed in the original spectra of intact ions (broadband and quad march).

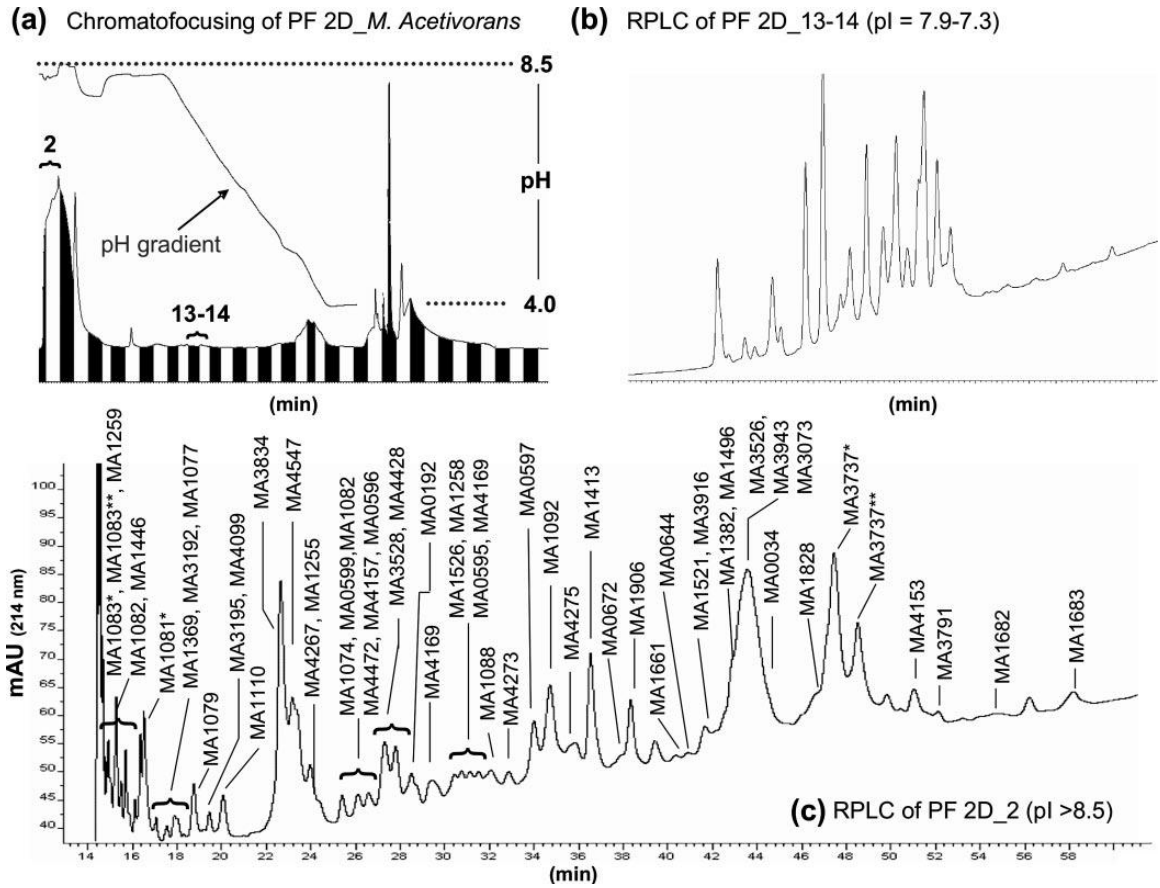


Figure 2.11. Processing of *M. acetivorans* whole cell lysates with chromatofocusing, RPLC and automated OCAD Q-FTMS. (a) PF 2D chromatogram for one *M. acetivorans* cell lysate. (b) RPLC chromatogram for fractions 13-14 (pH 7.3-7.9) from the PF 2D run. (c) RPLC chromatogram for fraction 2 (proteins with pI >8.5). Proteins identified using automated OCAD Q-FTMS are indicated.

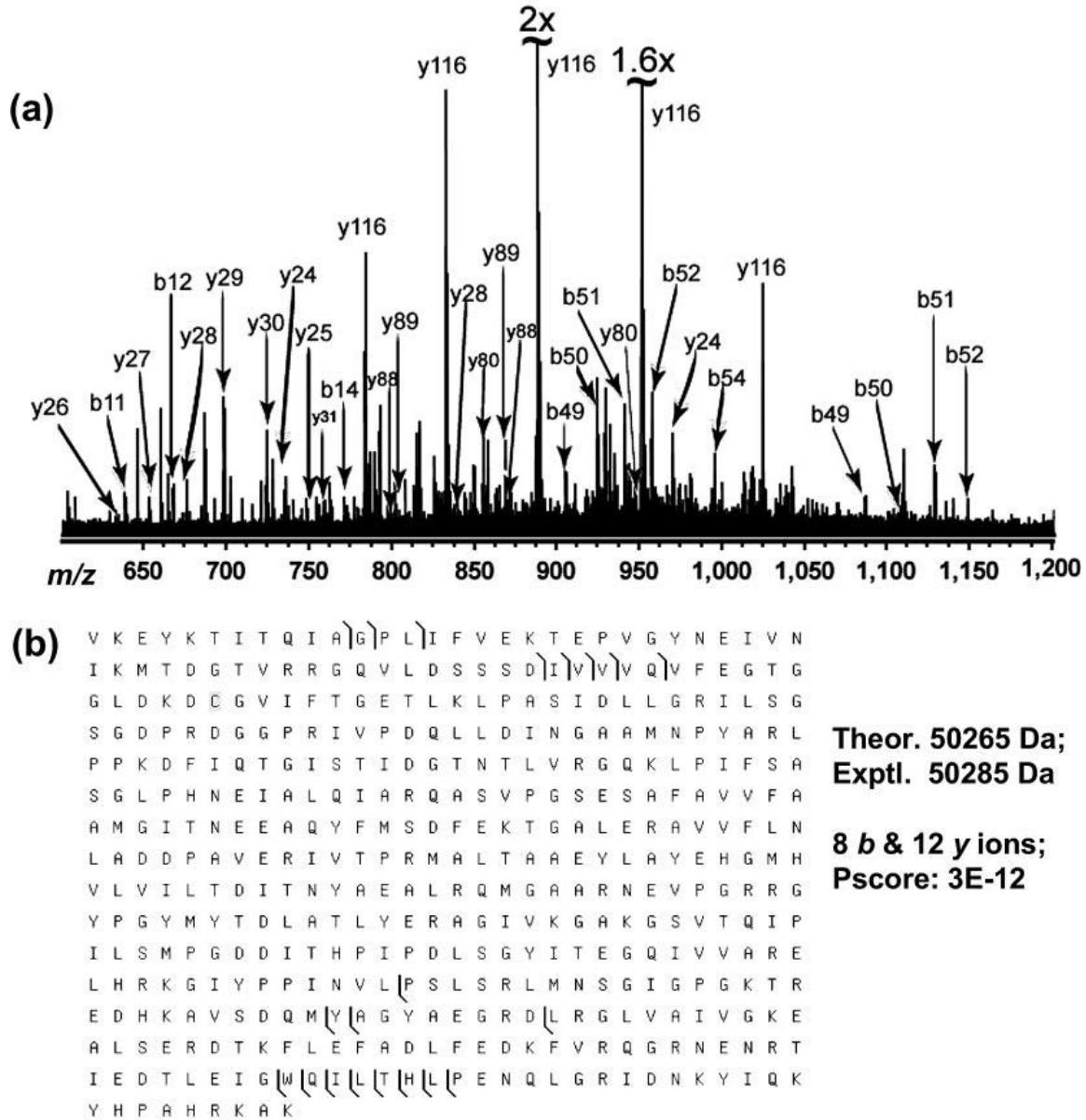


Figure 2.12. Identification of an *M. acetivorans* protein at high mass. (a) A 50.3 kDa protein identified using automated Q-FTMS. Fragmentation yielded 12 unique ions (eight *b*- and 12 *y*-ions). (b) The protein was identified as MA4159, an H⁺-transporting ATP synthase, subunit B.

Table 2.1. *S. cerevisiae* proteins identified from nine ALS-PAGE/RPLC fractions during the "quad march" validation.

Observed mass (Da)	Theoretical mass (Da)	Delta M	Name	b ions	y ions	P_{score}	Notes
6454.5	22540.4	x	60s ribosomal protein 113-a or -b	0	9	$3 \times 10(-5)$	G, PP
8319.4	8319.3	0.1	Cytochrome c oxidase copper chaperone. 7	7	0	0.003	M
11596.5	11596.6	-0.1	12kDa heat shock protein	28	18	$4 \times 10(-38)$	M, Ac
11612.6	11596.6	16.0	12kDa heat shock protein	11	6	$2 \times 10(-10)$	M, Ac, Ox
13535.6	13535.7	-0.1	40s ribosomal protein s26-b	12	1	$9 \times 10(-6)$	M
13593.6	13593.6	0	40s ribosomal protein s26-a	12	1	$9 \times 10(-6)$	M
13728.5	13727.9	0.6	Ribosomal protein L34.e.A	9	2	0.005	M
13728.5	13729.9	-1.4	Ribosomal protein L34.e.B	8	2	0.005	M
13827.2	13827.0	-0.2	60s ribosomal protein l35	6	4	0.0008	M
14094.9	14094.0	0.9	60s ribosomal protein l26-a	9	27	$3 \times 10(-10)$	M
14094.9	14095.0	-0.1	60s ribosomal protein l26-b	7	28	$1 \times 10(-9)$	M
14631.3	14631.2	0.1	Ribosomal protein L32.e	0	6	0.01	M
15230.4	15230.4	0	Ribosomal protein S24.e	5	14	$2 \times 10(-11)$	M, Ac
15605.9	15605.6	x	40s ribosomal protein s17-a or -b	11	0	0.01	M, C-TT
15704.7	15704.6	-0.1	40s ribosomal protein s17-a	15	0	$7 \times 10(-7)$	M
15719.8	15719.6	-0.2	40s ribosomal protein s17-b	15	0	$7 \times 10(-7)$	M
16303.9	16304.8	-0.3	Hypothetical protein YIL051c	5	3	0.0002	M, Ac
16579.6	16580.0	-0.4	Ribosomal protein L27a.e	15	0	$2 \times 10(-7)$	M
17081.2	22765.5	x	DNA polymerase epsilon subunit c	7	0	0.003	PP
17822.0	35707.5	x	Glyceraldehyde-3-phosphate dehydrogenase	8	1	$1 \times 10(-4)$	M, PP

Identified proteins from highly-automated Q-FTMS/MS analysis of nine ALS-PAGE/RPLC fractions processed with a THRASH-directed methodology. x or PP: proteolysis product; M: start methionine loss; Ac: N-terminal acetylation; C-TT: C-terminal truncation; Ox: Oxidation; G: could not differentiate duplicate genes; (): P_{score} exponent.

Table 2.2. Partial list of *M. acetivorans* protein identifications used to validate the automated OCAD fragmentation.

Protein	Description	Mass		Δm	pI	Ions		P-score ^b	Residue localization
		Observed	Theoretical			b ^a	y ^a		
MA3195	Sm protein	7,902.9	7,904.2	0.1	5.21	33	1	1.1E-09	N/A ^c
MA1110	Ribosomal protein S11p	13,161.9	13,162.9	-1.0	10.69	13	23	2.5E-07	Asp to Asn at 64
MA4099	Conserved hypothetical protein	15,590.8	15,593.8	-3.0	8.54	8	9	9.0E-07	Glu ⁴⁴ -His ⁶¹
MA1077	Ribosomal protein L22p	16,586.1	16,589.7	-3.6	9.81	3	7	1.6E-05	Ile ¹²⁰ -Ile ¹⁴²
MA3896	Nitrogen-regulatory protein	11,844.2	12,320.4	MSS ^d	8.55	1	11	1.2E-07	
MA1083	Ribosomal protein L24p	12,596.3	13,043.0	MSS	10.4	0	12	3.0E-06	
MA1108	Ribosomal protein S13p	16,788.0	20,513.0	MSS	10.12	1	14	1.9E-05	
MA1089	Ribosomal protein L32e	17,391.1	17,897.3	MSS	9.32	1	8	4.4E-05	

^a For MA3195, MA1110, and MA1077, fragment ions reflect those obtained from Δm mode in ProSight PTM.

^b All P-scores are based upon fragmentation data obtained without Δm mode in ProSight PTM.

^c N/A, not applicable.

^d MSS, mispredicted start site.

Table 2.3. List of proteins identified from a single PF 2D chromatofocusing fraction *M. acetivorans* using the automated platform.

Protein	Description	Mass		Δm^a	pI	Ions		P-score
		Observed	Theoretical			b	y	
MA0456	6-Methanol-5-hydroxybenzimidazolycobamide co-methyltransferase (methanol-specific corrinoid-binding protein)	27,800.6	27,802.9	DS	5.97	24	24	8.5E-46
MA4516	Nonhistone chromosomal protein MC1	10,686.7	10,686.8	0	10.4	26	16	4.9E-40
MA3631	Cobalamin biosynthesis protein	13,749.9	13,748.3	2	5.09	29	20	3.0E-38
MA0597	Ribosomal protein S9p	14,349.0	14,350.6	-2	10.05	19	19	3.0E-34
MA4472	Ribosomal protein S17e	7,328.3	7,329.0	-1	10.1	12	23	1.2E-31
MA1521	Ribosomal protein L7ae	12,642.9	12,642.8	0	5.26	17	21	1.5E-31
MA0672	Archaeal histone	7,438.1	7,438.1	0	8.08	8	18	7.1E-30
MA0903	hesB family protein	11,423.5	11,423.5	0	4.28	22	4	2.0E-26
MA1471	Translation elongation factor EF-1, subunit β	9,406.3	9,407.0	-1	4.45	27	6	1.8E-25
MA3195	Sm protein	7,904.2	7,904.2	0	5.21	26	0	9.4E-22
MA4277	Ribosomal protein L12p	10,418.3	10,418.1	0	3.85	28	2	2.8E-21
MA1079	Ribosomal protein L29p	7,476.0	7,476.0	0	5.38	16	6	5.7E-21
MA3693	Conserved hypothetical protein	N/A	7,067.6	N/A	5.71	1	17	4.8E-20
MA1093	Ribosomal protein L30p	17,654.3	17,654.3	0	9.52	18	20	1.0E-18
MA4547	Methyl coenzyme M reductase, subunit γ	27,557.3	27,557.6	0	5.91	2	18	2.2E-15
MA4546	Methyl coenzyme M reductase, subunit α	7,613.0	62,401.6	TR	5.14	0	15	3.1E-15
MA1074	Ribosomal protein L23p	9,295.4	9,295.8	0	8.96	1	24	5.8E-15
MA4169	Predicted protein	15,924.9	15,925.5	-1	7.82	9	11	1.5E-14
MA1266	DNA-directed RNA polymerase, subunit H	8,674.9	8,674.7	0	6.11	10	10	1.0E-13
MA2059	Predicted protein	13,321.5	13,320.8	1	5.48	10	10	1.7E-13
MA3619	Nonhistone chromosomal protein MC1	10,600.8	10,600.8	0	10.7	5	12	6.0E-13
MA1076	Ribosomal protein S19p	15,293.3	15,293.2	0	10.2	7	13	6.7E-13
MA0818	Conserved hypothetical protein	9,778.2	9,778.3	0	6.23	4	13	2.2E-12
MA1526	Ribosomal protein S6e	14,385.9	14,385.7	0	9.24	18	1	3.0E-12
MA3528	Conserved hypothetical protein	9,354.2	9,355.0	-1	7.98	8	13	3.3E-11
MA0192	Conserved hypothetical protein	9,676.5	9,676.4	0	9.48	6	14	4.0E-11
MA4094	Universal stress protein	16,238.0	16,239.2	-1	4.81	5	8	9.0E-11
MA0923	Ribosomal protein S8e	13,765.6	13,765.5	0	11.6	13	12	2.0E-10
MA1109	Ribosomal protein S4p	24,366.0	24,366.0	0	10.1	1	15	5.6E-10
MA0034	Pyruvate synthase, subunit γ	19,953.0	19,954.4	-1	7.19	3	9	6.4E-10
MA1446	Ribosomal protein L21e	10,962.7	10,962.7	0	10	5	10	8.3E-09
MA3854	Acetolactate synthase, small subunit	15,991.3	15,992.4	-1	5.45	3	10	2.0E-08
MA3938	Conserved hypothetical protein	9,312.9	8,500.5	812	5.42	2	13	3.7E-08
MA1258	Ribosomal protein S7p	20,795	20,939	-144	10	2	10	5.4E-08
MA4116	Conserved hypothetical protein	N/A	14,105.4	N/A	9.69	2	12	7.4E-08
MA1255	Ribosomal protein S10p	11,411.2	11,411.2	0	9.8	17	2	1.6E-07
MA1818	Riboflavin synthase, subunit β	14,820.9	14,821.8	-1	5.79	1	10	2.8E-07
MA3136	Peptidylprolyl isomerase	17,613.6	17,613.7	0	4.39	6	12	3.0E-07
MA1775	Ribosomal protein L37ae	10,391.6	10,395.5	2DS	10.5	3	10	8.2E-07
MA0684	Conserved hypothetical protein	14,296.6	14,296.3	0	4.97	3	6	8.4E-07
MA0269	Tetrahydromethanopterin S-methyltransferase, subunit H	6,288.4	34,232.5	TR	4.86	0	10	1.2E-06
MA0721	DNA-directed RNA polymerase, subunit L	10,424.4	10,424.5	0	4.54	7	10	1.6E-06
MA4075	Conserved hypothetical protein	15,880.0	15,879.1	1	5.87	1	11	2.0E-06
MA3695	Ribosomal protein S24e	11,596.1	11,596.0	0	5.59	8	0	2.8E-06
MA1077	Ribosomal protein L22p	16,703.7	16,703.7	0	9.81	3	4	3.5E-06
MA0595	Ribosomal protein L18e	13,628.4	15,100.2	TR	9.67	1	15	5.0E-06
MA4157	H ⁺ -transporting ATP synthase, subunit F	10,959.8	10,959.8	0	5.34	2	8	1.3E-05
MA4515	Acetylglutamate kinase	N/A	32,280.1	N/A	5.47	4	3	4.8E-05
MA1110	Ribosomal protein S11p	13,162.9	13,162.9	0	10.7	6	9	5.0E-05
MA4213	Predicted protein	8,568.1	8,567.1	1	4.92	0	12	6.0E-05
MA4430	Methylenetetrahydromethanopterin dehydrogenase	29,833.0	29,833.0	0	5.69	1	10	8.0E-05
MA1462	Universal stress protein	15,505.0	15,506.2	-1	5.23	3	10	1.2E-04
MA1084	Ribosomal protein S4e	N/A	26,168.1	N/A	9.5	5	5	1.3E-04
MA0949	Hypothetical protein (inconsistent evidence)	17,459.1	17,458.5	1	10.00	2	11	1.9E-04
MA2879	Formylmethanofuran dehydrogenase	N/A	13,294.8	N/A	6.59	0	9	2.3E-04
MA4152	H(+)-transporting ATP synthase, subunit H	12,121.3	12,121.3	0	5.07	17	6	3.0E-04
MA3547	Predicted protein	12,425.5	12,425.4	0	4.58	6	5	3.5E-04
MA4328	Conserved hypothetical protein	10,959.2	44,311.8	TR	4.31	4	5	3.9E-04
MA2212	Conserved hypothetical protein	13,715.0	13,713.0	AC	5.55	7	4	5.5E-04
MA0810	Conserved hypothetical protein	N/A	8,306.4	N/A	4.54	5	6	6.1E-04
MA1075	Ribosomal protein L2p	24,971	25,544	-573	10.28	4	8	1.1E-03
MA3733	F420-dependent N ⁵ ,N ¹⁰ -methylene-tetrahydromethanopterin reductase	34,849	34,869	-20	6.03	2	5	1.3E-03
MA3850	KE2 protein	13,337.8	13,338.0	AC	5.59	1	8	4.5E-03
MA1090	Ribosomal protein L19e	N/A	16,876.3	N/A	10.41	4	8	5.9E-03
MA0917	Conserved hypothetical protein	6,593.3	79,054.3	TR	5.66	7	0	6.3E-03
MA4117	Ribosomal protein S19e	16,407.1	16,407.7	-1	9.25	3	11	7.0E-03

^a TR, observed intact mass corresponded to a proteolysis product of theoretical intact mass; N/A, intact mass too convoluted for proper assignment; DS, disulfide bonds; AC, probable acetylation.

2.6 Literature Cited

Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics, *Nature*, **422**, 198-207.

Chong, B.E., Yan, F., Lubman, D.M. and Miller, F.R. (2001) Chromatofocusing nonporous reversed-phase high-performance liquid chromatography/electrospray ionization time-of-flight mass spectrometry of proteins from human breast cancer whole cell lysates: a novel two-dimensional liquid chromatography/mass spectrometry method, *Rapid Commun Mass Sp*, **15**, 291-296.

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

Comisaró, M.B. and Marshall, A.G. (1974) Fourier-Transform Ion-Cyclotron Resonance Spectroscopy, *Chem Phys Lett*, **25**, 282-283.

Conrads, T.P., Anderson, G.A., Veenstra, T.D., Pasa-Tolic, L. and Smith, R.D. (2000) Utility of accurate mass tags for proteome-wide protein identification, *Anal Chem*, **72**, 3349-3354.

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

Fey, S.J. and Larsen, P.M. (2001) 2D or not 2D. Two-dimensional gel electrophoresis, *Curr Opin Chem Biol*, **5**, 26-33.

Fields, S. and Song, O. (1989) A novel genetic system to detect protein-protein interactions, *Nature*, **340**, 245-246.

Figeys, D., McBroom, L.D. and Moran, M.F. (2001) Mass spectrometry for the study of protein-protein interactions, *Methods*, **24**, 230-239.

Forbes, A.J., Mazur, M.T., Patel, H.M., Walsh, C.T. and Kelleher, N.L. (2001) Toward efficient analysis of > 70 kDa proteins with 100% sequence coverage, *Proteomics*, **1**, 927-933.

Forbes, A.J., Patrie, S.M., Taylor, G.K., Kim, Y.B., Jiang, L. and Kelleher, N.L. (2004) Targeted analysis and discovery of posttranslational modifications in proteins from methanogenic archaea by top-down MS, *Proc Natl Acad Sci U S A*, **101**, 2678-2683.

Freitas, M.A., King, E. and Shi, S.D.H. (2003) Tool command language automation of the modular ion cyclotron data acquisition system (MIDAS) for data-dependent tandem Fourier transform ion cyclotron resonance mass spectrometry, *Rapid Commun Mass Sp*, **17**, 363-370.

Galagan, J.E., Nusbaum, C., Roy, A., Endrizzi, M.G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., Brown, A., Allen, N., Naylor, J., Stange-Thomann, N., DeArellano, K., Johnson, R., Linton, L., McEwan, P., McKernan, K., Talamas, J., Tirrell, A., Ye, W.J., Zimmer, A., Barber, R.D., Cann, I., Graham, D.E., Grahame, D.A., Guss, A.M., Hedderich, R., Ingram-Smith, C., Kuettner, H.C., Krzycki, J.A., Leigh, J.A., Li, W.X., Liu, J.F., Mukhopadhyay, B., Reeve, J.N., Smith, K., Springer, T.A., Umayam, L.A., White, O., White, R.H., de Macario, E.C., Ferry, J.G., Jarrell, K.F., Jing, H., Macario, A.J.L., Paulsen, I., Pritchett, M., Sowers, K.R., Swanson, R.V., Zinder, S.H., Lander, E., Metcalf, W.W. and Birren, B. (2002) The genome of *M-acetivorans* reveals extensive metabolic and physiological diversity, *Genome Research*, **12**, 532-542.

Godovac-Zimmermann, J. and Brown, L.R. (2001) Perspectives for mass spectrometry and functional proteomics, *Mass Spectrometry Reviews*, **20**, 1-57.

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

Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology, *Proc Natl Acad Sci U S A*, **97**, 9390-9395.

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.

Gygi, S.P., Rist, B., Griffin, T.J., Eng, J. and Aebersold, R. (2002) Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags, *J Proteome Res*, **1**, 47-54.

Haab, B.B., Dunham, M.J. and Brown, P.O. (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions, *Genome Biol*, **2**, RESEARCH0004.

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 U S A*, **90**, 5011-5015.

Horn, D.M., Zubarev, R.A. and McLafferty, F.W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules, *J Am Soc Mass Spectr*, **11**, 320-332.

Johnson, J.R., Meng, F.Y., Forbes, A.J., Cargile, B.J. and Kelleher, N.L. (2002) Fourier-transform mass spectrometry for automated fragmentation and identification of 5-20 kDa proteins in mixtures, *Electrophoresis*, **23**, 3217-3223.

Kelleher, N.L. (2000) From primary structure to function: biological insights from large-molecule mass spectra, *Chem Biol*, **7**, R37-45.

- Kelleher, N.L., Lin, H.Y., Valaskovic, G.A., Aaserud, D.J., Fridriksson, E.K. and McLafferty, F.W. (1999) Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry, *Journal of the American Chemical Society*, **121**, 806-812.
- Kelleher, N.L., Senko, M.W., Siegel, M.M. and McLafferty, F.W. (1997) Unit resolution mass spectra of 112 kDa molecules with 3 Da accuracy, *J Am Soc Mass Spectr*, **8**, 380-383.
- Kelleher, N.L., Taylor, S.V., Grannis, D., Kinsland, C., Chiu, H.J., Begley, T.P. and McLafferty, F.W. (1998) Efficient sequence analysis of the six gene products (7-74 kDa) from the *Escherichia coli* thiamin biosynthetic operon by tandem high-resolution mass spectrometry, *Protein Science*, **7**, 1796-1801.
- Keller, A., Purvine, S., Nesvizhskii, A.I., Stolyar, S., Goodlett, D.R. and Kolker, E. (2002) Experimental protein mixture for validating tandem mass spectral analysis, *OMICS*, **6**, 207-212.
- Klose, J. and Kobalz, U. (1995) 2-Dimensional Electrophoresis of Proteins - an Updated Protocol and Implications for a Functional-Analysis of the Genome, *Electrophoresis*, **16**, 1034-1059.
- LeDuc, R.D., Taylor, G.K., Kim, Y.B., Januszyk, T.E., Bynum, L.H., Sola, J.V., Garavelli, J.S. and Kelleher, N.L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry, *Nucleic Acids Res*, **32**, W340-345.
- Li, Q.B., Li, L.Y., Rejtar, T., Karger, B.L. and Ferry, J.G. (2005) Proteome of *Methanosarcina acetivorans* Part I: An expanded view of the biology of the cell, *J Proteome Res*, **4**, 112-128.
- Li, Q.B., Li, L.Y., Rejtar, T., Karger, B.L. and Ferry, J.G. (2005) Proteome of *Methanosarcina acetivorans* Part II: Comparison of protein levels in acetate- and methanol-grown cells, *J Proteome Res*, **4**, 129-135.
- Lin, D., Tabb, D.L. and Yates, J.R., 3rd (2003) Large-scale protein identification using mass spectrometry, *Biochim Biophys Acta*, **1646**, 1-10.
- Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M. and Yates, J.R. (1999) Direct analysis of protein complexes using mass spectrometry, *Nat Biotechnol*, **17**, 676-682.
- 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 U S A*, **99**, 11049-11054.
- Mann, M. and Jensen, O.N. (2003) Proteomic analysis of post-translational modifications, *Nat Biotechnol*, **21**, 255-261.
- Mann, M. and Wilm, M. (1994) Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags, *Anal Chem*, **66**, 4390-4399.

- Marshall, A.G., Hendrickson, C.L. and Jackson, G.S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: A primer, *Mass Spectrometry Reviews*, **17**, 1-35.
- McDonald, W.H. and Yates, J.R., 3rd (2002) Shotgun proteomics and biomarker discovery, *Dis Markers*, **18**, 99-105.
- Meng, F., Cargile, B.J., Miller, L.M., Forbes, A.J., Johnson, J.R. and Kelleher, N.L. (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea, *Nat Biotechnol*, **19**, 952-957.
- Meng, F., Cargile, B.J., Patrie, S.M., Johnson, J.R., McLoughlin, S.M. and Kelleher, N.L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry, *Anal Chem*, **74**, 2923-2929.
- Meng, F., Du, Y., Miller, L.M., Patrie, S.M., Robinson, D.E. and Kelleher, N.L. (2004) Molecular-level description of proteins from *saccharomyces cerevisiae* using quadrupole FT hybrid mass spectrometry for top down proteomics, *Anal Chem*, **76**, 2852-2858.
- Meng, F.Y., Cargile, B.J., Miller, L.M., Forbes, A.J., Johnson, J.R. and Kelleher, N.L. (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea, *Nat Biotechnol*, **19**, 952-957.
- Meng, F.Y., Cargile, B.J., Patrie, S.M., Johnson, J.R., McLoughlin, S.M. and Kelleher, N.L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry, *Anal Chem*, **74**, 2923-2929.
- Meri, S. and Baumann, M. (2001) Proteomics: posttranslational modifications, immune responses and current analytical tools, *Biomol Eng*, **18**, 213-220.
- Mortz, E., OConnor, P.B., Roepstorff, P., Kelleher, N.L., Wood, T.D., McLafferty, F.W. and Mann, M. (1996) Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases, *P Natl Acad Sci USA*, **93**, 8264-8267.
- Nesvizhskii, A.I., Keller, A., Kolker, E. and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry, *Anal Chem*, **75**, 4646-4658.
- Ong, S.E. and Pandey, A. (2001) An evaluation of the use of two-dimensional gel electrophoresis in proteomics, *Biomol Eng*, **18**, 195-205.
- Patrie, S.M., Charlebois, J.P., Whipple, D., Kelleher, N.L., Hendrickson, C.L., Quinn, J.P., Marshall, A.G. and Mukhopadhyay, B. (2004) Construction of a hybrid quadrupole/Fourier Transform Ion Cyclotron Resonance Mass Spectrometer for versatile MS/MS above 10 kDa, *J Am Soc Mass Spectr*, **15**, 1099-1108.
- Patrie, S.M., Ferguson, J.T., Robinson, D.E., Whipple, D., Rother, M., Metcalf, W.W. and Kelleher, N.L. (2006) Top down mass spectrometry of < 60-kDa proteins from *Methanosarcina acetivorans* using quadrupole FTMS with automated octopole collisionally activated dissociation, *Molecular & Cellular Proteomics*, **5**, 14-25.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. and Gygi, S.P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome, *J Proteome Res*, **2**, 43-50.

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

Senko, M.W., Canterbury, J.D., Guan, S.H. and Marshall, A.G. (1996) A high-performance modular data system for Fourier transform ion cyclotron resonance mass spectrometry, *Rapid Commun Mass Sp*, **10**, 1839-1844.

Senko, M.W., Hendrickson, C.L., Emmett, M.R., Shi, S.D.H. and Marshall, A.G. (1997) External accumulation of ions for enhanced electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, *J Am Soc Mass Spectr*, **8**, 970-976.

Smith, R.D., Anderson, G.A., Lipton, M.S., Pasa-Tolic, L., Shen, Y., Conrads, T.P., Veenstra, T.D. and Udseth, H.R. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements, *Proteomics*, **2**, 513-523.

Taylor, G.K., Kim, Y.B., Forbes, A.J., Meng, F., McCarthy, R. and Kelleher, N.L. (2003) Web and database software for identification of intact proteins using "top down" mass spectrometry, *Anal Chem*, **75**, 4081-4086.

Templin, M.F., Stoll, D., Schwenk, J.M., Potz, O., Kramer, S. and Joos, T.O. (2003) Protein microarrays: promising tools for proteomic research, *Proteomics*, **3**, 2155-2166.

Tonge, R., Shaw, J., Middleton, B., Rowlinson, R., Rayner, S., Young, J., Pognan, F., Hawkins, E., Currie, I. and Davison, M. (2001) Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology, *Proteomics*, **1**, 377-396.

Unlu, M., Morgan, M.E. and Minden, J.S. (1997) Difference gel electrophoresis: A single gel method for detecting changes in protein extracts, *Electrophoresis*, **18**, 2071-2077.

VerBerkmoes, N.C., Bundy, J.L., Hauser, L., Asano, K.G., Razumovskaya, J., Larimer, F., Hettich, R.L. and Stephenson, J.L., Jr. (2002) Integrating "top-down" and "bottom-up" mass spectrometric approaches for proteomic analysis of *Shewanella oneidensis*, *J Proteome Res*, **1**, 239-252.

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

Wu, C.C. and MacCoss, M.J. (2002) Shotgun proteomics: tools for the analysis of complex biological systems, *Curr Opin Mol Ther*, **4**, 242-250.

Zhang, Z.Q. and Marshall, A.G. (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra, *J Am Soc Mass Spectr*, **9**, 225-233.

Zubarev, R.A., Kelleher, N.L. and McLafferty, F.W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process, *Journal of the American Chemical Society*, **120**, 3265-3266.

CHAPTER 3: AUTOMATION OF A 12 TESLA HYBRID LINEAR ION TRAP FOURIER TRANSFORM MASS SPECTROMETER

The information and contents of this chapter were partially adapted from the following article: Craig D. Wenger, Michael T. Boyne II, Jonathan T. Ferguson, Dana E. Robinson, Neil L. Kelleher (2008) “Versatile Online-Offline Engine for Automated Acquisition of High-Resolution Tandem Mass Spectrometry.” *Analytical Chemistry* **80**(21): 8055-63. Cell culture and protein purification were performed by Mike Boyne and Jon Ferguson. Craig Wenger wrote some of the software, including the `cRAWlers` and the data acquisition user interface. This work was supported by NIH grant GM 067193-06.

3.1 Introduction

Historically, there has been a trade-off in mass spectrometry between resolution and sensitivity. In contemporary MS-based proteomics (Aebersold and Mann, 2003), there has been a long-standing interest in increasing either spectral resolution, the size of peptides/proteins analyzed, or both. Such improvements will allow more protein complexity to be measured with greater certainty (Kelleher, 2004). Driven in part by new ion fragmentation approaches (Syka, et al., 2004; Zubarev, et al., 1998) and improving instrumentation (Good, et al., 2007; Le Blanc, et al., 2003; Macek, et al., 2006; Makarov, et al., 2006; McAlister, et al., 2007; Olsen, et al., 2005; Syka, et al., 2004; Yates, et al., 2006), the steadily advancing capabilities of MS are challenged by targeting polypeptides >3 kDa, such as intact proteins, non-tryptic peptides, and/or large endogenous peptides (Forbes, et al., 2001). Although proteomics has traditionally been a field

ripe for automation (Quadroni and James, 1999), data acquisition solutions for MS/MS of proteins or peptides at >50 000 resolving power remain relatively underdeveloped.

In a typical bottom-up LC-MS/MS experiment using the new breed of ion trap-Fourier transform (FT) hybrid instruments, intact peptide data are now routinely acquired at FTMS resolution (Bakalarski, et al., 2007; Haas, et al., 2006), substantially clarifying protein identifications by database retrieval. However, parameters that lead to increased MS/MS data quality (*e.g.*, long ion accumulation times, detection by FTMS, and spectral averaging) are sacrificed to increase the speed of MS/MS sampling in order to maximize proteome coverage. This increase in sampling rate is not compatible with acquisition of high-resolution MS/MS spectra by Fourier-transform mass analyzers, which are inherently slower than electron multiplier detection-based ion traps and time-of-flight (TOF) instruments more commonly employed for automated MS/MS on a chromatographic time scale (Hofstadler, et al., 1993; Marshall, et al., 1998; Solouki, et al., 1995; Solouki, et al., 1996; Valaskovic, et al., 1996).

In addition to the challenges of performing proteomics with high-resolution MS/MS data, the data acquisition approaches common on modern mass spectrometers are less effective for masses >3 kDa due to charge state distributions that mask less abundant species (Johnson, et al., 2002). Straightforward implementation of data-dependent experimental methods with larger parent ions produced by electrospray typically fragments several charge states of the same precursor, thus making it unlikely to acquire data on a species not even an order of magnitude less abundant on a chromatographic time scale. For small peptides this is not a major concern, since 1-3 kDa peptides will usually produce only one or two charge states within the m/z range of analysis. Although online top-down proteomics is now a reality, as recently demonstrated for

yeast proteins <40 kDa (Parks, et al., 2007), these factors continue to argue for an offline data acquisition strategy.

Much of the current research into increasing the biological dynamic range accessible by MS/MS involves parallel (Purvine, et al., 2003) or data-independent (Venable, et al., 2004) methods, where multiple precursors are selected for simultaneous fragmentation. This multiplexing methodology is very effective at improving offline throughput (Roth, et al., 2005), which is particularly critical for the acquisition of high-resolution tandem mass spectra, and is one of three distinct data acquisition strategies published by Patrie *et al.* in 2004 (Patrie, et al., 2004). These past solutions focused on automating more intelligent data acquisition for top-down proteomics following fractionation by RPLC and were based on the use of advanced data analysis algorithms for determination of highly accurate precursor masses. Initially, the ZSCORE algorithm for charge state and mass determination (Johnson, et al., 2002; Mann, et al., 1989; Zhang and Marshall, 1998) was used. Later, a modified version of the Thorough High-Resolution Analysis of Spectra by Horn (THRASH) algorithm (Horn, et al., 2000) was incorporated to improve sensitivity and accuracy (Patrie, et al., 2006).

Although most research on this front involves innovative mass spectrometers, there are alternatives rooted in the inlet and chromatography configuration. For targeted work performed online, “peak parking” has shown promise (Davis and Lee, 1997; Davis and Lee, 1998; Davis, et al., 1995). For proteomic work, however, a split-flow setup is necessary to extend analysis time for species eluting throughout the entire chromatographic gradient. Split-flow mass spectrometry has thus far been successfully applied to the detection of low-abundance phosphopeptides (Annan, et al., 2001; Zappacosta, et al., 2002) and bacterial signaling molecules (Li, et al., 2007), but has not yet been extended to whole proteome studies.

The work presented here is a natural extension of these past platforms and combines the Advion TriVersa NanoMate with a Thermo Scientific 12 T LTQ FT Ultra into an integrated system for collection of online LC-MS data with simultaneous fraction collection for intelligent acquisition of offline fragmentation data. Targets are sought online, re-infused, and fragmented offline in a highly automated fashion. This new data production engine is coupled with a streamlined version of the ProSight software suite (LeDuc and Kelleher, 2007; LeDuc, et al., 2004; Taylor, et al., 2003; Zamdborg, et al., 2007), advancing technology for high-resolution proteomics that allows for automated acquisition of high-quality MS/MS for bottom-up, middle-down or top-down proteome projects.

3.2 The Data Acquisition Platform

Overview. The workflow described here is shown in Figure 3.1, with the online portion shown in panel a (top) and the offline portion shown in panel b (bottom). The AUTOMATION WAREHOUSE database links the online and offline segments of the workflow, acting as a data repository for the entire proteome project. After converting all isotopic clusters into neutral masses, filtering and binning the results of the online analysis led to several orders of magnitude reduction in the number of species. This step minimized the number of precursor targets for offline interrogation by condensing masses observed at multiple charge states and masses eluting over multiple scans into a single target that was selected at its time of maximum elution. For complete data accountability, every peak found by THRASH was stored in the AUTOMATION WAREHOUSE database.

A typical experimental result of the workflow for top-down acquisition and analysis is shown in Figure 3.2. First, an intact-only LC-MS run with no fragmentation is performed

(Figure 3.2a), during which multiple species are isolated for detection (Figure 3.2b). Any given target is re-isolated with a narrower m/z window in offline mode (Figure 3.2c) and subsequently fragmented (Figure 3.2d). ProSight analysis shows excellent fragmentation of several modified forms of intact human histone H4, the most abundant of which is N-terminally acetylated and dimethylated at lysine 20 (Figure 3.2e).

Software. All non-ProSight software was written for the Microsoft Windows platform in C# using the Microsoft .NET 2.0 Framework, with the exception of THRASH which was written in ANSI C and compiled into a dynamic link library (DLL). Development was done primarily with Microsoft Visual Studio 2005. The AUTOMATION WAREHOUSE database was implemented using MySQL 5.0. For data acquisition, Component Object Model (COM) libraries were used for control of both the Thermo Scientific LTQ (LTQInstControl.dll, March 2007 release) and the Advion TriVersa NanoMate (CSVirDevice.tlb from Chip-Soft 8.1.0.901). Reading of Thermo Scientific .raw data files was performed with the XRawfile COM library (XRawfile2.dll installed with Xcalibur). Extensive .NET wrapper libraries were written to encapsulate the functionality and simplify the interface of all three COM libraries.

Online data in the Thermo Scientific .raw file format was analyzed with an application called ONLINE AUTOMATION cRAWLER, which converts isotopically resolved peaks in every FT scan into neutral masses using a modified version of the THRASH algorithm (Horn, et al., 2000). These peaks were then filtered on m/z , charge, mass, and mass shift relative to previously observed species and other peaks in the same data set. The filtered species were then “binned” with a 10 ppm mass tolerance and inserted into the AUTOMATION WAREHOUSE database. Targets for offline analysis were selected from the AUTOMATION WAREHOUSE database based on intensity, degree of characterization and priority, via an application called TARGET

EXTRACTOR, saved to an extensible markup language (XML) file, and loaded into the MSⁿ APPLICATION, which was responsible for all automated data collection, controlling both the Advion TriVersa NanoMate and the Thermo Scientific LTQ FT Ultra. The MSⁿ APPLICATION iterated through every user-selected target in the list and collected a user-specified number of scans of various types: FT broadband (optional), IT broadband (optional), isolation, and fragmentation. Before the main acquisition on each target occurred, the software first determined whether or not it had sufficient signal abundance in a preview isolation scan, with a cutoff of 1000 (arbitrary units) typically used. If the minimum signal threshold was met, this abundance was used to determine the number of isolation and fragmentation scans to average, otherwise the target was skipped. For each target precursor with enough signal abundance to compel MS/MS, a separate Thermo Scientific .raw file was produced; the collection of which was then batch processed by an application called OFFLINE AUTOMATION cRAWLER. This software determined the mass of the precursor(s) and fragments with a modified version of the THRASH algorithm (Horn, et al., 2000). This information was passed into an XML-based .puf file for searching by ProSightHT, a module within ProSightPC 2.0 (Thermo Fisher Scientific).

The Automation Warehouse Database. The hub of the automation platform is a database application known as the AUTOMATION WAREHOUSE. This database stores all the information acquired by the platform and correlates it with information from ProSight to provide a "state" for a protein or peptide project (figure 3.3). This state can have many functions, from simply acting as a project-wide inclusion or exclusion list to acting as the main repository for analysis results. Many of the features of this database were designed to be modifiable by the user to better serve custom data acquisition needs.

The main tables of the database are the thrash and species tables (figure 3.4). The thrash table stores data for all the isotopic distributions found in the LC-MS runs by THRASH. The species table stores binned data from the THRASH table and correlates it with MS/MS search results from ProSight. Other information about the LC-MS run such as the plate information, LC parameters and information about the data and THRASH output files is stored in the tables to the left of figure 3.3. The fragmentation table stores sets of fragmentation parameters for various methodologies (CAD, IRMPD, ECD) so that a different set of parameters can be selected if a species that did not produce adequate fragmentation data is seen in a subsequent LC-MS run.

The relationship between the thrash and species tables is an important one. Identical species identified in the mass spectrometer are unlikely to have identical exact masses and so some sort of collation or binning of these results must be performed to reduce the dataset to a manageable size and to avoid re-targeting the same species for MS/MS analysis. The solution implemented in the AUTOMATION WAREHOUSE is to store the raw, unfiltered data in the thrash table and the reduced data in the species table. Imported isotopic distributions from THRASH are automatically binned at a user-defined mass tolerance and these new bins are merged with the existing data in the species table. This binning filter also includes a mode where species that differ by a few integral Da, but otherwise match to high mass tolerance can be binned together. This mode handles the case where THRASH mis-assigns the monoisotopic peak which can become common above 5 kDa. New species table entries from this step are indicated as having an identified mass but no MS/MS characterization.

The species table is the table that is queried by the TARGET EXTRACTOR, integrated with the MS/MS analysis results from ProSight and gives the "state" of a proteome project. Each species entry is related to one or more entries in the thrash table which have been binned

together. Optionally, species can be entered in the table without being linked to any thrash entries, allowing priorities to be set for as-yet-unobserved targets. Two extra fields in the species entry allow for alternative prioritization of a species for MS/MS outside of the normal intensity-based criteria. The first of these is the prioritization field, which is a user-specified signed integer. If this value is positive, the species is selected for MS/MS without regard to intensity in order of priority. Species with negative priority values are ignored when creating MS/MS target lists and species with a priority of zero undergo normal selection. The second of these prioritization fields is the "degree of characterization" field. This field indicates how well the protein or peptide has been characterized, with zero indicating "unobserved" (for manual entries with no corresponding MS data) and positive values indicating increasing levels of knowledge about the primary structure of the protein or peptide. The semantics of the value can be determined by the user but the existing platform uses values such as "observed", "fragmented", "partially characterized" and "fully characterized" to indicate that a mass has been observed, that a species has undergone MS/MS, that a species has partially-localized modifications and that a species is fully characterized, including post-translational modifications, respectively. Interaction with ProSight is manual at this time. Users must enter the mass and identification from ProSight as well as the characterization level into the database species table by hand.

Although this work is centered around high-resolution FTMS data, the AUTOMATION WAREHOUSE is agnostic with respect to the mass analyzer and a platform based on a lower-resolution instrument such as a basic LTQ ion trap instrument could be constructed around the database. The WAREHOUSE, being based on an open-source relational database, is also easily extensible if the user desires to store other data alongside the canonical database tables and fields.

Protein Database Searching Using ProSight. ProSight .puf files were iteratively searched against the appropriate top-down (69 435 basic sequences, 1 565 945 protein forms, 978 MB) or middle-down (3 378 894 basic sequences, 6 051 898 peptide forms, 2.5 GB) ProSight database, both shotgun annotated (Pesavento, et al., 2004). For top-down experiments, two absolute mass searches were performed, followed by a biomarker search, both against a heavily annotated human database previously described (Roth, et al., 2008). The first absolute mass search used a 10 Da precursor mass tolerance, while the second used a 300 Da tolerance. The biomarker search was performed with a 1.1 Da precursor mass tolerance. Fragment tolerance for all three searches was 10 ppm, and the expectation value (probability score (Meng, et al., 2001) \times database size) threshold to define a positive identification was conservatively set at 10^{-5} . Final results were exported to a Microsoft Excel .xls file by ProSightPC. Due to extensive modifications in the database and experimental data, the top-down results were also manually curated to ensure that only a single protein form that shows the maximum support in the fragmentation data is reported per precursor. For middle-down experiments, two absolute mass searches were performed against an *in silico* digested human database that contained all Lys-C peptides from 1-50 kDa with up to 4 missed cleavages. The first absolute mass search used a 5 Da precursor mass tolerance and a 10 ppm fragment tolerance. MS/MS experiments that did not yield an expectation value within the strict confidence threshold of $\leq 10^{-5}$ were automatically re-searched with a 200 Da intact mass window. Final results were exported to a Microsoft Excel .xls file by ProSightPC.

3.3 Application of the Platform to Soluble Proteins From HeLa Nuclei

Cell Processing and Sample Preparation. Washed human HeLa cell pellets ($\sim 2 \times 10^7$ cells) were suspended in nuclear isolation buffer (NIB-250): 15 mM tris-hydrochloric acid (pH 7.5), 60 mM potassium chloride, 15 mM sodium chloride, 5 mM magnesium chloride, 1 mM calcium chloride, 250 mM sucrose, 1 mM dithiothreitol, 10 mM sodium butyrate, protease inhibitor cocktail set III (Calbiochem; San Diego, CA) at a 100:1 v:v ratio, and phosphatase inhibitor cocktail set II (Calbiochem) at a 100:1 v:v ratio plus 0.3% NP-40 at a 10:1 v:v ratio. Cells were lysed by gentle mixing and incubation on ice for 5 min. Nuclei were pelleted at $600 \times g$ for 5 min at 4 °C and then washed twice with NIB-250 without detergent.

For top-down MS/MS, 0.4 N sulfuric acid was added to HeLa nuclei to give a 3:1 v:v ratio. The acid-extracted nuclei were maintained at 4 °C for 30 min and centrifuged at $2000 \times g$. The supernatant was transferred to a 1.5 mL microcentrifuge tube and centrifuged again at 14 000 rpm for 20 min. This supernatant (200 μ L) was mixed with 150 μ L of chromatography solvent A - water + 0.2% formic acid and 0.01% trifluoroacetic acid (TFA) - prior to injection. For middle-down, isolated nuclei were suspended directly in lysis buffer containing 50 mM ammonium bicarbonate, 1 mM dithiothreitol, 10 mM sodium butyrate, 2 M urea, and 10 nM microcystin. Nuclei were lysed with pulsed sonication six times for 30 s each, and to the unclarified lysate, 20 ng of endoproteinase Lys-C (Wako Chemicals; Richmond, VA) was added to give roughly a 250:1 substrate-to-enzyme ratio. The nuclear lysate was digested overnight at 37 °C. Prior to injection, the digest was clarified at 14 000 rpm for 20 min. Chromatography solvent A was added to the supernatant to double the volume, and the sample was reclarified to remove any precipitate.

Liquid Chromatography. Top-down and middle-down samples were injected with a Gilson 235P autosampler (Middleton, WI) into an Agilent 1200 binary HPLC system with

degasser (Santa Clara, CA). A flow rate of 100 $\mu\text{L}/\text{min}$ was used with PLRP-S 1000 \AA , 5 μm , 150 mm \times 1.0 mm polymer columns (Higgins Analytical; Mountain View, CA). The gradient lasted 116 min; samples were injected with 95% solvent A (water with 0.2% formic acid and 0.01% TFA) and 5% B (90:10 acetonitrile:isopropyl alcohol with 0.2% formic acid and 0.01% TFA) as starting conditions for 5 min. The linear gradient ramped to 30% B at 10 min and to 50% B at 106 min. A majority of the proteins/peptides eluted between 30 and 50% B. At 111 min, the gradient reached 95% B and was maintained until 116 min. The TriVersa NanoMate (Advion BioSciences; Ithaca, NY) was used in LC-MS fraction collection mode with a split such that 300 nL/min was infused into the mass spectrometer via the chip-based nanoelectrospray ionization source and the remaining 99.7 $\mu\text{L}/\text{min}$ was collected for subsequent offline analyses. The first 15 min of the gradient were directed to waste. Electrospray started at 16 min, when both data acquisition and fraction collection began, and ended after fraction 96 at 111 min. The electrospray voltage was typically +2.0 kV.

Mass Spectrometry. The mass spectrometer used was a Thermo Scientific 12 T LTQ FT Ultra running LTQ Tune Plus 2.2 and Xcalibur 2.0.5 (San Jose, CA/Bremen, Germany). For top-down experiments, the instrument method consisted of nine steps of “zoom mapping”, or data-independent ion trap isolation windows, detected by FT and with no subsequent fragmentation. The center of the isolation windows progressed from m/z 700 to 1100, with an isolation width of 60 m/z and a step size of 50 m/z to ensure overlap at the edges of the isolation windows. The detection range for all FT events was m/z 600-1200. This was done to ensure that all scans have sufficient data past the region of interest for the data analysis software to function optimally. After the fifth ion trap window centered at m/z 900, a full ion trap scan from m/z 600-1600 was included to enable optional assessment of data quality, but it was not analyzed by the

software. Automatic gain control (AGC) targets were increased from the default of 2×10^5 to 1×10^6 for MSⁿ FTMS, while the full ion trap was left at the default of 3×10^4 . The number of microscans was 1 except where noted.

For middle-down experiments, the instrument method consisted of full FT scans (5 microscans) from m/z 500-1500, since zoom mapping fails to cover enough m/z space in which peptide precursors occur to be effective. AGC targets were increased from 5×10^5 to 2×10^6 for full FTMS. For both top-down and middle-down experiments, maximum injection times were increased from the default of 500 to 4000 ms for full FTMS, 1000 to 8000 ms for MSⁿ FTMS, and 10 to 80 ms for full ITMS. FT resolving power was always $\sim 171\,500$ (nominally 100 000 in the software, based on a 7 T ion cyclotron resonance (ICR) cell) at m/z 400. Source-induced dissociation voltages of +10-20 V were applied to all scan events to reduce adducts.

For offline experiments, the TriVersa NanoMate was switched to direct infusion mode. An electrospray voltage of +1.8 kV and a backing gas pressure of 0.6 psi were used. The isolation width was typically 5 m/z for middle-down and 8 m/z for top-down. Collision-induced dissociation (CID) parameters were: normalized collision energy (NCE) of 0.41, activation Q of 0.5, and activation time of 50 ms.

Top-Down Proteomics. For intact proteins obtained from acid-extracted HeLa nuclei, LC-MS (2 microscans) resulted in the system recognizing 535 targets above 25 signal-to-noise ratio (S/N) in the online, intact-only data. For the offline mode, the system set up an accurate mass list for these targets in 73 of the 99.7 μL fractions collected in the whole 96-well plate. Of the 535 species targeted, MS/MS experiments were actually performed on 382 by the instrument, yielding 305 top-down identifications with ProSight expectation scores below 10^{-5} . Identified proteins ranged from 4-16 kDa.

These 305 identifications from human HeLa cells collapse to 57 forms from 30 unique genes, including several for all core histones (H2A, H2B, H3, H4), high mobility group proteins (HMGA, HMG2, HMG1, HMGA1), ribosomal protein 40S, and small ribonucleoproteins. By comparison, an online zoom mapping with fragmentation run of the same sample yielded 16 identifications from 16 genes with expectation values ranging from 10^{-13} to 10^{-102} . Of these genes, 10 were unique to the online run.

Middle-Down Proteomics. Online RPLC was run directly on a Lys-C digest of HeLa nuclear lysate, with the column eluent automatically split and the TriVersa NanoMate collecting 99.7 μ L fractions (a total of 96 fractions). Of these, 80 fractions were analyzed by automated MS/MS with an accurate mass target list obtained from peptides observed in the online run. A typical fraction is displayed in Figure 3.5, with a single scan shown in the center and six typical isolation windows shown as insets. Of the seven peptides identified from the six MS/MS spectra, expectation values ranged from 10^{-6} to 10^{-14} , with one example of multiplexed identifications (bottom right).

Data from all 1233 MS/MS experiments performed over the 80 fractions were iteratively searched with multiplexed searching enabled to intelligently manage multiple hits per spectrum. This resulted in identification of 256 peptides ranging from 1-13 kDa, of which 147 were unique, with expectation values from 10^{-5} to 10^{-84} . By comparison, an online data-dependent LC-MS/MS experiment of the same sample yielded 77 peptide identifications, of which 66 were unique. Of the 66 peptide forms, 31 were unique to the online run.

In the most complex region of the chromatogram, 20-50 accurate mass targets were typically identified per well. Of the 147 unique peptides, 29 were modified with 25 of these being N-terminal acetylation. At 12 kDa, one exhaustive Lys-C peptide was particularly large

(Figure 3.6), and the fragmentation data suggested two forms of the protein hnRNP A2/B1 (P22626) from the ProSight database. The hnRNP A2/B1 was known to harbor a monomethylation at Arg203, partially characterized in this study at ~25% occupancy (Figure 3.6c) without any other modifications on this 130-residue segment of the protein.

Data Acquisition Times for High-Resolution MS/MS Spectra. For top-down, target abundances directed the system to choose between 25, 50, or 100 fragmentation scans. This, in turn, sets the overall data acquisition times (along with AGC and maximum injection time settings), which ranged from 2-13 min for intact protein samples. The data for a tray of 73 sample wells collected from the online LC-MS run took ~15 h of instrument time to attempt 535 top-down MS/MS experiments. For middle-down, target abundances directed the system to choose between 10, 25, or 50 fragmentation scans. This resulted in MS/MS spectral acquisition times per target ranging from 1-5 min, translating to 36 h of instrument time to run the 1233 targets from the 80 fractions noted above. The duty cycle for this platform in the offline mode - the fraction of time the instrument is either accumulating or detecting ions for high-resolution MS¹ or MS² data acquisition - is typically over 90%.

3.4 Conclusions

There is no system currently available capable of acquiring ultrahigh-resolution tandem mass spectra with the sensitivity of an ion trap. Custom data acquisition systems have previously been developed but only for online bottom-up proteomics with low-resolution instruments (*e.g.*, triple quadrupoles) (Ducret, et al., 1998; Stahl, et al., 1996). Therefore, we have constructed an online-offline data acquisition system representing a significant advance toward using high-value mass spectrometer time automatically and more efficiently, with the midrange goal of increasing

the number of unique proteins and peptides identified and characterized in complex mixtures. The AUTOMATION WAREHOUSE database functions as a high-resolution exclusion and inclusion list spanning multiple LC-MS runs and supporting an entire proteome project.

The power of the automated system is demonstrated by a comparison to more established online experiments for both top-down and middle-down human proteomics. This is illustrated with Venn diagrams in Figure 3.7. For top-down, the automated system identifies proteins from approximately twice as many genes than an online zoom mapping experiment (Figure 3.7a). However, when all protein forms were counted, the automated system identifies nearly 4 times as many (Figure 3.7b), exemplifying the superior characterization power of the platform. For middle-down, the automated system identifies well over twice as many peptide forms as a traditional data-dependent double-play experiment (Figure 3.7c).

Advanced Data Acquisition. The data acquisition system has a number of features uncommon in modern commercial software on both the source and mass spectrometer side. On the source side, the COM library included in Advion ChipSoft version 8 facilitates constant monitoring and dynamic control of electrospray conditions. Users can manually adjust the electrospray voltage and gas pressure in real time, but more importantly, the software automatically checks the electrospray current against user-specified thresholds before every scan. When the spray current is not within the user-specified range, the system executes a predefined sequence of actions until the current is restored to an acceptable level. This sequence of actions includes momentarily raising the gas pressure, obtaining more sample, retrieving a new tip, using a new electrospray nozzle, and finally skipping the current well, in that order. This represents an important advance in offline nanospray that is only possible due to the tight integration of the TriVersa NanoMate and the LTQ FT Ultra.

On the mass spectrometer side, the software performs a custom workflow to ensure the data acquisition system minimizes time wasted on targets unlikely to produce an identification. Before normal data acquisition begins, the system acquires a low number (usually one) of preview isolation scans on every target selected for offline interrogation. The purpose of this preview scan event is twofold: it allows the system to determine if the target detected online is present offline at sufficient absolute signal abundance to warrant further acquisition and to determine how many isolation and fragmentation scans should be acquired. Both of the solutions above are by no means foolproof, but represent a significant advance toward emulating manual, human-controlled offline data acquisition. The overall outcome is a fully automated acquisition system at offline run time, operating continually for several days without intervention at the current stage of development.

Expectation values for peptides that are below 10^{-5} allow direct and error-tolerant identification of a protein without resorting to decoy/reverse database construction and searching (Elias, et al., 2005) or identifying multiple peptides from the same protein. Therefore, future comparisons of proteome coverage obtained by different data acquisition strategies will be interesting, as obtaining high-resolution MS/MS is contrasted with the lower-resolution MS/MS experiments that now dominate data acquisition for shotgun proteomics.

Offline Advantages. In performing the bulk of the data acquisition offline, the increased spray time can be used to average multiple scans with more ions accumulated before detection. An offline mode of operation also allows the prior information of the entire chromatographic run to be available in determining what to fragment; therefore, more intelligent decisions can be made in terms of when precursors should be fragmented and which charge state should be selected for MS/MS. Data analysis is also simplified because each target is acquired in a

separate data file. Although the abundance of precursors is typically reduced versus the maximal instantaneous concentration during elution (due to dilution and possibly sample degradation in sample wells), automated offline acquisition is still able to collect high-quality data often far superior to online LC-MS/MS for targets identified in both modes. The greatly increased time available during offline acquisition yields improved fragmentation through averaging spectra, as demonstrated in Figure 3.8 with human high mobility group protein 17 (expectation value online 10^{-3} versus offline 10^{-98}). Additionally, automated offline acquisition expands the depth of the proteome's dynamic range accessible with the characterization power of FTMS.

The ability to average scans offline also allows for routine multiplexed identifications. This occurred in 6% of the offline middle-down experiments attempted. When present at low abundance, 3-50 kDa precursor ions are particularly challenging to identify and characterize with MS/MS data obtained on a chromatographic time scale. Until mass spectrometers can produce such high-quality data sets at the resolution of a FT but with the speed and sensitivity of an ion trap, the current system now stands as a viable option for large-scale proteome projects.

Future Development. A critical development planned for the future is the linking of the ProSightHT database and the AUTOMATION WAREHOUSE database. Currently this feedback loop connecting prior database hits to future data acquisition runs must be performed manually. This enhancement will facilitate automated population of the AUTOMATION WAREHOUSE database with confident protein/peptide identifications, enabling it to function as a true high-resolution exclusion list for a proteome project. Well-characterized species in the database will be low priorities in the target selection stage, furthering the goal of increased proteome coverage using top-down and middle-down strategies.

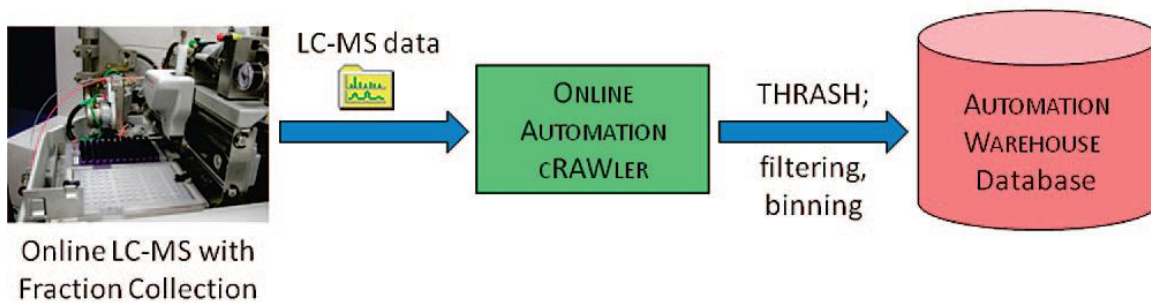
Although the current system represents the state-of-the-art for automated offline mass spectrometry, there are numerous opportunities for improvement. One concept that has successfully been applied to similar platforms in the past is automated determination of fragmentation parameters based on empirical data (Johnson, et al., 2002; Patrie, et al., 2006; Zamdborg, et al., 2007). An even more sophisticated possibility is automated dynamic adjustment of those parameters based on data surveyed in real time. Also promising is decision-making based on real-time spectral analysis in order to determine when averaging more scans is producing diminishing returns in terms of the number of new fragment ions or significant improvement in database retrieval scores (Johnson, et al., 2002), further optimizing the use of instrument time. Stahl *et al.* pioneered this concept using either the total ion current (TIC) of the most recent product ion spectrum or “spectrum reproducibility,” based on the abundance of the top three fragment peaks, depending on the sample levels (Stahl, et al., 1996).

In the future, it is critical that this system be compared to another promising route, the use of “smart” LC-MS/MS using data acquisition software that makes sophisticated decisions on-the-fly. Although commercial instrument firmware has progressed greatly in recent years, making concepts such as data-dependent acquisition, inclusion/exclusion lists, and neutral-loss experiments routine, there are several other advanced strategies to be implemented to better use high-value instrument time. Recent development of “decision tree” proteomics, where the fragmentation method is determined in real time based on precursor m/z and charge state, represents a significant first step toward this goal (Swaney, et al., 2008). In the future, such rapid experimental logic could be extended, for example, by querying a proteome project-wide database before deciding on fragmentation targets. The advantage of such a platform would mean workflows would be left relatively unchanged from current LC-MS/MS, although there is

undoubtedly a limit to the proteomic depth achievable with online MS/MS alone, particularly with top-down and middle-down using contemporary instrumentation. For some applications, the recently introduced concept of a “replay” run may be a feasible alternative that lies between a completely offline or completely online approach (Waanders, et al., 2008).

Extensibility. Although our focus is clearly top-down and middle-down analysis of proteomic samples with Fourier transform ion cyclotron resonance (FTICR) MS, the platform functions interchangeably with the Thermo Scientific LTQ Orbitrap or even standalone LTQ linear ion trap instruments. Additionally, the system could be adapted for other types of samples, such as small molecules or small peptides (*i.e.*, bottom-up proteomics; Luo *et al.* recently noted the limitations of data-dependent acquisition in a shotgun experiment (Luo, et al., 2008)), without much effort. We introduce this automated online-offline engine as a general approach to acquire high-quality, information-rich tandem mass spectra for species not identified or characterized on a chromatographic time scale.

a) Online Workflow



b) Offline Workflow

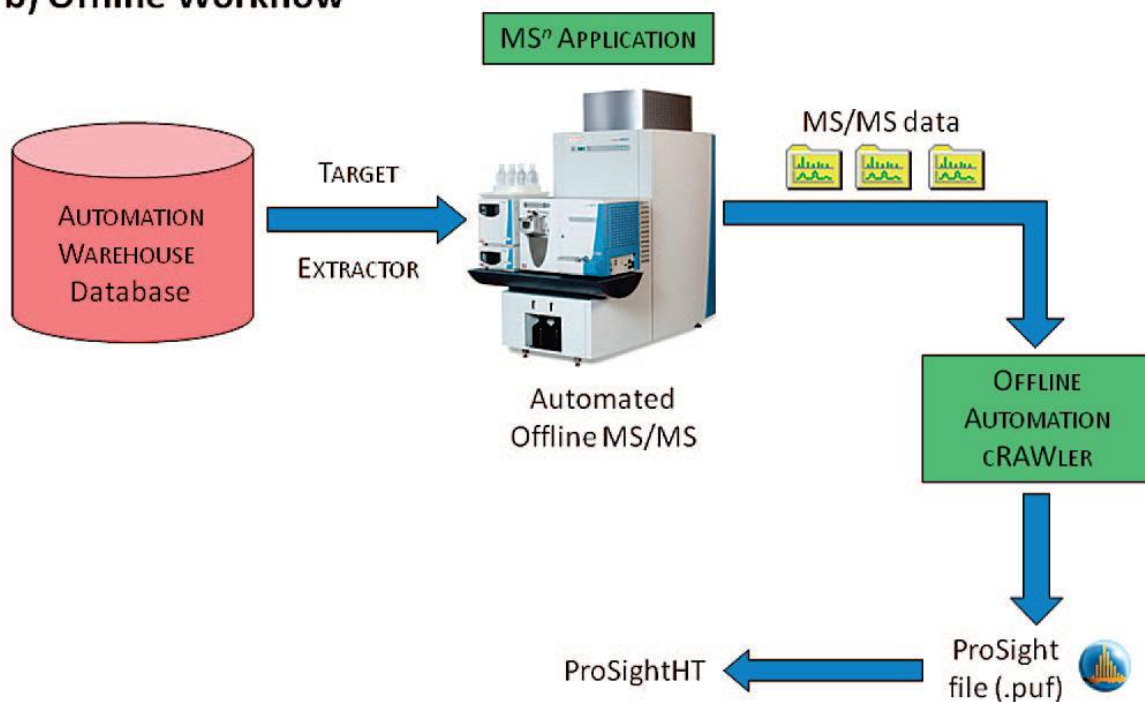
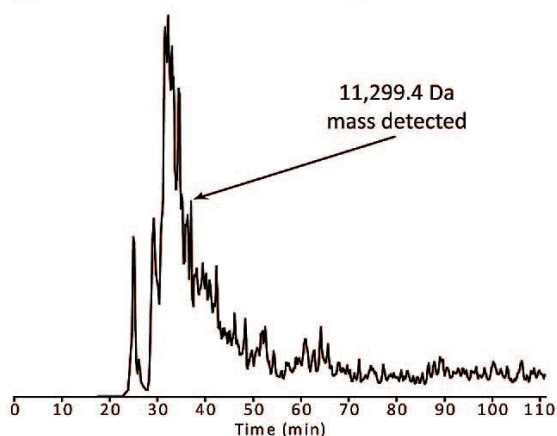
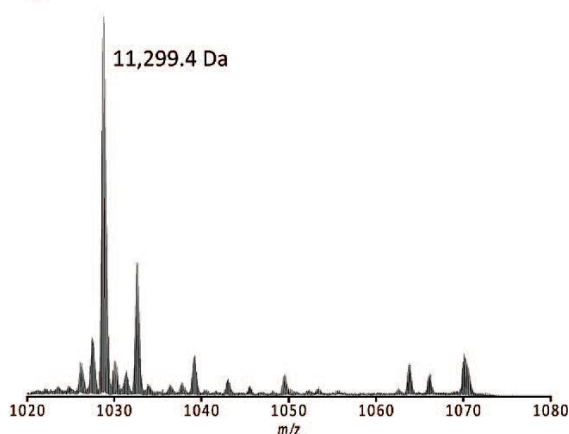


Figure 3.1. The workflow of the automation platform, with the online (a) and offline (b) portions separated. The Automation Warehouse database is the link between the online and offline segments of the experiment.

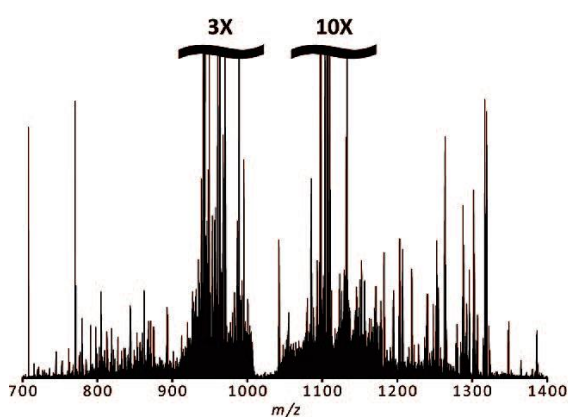
a) Base Peak Chromatogram



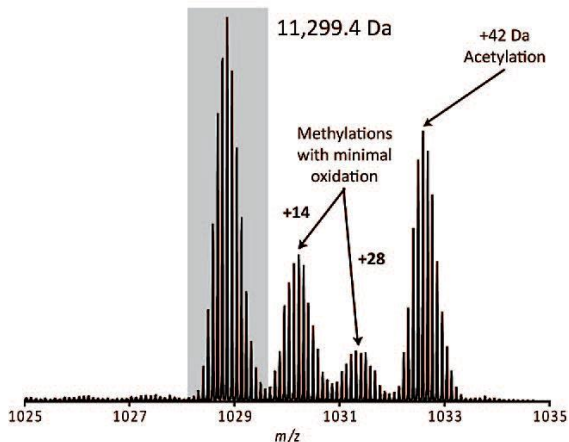
b) Online Isolation



d) Offline Fragmentation



c) Offline Isolation



e) ProSight Output

► >H4_HUMAN, P62805, Histone H4. | >H4_HUMAN, P62805, Histone H4.. (Type: *predicted*, Signal Peptide: *false*, Propep: *false*)

b1 - S - G - R - G - K } G - G - K } G - L } G - K } G - G - A - K - R } H - R } K - V - L - R } D } N - I } Q } G - I - T - y73

b31 - K } P - A - I - R } R } L - A - R } R } G } G - V - K - R } I } S } G - L - I - Y - E } E } T - R } G } V } L } K } V - y43

b61 } F - L - E } N - V - I - R } D } A } V } T } Y } T - E } H - A - K } R } K - T } V } T - A - M } D } V } V } Y } A - L - y13

b91 - K } R } Q } G - R - T - L } Y - G - F - G - G - y1

ID/Gene	Length	Mass	Mass Diff	PPM Diff	B Ions	Y Ions	Total Ions	PDE Score	Expectation	Lambda	P Score
531163 1130355	102	11299.4	-0.0314	-2.7777	16	43	59	74.2	1.49E-51	-7.60065	1.01E-57

Take to Sequence Gazer

RESID SEQ

Figure 3.2. An example of the online-offline strategy. A protein elutes approximately 37 min into the RPLC gradient, as depicted on the broadband ion trap base peak chromatogram (a). Zoom mapping generates an ion trap isolation window spectrum detected by FTMS, generating a high accuracy mass observation (b). After analysis of the online run, this species is automatically targeted for offline isolation (c) and fragmentation (d), both detected by FTMS. The top ProSight identification (e) shows robust fragmentation of human histone H4 that is acetylated at the N-terminus (or lysine 5) and dimethylated at lysine 20.

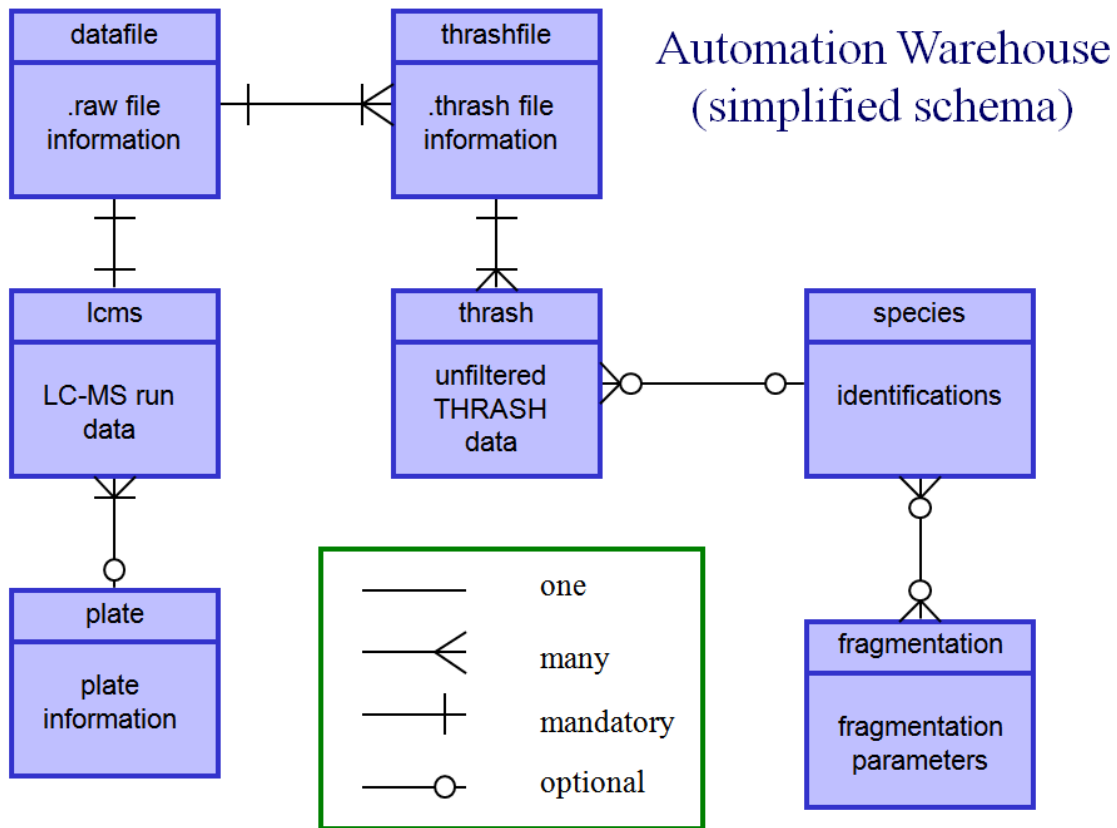


Figure 3.3. Entity-relationship (ER) diagram showing the schema of the AUTOMATION WAREHOUSE database. The thrash and species tables are the core of the database, with the thrash table storing the unfiltered isotopic distribution data and the species table containing protein/peptide identification data. The relation between these two tables correlates binned and filtered data from the thrash table with the identification data in the species table.

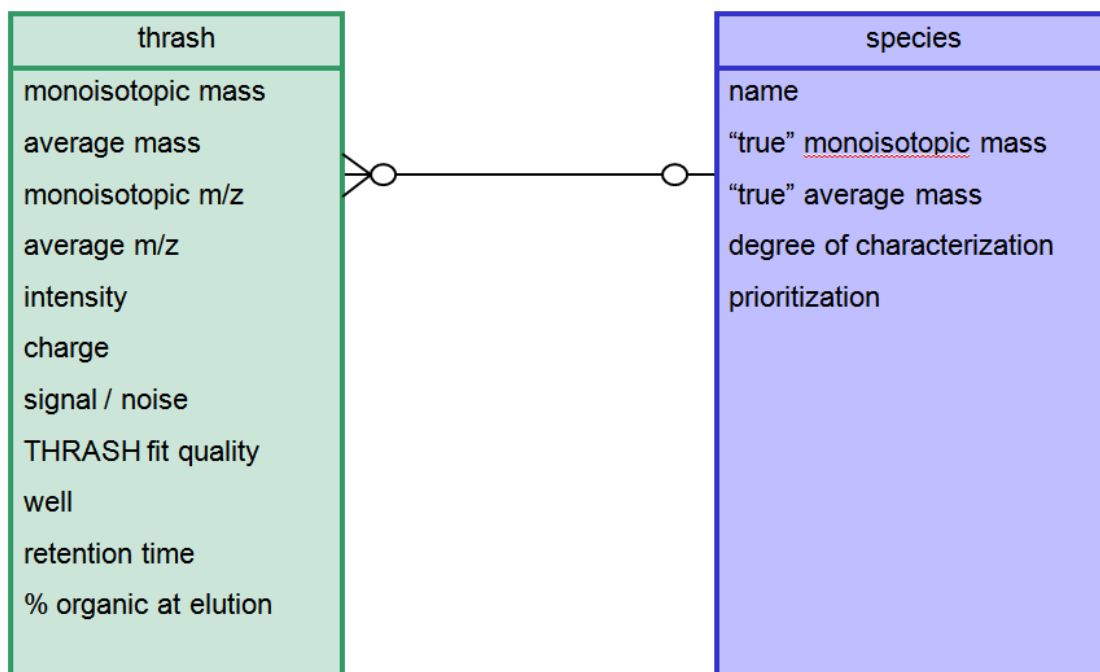


Figure 3.4. Data elements stored in the thrash and species tables of the AUTOMATION WAREHOUSE. The "true" mass values in the species table are the calculated masses of the protein or peptide from the sequence data, including any post-translational modifications or processing. The degree of characterization field is a user-specifiable integer value which indicates the level of characterization of a species. The prioritization field is a user-specified integer field that allows the software to flag particularly interesting species for processing out of the normal order.

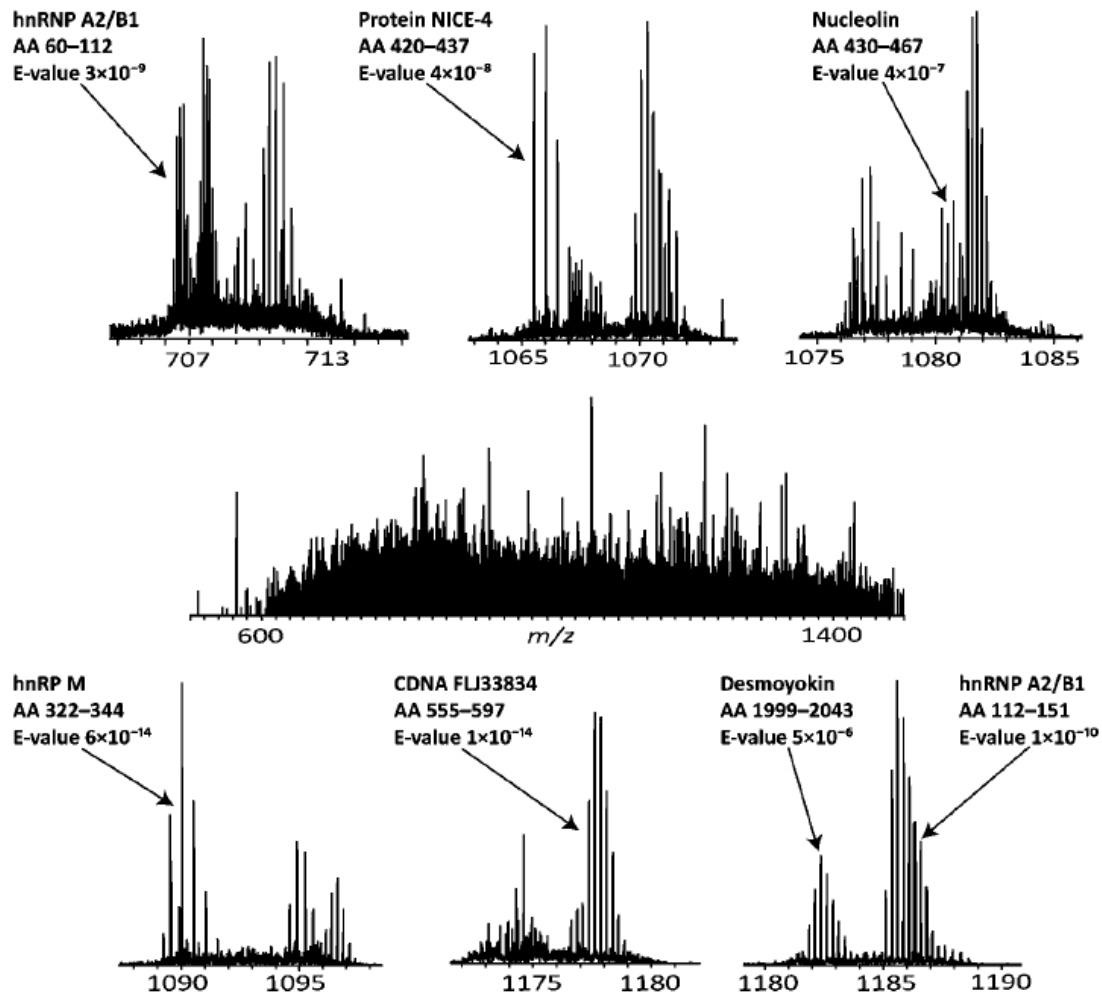
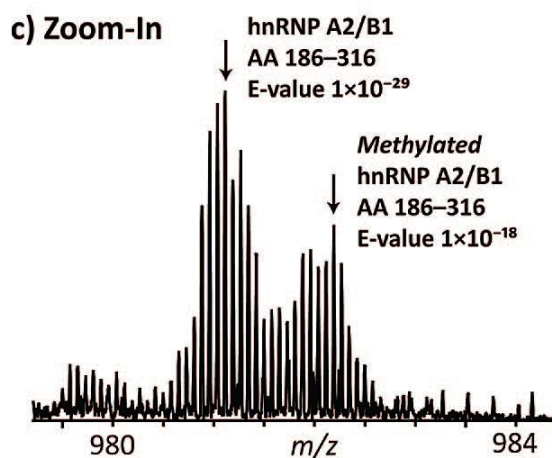
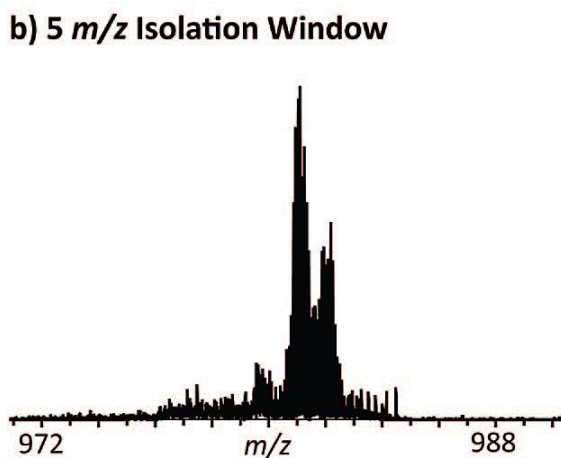
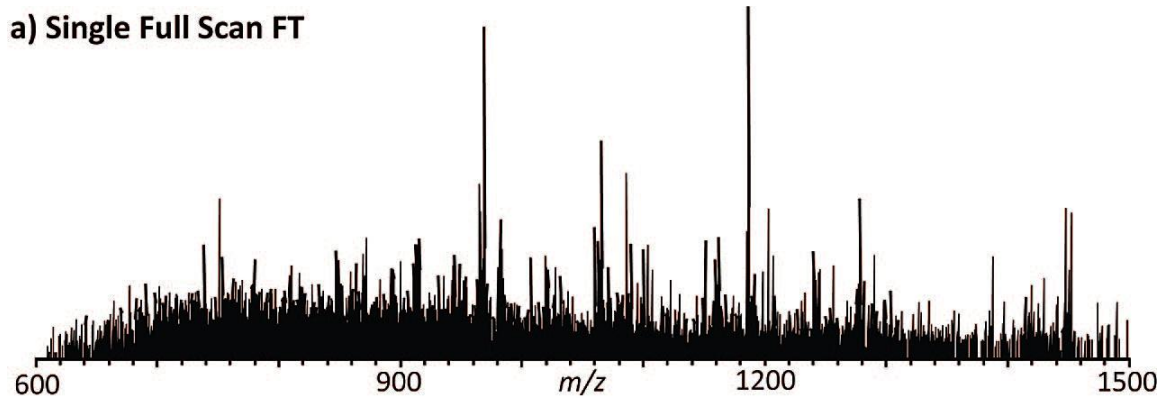


Figure 3.5. An example of the complexity in a typical middle-down human proteomics sample. Middle panel: offline FT mass spectrum (single scan) obtained from a 99.7 μL sample collected over 1 min of an LC-MS run. Seven peptides, all from unique proteins, were identified from these six isolation windows (insets), including an example of multiplexed identifications (bottom-right).



d) ProSight Output of Methylated Form

Methylation (mono)

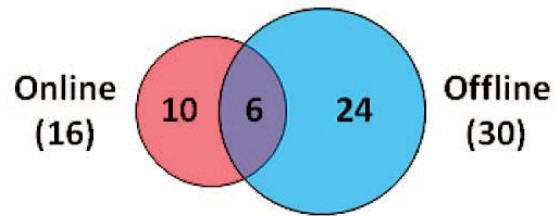
ID/Gene	Length	Mass	Mass Diff.	PPM Diff.	B Ions	Y Ions	Total Ions	PDE Score	Expectation	Lambda	P Score	
<p>>ROA2_HUMAN, P22626, Lys-C peptide from AA 187-317 in Heterogeneous nuclear ribonucleoproteins A2/B1 (hnRNP A2 / hnRNP B1) with 0 missed cleavage sites >ROA2_HUMAN, P22626, Lys-C peptide from AA 186-316 in Heterogeneous nuclear ribonucleoproteins A2/B1 (hnRNP A2 / hnRNP B1) with 0 missed cleavage sites. (Type: predicted, Signal Peptide: false, Propep: false)</p> <p>b1 · A · L · S · R · Q · E · M · Q · E · V · Q · S · S · R · S · G · R · G · G · N · F · G · F · G · D · S · R · G · G · G · y102 b31 · G · N · F · G · P · G · P · G · S · N · F · R · G · G · S · D · G · Y · G · S · G · R · G · F · G · D · G · Y · N · G · y72 b61 · Y { G { G { P · G · G · G · N · F · G · G · S } P · G · Y · G · G · G · R · G · G · Y · G · G · G · G { P · G · Y · y42 b91 { G · N · Q { G · G · G · Y { G { G { G · Y · D · N · Y { G · G · G · N · Y { G · S { G { N · Y · N { D · F · G · N · Y · y12 b121 { N } Q · Q { P · S · N } Y · G · P · M · K · y1</p>												
6567971	14712580	131	12747.4	.9651	75.7004	6	20	26	42.6	1.26E-18	0	2.11E-25

Take to Sequence Gazer

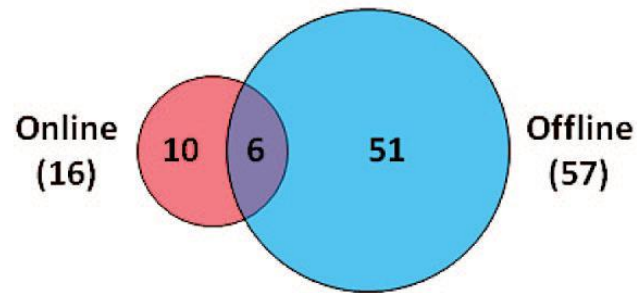
RESID SEQ

Figure 3.6. (a) A single scan FT mass spectrum of a fraction collected over 1 min of an LC-MS run. (b) FT mass spectrum (5 scans) of ions in an isolation window from targeting a 13 kDa species detected in the LC-MS run. (c) Expansion of the m/z 980-984 region of the data in part b, showing an exhaustive Lys-C peptide from the hnRNP A2/B1 protein identified in both its unmodified and monomethylated form. Of the 26 matching fragment ions observed, six were consistent with the known monomethylation at Arg203 that was stored in the ProSight database created for searching Lys-C peptide data. (d) ProSight fragment map of the methylated peptide form shown in part c.

a) Top-Down Genes



b) Top-Down Protein Forms



c) Middle-Down Peptide Forms

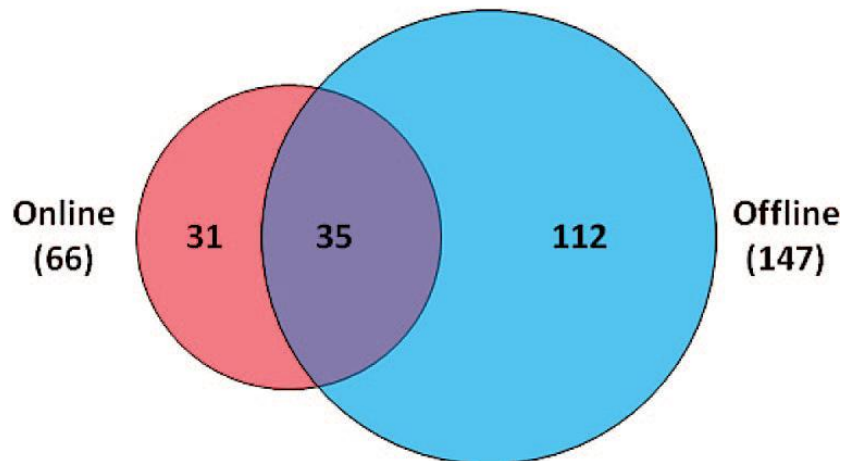
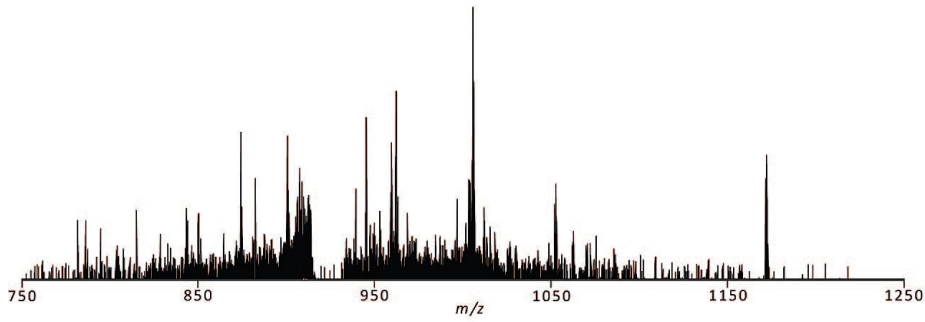


Figure 3.7. Venn diagrams (to scale) comparing traditional online experiments to the automated offline system. For top-down, approximately twice as many genes (a) and nearly 4 times as many protein forms (b) are identified with the new offline platform compared to an online-only approach. For middle-down (c), well over twice as many peptide forms are identified, including many co- and posttranslational modifications.

a) Online Fragmentation (6 scans)

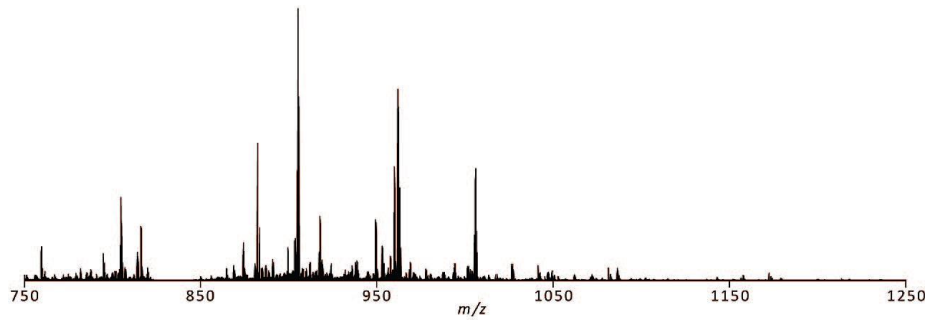


ID/Gene	Length	Mass	Mass Diff.	PPM Diff.	B Ions	Y Ions	Total Ions	PDE Score	Expectation	Lambda	P Score	
<p>► >HMG2_HUMAN, P05204, Nonhistone chromosomal protein HMG-17 (High-mobility group nucleosome-binding domain-containing protein 2). >Q0VGD5_HUMAN, Q0VGD5, High-mobility group nucleosomal binding domain 2.. (Type: <i>predicted</i>, Signal Peptide: <i>false</i>, Propep: <i>false</i>)</p> <p>b1 P-K-R-K-A-E-G-D-A-K-G-D-K-A-K-V-K-D E-P-Q-R-R-S-A-R-L-S-A-K· y60</p> <p>b31 P-A-P-P-K P-E P-K P-K-K-A-P-A-K K G-E-K-V-P-K-G-K-K-G-K-A-D· y30</p> <p>b61 A-G-K-E-G-N-N-P-A-E-N-G-D A-K-T-D Q-A-Q-K-A-E-G-A-G-D-A-K· y1</p>												
542356	1130641	89	9256.02	-0.729	-7.8762	7	0	7	126	0.00206	9.94544	1.4E-09

Take to Sequence Gazer

RESID SEQ

b) Offline Fragmentation (25 scans)



ID/Gene	Length	Mass	Mass Diff.	PPM Diff.	B Ions	Y Ions	Total Ions	PDE Score	Expectation	Lambda	P Score	
<p>► >HMG2_HUMAN, P05204, Nonhistone chromosomal protein HMG-17 (High-mobility group nucleosome-binding domain-containing protein 2). >Q0VGD5_HUMAN, Q0VGD5, High-mobility group nucleosomal binding domain 2.. (Type: <i>predicted</i>, Signal Peptide: <i>false</i>, Propep: <i>false</i>)</p> <p>b1 P-K-R-K-A-E G-D A K G-D K A K V-K D E P-Q-R-R-S-A-R-L-S-A-K· y60</p> <p>b31 P-A P P-K P-E P-K P-K K A P-A-K K G-E K V P-K G-K K G-K A-D y30</p> <p>b61 A-G-K E-G-N-N P-A E N G-D A K T-D Q-A Q-K A E G-A-G-D-A-K· y1</p>												
542356	1130641	89	9256.02	-1.08	-116.693	39	28	67	348	3.2E-98	6.37504	2.18E-104

Take to Sequence Gazer

RESID SEQ

Figure 3.8. A comparison of online (a; 6 scans) and offline (b; 25 scans) fragmentation for human high mobility group protein 17 (HMG-17), showing 95 orders-of-magnitude improvement in expectation value, turning a rejected hit into a confident identification. Offline acquisition affords the time for averaging multiple scans to improve fragmentation coverage in top-down and middle-down proteomics, facilitating protein/peptide identification and characterization by FTMS.

3.5 Literature Cited

Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics, *Nature*, **422**, 198-207.

Annan, R.S., Huddleston, M.J., Verma, R., Deshaies, R.J. and Carr, S.A. (2001) A multidimensional electrospray MS-based approach to phosphopeptide mapping, *Anal Chem*, **73**, 393-404.

Bakalarski, C.E., Haas, W., Dephoure, N.E. and Gygi, S.P. (2007) The effects of mass accuracy, data acquisition speed, and search algorithm choice on peptide identification rates in phosphoproteomics, *Anal Bioanal Chem*, **389**, 1409-1419.

Davis, M.T. and Lee, T.D. (1997) Variable flow liquid chromatography tandem mass spectrometry and the comprehensive analysis of complex protein digest mixtures, *J Am Soc Mass Spectr*, **8**, 1059-1069.

Davis, M.T. and Lee, T.D. (1998) Rapid protein identification using a microscale electrospray LC/MS system on an ion trap mass spectrometer, *J Am Soc Mass Spectr*, **9**, 194-201.

Davis, M.T., Stahl, D.C., Hefta, S.A. and Lee, T.D. (1995) A Microscale Electrospray Interface for Online, Capillary Liquid-Chromatography Tandem Mass-Spectrometry of Complex Peptide Mixtures, *Anal Chem*, **67**, 4549-4556.

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

Elias, J.E., Haas, W., Faherty, B.K. and Gygi, S.P. (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations, *Nature Methods*, **2**, 667-675.

Forbes, A.J., Mazur, M.T., Patel, H.M., Walsh, C.T. and Kelleher, N.L. (2001) Toward efficient analysis of >70 kDa proteins with 100% sequence coverage, *Proteomics*, **1**, 927-933.

Good, D.M., Wirtala, M., McAlister, G.C. and Coon, J.J. (2007) Performance characteristics of electron transfer dissociation mass spectrometry, *Mol Cell Proteomics*, **6**, 1942-1951.

Haas, W., Faherty, B.K., Gerber, S.A., Elias, J.E., Beausoleil, S.A., Bakalarski, C.E., Li, X., Villen, J. and Gygi, S.P. (2006) Optimization and use of peptide mass measurement accuracy in shotgun proteomics, *Mol Cell Proteomics*, **5**, 1326-1337.

Hofstadler, S.A., Wahl, J.H., Bruce, J.E. and Smith, R.D. (1993) Online Capillary Electrophoresis with Fourier-Transform Ion-Cyclotron Resonance Mass-Spectrometry, *Journal of the American Chemical Society*, **115**, 6983-6984.

Horn, D.M., Zubarev, R.A. and McLafferty, F.W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules, *J Am Soc Mass Spectr*, **11**, 320-332.

Johnson, J.R., Meng, F.Y., Forbes, A.J., Cargile, B.J. and Kelleher, N.L. (2002) Fourier-transform mass spectrometry for automated fragmentation and identification of 5-20 kDa proteins in mixtures, *Electrophoresis*, **23**, 3217-3223.

Kelleher, N.L. (2004) Top-down proteomics, *Anal Chem*, **76**, 196a-203a.

Le Blanc, J.C., Hager, J.W., Ilisiu, A.M., Hunter, C., Zhong, F. and Chu, I. (2003) Unique scanning capabilities of a new hybrid linear ion trap mass spectrometer (Q TRAP) used for high sensitivity proteomics applications, *Proteomics*, **3**, 859-869.

LeDuc, R.D. and Kelleher, N.L. (2007) Using ProSight PTM and related tools for targeted protein identification and characterization with high mass accuracy tandem MS data, *Curr Protoc Bioinformatics*, **Chapter 13**, Unit 13 16.

LeDuc, R.D., Taylor, G.K., Kim, Y.B., Januszyk, T.E., Bynum, L.H., Sola, J.V., Garavelli, J.S. and Kelleher, N.L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry, *Nucleic Acids Res*, **32**, W340-W345.

Li, X., Fekete, A., Englmann, M., Frommberger, M., Lv, S., Chen, G. and Schmitt-Kopplin, P. (2007) At-line coupling of UPLC to chip-electrospray-FTICR-MS, *Analytical and Bioanalytical Chemistry*, **389**, 1439-1446.

Luo, W., Slebos, R.J., Hill, S., Li, M., Brabek, J., Amanchy, R., Chaerkady, R., Pandey, A., Ham, A.J.L. and Hanks, S.K. (2008) Global impact of oncogenic Src on a phosphotyrosine proteome, *J Proteome Res*, **7**, 3447-3460.

Macek, B., Waanders, L.F., Olsen, J.V. and Mann, M. (2006) Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer, *Mol Cell Proteomics*, **5**, 949-958.

Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K. and Horning, S. (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer, *Anal Chem*, **78**, 2113-2120.

Mann, M., Meng, C.K. and Fenn, J.B. (1989) Interpreting Mass-Spectra of Multiply Charged Ions, *Anal Chem*, **61**, 1702-1708.

Marshall, A.G., Hendrickson, C.L. and Jackson, G.S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: a primer, *Mass Spectrom Rev*, **17**, 1-35.

McAlister, G.C., Phanstiel, D., Good, D.M., Berggren, W.T. and Coon, J.J. (2007) Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer, *Anal Chem*, **79**, 3525-3534.

- Meng, F., Cargile, B.J., Miller, L.M., Forbes, A.J., Johnson, J.R. and Kelleher, N.L. (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea, *Nat Biotechnol*, **19**, 952-957.
- Olsen, J.V., de Godoy, L.M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S. and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap, *Mol Cell Proteomics*, **4**, 2010-2021.
- Parks, B.A., Jiang, L., Thomas, P.M., Wenger, C.D., Roth, M.J., Boyne, M.T., Burke, P.V., Kwast, K.E. and Kelleher, N.L. (2007) Top-down proteomics on a chromatographic time scale using linear ion trap Fourier transform hybrid mass spectrometers, *Anal Chem*, **79**, 7984-7991.
- Patrie, S.M., Ferguson, J.T., Robinson, D.E., Whipple, D., Rother, M., Metcalf, W.W. and Kelleher, N.L. (2006) Top down mass spectrometry of < 60-kDa proteins from *Methanosarcina acetivorans* using quadrupole FTMS with automated octopole collisionally activated dissociation, *Molecular & Cellular Proteomics*, **5**, 14-25.
- Patrie, S.M., Robinson, D.E., Meng, F.Y., Du, Y. and Kelleher, N.L. (2004) Strategies for automating top-down protein analysis with Q-FTICR MS, *Int J Mass Spectrom*, **234**, 175-184.
- Pesavento, J.J., Kim, Y.B., Taylor, G.K. and Kelleher, N.L. (2004) Shotgun annotation of histone modifications: a new approach for streamlined characterization of proteins by top down mass spectrometry, *J Am Chem Soc*, **126**, 3386-3387.
- Purvine, S., Eppel, J.T., Yi, E.C. and Goodlett, D.R. (2003) Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer, *Proteomics*, **3**, 847-850.
- Quadroni, M. and James, P. (1999) Proteomics and automation, *Electrophoresis*, **20**, 664-677.
- Roth, M.J., Forbes, A.J., Boyne, M.T., Kim, Y.B., Robinson, D.E. and Kelleher, N.L. (2005) Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry, *Molecular & Cellular Proteomics*, **4**, 1002-1008.
- Roth, M.J., Parks, B.A., Ferguson, J.T., Boyne, M.T., 2nd and Kelleher, N.L. (2008) "Proteotyping": population proteomics of human leukocytes using top down mass spectrometry, *Anal Chem*, **80**, 2857-2866.
- Solouki, T., Marto, J.A., White, F.M., Guan, S.H. and Marshall, A.G. (1995) Attomole Biomolecule Mass Analysis by Matrix-Assisted Laser-Desorption Ionization Fourier-Transform Ion-Cyclotron Resonance, *Anal Chem*, **67**, 4139-4144.
- Solouki, T., PasaTolic, L., Jackson, G.S., Guan, S.G. and Marshall, A.G. (1996) High-resolution multistage MS, MS(2), and MS(3) matrix-assisted laser desorption/ionization FT-ICR mass spectra of peptides from a single laser shot, *Anal Chem*, **68**, 3718-3725.

Stahl, D.C., Swiderek, K.M., Davis, M.T. and Lee, T.D. (1996) Data-controlled automation of liquid chromatography tandem mass spectrometry analysis of peptide mixtures, *J Am Soc Mass Spectr*, **7**, 532-540.

Swaney, D.L., McAlister, G.C. and Coon, J.J. (2008) Decision tree-driven tandem mass spectrometry for shotgun proteomics, *Nature Methods*, **5**, 959-964.

Syka, J.E., Marto, J.A., Bai, D.L., Horning, S., Senko, M.W., Schwartz, J.C., Ueberheide, B., Garcia, B., Busby, S., Muratore, T., Shabanowitz, J. and Hunt, D.F. (2004) Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications, *J Proteome Res*, **3**, 621-626.

Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J. and Hunt, D.F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry, *P Natl Acad Sci USA*, **101**, 9528-9533.

Taylor, G.K., Kim, Y.B., Forbes, A.J., Meng, F.Y., McCarthy, R. and Kelleher, N.L. (2003) Web and database software for identification of intact proteins using "top down" mass spectrometry, *Anal Chem*, **75**, 4081-4086.

Valaskovic, G.A., Kelleher, N.L. and McLafferty, F.W. (1996) Attomole protein characterization by capillary electrophoresis mass spectrometry, *Science*, **273**, 1199-1202.

Venable, J.D., Dong, M.Q., Wohlschlegel, J., Dillin, A. and Yates, J.R. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra, *Nature Methods*, **1**, 39-45.

Waanders, L.F., Almeida, R., Prosser, S., Cox, J., Eikel, D., Allen, M.H., Schultz, G.A. and Mann, M. (2008) A novel chromatographic method allows on-line reanalysis of the proteome, *Molecular & Cellular Proteomics*, **7**, 1452-1459.

Yates, J.R., Cociorva, D., Liao, L. and Zabrouskov, V. (2006) Performance of a linear ion trap-Orbitrap hybrid for peptide analysis, *Anal Chem*, **78**, 493-500.

Zamdborg, L., LeDuc, R.D., Glowacz, K.J., Kim, Y.B., Viswanathan, V., Spaulding, I.T., Early, B.P., Bluhm, E.J., Babai, S. and Kelleher, N.L. (2007) ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry, *Nucleic Acids Res*, **35**, W701-W706.

Zappacosta, F., Huddleston, M.J., Karcher, R.L., Gelfand, V.I., Carr, S.A. and Annan, R.S. (2002) Improved sensitivity for phosphopeptide mapping using capillary column HPLC and microionspray mass spectrometry: Comparative phosphorylation site mapping from gel-derived proteins, *Anal Chem*, **74**, 3221-3231.

Zhang, Z.Q. and Marshall, A.G. (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra, *J Am Soc Mass Spectr*, **9**, 225-233.

Zubarev, R.A., Kelleher, N.L. and McLafferty, F.W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process, *Journal of the American Chemical Society*, **120**, 3265-3266.

CHAPTER 4: EVOLUTIONARY MIGRATION OF THE HISTIDINE-TYROSINE CROSS-LINK IN THE HEME-COPPER OXYGEN REDUCTASES

The contents of this chapter were adapted from the following article: James Hemp, Dana E. Robinson, Krithika B. Ganesan, Todd J. Martinez, Neil L. Kelleher, Robert B. Gennis (2006) “Evolutionary Migration of a Post-Translationally Modified Active Site Residue in the Proton-Pumping Heme-Copper Oxygen Reductases.” Biochemistry **45**(51): 15405-10. Cell culture, phylogenetic tree construction and computational studies were performed by James Hemp and Krithika Ganesan. This work was supported by NIH grants GM 067193-04 (NLK) and HL 16101 (RBG) and NSF grant NSF-BES-04-03846 (TJM).

4.1 Introduction

Aerobic respiration plays a fundamental role in Earth’s biogeochemical oxygen cycle. It has been estimated that ~75% of the O₂ produced by oxygenic photosynthesis is reduced to water via this enzymatically catalyzed process, tightly coupling two of the most widespread metabolisms on earth. Aerobic respiration is also the most exergonic metabolism known and appears to be a requirement for multicellular life. Respiration is performed by a series of integral membrane protein complexes that form electron transfer chains, found within the inner mitochondrial membrane of aerobic eukaryotes and the cytoplasmic membrane of many prokaryotic organisms (Garcia-Horsman, et al., 1994; Pereira, et al., 2001). Mitochondria have a linear electron transfer chain terminating with cytochrome *c* oxidase, a proton-pumping oxygen reductase which reduces O₂ to water. Prokaryotes have more complicated electron transfer

chains with branches leading to different terminal electron acceptors (*e.g.*, fumarate, nitrate, Fe^{3+} , O_2), allowing for metabolic flexibility when different environments are encountered.

Most aerobic prokaryotes utilize respiratory oxidases (*i.e.*, oxygen reductases) that are members of the heme-copper superfamily, which is structurally and catalytically diverse, containing both oxygen reductases and nitric oxide reductases. The mitochondrial cytochrome *c* oxidase is also a member of the heme-copper superfamily. Heme-copper oxygen reductases catalyze the reduction of O_2 to water with the concomitant electrogenic translocation of protons across the membrane, contributing to the generation of a proton electrochemical gradient that can be coupled to energy-requiring cellular processes (Garcia-Horsman, et al., 1994; Pereira, et al., 2001). The oxygen reductases are all multisubunit protein complexes that span the membrane bilayer. They are classified by type on the basis of genomic, phylogenetic, and structural analyses with the A-, B-, and C-types being the most well known (Pereira, et al., 2001). These A-, B-, and C-type oxygen reductase families have been shown to pump protons coupled to the reduction of oxygen; however, they differ in biochemical properties such as reaction rate and oxygen affinity. Members of the other families are assumed to also pump protons due to the presence of conserved features such as the D- and K-proton transfer channels. Many prokaryotic genomes encode several heme-copper oxygen reductases which are differentially expressed depending on the environmental conditions.

Subunit I is the core protein in the enzyme complex and is the only subunit shared by all three families of the oxygen reductases. All of the amino acid residues and cofactors necessary for catalysis and proton pumping are within subunit I. The active site of the enzyme is a bimetallic center composed of a copper ion (Cu_B) and a high-spin heme, together ligated by four conserved histidines (three to Cu_B and one to the heme Fe). X-ray structures of members of the

A- and B-type heme-copper oxygen reductases reveal a unique cross-linked histidine-tyrosine cofactor in the active site between one of the Cu_B ligands and a tyrosine that is essential for enzyme function (Ostermeier, et al., 1997; Soulimane, et al., 2000; Tsukihara, et al., 1996). This tyrosine is postulated to be oxidized to a tyrosyl radical during turnover and to donate a hydrogen atom to facilitate breaking of the O-O bond during catalysis (Babcock, 1999; Gennis, 1998; Proshlyakov, et al., 2000). The cross-link has been verified by mass spectrometry in the B-type oxygen reductase from *Thermus thermophilus* (Buse, et al., 1999). Figure 4.6 shows the structure of the cross-linked residues at the active site of the A-type oxygen reductases.

Sequence alignments have shown that the active-site tyrosine present in all of the A- and B-type oxygen reductases is absent in the C-type oxygen reductases. A crystal structure for a C-type oxygen reductase is not available, though structural models of subunit I for the C-type oxygen reductases from *Vibrio cholerae* (Hemp, et al., 2005) and *Rhodobacter sphaeroides* (Sharma, et al., 2006) have been built utilizing the X-ray structures of the A- and B-type oxygen reductases as templates. A surprising result was the prediction that a completely conserved tyrosine (Y255 in *V. cholerae*) from transmembrane helix VII in the C-type oxygen reductases occupies the same physical position in the active site as the tyrosine located in transmembrane helix VI of the A- and B-type oxygen reductases (Hemp, et al., 2005; Sharma, et al., 2006). It was also shown by modeling that it is geometrically feasible for a cross-link to be formed with the equivalent histidine ligand to Cu_B (H211 in *V. cholerae*). In this work, mass spectrometry was used to show that the predicted cross-link is indeed present in subunit I of the *V. cholerae* C-type oxygen reductase.

4.2 Preparation and Analysis of HCOR

Overexpression of C-Type Oxygen Reductase from V. cholerae. Protein was overexpressed and collected as previously reported (Hemp, et al., 2005). Briefly, *V. cholerae* cells were grown in LB medium (USB Corp.) with 100 mg/L ampicillin (Fisher Biotech) and 100 mg/L streptomycin at 37 °C. Gene expression was induced with 0.2% L-(+)-arabinose. The cells were lysed and centrifuged at 40 000 rpm to collect the membranes. Membrane proteins were solubilized by adding 0.5% dodecyl-D-maltoside (DDM) (Anatrace). Nonsolubilized membranes were removed by centrifugation at 40 000 rpm for 30 min.

Purification of Oxygen Reductase. To obtain a preparation sufficiently pure for mass spectrometry, the enzyme was first purified using immobilized metal affinity chromatography (IMAC) followed by weak anion exchange (WAX) on DEAE-Sepharose. IMAC was performed as previously reported (Hemp, et al., 2005), using a nickel affinity column (Qiagen, Valencia, CA) in a cold room (4 °C) at low pressure in 0.05% DDM and eluting the His-tagged protein using a stepped gradient of imidazole. WAX was performed using fast protein liquid chromatography (FPLC) (GE Healthcare, Piscataway, NJ) in a cold room using 10 mM ammonium bicarbonate (pH 8.0) and 0.05% DDM as solvent A and 1 M ammonium bicarbonate and 0.05% DDM as solvent B. Samples were loaded at a percentage of solvent B that was approximately 10% below the expected elution concentration of solvent B for the protein complex as determined by test gradients. A 1 h gradient to 100% solvent B was then utilized, and fractions containing the purified protein were combined. After each chromatography step, the sample was concentrated using a centrifugal filter with a mass cutoff of 50 kDa (Millipore, Billerica, MA).

Trypsin Digestion of the Oxygen Reductase and Sample Preparation. Ten microliters of purified enzyme (approximately 25 mg/mL) was digested overnight with 20 µg of sequencing-

grade trypsin (Bio-Rad, Hercules, CA) in a 90/10 100 mM ammonium bicarbonate (pH 8.0)/acetonitrile mixture at 37 °C. Immediately following trypsin digestion, 50 µL of the sample was applied to a gel filtration spin column with a 6 kDa mass cutoff (Micro Bio-Spin P6, Bio-Rad) to remove low-mass peptides. The spin column was equilibrated four times in 0.05% DDM prior to use. A methanol/chloroform precipitation (Wessel and Flugge, 1984; Whitelegge, et al., 1999) was then used to separate the remaining peptides from the detergent and soluble peptides. The resulting pellet was resuspended in 500 µL of 75% acetic acid and immediately subjected to analysis via mass spectrometry.

Mass Spectrometry. Samples were analyzed on a custom-built 8.5 T quadrupole Fourier-transform ion cyclotron resonance mass spectrometer (Q-FTICR MS) (Patrie, et al., 2004) using the MIDAS data station for data acquisition (Senko, et al., 1996). Introduction of the samples was performed using electrospray ionization (ESI) from a nanospray robot (Advion BioSciences, Ithaca, NY) at 1.2 kV with a backing gas pressure of 0.5 psi. Broadband scans were obtained to identify species of interest for fragmentation followed by quadrupole isolation (2 m/z window) and MS/MS using collisionally activated dissociation (CAD) fragmentation in the external accumulation octopole (Patrie, et al., 2006; Senko, et al., 1997). In these MS/MS experiments, several CAD acceleration voltages were used to generate a wider variety of fragment ions and these fragment lists were combined to create a final master fragment list. The time for transfer into the ICR cell was also varied to compensate for time-of-flight effects.

Data Analysis. Data from the broadband and MS/MS experiments were processed using an in-house developed version of the THRASH algorithm (Horn, et al., 2000) to detect isotopic distributions and the resulting mass lists were analyzed with ProSightPTM (LeDuc, et al., 2004; Taylor, et al., 2003). The presence of the cross-link required a separate fragmentation analysis of

each of the two peptides, with the opposite peptide modeled as a single large posttranslational modification. C-Terminal *b*- and N-terminal *y*-type fragment ions (Roepstorff and Fohlman, 1984) were matched at 20 ppm for the cross-linked peptide from the *V. cholerae* C-type oxygen reductase and 10 ppm for all other peptides.

4.3 MS Analysis of the C-Type HCOR From *V. cholerae*

Mass Spectrometry of a Tryptic Digest of a C-Type Oxygen Reductase. Mass spectrometry (MS) of a tryptic digest of subunit I from the *V. cholerae* C-type oxygen reductase was performed to discern the presence of the predicted cross-link. Analysis of the protein sequence predicted that if a cross-link had formed then complete trypsin digestion would result in a peptide containing residues S193-K232 cross-linked to residues L243-K304, but missing the region from residue Q233 to R242. This “H-shaped” tryptic peptide would not be present if the cross-link did not exist and was predicted to have a mass equal to that of the two individual peptides and subtracting 2 Da for the two protons lost during the formation of the cross-link. A peptide of this expected molecular mass (monoisotopic mass of 11 478.7 Da) was present in the mass spectrum of the trypsin digest. This putatively cross-linked tryptic fragment was isolated and analyzed by tandem MS (MS/MS) using collisionally activated dissociation (CAD) fragmentation. Figure 4.1 shows the MS/MS fragment map of the cross-linked peptide. Multiple N- and C-terminal fragment ions were detected from the peptides on either end of the cross-link, demonstrating the existence of a cross-link between the two. In addition to the MS/MS fragment ions containing the cross-linked peptide, several of the MS/MS fragments spanned cross-linked residue Y255 but did not include the crosslink (Figure 4.2).

Presence of the Non-Cross-Linked Protein in the Digest. The trypsin digest also contained non-cross-linked S193-K232 and L243-K304 peptides, and detailed MS/MS fragmentation confirmed their identity (Figures 4.3 and 4.4). It is unclear whether the His-Tyr cross-link is normally absent in a portion of the population of the protein or if the non-cross-linked species is an artifact of the recombinant protein expression or sample preparation. To address the lability of the cross-link, the same protocol was used to investigate the cross-link in the A-type oxygen reductase from *Rhodobacter sphaeroides*. In the A-type oxygen reductases, the cross-link is between residues that are only four amino acids apart on the same transmembrane helix (H284-Y288 in the *R. sphaeroides* A-type oxygen reductase), whereas there are 44 amino acids between the cross-linked residues in the C-type oxygen reductase from *V. cholerae*, which span two helices (H211-Y255). The data show definitively that the cross-link is present between His284 and Tyr288 in subunit I of the *R. sphaeroides* A-type oxygen reductase, and no fragments with the molecular mass expected for the non-cross-linked peptide were detected (Figure 4.5). In agreement with the latest X-ray structure of the *R. sphaeroides* A-type oxygen reductase (S. Ferguson-Miller, personal communication), it is concluded that the His-Tyr crosslink is present in the entire population of this A-type oxygen reductase. At this time, it is unclear whether the occupancy of the crosslink in the C-type reductase is less than 100% *in vivo* or if the non-cross-linked peptides are an artifact of the sample preparation or mass spectrometry. Although the A-type reductase appears to be entirely cross-linked, the His-Tyr bond lability may be higher in the C-type under the analysis conditions.

Presence of an Active-Site Cross-Linked Cofactor in C-Type Heme-Copper Oxygen Reductases. This work demonstrates that a novel cross-linked cofactor is present in all three families of the heme-copper oxygen reductases (Figure 4.6). This verifies the prediction by

molecular modeling (Hemp, et al., 2005; Sharma, et al., 2006) of the presence of an active-site tyrosine in the C-type oxygen reductases that is structurally and functionally equivalent to the active-site tyrosine in the A- and B-type oxygen reductases. It is also a unique structural feature which separates the oxygen reductases from other members of the heme-copper superfamily, notably NO reductases. The analysis of the C-type oxidase also revealed the presence of a population of the enzyme without the His-Tyr cross-link, but this could be an artifact either of the conditions of the expression of the enzyme (*e.g.*, incomplete incorporation of copper) or of the sample preparation. It is clear that collisionally activated fragmentation of the isolated crosslinked peptide during MS/MS analysis can result in scission of the cross-linking bond. However, the non-cross-linked peptide is also apparent in the absence of collisionally activated fragmentation in the mass spectrometer. This suggests that either the non-cross-linked enzyme is present within the trypsin digest or the cross-link in the C-type enzyme is more labile than that of the A-type enzyme either during the electrospray process or in the trapping and cooling of the ions in the mass spectrometer. Rauhamaki *et al.* have published a paper which demonstrates by MALDI mass spectrometry the presence of the His-Tyr cross-link in the C-type oxygen reductase from *R. sphaeroides* (Rauhamaki, et al., 2006). This report does not indicate any non-cross-linked protein, so it is very likely that the cross-link is present in all properly assembled enzymes. It can be concluded from our work and that of Rauhamaki *et al.* that the presence of the active-site His-Tyr cross-link is a universal feature of the C-type oxygen reductases and, by extension, all heme-copper oxygen reductases.

4.4 Conclusions

A Unified Catalytic Mechanism for All Oxygen Reductase Families. The novel cross-linked cofactor is thought to form as a result of the generation of a tyrosine radical in the active site, presumably upon the initial turnover of the reduced enzyme with O₂. Conceivably, this could be a side reaction and not essential for enzyme function. However, replacement of the active-site tyrosine with a phenylalanine in the *R. sphaeroides* A-type oxygen reductase not only resulted in an inactive enzyme but also altered the metal ligation in the active site (Das, et al., 1998). This suggests that the cross-link is needed to maintain the structure of the active site. Furthermore, work by Uchida *et al.* (Uchida, et al., 2004) demonstrated that substituting d⁴-Tyr for tyrosine resulted in a large decrease in the enzymatic activity of an *Escherichia coli* A-type oxygen reductase, and the spectroscopic properties suggested that the His-Tyr cross-link was not formed. These observations suggest that the cross-link is not simply an irrelevant side product of the chemistry at the active site but is essential for the function of the heme-copper oxygen reductases. The active-site tyrosine is proposed to donate both a proton and an electron to facilitate cleavage of the O-O bond (Babcock, 1999; Blomberg, et al., 2000; Blomberg, et al., 2003; Bu and Cukier, 2005; Gennis, 1998; Proshlyakov, et al., 2000). It is clear that an amino acid radical does form during the catalytic cycle of the oxygen reductase (Budiman, et al., 2004; MacMillan, et al., 2006; MacMillan, et al., 1999; Rich, et al., 2002; Wiertz, et al., 2004), and the active-site tyrosine is a logical primary electron donor for the chemistry. There is no evidence from rapid-quench electron paramagnetic resonance (EPR) spectroscopy for rapid formation of a tyrosine radical (Wiertz and de Vries, 2006; Wiertz, et al., 2004), but it would likely be EPR-silent due to the proximity of the metals at the active site. Attempts to demonstrate the presence of a radical by Fourier-transform infrared (FTIR) spectroscopy (Iwaki, et al., 2004; Nyquist, et al., 2003) and by iodination of the amino acid radical (Proshlyakov, et al., 2000) have provided

data consistent with the formation of neutral tyrosyl radical, but these data are acquired over a longer time period allowing for radical migration. Indeed, the strongest argument for the formation of a radical at the active-site tyrosine may be the fact that the His-Tyr cross-link is present. Presumably, the cross-link is a consequence of radical-based chemistry that occurs during the initial turnovers of the enzyme. The data strongly suggest that all heme-copper oxygen reductases utilize the same catalytic mechanism of hydrogen atom donation for oxygen bond scission.

Functional Role of the His-Tyr Cross-Link. The function of the cross-link has been the subject of considerable speculation as well as investigation. Recent studies with model compounds (Cappuccio, et al., 2002; Kim, et al., 2005; Pesavento, et al., 2006; Pratt, et al., 2005; Tomson, et al., 2002) as well as computational studies (Bu and Cukier, 2005; Colbran and Paddon-Row, 2003) have suggested a possible functional significance for the cross-link. The cross-linked histidine withdraws electrons from the tyrosine, resulting in a lower pK_a and a higher midpoint potential of the tyrosine (McCauley, et al., 2000). Conversely, the redox state and protonation state of the tyrosine influence the electron donating capacity of the imidazole as a metal ligand, thus controlling the preferred ligand geometry about Cu_B (Pesavento, et al., 2006). It has also been suggested that, due to the presence of the cross-linked tyrosine, the histidine ligand to Cu_B might be labile and move away from the metal during turnover, playing a key role in the proton pump mechanism (Colbran and Paddon-Row, 2003). These studies, in conjunction with the presence of the cross-linked cofactor in all oxygen reductase families, suggest that the cofactor may be a required component for proton pumping. Further work is necessary to elucidate its role.

Evolutionary Migration of the Post-Translationally Modified Tyrosine. The post-translational modification of active site amino acid residues to form novel cofactors in situ has been observed in a number of redox active enzymes (Okeley and van der Donk, 2000). Some cofactors are produced via chemical modification of amino acid side chains (*e.g.*, oxidation, methylation, and hydroxylation), whereas other cofactors are formed by cross-linking two or more amino acids together. These post-translationally cross-linked active-site amino acids can be found in tyrosinase, hemocyanin, and catechol oxidase (Cys-His), catalase-peroxidase (Met-Tyr-Trp), galactose oxidase (Tyr-Cys), catalase (His-Tyr-RC), and the A- and B-type heme-copper oxygen reductases (His-Tyr) (Okeley and van der Donk, 2000). The evolutionary migration of amino acids within a protein family can be defined as the situation in which residues that have the same structural or functional role and which share the same spatial location derive from different positions within their respective protein sequences. Evolutionary migration has been reported for active-site residues (Hasson, et al., 1998; Todd, et al., 2002), but this is the first report of a post-translationally modified activesite residue. The active-site tyrosine forming the cross-linked cofactor is located within a different transmembrane helix in the C-type oxygen reductases (helix VII) in comparison to that of the A-type and B-type oxygen reductases (helix VI). It is currently unknown which state (the tyrosine being located in helix VI or helix VII) is ancestral.

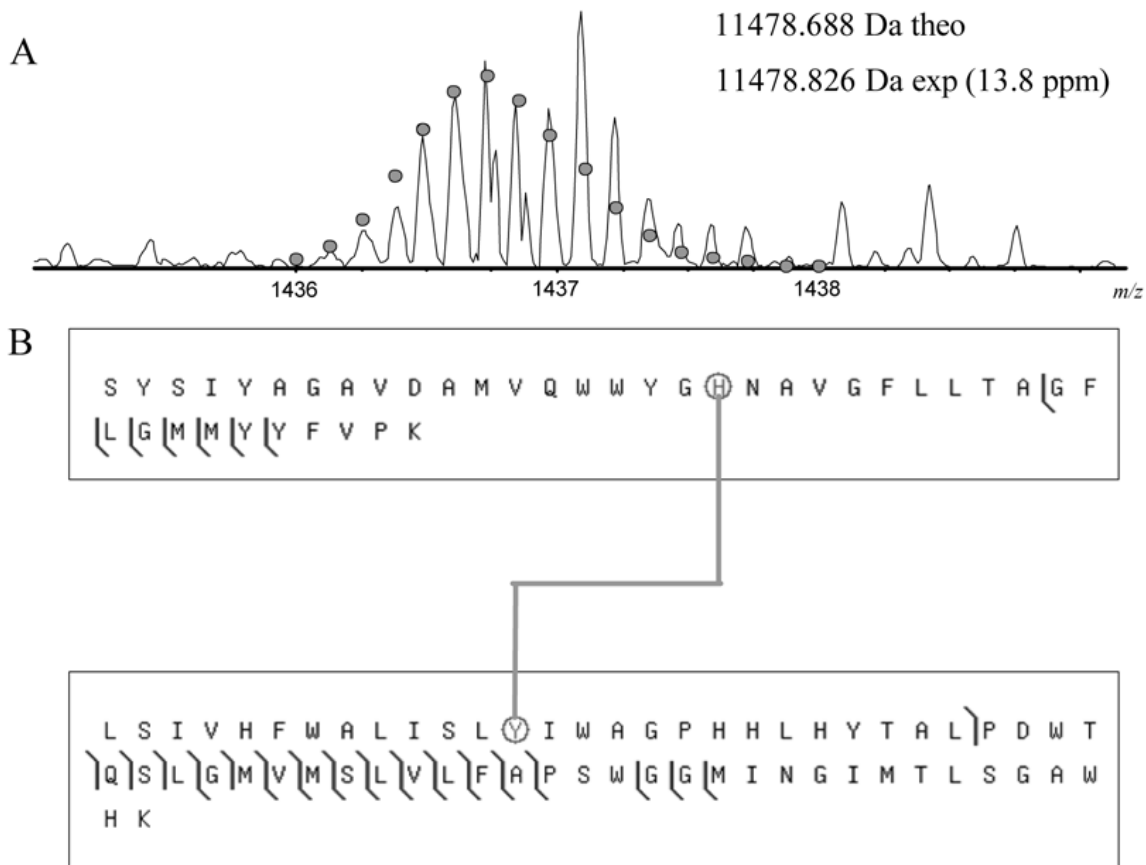


Figure 4.1. CAD fragmentation map for the cross-linked peptides S193-K232 and L243-K304. (A) The mass spectrum of the cross-linked tryptic peptide. The circles show the theoretical isotope heights for a peptide of the given mass. (B) CAD MS/MS fragmentation results. The cross-linked histidine and tyrosine are circled. Spectra were processed using a modified version of the THRASH algorithm (Horn, et al., 2000) and fragments were matched using ProSight PTM (LeDuc, et al., 2004; Taylor, et al., 2003) with a match tolerance of 20 ppm.

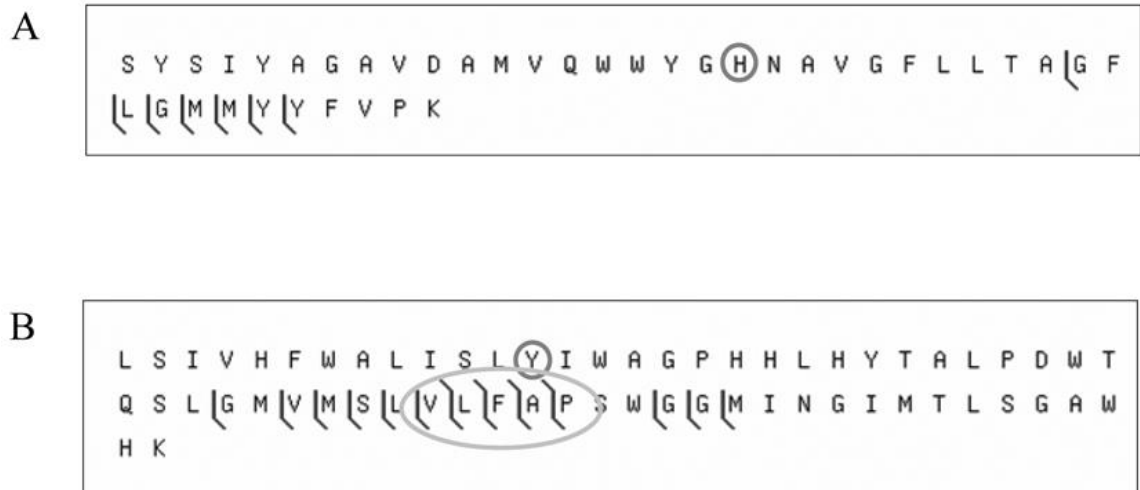


Figure 4.2. CAD fragmentation of the crosslinked tryptic peptide from the C-type oxygen reductase assuming the absence of the crosslink. The CAD fragmentation dataset used to generate figure 4.1 was also analyzed as if the cross-link were not present. (A) shows the matches for the tryptic peptide S193-K232 and (B) for L243-K304. The histidine and tyrosine normally involved in the cross-link are circled. The oval highlights several fragments that were found to span the cross-linked residue but do not include the mass of the opposite cross-linked peptide.

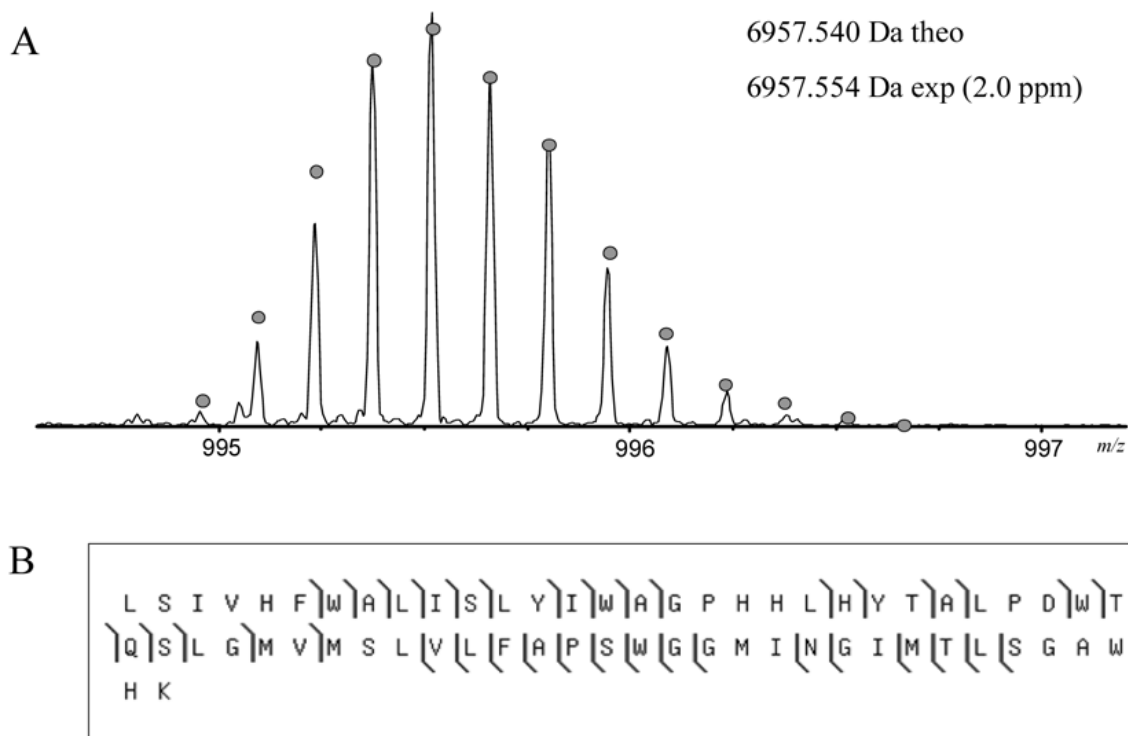


Figure 4.3. Mass spectrum and fragmentation of the tryptic peptide L243-K304 from the C-type oxygen reductase from *V. cholerae*. (A) The mass spectrum of the peptide. The circles show the theoretical isotope heights for a peptide of the given mass. (B) CAD MS/MS fragmentation results.

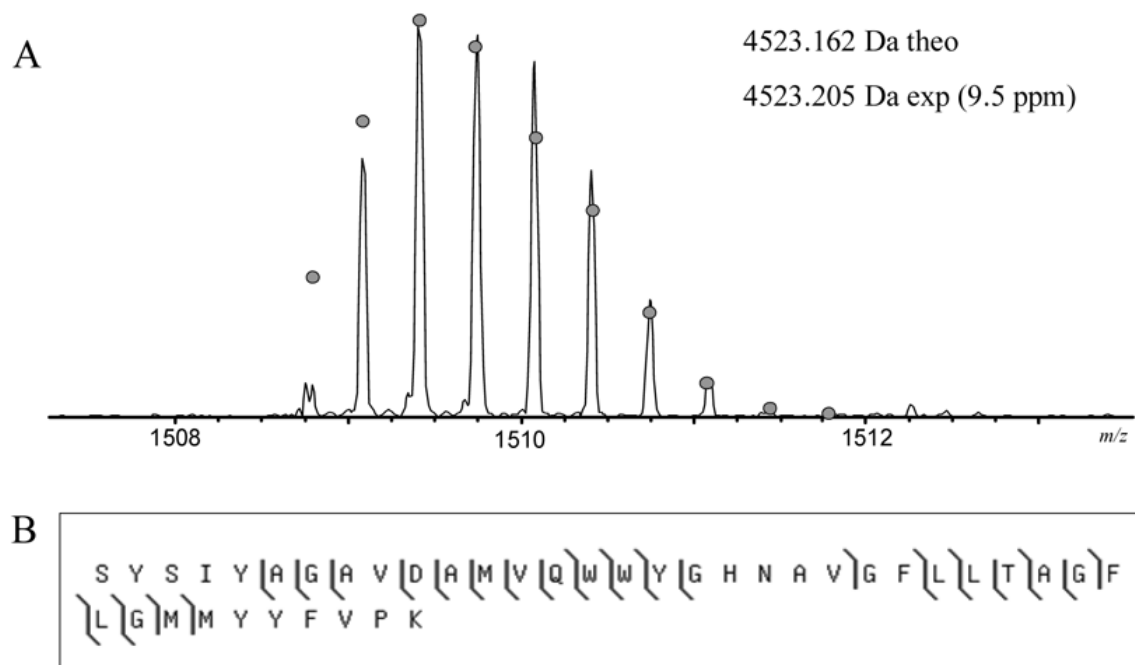


Figure 4.4. Mass spectrum and fragmentation of the tryptic peptide S193-K232 from the C-type oxygen reductase from *V. cholerae*. (A) The mass spectrum of the peptide. The circles show the theoretical isotope heights for a peptide of the given mass. (B) CAD MS/MS fragmentation results

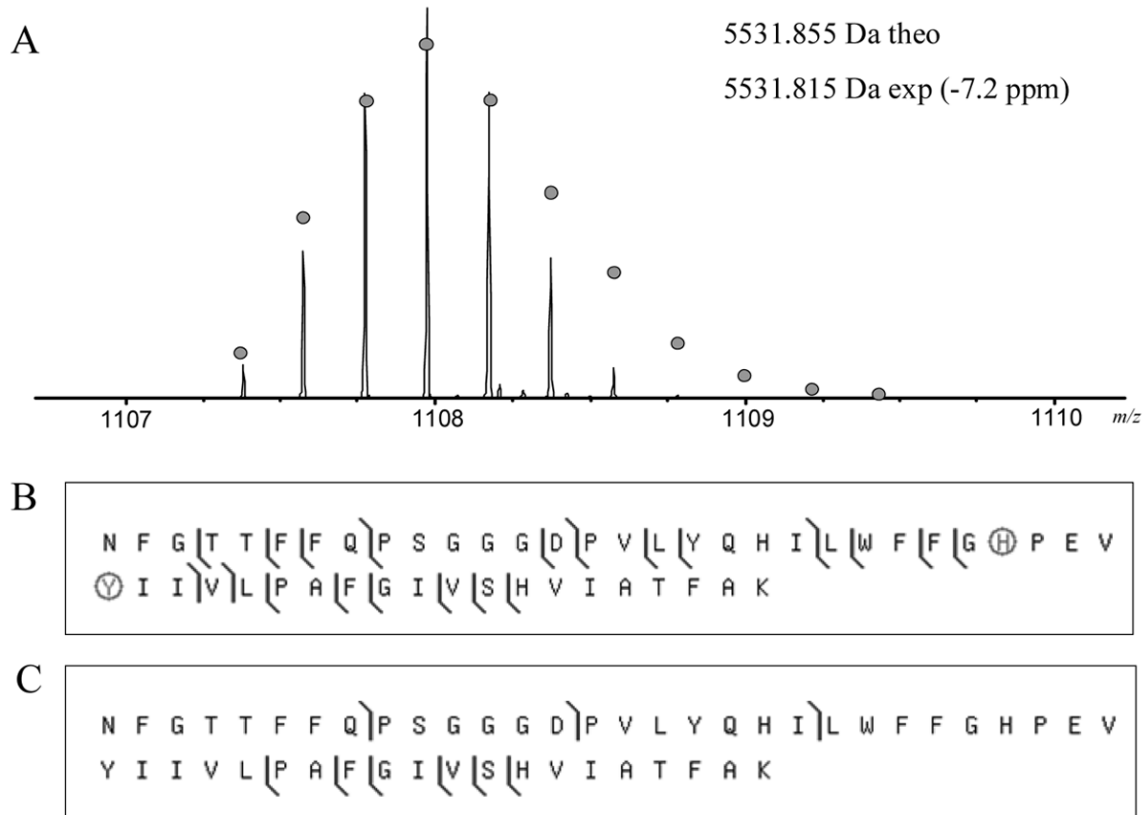


Figure 4.5. Mass spectrum and fragmentation of the crosslinked peptide from the A-type oxygen reductase from *R. sphaeroides*. (A) The mass spectrum of the cross-linked tryptic peptide N258-K307. The circles show the theoretical isotope heights for a peptide of the given mass. The mass spectrum shows that only the cross-linked species is detected in the tryptic digest. (B) CAD MS/MS fragmentation results when analyzed with the cross-link present (the histidine and tyrosine are circled) and (C) without the cross-link. Multiple fragments containing the cross-link are detected whereas no fragments are detected which do not contain it.

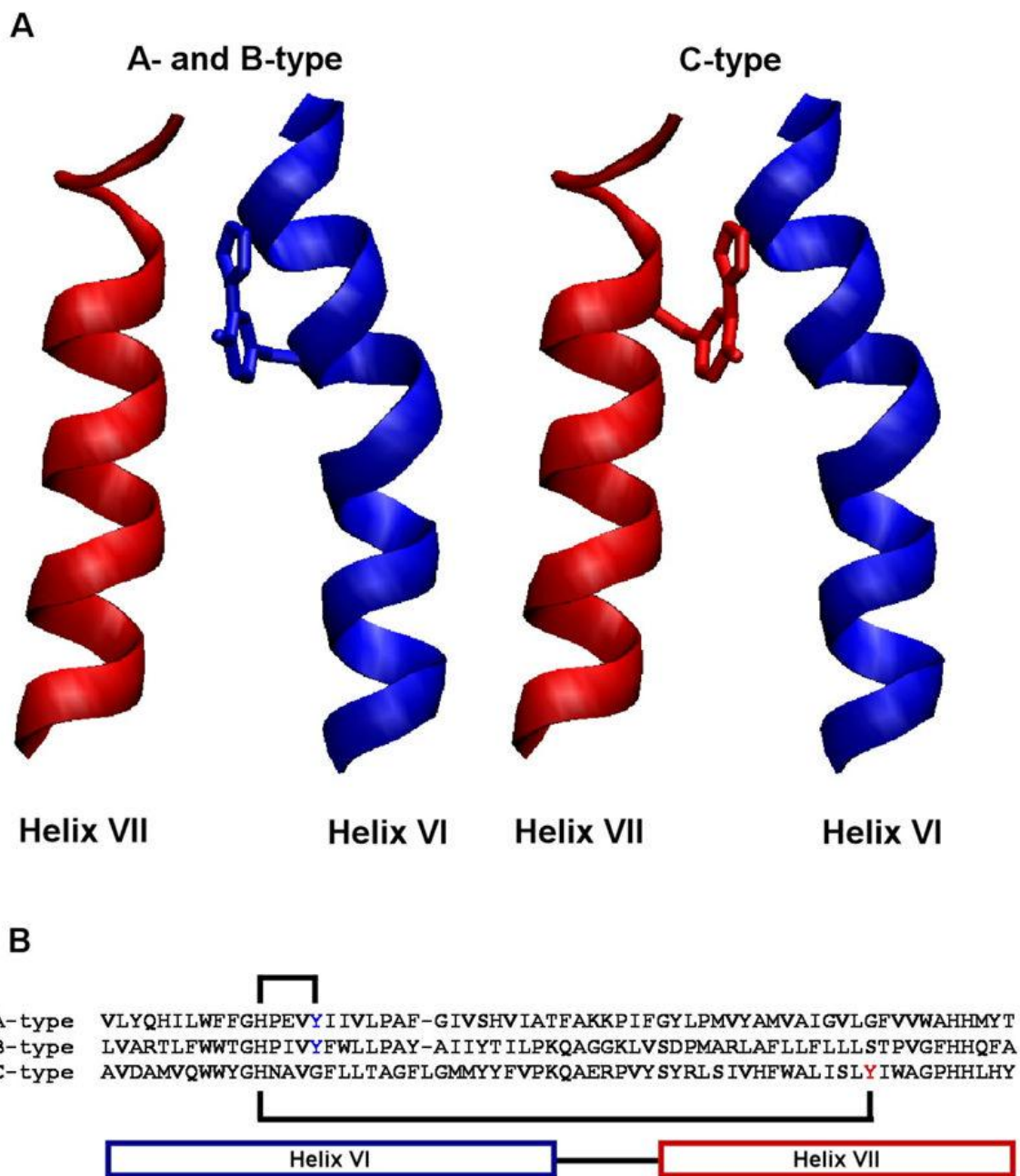


Figure 4.6. A novel cross-linked cofactor is present in all three heme-copper oxygen reductase families. (A) The active-site tyrosine forming the cofactor originates from helix VI in the A- and B-type oxygen reductases, while in the C-type oxygen reductases it originates from helix VII. This is the first example of the evolutionary migration of a post-translationally modified active-site residue. (B) The cross-link is formed between a histidine and tyrosine within helix VI in the A- and B-type oxygen reductases. In the C-type oxygen reductases the crosslink is formed between a histidine in helix VI and a tyrosine in helix VII, covalently coupling the helices together. This figure was generated using VMD software (Humphrey, et al., 1996).

4.5 Literature Cited

Babcock, G.T. (1999) How oxygen is activated and reduced in respiration, *P Natl Acad Sci USA*, **96**, 12971-12973.

Blomberg, M.R.A., Siegbahn, P.E.M., Babcock, G.T. and Wikstrom, M. (2000) O-O bond splitting mechanism in cytochrome oxidase, *J Inorg Biochem*, **80**, 261-269.

Blomberg, M.R.A., Siegbahn, P.E.M. and Wikstrom, M. (2003) Metal-bridging mechanism for O-O bond cleavage in cytochrome c oxidase, *Inorg Chem*, **42**, 5231-5243.

Bu, Y.X. and Cukier, R.I. (2005) Structural character and energetics of tyrosyl radical formation by electron/proton transfers of a covalently linked histidine-tyrosine: A model for cytochrome c oxidase, *J Phys Chem B*, **109**, 22013-22026.

Budiman, K., Kannt, A., Lyubenova, S., Richter, O.M.H., Ludwig, B., Michel, H. and MacMillan, F. (2004) Tyrosine 167: The origin of the radical species observed in the reaction of cytochrome c oxidase with hydrogen peroxide in *Paracoccus denitrificans*, *Biochemistry*, **43**, 11709-11716.

Buse, G., Soulimane, T., Dewor, M., Meyer, H.E. and Bluggel, M. (1999) Evidence for a copper-coordinated histidine-tyrosine cross-link in the active site of cytochrome oxidase, *Protein Sci*, **8**, 985-990.

Cappuccio, J.A., Ayala, I., Elliott, G.I., Szundi, I., Lewis, J., Konopelski, J.P., Barry, B.A. and Einarsdottir, O. (2002) Modeling the active site of cytochrome oxidase: Synthesis and characterization of a cross-linked histidine-phenol, *Journal of the American Chemical Society*, **124**, 1750-1760.

Colbran, S.B. and Paddon-Row, M.N. (2003) Could the tyrosine-histidine ligand to Cu-B in cytochrome c oxidase be coordinatively labile? Implications from a quantum chemical model study of histidine substitutional lability and the effects of the covalent tyrosine-histidine cross-link, *J Biol Inorg Chem*, **8**, 855-865.

Das, T.K., Pecoraro, C., Tomson, F.L., Gennis, R.B. and Rousseau, D.L. (1998) The post-translational modification cytochrome c oxidase is required to establish a functional environment of the catalytic site, *Biochemistry*, **37**, 14471-14476.

Garcia-Horsman, J.A., Barquera, B., Rumbley, J., Ma, J. and Gennis, R.B. (1994) The superfamily of heme-copper respiratory oxidases, *J Bacteriol*, **176**, 5587-5600.

Gennis, R.B. (1998) Multiple proton-conducting pathways in cytochrome oxidase and a proposed role for the active-site tyrosine, *Bba-Bioenergetics*, **1365**, 241-248.

Hasson, M.S., Schlichting, I., Moulai, J., Taylor, K., Barrett, W., Kenyon, G.L., Babbitt, P.C., Gerlt, J.A., Petsko, G.A. and Ringe, D. (1998) Evolution of an enzyme active site: The structure

of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase, *P Natl Acad Sci USA*, **95**, 10396-10401.

Hemp, J., Christian, C., Barquera, B., Gennis, R.B. and Martinez, T.J. (2005) Helix switching of a key active-site residue in the cytochrome cbb3 oxidases, *Biochemistry*, **44**, 10766-10775.

Horn, D.M., Zubarev, R.A. and McLafferty, F.W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules, *J Am Soc Mass Spectr*, **11**, 320-332.

Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: Visual molecular dynamics, *J Mol Graphics*, **14**, 33-&.

Iwaki, M., Puustinen, A., Wikstrom, M. and Rich, P.R. (2004) ATR-FTIR spectroscopy and isotope labeling of the P-M intermediate of *Paracoccus denitrificans* cytochrome c oxidase, *Biochemistry*, **43**, 14370-14378.

Kim, E., Kamaraj, K., Galliker, B., Rubie, N.D., Moenne-Loccoz, P., Kaderli, S., Zuberbuhler, A.D. and Karlin, K.D. (2005) Dioxygen reactivity of copper and heme-copper complexes possessing an imidazole-phenol cross-link, *Inorg Chem*, **44**, 1238-1247.

LeDuc, R.D., Taylor, G.K., Kim, Y.B., Januszyk, T.E., Bynum, L.H., Sola, J.V., Garavelli, J.S. and Kelleher, N.L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry, *Nucleic Acids Res*, **32**, W340-W345.

MacMillan, F., Budiman, K., Angerer, H. and Michel, H. (2006) The role of tryptophan 272 in the *Paracoccus denitrificans* cytochrome c oxidase, *Febs Letters*, **580**, 1345-1349.

MacMillan, F., Kannt, A., Behr, J., Prisner, T. and Michel, H. (1999) Direct evidence for a tyrosine radical in the reaction of cytochrome c oxidase with hydrogen peroxide, *Biochemistry*, **38**, 9179-9184.

McCauley, K.M., Vrtis, J.M., Dupont, J. and van der Donk, W.A. (2000) Insights into the functional role of the tyrosine-histidine linkage in cytochrome c oxidase, *Journal of the American Chemical Society*, **122**, 2403-2404.

Nyquist, R.M., Heitbrink, D., Bolwien, C., Gennis, R.B. and Heberle, J. (2003) Direct observation of protonation reactions during the catalytic cycle of cytochrome c oxidase, *P Natl Acad Sci USA*, **100**, 8715-8720.

Okeley, N.M. and van der Donk, W.A. (2000) Novel cofactors via post-translational modifications of enzyme active sites, *Chem Biol*, **7**, R159-R171.

Ostermeier, C., Harrenga, A., Ermler, U. and Michel, H. (1997) Structure at 2.7 Å resolution of the *Paracoccus denitrificans* two-subunit cytochrome c oxidase complexed with an antibody FV fragment, *Proc Natl Acad Sci U S A*, **94**, 10547-10553.

Patrie, S.M., Charlebois, J.P., Whipple, D., Kelleher, N.L., Hendrickson, C.L., Quinn, J.P., Marshall, A.G. and Mukhopadhyay, B. (2004) Construction of a hybrid quadrupole/Fourier transform ion cyclotron resonance mass spectrometer for versatile MS/MS above 10 kDa, *J Am Soc Mass Spectrom*, **15**, 1099-1108.

Patrie, S.M., Ferguson, J.T., Robinson, D.E., Whipple, D., Rother, M., Metcalf, W.W. and Kelleher, N.L. (2006) Top down mass spectrometry of < 60-kDa proteins from *Methanosarcina acetivorans* using quadrupole FRMS with automated octopole collisionally activated dissociation, *Mol Cell Proteomics*, **5**, 14-25.

Pereira, M.M., Santana, M. and Teixeira, M. (2001) A novel scenario for the evolution of haem-copper oxygen reductases, *Biochim Biophys Acta*, **1505**, 185-208.

Pesavento, R.P., Pratt, D.A., Jeffers, J. and van der Donk, W.A. (2006) Model studies of the Cu-B site of cytochrome c oxidase utilizing a Zn(II) complex containing an imidazole-phenol cross-linked ligand, *Dalton T*, 3326-3337.

Pratt, D.A., Pesavento, R.P. and van der Donk, W.A. (2005) Model studies of the histidine-tyrosine cross-link in cytochrome c oxidase reveal the flexible substituent effect of the imidazole moiety, *Organic Letters*, **7**, 2735-2738.

Proshlyakov, D.A., Pressler, M.A., DeMaso, C., Leykam, J.F., DeWitt, D.L. and Babcock, G.T. (2000) Oxygen activation and reduction in respiration: involvement of redox-active tyrosine 244, *Science*, **290**, 1588-1591.

Rauhamaeki, V., Baumann, M., Soliymani, R., Puustinen, A. and Wikstrom, M. (2006) Identification of a histidine-tyrosine cross-link in the active site of the cbb(3)-type cytochrome c oxidase from *Rhodobacter sphaeroides*, *P Natl Acad Sci USA*, **103**, 16135-16140.

Rich, P.R., Rigby, S.E.J. and Heathcote, P. (2002) Radicals associated with the catalytic intermediates of bovine cytochrome c oxidase, *Bba-Bioenergetics*, **1554**, 137-146.

Roepstorff, P. and Fohlman, J. (1984) Proposal for a Common Nomenclature for Sequence Ions in Mass-Spectra of Peptides, *Biomed Mass Spectrom*, **11**, 601-601.

Senko, M.W., Canterbury, J.D., Guan, S. and Marshall, A.G. (1996) A high-performance modular data system for Fourier transform ion cyclotron resonance mass spectrometry, *Rapid Commun Mass Spectrom*, **10**, 1839-1844.

Senko, M.W., Hendrickson, C.L., Emmett, M.R., Shi, S.D.H. and Marshall, A.G. (1997) External accumulation of ions for enhanced electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, *J Am Soc Mass Spectr*, **8**, 970-976.

Sharma, V., Puustinen, A., Wikstrom, M. and Laakkonen, L. (2006) Sequence analysis of the cbb3 oxidases and an atomic model for the *Rhodobacter sphaeroides* enzyme, *Biochemistry*, **45**, 5754-5765.

Soulimane, T., Buse, G., Bourenkov, G.P., Bartunik, H.D., Huber, R. and Than, M.E. (2000) Structure and mechanism of the aberrant ba(3)-cytochrome c oxidase from thermophilus, *EMBO J*, **19**, 1766-1776.

Taylor, G.K., Kim, Y.B., Forbes, A.J., Meng, F.Y., McCarthy, R. and Kelleher, N.L. (2003) Web and database software for identification of intact proteins using "top down" mass spectrometry, *Anal Chem*, **75**, 4081-4086.

Todd, A.E., Orengo, C.A. and Thornton, J.M. (2002) Plasticity of enzyme active sites, *Trends Biochem Sci*, **27**, 419-426.

Tomson, F., Bailey, J.A., Gennis, R.B., Unkefer, C.J., Li, Z.H., Silks, L.A., Martinez, R.A., Donohoe, R.J., Dyer, R.B. and Woodruff, W.H. (2002) Direct infrared detection of the covalently ring linked His-Tyr structure in the active site of the heme-copper oxidases, *Biochemistry*, **41**, 14383-14390.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. and Yoshikawa, S. (1996) The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å, *Science*, **272**, 1136-1144.

Uchida, T., Mogi, T., Nakamura, H. and Kitagawa, T. (2004) Role of Tyr-288 at the dioxygen reduction site of cytochrome bo studied by stable isotope labeling and resonance Raman spectroscopy, *Journal of Biological Chemistry*, **279**, 53613-53620.

Wessel, D. and Flugge, U.I. (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids, *Anal Biochem*, **138**, 141-143.

Whitelegge, J.P., le Coutre, J., Lee, J.C., Engel, C.K., Prive, G.G., Faull, K.F. and Kaback, H.R. (1999) Toward the bilayer proteome, electrospray ionization-mass spectrometry of large, intact transmembrane proteins, *Proc Natl Acad Sci U S A*, **96**, 10695-10698.

Wiertz, F.G.M. and de Vries, S. (2006) Low-temperature kinetic measurements of microsecond freeze-hyperquench (MHQ) cytochrome oxidase monitored by UV-visible spectroscopy with a newly designed cuvette, *Biochem Soc T*, **34**, 136-138.

Wiertz, F.G.M., Richter, O.M.H., Cherepanov, A.V., MacMillan, F., Ludwig, B. and de Vries, S. (2004) An oxo-ferryl tryptophan radical catalytic intermediate in cytochrome c and quinol oxidases trapped by microsecond freeze-hyperquenching (MHQ), *Febs Letters*, **575**, 127-130.

**CHAPTER 5: PROTON CHANNEL AND ELECTRON DELIVERY MUTANTS
PROVIDE INSIGHT INTO THE FORMATION OF THE HIS-TYR COVALENT
CROSSLINK IN CYTOCHROME C OXIDASE**

This chapter is derived from an article written with Krithika Ganesan which is being prepared for publication. Krithika performed the cell culture, created the mutant strains and assisted in the analysis.

5.1 Introduction

Cytochrome *c* oxidase couples the one-electron oxidation of cytochrome *c* to the four-electron reduction of oxygen to water and conserves the free energy by translocating protons across the inner mitochondrial or the bacterial cytoplasmic membrane. The catalytic site where the reduction of oxygen is carried out is a binuclear center, which is composed of a high-spin heme, heme *a*₃, and a copper center, Cu_B, which has three histidine ligands (H284, H333 and H334, *R. sphaeroides* numbering, figure 5.1). The X-ray crystal structures of the A-type bovine and bacterial oxidases also revealed a post-translational modification in this active site that is unique to the heme-copper superfamily (Iwata, et al., 1995; Tsukihara, et al., 1996) and was later verified experimentally by mass-spectrometry (Buse, et al., 1999).

One of the Cu_B ligands, H284, was found to be covalently crosslinked to a nearby tyrosine residue, Y288, and this modification was subsequently observed in the crystal structure of the B-type oxidase from *Thermus thermophilus* (Soulimane, et al., 2000). Until recently, it was believed that the C-type oxidases function without the crosslink owing to the absence of the conserved motif Gly-His-Pro-X-Val-Tyr in the sequence alignments. Recently, however,

computational modeling studies (Hemp, et al., 2005) in conjunction with mass spectrometry (Hemp, et al., 2006; Rauhamaki, et al., 2006) have revealed that the conserved tyrosine has been moved to a different helix but is still in a position which allows the formation of the crosslink to the histidine residue (H284 equivalent in the C-type oxidases).

This covalent linkage between the C ϵ_2 of the Y288 side-chain and N ϵ_2 of the H284 side chain is seen in the crystal structures of oxidized, fully reduced and ligand-bound forms of the A-type enzyme (Iwata, et al., 1995; Ostermeier, et al., 1995; Qin, et al., 2009; Tsukihara, et al., 1996) and is therefore considered an essential part of the active site. Mutation of tyrosine to phenylalanine has been reported to abolish the formation of crosslink and result in the loss of Cu_B from the binuclear center (Das, et al., 1998; Pinakoulaki, et al., 2002). This covalently linked Y288, which is conserved across the proton-pumping oxidases, is also at the end of the proton-input pathway, the K-channel. Therefore, this residue has been proposed to play a structural as well as a functional role in the oxidases.

As soon as oxygen binds to the reduced binuclear center, the O=O bond is irreversibly broken without the release of reactive oxygen intermediates. Out of the four electrons supplied by the oxidase in one turnover, two are taken from heme *a*₃, one from Cu_B and one proton and an electron presumably from the crosslinked Y288, leaving behind a neutral tyrosine radical (Blomberg, et al., 2000; Proshlyakov, et al., 1998; Proshlyakov, et al., 2000). EPR signals from the radical seen during *P_m* formation or upon treatment of oxidase with H₂O₂ have generally been assigned to Y288 rather than residues like W280 and Y167 which have also been implicated to be source of the radical in the past (Budiman, et al., 2004; MacMillan, et al., 2006; MacMillan, et al., 1999). The proximity of Y288 to the catalytic site and its unique covalent modification with a Cu_B ligand make it the likeliest donor of the electron for O=O bond splitting (Hemp, et al., 2006;

Proshlyakov, et al., 1998; Proshlyakov, et al., 2000; Rauhamaki, et al., 2006). The His-Tyr covalent linkage has been studied *in vitro* using model organic compounds and the findings range from the covalent crosslink lowering the pK_a of the tyrosine residue, increasing its midpoint potential in the radical state and forming phenoxyl radicals during single turnover (Collman, et al., 2006; Collman, et al., 2007; Collman, et al., 2007; McCauley, et al., 2000). Although tyrosyl radicals have been detected in EPR during the catalysis, not much is known about the regeneration of tyrosine or tyrosinate. It is also not known if, during *P_m* formation, a hydrogen-atom transfer or an electron transfer followed by proton transfer occurs or vice-versa. The role of Y288 in proton delivery is undisputed and has been clearly shown by recent FTIR studies in conjunction with electrometry, where the deprotonation of this tyrosine residue during *P_m* formation in E286Q mutants was reported (Gorbikova, et al., 2008).

The properties of the crosslink that convert the Y288 residue into a facile proton and electron donor have rightly received a lot of attention but very little is known about the origin of the crosslink. It is generally believed to be a post-translational modification that is formed after the very first or initial few turnovers of the oxidase (Babcock, 1999; Buse, et al., 1999; Gennis, 1998; Rogers and Dooley, 2001). In other enzymatic systems with covalently linked amino acids, the metals in the active site in the presence of oxygen auto-catalytically induce covalent linkages among susceptible residues like Tyr, Cys, His, Trp and Met to form the functional enzyme (Davidson, 2007; Rogers and Dooley, 2003; Stubbe and van Der Donk, 1998). In enzymatic systems like catalase peroxidase and galactose oxidase, holoproteins lacking the covalent linkage were synthesized by depleting the growth medium of the active site metals, Fe or Cu, respectively. Once the metal centers were reconstituted *in-vitro*, cross-links formed

spontaneously in both systems, establishing that their post-translational event is a self-processing one (Ghiladi, et al., 2005; Rogers, et al., 2000).

In the heme-copper oxidases the cross-link formation is also perceived as a self-processing event. Unfortunately, there is no direct way to test this since in-vitro reconstitution of metals has been shown to be unsuccessful for oxidases (Hiser, et al., 2000; Zhen, et al., 2002). Instead, since a turnover of the enzyme requires a supply of both protons and electrons, this work investigates the crosslink processing reaction by manipulating the proton and electron transfer pathways in the enzyme, allowing us to understand the conditions required for the formation of the covalent linkage.

5.2 Preparation and Analysis of Cytochrome *c* Oxidase

All reagents are from Sigma (St. Louis, MO) unless otherwise noted.

Overexpression, Purification and Preparation of Oxidases. The pRK415-expression plasmids with the corresponding mutations for H260N and LpM in subunit II were transferred by conjugation to the *Rhodobacter* expression strain YZ200. The expression vectors bearing mutations of H260N and LpM were kindly sent to us by Dr. Shelagh Ferguson-Miller at the Michigan State University, East Lansing. The ΔCu_B mutant protein was gifted by Dr. Jon Hosler at the Mississippi Medical Center. Other mutants were created and protein was overexpressed, collected, purified, digested and prepared for mass spectrometry as previously reported (Ganesan and Gennis, 2010; Hemp, et al., 2006). Subunit I of the *aa*₃-type oxidase in *Rhodobacter sphaeroides* has 25 cleavage sites for trypsin and the tryptic peptide of interest bearing the post-translational modification is N258-K307 (5531.853 Da with crosslink, 5531.869 Da without).

Mass Spectrometry and Data Analysis. Mass spectra of this peptide were acquired on a commercial LTQ-FT Ultra 12 Tesla linear ion trap-Fourier transform ion cyclotron resonance hybrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA). The detergent-free trypsin digests were introduced into the spectrometer using direct-infusion electrospray ionization using an Advion Triversa NanoMate source (Advion BioSciences, Ithaca, NY). For each peptide, three mass spectra were obtained. First, a mass spectrum of the tryptic peptide mixture covering a 500-2000 m/z range was acquired (25 summed scans). The peptide of interest was then identified and the mass spectrum of the isolated intact peptide after isolation in the linear ion trap (10 scans), followed by a collisionally activated dissociation (CAD) fragmentation spectrum (100 scans).

The ΔCu_B mutant was processed differently as a highly abundant contaminating peptide (similar to that observed in figure 5.3) was present in the digest mixture. In this case, the tryptic digest was subjected to LCMS, producing a lower-quality intact mass spectrum (due to only one scan being collected with no ion trap isolation) and a CAD MS/MS spectrum. Approximately 200 nM digest was injected using a Gilson 235P autosampler () onto a 1 mm \times 100 mm polymeric reverse phase column (PLRP-S, Higgins-Analytical, Mountain View, CA) connected to an Agilent 1200 HPLC pump flowing at 100 $\mu\text{L}/\text{min}$ following a linear gradient (5% B for 10 min, increasing to 60% B in 90 min, 95% B in 100 min, Buffer A) 10% acetonitrile/water with 0.2% formic acid; Buffer B) 90% acetonitrile/isopropanol with 0.2% formic acid). Eluent flowed through the NanoMate with a 300:1 split resulting in ~ 300 nL/min flow rate at the nanospray source. Data-dependent LC-MS/MS data (3 most intense precursors, 3 m/z isolation window) utilizing dynamic exclusion (2 times, 120 s exclusion, 300 max exclusion list size) were

acquired on the 12 T LTQ-FT Ultra with both the precursor and fragmentation scans analyzed in the ICR cell at a resolution setting of 171 000 at m/z 400.

Isotopic distributions in all spectra were identified using the THRASH algorithm (Horn, et al., 2000) and analyzed using the single protein mode of ProSightPTM (LeDuc, et al., 2004), using a 10 ppm fragment match tolerance with the crosslink represented as a -2.01565 Da mass shift (- the mass of ^1H on both the His and Tyr residues). All masses in the figures and analysis are monoisotopic.

5.3 MS Analysis of Proton Channel and Electron Delivery Mutants

To determine the presence or absence of the crosslink, the experimental mass of the intact peptide was compared with the theoretical mass and the MS/MS data were mapped to theoretical fragment ion masses both containing and lacking the putative crosslink. In three of the four samples, the mass of the intact peptide supports a particular crosslink state. In the case of the ΔCu_B mutant, the intact mass spectrum was not of high enough signal-to-noise to determine the accurate mass, though it was of high enough signal-to-noise to trigger MS/MS data acquisition during the LCMS run. In the MS/MS fragmentation analysis, the masses of the experimental peptide fragments are mapped against two different sets of theoretical fragment masses; one set lacking the predicted crosslink mass shift and one set including it. In all four mutants, the MS/MS fragmentation map results overwhelmingly favor one state over the other. Although neither analysis rules out the possibility of the other state existing at a low, unobserved level, they do demonstrate that the enzyme predominantly exists in a particular state.

Using this technique, our group had shown earlier that the crosslink was present in the wild-type *R. sphaeroides aa₃* enzyme population and that no non-crosslinked peptides were

detected (Hemp, et al., 2006). In this work, the same technique was used to determine the presence of crosslink in the tryptic digests of the subunit I of several mutant *aa*₃-type oxidases.

H260N The residue H260 in subunit II serves as a ligand to the dinuclear Cu_A-site. Electron transfer from Cu_A to heme *a* is believed to occur via a route involving H260, the peptide bond between R481 and R482 and several hydrogen bonds leading up to the heme *a* propionates (Zhen, et al., 2002). When H260 is substituted with asparagine, oxygen reduction activity is reduced to 1% of wild-type and the Cu_A band at 830 nm is diminished. The redox-potential of Cu_A was raised by 118 mV compared to the wild-type resulting in very slow electron transfer from Cu_A to heme *a* (Wang, et al., 2002).

In this mutant, electron transfer to the binuclear center is slowed to a significant extent but is sufficient to allow enzyme-turnover. The -2 dalton mass shift in the intact mass and the MS/MS results shown in figure 5.2 reveals that the H284-Y288 crosslink is present in this mutant.

LpM This is the “loop mutant” in the subunit II of the *aa*₃-type oxidase, where the entire Cu_A-binding loop from residues 252-265 (**Cys-Ser-Glu-Ile-Cys-Gly-Ile-Ser-His-Ala-Tyr-Met-Pro-Ile**) is replaced with blue-copper binding loop from azurin (**Cys-Ser-Glu-Pro-Gly-His-Ser-Ala-Leu-Met-Lys-Gly**) (Zhen, et al., 2002). This substitution blocks the incorporation of copper into the oxidase, hence this oxidase lacks the Cu_A-site to receive electrons from cytochrome *c* and, as expected, has no measurable oxygen reduction activity.

The mass spectrometry results show that the tryptic peptide N258-K307 has a mass of 5533.87 Da, which is the identical to the predicted mass of the non-crosslinked wild-type

protein, indicating a lack of a crosslink (see figure 5.3). The MS/MS fragments confirm the absence of the crosslink.

D132N/K362M D132 is the residue at the entrance of the D-channel and K362 is a key residue for the K-channel. This double mutant blocks the entry of proton through either channel and the oxygen reduction activity of the enzyme is negligible (Pawate, 2005).

The mass spectrometry data on this double mutant (figure 5.4) shows that an intact crosslink is present at the active site of this enzyme. Both the intact mass and the fragment map confirm the presence of the crosslink. Although the proton transfer is blocked for any measurable oxygen reduction activity during steady-state, the post-translational modification has not been affected.

ΔCu_B mutant The wild-type *aa₃*-oxidase was expressed and purified from a *Δcox11* strain. The product of the *cox11* gene is the protein cox11p, which is involved in the assembly of Cu_B into the active site of the oxidase. When the wild-type expression plasmid was expressed in the *Δcox11* strain, the resulting enzyme did not have Cu_B in the active site, as confirmed by metal analysis and EPR spectroscopy (Hiser, et al., 2000). Although the intact mass of the peptide could not be determined due to low signal-to-noise, MS/MS fragment maps from the *ΔCu_B* mutant confirm the absence of the crosslink in any observed fragments (figure 5.5). It can therefore be concluded that the active site in this enzyme lacks the post-translational crosslink modification.

5.4 Conclusions

From the MS/MS results, it is obvious that simply slowing down the electron or proton transfer to the binuclear center, as exemplified in the cases of the H260N and D132N/K362M mutants, does not prevent the formation of the crosslinked cofactor at the active site. Although these enzymes have negligible oxygen reduction activity, the presence of the crosslink in the enzyme suggests that this level of activity is sufficient for the post-translational modification to form. Unlike electron delivery, which utilizes a more structured delivery chain and can be more easily blocked by mutating the electron-carrying amino acids, it is much harder to block proton delivery via the proton channels as the active site of the oxidase is connected to the bulk solution through both proton input and exit channels (Mills, et al., 2000). Also, since radical chemistry involves the oxidation of amino acids in the vicinity of the catalytic site, specific proton delivery for the crosslink formation does not seem essential.

In the case of the LpM (ΔCu_A) mutant, electrons are entirely unable to be transferred to heme *a* and this likely prevents the formation of the covalent crosslink at the active site. The complete loss of the crosslink due to a block in the electron-transfer pathway also negates the possibility of the involvement of any assembly enzyme in catalyzing the post-translational modification. The H-Y crosslink is also not seen in the ΔCu_B mutant and this indicates that the presence of Cu_B at the active site is vital for the covalent modification. This puts the heme-copper oxidases in the company of a host of other metalloproteins that undergo auto-catalyzed radical chemistry to form one or more covalent crosslinks among residues in the active site vicinity (Davidson, 2007). The amino acids most susceptible to form these cross-linked cofactors are cysteine, methionine, tryptophan, tyrosine and glycine and the formation is

generally catalyzed by a metal (Fe, Co, Cu and Mn) in the presence of oxygen. Upon modification, these protein-derived cofactors contribute to the catalysis in many ways the unmodified amino acids cannot. Commonly, they have been found to provide electrophilic sites for substrate interaction or to stabilize free-radical intermediates (Stubbe and van Der Donk, 1998).

This is significant because we can then be confident that cytochrome *c* oxidase also uses the covalent modification after radical chemistry converts the Y288 residue to form a stable tyrosyl radical during the catalytic cycle. Many research groups have localized the radical detected during the P_m state formation to the Y288 residue although it is not yet universally accepted (Budiman, et al., 2004; MacMillan, et al., 2006). The fact that the active-site tyrosine is conserved across the entire heme-copper superfamily and that mutations at the site affect both structure and function is indicative of involved contribution to catalysis beyond the delivery of protons.

Radical chemistry at the active site of the oxidase In heme proteins and copper proteins, as in the instances of catalase peroxidase (KatG)(Ghiladi, et al., 2005) and galactose oxidase (Rogers, et al., 2000; Rogers and Dooley, 2001) respectively, the active site heme or copper individually could generate the cross-linked amino acids during biosynthesis. From the results presented in this work, it is clear that Cu_B is important to the radical chemistry. Despite the lack of direct evidence, it stands to reason that heme *a*₃ is also essential for the biosynthesis of cross-linked amino acids in the heme-copper oxidases owing to the nature of the reaction in the the active site.

During the catalysis of cytochrome *c* oxidase, Cu_B and heme *a*₃ are known to provide one and two electrons respectively. It is possible that during the very first turnover, when oxygen

binds to the reduced binuclear center, it gets reduced with three electrons from these metal centers, resulting in the formation of a reactive hydroxyl radical. This hydroxyl radical may then oxidize the neighboring tyrosine and histidine residues enabling the formation of the crosslink. The crosslink thus formed may help in the catalytic function of the mature enzyme by facilitating electron donation and stabilizing the radical formed at Y288 (figure 5.6). The above hypothesis is quite reasonable since examples of this chemistry exist in the literature (Boguta and Dancewicz, 1983; Gross and Sizer, 1959). In general, it is accepted that $\bullet\text{OH}$ is capable of causing modifications to the primary structure of proteins and it has been demonstrated that hydrogen abstraction by hydroxyl radical produces tyrosyl radical (Davies, et al., 1987). These tyrosyl radicals may then react with nearby amino acids to form stable covalent bonds.

It is evident that it is the His-Tyr crosslink that makes the oxidase competent in the one-step reduction of oxygen without the release of partially reduced oxygen species (PROS). Such a protective mechanism is perhaps necessary since the physiological electron delivery by cytochrome *c* is significantly slower than intramolecular electron transfer in the oxidase. In a very elegant biomimetic experiment, it has been shown that, under limiting electron supply, the presence of Fe_{a3} , Cu_B and tyrosine were essential to the reduce oxygen without releasing PROS. The absence of any one of these resulted in a marked increase in the release of toxic PROS (Collman, et al., 2007).

In-vitro generation of the crosslink: Cytochrome oxidases do not lend themselves well to the incorporation of active site metals *in vitro* and hence it is difficult to externally induce the covalent modification as has been reported for many other enzymes. However, a mutant described in this study, the LpM (or ΔCu_A) mutant may be a suitable candidate for testing if the crosslink formation can be induced with exposure to H_2O_2 in the presence of a reductant.

This work has described some of the conditions found to be important for the formation of the post-translational crosslink modification between H284 and Y288 at the active site of cytochrome *c* oxidase. It has been demonstrated that active site metals, Fe_{a3} and Cu_B, in the presence of reductants and oxygen are sufficient to cause the formation of the crosslink. Parallels with other crosslinked metalloproteins imply that the formation of the crosslink is mediated by a hydroxyl radical causing the formation of the tyrosyl radical which, in turn, leads to the oxidation of His-284 to form the covalent linkage.

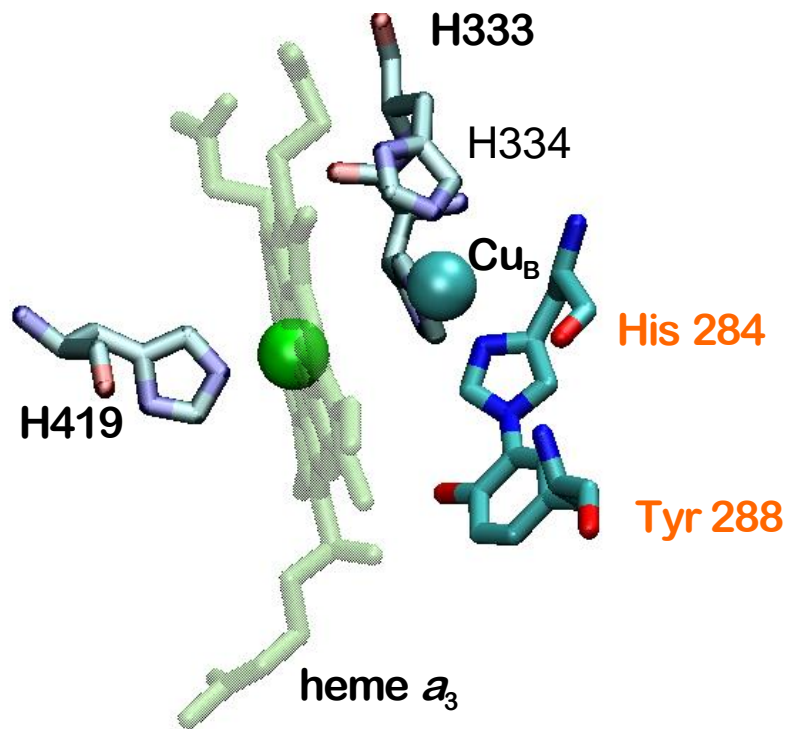


Figure 5.1. This schematic shows heme a_3 and Cu_B coordinated to their histidine ligands. The unique covalent modification between one of the Cu_B ligands and Y288 is shown in the figure. Y288 is at the end of the K-channel. The conserved motif for the crosslink in the A- and B-type oxidases is -Gly-His-Pro-X-Val-Tyr-. This figure was generated using VMD (Humphrey, et al., 1996).

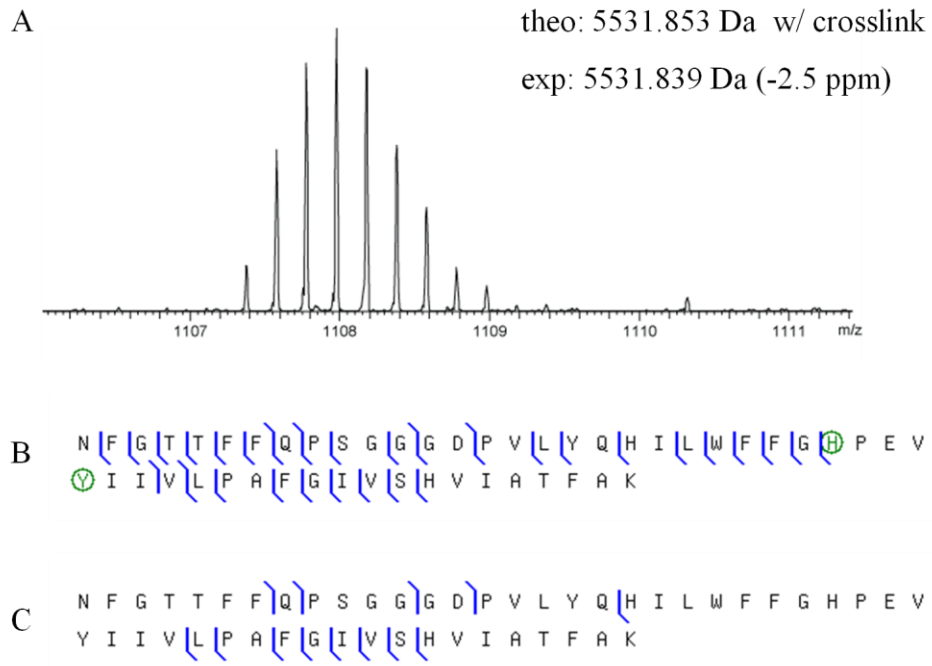


Figure 5.2. Mass spectrum and fragmentation maps of the N258-K307 tryptic peptide from the electron-path mutant H260N from *Rhodobacter sphaeroides*. (A) Mass spectrum of the intact, isolated peptide. (B) Fragmentation map of CAD MS/MS results when analyzed with assumption of a crosslink present (green circles) and (C) no crosslink. Although a mass which maps to a single non-crosslinked fragment exists, the fragmentation map results show that the species containing the His-Tyr crosslink is the dominant form in H260N.

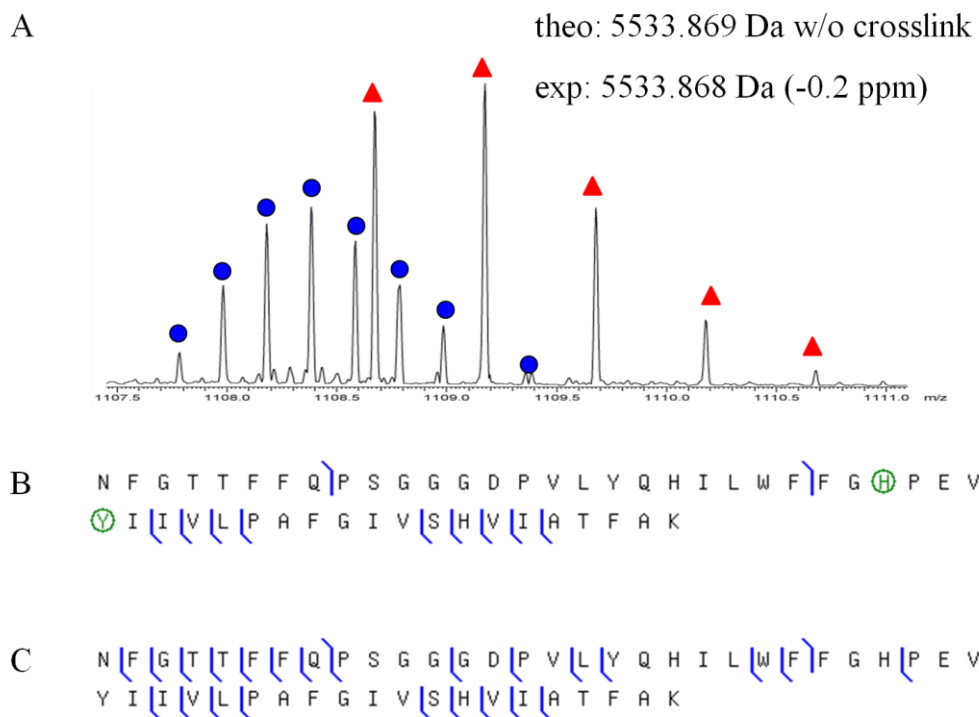


Figure 5.3. Mass spectrum and fragmentation maps of the N258-K307 tryptic peptide from the Cu_A mutant LpM from *Rhodobacter sphaeroides* (blue circles). The isotopic peaks identified with the red triangles are from a contaminating peptide. (A) Mass spectrum of the intact, isolated peptide. (B) Fragmentation map of CAD MS/MS results when analyzed with the assumption of a crosslink present (green circles) and (C) no crosslink. The fragmentation map results show that the His-Tyr crosslink is not present in LpM.

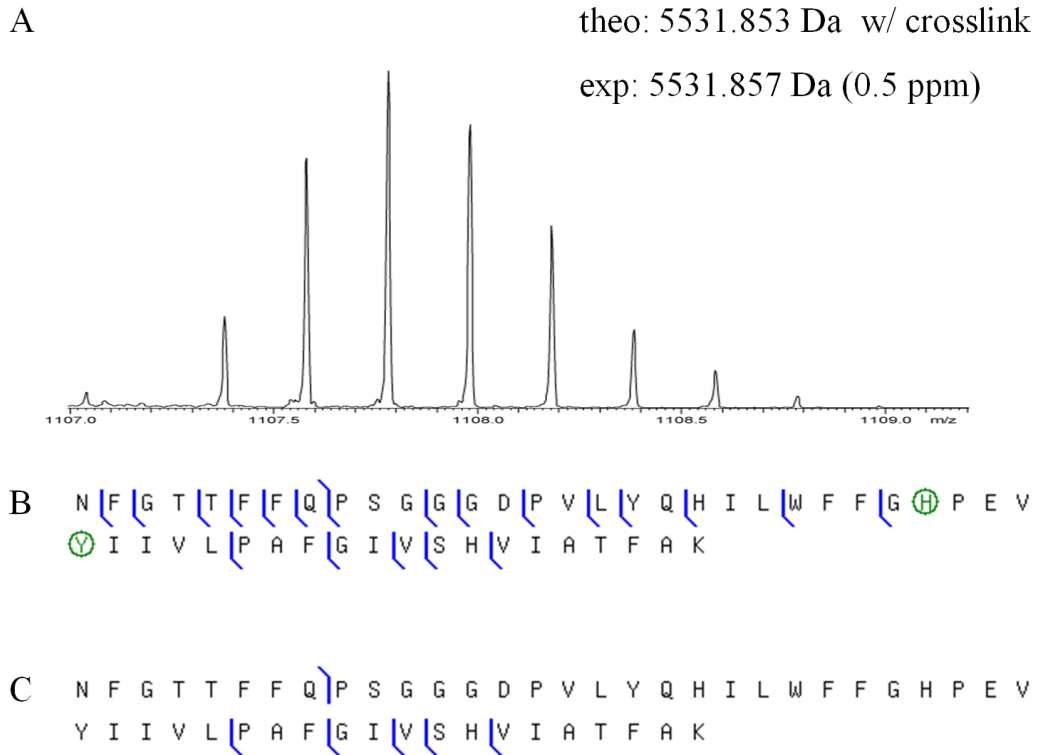


Figure 5.4. Mass spectrum and fragmentation maps of the N258-K307 tryptic peptide from the proton-channel double mutant D132N/K362M from *Rhodobacter sphaeroides*. (A) Mass spectrum of the intact, isolated peptide. (B) Fragmentation map of CAD MS/MS results when analyzed with the assumption of a crosslink present and (C) no crosslink. The fragmentation map results show that the His-Tyr crosslink is present in D132N/K362M.

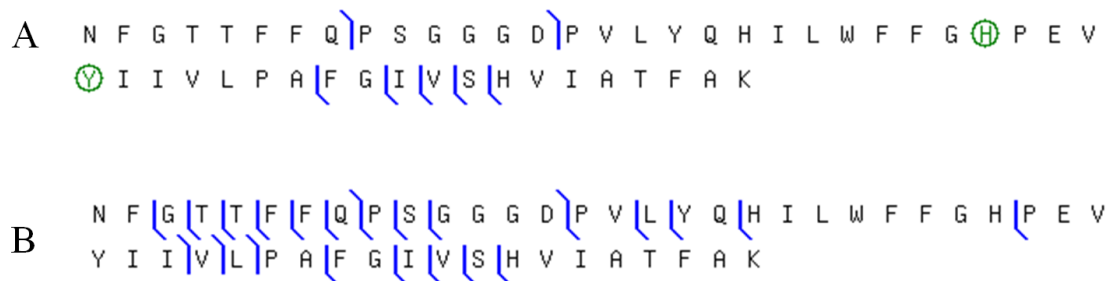


Figure 5.5. Fragmentation maps of the N258-K307 tryptic peptide from the ΔCu_B mutant from *Rhodobacter sphaeroides*. (A) Fragmentation map of CAD MS/MS results when analyzed with the assumption of a crosslink present and (B) no crosslink. The fragmentation map results show that the His-Tyr crosslink is not present in the ΔCu_B mutant.

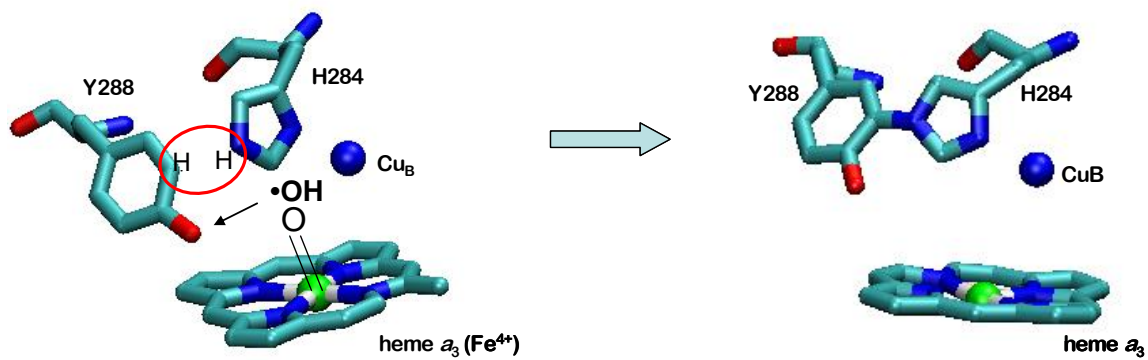


Figure 5.6. The active site of the oxidase before and after the post-translational modification. The formation of the crosslink is mediated by the hydroxyl radical. The heme a_3 structure has been truncated for clarity. This figure was generated using VMD (Humphrey, et al., 1996).

5.5 Literature Cited

- Babcock, G.T. (1999) How Oxygen is Activated and Reduced in Respiration, *Proc. Natl. Acad. Sci. USA*, **96**, 12971-12973.
- Blomberg, M.R., Siegbahn, P.E., Babcock, G.T. and Wikstrom, M. (2000) O-O bond splitting mechanism in cytochrome oxidase, *J Inorg Biochem*, **80**, 261-269.
- Boguta, G. and Danciewicz, A.M. (1983) Radiolytic and Enzymatic Dimerization of Tyrosyl Residues in Insulin, Ribonuclease, Papain and Collagen, *International Journal of Radiation Biology*, **43**, 249-265.
- Budiman, K., Kannt, A., Lyubenova, S., Richter, O.M., Ludwig, B., Michel, H. and MacMillan, F. (2004) Tyrosine 167: the origin of the radical species observed in the reaction of cytochrome c oxidase with hydrogen peroxide in *Paracoccus denitrificans*, *Biochemistry*, **43**, 11709-11716.
- Buse, G., Soulimane, T., Dewor, M., Meyer, H.E. and Bluggel, M. (1999) Evidence for a copper-coordinated histidine-tyrosine cross-link in the active site of cytochrome oxidase, *Protein Sci*, **8**, 985-990.
- Collman, J.P., Decreau, R.A. and Sunderland, C.J. (2006) Single-turnover intermolecular reaction between a Fe-III-superoxide-Cu-I cytochrome c oxidase model and exogeneous Tyr244 mimics, *Chemical Communications*, 3894-3896.
- Collman, J.P., Decreau, R.A., Yan, Y.L., Yoon, J. and Solomon, E.I. (2007) Intramolecular single-turnover reaction in a cytochrome c oxidase model bearing a Tyr244 mimic, *Journal of the American Chemical Society*, **129**, 5794-+.
- Collman, J.P., Devaraj, N.K., Decreau, R.A., Yang, Y., Yan, Y.L., Ebina, W., Eberspacher, T.A. and Chidsey, C.E. (2007) A cytochrome C oxidase model catalyzes oxygen to water reduction under rate-limiting electron flux, *Science*, **315**, 1565-1568.
- Das, T.K., Pecoraro, C., Tomson, F.L., Gennis, R.B. and Rousseau, D.L. (1998) The post-translational modification in cytochrome c oxidase is required to establish a functional environment of the catalytic site, *Biochemistry*, **37**, 14471-14476.
- Davidson, V.L. (2007) Protein-derived cofactors. Expanding the scope of post-translational modifications, *Biochemistry*, **46**, 5283-5292.
- Davies, K.J.A., Delsignore, M.E. and Lin, S.W. (1987) Protein Damage and Degradation by Oxygen Radicals .2. Modification of Amino-Acids, *Journal of Biological Chemistry*, **262**, 9902-9907.

- Ganesan, K. and Gennis, R.B. (2010) Blocking the K-pathway still allows rapid one-electron reduction of the binuclear center during the anaerobic reduction of the aa3-type cytochrome c oxidase from *Rhodobacter sphaeroides*, *Biochim Biophys Acta*, **1797**, 619-624.
- Gennis, R.B. (1998) Multiple Proton-conducting Pathways in Cytochrome Oxidase and a Proposed Role for the Active-site Tyrosine, *Biochim. Biophys. Acta*, **1365**, 241-248.
- Ghiladi, R.A., Medzihradzky, K.F. and Ortiz de Montellano, P.R. (2005) Role of the Met-Tyr-Trp cross-link in *Mycobacterium tuberculosis* catalase-peroxidase (KatG) as revealed by KatG(M255I), *Biochemistry*, **44**, 15093-15105.
- Gorbikova, E.A., Belevich, I., Wikstrom, M. and Verkhovsky, M.I. (2008) The proton donor for O-O bond scission by cytochrome c oxidase, *Proc Natl Acad Sci U S A*, **105**, 10733-10737.
- Gross, A.J. and Sizer, I.W. (1959) Oxidation of Tyramine, Tyrosine, and Related Compounds by Peroxidase, *Journal of Biological Chemistry*, **234**, 1611-1614.
- Hemp, J., Christian, C., Barquera, B., Gennis, R.B. and Martinez, T.J. (2005) Helix switching of a key active-site residue in the cytochrome cbb3 oxidases, *Biochemistry*, **44**, 10766-10775.
- Hemp, J., Robinson, D.E., Ganesan, K.B., Martinez, T.J., Kelleher, N.L. and Gennis, R.B. (2006) Evolutionary migration of a post-translationally modified active-site residue in the proton-pumping heme-copper oxygen reductases, *Biochemistry*, **45**, 15405-15410.
- Hiser, L., Di Valentin, M., Hamer, A.G. and Hosler, J.P. (2000) Cox11p is required for stable formation of the Cu(B) and magnesium centers of cytochrome c oxidase, *J Biol Chem*, **275**, 619-623.
- Horn, D.M., Zubarev, R.A. and McLafferty, F.W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules, *Journal of the American Society for Mass Spectrometry*, **11**, 320-332.
- Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: visual molecular dynamics, *J Mol Graph*, **14**, 33-38, 27-38.
- Iwata, S., Ostermeier, C., Ludwig, B. and Michel, H. (1995) Structure at 2.8 Å resolution of cytochrome c oxidase from *Paracoccus denitrificans*, *Nature*, **376**, 660-669.
- LeDuc, R.D., Taylor, G.K., Kim, Y.B., Januszyk, T.E., Bynum, L.H., Sola, J.V., Garavelli, J.S. and Kelleher, N.L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry, *Nucleic Acids Res*, **32**, W340-W345.
- MacMillan, F., Budiman, K., Angerer, H. and Michel, H. (2006) The role of tryptophan 272 in the *Paracoccus denitrificans* cytochrome c oxidase, *FEBS Lett*, **580**, 1345-1349.
- MacMillan, F., Kannt, A., Behr, J., Prisner, T. and Michel, H. (1999) Direct evidence for a tyrosine radical in the reaction of cytochrome c oxidase with hydrogen peroxide, *Biochemistry*, **38**, 9179-9184.

- McCauley, K.M., Vrtis, J.M., Dupont, J. and van der Donk, W.A. (2000) Insights into the functional role of the tyrosine-histidine linkage in cytochrome c oxidase, *Journal of the American Chemical Society*, **122**, 2403-2404.
- Mills, D.A., Florens, L., Hiser, C., Qian, J. and Ferguson-Miller, S. (2000) Where is 'outside' in cytochrome c oxidase and how and when do protons get there?, *Biochim Biophys Acta*, **1458**, 180-187.
- Ostermeier, C., Iwata, S., Ludwig, B. and Michel, H. (1995) Fv fragment-mediated crystallization of the membrane protein bacterial cytochrome c oxidase, *Nat Struct Biol*, **2**, 842-846.
- Pawate, A. (2005) The aa₃ type cytochrome c oxidase from *Rhodobacter sphaeroides*: Insights into the proton pumping mechanism from the N139D mutation in the D-channel.
- Pinakoulaki, E., Pfitzner, U., Ludwig, B. and Varotsis, C. (2002) The role of the cross-link His-Tyr in the functional properties of the binuclear center in cytochrome c oxidase, *J Biol Chem*, **277**, 13563-13568.
- Proshlyakov, D.A., Pressler, M.A. and Babcock, G.T. (1998) Dioxygen activation and bond cleavage by mixed-valence cytochrome c oxidase, *Proc Natl Acad Sci U S A*, **95**, 8020-8025.
- Proshlyakov, D.A., Pressler, M.A., DeMaso, C., Leykam, J.F., DeWitt, D.L. and Babcock, G.T. (2000) Oxygen activation and reduction in respiration: involvement of redox-active tyrosine 244, *Science*, **290**, 1588-1591.
- Qin, L., Liu, J., Mills, D.A., Proshlyakov, D.A., Hiser, C. and Ferguson-Miller, S. (2009) Redox-dependent conformational changes in cytochrome C oxidase suggest a gating mechanism for proton uptake, *Biochemistry*, **48**, 5121-5130.
- Rauhamaeki, V., Baumann, M., Soliymani, R., Puustinen, A. and Wikstrom, M. (2006) Identification of a histidine-tyrosine cross-link in the active site of the cbb3-type cytochrome c oxidase from *Rhodobacter sphaeroides*, *Proc Natl Acad Sci U S A*, **103**, 16135-16140.
- Rogers, M.S., Baron, A.J., McPherson, M.J., Knowles, P.F. and Dooley, D.M. (2000) Galactose oxidase pro-sequence cleavage and cofactor assembly are self-processing reactions, *Journal of the American Chemical Society*, **122**, 990-991.
- Rogers, M.S. and Dooley, D.M. (2001) Posttranslationally modified tyrosines from galactose oxidase and cytochrome c oxidase, *Adv Protein Chem*, **58**, 387-436.
- Rogers, M.S. and Dooley, D.M. (2003) Copper-tyrosyl radical enzymes, *Curr Opin Chem Biol*, **7**, 189-196.
- Soulimane, T., Buse, G., Bourenkov, G.P., Bartunik, H.D., Huber, R. and Than, M.E. (2000) Structure and Mechanism of the Aberrant ba₃-cytochrome c Oxidase from *Thermus thermophilus*, *EMBO J*, **19**, 1766-1776.

Stubbe, J. and van Der Donk, W.A. (1998) Protein Radicals in Enzyme Catalysis, *Chem Rev*, **98**, 705-762.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. and Yoshikawa, S. (1996) The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å, *Science*, **272**, 1136-1144.

Wang, K., Geren, L., Zhen, Y., Ma, L., Ferguson-Miller, S., Durham, B. and Millett, F. (2002) Mutants of the CuA site in cytochrome c oxidase of *Rhodobacter sphaeroides*: II. Rapid kinetic analysis of electron transfer, *Biochemistry*, **41**, 2298-2304.

Zhen, Y., Schmidt, B., Kang, U.G., Antholine, W. and Ferguson-Miller, S. (2002) Mutants of the CuA site in cytochrome c oxidase of *Rhodobacter sphaeroides*: I. Spectral and functional properties, *Biochemistry*, **41**, 2288-2297.