Pushing the Limits of Structural Mass Spectrometry to Characterize the Viral Packaging Motor

Lucas Monkkonen

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Carlos E. Catalano, Chair

David R. Goodlett

Kelly K. Lee

Program Authorized to Offer Degree:

Pharmacy-Medicinal Chemistry

University of Washington

Abstract

Pushing the Limits of Structural Mass Spectrometry to Characterize the Viral Packaging Motor

Lucas Monkkonen

Chair of the Supervisory Committee:
Professor Carlos E. Catalano
Medicinal Chemistry

Assessing protein structure is a difficult undertaking, yet is often required for a complete mechanistic understand of how a protein functions. The traditional techniques used to assess protein structure are X-ray crystallography and NMR, which both yield high-resolution data. Despite the success of these techniques, NMR and crystallography are not high-throughput techniques, have size limitations, require protein crystals, and/or require high concentrations of protein purified to homogeneity. Alternative techniques are required to probe the structures of proteins that are difficult to purify and rapidly characterize protein structure. Toward this end, structural mass spectrometry techniques provide a powerful alternative due to high sensitivity, versatility, and potential for fairly detailed structural information. All of the work described in this thesis concerns structural MS in some capacity.

A major focus of this work is the bacteriophage lambda packaging motor, called terminase. Terminase is a complex, multi-subunit viral enzyme that is one of the most powerful molecular motors known. This enzyme packages a double stranded DNA genome into empty

viral capsids at a rate of 600 base pairs per second, which generates over 20 atmospheres of internal capsid pressure. The packaging of the viral genome by a terminase enzyme is crucial for producing new infectious virus for many dsDNA viruses such as adenovirus, herpesvirus, and several bacteriophages.

In the first chapter, new structural MS technologies are developed and discussed. A new ionization technique, called surface acoustic wave nebulization (SAWN), was coupled with digital microfluidics (DMF) to show a novel workflow that could lead to a lab-on-a-chip type workflow for hydrogen-deuterium exchange (HDX). In DMF-SAWN-HDX, a 5 µl peptide sample was successful merged with deuterated water, transferred to the SAWN chip for nebulization and analyszed by MS. Additionally, SAWN-HDX alone was shown to be effective at measuring HDX of ubiquitin, and the SAWN-HDX data of ubiquitin were corroborated with NMR HDX ubiquitin data.

I then demonstrate the utility of SAWN-MS for a novel application: the structural characterization of noncovalent interactions. Comparisons of SAWN with ESI showed several potential advantages of SAWN over ESI. SAWN-MS analysis of myoglobin resulted in less heme dissociation when analyzing myoglobin in comparison to ESI, and SAWN had less evidence of nonspecific noncovalent interactions. SAWN of the TerS-A55 deletion mutant containing the DNA binding domain of the small terminase subunit led to little to no dissociation of the dimer, while ESI disrupted significantly more of noncovalent interactions. SAWN-HDX studies of TerS-A55 were found to exhibit solvent protection at the dimer interface and SAWN-HDX demonstrated a clear decrease in deuterium uptake in the presence of DNA.

Finally, I discuss the development and application of cross-linking MS (XLMS) to study terminase. I validated a new means of analyzing XLMS data, which resulted in the identification

of more cross-links in comparison to other methods. XLMS data was further validated with cross-linking data from proteins with known structures, and this data showed that cross-links accurately reflect the native solution structure. XLMS was then applied to lambda terminase, a protein with unknown structure, in combination with homology modeling. These results provide valuable structural insight into the structure of TerS and the packaging mechanism.

Overall, the work described here advances structural MS and our understanding of lambda terminase. Novel structural MS technologies and workflows are described, which provide a future means to streamline workflows, improve analysis, yield higher quality data, increase the accessibility of certain structural MS techniques, and provide several avenues for continued future development. The last chapter provides valuable structural insight into the assembly of the TerS dimer and the structural plasticity within the TerL subunit, which has important mechanistic implications. This work also more broadly presents a detailed means for how to approach structural characterization of proteins that are difficult to purify and analyze by traditional structural methods.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

Chapter 1

Introduction to Structural Mass Spectrometry and the Viral Packaging Motor

Part 1: Introduction to Structural MS

1.1.1 Introduction to Protein Structure

Proteins perform a vast majority of all cellular processes, are chemically very complex, and have an intricate architecture [1].To name a few functions, proteins carry out signal transduction, catalysis, and genetic regulation [2]. Proteins are essential to the growth and homeostasis of cells and organisms, and even slight disturbances in a protein's activity can drastically harm a cell and organism, as evidenced by many different diseases [2]. Single amino acid substitutions, caused by point mutations, can lead to cystic fibrosis [3], sickle cell anemia [4], and cancer [5], and even slight changes such as post-translational modification of proteins have been implicated in diseases such as Parkinson's and Alzheimer's disease [6]. Individual proteins are largely constructed from the same twenty different amino acids, but this affords an extraordinary chemical diversity. For example, a simple 100 amino acid protein can have 10^{130} possible sequences, yet many proteins exceed this length [7]. It is now well understood that the chemistry of the amino acids typically determine two levels of protein structure (secondary and tertiary): smaller sub-features within the greater protein structure such as alpha helices and beta sheets, and an amalgam of these sub-structures that interact and form a unique overall shape [2]. To complicate matters further, many proteins do not act alone but rather via noncovalent complexes of varying size. For example, RuBisCo [8], an octomer responsible for carbon fixation in plants, and the spliceosome [9], which processes mRNA and has 10 types of subunits

in varying amounts. In humans, it is estimated that there are 650,000 protein-protein interactions [10].

The vast chemical and structural diversity in proteins is ultimately what allows proteins to accomplish a myriad of cellular tasks. For instance, many enzymes have intricate active sites that not only provide a specific pocket contoured to the shape of its substrate, but also add or withdraw electron density at specific times during a chemical reaction to greatly enhance the rate of a chemical reaction, as with nitric oxide synthase [11, 12]. Also, proteins can adopt multiple conformations, which can be triggered by binding specific substrates. For instance, the influenza viral fusion protein, hemagglutinin, undergoes a dramatic conformational change upon binding to a host receptor that converts a more compact surface protein into a harpoon-like structure that allows the virus to attach to the host cell and initiate viral infection [13]. Furthermore, the discovery of numerous structural motifs with specific functions across different proteins further highlight the relationship between protein structure and function. To name a few, Walker A nucleotide binding motifs [14], DNA binding motifs such as zinc-fingers [15], winged helix-turn-helix folds [16] are all common protein structural features that have been characterized. Additionally, proteins form heterogeneous, multi-protein complexes, which lead to greater diversity in function. Hemoglobin and GroEL-ES complexes are classic examples of large, multiprotein complexes. In hemoglobin, there are four different subunits with oxygen binding sites that cooperatively help the other sites bind oxygen, which increases the efficiency of hemoglobin to carry and deliver oxygen [17, 18]. The GroEL-ES complex is a very large chaperone complex, and put simplistically, the subunits collectively form a cylindrical folding chamber with a lid to correctly fold a protein [19]. Over the years, a tremendous amount of research has been done to characterize protein structures to understand their function, yet due to

the vast chemical diversity and sheer number of proteins in organisms (e.g. 20,000 - 25,000 proteins in humans [20]), it is difficult to get a comprehensive understanding as to how proteins carry out their vast array of biological processes within a cell.

The advent of protein X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) were crucial to our understanding of protein structure-function relationships. In 1958, the first crystal structure was published by Sir John Kendrew of sperm whale myoglobin [21], and two years later a crystal structure of hemoglobin (5.5 Å resolution) was published by Dr. Max Perutz [22], who were jointly awarded the Nobel Prize in 1962. In 1985, the first protein NMR structure was published of proteinase inhibitor IIA by Dr. Kurt Wüthrich's laboratory [23], and in 2002, Dr. Wüthrich was awarded a nobel prize. These techniques fueled the search for more protein structures and gave momentum to biochemical research. Since then, many NMR and crystal structures have been published as seen by the rapidly growing RCSB protein structure database (>77,000 as of 2011).

1.1.2 Beyond NMR and Crystallography

NMR and crystallography are the traditional methods used in structural biology. Both techniques yield very high-resolution data that can be complementary to one another [24]. NMR yields native solution structure of proteins yet crystallography requires protein crystals, which is challenging since the appropriate crystallization conditions can vary widely and cannot be predicted for a protein. Crystallography can be used to obtain structures for macromolecules quite large in size (such as the ribosome), yet characterizing proteins >35kDa by NMR presents a significant challenge. Both techniques also have shared disadvantages: they are not high-

throughput, require very pure and concentrated protein sample, and have complex data analysis [25].

Due to these constraints, a number of different techniques have been developed to investigate the structure of protein complexes[26–28]. Molecular biology techniques including yeast-two-hybrid, affinity purification, and site-directed mutagenesis, along with spectroscopic techniques such as surface plasmon-resonance and fluorescence resonance energy transfer have been used to locate subunit binding interfaces, establish the spatial arrangement of subunits within a complex, and/or identify interacting proteins. Microscopy techniques such as electron microscopy, atomic force microscopy, and cryoelectron microscopy, have been used in one way or another to establish overall protein complex three-dimensional shape, symmetry, subunit binding interfaces, and/or subunit packing, sometimes with a very high resolution [29]. Additionally, analytical ultracentrifugation, ion mobility separations, and small angle x-ray scattering can be used (much like microscopy techniques), to determine overall protein complex shape. Subunit stoichiometry can be probed using quantitative immunoblotting and analytical ultracentrifugation. This list of structural techniques is by no means comprehensive, and is mainly meant to highlight the diversity of techniques and mention some of the main techniques playing a large role in probing secondary, tertiary, and quaternary protein structure**. Figure 1.1** shows many structural biology techniques available and what information can be obtained from them.

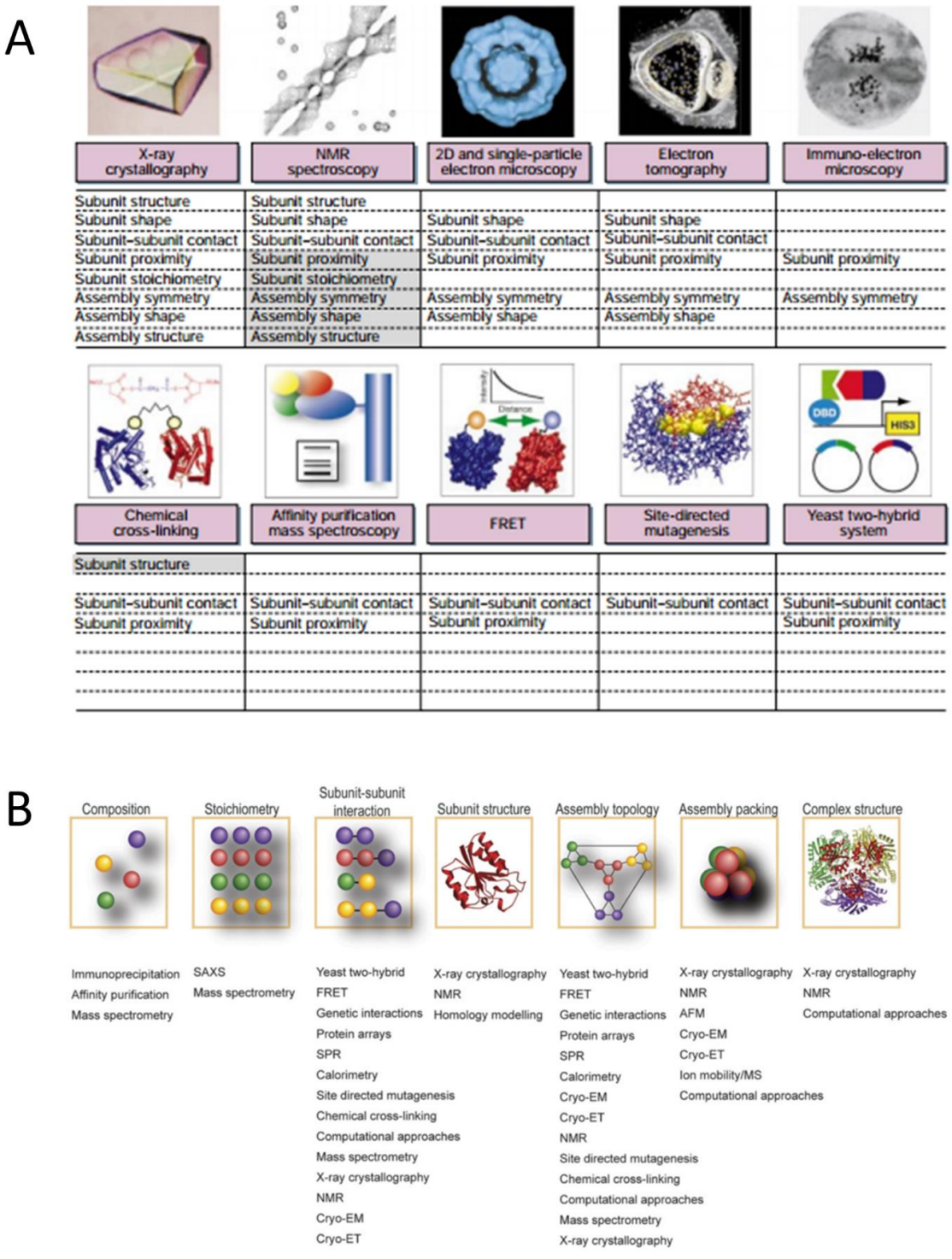


Figure 1.1. Overview of methods used in structural biology. **Panel A** is from Sali et al. [27].

Panel B is from Sharon [28].

In structural biology, it is now clear that the array of techniques make it possible to use multiple methods to establish protein structure with greater consistency and accuracy. No one technique is all encompassing and completely informative of the native solution protein structure, and combining techniques can make up for the disadvantages of a particular technique. Several examples of studies combing techniques are of the 26S proteasome, nuclear pore complex, and the 70S ribosome [26].

1.1.3 Computational Modeling of Proteins

In the post-genomic era, access to genetic information, the Uniprot protein database, and the large repository of protein structures found on RCSB led to a paradigm shift in structural biology. Instead of exclusive reliance on experimental techniques, a concerted effort has been put into mining the wealth of information in these databases and utilizing it to create computational approaches to predict protein structure. Predicting protein structure is now common, and little expertise is required to do this. Structure prediction referred to as homology modeling, is an incredibly powerful tool that is a good starting point for any structural biology problem. In this technique, a protein sequence is submitted to an online server such as ITASSER or Robetta [30–33]. Then, sophisticated computer algorithms analyze the protein sequence in order to search for proteins with known structure having similar sequences. The highly homologous sequence regions are aligned and structural features from the aligned portion of the protein contribute to the structural model of a protein. A final model is then produced after a computationally based energy minimization. In the absence of high-resolution structural information from crystallography and NMR, using homology modeling is the best means for

obtaining a sense for what the overall secondary, tertiary, and quaternary structure may be. Homology modeling efforts are not limited to targeted protein structure prediction, and efforts are ongoing to predict protein structures for all proteins in the human proteome [34]. Homology modeling will be utilized and discussed in chapter 4.

1.1.4 The Increasing Prominence of MS in Structural Biology

While half of the 2002 Nobel Prize was awarded to Kurt Wüthrich for developments in protein NMR analysis, the other two quarters of the prize was shared by mass spectrometrists John Fenn and Koichi Tanaka. This Nobel prize was awarded for the "development of methods for identification and structure analyses of biological macromolecules". Since these awards, mass spectrometry has come to play a large role in analyzing protein structure. A brief discussion of the history of MS is presented here.

John Fenn and Koichi Tanka are famous for their invaluable contribution to MS. They developed electrospray ionization and matrix assisted laser desorption ionization techniques, respectively, as a means for taking a solution phase biological molecule and transforming it into a charged gaseous particle [35, 36]. Combining these ionization methods with the development of mass analyzers such as the ion trap [37], quadrupole [38], ion cyclotron resonance [39], orbitrap [40], and time of flight [41] proved to be a powerful combination. These mass spectrometers function as sophisticated molecular scales that are capable of manipulating ions, selecting ions, and measuring their mass to charge ratio. The power of mass spectrometry is greatly increased with the use of hybrid instrumentation [42] and gas phase fragmentation techniques such as collision induce dissociation [43]. Hybrid instruments allow for greater versatility and fragmentation provides an indispensable means for analyzing the composition of

the incoming ions and increasing confidence of the identity of the analyte [42]. The utility of mass spectrometry is undeniable with the MS market today being a multibillion dollar industry, growing around seven percent per year [44, 45].

Arguably, mass spectrometry (MS) has risen to prominence mainly due to the field of proteomics [46–48]. The speed, sensitivity, selectivity, dynamic range, and mass accuracy of commercially available instruments are astounding, and MS is the most capable instrument for identifying and quantifying hundreds to thousands of proteins at a time [48, 49]. The success of MS based proteomics is in part due to the use of proteases, and reliability and predictability in peptide fragmentation [50, 51]. Proteases such as trypsin can take proteins and digest them into peptide fragments with high selectivity after certain amino acids. This is advantageous since peptides are much smaller and easier to analyze and ionize, and generally remain solvated while many proteins tend precipitate during the analysis [52]. Also, when peptides are subjected to regular collision induced dissociation, the peptides fragment along the backbone in a predictable manner, generating fragments that can be used to deduce the amino acid sequence of the original peptide [50, 53]. The most widely used fragments to sequence peptides are shown in **Figure 1.2**, where fragmentation occurs across the amide bond, creating a so-called "ladder" where the spacing between peaks corresponds to the mass to charge ratio of an amino acid. These fragmentation patterns allow researchers to take genomic data, predict proteins that should be present in a sample, and match experimental data against the thousands of predicted tryptic peptides and their tandem mass spectra [54].

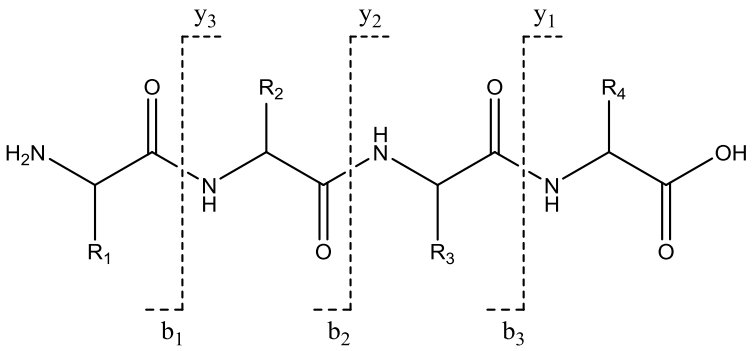


Figure 1.2. Diagram showing fragmentation of a peptide into b and y ions. The long vertical dashed line indicates that fragmentation occurs between the amide bond. The directionality of the short lines near the ion labels indicates which side of the cleaved peptide bond retains the charge and is recorded by the mass spectrometer.

The more sophisticated and powerful mass spectrometers also led to the development of other structural biology techniques. Many structural MS-based methods are listed in **Figure 1.1B**. As with almost all structural MS techniques, there are three major advantages: i) very little protein is required, often 2-3 orders of magnitude lower than NMR and crystallography [55], ii) MS often does not require high purity protein samples, iii) MS is relatively high throughput while providing detailed structural information, which is a large contrast to NMR and crystallography [28]. There are several structural approaches that use a similar "bottom-up" style approach as proteomics. Hydrogen-deuterium exchange (HDX) is one of the oldest MS techniques for studying protein structure, and is a useful approach for probing secondary through quaternary structure [56]. In this technique a protein sample is incubated at various time points with deuterium oxide, and the backbone amide hydrogens exchange with deuteriums at a rate that reflects the solvent accessibility of that particular hydrogen and the reactivity of the

hydrogen. HDX is greatly decreased when the proton is involved in alpha helices and beta sheets due to hydrogen bonding. The samples are flash frozen, quenched in acid, digested by pepsin, and analyzed by high resolution MS in order to resolve the isotopic distributions of deuterated peptides [56]. HDX has been used to probe secondary structure, protein conformations and dynamics, binding interfaces, and ligand binding faces. This method has proven to be quite successful, yet it can be challenging to retain the deuterium labels since the MS analysis exposes the sample to non-deuterated solvents and extreme conditions during ionization [57]. A similar method has recently been used to probe the solvent accessible space with fast photochemical oxidation techniques [58]. The essential principle of this technique is very similar to HDX in that it probes solvent accessible space, but instead of a temporary deuterium label, a permanent oxidation occurs when the protein comes in contact with an oxidizing agent such as hydrogen-peroxide. HDX will be discussed in chapters 2 and 3 of this thesis, where a novel workflow for HDX is established and used to study the bacteriophage lambda viral packaging motor.

Another bottom-up based structural approach is cross-linking in combination with mass spectrometry (XLMS). This method involves covalently modifying a protein with a bifunctional reagent reactive to specific amino acids (e.g, lysine or acidic residues). After the cross-linking reaction, the proteins are digested and analyzed by high-resolution tandem MS to assess which amino acids are physically cross-linked. This method is useful, and provides low-resolution distance constraints in a protein complex and can also reveal subunit binding interfaces [25, 55, 59, 60]. When used in combination with homology modeling, XLMS is quite valuable in validating that the homology model reflects the native protein solution. It is a unique approach that holds much promise for achieving 3-dmensional models of proteins relatively quickly while avoiding challenges of NMR and crystallography. The main challenge with this technique lies in

the data analysis. The unequivocal assignment of cross-linked peptides and their exact location remains difficult due to the inability of most software to annotate a significant portion of a tandem mass spectrum and the availability of few open source, user-friendly programs to analyze cross-linking data visually [61]. New advances in XLMS software analysis will be discussed in chapter 4, along with the application of XLMS to study terminase.

A third category of structural mass spectrometry is native MS. The goal of native MS is to place the protein complex of interest in quasi-native conditions in buffers amenable to MS, and ionize the intact complex [62–65]. This method is useful for measuring complex stoichiometry, observing ligand binding, and probing subunit assembly [66]. Complexes as large as the GroEL-ES complex and whole viral capsids have been ionized and characterized by MS [67, 68]. Currently, analysis of complexes by native MS only requires a few microliters for stable spray. Sophisticated time-of-flight mass spectrometers such as the Waters synapt (Manchester, UK) and the Thermo Orbitrap (Bremen, Germany) have been designed to optimally desolvate and manipulate large protein complexes [69]. The main challenges with this structural MS method is that the electrospray ionization process can cause complex dissociation and it is still not clear to what extent protein complexes retain their native solution structure in the gas phase. Chapter 3 will discuss the use of a novel ionization technique called surface acoustic wave nebulization as a gentler means to ionize noncovalent protein complexes.

Part 2: Introduction to Terminase, the Viral Packaging Motor

1.2.1 What is Terminase?

Bacteriophage lambda terminase is a complex, multi-subunit viral enzyme that is one of the most powerful molecular motors known [70]. This enzyme is responsible for packaging double stranded DNA (dsDNA) inside empty viral capsids at a rate of 600 base pairs per second. DNA is packaged to liquid crystalline densities, which results in an internal pressure of 20 atmospheres, making this enzyme 25 times more powerful than myosin [71, 72]. Viral packaging is a process crucial to producing new, infectious viruses for many dsDNA viruses. In fact, human dsDNA viruses such as adenoviruses and herpes viruses undergo viral packaging through a terminase enzyme, thus making lambda terminase a good model system for studying human dsDNA terminase [73]. A structural model for lambda terminase is needed for two reasons: i) antiviral drug development for human viruses that undergo a terminase mediate packaging process and ii) basic understanding of the mechanism for how lambda terminase can package capsids with such force.

Terminase enzyme performs two main DNA-centered processes necessary for building new viruses [74, 75]. First, terminase facilitates a process called genome maturation. Terminase binds to the end/beginning of a lambda genome within a concatemer of newly synthesized viral genomes and prepares a single-stranded mature 12 base-pair overhang. Second, terminase carries out DNA packaging. In this process, terminase must bind to an empty capsid (i.e., a procapsid) and efficiently package one lambda genome inside of it. This involves hydrolyzing ATP to power the packaging of the entire 48.5 kbp lambda genome through a small opening in the capsid. After packaging the full genome length, terminase then must cut the DNA and release

from the capsid. Terminase is then free to bind another empty capsid for packaging. The cycle continues to produce hundreds of DNA filled capsids [71, 76].

Lambda terminase is a hetero-oligomer comprised of a small subunit (TerS) and a large subunit (TerL); TerS and TerL are often referred to as gpNu1 and gpA in the literature, respectively. TerS is the small subunit that contains the DNA binding domain and TerL is the large subunit that contains the catalytic domains. Shown in **Figure 1.3** are the two main assembly states of terminase. The protomer is the smallest repeating unit of terminase and contains two TerS and one TerL subunit. Four protomer units further assembles into a ring tetramer and this assembly is capable of all DNA maturation and packaging activities. *In vitro* solution conditions favoring oligomer disassembly and assembly have been elucidated by Andrews *et al*. as shown in **Figure 1.3**.

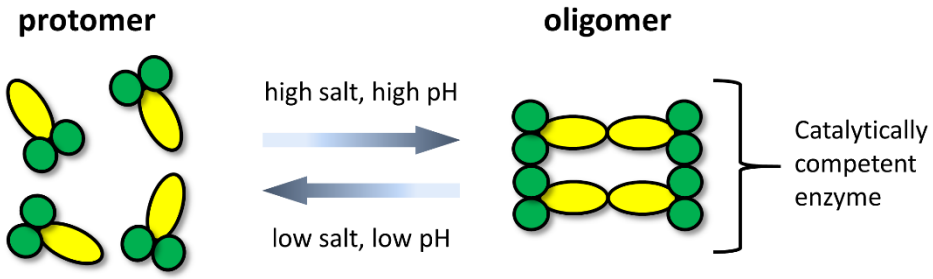


Figure 1.3. Diagram of the two main assembly states of lambda terminase. Green circles represent TerS and yellow elipses represent TerL.

1.2.2 Relevant Structural Information

Currently, there is limited structural data available for lambda terminase. This is largely because purifying the individual small and large subunits is not possible since they form aggregates, and the whole oligomeric assembly responsible for viral packaging is difficult to characterize due to its large size. Nevertheless, mutagenesis, biochemical and kinetics studies, azido ATP labelling studies, and other approaches have been used to generate a rough map of the domains as shown in **Figure 1.4** [71, 77–80]. The most striking feature of terminase from **Figure 1.4** is the fact that TerS is responsible for specific binding to the lambda genome only, and TerL carries out the bulk of the catalytic activities including genome maturation and packaging.

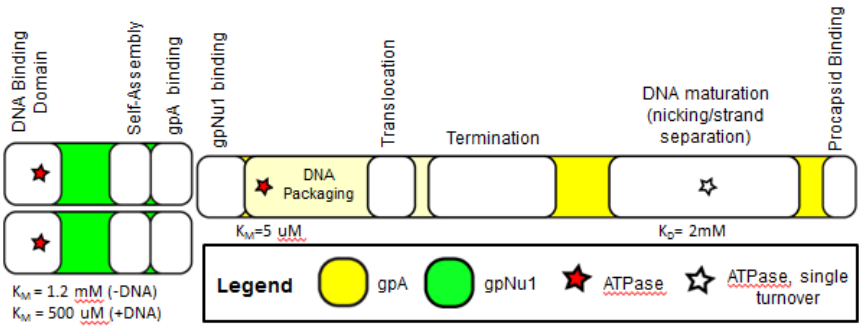


Figure 1.4. Putative domain organization of lambda terminase. Protomer is shown above stretched lengthwise from N- to C-terminus. (Figure courtesy of Ben Andrews)

An NMR structure for the first 55 amino acids of TerS, containing the portion responsible for specific DNA binding activity has been determined (**Figure 1.5A**) [81]. This TerS-A55 deletion mutant could be purified and concentrated in large quantities, since this mutant excludes the highly hydrophobic region most likely responsible for cause full-length TerS to aggregate [82]. TerS-A55 forms an extremely stable homodimer in solution and contains two winged helix-

turn-helix motifs that atypically face 180 degrees from one another. However, the DNA binding domain orientation was consistent within the context that a native *E. Coli* protein called integration host factor (IHF) has a cooperative effect on terminase binding to the genome [83]. IHF is known to introduce a 180 degree bend in the DNA, so it was hypothesized that the TerS-A55 complex binds across the bent substrate, as depicted in the model shown in **Figure 1.5B**. Structural MS data concerning TerS-A55 is included in chapter 3.

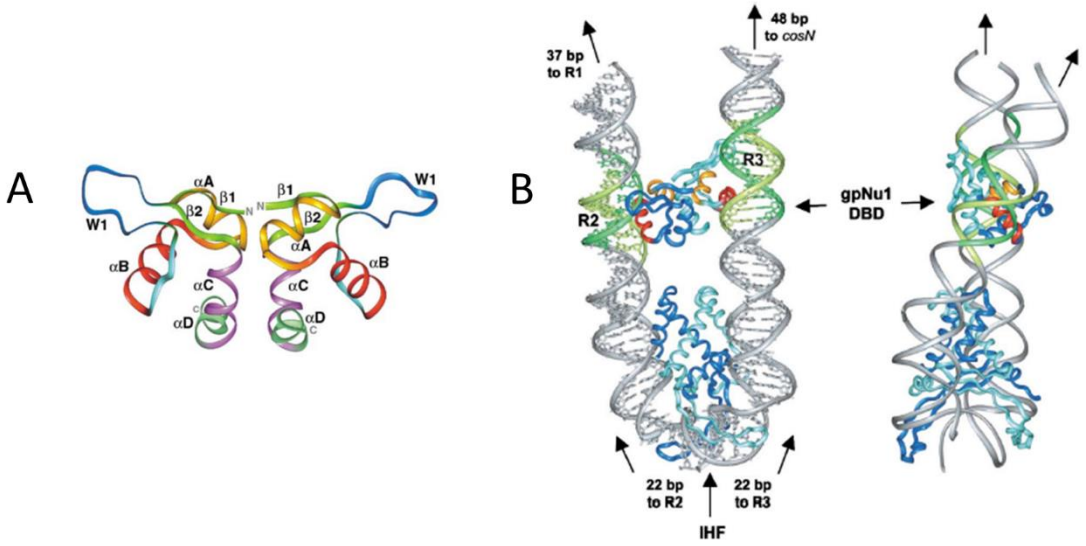


Figure 1.5. **Panel A** shows the ribbon structure of Ters-A55 derived from the NMR data. **Panel B** shows the proposed model of how the TerS-A55 interacts with λ genome bent 180° by integration host factor. Both diagrams are from de Beer et al. [81].

There is no structure available for lambda TerL. Structures of full length TerL and TerS for other dsDNA bacteriophages have been published. Full-length TerS structures have been published for bacteriophages SF6, 44RR, and P22 [84, 85]. TerS in SF6 forms an octamer and while that of 44RR forms a 12-mer; both subunits are highly helical and assemble into a ring.

Electron microscopy studies of P22 TerS reveal a similar overall structure for a nonameric ring assembly [86]. There are also several structures of TerL available. The only available full-length TerL subunit crystal structure is from bacteriophage Sf6 and T4 [87]. Generally speaking, the results from the crystal structure reveal two distinct domains separated by a linker region; an ATPase domain that is largely involved in the motor activity and a second globular DNA maturation domain. It is hypothesized that the ATPase activity contained within the ATPase domain leads to a conformational chane in TerL, which in turn leads to the translocation of DNA. Other high resolution structures of the TerL subunit have been published for T4 terminase [88, 89], which shows a similarly structured ATPase domain, nuclease domain, and linker region. Cryo-EM studies of the large terminase subunit in bacteriophage T7 have been completed, which have been interpreted as support for a conformational change within the large subunit in order to translocate the DNA into the capsid [90]. It will be informative to compare more advanced structural data from lambda terminase to the aforementioned proteins, since lambda terminase has low sequence alignment (**Table 1.1**) and different stoichiometries in comparison to other terminases. XLMS structural studies of full-length lambda terminase are discussed in chapter 4.

	P22	44rr	Sf6	T4	T7
lambda TerS	no significant similarity found	25%	7%	no significant similarity found	no significant similarity found
lambda TerL	no structure available	no structure available	23%	4%	no structure available

Table 1.1. Amino acid sequence alignment by BLAST search of lambda terminase with terminase subunits from other bacteriophages with published crystal structures or EM data. Percentages indicate what percentage of the whole sequence was aligned with a given terminase subunit from one of the other bacteriophages.

More than the lack of structural data for lambda terminase provides motivation for studying terminase. Bacteriophage lambda terminase is a model system for other human dsDNA viruses that undergo a packaging mechanism by a terminase enzyme, such as herpesvirus and adenovirus. Structural insight into lambda terminase may provide valuable structural information necessary to allow for the development of new antiviral drugs to abolish terminase activity and therefore halt the production of new, infectious viruses. It is not only my goal to use XLMS to make progress towards the structural and mechanistic understanding of terminase, but also provide an example and an accessible XLMS workflow for other researchers such that they can pursue XLMS studies on the their own proteins that require structural characterization in the absence of NMR and X-ray crystallography data.

1.3 References for Chapter 1

1. Alberts, B.: The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. *Cell* **92**(3), 291–294 (1998)

2. Alberts, B., Johnson, A., Lewis, J., Raff, M. et al.: Molecular Biology of the Cell (2002)

3. Kerem, B., Rommens, J.M., Buchanan, J. a, Markiewicz, D. et al.: of the Cystic Fibrosis Gene : Identification Genetic Analysis. *Science (80-.).* **245**(1984), 1073–1080 (1989)

4. Pauling, L., Itano, H.: Sickle cell anemia, a molecular disease. *Science (80-.).* **110**(2865), 543–548 (1949)

5. Goh, H., Yao, J., Smith, D.R.: p53 Point Mutation and Survival in Colorectal Cancer Patients and Survival in Colorectal Cancer Patients ' 5217–5221 (1995)

6. Karve, T.M., Cheema, A.K.: Small Changes Huge Impact: The Role of Protein Posttranslational Modifications in Cellular Homeostasis and Disease. *J. Amino Acids* **2011**Article ID: 207691 (2011)

7. Blackman, D.: Amino acid sequence diversity in proteins. *J. Chem. Educ.* **54**170–171 (1977)

8. Taylor, T.C., Backlund, a, Bjorhall, K., Spreitzer, R.J., Andersson, I.: First crystal structure of Rubisco from a green alga, Chlamydomonas reinhardtii. *J. Biol. Chem.* **276**(51), 48159–48164 (2001)

9. Hochleitner, E.O., Kastner, B., Fröhlich, T., Schmidt, A. et al.: Protein stoichiometry of a multiprotein complex, the human spliceosomal U1 small nuclear ribonucleoprotein: Absolute quantification using isotope-coded tags and mass spectrometry. *J. Biol. Chem.* **280**(4), 2536–2542 (2005)

10. Stumpf, M.P.H., Thorne, T., de Silva, E., Stewart, R. et al.: Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U. S. A.* **105**6959–6964 (2008)

11. Wolfenden, R., Snider, M.J.: The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* **34**(12), 938–945 (2001)

12. Davydov, R., Labby, K.J., Chobot, S.E., Lukoyanov, D.A. et al.: Enzymatic and Cryoreduction EPR Studies of the Hydroxylation of Methylated N. *Biochemistry* **53**6511–6519 (2014)

13. Carr, C.M., Kim, P.S.: A spring-loaded mechanism for the conformational change of influenza hemagglutinin. *Cell* **73**823–832 (1993)

14. Walker, J.E., Saraste, M., Runswick, M.J., Gay, N.J.: Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**(8), 945–951 (1982)

15. Miller, J., McLachlan, a. D., Klug, a.: Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *J. Trace Elem. Exp. Med.* **14**(6), 157–169 (2001)

16. Lomenick, B., Hao, R., Jonai, N., Chin, R.M. et al.: Target identification using drug affinity responsive target stability (DARTS). *Proc. Natl. Acad. Sci. U. S. A.* **106**(51), 21984–9 (2009)

17. Bohr, C.: Die Sauerstoffaufnahme des genuinen Blutfarbstoffes und des aus dem Blute dargestellten Hämoglobins. *Zentralblatt PHysiol.* **23**688–690 (1904)

18. Bohr, C., Hasselbalch, K., Krogh, A.: Ueber einen in biologischer Beziehung wichtigen Einfluss, den die Kohlensäurespannung des Blutes auf dessen Sauerstoffbindung übt. *Skand. Arch. Physiol.* **16**(2), 402–412 (1904)

19. Wang, J., Boisvert, D.C.: Structural basis for GroEL-assisted protein folding from the crystal structure of (GroEL-KMgATP)14 at 2.0 ?? resolution. *J. Mol. Biol.* **327**(03), 843–855 (2003)

20. Uhlen, M., Ponten, F.: Antibody-based proteomics for human tissue profiling. *Mol. Cell. Proteomics* **4**384–393 (2005)

21. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G. et al.: A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**662–666 (1958)

22. Perutz, M.F.: Structure of hemoglobin. *Brookhaven Symp. Biol.* **13**165–183 (1960)

23. Williamson, M.P., Havel, T.F., Wüthrich, K.: Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.* **182**295–315 (1985)

24. Yee, A. a., Savchenko, A., Ignachenko, A., Lukin, J. et al.: NMR and X-ray crystallography, complementary tools in structural proteomics of small proteins. *J. Am. Chem. Soc.* **127**(4), 16512–16517 (2005)

25. Sinz, A.: Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *J. Mass Spectrom.* **38**(12), 1225–37 (2003)

26. Robinson, C. V, Sali, A., Baumeister, W.: The molecular sociology of the cell. *Nature* **450**(December), 973–982 (2007)

27. Sali, A., Glaeser, R., Earnest, T., Baumeister, W.: From words to literature in structural proteomics. *Nature* **422**(March), 216–225 (2003)

28. Sharon, M.: How far can we go with structural mass spectrometry of protein complexes? *J. Am. Soc. Mass Spectrom.* **21**(4), 487–500 (2010)

29. Lander, G.C., Evilevitch, A., Jeembaeva, M., Potter, C.S. et al.: Bacteriophage Lambda Stabilization by Auxiliary Protein gpD: Timing, Location, and Mechanism of Attachment Determined by Cryo-EM. *Structure* **16**1399–1406 (2008)

30. Roy, A., Kucukural, A., Zhang, Y.: I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**(4), 725–738 (2010)

31. Yang, J., Yan, R., Roy, A., Xu, D. et al.: The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**(1), 7–8 (2014)

32. Zhang, Y.: I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**40 (2008)

33. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., Baker, D.: Protein structure prediction using Rosetta. *Methods Enzymol.* **383**66–93 (2004)

34. Montelione, G.T., Anderson, S.: Structural genomics: keystone for a Human Proteome Project. *Nat. Struct. Biol.* **6**(1), 11–12 (1999)

35. Fenn, J., Mann, M., Meng, C., Wong, S., Whitehouse, C.: Electrospray ionization for mass spectrometry of large biomolecules. *Science (80-.).* **246**(4926), 64–71 (1989)

36. Tanaka, K., Waki, H., Ido, Y., Akita, S. et al.: Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2**151–153 (1988)

37. Paul, W., Steinwedel, H.: Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift Naturforsch.* **8**(1953)

38. Yost, R.A., Enke, C.G.: Selected ion fragmentation with a tandem quadrupole mass spectrometer. *J. Am. Chem. Soc.* **100**(7), 2274–2275 (1978)

39. Comisarow, M.B., Marshall, A.G.: Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.* **25**(2), 282–283 (1974)

40. Makarov, A.: Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* **72**(6), 1156–1162 (2000)

41. Cameron, A.E., Eggers, D.F.: An Ion ``Velocitron''. *Rev. Sci. Instrum.* **19**(9), 605 (1948)

42. Glish, G.L., Burinsky, D.J.: Hybrid mass spectrometers for tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **19**(2), 161–72 (2008)

43. Sleno, L., Volmer, D. a.: Ion activation methods for tandem mass spectrometry. *J. Mass Spectrom.* **39**1091–1112 (2004)

44. Pressure BioSciences, Inc. Announces the Market Release of its PCT-Enhanced Trypsin Digestion Application for Proteomics (n.d.)

45. Mass Spectrometry Market worth $5.9 Billion by 2018 (n.d.)

46. Aebersold, R., Goodlett, D.R.: Mass Spectrometry in Proteomics. *Chem. Rev.* **101**(2), 269–296 (2001)

47. Domon, B., Aebersold, R.: Mass spectrometry and protein analysis. *Science* **312**(5771), 212–7 (2006)

48. Aebersold, R., Mann, M.: Mass spectrometry-based proteomics. *Nature* **422**(March), 198–207 (2003)

49. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B.: Quantitative mass spectrometry in proteomics: A critical review. *Anal. Bioanal. Chem.* **389**1017–1031 (2007)

50. Hunt, D.F., Yates, J.R., Shabanowitz, J., Winston, S. et al.: Protein sequencing by tandem mass spectrometry **83**(17), 6233–6237 (2007)

51. Olsen, J. V, Ong, S.-E., Mann, M.: Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3**(6), 608–14 (2004)

52. Chait, B.T.: Mass Spectrometry: Bottom-Up or Top-Down. *October* **314**(October), 65–66 (2006)

53. Steen, H., Mann, M.: The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**(9), 699–711 (2004)

54. Yates III, J.R.: Mass spectrometry: from Genomics to Proteomics. *Trends Genet.* **16**(1), 5–8 (2000)

55. Singh, P., Panchaud, A., Goodlett, D.R.: Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. *Anal. Chem.* **82**(7), 2636–42 (2010)

56. Engen, J.R.: Analysis of protein conformation and dynamics by hydrogen/deuterium exchange MS. *Anal. Chem.* **81**(19), 7870–5 (2009)

57. Kipping, M., Schierhorn, A.: Improving hydrogen/deuterium exchange mass spectrometry by reduction of the back-exchange effect. *J. Mass Spectrom.* **38**(3), 271–6 (2003)

58. Gau, B.C., Sharp, J.S., Rempel, D.L., Gross, M.L.: Fast photochemical oxidation of protein footprints faster than protein unfolding. *Anal. Chem.* **81**(16), 6563–71 (2009)

59. Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F. et al.: Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol. Cell. Proteomics* **9**1634–1649 (2010)

60. Maiolica, A., Cittaro, D., Borsotti, D., Sennels, L. et al.: Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol. Cell. Proteomics* **6**(12), 2200–11 (2007)

61. Götze, M., Pettelkau, J., Schaks, S., Bosse, K. et al.: StavroX-A software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom.* **23**(July 2011), 76–87 (2012)

62. Loo, J.A.: Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrom. Rev.* **16**(1), 1–23 (1997)

63. Benesch, J.L., Robinson, C. V.: Mass spectrometry of macromolecular assemblies: preservation and dissociation. *Curr. Opin. Struct. Biol.* **16**245–251 (2006)

64. Heck, A.J.R., van den Heuvel, R.H.H.: Investigation of intact protein complexes by mass spectrometry. *Mass Spectrom. Rev.* **23**(5), 368–389 (2004)

65. Heck, A.J.R.: Native mass spectrometry: a bridge between interactomics and structural biology. *Nat Meth* **5**(11), 927–933 (2008)

66. Scarff, C.A., Thalassinos, K., Hilton, G.R., Scrivens, J.H.: Travelling wave ion mobility mass spectrometry studies of protein structure: biological significance and comparison with X-ray crystallography and nuclear magnetic resonance spectroscopy measurements. *Rapid Commun. Mass Spectrom.* **22**(20), 3297–304 (2008)

67. Van Duijn, E., Simmons, D.A., van den Heuvel, R.H.H., Bakkes, P.J. et al.: Tandem mass spectrometry of intact GroEL-substrate complexes reveals substrate-specific conformational changes in the trans ring. *J. Am. Chem. Soc.* **128**(14), 4694–702 (2006)

68. Uetrecht, C., Versluis, C., Watts, N.R., Roos, W.H. et al.: High-resolution mass spectrometry of viral assemblies: molecular composition and stability of dimorphic hepatitis B virus capsids. *Proc. Natl. Acad. Sci. U. S. A.* **105**(27), 9216–20 (2008)

69. Ruotolo, B.T., Benesch, J.L.P., Sandercock, A.M., Hyung, S.-J., Robinson, C. V: Ion mobility-mass spectrometry analysis of large protein complexes. *Nat. Protoc.* **3**(7), 1139–52 (2008)

70. Casjens, S.R.: The DNA-packaging nanomotor of tailed bacteriophages. *Nat. Rev. Microbiol.* **9**(9), 647–57 (2011)

71. Feiss, M., Catalano, C.E.: Bacteriophage lambda terminase and the mechanism of viral DNA packaging. *Viral Genome Packag. Mach. Genet. Struct. Mech.* 5–39 (2005)

72. Fuller, D.N., Raymer, D.M., Rickgauer, J.P., Robertson, R.M. et al.: Measurements of single DNA molecule packaging dynamics in bacteriophage lambda reveal high forces, high motor processivity, and capsid transformations. *J. Mol. Biol.* **373**(5), 1113–22 (2007)

73. Catalano, C.E.: The terminase enzyme from bacteriophage lambda: A DNA-packaging machine. *Cell. Mol. Life Sci.* **57**128–148 (2000)

74. Murialdo, H.: Bacteriophage Lambda DNA Maturation and Packaging (2003)

75. Maluf, N.K., Gaussier, H., Bogner, E., Feiss, M., Catalano, C.E.: Assembly of bacteriophage lambda terminase into a viral DNA maturation and packaging machine. *Biochemistry* **45**15259–15268 (2006)

76. Andrews, B.T., Catalano, C.E.: Strong subunit coordination drives a powerful viral DNA packaging motor. *Proc. Natl. Acad. Sci. U. S. A.* **110**(15), 5909–14 (2013)

77. Andrews, B.T., Catalano, C.E.: The enzymology of a viral genome packaging motor is influenced by the assembly state of the motor subunits. *Biochemistry* **51**9342–9353 (2012)

78. Meisenheimer, K.M., Koch, T.H.: Photocross-linking of nucleic acids to associated proteins. *Crit. Rev. Biochem. Mol. Biol.* **32**101–140 (1997)

79. Babbar, B.K., Gold, M.: ATP-reactive sites in the bacteriophage lambda packaging protein terminase lie in the N-termini of its subunits, gpA and gpNu1. *Virology* **247**(2), 251–64 (1998)

80. Hwang, Y., Catalano, C.E., Feiss, M.: Kinetic and mutational dissection of the two ATPase activities of terminase, the DNA packaging enzyme of bacteriophage Chi. *Biochemistry* **35**(8), 2796–803 (1996)

81. De Beer, T., Fang, J., Ortega, M., Yang, Q. et al.: Insights into Specific DNA Recognition during the Assembly of a Viral Genome Packaging Machine. *Mol. Cell* **9**(5), 981–991 (2002)

82. Yang, Q., de Beer, T., Woods, L., Meyer, J.D. et al.: Cloning, expression, and characterization of a DNA binding domain of gpNu1, a phage lambda DNA packaging protein. *Biochemistry* **38**(1), 465–77 (1999)

83. Ortega, M.E., Catalano, C.E.: Bacteriophage Lambda gpNu1 and Escherichia coli IHF Proteins Cooperatively Bind and Bend Viral DNA : Implications for the Assembly of a Genome-Packaging 5180–5189 (2006)

84. Zhao, H., Kamau, Y.N., Christensen, T.E., Tang, L.: Structural and functional studies of the phage Sf6 terminase small subunit reveal a DNA-spooling device facilitated by structural plasticity. *J. Mol. Biol.* **423**(3), 413–426 (2012)

85. Sun, S., Gao, S., Kondabagil, K., Xiang, Y. et al.: Structure and function of the small terminase component of the DNA packaging machine in T4-like bacteriophages. *Proc. Natl. Acad. Sci.* **109**(3), 817–822 (2012)

86. Němeček, D., Lander, G.C., Johnson, J.E., Casjens, S.R., Thomas, G.J.: Assembly Architecture and DNA Binding of the Bacteriophage P22 Terminase Small Subunit. *J. Mol. Biol.* **383**494–501 (2008)

87. Zhao, H., Christensen, T.E., Kamau, Y.N., Tang, L.: Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. *Proc. Natl. Acad. Sci. U. S. A.* **110**(20), 8075–80 (2013)

88. Sun, S., Kondabagil, K., Draper, B., Alam, T.I. et al.: The Structure of the Phage T4 DNA Packaging Motor Suggests a Mechanism Dependent on Electrostatic Forces. *Cell* **135**(7), 1251–1262 (2008)

89. Sun, S., Kondabagil, K., Gentz, P.M., Rossmann, M.G., Rao, V.B.: The Structure of the ATPase that Powers DNA Packaging into Bacteriophage T4 Procapsids. *Mol. Cell* **25**943–949 (2007)

90. Daudén, M.I., Martiń-Benito, J., Sánchez-Ferrero, J.C., Pulido-Cid, M. et al.: Large terminase conformational change induced by connector binding in bacteriophage T7. *J. Biol. Chem.* **288**16998–17007 (2013)

Chapter 2

Novel Digital Microfluidics Combined with SAWN for HDX Measurements

2.1 Motivation and Background for a new HDX Workflow

Hydrogen deuterium exchange (HDX) is a powerful technique for studying protein structure [1]. It is widely used to probe changes in secondary protein structure, protein-ligand interactions, structural differences between protein isoforms and mutants, and protein conformational changes [1]. As evidenced by the rapidly growing body of HDX literature, HDX has seen steady growth and utility. For the most part, HDX workflows have deviated little from HPLC or a direct infusion apparatus coupled to ESI-MS. In this study, I investigate the use of two fundamental changes to the HDX workflow. First, I examine the use of an alternative ionization method called surface acoustic wave nebulization (SAWN) and second, I explore the use of microfluidics as an alternative to manual or an automated LC-type unit for sample preparation to enhance the ease and speed at which HDX experiments can be performed.

While ESI has been critical to HDX success and many other MS based assays [2, 3], it is not without limitations. For instance, while ESI is highly sensitive, it can lead to in-source fragmentation and/or the oxidation of small molecules and proteins [4–6], and ESI requires its own charged, continuous flow apparatus. To address the analytical short-comings of ESI, many alternative ionization techniques have been developed, such as Desorption Ionization on Silicon [7], Desorption-ESI [8] and Laser Ablation-ESI [9].

In this study, I employed surface acoustic wave nebulization (SAWN) for HDX analysis. SAWN generates ions from a planar piezoelectric surface and delivers them to the inlet of a mass spectrometer [10]. To accomplish this, an alternating current is applied to an interdigitated transducer (IDT) on a piezoelectric $LiNbO_3$ wafer to generate a high frequency surface acoustic

wave [11]. When the SAW reaches the droplet on the surface of the chip, the wave energizes the droplet and leads to the nebulization of the liquid sample within seconds. To date, a range of analytes of small molecules have been analyzed by SAWN-MS [5, 6, 10].

SAWN has several advantages over ESI and other ionization techniques for HDX. First, SAWN has been found to generate ions of lower energy than ESI [6], which has the potential to maintain the structural integrity of more analytes during the ionization process [5]. Additionally, given its relative "softness" compared to ESI, there is the potential advantage of reduced back exchange during HDX, which I test here. Second, in contrast to ESI, SAWN is very simple to operate and its planar design obviates the need capillary tubing altogether. Specifically, SAWN operation only involves transferring a droplet of sample directly onto the chip and activation of the chip, which leads to the immediate nebulization of the sample. In this chapter, I validated the use of HDX via SAWN-MS using ubiquitin, a well-characterized protein [12–14].

In addition to the ionization source, I explored the use of a sophisticated fluid handling platform called digital microfluidics (DMF) for sample manipulation within the HDX workflow. The concept of DMF is representative of a class of techniques using the relatively weak interactions between electric fields and polarizable droplets of liquid on a planar surface. A typical DMF device consists of a flat substrate covered in a microelectrode pattern which is covered by subsequent layers of a suitable dielectric and a hydrophobic layer [15]. This technology has been referred to by many names during its development including; metal-insulator-solution-transport [16], digital microfluidic system [17], electrowetting on insulator coated electrodes [18], electrowetting on dielectric [19] and finally DMF [20].

Due to the promise of automated sample handling, small sample volumes, lossless sample preparation, and miniaturized devices that can forgo large and costly LC systems, much effort

has been placed into connecting microfluidics with ESI-MS as described in prior publications [21, 22]. However, coupling ESI-MS to microfluidic systems remains a challenge due to the requirement of maintaining constant fluid flow for steady electrospray [23]. The most popular means for coupling microfluidics with electrospray is simply attaching a micro-electrospray emitter [21], such as the device manufactured by Rob *et al*. that integrated microfluidics with HDX-MS [24]. However, this approach leads to additional fabrication challenges, increased adsorption of the sample onto the capillary walls, and potential for clogs. Microfluidics has been coupled with MALDI-MS to circumvent these problems, but MALDI is susceptible to matrix effects [25]. Alternatively, integrating SAWN and DMF is quite simple and straightforward since both are planar platforms which manipulate small droplets [20].

Here, I demonstrate the ease of use of DMF-SAWN to perform HDX on an inexpensive DMF device. Briefly, the DMF device was made by screen-printing conductive ink on top of a flexible polyimide substrate and coating the device with hydrophobic materials. The motivation for utilizing an inexpensive DMF design was to obviate the problem of biofouling by making the device disposable. I report results obtained from fusing droplets of D_2O and angiotensin II on such a disposable DMF device and analyzing the sample immediately by SAWN-MS.

The significance of this report is three-fold. Firstly, it represents the first use of SAWN to ionize HDX samples for MS analysis. Secondly, I demonstrated that SAWN-MS can yield high-resolution, reproducible data for whole proteins with relatively low back exchange. Thirdly, SAWN was successfully coupled with a disposable, screen-printed DMF device to provide a means for sample preparation at the mass spectrometer. This combination points the way toward monitoring more complex chemical reactions in real time by DMF-SAWN-MS with many different types of analytes for a new “lab on a chip” or micro total analysis system [22, 26].

2.2 Materials and Methods

Fabrication and Operation of SAWN Chips. Fabrication of SAWN chips has been reported in detail previously and was done by Scott Heron and co-workers [6, 10]. In summary, a SAW transducer consisting of 20 pairs of 100 μm interdigitated (IDT) electrodes (40 in total) with 100 μm spacing and 10 mm aperture, along with a secondary electrode to apply external electrical potential, were patterned onto the surface of 128 Y-cut X-propagating 3 inch $LiNbO_3$ wafers purchased from Crystal Technology, Inc. (Palo Alto, Ca). The SAWN configuration was first designed in AutoCAD before being written into a chrome mask by a Heidelberg μPG 101 Laser Pattern Generator (Heidelberg Instruments Mikrotechnik GmbH Tullastrasse 2, D-69126 Heidelberg, Germany) at the University of Washington Nanotech User Facility (https://depts.washington.edu/ntuf/). Wafers were then coated using AZ 1512 positive photoresist (AZ Electronic Materials, Somerville, NJ) spun at 4000 rpm for 30 seconds creating a 1-1.2 μm thick resist layer. Exposure of the photo resist was done for 5 seconds using an Oriel mask aligner (Newport Corporation, CA). The exposed wafers were placed in a development bath for 60 seconds in AZ 351 (AZ Electronic Materials, Somerville, NJ). To pattern conductive IDT electrodes a 20 nm chrome adhesion layer was deposited by heated vapor deposition followed by a 60 nm layer of gold, followed by lift-off in acetone for 30 minutes. The resulting SAWN IDT has a resonance frequency of 9.56 MHz. To operate the SAWN chip a MXG analog signal generator (Agilent N5181A, Santa Clara, CA) and a Mini Circuits ZHL-5W-1, 5-500 MHz RF amplifier (GwInstek GPS-2303, New York, NY) were used to generate and amplify the RF signal.

Fabrication of Screen-Printed DMF Chips. Flexible, disposable, DMF chips were produced in a proprietary low-cost printing process using carbon-containing conductive ink on 50 µm thick flexible polyimide foils (DuPont Kapton) to pattern electrodes with 100 µm spacing, connection leads and external contact pads, followed by a 7 µm layer of ink acting as dielectric layer. After the printing process, a fluoropolymer solution, either Teflon-AF (DuPont) or Cytop (Asahi Glass) solution was spin-coated onto the devices and left to dry in ambient condition to create a 150 nm thick hydrophobic layer. The finished devices could be stored in normal laboratory conditions over the time period of several months without influence on performance. The devices used in this work were developed, manufactured and provided by Dr. Patrick R. R. Langridge-Smith of the University of Edinburgh, Edinburgh, UK.

Operation of Screen-Printed DMF Chips. Droplets were moved, merged and transferred to the SAWN chip in AC mode [26, 27] with a driving voltage of 500 Vpp and a frequency of 20 kHz. The devices performed flawlessly and without any noticeable degradation over several experimental cycles. The electrode geometry used was an arrangement of two parallel columns of square electrodes. The droplet was held between two adjacent electrodes from the respective columns by applying the driving voltage between them. To move the droplet to the next electrode pair, the driving voltage was first applied to the new pair, pulling the droplet in between the pairs, and then switching off the voltage between the first pair. DMF operation was controlled using control software running on a PC that was connected to custom drive electronics. For this, the 2.5 Vpp AC sine output of a 20 kHz signal generator was amplified by a high-voltage amplifier (Trek) to 500 Vpp and distributed to the DMF electrode pads using a distribution circuit based on solid-state relays. Connection to the DMF devices was through 2.54

mm pitch edge connectors. The work was done by Scott Edgar and Daniel Winters (University of Edinburgh).

HDX of Ubiquitin. In an eppendorf tube, 1µl of 200 µM ubiquitin in 40 mM ammonium bicarbonate (pH 7.4) buffer were mixed with 19 µl of D_2O to afford a final concentration of 10 µM ubiquitin, 2mM ammonium bicarbonate. Exposure to deuterium was varied so as to obtain many time points to properly visualize the rate of deuterium uptake in ubiquitin over time. To quench the reaction, 1.5 µl of 10% formic acid was added to the mixture. Three replicate spectra were obtained for each time point. To obtain a mass spectrum of the mixture, 1 µl of the quenched reaction mixture was placed on the SAWN chip surface in front of the interdigitated transducer. An ultrazoom scan of this droplet was recorded with a Thermo LTQ-Velos (Bremen, Germany) using the s-lens at 60% voltage and a capillary temperature of 200C after activating the SAWN chip with 25 dBm at 9.56 Mhz. To calculate the overall percent deuteration, the percent deuteration of the +6, +7, and +8 ions was calculated as per the equation 1 reported in Zhang and Smith [28] with the centroided mass obtained using HX express [29, 30]. These values were averaged for the final percent deuteration value. Estimated percent deuteration was achieved by using the method described in Pan *et al.* [13] and the rate constants from Bougalt *et al.* [12] adjusted for pH differences. This was done by Lucas Monkkonen.

Measuring Back Exchange of Ubiquitin. As before, 1µl of stock solution was pipetted into 19 µl of D_2O. The reaction mixture was boiled for 10 minutes and then dried by speed vac. Then, 19 µl of D_2O and 1.5 µl of 10% formic acid was added to the dried sample, mixed briefly, and then pipetted onto the SAWN chip for analysis by SAWN. First, in order to calculate the percent back exchange, the number of backbone hydrogenations was calculated. This was done by subtracting the measured molecular weight by the mass of undeuterated ubiquitin and the

number of fast exchanges in ubiquitin (73). This quantity was divided by the total theoretical number of backbone deuterations that can occur in 95% D_2O (63.8) to obtain the final percent back exchange. This was done by Lucas Monkkonen.

HDX Mass Spectrometry of Angiotensin. 4 μL droplet of angiotensin II (10μM, Proteomass) in H_2O and a 4 μL droplet of D_2O were pipetted onto separate regions of the flexible DMF chip. Then, an AC field to the electrodes adjacent to the droplets was applied to transport and fuse the two droplets. The fused droplets were transferred directly from the hydrophobic DMF chip to the hydrophilic SAWN chip by DMF and wetting. After transfer, SAWN was activated to aerosolize the droplet and HDX was monitored by a Thermo LTQ linear ion trap. This was done as a collaborative effort with Daniel Winters, Scott Edgar, Lucas Monkkonen, and Christophe Masselon (CEA Grenoble, France).

2.3 Results and Discussion

HDX of Ubiquitin.

To demonstrate the utility of SAWN for HDX analysis of proteins, global HDX studies were carried out on ubiquitin, a well characterized protein. High-resolution mass spectra showing the isotopic envelope of deuterated ubiquitin were readily obtained by SAWN-MS on an ion trap mass spectrometer, and the mass spectra clearly demonstrate a shift of the ubiquitin ions towards higher m/z values with increased exposure to D_2O (**Figure 2.1**). The maximum deuteration achieved was within 5 exchanges of the theoretical maximum ubiquitin backbone deuteration

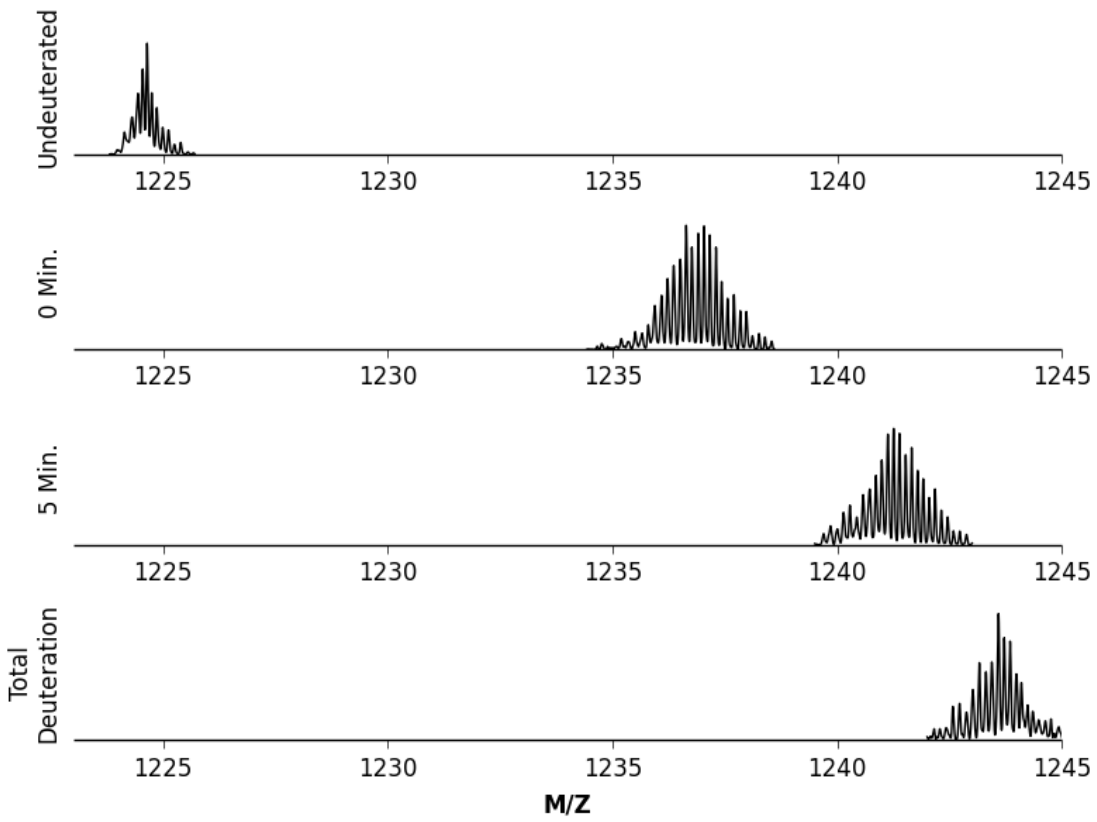


Figure 2.1. SAWN-MS Monitored HDX of Ubiquitin (+7 charge state). The time zero mass spectrum was obtained from ubiquitin exposed to quench solution. The five minute mass spectrum was obtained from ubiquitin exposed to D_2O for five minutes. The per-deuterated mass spectrum was obtained by boiling in D_2O for 10 minutes. All spectra were collected by an LTQ-Velos operating in ultrazoom scan mode.

Due to these promising results, I set out to record a full time course of the HDX reaction by SAWN to show that this approach can be used to rigorously measure global HDX rates. These results are displayed in **Figure 2.2**. Each data point represents triplicate measurements that highlight the reproducibility of this technique. Additionally, the experimental per-deuterated time-point of 94.6% is very close to the maximum possible 95% deuteration of ubiquitin under

these conditions. Most importantly, however, the data from SAWN-MS correlates well with global estimated percent deuteration values based on HDX-NMR data obtained by Bougault, especially for the later timepoints [12] (**Figure 2.2**).

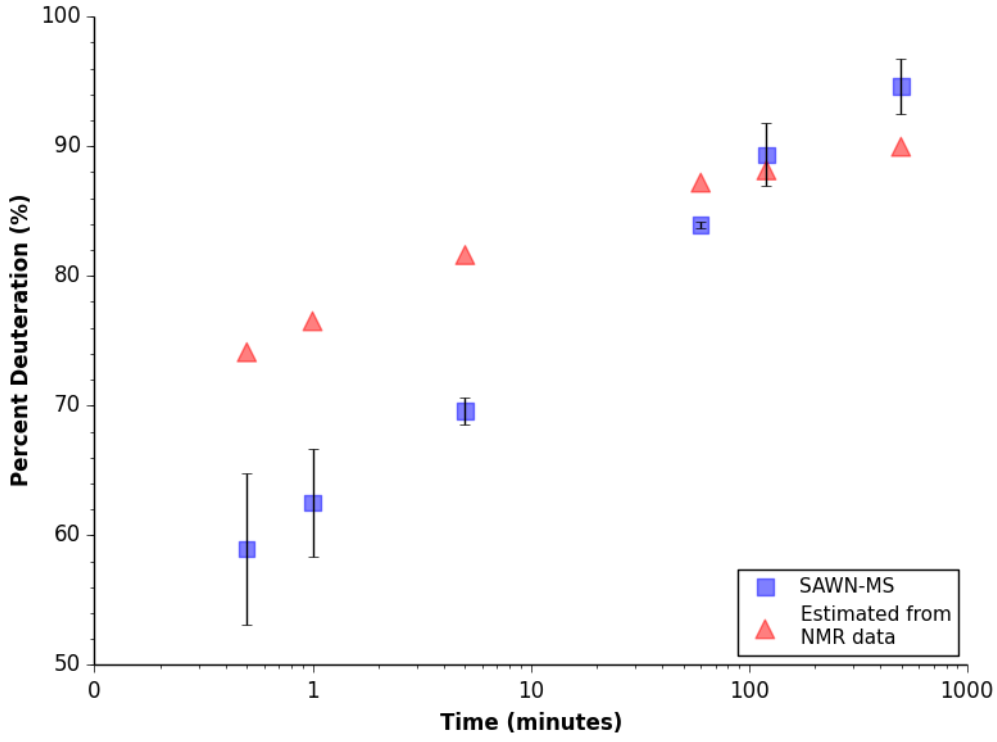


Figure 2.2. Time-course of Deuterium Uptake in Ubiquitin. Time refers to the incubation time of ubiquitin in 95% D_2O at room temperature. The blue squares represent the experimentally derived percent deuteration. This was obtained by calculating the percent deuteration for each of the +6, +7, and +8 charge state ions with the aid of equation #1 [28] and HX-express [29, 30] and averaging these values. Error bars represent the standard deviation for the average of the three replicates of the percent deuteration levels found with each of the three charge states. The red triangles represent the estimated global percent deuteration based on combining NMR kinetic data of ubiqutin [12] with the equation by Pan *et al.* [13].

An additional point of validation is that my SAWN-MS data compare well to the NMR studies of Pan *et al*. [13] who measured a global percent deuteration of 83% at 30 minutes of incubation for ubiquitin under similar conditions. When a simple natural log regression is fitted to the SAWN-MS data (r-squared value 0.9957), the estimated percent deuteration measured by SAWN-MS at 30 minutes incubation would be 80.4%, which agrees well with the published NMR data (**Figure 2.2**) [13]. Finally, the back exchange of the SAWN-HDX setup was measured. At room temperature and relative high capillary inlet temperatures (300ºC), 50% back exchange was measured (**Figure 2.3**). While this is somewhat high, it is comparable to some back exchange results using MALDI [31].

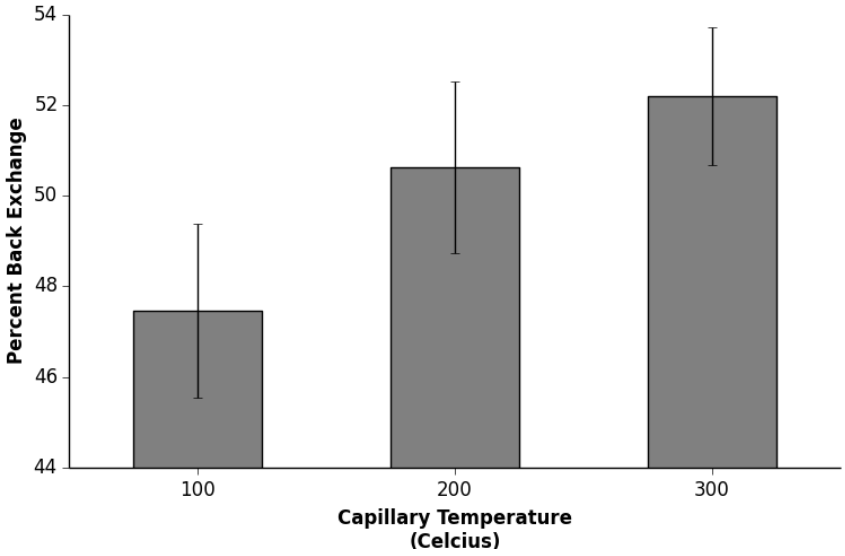
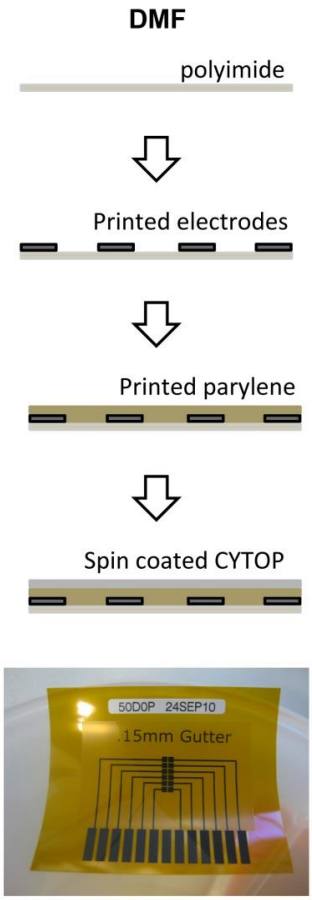


Figure 2.3. SAWN-MS Measured Back Exchange. Percent back exchange obtained from SAWN-MS HDX at room temperature with various inlet capillary temperatures of ubiquitin. Back exchange values for each temperature are based on the average of three different samples.

DMF Chip Fabrication.

Though SAWN chips are simple to operate they remain rudimentary in their ability to accommodate fluid handling or chromatography. To address this issue we developed a technique for fabrication of low-cost screen-printed DMF chips on polyimide that can easily be coupled with SAWN and may be considered disposable (**Figure 2.4**). These inexpensive DMF chips were prepared by screen-printing [32, 33] on to flexible polyimide substrates. Specifically, screen-printed electrodes were developed using a carbon-containing conductive ink on top of which was printed a dielectric with a layer of hydrophobic Cytop on top of these first two layers.

Figure 2.4. Fabrication of Digital Microfluidic (DMF) chips. DMF chips were fabricated by screening printing conductive ink onto 50 □m thick flexible polyimide foils. The DMF chips used herein were gifts from Dr. Patrick R. R. Langridge-Smith of the University of Edinburgh, Edinburgh, UK.



HDX via DMF-SAWN-MS.

HDX was carried out on the signal peptide angiotensin II to demonstrate the functionality of DMF-SAWN-HDX. The movement of droplets on the DMF chip and transfer to SAWN is shown in **Figure 2.5**. The design of the DMF chip used a series of electrodes spaced 100 μm apart (**Figure 2.5A**). Samples were pipetted onto the surface of the DMF chip across from an equal volume of D_2O as shown in **Figure 2.5B**. Once both droplets were in place on the DMF chip, the sample droplet consisting of a 4 μl droplet of angiotensin II in H_2O was actuated towards the 4 μl droplet of D_2O by applying a field on the adjacent electrode while simultaneously releasing the field on the initial electrode. In this way, one can actuate droplets back and forth across the DMF chip and combine droplets to initiate HDX as shown in **Figure 2.5C**. Transfer of the post-HDX reaction sample to the SAWN chip was performed by placing the droplet in contact with the surface of the SAWN wafer, which allowed the droplet to wick easily over to the natively hydrophilic lithium niobate SAWN chip from the more hydrophobic coated Cytop DMF chip **(Figure 2.5D-E**).

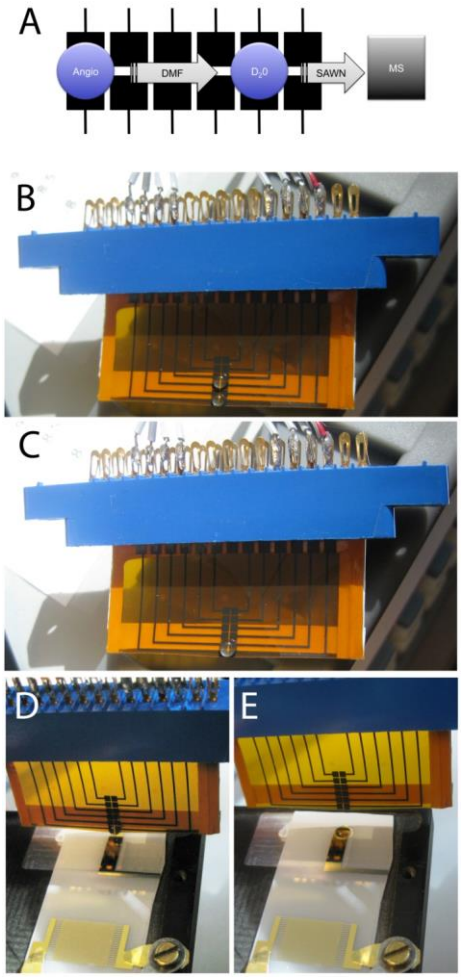


Figure 2.5. DMF Experimental Workflow. **Panel A**: cartoon of HDX workflow. **Panel B**: picture before droplets were merged on the DMF chip. **Panel C**: picture showing merged droplets after applying an AC field to adjacent electrodes. **Panel D**: picture of merged droplet prior to transfer to SAWN chip. **Panel E**: picture of droplet after being transferred from hydrophobic DMF chip to the natively hydrophilic lithium niobate SAWN wafer by wicking action.

The mass spectra of angiotensin II before (**Figure 2.6A**) and after (**Figure 2.6B**) a one minute HDX are shown in **Figure 2.6**. In addition to the expected $[M+H]^+$ ion of angiotensin II, the presence of a small amount of sodiated angiotensin II, $[M+Na]^+$, is observed. The post-HDX mass spectrum is shown in **Figure 2.6B** which shows that eight deuteriums were found to be present on angiotensin II, which is the maximum possible deuteration for angiotensin II in 50% D_2O. These results are expected for a small peptide like angiotensin II (1.046 KDa), since all the sites are highly solvent accessible.

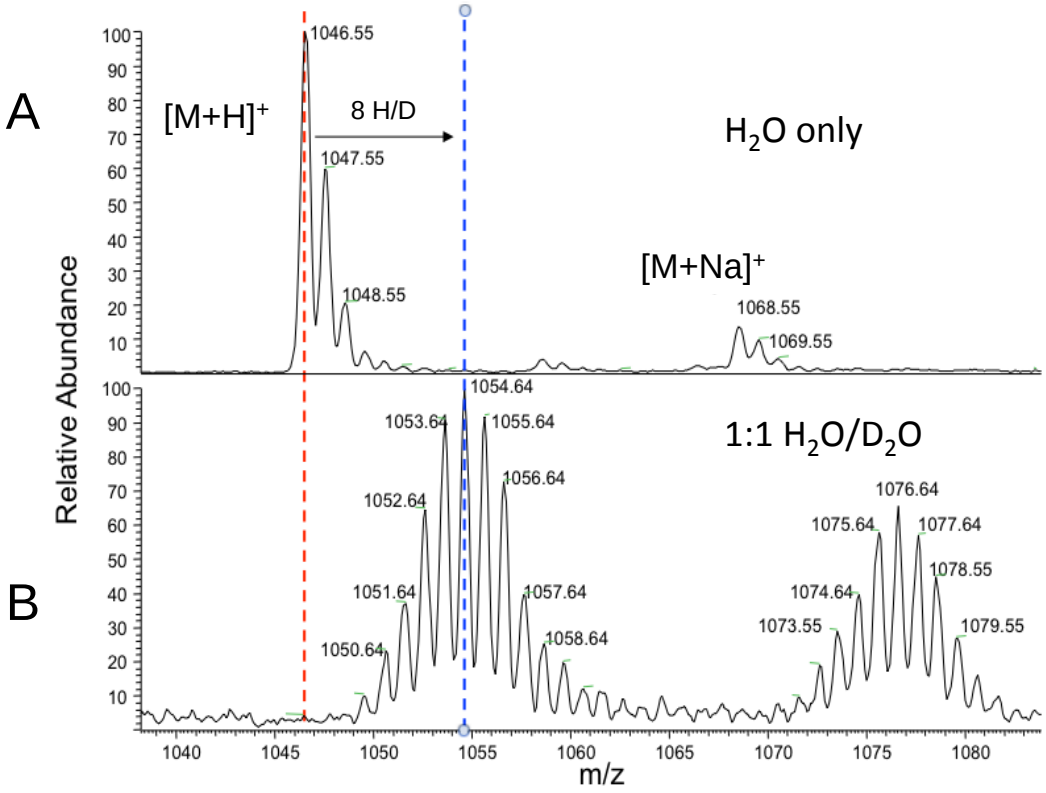


Figure 2.6. HDX of angiotensin II Measured by DMF-SAWN-MS. Panel A shows an angiotensin II mass spectrum is shown prior to HDX. **Panel B** shows one minute post- mixing by DMF manipulation of a droplet of angiotensin in water with a droplet of D_2O. One minute of exposure of angiotensin II in 50% D_2O resulted in eight deuterium exchanges.

2.4 Conclusions

In summary, I have demonstrated the use of SAWN-MS for monitoring HDX reactions of small proteins. Additionally, and importantly for future automation of the method, I have shown it is feasible to combine a DMF sample preparation chip with the SAWN chip to carry out HDX monitoring by MS. The DMF chips were produced by screen-printing onto a flexible, polyimide substrate making the chip effectively disposable. These HDX methods using SAWN-MS and DMF-SAWN-MS offer a new, facile means of elucidating protein structure. Given the simplicity of these methods, I envision them being used to carry out rapid pilot studies at the MS interface.

2.5 References for Chapter 2

1. Engen, J.R.: Analysis of protein conformation and dynamics by hydrogen/deuterium exchange MS. *Anal. Chem.* **81**(19), 7870–5 (2009)

2. Fenn, J., Mann, M., Meng, C., Wong, S., Whitehouse, C.: Electrospray ionization for mass spectrometry of large biomolecules. *Science (80-.).* **246**(4926), 64–71 (1989)

3. Whitehouse, C.M., Dreyer, R.N., Yamashita, M., Fenn, J.B.: Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.* **57**(3), 675–679 (1985)

4. Boys, B.L., Kuprowski, M.C., Noël, J.J., Konermann, L.: Protein oxidative modifications during electrospray ionization: solution phase electrochemistry or corona discharge-induced radical attack? *Anal. Chem.* **81**(10), 4027–34 (2009)

5. Yoon, S.H., Huang, Y., Edgar, J.S., Ting, Y.S. et al.: Surface acoustic wave nebulization facilitating lipid mass spectrometric analysis. *Anal. Chem.* **84**(15), 6530–7 (2012)

6. Huang, Y., Yoon, S.H., Heron, S.R., Masselon, C.D. et al.: Surface acoustic wave nebulization produces ions with lower internal energy than electrospray ionization. *J. Am. Soc. Mass Spectrom.* **23**(6), 1062–70 (2012)

7. Wei, J., Buriak, J., Siuzdak, G.: Desorption–ionization mass spectrometry on porous silicon. *Nature* **399**(May), 243–246 (1999)

8. Takáts, Z., Wiseman, J.M., Gologan, B., Cooks, R.G.: Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science* **306**(5695), 471–3 (2004)

9. Nemes, P., Vertes, A.: Laser ablation electrospray ionization for atmospheric pressure, in vivo, and imaging mass spectrometry. *Anal. Chem.* **79**(21), 8098–106 (2007)

10. Heron, S.R., Wilson, R., Shaffer, S.A., Goodlett, D.R., Cooper, J.M.: Surface acoustic wave nebulization of peptides as a microfluidic interface for mass spectrometry. *Anal. Chem.* **82**(10), 3985–9 (2010)

11. Tanaka, Y., Tamura, S.: Surface acoustic waves in two-dimensional periodic elastic structures. *Phys. Rev. B* **58**(12), 7958–7965 (1998)

12. Bougault, C., Feng, L., Glushka, J., Kupce, E., Prestegard, J.H.: Quantitation of rapid proton-deuteron amide exchange using hadamard spectroscopy. *J. Biomol. NMR* **28**385–390 (2004)

13. Pan, J., Han, J., Borchers, C.H., Konermann, L.: Electron Capture Dissociation of Electrosprayed Protein Ions for Spatially Resolved Hydrogen Exchange Measurements (Figure 1), 11574–11575 (2008)

14. Pan, Y., Briggs, M.: Hydrogen exchange in native and alcohol forms of ubiquitin. *Biochemistry* **31**(46), 11405–11412 (1992)

15. Abdelgawad, M., Wheeler, A.R.: The Digital Revolution: A New Paradigm for Microfluidics. *Adv. Mater.* **21**(8), 920–925 (2009)

16. Fair, R.B., Pollack, M.G., Woo, R., Pamula, V.K. et al.: A micro-watt metal-insulator-solution-transport (MIST) device for scalable digital bio-microfluidic systems. *Int. Electron Devices Meet. Tech. Dig.* 16.4.1–16.4.4 (2001)

17. Bohringer, K.F., Boehringer, K.F.: in:, *IEEE Int. Conf. Robot. Autom. 2004. Proceedings. ICRA '04. 2004*, vol. 2, IEEE, 2004, pp. 1468–1474 Vol.2

18. Quilliet, C., Berge, B.: Electrowetting: a recent outbreak. *Curr. Opin. Colloid Interface Sci.* **6**(1), 34–39 (2001)

19. Li, Y., Fu, R., Winters, D.: Test Structures for Characterizing the Integration of EWOD and SAW Technologies for Microfluidics. *IEEE Trans. Semicond. Manuf.* **25**(3), 323–330 (2012)

20. Kirby, A., Wheeler, A.: Digital microfluidics: an emerging sample preparation platform for mass spectrometry. *Anal. Chem.* (2013)

21. Koster, S., Verpoorte, E.: A decade of microfluidic analysis coupled with electrospray mass spectrometry: an overview. *Lab Chip* **7**(11), 1394–412 (2007)

22. Oedit, A., Vulto, P., Ramautar, R., Lindenburg, P.W., Hankemeier, T.: Lab-on-a-Chip hyphenation with mass spectrometry: strategies for bioanalytical applications. *Curr. Opin. Biotechnol.* **31**79–85 (2015)

23. Limbach, P. a., Meng, Z.: Integrating micromachined devices with modern mass spectrometry. *Analyst* **127**(6), 693–700 (2002)

24. Rob, T., Gill, P.K., Golemi-Kotra, D., Wilson, D.J.: An electrospray ms-coupled microfluidic device for sub-second hydrogen/deuterium exchange pulse-labelling reveals allosteric effects in enzyme inhibition. *Lab Chip* **13**(13), 2528–32 (2013)

25. Cohen, S.L., Chait, B.T.: Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins. *Anal. Chem.* **68**(1), 31–7 (1996)

26. Cho, S.K., Moon, H., Kim, C.: Creating, transporting, cutting, and merging liquid droplets by electrowetting-based actuation for digital microfluidic circuits. *J. Microelectromechanical Syst.* **12**(1), 70–80 (2003)

27. Choi, K., Ng, A.H.C., Fobel, R., Wheeler, A.R.: Digital microfluidics. *Annu. Rev. Anal. Chem.* **5**413–40 (2012)

28. Zhang, Z., Smith, D.L.: Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. *Protein Sci.* **2**(4), 522–31 (1993)

29. Guttman, M., Weis, D.D., Engen, J.R., Lee, K.K.: Analysis of overlapped and noisy hydrogen/deuterium exchange mass spectra. *J. Am. Soc. Mass Spectrom.* **24**(12), 1906–12 (2013)

30. Weis, D.D., Engen, J.R., Kass, I.J.: Semi-automated data processing of hydrogen exchange mass spectra using HX-Express. *J. Am. Soc. Mass Spectrom.* **17**(12), 1700–3 (2006)

31. Kipping, M., Schierhorn, A.: Improving hydrogen/deuterium exchange mass spectrometry by reduction of the back-exchange effect. *J. Mass Spectrom.* **38**(3), 271–6 (2003)

32. Tudorache, M., Bala, C.: Biosensors based on screen-printing technology, and their applications in environmental and food analysis. *Anal. Bioanal. Chem.* **388**(3), 565–78 (2007)

33. Krebs, F.C., Jørgensen, M., Norrman, K., Hagemann, O. et al.: A complete process for production of flexible large area polymer solar cells entirely using screen printing—First public demonstration. *Sol. Energy Mater. Sol. Cells* **93**(4), 422–441 (2009)

Chapter 3

Use of Surface Acoustic Wave Nebulization to Probe Native Protein Structure

3.1 Ionization Methods for Probing Noncovalent Interactions

Noncovalent interactions are crucial to all biological processes. Despite the vast diversity in biological proccesses that occur in different cell types, tissues, organisms, etc., these pathways share a common operational principle; they all rely on the noncovalent interactions between biomolecules. Thus, it is of interest to find suitable methods to study these interactions in order to understand biological processes occuring within virus particles, living cells, and disease pathways [1–3].

Characterizing noncovalent protein-protein, protein-DNA, and protein-small molecule interactions is difficult since they are often transient, have weak binding affinities, and can be hard to isolate at concentrations sufficient to characterize. While many techniques such as small angle x-ray scattering, analytical ultracentrifugation, size-exclusion chromatography, fluorescence studies, NMR, and x-ray crystallography have been used successfully to study protein complexes [4–9], they have limitations that include lack of sensitivity and/or longer experimental times. Electrospray ionization-mass spectrometry (ESI-MS) has emerged as a powerful technique to analyze native protein complexes in part because it overcomes the need for large quantities of protein complexes for characterization and routinely requires protein concentrations only in the nano to micromolar range. ESI-MS is also rapid, versatile, and can simultaneously measure several species in a mixture [10].

There is a sizeable body of literature that demonstrates the utility of ESI-MS experiments for studying noncovalent complexes. The use of MS-based structural studies of intact protein complexes by ESI under semi-native buffer conditions is now commonly referred to as native MS. Such studies may include the use of techniques like hydrogen/deuterium exchange (HDX) to probe the structure and stoichiometry of protein complexes [11–17]. While ESI is the gold standard for native MS, ESI has known limitations. For example, the often fragile protein complex is exposed to commonly used voltages (~1.5-4kV), which present the potential for significant structural perturbation and complex dissociation. Additionally, ESI can lead to the oxidation of the solution components [18] and acidification of the solution [19]. Konermann *et al.* (2001) published further studies showing that ESI can cause a significant amount of protein unfolding due to the electrolysis of water, which indicates that ESI could destroy the structural integrity of noncovalent protein complexes [20]. Finally, ESI can also lead to the formation of nonspecific protein complexes that are an artifact of ionization and that do not relfect in-solution interactions [38].

Very few ionization techniques are able to simultaneously ionize proteins in near native solution conditions and significantly minimize the exposure of the protein sample to high voltage. To my knowledge, desorption electrospray ionization (DESI) and laser induced liquid bead ion desorption (LILBID) are the only ionization methods that satisfy these two conditions [21, 22]. In the DESI experiments, protein samples are sprayed through an uncharged capillary and collides with a separate spray of charged microdroplets emitted by a high voltage DESI probe. Doing so allowed Liu *et al*. and Ferguson *et al*. to directly monitor intact, noncovalent complexes such as RNase A-cytidine nucleotide complexes, lysozyme-N-acetylglucosamine complexes, superoxide dismutase holoenzyme, enolase holoenzyme, and hemoglobin

holoenzyme [21, 23]. In contrast, LILBID involves loading a protein sample into a piezoelectric droplet emitter and irradiating the droplets with an IR laser tuned to the absorption of water molecules to desolvate and ionize the proteins. This method has successfully detected very large and fragile protein complexes such as bacterial cytochrome bc1 and cytochrome c oxidase [24], several large mitochondrial membrane protein complexes [25], and even DNA dodecamers [26].

Another recently developed ambient ionization technique, surface acoustic wave nebulization (SAWN), has been shown to produce ions of lower energy than ESI [27]. This finding could be beneficial to native MS since fragile noncovalent interactions can be perturbed by harsher ionization. Briefly, in SAWN, a small (~1-10 µl) volume of sample is deposited onto the surface of a $LiNbO_3$ piezoelectric substrate (SAWN chip) [28]. Next, the liquid sample droplet on the SAWN chip is nebulized producing a plume of microdroplets by actuation of the surface acoustic wave (SAW). Within a few seconds of being placed on the chip, the nebulized sample is analyzed by an atmospheric pressure ionization (API) mass spectrometer. Since the SAWN device does not require the addition of a high DC voltage on the sample droplet for nebulization to occur, the SAWN process is not likely to acidify the samples and induce oxidative damage to proteins like ESI. While SAWN has been used to ionize a host of analytes including small molecules and peptides, SAWN-MS of proteins and noncovalent protein complexes in native buffer has not been reported. One other potential benefit of SAWN over ESI, DESI or LILBID is that it is a planar device like MALDI. This provides the advantage that there are no capillaries or connectors to clog or set up, and liquid sample is simply pipetted onto the chip surface.

The goal of this study was to assess SAWN for analyzing small proteins and native protein complexes. Several comparisons of SAWN and ESI were done to investigate the impact

of ionization small molecule-protein interactions, and nonspecific protein-protein complex formation. I also examined use of global HDX with SAWN-MS on the homodimeric protein complex, TerS-A55, the DNA binding domain of a viral genome packaging motor complex, alone and in the presence of duplex DNA. The studies reported here demonstrate the value of SAWN for analyzing protein complexes and the ability for SAWN to rapidly assess protein-DNA interactions by global HDX studies.

3.2 Materials and Methods

SAWN-MS. SAWN chips were fabricated and the SAWN electronics were connnected in the same fashion as Heron et al. [28]. The SAWN chip is positioned approximately 0.5 inches below the orifice of the mass spectrometer and the front edge of the chip is positioned as close to the edge of the inlet without touching. To obtain mass spectra, a droplet (usually 1 µl) is pipetted onto the surface of the chip in front of the transducer. With the SAWN device operating (9.56 MHz, 26 dBm) in continuous mode the liquid droplet nebulizes immediately and the mass spectrometer aquires spectra. Most studies used a Thermo LTQ-Velos mass spectrometer (San Jose, CA), aside from the ubiquitin mass spectra shown in **Figure 3.3A**, which were acquired with a Waters Synapt G2S (Manchester, UK).

ESI direct infusion. A syringe pump (Harvard Apparatus, Holliston, MA) was used to deliver syringe sample in a 200 µl syringe (Hamilton company, Reno, NV) through a 50 µm i.d. fused silica capillary (Polymicro Technologies, Phoenix, AZ) which was joined to a junction connected to an electrode and a short 3.5 inch capillary tapered by a laser puller (P2000 model, Sutter Instrument Company, Novato CA). A flow rate of 1.5 µl/min was used and an electrospray voltage of 1.7kV was applied directly behind the tip to produce an electrospray

plume. Polyethylene ketone fittings used to construct the direct infusion line were purchased from Upchurch Scientific (Oak Harbor, WA).

Protein and DNA Constructs. Myoglobin (M-0630, Sigma) and ubiquitin (U100SC05M, Fisher Scientific R&D systems) was purchased and used without further purification. The TerS-A55 dimer was purified as previously described [29]. The 16 base pair DNA duplex substrates used in this study were obtained from Integrated DNA Technologies (IDT, Coralville, IA). The specific "R3" duplex encompasses the sequence of the R3 element ($^{5'}$-GGCGTTTCCGTTCTTC-$^{3'}$), to which TerS binds with high specificity [30]. The 16 base pair "NS" duplex was of random sequence ($^{5'}$-AAGACCTTAGAGACAT-^{3}).

SAWN-HDX protocol for time course studies with TerS-A55 with and without DNA. The HDX stock solution contained 100mM ammonium bicarbonate and 300µM TerS-A55 (pH 7.4). For a single time point, 1 µl of stock solution was combined with 19 µl of D_2O for the prescribed amount of time. The final concentration for HDX reactions was 15µM TerS-A55 in 5mM ammonium bicarbonate. Deuterium exchange times were 30 sec., 1, 5, 30, 60, 90, and 120 min. To quench the reaction, a 20% formic acid (1.25 µl) was added to achieve a pH of 2.5. For SAWN-MS, one µl of sample was then manually pipetted immediately onto the SAWN chip to obtain SAWN-MS spectra of the deuterated sample using an LTQ-Velos. All steps were completed at room temperature. The estimated time between quench and recording the first spectrum by SAWN was 4 seconds. For the HDX samples containing DNA the following modifications to the prior protocol were made. First, the stock solution contained 1mM DNA, so that dilution of the stock solution achieved a concentration of 50µM DNA. The quench solution was exactly the same as the one described previously, but contained and additional 2mg/ml of protamine sulfate to precipitate the DNA and facilitate the detection of protein by MS [31].

Maximum achievable deuteration was measured by drying the stock sample by vacuum concentrator, resuspending in 100% D_2O, boiling for 10 min, and recording the SAWN-MS spectrum. Peak centroids were determined using HX-Express v2 [32, 33].

3.3 Results and Discussion

Comparison of ESI and SAWN Mass Spectra of Myoglobin.

Myoglobin was anlyzed by both ESI- and SAWN-MS. We previoulsy demonstrated that SAWN produced a shift of the average charge state of peptides to higher m/z relative to electrospray ionization [28]. Consistently, **Figure 3.1** demonstrates that SAWN-MS analyis of myoglobin produces similar spectral shift relative to ESI as seen previously for peptides. Specifically, the base peak in the SAWN mass spectrum was the 10+ ion while it was the 12+ ion in the ESI mass spectrum. Given prior reports of SAWN producing ions with lower internal energy than ESI, I investigated the degree to which SAWN might preserve the association of heme with myoglobin [27].

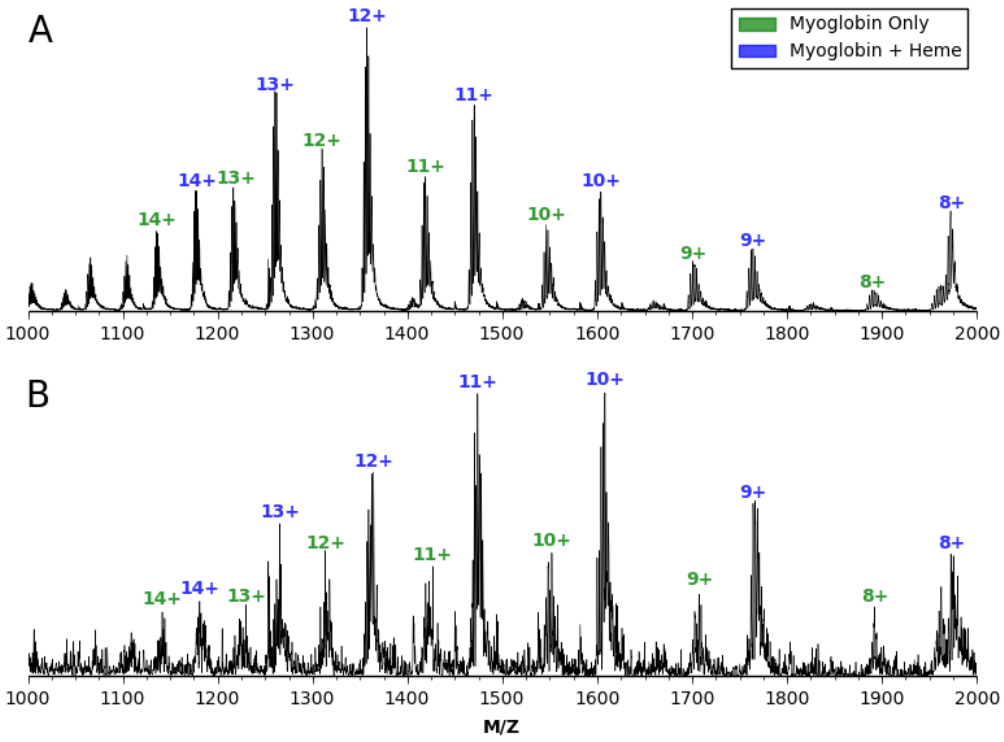


Figure 3.1. Native mass spectra of myoglobin. Panel A. Mass spectrum was produced by averaging mass spectra acquired over 1 minute of time generated by ESI direct infusion at 1 µl/min of an aqueous myoglobin solution. **Panel B.** Mass spectrum produced by nebulization of a 1 µl droplet of the same solution from the SAWN chip. Green and blue ion labeling corresponds to apo- and holo-myoglobin, respectively.

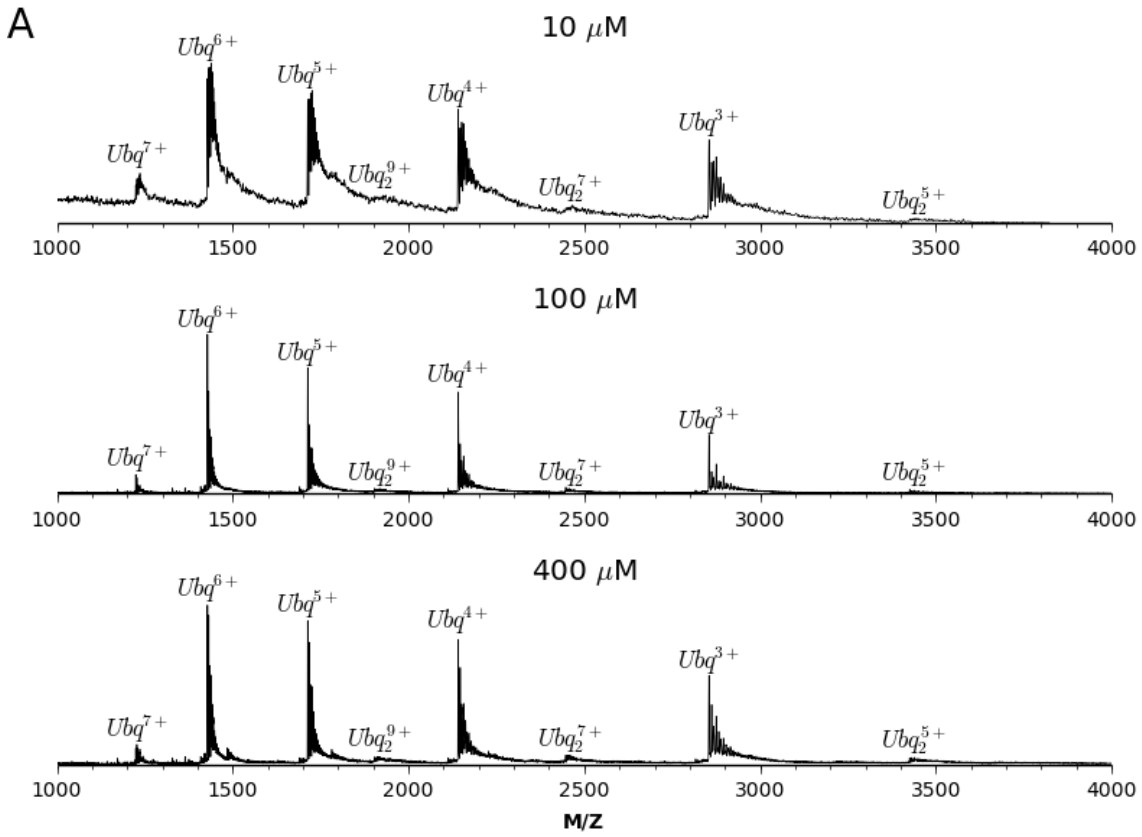
A comparison of the relative amounts of apo and holomyoglobin was used to assess the energetic softness of SAWN relative to ESI. It is well known that myoglobin noncovalently binds a heme molecule as a prosthetic group, and that holomyoglobin is readily visible in mass spectra of myoglobin if the native complex is not perturbed [34]. Both ESI and SAWN mass spectra of myoglobin clearly show both apo and holo forms of myoglobin, with holomyoglobin being the predominant species (**Figure 3.1**). However, when the sum of the peak area from the

extracted ion chromatograms for the +8 to +14 ions for holomyoglobin is divided by the sum of the peak area from the extracted ion chromatograms for the +8 to +14 ions for apomyoglobin, there is a difference between the SAWN and ESI spectra. SAWN showed 2.3 fold more signal for holo than apomyglobin, while ESI showed only 1.7 fold difference. This result indicates that SAWN perturbed the noncovalent association between the protein and heme less than ESI, which suggests that SAWN may advantages for analysis of intact noncovalent protein complexes.

Influence of SAWN on Nonspecific Protein-Protein Interactions.

Previous studies have shown that ESI can generate nonspecific protein-protein complexes [35–37]. Since SAWN can produce ions with lower internal ion energeties (i.e. it is softer) than ESI for small molecules [27], I hypothesized that SAWN may produce fewer nonspecific protein-protein complexes than ESI. Sun *et al*. examined non-specific dimerization of ubiquitin by ESI [35] and I repeated this study with SAWN-MS to directly test our hypothesis. Three different ubiquitin concentrations (10 µM, 100 µM, and 400 µM) were prepared as described by Sun *et al*. and analyzed by SAWN-MS. When these samples were ionized by SAWN, the spectra showed very little visual evidence of ubiquitin multimers, except for three very small dimer peaks (**Figure 3.2A**). In contrast, a significant population of dimers and even tetramers of ubiquitin were visible in the ESI spectra, (see **Figure 3.1 of reference** [35]). This observation was quantified by computational analysis of the mass spectrum. The SAWN mass spectrum was analyzed by the MaxEnt tool (Massylnx software, Waters) to derive a molecular weight spectrum, which revealed evidence of dimer, trimer, and tetramer . The amounts of these oligomeric states of ubiquitin were then compared to the ESI data collected by Sun *et al*. (see **Figure 3.2B** and **Figure 3.3A of reference** [35]), which demonstrates that there is a notable

difference between SAWN and ESI at 100 µM and 400 µM ubiquitin. In SAWN, the percentages of ubiquitin trimers and tetramers remained roughly constant between 100 µM and 400 µM, while in ESI, where there was a greater increase in the amount of ubiquitin trimers and tetramers present in the 400 µM versus the 100 µM samples (**Figure 3.3B)**. This demonstrates that even at elevated protein concentrations, SAWN induces less nonspecific protein-protein interactions than ESI, which minimizes the likelihood that spectral evidence of complexes are due to ionization-induced artifacts.



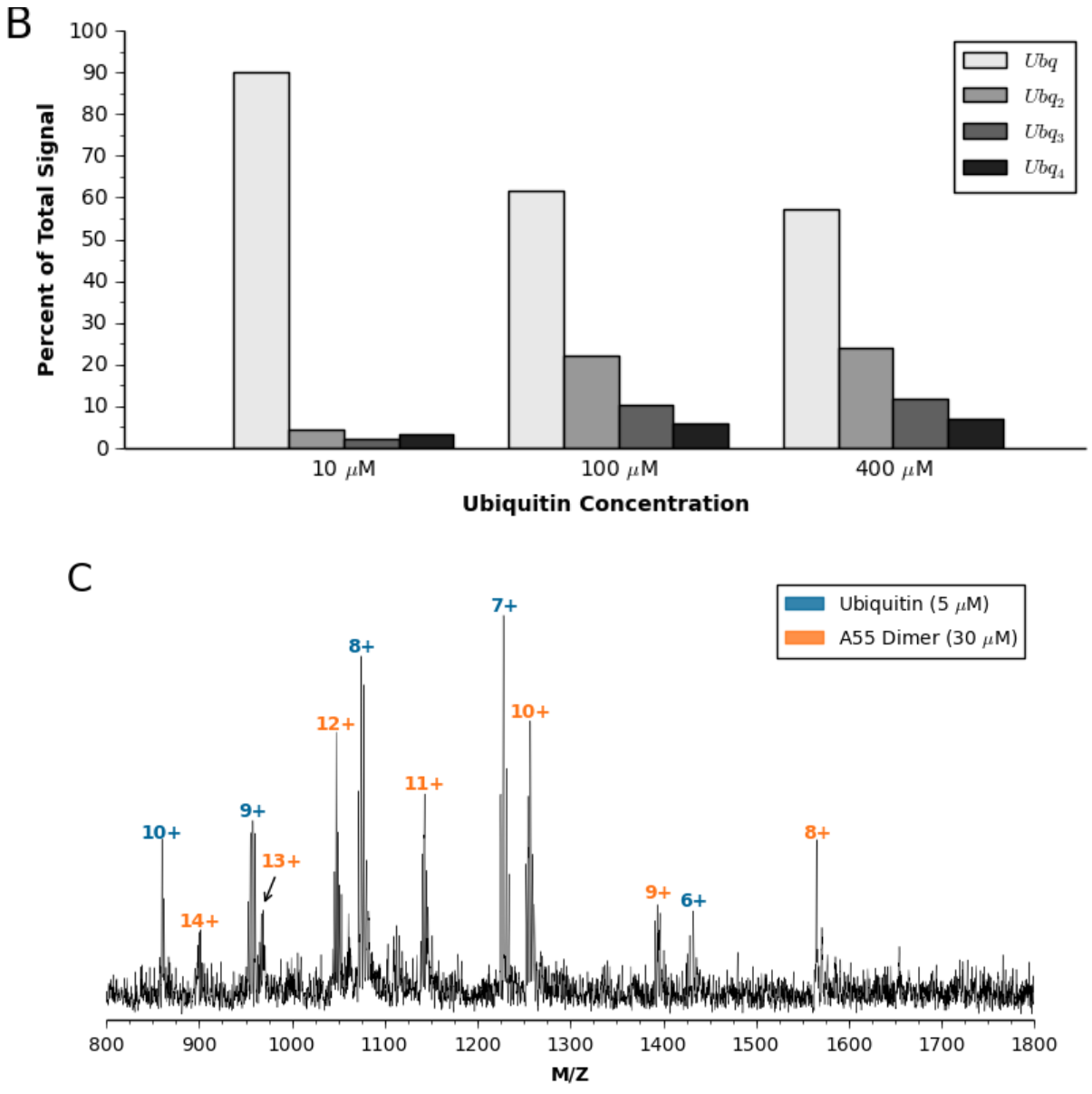


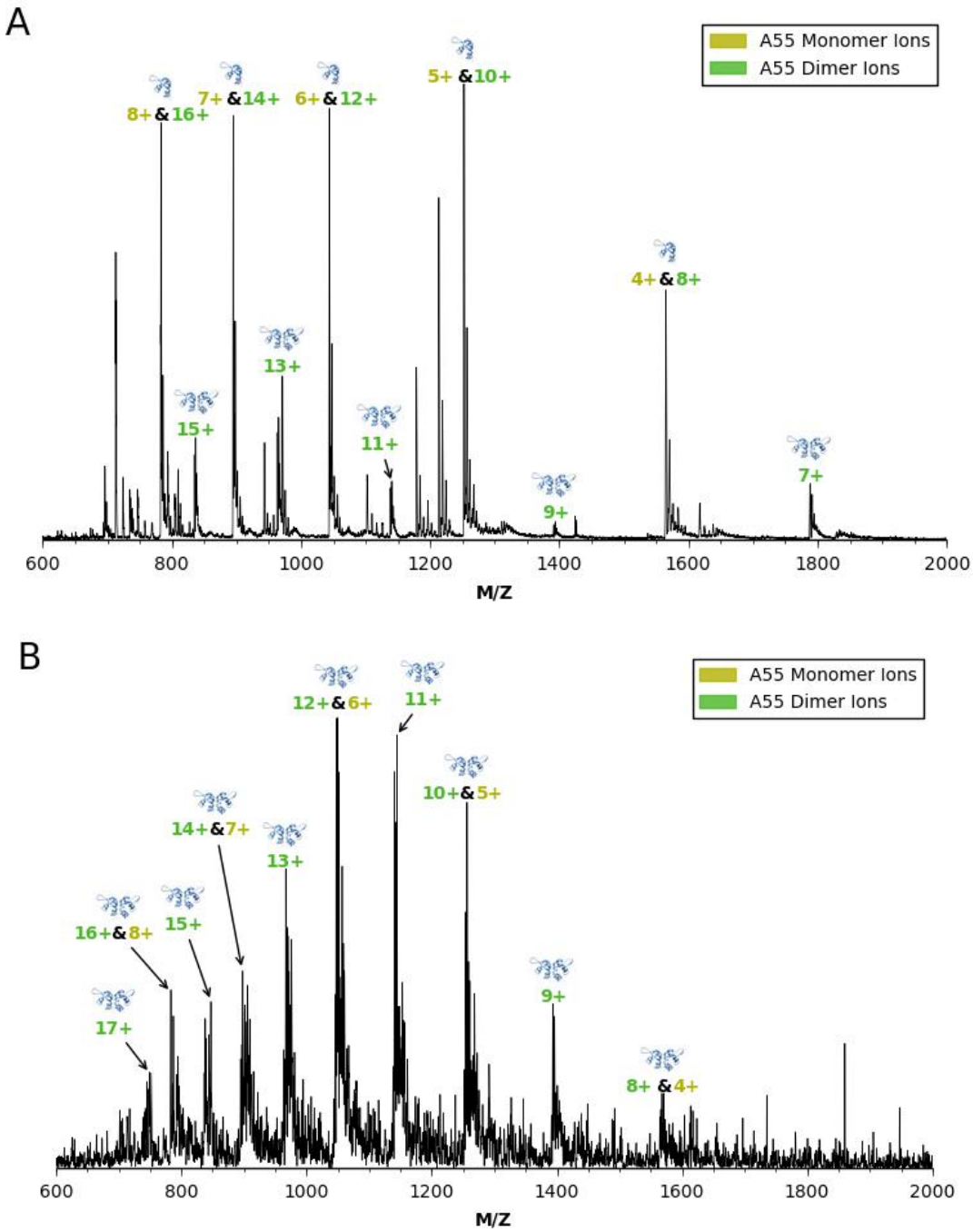
Figure 3.2. SAWN-MS of ubiquitin. Panel A. Mass spectra of 10, 100, and 400 µM ubiquitin recorded by SAWN-MS with a Waters SYNAPT G2S. **Panel B.** The degree of oligomerization present in the mass spectra in panel A was calculated using Waters' MaxEnt tool found in Massylnx software. **Panel C.** A 1.5 µl droplet containing a mixture of 30 µM TerS-A55 and 5 µM ubiquitin was nebulized by SAWN and recorded on an LTQ-Velos.

To further investigate if SAWN would produce nonspecific interactions, I analyzed a mixture of two non-interacting proteins. Specifically, a sample containing a six-fold stoichiometric excess of ubiquitin to TerS-A55, a recombinant viral protein construct with molecular weight 5500 Da [29], was analyzed by SAWN-MS. The results in **Figure 3.2C** indicate that no TerS-A55-ubiquitin complexes were detected, which further supports the claim that SAWN generates few non-native nonspecific protein complexes.

Comparison of ESI and SAWN Mass Spectra of TerS-A55 homodimer.

ESI and SAWN mass spectra of the viral protein construct TerS-A55, which is known to exclusively form a homodimer in solution [38], were acquired to test the utility of SAWN for native MS analysis. The first comparison shown is ESI and SAWN mass spectra of TerS-A55 in native conditions. In the ESI mass spectrum, there is no single semi-Gaussian charge state distribution present, but rather two distributions: one each for monomer and dimer (**Figure 3.3A**). In a homodimeric native mass spectrum as shown here, the intensity of the odd dimer ion series is diagnostic of the amount of dimer present in the mass spectrum, since the evenly charged dimer ions overlap with monomeric ions. In the ESI generated mass spectrum (**Figure 3.3A**), these dimer ions have a very low average peak height of 17% of the maximum intensity. In the same mass spectrum, the even ion series, which consists of both dimers and monomers, have an average peak height of 86% of the maximum intensity. This suggests that the ESI generated mass spectrum of A55 is virtually all monomer. In contrast, in the SAWN mass spectrum, the odd ion series of dimer ions forms a single charge state distribution along with the

other evenly charged ions (**Figure 3.3B**) suggesting that there is little to no monomer detected and that SAWN preserved the TerS-A55 dimer structure.



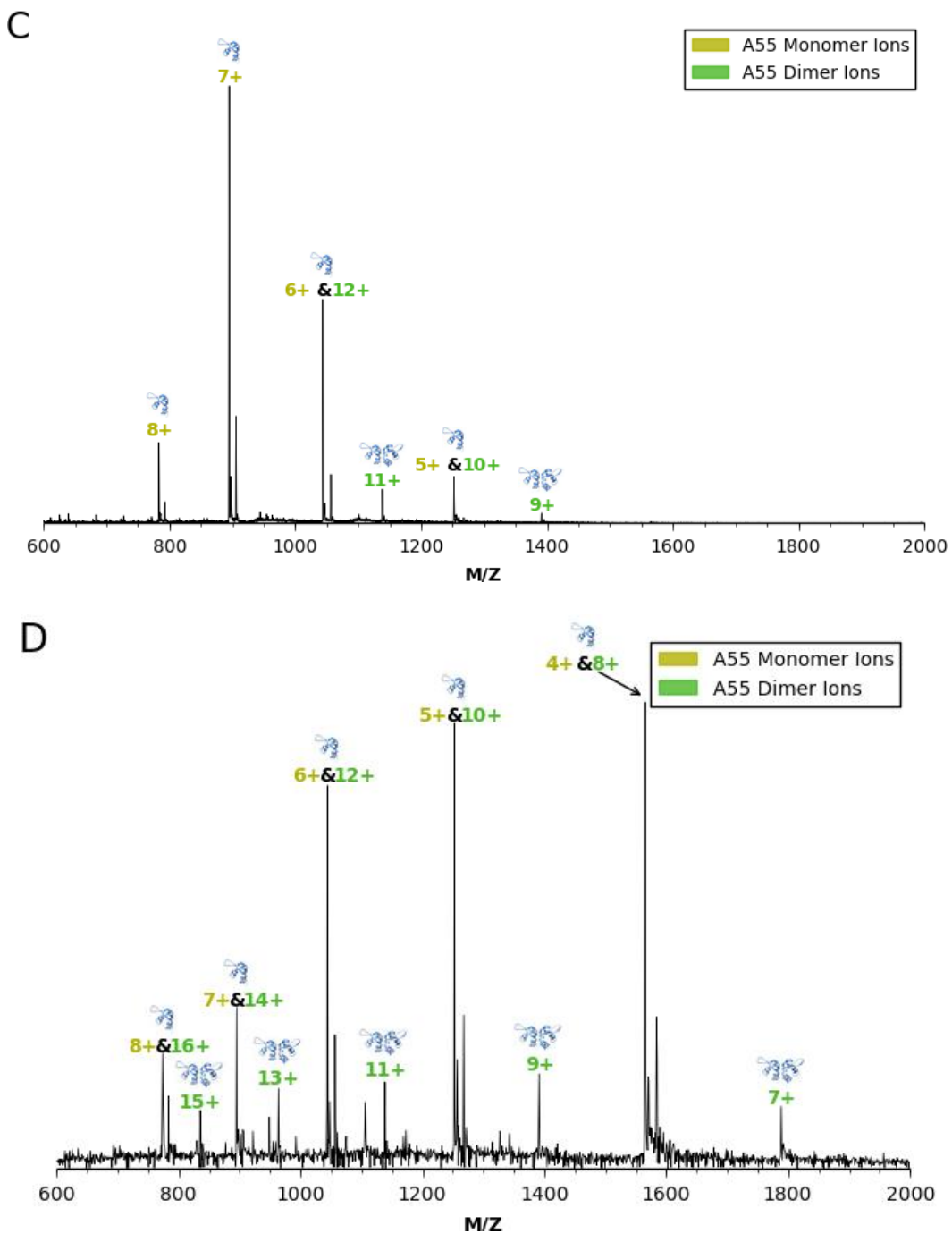


Figure 3.3. Native mass spectra of TerS-A55. Panel A. ESI mass spectrum of TerS-A55 (30µM) in native conditions. **Panel B**. SAWN mass spectrum of TerS-A55 in native conditions. **Panel C**. ESI mass spectrum of TerS-A55 (30µM) in denaturing conditions. **Panel D**. SAWN

mass spectrum of TerS-A55 (10μM) in denaturing conditions. Native buffer used was 10 mM, ammonium bicarbonate, pH 7.4, and denaturing buffer used was 5% Acetonitrile and 0.1% formic acid. The ESI generated mass spectra in **panels A** and **C** were achieved by direct infusion (0.5 μl/min) and the SAWN generated mass spectra in **panels B** and **D** were achieved by nebulizing 1 μl of the same solutions described above with no added voltage. The blue structures above the peaks indicate the predominant dimerization state of that ion, while yellow charge states refer to the monomer and green to the dimer. When monomer and dimer overlap, the predominant species is listed first in the charge state label.

Even under denaturing conditions, some of the TerS-A55 dimer remained intact (**Figure 3.3C and 3.3D**) [30]. As in the previous spectra under native conditions, it was observed that SAWN retained more TerS-A55 dimer than ESI. The ESI spectrum (**Figure 3.3C**) only had two odd charge dimer only peaks present with a relative intensity less than ten percent, while the SAWN spectrum (**Figure 3.3D**) had five oddly charged dimer only peaks with an intensity greater than ten percent. Since SAWN maintained a greater relative amount of TerS-A55 dimer than ESI under several conditions, this suggests that SAWN may be advantageous to use for native MS studies of other noncovalent complexes.

Hydrogen/Deuterium Exchange Studies with TerS-A55.

To further define the utility of SAWN-MS, I examined HDX of TerS-A55 in the absence and presence of DNA. ESI was not used for these studies because SAWN-MS was shown to retain the dimer much better than ESI making it a more appropriate method to study the solution based interactions of this complex. This is important because sedimentation equilibrium

centrifugation methods showed that TerS-A55 is almost exclusively a stable dimer in solution in the concentration range of 5 µM to 1 mM [38]. Another motivation for using SAWN over ESI for these experiments is that the HDX samples with DNA will contain considerable precipitate due to the addition of protamine sulfate [31]. This precipitate formation will impede ESI ionization by obstructing sample flow through the direct infusion line, whereas in SAWN, the precipitate will remain on the chip surface without hindering ionization.

As a first step, HDX analysis of TerS-A55 in mild denaturing conditions (5% ACN, 0.1% formic acid) was completed with SAWN-MS. This served to disrupt the dimer in solution and allow for the simultaneous analysis of the TerS-A55 monomer and dimer (as shown by **Figure 3.3C** and **D**). In **Figure 3.4**, the zoom scan of the +8 dimer and +4 monomer peak clearly indicates that the dimer has solvent protection versus the monomer. At the exact same time point of 5 minutes, roughly 74 total exchanges have occurred in two individual monomer units, whereas the dimer has only 50 exchanges, meaning that 24 fewer exchanges have occurred in the dimer. This indicates that the dimerization interface in the TerS-A55 dimer is less solvent accessible, which prevents deuterium from accessing exchangeable sites. It is further noted that these mass spectra were acquired by simply pipeting one microliter of the TerS-A55 solution onto the SAWN chip after initiating the HDX experiment. This shows that SAWN can rapidly obtain structurally relevant data for protein complexes.

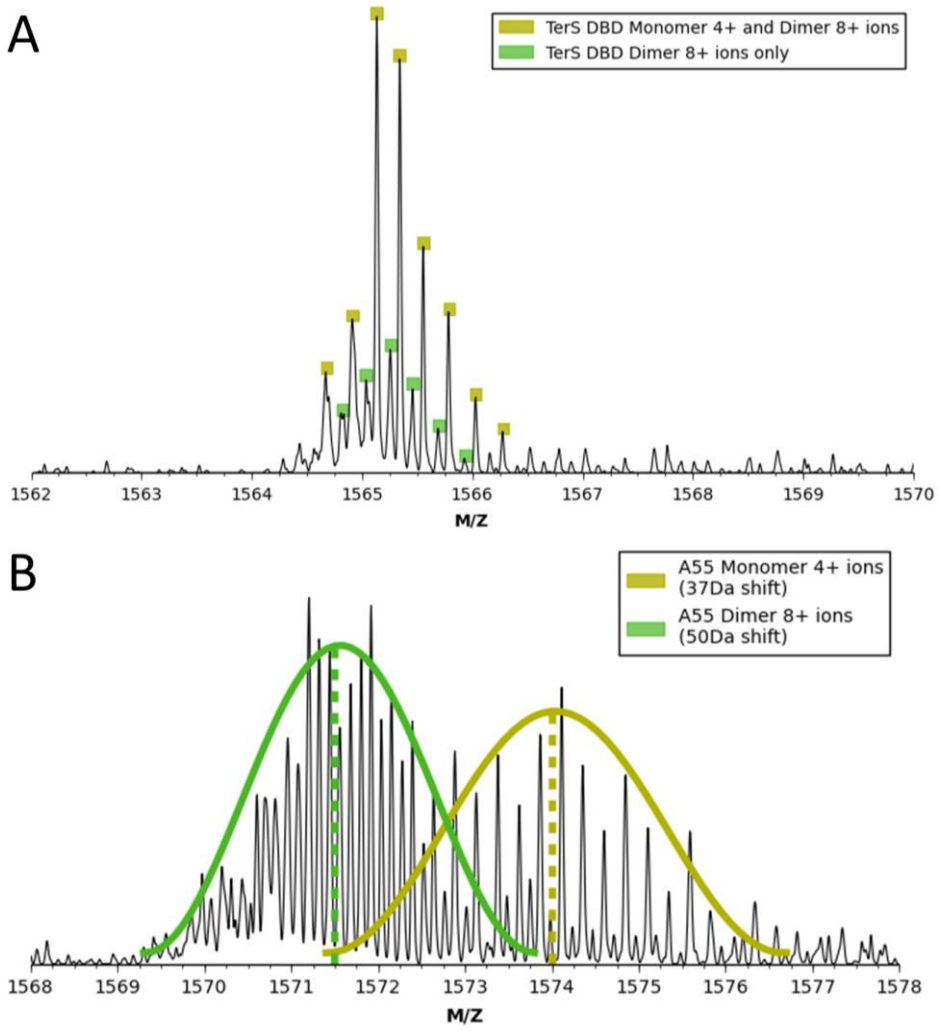


Figure 3.4. SAWN-MS ultra-zoom scans of TerS-A55 HDX. Panel A. Mass spectrum shows the unexchanged zoom scan of the 4+ monomomer and 8+ dimer ion in 5% Acetonitrile, 0.1% formic acid. **Panel B**. Mass spectrum shows an ultra-zoom scan of TerS-A55 +8 dimer and +4 monomer ions produced by SAWN before and after HDX performed in 70% D_2O under denaturing conditions (5% Acetonitrile, 0.1% Formic Acid) for 5 minutes. Solid lines outline the isotopic envelope for the particular species defined in the upper right hand legend. Dashed lines represent the estimated centroid for its respective isotopic distribution.

Finally, HDX studies of TerS-A55 were completed in the presence and absence of different DNA substrates in native buffer. The goal of this study was to gain insight into the dimerization properties of TerS-A55 and to probe protein-DNA interactions of TerS-A55 with a DNA substrate. In **Figure 3.5**, an example of acquired raw mass spectrometry data is displayed. While the data are not obtained from higher-resolution zoom scans like **Figure 3.4**, the resulting peaks could still be analyzed by HX-Express v2, which is capable of accurately assigning a centroid mass to complex peak shapes that are not gaussian in nature. The centroided masses of the peaks in **Figure 3.5** as determined by HX-Express v2 are displayed as color coded dashed lines on the figure, and these values were used to determine the average level of deuteration over the entire isotopic distribution of a given sample. As expected, the dimer only samples exchanged faster than TerS with DNA present, which is mostlikely due to protection of the DNA binding interface of the protein. Another important result was that the signal intensities of the TerS-A55 only peaks and the TerS-A55 with DNA peaks were similar. This is unusual for most HDX experiments with DNA and a DNA binding protein, since the acidifed complex has been observed to precipitate out of solution and evade detection by MS [31]. However, the addition of protamine sulfate in the quench, as done by Polikav et al [31], successfully precipitated the DNA only and prevented the precipitation of the acidified protein-DNA complex.

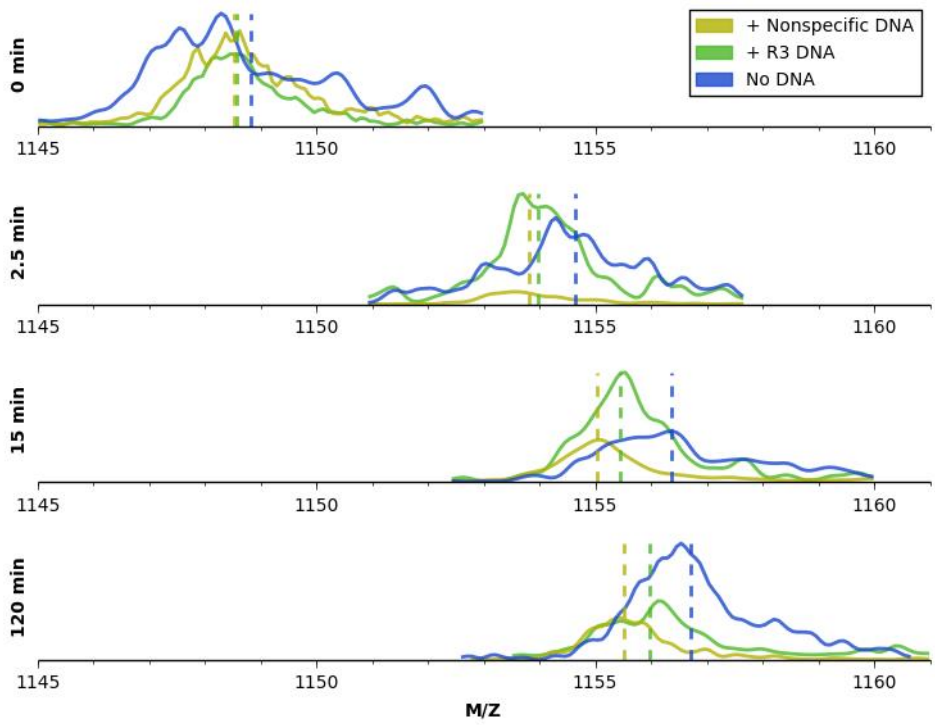


Figure 3.5. Time course of the 11+ dimer ion. The three traces represent a different SAWN-HDX experiment in the presence or absence of a DNA substrate (see legend upper right corner). Time zero refers to exposing the sample to D_2O with quench already present. All of the relative intensity values are true to the original spectrum. Vertical dashed lines indicate the average m/z for the peak as determined by HX-Express v2 [32, 33]. The undeuterated TerS-A55 11+ ion is 1142 *m*/*z* for reference.

To gather further evidence that TerS-A55 deuterium uptake differs when DNA is present, more rigorous HDX kinetic data was collected in triplicate, which is displayed in **Figure 3.6**. Each data point was calculated from the peak centroid obtained with HX-Express v2 as in **Figure 3.5**. Based on this data, the SAWN-HDX workflow still seemed effective and reproducible, due to the relatively low error bars and consistency in the kinetic trends across all

five ions analyzed. The data in **Figure 3.6** meets several important expectations: i) the backbone dueteration for all ions increases as incubation time increases for all ions, and as anticipated from **Figure 3.5**; ii) the SAWN HDX data could still distinguish free TerS-A55 versus TerS-A55 with DNA since the amount of backbone deuteration for TerS-A55 only was consistently higher than TerS-A55 with DNA present (statistically significant). On a final note, in order to highlight the ease and efficiency of this workflow, all triplicate HDX timepoints were collected on a single afternoon, using only one microliter of solution to collect a single spectrum.

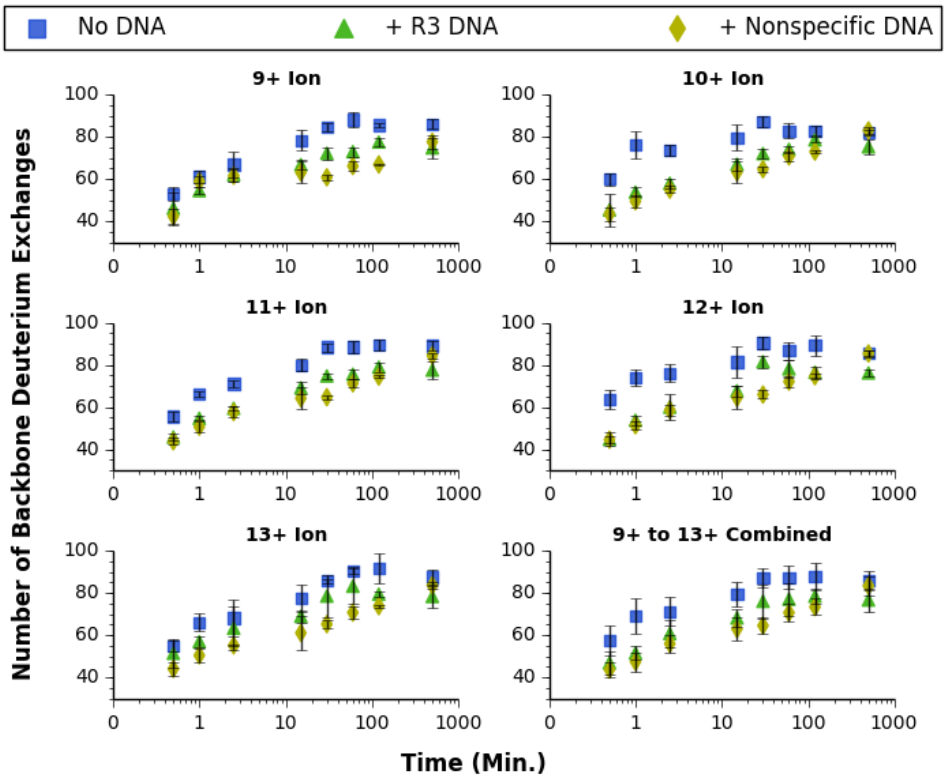


Figure 3.6. TerS-A55 time course amide backbone deuterium uptake. Shown are a series of time course measurement of amide backbone deuterium uptake of TerS-A55 under native conditions (pH 7.4, 5 mM ammonium bicarbonate) with and without DNA (see legend at top). Exchanges took place in 95% D_2O and was analyzed by SAWN-MS. Each plot corresponds to a different ion, and each data point was obtained by using HX-Express v2 to find the centroided mass [32, 33], calculating the mass, and subtracting it from the zero time point average weight to exclude side chain exchanges. All DNA HDX experiments were with 3 and 1/3 molar excess of DNA versus TerS-A55. All quenches contained 1.25 µl of 20% formic acid, and all HDX experiments with DNA also contained 2 mg/ml protamine sulfate to precipitate the DNA. The final time point in each plot does not correspond to a 500 min exchange: this data point corresponds to the maximum achievable deuteration.

3.4 Conclusions

While ESI is effective for ionizing protein complexes [11–17], in these studies SAWN performed better than ESI. Specifically, SAWN produced mass spectra with a higher abundance of holomyglobin versus apomyoglobin than ESI. SAWN also provided structural data more relevant to the in solution protein structure than ESI, judging by the fact that fewer nonspecific protein interacts occurred with ubiquitin and that known non-interacting proteins showed no spectral evidence of dimer ions. Additionally, SAWN produced mainly TerS-A55 noncovalent dimers in the gas phase whereas ESI produced mainly monomers of a solution known to contain 100% dimer. These encouraging results provide motivation to further explore the use of SAWN with additional protein complexes to establish the degree to which SAWN might be more universaly advantageous for Native MS analysis.

I also demonstrated the use of SAWN-HDX for obtaining relevant structral information for protein-protein and protein-DNA interactions. HDX-SAWN-MS analysis of denatured TerS-A55 showed a solvent protective effect in the dimer versus the monomer, which is most likely due to the subunit interface. The same analysis done in a non-denaturing buffer exhibited a marked decrease in deuterium uptake of TerS-A55 with DNA present versus no DNA present, which is consistent with DNA binding to the dimer. The results presented here lay the foundation for use of an automated digital microfluidics-SAWN platform, which is something we are currently pursuing [39].

3.5 References for Chapter 3

1. Alberts, B.: The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. *Cell* **92**(3), 291–294 (1998)

2. Jeong, H., Mason, S.P., Barabási, a L., Oltvai, Z.N., Barabasi, A.-L.: Lethality and centrality in protein networks. *Nature* **411**(6833), 41–42 (2001)

3. Johnson, J.E.: Functional implications of protein-protein interactions in icosahedral viruses. *Proc. Natl. Acad. Sci.* **93** (1), 27–33 (1996)

4. Yan, Y., Marriott, G.: Analysis of protein interactions using fluorescence technologies. *Curr. Opin. Chem. Biol.* **7**(5), 635–640 (2003)

5. Balbo, A., Minor, K.H., Velikovsky, C.A., Mariuzza, R.A. et al.: Studying multiprotein complexes by multisignal sedimentation velocity analytical ultracentrifugation. *Proc. Natl. Acad. Sci. United States Am.* **102** (1), 81–86 (2005)

6. Clore, G.M., Gronenborn, A.M.: NMR structure determination of proteins and protein complexes larger than 20 kDa. *Curr. Opin. Chem. Biol.* **2**(5), 564–570 (1998)

7. Mertens, H.D.T., Svergun, D.I.: Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J. Struct. Biol.* **172**(1), 128–141 (2010)

8. Wen, J., Arakawa, T., Philo, J.S.: Size-Exclusion Chromatography with On-Line Light-Scattering, Absorbance, and Refractive Index Detectors for Studying Proteins and Their Interactions. *Anal. Biochem.* **240**(2), 155–166 (1996)

9. van Rooyen, J.M., Abratt, V.R., Belrhali, H., Sewell, T.: Crystal Structure of Type III Glutamine Synthetase: Surprising Reversal of the Inter-Ring Interface. *Structure* **19**(4), 471–483 (2011)

10. Heck, A.J.R.: Native mass spectrometry: a bridge between interactomics and structural biology. *Nat Meth* **5**(11), 927–933 (2008)

11. Loo, J.A.: Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrom. Rev.* **16**(1), 1–23 (1997)

12. Smith, D.L., Deng, Y., Zhang, Z.: Probing the Non-covalent Structure of Proteins by Amide Hydrogen Exchange and Mass Spectrometry. *J. Mass Spectrom.* **32**(2), 135–146 (1997)

13. Hilton, G.R., Benesch, J.L.P.: Two decades of studying non-covalent biomolecular assemblies by means of electrospray ionization mass spectrometry. *J. R. Soc. Interface* **9** (70), 801–816 (2012)

14. Heck, A.J.R., van den Heuvel, R.H.H.: Investigation of intact protein complexes by mass spectrometry. *Mass Spectrom. Rev.* **23**(5), 368–389 (2004)

15. Sharon, M., Robinson, C. V: The Role of Mass Spectrometry in Structure Elucidation of Dynamic Protein Complexes. *Annu. Rev. Biochem.* **76**(1), 167–193 (2007)

16. Ashcroft, A.E.: Recent developments in electrospray ionisation mass spectrometry: noncovalently bound protein complexes. *Nat. Prod. Rep.* **22**(4), 452–464 (2005)

17. Hyung, S.-J., Ruotolo, B.T.: Integrating mass spectrometry of intact protein complexes into structural proteomics. *Proteomics* **12**(10), 1547–1564 (2012)

18. Boys, B.L., Kuprowski, M.C., Noël, J.J., Konermann, L.: Protein oxidative modifications during electrospray ionization: solution phase electrochemistry or corona discharge-induced radical attack? *Anal. Chem.* **81**(10), 4027–34 (2009)

19. Van Berkel, G.J., Zhou, F., Aronson, J.T.: Changes in bulk solution pH caused by the inherent controlled-current electrolytic process of an electrospray ion source. *Int. J. Mass Spectrom. Ion Process.* **162**(1–3), 55–67 (1997)

20. Konermann, L., Silva, E.A., Sogbein, O.F.: Electrochemically Induced pH Changes Resulting in Protein Unfolding in the Ion Source of an Electrospray Mass Spectrometer. *Anal. Chem.* **73**(20), 4836–4844 (2001)

21. Spectrometry, D.E.I., Ferguson, C.N., Benchaar, S.A., Miao, Z. et al.: Direct Ionization of Large Proteins and Protein Complexes by. *Anal. Chem.* **83**(17), 6468–6473 (2011)

22. Sobott, F., Wattenberg, A., Kleinekofort, W., Pfenninger, A., Brutschy, B.: Laser desorption mass spectrometry on thin liquid jets. *Fresenius. J. Anal. Chem.* **360**(7-8), 745–749 (1998)

23. Liu, P., Zhang, J., Ferguson, C.N., Chen, H., Loo, J. a: Measuring protein-ligand interactions using liquid sample desorption electrospray ionization mass spectrometry. *Anal. Chem.* **85**(24), 11966–72 (2013)

24. Morgner, N., Kleinschroth, T., Barth, H.-D., Ludwig, B., Brutschy, B.: A novel approach to analyze membrane proteins by laser mass spectrometry: from protein subunits to the integral complex. *J. Am. Soc. Mass Spectrom.* **18**(8), 1429–38 (2007)

25. Mager, F., Sokolova, L., Lintzel, J., Brutschy, B., Nussberger, S.: LILBID-mass spectrometry of the mitochondrial preprotein translocase TOM. *J. Phys. Condens. Matter* **22**(45), 454132 (2010)

26. Morgner, N., Barth, H.-D., Brutschy, B.: A New Way To Detect Noncovalently Bonded Complexes of Biomolecules from Liquid Micro-Droplets by Laser Mass Spectrometry. *Aust. J. Chem.* **59**(2), 109 (2006)

27. Huang, Y., Yoon, S.H., Heron, S.R., Masselon, C.D. et al.: Surface acoustic wave nebulization produces ions with lower internal energy than electrospray ionization. *J. Am. Soc. Mass Spectrom.* **23**(6), 1062–70 (2012)

28. Heron, S.R., Wilson, R., Shaffer, S.A., Goodlett, D.R., Cooper, J.M.: Surface acoustic wave nebulization of peptides as a microfluidic interface for mass spectrometry. *Anal. Chem.* **82**(10), 3985–9 (2010)

29. Yang, Q., de Beer, T., Woods, L., Meyer, J.D. et al.: Cloning, expression, and characterization of a DNA binding domain of gpNu1, a phage lambda DNA packaging protein. *Biochemistry* **38**(1), 465–77 (1999)

30. Ortega, M.E., Catalano, C.E.: Bacteriophage Lambda gpNu1 and Escherichia coli IHF Proteins Cooperatively Bind and Bend Viral DNA : Implications for the Assembly of a Genome-Packaging 5180–5189 (2006)

31. Poliakov, A., Jardine, P., Prevelige, P.E.: Hydrogen/deuterium exchange on protein solutions containing nucleic acids: utility of protamine sulfate. *Rapid Commun. Mass Spectrom.* **22**(16), 2423–2428 (2008)

32. Guttman, M., Weis, D.D., Engen, J.R., Lee, K.K.: Analysis of overlapped and noisy hydrogen/deuterium exchange mass spectra. *J. Am. Soc. Mass Spectrom.* **24**(12), 1906–12 (2013)

33. Weis, D.D., Engen, J.R., Kass, I.J.: Semi-automated data processing of hydrogen exchange mass spectra using HX-Express. *J. Am. Soc. Mass Spectrom.* **17**(12), 1700–3 (2006)

34. Katta, V., Chait, B.T.: Observation of the heme-globin complex in native myoglobin by electrospray-ionization mass spectrometry. *J. Am. Chem. Soc.* **113**(22), 8534–8535 (1991)

35. Sun, J., Kitova, E.N., Sun, N., Klassen, J.S.: Method for Identifying Nonspecific Protein−Protein Interactions in Nanoelectrospray Ionization Mass Spectrometry. *Anal. Chem.* **79**(21), 8301–8311 (2007)

36. Lane, L.A., Ruotolo, B.T., Robinson, C. V, Favrin, G., Benesch, J.L.P.: A Monte Carlo approach for assessing the specificity of protein oligomers observed in nano-electrospray mass spectra. *Int. J. Mass Spectrom.* **283**(1–3), 169–177 (2009)

37. Hossain, B.M., Simmons, D.A., Konermann, L.: Do electrospray mass spectra reflect the ligand binding state of proteins in solution? *Can. J. Chem.* **83**(11), 1953–1960 (2005)

38. Bain, D.L., Berton, N., Ortega, M., Baran, J. et al.: Biophysical characterization of the DNA binding domain of gpNu1, a viral DNA packaging protein. *J. Biol. Chem.* **276**(23), 20175–81 (2001)

39. Monkkonen, L., Edgar, J.S., Winters, D., Heron, S.R. et al.: Screen-Printed Digital Microfluidics Combined with Surface Acoustic Wave Nebulization for Hydrogen-Deuterium Exchange Measurements. *Proteomics* (2015)

Chapter 4

Development and Application of XLMS Studies to Lambda Terminase

4.1 XLMS Approach and Goals with Terminase

Cross-linking mass spectrometry (XLMS) is a powerful approach for characterizing protein structure because it is a high-throughput technique, requires minimal protein (30-100 micrograms), reflects relevant native solution structure, and is compatible with large, highly flexible proteins. While other techniques such as nuclear magnetic resonance and X-ray crystallography provide much higher resolution structural data, these methods require more concentrated protein, more expertise to analyze the data, and specialized sample preparation (such as isotope labeling or crystals) [1–3]. Despite the lower resolution distance constraint information obtained by XLMS, combing this approach with homology modeling is quite successful. For instance, models for bacterial type II secretion systems, soluble guanylyl cyclase [4], bacteriophage lambda capsids [5], and many other proteins and protein complexes have been generated using a combination of XLMS, homology modeling, and other experimental techniques [6–8].

While XLMS is increasingly being used to characterize protein structure, validation of the identified cross-linking data remains a challenge. This is due to the complexity of the digested cross-linked protein samples and the low abundance of *bone fide* cross-linked peptides. To address these issues, designer cross-linking reagents were developed to yield diagnostic ions during special types of fragmentation [9]. In addition, various crosslinking enrichment strategies were employed. These include gel electrophoresis of the pre-digested crosslinking reaction and the addition of functional groups to the cross-linker (such as biotin) to allow for enrichment of the modified peptides [9–13]. However, these methods can lead to the loss of cross-linked

peptides due to over-enrichment and prevent the use of conventional cross-linkers such as EDC and BS2. To maintain a simple workflow and allow XLMS to be accessible to researchers outside of their expertise, one approach is to simply adapt the informatics approach. To be useful, the software for data analysis must confidently identify the cross-linked peptides with understandable statistics, provide helpful graphics to help evaluate the data, and utilize open source software that can be installed on personal computers. In this report, I demonstrate the use of an integrated cross-linking approach that combines i) xComb, a cross-linked database search strategy [14], ii) EasyProt, an open-source database search software [15] and iii) custom written python code called PyXL. PyXL takes EasyProt search results, de-isotopes the MS2 spectrum, annotates the spectrum for all possible cross-link locations, provides intuitive and meaningful statistics to evaluate the potential cross-links, and displays the annotated cross-linked spectra. PyXL scripts are available upon request.

To illustrate this approach, the second focus of this paper is to use XLMS to provide structural information on the bacteriophage lambda terminase enzyme. Terminase enzymes are common to both prokaryotic and eukaryotic complex double-stranded DNA viruses and serve as powerful molecular motors that package viral genomes into pre-assembled capsids [16]. Lambda terminase is a hetero-oligomer composed of large (TerL, 73.3kDa) and small (TerS, 20.4kDa) subunits [17]. While there is substantial information on the enzymology of the enzyme, there is little structural information available. Given that viral DNA packaging in pathogenic human viruses is performed by analogous terminase enzymes, structural information on lambda terminase will provide the foundation towards the development of novel anti-viral therapeutics. Furthermore, terminase enzymes are about twenty times as powerful as myosin muscle fibers and are among the most powerful molecular motors known in biology [18]. Lambda terminase

packages DNA into empty viral capsids at a rate of ~600 base pairs per minute, ultimately to liquid crystalline density which generates over 20 atmospheres of internal capsid pressure. The mechanism for how terminase can package DNA inside capsids with such speed and force is unclear, in part due to the lack of structural information. The only structural information available for lambda terminase is a Nuclear Magnetic Resonance (NMR) structure of the N-terminal DNA binding domain of the TerS subunit.

Lambda terminase is an ideal complex for XLMS studies. Limited structural information exists because it is large (four TerL subunits and eight TerS subunits in the 456.4 kDa functional motor), it has multiple assembly states, it is dynamic, and has solubility that is sensitive to both pH and salinity. All of these factors make the characterization of terminase by traditional high-resolution NMR and X-Ray crystallography quite difficult. In contrast, XLMS is insensitive to conformational heterogeneity in the preparation and can yield important information with small amounts of protein sample. The chemical-crosslinking data and homology models of terminase reported here suggest that lambda TerL has significant mobility between the two globular domains within the protein, which has precedence in other bacteriophage terminase proteins and is relevant to the DNA packaging mechanism [19, 20]. The implications of my results with respect to the DNA packaging mechanism by the enzyme are discussed.

4.2 Materials and Methods

Protein Purification. TerS-A55, IHF, terminase protomer, and a mixture of terminase oligomer and protomer (mix) were purified by published procedures [21–24].

XLMS sample preparation and data acquisition. Prior to cross-linking, TerS-A55 and IHF were dialyzed into 50mM sodium phosphate buffer, pH 7.2. The terminase protomer was

maintained in native buffer (20mM sodium phosphate pH 6.8, 100mM NaCl, 10% glycerol, 1 mM EDTA, 7mM beta-mercaptoethanol). All cross-linking reactions contained 30µg of protein and 10mM bis[sulfosuccinimidyl]glutarate (BS2, solubilized in DMSO, Thermo Scientific) or 1-ethyl-3-(3-dimethylamino- propyl)carbodiimide hydrochloride (EDC, Thermo Scientific) as indicated in each experiment. Cross-linking was completed at room temperature for 2 hours.

At the end of the reaction period, samples were dialyzed into their respective cross-linking reaction buffer to remove excess cross-linking reagent and the proteins were denatured using 6M urea, reduced with DTT, and alkylated with iodoacetamide. The samples were then digested with sequencing grade modified trypsin from Roche Applied Science (Indianapolis, IN) added at a 50:1 ratio (protein to trypsin) for 2 hours at 37°C and then overnight at room temperature. The resulting peptide mixtures were dried, desalted (using microspin columns from the Nest group, Inc.) and resuspended in 5% HPLC grade acetonitrile in water (Burdick and Jackson). Data acquisition was performed as described by Singh *et al.* [25].

XLMS data analysis. First, thermo raw files were converted by EasyProtConvert into an mgf file format. The resulting mgf files were then modified with a short python code to include scan titles that are recognized by EasyProt. Prior to the Easyprot search, a cross-linked peptide database was created using xComb. Tolerant xComb parameters were used: a 4 amino acid minimum per peptide in the cross-link, trypsin with no proline rule, no missed cleavages required. Both the mgf file and the cxdb file were uploaded onto EasyProt and an initial screen removed linear peptides, using the full-length protein. Important settings include the following: scoring model CID_LTQ_scan_Orbitrap_6ppm, turbo was off, trypsin with no proline rule, precursor ppm set to 10, min peptide score 4, min peptide length 6, missed cleavage was 3, with variable modifications for cysteines treated with iodoacetamide and oxidized methionines.

Unmatched peaks from this search were exported and another EasyProt search was completed with the xComb database. Many of the same search parameters were used, except this time, "do not cleave" was selected, min scores were set to 1, and PTMs were specified in accordance to Panchaud *et al.* to correct for cross-link linearization [14]. Finally, results were exported into a file, with all columns checked and the false discovery rate set to 100%. Easyprot output was then submitted to PyXL along with the modified mgf file to annotate the spectra for all possible cross-link locations and save visuals of plots of annotated mass spectra.

Homology modeling. Both TerS and TerL homology modeling was achieved using the ITASSER online structure prediction server [26–28]. For lambda TerS, three searches were submitted using three different template parameters: i) none (open search), ii) the phage 44rr TerS crystal structure as a template (PDB 3TXQ) [29] and iii) the phage sf6 TerS crystal structure (PDB 3HEF) [30]. These searches generated a pool of 15 potential homology models.

Candidate lambda TerL homology models were made as follows. Three ITASSER searches were performed: i) none (open search), ii) the T4 TerL crystal structure as a template (PDB 3CPE) [29], and iii) the Sf6 TerL crystal structure as a template (PDB 4IDH). Models returned by ITASSER in the unconstrained search either closely resembled T4 TerL, or diverged significantly from T4 TerL. The divergent models often had large areas of disorganization or long loops protruding from the main part of the protein, characteristics that are energetically disfavored. These were not considered further.

4.3 Results and Discussion

A New Approach to Identify Cross-Linked Peptides.

Identifying cross-linked peptide pairs and accurately defining the specific cross-link location in the pair from XLMS data using conventional cross-linkers like EDC and BS2 (Thermo-Pierce) is a difficult undertaking. Many programs and software packages were developed to analyze chemical cross-linking data to automate and aid XLMS data analysis. Informatics tools like Crux, the open modification search pipeline, SQUID, and many other programs simply output a large spreadsheet containing potential cross-linked pairs, a list of cross-linked fragments that were matched in the tandem mass spectrum (MS2), and a score that provides an indication of how good a match the candidate is to the theoretical cross-link. However, it is often difficult to confidently assign the cross-link without a visual presentation of the annotated candidate cross-linked mass spectra. In many cases, cross-links that score well in these programs can i) derive from poor quality spectra with low abundance and low intensity ions, ii) reflect unmodified peptides that are adjacent to each other in the protein, and/or iii) fail to assign the higher abundant fragments in the spectrum. The best way to accurately assign a cross-linked peptide pair and pinpoint the location is to manually view a thoroughly annotated cross-linked spectrum.

Here I report a new, open source, integrated approach for analyzing CXMS data to accurately assign cross-linked amino acids. The workflow for this approach is detailed in **Figure 4.1**. In the first portion of the CXMS data analysis pipeline, Easyprot tools, xComb, and short python scripts are combined to perform the initial screen to select a pool of cross-linked candidates. Then, the candidates are analyzed by PyCXL, a python code that I developed that first calculates all possible cross-link locations within a peptide pair, assigns charge states to as

many peaks as possible with a deconvolution algorithm, displays and annotates all possible cross-linked locations with the same mass spectra in visual and list presentations, and finally displays several intuitive metrics to allow the user to best assign the cross-link location.

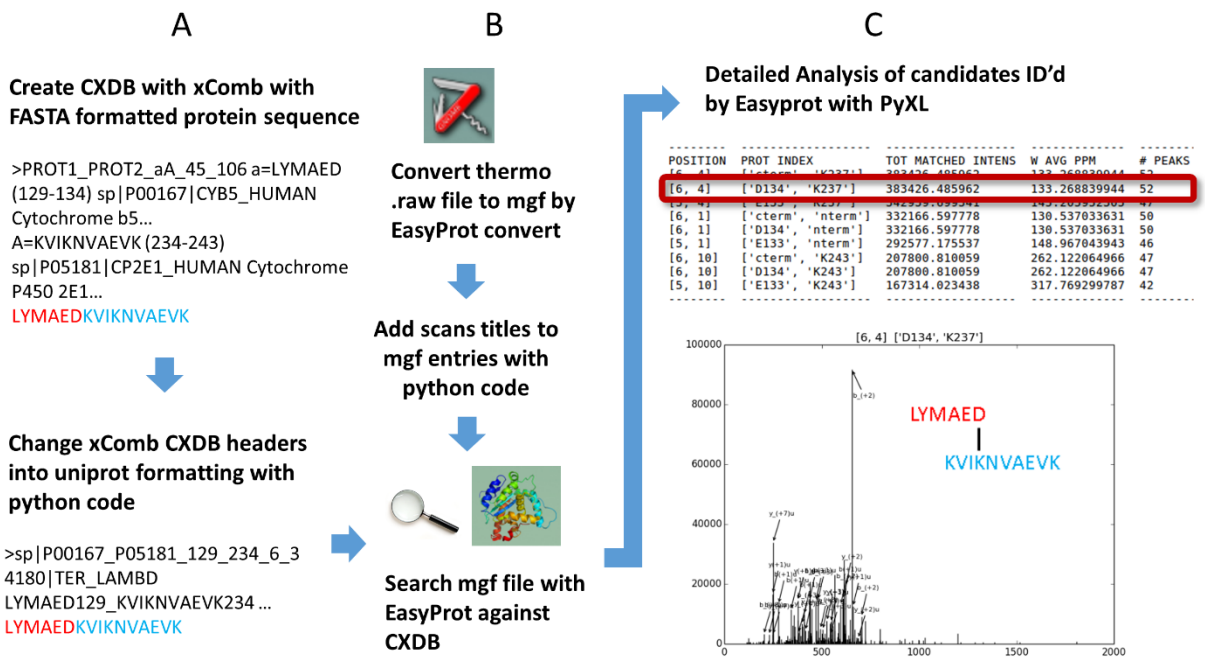


Figure 4.1. Diagram of the complete informatics workflow to identify cross-links from high resolution LC-MS data. **Panels A** and **B** show the processing of the two inputs required for a cross-link search. **Panel A** shows how a cross-linked database generated by xComb is converted by custom python code into uniprot formatting. This allows for EasyProt to parse the database correctly when it is uploaded. The header also contains the protein index for each peptide in the crosslink, and denotes where the first peptide begins and ends. **Panel B** shows how the raw data is converted into an mgf file with the proper scans title to allow for the PyCrux CX-Viewer in **Panel C** to retrieve the appropriate scan. **Panel C** shows how the cross-link location is determined by visual inspection of all annotated mass spectra of chemically possible cross-link locations and fragments assigned by PyXL, a python program using matplotlib, and ipython notebook.

To begin validating this informatics approach, a previously published dataset of cytochrome 2e1-b5 cross-linked with EDC was reanalyzed using my EasyProt/xComb/PyXL approach. In **Figure 4.2**, a good candidate for a cross-linked peptide pair identified by EasyProt is shown in the Easyprot spectrum viewer after the search against the xComb cross-linked database. Few fragments are identified since the xComb approach can only match fragments to the spectrum that exclude the crosslink. However, the ease and speed of the Easyprot/xComb approach are advantageous for generating a pool of quality cross-link candidates. When a cross-linked candidate like in **Figure 4.2** is found, a more thorough analysis of the candidate cross-link is performed via PyXL and all fragment ions are considered when annotating the spectrum, as shown in **Figure 4.3**. To perform the PyXL analysis, an exported text file of the EasyProt search and the mgf file with the scan titles are inputted into the code. Then, PyXL systematically takes each search entry, takes the corresponding candidate tandem mass spectrum and deconvolutes it, calculates all possible cross-link locations based on the cross-linked sequence, and annotates the de-isotoped spectrum with fragments of all of these possible cross-link locations. For each cross-link location, PyXL records all of the fragments and the mass error (e.g. **Figure 4.3C**), and then calculates the sum of the total matched peak intensity, average ppm mass error for the assigned peaks weighted by peak intensity, and the total peaks matched within a spectrum for each cross-link location, which are summarized in a table (such as **Figure 4.3A**). For additional confidence in annotation of the MS2 spectrum, the last column in the fragment ion tables describes whether the charge state could be assigned by the deconvolution process. Finally, the top three locations that have the best matches saves an image of the respective annotated mass spectrum in a separate document (3 of which are shown in **Figure 4.3B**).

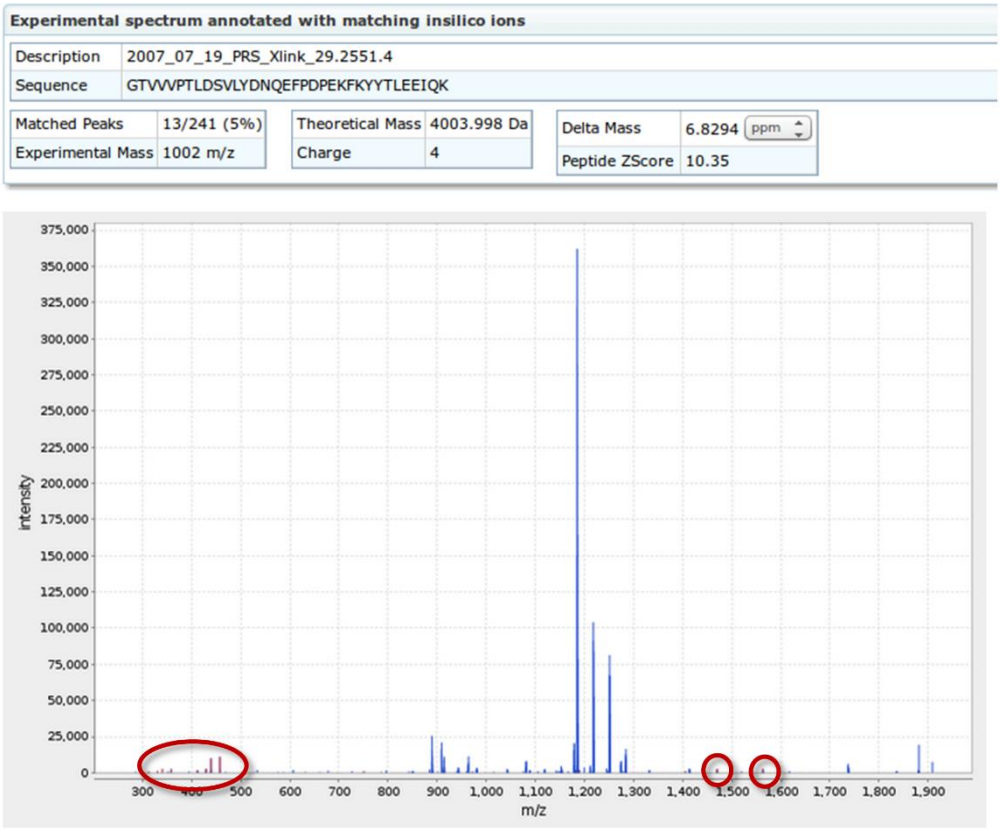


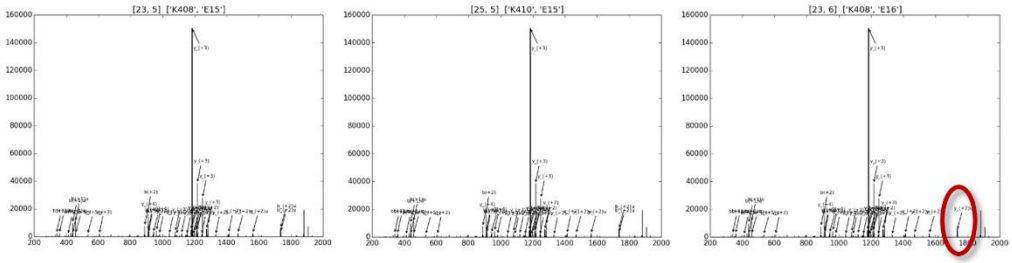
Figure 4.2. Screenshot from a cytochrome 2e1-b5 cross-linked candidate identified with EasyProt with an xComb cross-linked database. Red circles indicate fragment ion matches from theoretical ions generated from a concatenated version of the crosslinked peptide pair. Even though Easyprot can identify very few fragments, the quality of the spectrum and low precursor mass error make this a good candidate to analyze by PyXL.

A

scan number = 2551.4
peptide A = GTVVVPTLDSVLYDNQEFPDPEKFK, length = 25, prot index = 385, protein = Cyt2E1
peptide B = YYTLEEIQK, length = 9, prot index = 10, protein = CytB5

POSITION	PROT INDEX	TOT MATCHED INTENS	W AVG PPM	# PEAKS MATCHED
[23, 5]	['K408', 'E15']	408167.840637	459.219159714	82
[25, 5]	['K410', 'E15']	408167.840637	455.637555581	82
[23, 6]	['K408', 'E16']	404185.892517	478.024623232	82
[25, 6]	['K410', 'E16']	404185.892517	474.407733944	82
[22, 9]	['E407', 'K19']	398558.449402	460.125231334	79
[20, 9]	['D405', 'K19']	394347.11792	458.487522047	78
[17, 9]	['E402', 'K19']	370077.228638	478.514953652	71
[14, 9]	['D399', 'K19']	362085.70282	478.801668434	68
[9, 9]	['D394', 'K19']	343907.306702	503.525526032	58

B



C

[25, 5]

M/Z	INTENSITY	PPM	ION	SEQUENCE	KNOWN CS
1183.23901517	150517.25	571.630903897	y(+3)	PTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(20,5)	
1216.2607876	38689.8046875	556.967373991	y(+3)	VPTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(21,5)	
1249.28220957	28155.0644531	543.359006718	y(+3)	VVPTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(22,5)	
907.464913745	16219.6474609	19.8493859456	b(+2)	GTVVVPTLDSVLYDNQE-NH3	
456.281900974	11002.0390625	0.677339557901	b(+1)	GTVVV	u
438.271140407	10036.7607422	1.15125235325	b(+1)	GTVVV-H2O	u
1283.64703691	9112.07226562	516.643066338	y(+3)	VVVPTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(23,5)	
887.684326226	7821.75537109	567.967414814	y(+4)	PTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(20,5)	
1735.79027117	6181.49609375	38.0327177478	b(+2)	GTVVVPTLDSVLYDNQEFPDPEKFK, YYTLE(25,5)-H2O	u
1274.11296803	5452.82275391	381.839133909	y(+2)	NQEFPDPEKFK, YYTLEEIQK(11,5)	
1179.24449912	5164.22265625	852.319767614	y(+3)	PTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(20,5)-NH3	u
1210.91482152	5107.33251953	16.713931931	y(+3)	VPTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(21,5)-H2O	u
1177.2419477	4975.91503906	569.064991846	y(+3)	PTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(20,5)-H2O	
961.982356244	4518.30419922	527.530384215	y(+4)	VVVPTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(23,5)	
1211.25768149	4236.89746094	4.44563050443	y(+3)	VPTLDSVLYDNQEFPDPEKFK, YYTLEEIQK(21,5)-NH3	u
908.972298944	4146.99023438	541.300436857	y(+2)	PEKFK, YYTLEEIQK(5,5)	
1079.5093693	3784.31298828	627.890672657	y(+3)	DSVLYDNQEFPDPEKFK, YYTLEEIQK(17,5)	
1198.2505569	3598.76708984	1946.63789173	b(+3)	GTVVVPTLDSVLYDNQEFPDPEKFK, YYTLEE(25,5)-H2O	u

Figure 4.3. Screenshots from PyXL analysis of an EDC-linked cytochrome 2e1-b5 cross-link pair in **Figure 4.2**. PyXL tries to match all possible b and y fragment ions in the spectrum. **Panel A** shows a summary of all of the possible cross-link locations, the total matched intensity, the average ppm mass error for the matched fragments (weighted by intensity), and the number of peaks matched. **Panel B** shows how the annotated spectra for a specific possible crosslinker location are displayed by PyXL using matplotlib. As expected, the first two spectra look virtually the same, but the third has one less peak than the first two spectra, which is highlighted with a red circle. Additionally, in PyXL, all of the matched fragments are listed, such as the excerpt

shown in **C**. Matched fragments are ordered by intensity, and in the last column, the absence of a "u" (for unknown) indicates that the charge state for the annotated peak was assigned through the deconvolution algorithm. The most likely position for this crosslink is at amino acid 25 in peptide A and 5 in peptide B, since this option has the highest matched intensity and lowest weighted ppm fragment mass error.

The particular cross-linked example shown in **Figures 4.2** and **4.3** demonstrates the need to have detailed information to assign a true cross-link location. As seen in **Figure 4.3A**, there are nine possibilities where the cross-link can occur and the top four potential cross-link sites identify the same number of peaks. To complicate matters further, the top two choices match the same total peak intensity, but ultimately, the best cross-link location that describes this mass spectrum is amino acid 25 in peptide A and amino acid 5 in peptide B since the average weighted ppm is lowest (25,3).

Recently, other labs have recognized the need for visualization of the mass spectrum as well. Stavrox, plink, and crosswork are all XLMS software that are open-source, windows compatible, and allow for visualization of the MS2 spectrum [31–33]. Although the method described here is not windows compatible, EasyProt has recently provided a user-friendly virtualbox download with EasyProt and the raw file conversion programs already installed so that a linux virtual machine can be used on any type of personal computer. The only additional installations required on the virtual machine for PyXL is ipython notebook and the scipy packages (to allow for use of matplotlib), which are relatively straightforward. While there are many other programs available for analyzing XLMS data, most do not allow for visualization of the tandem mass spectrum.

Validation of the xComb/PyXL Approach.

A comparison of the cross-linking results obtained from the xComb/PyXL method, Stravrox, and a program called Crux was performed to assess the utility of the integrated method described here. The results are shown in **Table 4.1**. More cross-links were identified with the integrated xComb/PyXL method versus the other two published methods. Specifically, five more cross-links were identified with my integrated method versus StavroX and ten of my cross-links were not identified in the Crux analysis completed by McIlwain *et al*. [34]. Importantly, the best cross-link identifications by my integrated method typically had a high percentage of matched peak intensity in the tandem mass spectrum, and low average fragment mass error (weighted by intensity). I further note that many of the high quality matches possessing these characteristics did not have a consistent correlation with the EasyProt zscore assigned (e.g., cross-link with spectrum 2551.4) or the StavroX score (e.g., cross-link with spectrum 1793.4). This shows that reliance on the scores assigned by the cross-link identification tools can be detrimental and a more careful and detailed analysis of the tandem mass spectrum, as described here with PyXL, is necessary to optimize cross-link identifications. Also, this analysis shows that my integrated method will identify more cross-links from the XLMS data using stringent mass spectral metrics.

peptide 1	peptide 2	scan #	CXL loc PyCXL	StavroX	Crux	EasyProt zscore	Stavrox Score	precur. ppm	% matched pk intens.	avg frag ppm (weight)	# peaks ID'd
LYMAED	KVIKNVAEVK	1267.4	6, 4	(5/6), 4	6, 4	19.96	161	-8.78	70.5	96.90	103
EQAGGDATENFED VGHSTDAR	NIPK	1298.4	19,4	9,4	x	12.01	58	3.43	68.7	137.22	88
HNHSKSTWLILHHK	FLEEHPGGEEVLR	1350.6	14, 3	14, 3		5.54	49	-9.31	48.5	255.53	116
HNHSKSTWLILHHK	FLEEHPGGEEVLR	1358.6	5, 3	5,9		5.4	46	-9.77	51.4	242.60	115
EQAGGDATENFED VGHSTDAR	IPAIK	1599.4	13, 5	12,5	x	10.15	48	-8.52	66.0	121.47	78
YSDYFKPFSTGKR	EQAGGDATENFE DVGHSTDAR	1605.5	6, 1	6, 1	6, 1	25.94	51	-9.31	61.4	240.76	155
FLEEHPGGEEVLR	YKLCVIPR	1657.5	3,2		3,2	6.64		2.74	63.4	111.08	93
FLEEHPGGEEVLR	YKLCVIPR	1658.4	3,2		3,2	13		2.09	59.3	138.08	71
EQAGGDATENFED VGHSTDAR	GYLIPK	1665.4	19, 6	19, 6	x	8.34	34	1.54	52.3	216.78	101
FITLVPSNLPHEATR	NVAEVK	1793.4	12, 6	12, 6	x	12.63	1	2.09	52.9	279.13	58
FITLVPSNLPHEATR	NIPK	1798.4	12, 4	12, 4	x	10.66	40	2.60	38.3	294.81	44
LYMAED	EMSKTFIIGELHPD DRPK	2067.5	6, 4	6, 4	x	27.14	92	-0.94	72.4	139.50	113
PEHFLNENGK	DLTDCLLVEMEKE KHSAER	2246.4	10, 18		x	4.14		-6.94	49.2	882.19	75
GTVVVPTLDSVLYD NQEFPDPEKFK	YYTLEEIQK	2551.4	25,5	23, 5	x	10.62	34	6.83	69.2	455.64	82
GTVVVPTLDSVLYD NQEFPDPEK	FSLTTLRNYGMGK	2746.4	9, 13		x	2.94		6.76	49.5	433.69	56
GTVVVPTLDSVLYD NQEFPDPEK	RVCAGEGLAR	2749.4	23, 6		x	5.14		4.24	52.2	219.71	43

Table 4.1. Comparison of cross-linking results from cyt2e1-b5 cross-linked with EDC (data collected by [25]) analyzed with the integrated workflow described here, StavroX, and the Crux approach reported in [34]. The cross-linker location for each cross-linked pair is listed for each program tested. Easyprot zscore is obtained by searching the data against the xComb database. In the Crux location column, a blank means that the cross-linked pair was mentioned in [34] but no scan number was reported to allow for a direct comparison. An "x" in the same column means that the cross-linked pair was not identified in [34]. As a final note, the second to last column, average fragment ppm (weight), refers to the average mass error for the assigned fragment weighted by intensity (i.e. higher intensity fragments are weighted more). For this analysis, the

MS2 spectrum must have a base peak intensity higher than 10000, and the cross-linked peptide must be at least 4 amino acids long. Cross-linked peptide pairs that are adjacent to each other in the protein sequence are discounted. In theory, the higher the EasyProt zscore, the less likely the candidate should be due to chance. The reverse is true for the StavroX score: the higher this score, the less likely the candidate identified should be due to chance.

Of the cross-links identified by my integrated method and Crux, all of the locations assigned were in agreement, as seen in **Table 4.1**. Interestingly, Stavrox and PyXL had four discrepancies in cross-linker assignment and one cross-link where Stavrox could not distinguish two possible locations. These particular cross-links are explored in more detail in **Table 4.2**, where fragments identified by PyXL that are diagnostic of the cross-linker location are displayed. The first, second, and last candidate cross-link location have few low mass error fragments assigned, giving confidence in the location assigned by PyXL (**Table 4.2**). In contrast, cross-link with spectrum number 1599.4 has only 2 fragments with low abundance that distinguish this particular cross-link location and make this assignment less certain than the others. The following cross-link with spectrum number 2551.4 is an even less certain case. The particular fragment listed can be annotated in two different ways, and the fragment matched with the 25,5 location yields the lowest mass error. These last two examples highlight the need for high accuracy and comprehensive annotation of the MS2 spectrum as with PyXL, since small peaks can influence the cross-link location greatly.

(EQAGGDATENFEDVGHSTDAR, NIPK)	**Scan 1298.4**	**PyCXL loc 19,4**		**StavroX loc 9,4**		
theor m/z	exp m/z	ppm	intensity	ion	sequence	ch. known
616.73929	617.29515	901.29	28350.4	b(+2)	EQAGGDATENFE-NH3	
843.41627	843.41292	3.97	3375.97	y(+2)	FEDVGHSTDAR, NIPK(9,4)	
900.43774	900.43535	2.65	5477.99	y(+2)	NFEDVGHSTDAR, NIPK(10,4)	
600.62777	600.62733	0.72	3786.23	y(+3)	NFEDVGHSTDAR, NIPK(10,4)	
(HNHSKSTWLILHHK, FLEEHPGGEEVLR)	**Scan 1358.6**	**PyCXL loc 5,3**		**StavroX loc 5,9**		
theor m/z	exp m/z	ppm	intensity	ion	sequence	ch. known
428.73034	428.72944	2.08	7928.24	y(+2)	PGGEEVLR	
594.55990	594.55773	3.65	4122.93	b(+4)	HNHSKSTWLILHHK, FLEEH(5,3)	u
590.05726	590.05621	1.78	2816.43	b(+4)	HNHSKSTWLILHHK, FLEEH(5,3)-H2O	
419.72506	419.72508	0.05	2782.03	y(+2)	PGGEEVLR-H2O	u
(EQAGGDATENFEDVGHSTDAR, IPAIK)	**Scan 1599.4**	**PyCXL loc 13,5**		**StavroX loc 12,5**		
theor m/z	exp m/z	ppm	intensity	ion	sequence	ch. known
625.25256	624.99358	414.2	2762.1	b(+2)	EQAGGDATENFE	u
493.93587	493.93801	4.34	1493.57	y(+3)	DVGHSTDAR, IPAIK(1,5)	u
(GTVVVPTLDSVLYDNQEFPDPEKFK, YYTLEEIQK)	**Scan 2551.4**	**PyCXL loc 25,5**	**StavroX loc 23,5**			
theor m/z	exp m/z	ppm	intensity	ion	sequence	ch. known
329.80329	329.21766	1775.7	1290.69	b(+11)	GTVVVPTLDSVLYDNQEFP DPEKFK, YYTLEE(23,6)	**From 23,5
329.42950	329.21766	643.05	1290.69	y(+4)	K, YYTLEEIQK(1,5)	**From 25,5
(LYMAED, KVIKNVAEVK)	**Scan 1267.4**	**PyCXL loc 6,4**	**StavroX loc 5,4**			
theor m/z	exp m/z	ppm	intensity	ion	sequence	ch. known
613.36172	613.35919	4.11	28008.8	y(+2)	D, KVIKNVAEVK(1,4)-NH3	
621.87499	621.87242	4.14	19523.7	y(+2)	D, KVIKNVAEVK(1,4)	
608.27540	608.27302	3.91	15003.7	b(+1)	LYMAE	u

Table 4.2. Detailed analyses of cyt2E1-B5 cross-linked pairs from **Table 4.1** that had cross-link location discrepancies with StavroX. Below each cross-link pair are fragments that are listed that support the assignment of the location for PyXL over the location assigned by StavroX. Cross-link with MS2 scan 2551.4 is a special case. The following fragment listed could be assigned in two different ways, but the fragment assigned in 25, 5 results in a lower mass error.

Analysis of Proteins with Known Structures with Integrated Informatics Method.

In order to assess if the cross-links identified by the integrated informatics method can accurately reflect protein structure, two proteins for which high-resolution structural information is available were cross-linked and analyzed with the PyXL approach. The first protein, the TerS-A55 homodimer, has a published NMR structure [35] and the second protein, the *Escherichia coli* integration host factor protein (IHF), a DNA-binding heterodimer, has a published crystal structure [36]. The latter structure was obtained for the DNA bound complex. Each protein was cross-linked with both EDC and BS2. My criteria for a "true" cross-link are a precursor mass error less than ten ppm and at least 50% of the total signal intensity annotated.

Analysis of TerS-A55.

TerS A55 cross-linking and data analysis was performed in duplicate as described in Materials and Methods and a summary of the results is presented in **Tables 4.3** and **4.4** for BS2 and EDC, respectively. The identified cross-links had a relatively low fragment ion mass error (<300ppm) and the majority of peaks were annotated, except a few cases where only a third of the peaks were matched. Generally in these cases, the base peak was unlabeled, but the second most intense peaks were identified with low mass accuracy. Typically, several of these peaks represented a fragment ion that contained a cross-link. Most all of the cross-links identified here had a precursor charge state of 4+, although a few of the EDC crosslinks with repeat matches in the second replicate contained precursor ions with a 3+ charge. Repeat identification of cross-links in multiple different spectra occurred frequently, which increases the confidence in identifying true cross-links and minimizing false positives. Moreover, identification of the same

cross-link in two independent experiments provides additional confidence that this represents accurate structural information.

BS2 Replicate 1										
pepA	pepB	loc	aa #	tot match intensity	% match intensity	w avg ppm	# pks ID'd	prec ppm	prec charge	zscore
MEVNKKQLADIFGASIR	KQLADIFGASIR	[1, 1]	[nterm, K6]	9.15E+04	45.3	226.21	68	6.737	4	9.33
MEVNKK	GGGKGNEVLYDSAAVIK	[1, 4]	[nterm, K35]	7.26E+04	64.0	241.79	73	-4.315	4	22.23
MEVNK	TIQNWQEQGMPVLRGGGKGNEVLYDSAAVIKWYAER	[1, 18]	[nterm, K35]	3.44E+05	55.9	418.35	103	-2.782	4	8.52
MEVNKKQLADIFGASIR	KQLADIFGASIR	[5, 1]	[K5, K6]	1.28E+05	41.4	180.19	69	6.591	4	8.52
MEVNKK	GGGKGNEVLYDSAAVIK	[6, 4]	[K6, K35]	1.81E+05	69.6	130.82	117	-1.990	4	25.07
BS2 Replicate 2										
pepA	pepB	loc	aa #	tot match intensity	% match intensity	w avg ppm	# pks ID'd	prec ppm	prec charge	zscore
MEVNKKQLADIFGASIR	KQLADIFGASIR	[1, 1]	[nterm, K6]	3.46E+05	47.0	275.05	71	7.177	4	10.37
MEVNKKQLADIFGASIR	KQLADIFGASIR	[1, 1]	[nterm, K6]	3.46E+05	47.0	275.05	71	7.177	4	9.88
MEVNKKQLADIFGASIR	KQLADIFGASIR	[1, 1]	[nterm, K6]	3.34E+05	40.2	254.21	74	8.056	4	10.04
MEVNKK	GGGKGNEVLYDSAAVIK	[1, 4]	[nterm, K35]	1.23E+05	61.2	210.58	120	-1.100	5	15.49
MEVNKK	GGGKGNEVLYDSAAVIK	[1, 4]	[nterm, K35]	3.01E+05	49.9	308.96	88	-0.138	4	23.26
MEVNKK	GGGKGNEVLYDSAAVIK	[1, 4]	[nterm, K35]	3.82E+05	50.8	330.15	94	-0.523	4	24.31
MEVNKK	GGGKGNEVLYDSAAVIK	[1, 4]	[nterm, K35]	1.42E+06	49.0	299.62	89	-1.871	4	21.49
MEVNKK	GGGKGNEVLYDSAAVIK	[6, 4]	[K6, K35]	6.43E+06	66.8	108.54	111	-1.797	4	23.30
MEVNKK	GGGKGNEVLYDSAAVIK	[6, 4]	[K6, K35]	2.16E+07	70.7	140.13	98	-2.281	4	23.54
MEVNKK	GGGKGNEVLYDSAAVIK	[6, 4]	[K6, K35]	5.46E+06	69.2	140.66	116	-1.119	4	21.71
GGGKGNEVLYDSAAVIK	GGGKGNEVLYDSAAVIK	[4, 17]	[K35, K48]	8.33E+04	54.5	225.21	95	-1.912	4	11.77
GGGKGNEVLYDSAAVIK	GGGKGNEVLYDSAAVIK	[4, 17]	[K35, K48]	1.00E+05	57.0	183.75	98	-0.709	4	9.10
GGGKGNEVLYDSAAVIK	GGGKGNEVLYDSAAVIK	[4, 17]	[K35, K48]	6.47E+04	45.9	271.65	85	-1.912	4	8.33

Table 4.3. Summary of BS2 TerS A55 cross-links identified with the integrated informatics approach.

EDC Replicate 1

pepA	pepB	loc	aa #	tot match intensity	% match intensity	w avg ppm	# pks ID'd	prec ppm	prec charge	zscore
MEVNKK	TIQNWQEQGMPVLR	[5, 7]	[K5, E24]	9.29E+04	58.8	247.59	67	-1.313	4	13.66
MEVNKK	GNEVLYDSAAVIK	[6, 3]	[K6, E38]	1.50E+05	66.6	118.41	90	-0.003	4	31.02
MEVNKK	GNEVLYDSAAVIK	[6, 3]	[K6, E38]	1.12E+05	61.7	103.08	86	0.113	4	24.23
KQLADIFGASIR	WYAER	[1, 4]	[K6, E52]	8.38E+05	77.6	51.32	84	6.443	4	13.25
GGGKGNEVLYDSAAVIK	MEVNKK	[4, 2]	[K35, E2]	2.25E+05	64.4	112.78	102	-0.139	4	18.71
GGGKGNEVLYDSAAVIK	MEVNKK	[4, 2]	[K35, E2]	1.78E+05	62.1	124.89	101	-0.139	4	16.53
GGGKGNEVLYDSAAVIK	WYAER	[4, 4]	[K35, E52]	8.73E+05	69.8	154.39	115	-0.797	4	23.25
GGGKGNEVLYDSAAVIK	WYAER	[4, 4]	[K35, E52]	9.02E+05	72.1	160.07	104	-1.514	4	21.29
GNEVLYDSAAVIKWYAER	WYAER	[13, 4]	[K48, E52]	8.78E+05	76.5	124.66	82	-1.277	4	20.19

EDC Replicate 2

pepA	pepB	loc	aa #	tot match intensity	% match intensity	w avg ppm	# pks ID'd	prec ppm	prec charge	zscore
KQLADIFGASIR	TIQNWQEQGMPVLR	[1, 7]	[K6, E24]	7.34E+05	66.0	119.79	100	-2.091	4	13.97
MEVNKK	GNEVLYDSAAVIK	[6, 3]	[K6, E38]	1.49E+06	66.7	134.06	130	-0.698	4	22.54
MEVNKK	GNEVLYDSAAVIK	[6, 3]	[K6, E38]	1.40E+06	67.5	123.93	109	-1.046	4	21.74
MEVNKK	GNEVLYDSAAVIK	[6, 3]	[K6, E38]	5.60E+05	67.7	99.78	81	-1.257	3	27.71
KQLADIFGASIR	WYAER	[1, 4]	[K6, E52]	3.36E+06	77.7	50.08	88	0.551	4	15.15
KQLADIFGASIR	WYAER	[1, 4]	[K6, E52]	2.91E+06	71.9	67.61	80	0.513	3	29.03
KQLADIFGASIR	WYAER	[1, 4]	[K6, E52]	2.24E+06	69.0	82.80	77	-1.750	3	28.66
MEVNKK	WYAER	[6, 4]	[K6, E52]	4.64E+05	37.9	212.58	44	-2.908	4	8.02
MEVNKK	WYAER	[6, 4]	[K6, E52]	5.36E+05	64.7	67.01	42	-1.197	3	8.02
GGGKGNEVLYDSAAVIK	MEVNKK	[4, 2]	[K35, E2]	1.89E+06	70.3	133.19	95	-0.038	4	16.59
GGGKGNEVLYDSAAVIK	MEVNKK	[4, 2]	[K35, E2]	2.04E+06	71.8	160.29	103	-0.748	4	16.17
GGGKGNEVLYDSAAVIK	MEVNKK	[4, 2]	[K35, E2]	5.48E+05	76.4	113.08	106	-0.842	5	20.93
GGGKGNEVLYDSAAVIK	WYAER	[4, 4]	[K35, E52]	1.36E+06	62.8	255.73	119	-1.924	4	19.32
GGGKGNEVLYDSAAVIK	WYAER	[4, 4]	[K35, E52]	1.78E+06	62.7	216.56	103	-3.257	4	21.09
GGGKGNEVLYDSAAVIK	WYAER	[4, 4]	[K35, E52]	1.42E+06	67.5	111.90	74	-0.855	3	32.09
GGGKGNEVLYDSAAVIK	WYAER	[4, 4]	[K35, E52]	1.45E+06	67.8	124.31	91	-2.008	3	29.75
GNEVLYDSAAVIKWYAER	TIQNWQEQGMPVLR	[13, 7]	[K48, E24]	4.70E+05	59.4	164.49	94	-2.357	4	11.23
GNEVLYDSAAVIKWYAER	TIQNWQEQGMPVLR	[13, 7]	[K48, E24]	4.82E+05	60.0	178.77	108	-1.838	4	11.84
GNEVLYDSAAVIKWYAER	WYAER	[13, 4]	[K48, E52]	2.13E+06	74.9	116.52	92	-2.327	4	20.13
GNEVLYDSAAVIKWYAER	WYAER	[13, 4]	[K48, E52]	3.17E+06	74.8	136.76	101	-2.678	4	17.82
GNEVLYDSAAVIKWYAER	WYAER	[13, 4]	[K48, E52]	1.14E+06	77.3	72.64	76	-2.793	3	26.90
GNEVLYDSAAVIKWYAER	WYAER	[13, 4]	[K48, E52]	1.35E+06	73.7	74.34	76	-2.990	3	19.31

Table 4.4. Summary of EDC TerS-A55 cross-links identified with the integrated informatics approach.

Cross-links identified by the integrated method were then mapped onto the NMR structure of TerS A55 as shown in **Figures 4.4** and **Figure 4.5** for EDC and BS2, respectively. The physical distance limit of a cross-link was calculated based on the distance between the alpha carbon coordinates of the reactive residues plus the length of the cross-linker (22Å and 11Å for BS2 and EDC, respectively [5]).

	BS2*			distance in Å	
pos1	pos2	rep1	rep2	chain A	chain B
nterm	K6	1	3	12.9	12.9
K6	K35	2	3	18.8	16.5
nterm	K35	2	4	21.0	20.4
K35	K48	0	3	25.2	24.0

*max theor. C_α dist. ~22Å

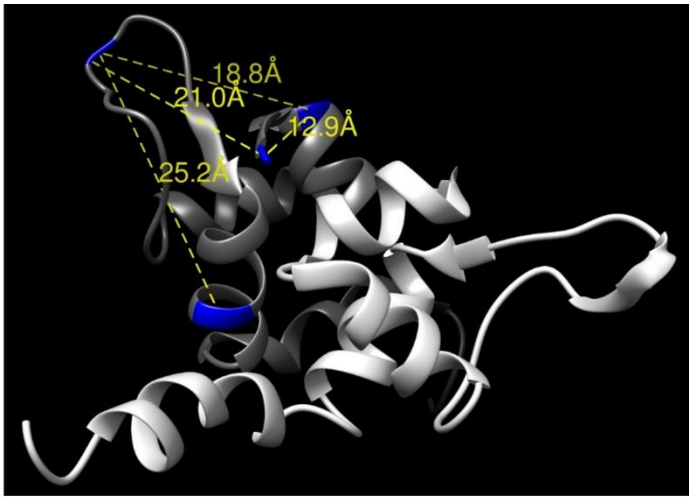


Figure 4.4. Summary of results from cross-linking TerS A55 with BS2. The table at left displays the spectral counts for the listed cross-link found in each replicate. The distances are estimated alpha carbon distances between the two resides (or N- and C-terminus where applicable) as determined by Chimera [37]. The picture on the right displays these crosslinks mapped on the NMR structure of TerS [35]. Blue residues are lysines or the N-terminus. The individual subunits are colored white (chain B) and grey (chain A).

	EDC*			distance in Å	
pos1	pos2	rep1	rep2	chain A	chain B
K48	E52	1	4	6.2	6.2
K6	E38	2	3	10.5	11.9
K48	E24	0	2	16.1	16.3
K35	E2	2	3	18.4	17.5
K6	E52	1	5	20.9	21.1
K35	E52	2	4	27.7	26.0

*max theor. C_α dist. ~11Å

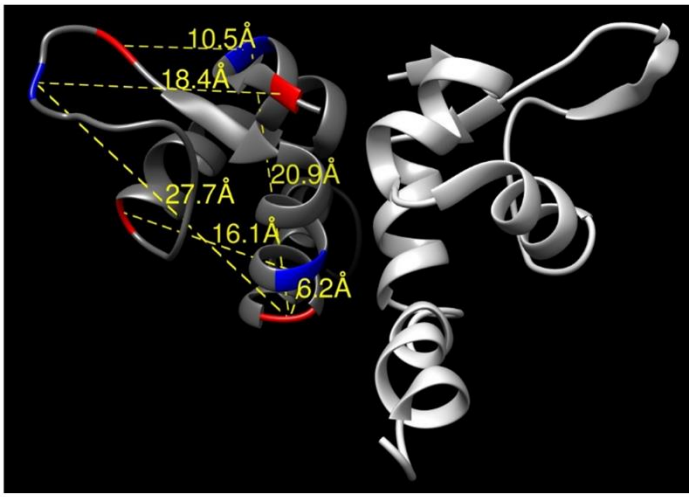


Figure 4.5. Summary of results from cross-linking TerS A55 with EDC. The table at left displays the spectral counts for the listed cross-link found in each replicate. The distances are estimated alpha carbon distances between the two residues (or N- and C-terminus where applicable) as determined by Chimera. The picture on the right displays these crosslinks on the NMR structure of TerS [35]. Blue residues are basic (lysines and N-terminus), and red residues are acidic residues (glutamate, aspartate, and the C-terminus). The individual subunits are colored white (chain B) and grey (chain A).

To assess the quality of the XLMS data, false positives and false negatives were assigned to the BS2 and EDC data sets and all additional data discussed below. False positives are defined as the percentage of cross-links that have been identified as true hits based on mass spectra data and metrics, but the measured distance based on the structural model is outside of the range of the cross-linking reagent. False negatives are defined as the percentage of cross-links that fall within the expected distance constraints of the cross-linking reagent, based on the NMR structure, that do not appear within the experimental data. The TerS-A55 BS2 data has a 25%

false positive rate (**Figure 4.4**) and a 65% false negative rate (**Table 4.5A**) based on the 22Å limit. The TerS-A55 EDC data has a 66% false positive rate (**Figure 4.5**) and an 85% false negative rate (**Table 4.5B**) based on the 11Å limit. Based on these metrics, it is clear that the BS2 XLMS cross-links identified are more consistent with the TerS-A55 NMR structure than the EDC data, and fewer expected cross-links are missed by BS2 than EDC. Both data sets have relatively high false negatives and false positives. The false negatives might simply be due to the fact that i) trypsin has difficulty cleaving the cross-linked protein, ii) some cross-links were simply not abundant enough to record a quality tandem mass spectrum, iii) these residues were simply not reactive enough (due to microenvironment effects) to form a cross-link, and/or iv) the cross-linking reagent was unable to access the cross-linking site due to polarity or solvent accessibility. In this case, false negatives are not due to the strict criteria for cross-link identification; all cross-links identified had quality mass spectra which pass the MS criteria, and all cross-links were identified at least twice among the two replicates.

A

BS2 expected cross-links		
pos1	pos2	distance (Å)
nterm	K48	11.6
nterm	K5	13.5
nterm	**K6**	**14.6**
K5	K35	15.8
K5	K48	16.2
K6	K48	18.2
K6	**K35**	**18.8**
nterm	**K35**	**22.7**

B

EDC expected cross-links		
pos1	pos2	distance (Å)
nterm	E2	4.8
nterm	D42	5.5
K6	D10	5.7
K48	**E52**	**6.2**
K5	E38	7.9
K5	D10	8.0
K35	E38	8.3
K48	D54	9.5
K48	D42	9.6
K5	E2	9.8
K5	D42	10.2
K6	**E38**	**10.5**
K6	E2	11.3

Table 4.5. Tables containing all theoretical in-range cross-links for BS2 and EDC based on the NMR structure of TerS-A55. Bolded entries indicate cross-links that were identified in the experimental data set.

Upon closer analysis of the false positives from both data sets, the observed false positive cross-links generally fall within known flexible regions within the TerS-A55 structure. When all structures of TerS-A55 obtained by NMR are displayed (**Figure 4.6**), flexibility is particularly apparent in the "wing" residues (residues 31-39) of the winged helix-turn-helix DNA binding motif. K35 which lies at the tip of the wing, is cross-linked in three out of the five false positive cross-links observed in the BS2 and EDC data, suggesting that these false positives reflect flexibility in the protein structure. Due to the variability in the TerS A55 structure observed in **Figure 4.6**, the distances for all outlier crosslinks were calculated for all structural models obtained by NMR and are summarized in **Table 4.6**; it becomes immediately apparent that the

BS2 outlier is actually comfortably in the 22Å range. Similarly, the EDC K35-E2 cross-link is much closer to the 11Å limit. Of the other three remaining EDC outliers, the cross-link from K48-E24 exhibits minimal distance variation, however, the minimum distance of 15.9Å is reasonably close to the 11Å limit. The remaining two EDC false positives, K35-E2 and K6-E52, lie 20Å or greater in distance, and there is no clear explanation for these cross-links. I interpret the data to indicate that the K35-E2 cross-link may arise due to further structural mobility in the wing region beyond what was detected by NMR. The K6-E52 cross-link is unusual, and may reflect and intermolecular cross-link between another TerS-A55 dimer. Overall, from a close inspection of the cross-link data, there are two conclusions to be drawn: i) the BS2 cross-links are more accurate and are all consistent with the NMR data, while EDC cross-links are more variable, and ii) the false positive cross-links may be explained by the known mobility of residues in these regions.

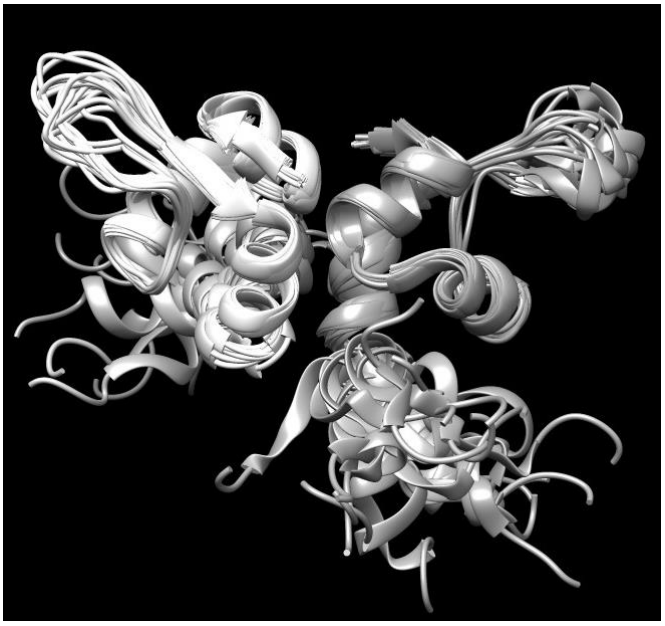


Figure 4.6. Illustration of all NMR overlays displayed for TerS-A55.

			chain A distance (Å)		chain B distance (Å)	
reagent	pos1	pos2	min	max	min	max
BS2	K35	K48	22.0	27.0	22.7	27.0
EDC	K35	E2	16.7	19.6	16.6	19.0
EDC	K35	E52	23.7	28.9	24.4	30.1
EDC	K48	E24	15.9	16.4	15.9	16.6
EDC	K6	E52	20.2	21.2	20.4	21.4

Table 4.6. Distance ranges in cross-link distances for each outlier cross-link across all NMR structures reported for TerS-A55.

Analysis of IHF.

The next XLMS control study utilized *Escherichia coli* Integration Host Factor protein (IHF). This protein is a basic heterodimer composed of alpha (11.4 kDa) and beta (10.6 kDa) subunits and provides a more complex control protein due to its increased size and abundance of cross-linkable residues. Cross-linking of IHF and data analysis was performed in duplicate as described in Materials and Methods and a summary of the results is presented in **Tables 4.7** and **4.8**. The cross-links assigned to the tandem mass spectra with good confidence, since the majority of the tandem mass spectra have greater than half of the total peak intensity assigned, a low average weighted fragment mass no higher than 350ppm, a low precursor mass error, and a high average matched signal intensity near 1E6 (**Table 4.7D** and **Table 4.8D**). The identified cross-links had modest reproducibility, where 43% of cross-links identified were found in at least two independent cross-linking experiments.

A

IHFA	IHFB	rep1	rep2	distance (Å)
K86	K27	1	0	14.846
nterm	K81	1	0	15.251
K20	K27	1	0	17.603
nterm	K27	1	0	20.746
K24	K81	1	0	22.912
K45	K75	2	0	24.266
K97	K27	1	0	24.472
K57	K3	2	2	26.823
K66	K75	1	1	32.045
K5	K65	3	0	43.306

B

IHFB		rep1	rep2	distance (Å)
K27	K3	1	0	8.413
K27	K20	13	2	10.717
K27	nterm	2	0	16.115
K84	K3	1	1	23.737
K69	K27	1	0	38.843

C

IHFA		rep1	rep2	distance (Å)
K24	K20	23	16	6.029
K20	K15	4	1	11.791
K45	K88	1	0	15.067
K15	K5	3	1	15.197
K24	nterm	2	0	18.121
K15	nterm	1	0	19.917
K45	K57	3	3	20.959
K66	K57	14	15	23.933
K15	K88	1	0	25.01
K66	nterm	1	0	30.198
K66	K97	3	2	31.692
K66	K88	2	1	35.278
K66	K45	2	2	38.217
K66	K24	1	0	41.631
K66	K20	1	0	46.495
K24	K20	23	16	6.029

D

Cross-link statistics							
avg cxl pep len	avg prec charge	avg tot match inten	avg % intens match	avg w avg ppm	avg # pks ID'd	avg prec ppm (abs)	avg zscore
12.0	4.11	9.21E+05	58.17	251.93	81.7	1.644	15.995

Table 4.7. Result from cross-linking IHF with BS2 (maximal theoretical Cα distance ~ 22 Å). **Panel A** shows intermolecular crosslinks between IHF alpha and beta, **Panel B** shows IHF beta only cross-links, and **Panel C** shows IHF alpha only cross-links, **Panel D** shows averaged statistics for the identified cross-links. Distances are estimated by Chimera using the IHF-DNA co-crystal structure (PDB 1IHF) [36].

A

IHFA		rep1	rep2	distance (Å)
E7	K5	1	3	5.421
E92	K88	2	1	5.839
D53	K57	2	4	12.618
E99	K88	4	3	18.795
E99	K86	7	1	22.588
D56	K66	1	1	27.093
E25	K57	4	0	34.073
D70	K45	2	0	37.268

B

IHFB		rep1	rep2	distance (Å)
E8	K84	1	4	26.912
E28	K65	0	1	38.175
E28	K69	0	1	39.971
E31	K69	0	1	34.82
D68	K84	2	1	43.207
D88	K75	0	1	32.318

C

IHFA	IHFB	rep1	rep2	distance (Å)
E92	K75	1	0	5.08
K15	E28	3	0	8.44
E39	M1	2	0	11.857
K5	E28	1	0	13.048
K5	E31	1	0	14.087
K45	E8	1	0	14.229
E99	K75	3	1	15.523
E39	K3	1	0	17.235
E92	K81	12	4	18.614
E99	K81	8	3	19.365
K97	E73	1	0	19.84
E92	K69	2	0	20.182
D70	K84	1	0	24.144
E69	K84	1	2	24.68
D70	K81	2	0	24.815
E99	K84	1	1	25.797
E92	K84	3	0	28.056
E92	K75	1	0	5.08

D

Cross-link statistics							
avg cxl pep len	avg prec charge	avg tot match inten	avg % intens match	avg w avg ppm	avg # pks ID'd	avg prec ppm (abs)	avg zscore
11.3	4.46	2.08E+05	57.83	345.86	73.4	2.180	8.84876

Table 4.8. Result from cross-linking IHF with EDC (maximal theoretical Cα distance ~ 11 Å). **Panel A** shows IHF alpha only cross-links, **Panel B** show IHF beta only cross-links, and **Panel C** shows intermolecular crosslinks between IHF alpha and beta. Distances are estimated by Chimera using the IHF crystal structure (PDB 1IHF) [36].

Overall, the XLMS data has a high rate of false-negatives (**Table 4.9**), which is similar to the TerS-A55 XLMS data. The TerS-A55 false-positive rate was low for BS2 and increased in the IHF data sets, yet the EDC false-positive rates remained high for both TerS-A55 and IHF. The IHF data shares a common feature with the TerS-A55 data: the false-positive cross-links

tend to localize in regions of high structural mobility in the proteins. Crystallography and NMR data of IHF and of a highly homologous DNA bending protein, HU, suggest that the wing-like protrusions from the main body of the protein, or arms, are highly flexible in solution, especially in the absence of DNA [38]. The crystal structure coordinates of IHF used to calculate cross-link distances presented in **Tables 4.7** and **4.8** was obtained from a crystal structure of IHF bound to DNA. Importantly, the mobility of the flexible arm regions is restricted in this complex (PDB1IHF) [36]. Within this context, all of the false positive BS2 cross-links and 55% (11/20) of EDC cross-links that lie outside of the maximal theoretical distance originate from the arm region of the IHF molecule, which may allow for cross-links to form beyond the theoretical reagent distance limit (**Figure 4.7**). The remaining 45% of the EDC false positive cross-links originate primarily from the C-termini of the alpha and beta subunit, but the reason for this is unclear (**Figure 4.7**). I note that the last few amino acids from C-terminal end of the alpha subunit are disordered in the crystal structure, providing an indication that these residues may be flexible in solution; however, there are no direct reports of high mobility in this region.

BS2	false positives	false negatives	total # expected x-links
IHF inter	50.0%	100.0%	47
IHF α intra	56.3%	68.2%	22
IHF β intra	40.0%	80.0%	15
average	48.8%	82.7%	

EDC	false positives	false negatives	total # expected x-links
IHF inter	83.3%	87.5%	16
IHF α intra	75.0%	92.3%	26
IHF β intra	100.0%	100.0%	22
average	86.1%	93.3%	

Table 4.9. Summary of false positive and false negative rates for the IHF XLMS data, based on the IHF structural data.

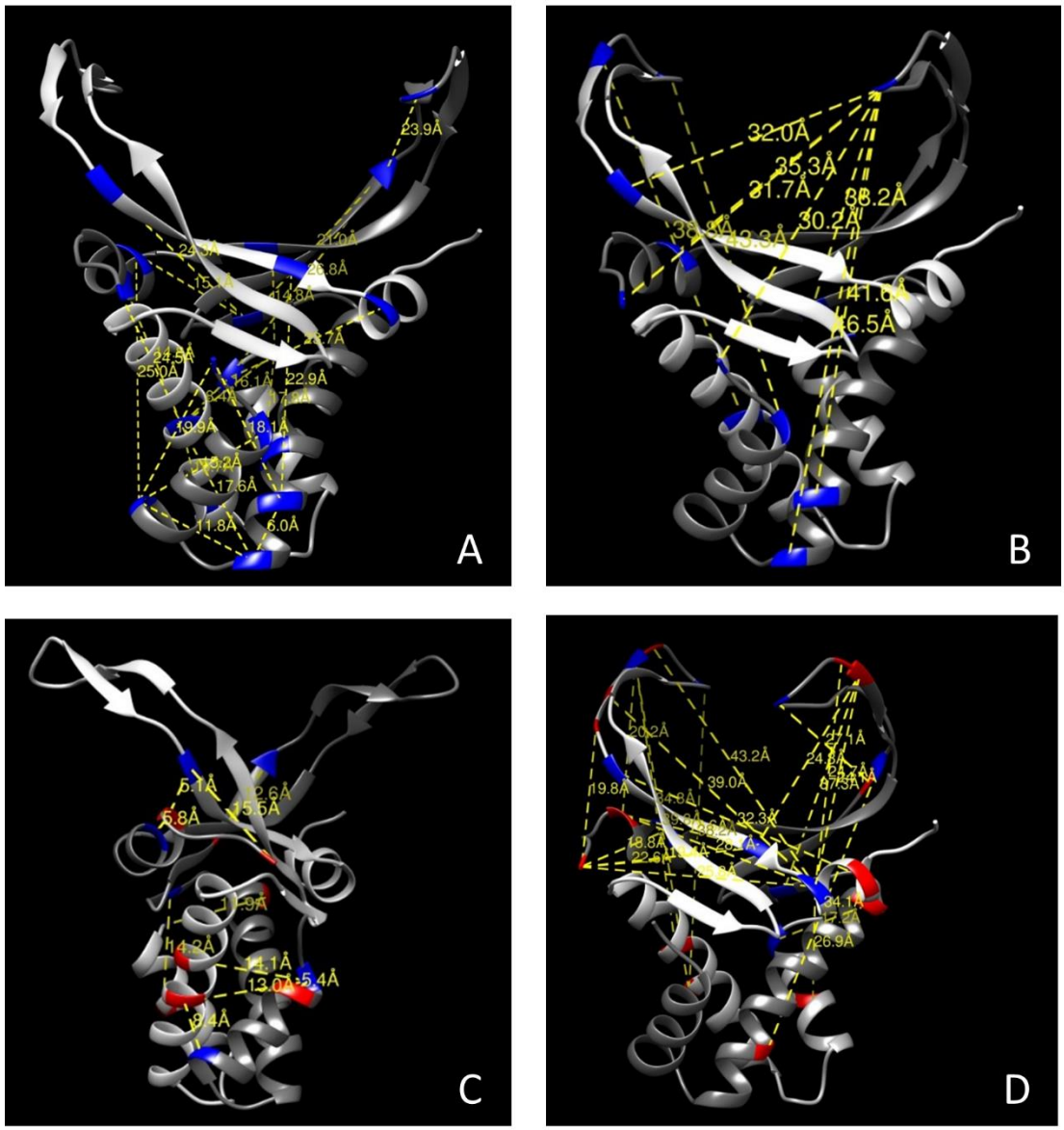


Figure 4.7. **Panel A**. BS2 cross-links for IHF within the 22Å limit. **Panel B.** IHF BS2 cross-links outside of the 22Å limit. **Panel C.** IHF EDC cross-links within the 11Å limit. **Panel D.** IHF EDC cross-links outside the 11Å limit. Structure shown is the IHF crystal structure (PDB 1IHF) [32].

The IHF data show that BS2 cross-links more accurately capture cross-links that are consistent with the static and flexible regions of IHF than EDC. IHF BS2 cross-links largely provide accurate distance constraints within the less-dynamic central core of the dimer and the false positives are consistent with flexible arm regions within the protein. In contrast, there are few EDC cross-links consistent with the static portions of IHF, and the majority of the cross-links are false positives in the arm regions and areas of the protein that are not documented to have high mobility. These observations indicate that more emphasis should be placed on BS2 cross-linking data and that EDC data is more unpredictable. These conclusions provide a foundation for XLMS interpretation of cross-linking studies of full-length TerS and TerL, which will be discussed below.

XLMS and Homology Modeling of Full-length TerS.

After analyzing XLMS data of proteins with known structures, I sought to utilize XLMS to validate a structural model for full-length lambda TerS. To achieve this, a homology model of full-length TerS dimer needed to be made due to the lack of C-terminal structural data of lambda TerS, which was guided by the TerS-A55 NMR structure. Ultimately, XLMS data will validate the homology model only if the false positive and negative rates are relatively low.

As mentioned in the introduction, terminase can assemble into two different assembly states: the protomer and the ring tetramer (Chapter 1, **Figure 1.3**). The protomer is the smallest repeated unit composed of one TerL subunit tightly associated with a dimer of TerS ($TerL_1$•$TerS_2$ heterotrimer). Four protomers assemble into a ring-like tetramer [($TerL_1$•$TerS_2$)$_4$], which represents the fully functional enzyme. Here I examine cross-linking of the isolated

protomer, since this simplifies data interpretation. In these samples, TerS only forms a dimer so it contains no higher-order interactions, and TerL cross-links must be intramolecular only.

Selection of the best homology model was straight-forward, due to the previously published TerS-A55 NMR structure. In order to choose the best model, I took all 15 possible models, examined the N-terminal portion for similarity to the TerS-A55 NMR structure, and then aligned two copies of the homology model with the NMR structure to generate a model of the lambda TerS dimer. The model which fit the best is shown in **Figure 4.8**. Out of all the models, TerS-A55 NMR matched the closest to this model. Also, when the two models were aligned with the NMR structure to generate the dimer, the resulting dimer had the least steric clash compared to other models. Interestingly, the model of best fit resembles the TerS Sf6 crystal structure fairly closely (**Figure 4.8B**). Nevertheless, this dimer model still only loosely reproduces the TerS-A55 NMR structure and significant deviation in the winged helix-turn-helix domains is observed (**Figure 4.8C** and **D**). This suggests that this model is not an accurate representation of the full-length lambda TerS dimer structure.

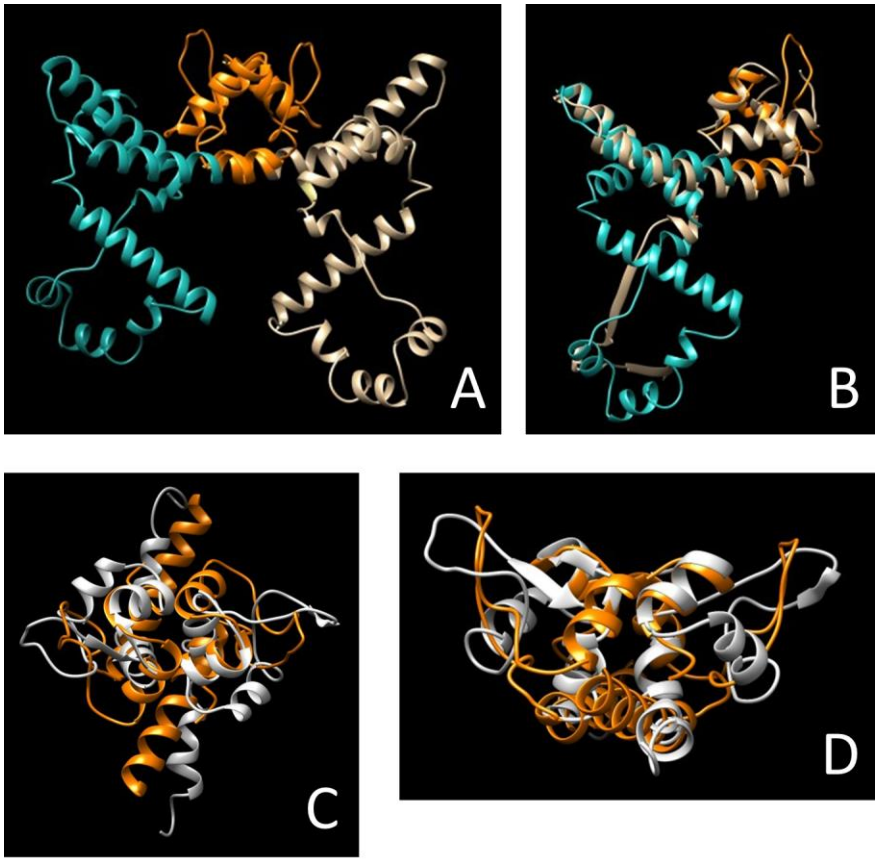


Figure 4.8. Homology model for lambda TerS. **Panel A** shows the lambda TerS dimer model selected. The N-terminal DBD (residues 1-55) was colored orange and the C-terminal domains are colored teal and light brown. **Panel B** shows a lambda TerS subunit overlayed on top of the TerS Sf6 crystal structure (light brown). **Panel C** shows a top view of the TerS-A55 NMR structure (white) overlayed with the N-terminal DBD residues of the lambda TerS dimer homology model (orange). **Panel D** shows a side view of **Panel C**.

Next, I proceeded to validate the full-length TerS model in **Figure 4.8A** with BS2 and EDC cross-linking data from terminase protomer ($TerL_1$•$TerS_2$). The data is summarized in **Figure 4.9**. Unexpectedly, none of the identified TerS-A55 cross-links were observed in the full-length protein (see Figures 4.4 and 4.5). This suggests that either i) there is some feature in the

protomer assembly that prevents the TerS-A55 cross-links from being recapitulated (i.e., altered conformation) or ii) these cross-link are present, but in much lower yield in relation to other cross-links and are not recorded by the mass spectrometer. Nevertheless, these data provide many distance constraints for model validation.

A

BS2

pos1	pos2	rep1	rep2	rep3	distance (Å)	
K48	K61	0	0	1	17.7	inter
K100	K100	6	5	5	22.6	inter
K35	K61	2	2	1	27.8	intra
K35	K152	0	0	2	43.2	inter

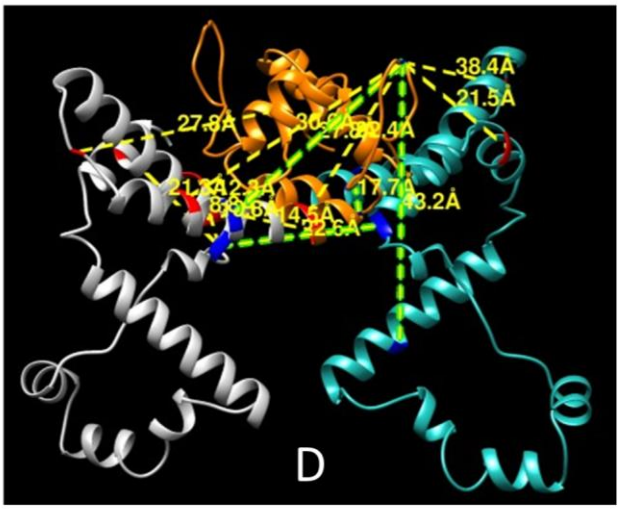
B

EDC

pos 1	pos2	rep1	rep2	rep3	distance (Å)	
K100	D95	2	0	0	8.8	intra
K61	D95	0	2	0	12.3	intra
K61	E98	2	0	0	13.8	intra
E52	K61	1	2	0	14.5	intra
K100	E178	2	1	0	21.3	intra
K35	E178	2	0	0	21.5	inter
K35	D54	1	2	0	22.4	intra
K6	E74	2	0	0	27.8	intra
K35	E65	0	2	0	30.6	intra
K35	D76	3	3	0	38.4	inter

C

	false positives	false negatives	total # expected x-links
BS2	50.0%	96%	20
EDC	90.0%	98%	28



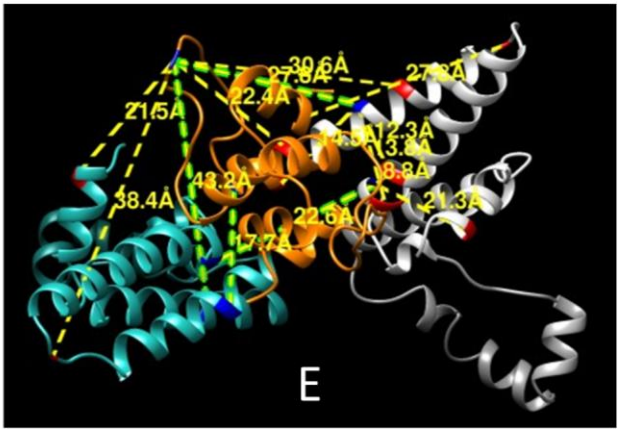
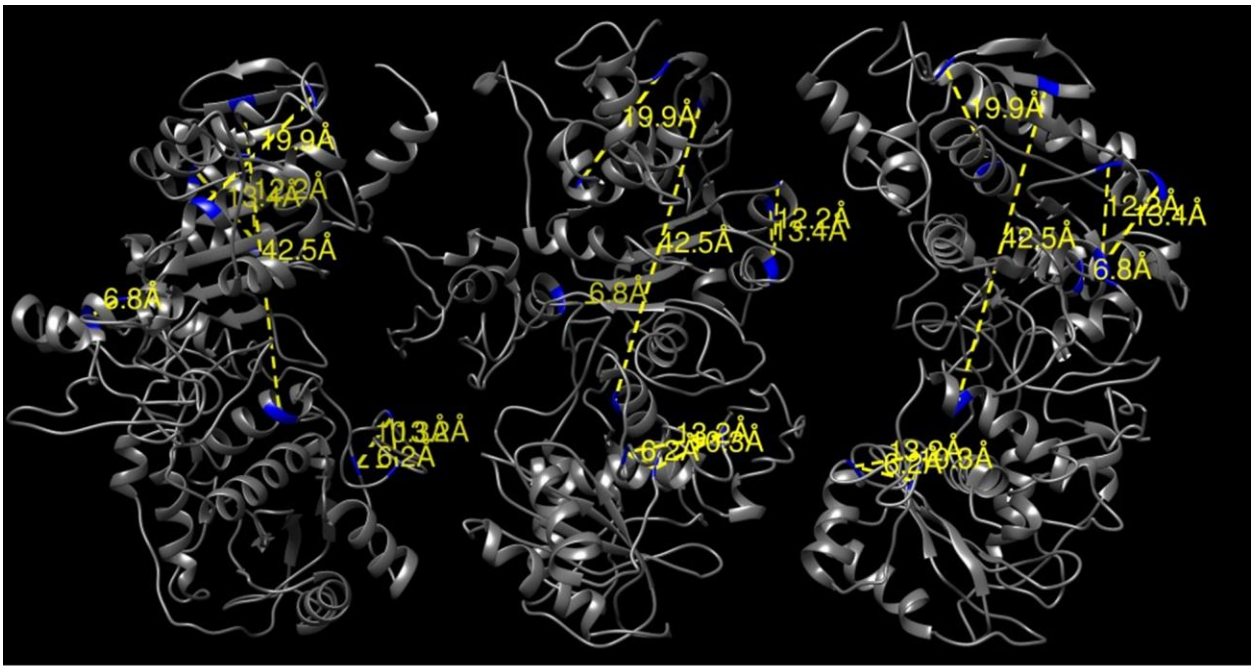


Figure 4.9. **Panel A.** TerS BS2 cross-links with distances based on the model presented in Figure **4.8A**. **Panel B.** TerS EDC cross-links with distances based on the model presented in Figure **4.8A**. **Panel C.** Summary of the false positive and negative rates for the EDC and BS2 TerS cross-links. **Panel D.** Side view of model shown in **4.8A** with all cross-links displayed. **Panel E.** Top view of model shown in **4.8A** with all cross-links displayed. In both **Panel D** and **Panel E**, BS2 cross-links are shown by yellow dashed lines bordered by light green, and EDC cross-links have no light green border around the dashed lines.

Overall, the cross-linking data do not seem to validate this model (**Figure 4.9**). While many BS2 cross-links lie within the 22Å limit, the K35 wing residue is nearly twice the allowed 22Å limit from the downstream C-terminal E152 residue. Also, BS2 has the highest false negative rate of all XLMS data sets discussed, which disfavors this model. While the EDC cross-links have previously been shown to be less predictable, these cross-links do not seem to validate the structure either. Only a small portion of the structure within approximately residues 50-100 generate cross-links close to the 11Å EDC limit. The rest of the false positive cross-links indicate that the wing DNA binding domain portion does not seem to be accurately positioned in relation to the remaining C-terminal portion of the model, and some of these cross-links have very high distances of 30Å or greater. While many of these cross-links involve the mobile wing region, this explanation is not sufficient to account for all false positives. In this case, it appears that homology modeling was unable to generate a model that was consistent with the cross-linking data.

XLMS and Homology Modeling of the Lambda Terminase TerL Subunit.

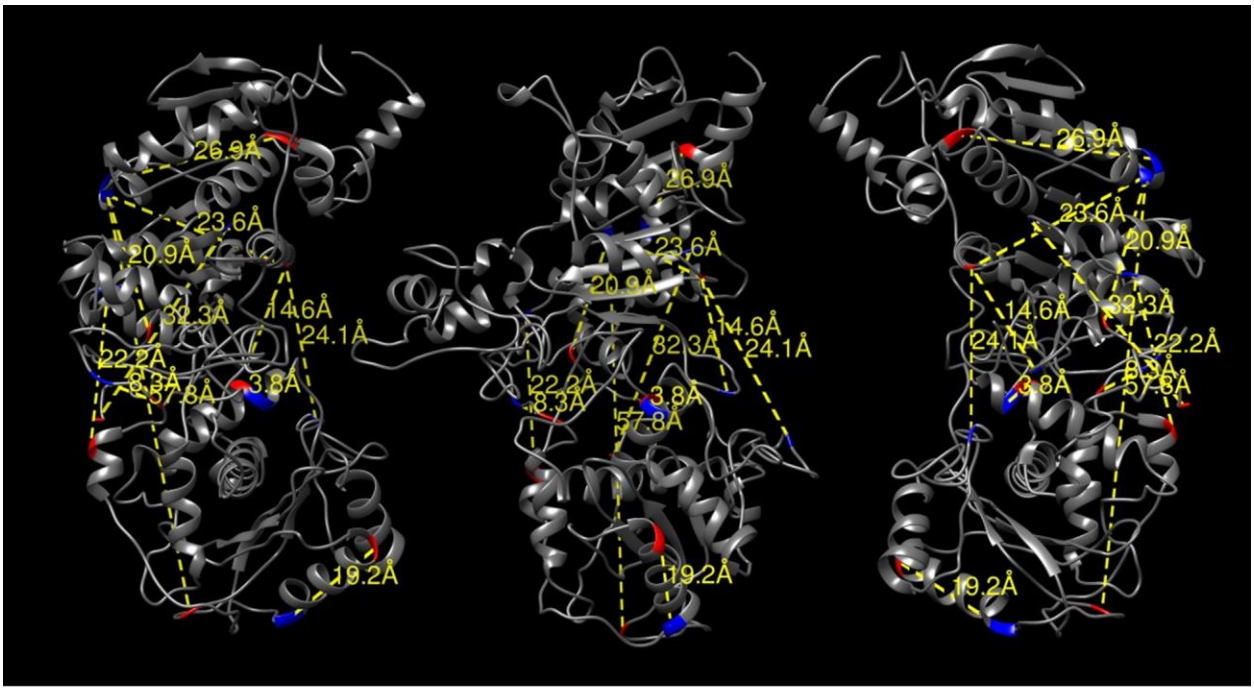
A similar approach used with TerS was applied to TerL. A candidate pool of TerL homology models were generated for TerL as specified in the Materials and Methods and the model that had the lowest average cross-link distances was the model selected. Ultimately, the best model was generated with no template specified, and the BS2 and EDC cross-links are shown (with distances) and displayed on the TerL homology structure that was the best fit (**Figure 4.10** and **4.11**). For these XLMS studies, only cross-links that had a spectral count greater than two for all replicates were considered to further enhance confidence in the cross-links identified.



Protomer, BS2					
pos1	pos2	rep1	rep2	rep3	distance (Å)
K490	K508	15	7	11	6.2
K216	K354	2	3	2	6.774
K490	K497	2	2	2	10.279
K170	K199	3	1	3	12.224
K497	K508	1	1	0	13.217
K165	K199	1	3	3	13.407
K138	K84	4	1	1	19.886
K148	K522	0	2	1	42.54

false positives	false negatives	total # expected x-links
12.5%	96.6%	203

Figure 4.10. TerL protomer BS2 cross-links summary using the homology model with the lowest average distance constraints. Three different orientations of TerL are shown with the cross-links highlighted where the protein has been rotated 45 degrees with respect to the longest axis. Distances are listed for each crosslink in the table. The BS2 distance limit is 22Å.



Protomer, EDC					
pos1	pos2	rep1	rep2	rep3	distance (Å)
K522	E523	4	0	0	3.8
K368	E371	2	0	0	8.3
K261	D67	0	1	2	14.6
E435	K447	3	0	0	19.2
K162	E245	0	1	1	20.9
K216	E376	2	0	0	22.2
K162	D67	0	2	0	23.6
K497	D67	1	1	0	24.1
K165	E95	1	1	0	26.9
K208	E641	0	2	0	32.3
K162	E456	1	1	0	57.8

false positives	false negatives	total # expected x-links
81.8%	98.9%	180

Figure 4.11. TerL protomer EDC cross-links summary using the homology model with the lowest average distance constraints. Three different orientations of TerL are shown with the cross-links highlighted where the protein has been rotated 45 degrees with respect to the longest axis. Distances are listed for each crosslink in the table and the theoretical distance limit is 11Å.

The BS2 cross-links provide significant evidence that the chosen model is a good representation of the lambda TerL structure. BS2 cross-links had a low false positive rate of 12.5% showing that the BS2 cross-links have strong agreement with the homology model (**Figure 4.10**). The positive hits all occur either within the ATPase or nuclease domain, which shows that these individual domains are modelled accurately. Despite the high false positive rate of 81.8% for EDC, (**Figure 4.11**), I would argue that the model still agrees with cross-linking data well since EDC cross-links have proven to be more unpredictable in control studies. The high false negative rate is not surprising, since the stringent requirements of a *bona fide* cross-link results in discarding in-range cross-links. I also note that the TerL protein in general suffers from poor trypsin digestion even in the absence of cross-linking (sequence coverage varies between 60-70% without cross-linking), which further contributes to false-negative results.

The false positive BS2 cross-link between K142 and K522 is indicative of mobility in the TerL structure (**Figure 4.10**). In the previously analyzed control structures, BS2 false positive cross-links were highly correlated to documented areas of mobility within the TerS-A55 and IHF structures. Therefore, it is highly like that this particular BS2 cross-link between K148 to K522 spans a distance greater than two times the 22Å limit is indicative of mobility within the TerL structure. Additionally, published TerL structural data concerning bacteriophage T4 and T7 terminase suggests that there is significant movement of these domains relative to each other [20, 39]. Therefore, I hypothesize that similar conformational motions are present in the lambda TerL subunit, which could explain this long-range cross-links. Specifically, the ATPase and nuclease domains rotate from their position in the model to afford a “closed” conformation of the protein, as suggested by cross-linking data.

While only one BS2 false positive cross-link indicates mobility within the TerL structure, several EDC false positive cross-links provide additional that supports the possibility of movement between the two domains (**Figure 4.11**). Four out of the eleven total EDC cross-links span between the ATPase and nuclease domains, and these can only form a viable cross-link if the domains move in relation to one another, resulting in a more compact conformation.

The false positive cross-links in BS2 and EDC which suggest conformational motion between the ATPase and nuclease domain of lambda TerL are illustrated in **Figure 4.12**. The BS2 false positive, which is more reliable than the EDC cross-links, suggest a rotation illustrated by **Arrow A** (**Figure 4.12**), where the two lobes of the roughly crescent-shaped protein rotate closer to one another in a "closed" or compact conformation. The motion suggested by **Arrow A** (**Figure 4.12**) would also serve to satisfy the 24.1Å EDC false positive as well. Additionally, the 22.2Å and 57.8Å EDC false positives suggest the motion illustrated by **Arrow B** (**Figure 4.12**), but considering EDC cross-links are less reliable, I believe that rotation shown in **Arrow B** is less likely (**Figure 4.12**). These motions will form extended and compact TerL conformations, and this finding is supported by the electron microscopy (EM) data obtained by Gabriel Lander shown in **Figure 4.13**. **Panel B** (**Figure 4.13**) most likely shows that TerL can adopt an extended conformation, because the appearance of four extended "strands" from the main EM density most likely corresponds to the 4 TerL subunits within the terminase oligomer. The disappearance of these extended strands as shown in **Panel A** (**Figure 4.13**), provides supporting evidence for a more compact lamda TerL conformation. These conformational changes have important implications for the terminase mechanism.

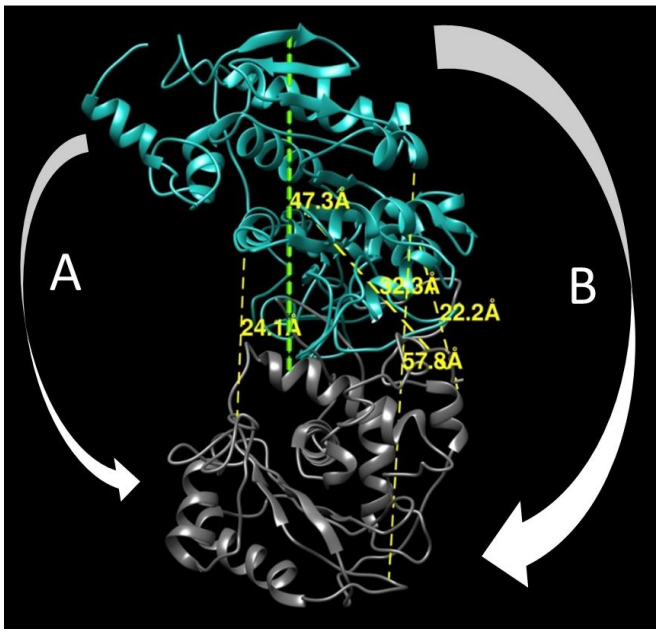


Figure 4.12. Diagram displaying putative motion within TerL. **Arrows A** and **B** indicate potential movement that may occur between the ATPase and nuclease domains so as to satisfy the distance constraints for the cross-link. The cross-link with a yellow dashed line bordered with a bright green represents the BS2 false-positive, all other cross-links are EDC false positives.

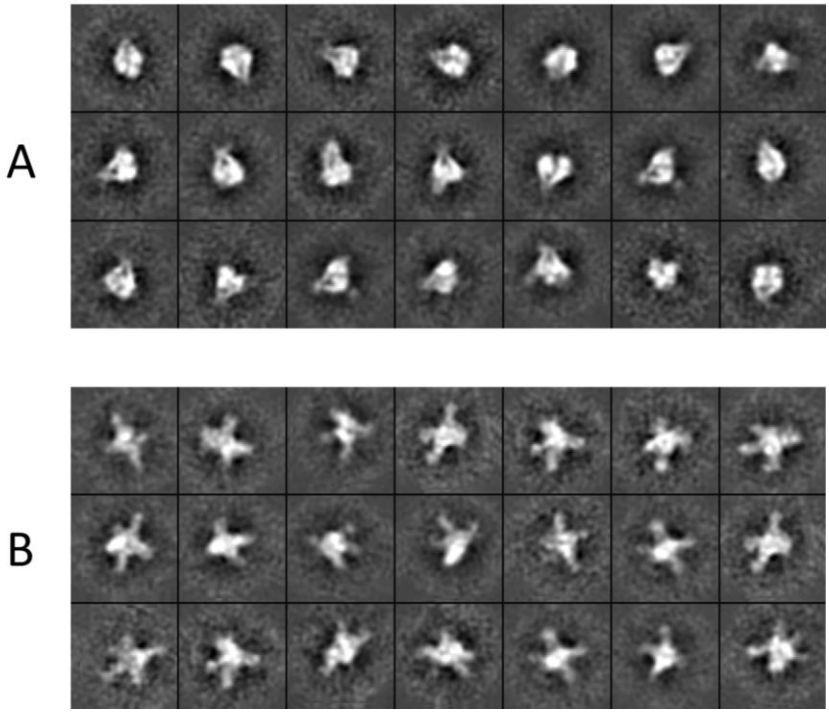


Figure 4.13. Electron microscopy data of lambda terminase collected by Gabriel Lander (unpublished results). **Panel A** shows terminase in a more compact structure, and **Panel B** shows 4 extended strands extending away from the terminase protein.

The motion proposed in **arrow A** of **Figure 4.12** is similar to that discussed in bacteriophage T7 TerL. Reconstructed Cryo-EM images of a ring-like T7 TerL in the oligomeric packaging motor appeared to show extended and compact conformations of terminase that involve rotation movements between the ATPase and nuclease domains [20]. Specifically, a crude estimation of the distance between the ATPase and nuclease domains in **Figure 5** of this publication [40] seems to approximately suggest a 20Å movement between the two domains, much like the BS2 false positive cross-link. This provides further supporting evidence for the conformational between the ATPase and nuclease domain inferred from the lambda XLMS data.

4.4 Conclusions

New developments in XLMS data analysis and the application of XLMS has the potential to increase the use and accessibility of this technique to other labs. In the past, the challenge of XLMS data analysis has lessened the use of this technique. My integrated approach described here addresses this issue in several ways. First, the software described here allows others to use and install a comprehensive and semi-automated XLMS analysis toolkit on their own personal computers that encompasses all steps from processing the raw data to cross-link identification. Secondly, the software provides the user detailed information and visual aids to allow the user to thoroughly judge the quality of their crosslink identifications. Thirdly, the approach maximizes the cross-link identifications as shown by the comparative analysis of the cyt2E1-B5 data.

XLMS data was also validated using known structures, which not only demonstrates the utility of this technique, but also allows for a more thorough understanding of what XLMS data means. It not only provides valuable distance constrain information for assessing tertiary and quaternary protein structure, but also has the potential to demonstrate structural plasticity within the protein. This aspect of XLMS is fairly new and unexplored. It was also demonstrated that BS2 cross-links proteins in a more predictable manner than EDC.

XLMS data in combination with homology modeling revealed valuable insight into lambda terminase. While homology modeling was unable to generate a valid lambda TerS dimer model that was consistent with cross-linking data, XLMS of lambda TerL suggested dynamic motion between the ATPase and nuclease domain. This motion has been proposed based on structural data of related bacteriophage terminases and has important mechanistic implications for the lambda terminase packaging activity.

Like many motor proteins, conformational changes are usually the linchpin for understanding the mechanism. For the lambda TerL packaging mechanisms, it is unclear how the movement between the two domains is translated into physical motion that packages DNA. A detailed putative mechanism for bacteriophage T4 terminase described by Draper *et al.* provides some insight as to how the motion between the domains translates into DNA packaging [39]. In this mechanism, it is postulated that ATPase activity drives a conformational change that results in an "open" and "tight" conformation, which are states where the ATPase and nuclease domains are further apart, or closer together, after a quasi-hinging action between the domains which drives DNA translocation. XLMS and EM data seem to corroborate these hypothesized conformations. Interestingly, a competing packaging mechanism has been proposed where motions between the domains are explicitly disfavored and instead invoke a localized conformational change within the linker domain to drive DNA translocation [19]. The XLMS and EM data from lambda TerL do not seem to support these claims due to the suggested conformational movement of these data and instead support the mechanism proposed by Draper *et al.* [39]. The lambda TerL XLMS data provides valuable evidence of these conformational changes within TerL since this methodology reflects native solution-based structural changes within a protein structure.

4.5 References for Chapter 4

1. Singh, P., Panchaud, A., Goodlett, D.R.: Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. *Anal. Chem.* **82**(7), 2636–42 (2010)

2. Sinz, A.: Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *J. Mass Spectrom.* **38**(12), 1225–37 (2003)

3. Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F.: Review / Perspective Probing native protein structures by chemical cross-linking , mass spectrometry and bioinformatics * 1–47 (2010)

4. Fritz, B.G., Roberts, S. a, Ahmed, A., Breci, L. et al.: Molecular model of a soluble guanylyl cyclase fragment determined by small-angle X-ray scattering and chemical cross-linking. *Biochemistry* **52**(9), 1568–82 (2013)

5. Singh, P., Nakatani, E., Goodlett, D.R., Catalano, C.E.: A pseudo-atomic model for the capsid shell of bacteriophage lambda using chemical cross-linking/mass spectrometry and molecular modeling. *J. Mol. Biol.* **425**(18), 3378–88 (2013)

6. Silva, R. a G.D., Hilliard, G.M., Fang, J., Macha, S., Davidson, W.S.: A three-dimensional molecular model of lipid-free apolipoprotein A-I determined by cross-linking/mass spectrometry and sequence threading. *Biochemistry* **44**2759–2769 (2005)

7. Zeng-elmore, X., Gao, X., Pellarin, R., Schneidman-duhovny, D. et al.: Molecular Architecture of Photoreceptor Phosphodiesterase Elucidated by Chemical Cross-Linking and Integrative Modeling. *J. Mol. Biol.* **426**(22), 3713–3728 (2014)

8. Dimova, K., Kalkhof, S., Pottratz, I., Ihling, C. et al.: Structural insights into the calmodulin-Munc13 interaction obtained by cross-linking and mass spectrometry. *Biochemistry* **48**(25), 5908–21 (2009)

9. Politis, A., Stengel, F., Hall, Z., Hernández, H. et al.: A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nat. Methods* **11**(4), 403–6 (2014)

10. Sinz, A., Kalkhof, S., Ihling, C.: Mapping protein interfaces by a trifunctional cross-linker combined with MALDI-TOF and ESI-FTICR mass spectrometry. *J. Am. Soc. Mass Spectrom.* **16**(12), 1921–31 (2005)

11. Liu, H., Zhang, H., Weisz, D. a, Vidavsky, I. et al.: MS-based cross-linking analysis reveals the location of the PsbQ protein in cyanobacterial photosystem II. *Proc Natl Acad Sci USA* **2014**(12), 1–6 (2014)

12. Müller, D.R., Schindler, P., Towbin, H., Wirth, U. et al.: Isotope-Tagged Cross-Linking Reagents. A New Tool in Mass Spectrometric Protein Interaction Analysis. *Anal. Chem.* **73**(9), 1927–1934 (2001)

13. Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F. et al.: Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol. Cell. Proteomics* **9**(8), 1634–49 (2010)

14. Panchaud, A., Singh, P.: xComb: A Cross-Linked Peptide Database Approach to Protein−Protein Interaction Analysis. *… proteome Res.* 2508–2515 (2010)

15. Gluck, F., Hoogland, C., Antinori, P., Robin, X. et al.: EasyProt--an easy-to-use graphical platform for proteomics data analysis. *J. Proteomics* **79**146–60 (2013)

16. Gaussier, H., Yang, Q., Catalano, C.E.: Building a virus from scratch: Assembly of an infectious virus using purified components in a rigorously defined biochemical assay system. *J. Mol. Biol.* **357**1154–1166 (2006)

17. Feiss, M., Catalano, C.E.: Bacteriophage lambda terminase and the mechanism of viral DNA packaging. *Viral Genome Packag. Mach. Genet. Struct. Mech.* 5–39 (2005)

18. Fuller, D.N., Raymer, D.M., Rickgauer, J.P., Robertson, R.M. et al.: Measurements of single DNA molecule packaging dynamics in bacteriophage lambda reveal high forces, high motor processivity, and capsid transformations. *J. Mol. Biol.* **373**(5), 1113–22 (2007)

19. Zhao, H., Christensen, T.E., Kamau, Y.N., Tang, L.: Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. *Proc. Natl. Acad. Sci. U. S. A.* **110**(20), 8075–80 (2013)

20. Daudén, M.I., Martiń-Benito, J., Sánchez-Ferrero, J.C., Pulido-Cid, M. et al.: Large terminase conformational change induced by connector binding in bacteriophage T7. *J. Biol. Chem.* **288**16998–17007 (2013)

21. Yang, Q., de Beer, T., Woods, L., Meyer, J.D. et al.: Cloning, expression, and characterization of a DNA binding domain of gpNu1, a phage lambda DNA packaging protein. *Biochemistry* **38**(1), 465–77 (1999)

22. Maluf, N.K., Gaussier, H., Bogner, E., Feiss, M., Catalano, C.E.: Assembly of bacteriophage lambda terminase into a viral DNA maturation and packaging machine. *Biochemistry* **45**15259–15268 (2006)

23. Sanyal, S.J., Yang, T.-C., Catalano, C.E.: Integration host factor assembly at the cohesive end site of the bacteriophage lambda genome: implications for viral DNA packaging and bacterial gene regulation. *Biochemistry* **53**(48), 7459–70 (2014)

24. Andrews, B.T., Catalano, C.E.: The enzymology of a viral genome packaging motor is influenced by the assembly state of the motor subunits. *Biochemistry* **51**9342–9353 (2012)

25. Singh, P., Shaffer, S. a, Scherl, A., Holman, C. et al.: Characterization of protein cross-links via mass spectrometry and an open-modification search strategy. *Anal. Chem.* **80**(22), 8799–806 (2008)

26. Roy, A., Kucukural, A., Zhang, Y.: I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**(4), 725–738 (2010)

27. Yang, J., Yan, R., Roy, A., Xu, D. et al.: The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**(1), 7–8 (2014)

28. Zhang, Y.: I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**40 (2008)

29. Sun, S., Gao, S., Kondabagil, K., Xiang, Y. et al.: Structure and function of the small terminase component of the DNA packaging machine in T4-like bacteriophages. *Proc. Natl. Acad. Sci.* **109**(3), 817–822 (2012)

30. Zhao, H., Kamau, Y.N., Christensen, T.E., Tang, L.: Structural and functional studies of the phage Sf6 terminase small subunit reveal a DNA-spooling device facilitated by structural plasticity. *J. Mol. Biol.* **423**(3), 413–426 (2012)

31. Götze, M., Pettelkau, J., Schaks, S., Bosse, K. et al.: StavroX-A software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom.* **23**(July 2011), 76–87 (2012)

32. Rasmussen, M.I., Refsgaard, J.C., Peng, L., Houen, G., Højrup, P.: CrossWork: Software-assisted identification of cross-linked peptides. *J. Proteomics* **74**(10), 1871–1883 (2011)

33. Yang, B., Wu, Y.-J., Zhu, M., Fan, S.-B. et al.: Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**(9), 904–906 (2012)

34. Mcilwain, S., Draghicescu, P., Singh, P., Goodlett, D.R., Noble, W.S.: Detecting Cross-Linked Peptides by Searching against a Database of Cross-Linked Peptide Pairs research articles 2488–2495 (2010)

35. De Beer, T., Fang, J., Ortega, M., Yang, Q. et al.: Insights into Specific DNA Recognition during the Assembly of a Viral Genome Packaging Machine. *Mol. Cell* **9**(5), 981–991 (2002)

36. Rice, P. a., Yang, S.W., Mizuuchi, K., Nash, H. a.: Crystal structure of an IHF-DNA complex: A protein-induced DNA U-turn. *Cell* **87**1295–1306 (1996)

37. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S. et al.: UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**(13), 1605–12 (2004)

38. Swinger, K.K., Rice, P.A.: IHF and HU: flexible architects of bent DNA. *Curr. Opin. Struct. Biol.* **14**(1), 28–35 (2004)

39. Draper, B., Rao, V.B.: An ATP Hydrolysis Sensor in the DNA Packaging Motor from Bacteriophage T4 Suggests an Inchworm-Type Translocation Mechanism. *J. Mol. Biol.* **369**79–94 (2007)

40. Sastry, S.S.: Identification of the Template-Binding Cleft of T7 RNA Polymerase as the Site for Promoter Binding by Photochemical Cross-linking with Psoralen † **2960**(96), 13519–13530 (1996)