ABSTRACT

STRUBLE, JACKSON. Structural Determination of Proteins by Chemical Crosslinking Analysis. (Under the direction of Dr. Michael B. Goshe).

This document outlines briefly the history of various techniques of crosslinking and their applications to protein studies. More modern methods are also discussed, with focus primarily on homobifunctional chemical crosslinking agents, especially disuccinimidyl-succinamyl-aspartyl-proline (SuDP). It is hypothesized that these reagents can be used to supply structural data about substrate protein molecules. Their binding is further hypothesized to be biased towards surface residues of the host proteins, which can be tested. The chemical crosslinking agent SuDP must first be synthesized, using solid-phase peptide synthesis methodology. Quantification of the synthesized crosslinking agent was performed using an ester quantification assay which exploits the ultraviolet absorbance of the dissociated succinimidyl rings in basic solution. The crosslinking reaction itself was performed on standard stocks of bovine serum albumin (BSA) at various ratios of crosslinking agent to protein, with acetic acid N-hydroxysuccinimide used in place of SuDP as a control. The digested product was analyzed with liquid chromatography coupled with multistage mass spectroscopy to produce raw data which was interpreted by a number of software suites, including StavroX and Chimera, the relative merits of which are also enumerated in this thesis, to yield a list of purported crosslinked polypeptide fragments. The sequences of these polypeptides were then analyzed with house-written scripts to collate the representation of each residue of BSA in crosslinking events. Deeper analysis was performed using a novel script to compare the average solvent accessibility of residues determined from the protein structure to the set of crosslinked polypeptides to assess the extent of identified crosslinks as a sole means to probe structure.

Structural Determination of Proteins by Chemical Crosslinking Analysis

by
Jackson Struble

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Biochemistry

Raleigh, North Carolina

2018

APPROVED BY:

_____	_____
Dr. Michael B. Goshe Committee Chair	Dr. Robert Rose
_____	
Dr. Guozhou Xu	

DEDICATION

For The Power

BIOGRAPHY

Jack Struble was born a gifted writer in Massachusetts, where he then grew up. He attended Macalester College for his undergraduate education, earning a bachelor's degree in biology. He moved to North Carolina in 2014 to pursue a higher level of more chemistry-focused biology, for which he had developed a keen interest. His publication output has remained prolific at an average of one master's thesis every 27 years.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1 – An Introduction to Chemical Crosslinking

I: Beginnings

Crosslinking reagents find wide use in many biochemical applications. Conventional use of crosslinking uses chemical agents to bond two or more interacting moieties. Making transient interactions permanent, within the time bounds defined by the experiment, or merely more lasting has value in permitting the study of the relationships between these interacting molecules. Many compounds have a crosslinking effect. Formaldehyde is perhaps the commonest of these; its use in crosslinking proteins to other-protein or DNA substrates is well explored, and it finds frequent use in studies of DNA- and RNA-binding proteins.

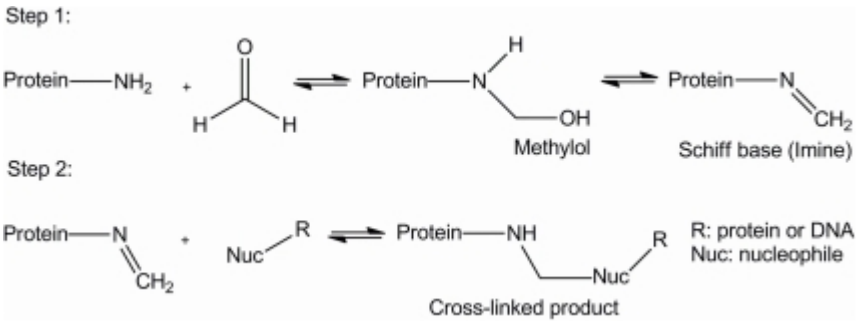
Crosslinking agents are also usable as a tool for probing protein surface structure. Many chemical crosslinking agents possess preference or specificity for certain substrate sites. Predictable attachment to primary amines or specific amino acid side chains are common, and can be exploited for information about the possessing protein's secondary structure. Upon fragmentation and mass spectrometry to assess the resultant peptides' sequence and the presence of characteristic mass additions from the crosslinking agents, some conclusions can be drawn about the structure of the originally crosslinked proteins from the apparent proximity of the crosslinked peptides. A variety of software exists to detect characteristic mass additions to single peptides, but many crosslinking agents possess more than one functional group and may link variably distant peptides. The mass spectrometry data in these cases can be more difficult to interpret, and software programs that can perform this analytical task are fewer and more specialized.

For greater analytical flexibility, crosslinkers with more than one functional group within the linker region may be classified as “cleavable” or “non-cleavable,” reflecting their respective tendency to break apart, or not, at a predictable location during manipulation of the ion in the gas phase during mass spectrometry analysis. For cleavable crosslinkers, each portion of the crosslinker should retain a characteristic mass due to the cleavage event, to assist in post-acquisition analysis by software. Tendency towards or against cleavage of a cleavable crosslinker can furnish data indicative of forces experienced between the two joined peptides.

II: An Historical Perspective

As a technique, crosslinking has evolved in form and function across many years. Simply described, crosslinking refers to the process of utilizing chemical means to hold two substrate moieties together. Inflicting greater permanency on transient interactions has many applications, and crosslinking in one or another form has been used to probe interactions too transient for easy study. Fleeting interactions between macromolecules in many permutations are susceptible to crosslinking by the appropriate agents. Proteins which bind nucleic acids, an important category encompassing the machinery of DNA replication, transcription, and other less-central cellular functions, are able to be preserved for study. Protein-protein interactions may also be linked for study. This is a boon to investigators of both transient interactions such as ligand-receptor pairs, which can be fleeting enough to defy other means of study, as well as to researchers investigating multi-subunit proteins for which dissociation of the constituent components can interfere with analysis. As a direct consequence of this usefulness, various crosslinking agents have been employed across many subfields of research over the years.

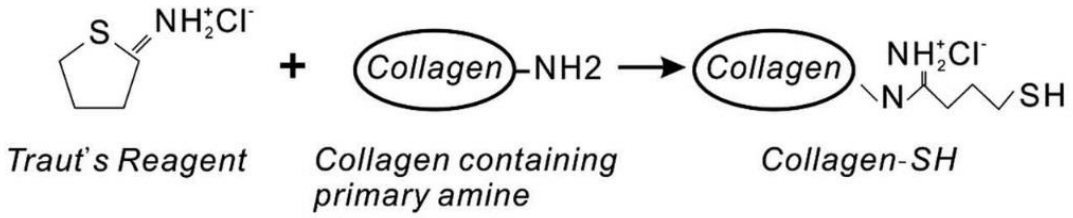
Use of crosslinkers as a tool to study biological systems predates the specific use of designed crosslinking agents such as disuccinimidyl-succinamyl-aspartyl-proline (SuDP) (Soderblom & Goshe 2006) as tools of intraprotein (peptide-to-peptide links within the same particle) structure determination. Formaldehyde's reaction with protein has been categorized for decades. In fact, knowledge of the mechanism of the formaldehyde crosslinking reaction predates definite knowledge of DNA as the molecule of inheritance (Hershey 1952). The compound effects its crosslinking by binding to the primary amine group of amino acid residues as well as basic groups. It lacks the characteristic twin n-hydroxy-succinimidyl rings of the synthetic crosslinking reagents, but formaldehyde still binds to two sites on either the same protein or different proteins and/or DNA molecules, as seen in Figure 1.1 (Fraenkel-Conrat, Cooper, and Olcott 1945). Use of the crosslinking and fixative properties of formaldehyde predate knowledge of the mechanism of its action by decades as well, as its protein-protein crosslinking capabilities have given it applications such as in tanning of leather for strength and long-life durability, and in production of textiles with stronger fibers. Compared to such agents as SuDP, formaldehyde is very small. Many modern dedicated peptide-to-peptide crosslinking agents consist of two functional groups separated by a linking region, often made of several peptides, of a significant, known length, one that can be varied according to needs by selecting different crosslinking agents, but formaldehyde by comparison links only two moieties in very close proximity.



(Figure from reference Wu et al. 2011)

Figure 1.1 Formaldehyde Crosslinking Mechanism: A genericized depiction of the mechanism by which formaldehyde effects its crosslinking of two disparate substrate protein moieties.

Expanding the use of crosslinking agents to a capacity as a tool to probe intraprotein geometry began with their application in studying multi-subunit proteins. In proteins with multiple subunits, crosslinking analysis can supply reliable data on which subunits are habitually found in proximity, which can be in turn interpreted into understanding of the organization of the protein. Proving this capability, crosslinking was used to investigate and ultimately elucidate the structure of the ribosome (Nygard and Nika 1982). The ribosome being a complex of discrete protein and RNA components, crosslinking is an eminently suitable investigative technique to discern its assembly, since the constituent components of each subunit are known with a high confidence. This reduces potential complexity of assembly dramatically, if taken relative to the potential complexity of a single polypeptide chain of equivalent mass. Though ribosomes are a nucleic acid-protein complex, formaldehyde was replaced with the alternative crosslinking agent 2-iminothiolane (Lambert, Boileau, Cover, and Traut 1983). Iminothiolane does not resemble contemporary crosslinking agents in that it possesses only one conventionally-functional end.



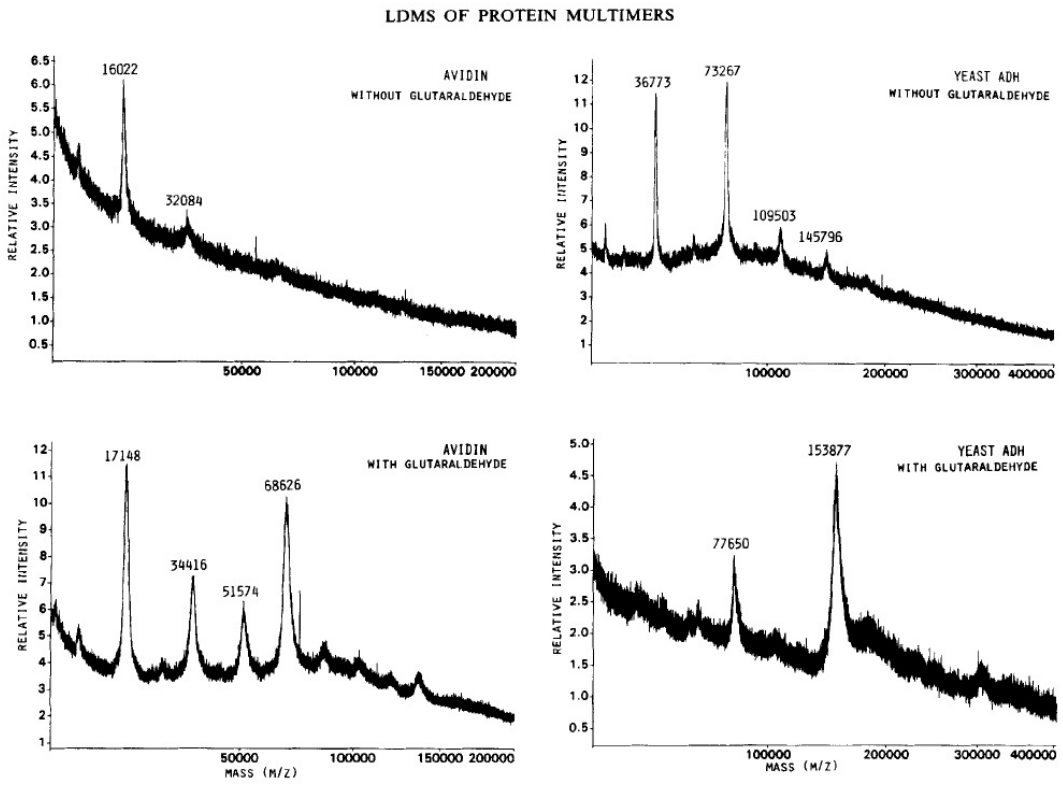
(Figure from reference Sun, Haipeng, et al. 2016)

Figure 1.2 Traut's Reagent: Also known as 2-iminothiolane, this reagent decircularizes to bind its substrate, as seen in this mechanism diagram. It must be treated with ultraviolet light to bind a second site, making it a heterobifunctional reagent.

This compound binds to polypeptides with lysine specificity. This sets it apart from formaldehyde, which is largely nonspecific to any specific residue or side chain, binding instead to universal primary amines. For further dissimilarity, crosslinking, i.e. binding to two distinct sites, with 2-iminothiolane requires more additional steps than crosslinking with formaldehyde or, later, SuDP and similar reagents; not only is initial mixture with the desired target crosslinking substrate required, but an additional application of ultraviolet radiation is needed to bind the other functional end of each 2-iminothiolane moiety to a second molecule (Wower, Meinke, and Brimacombe 1981).

In Lambert et al.'s 1983 experiment, identification of crosslinked ribosomal subunit components was accomplished by that traditional tool of crosslinking analysis, diagonal gel electrophoresis of the crosslinked subunits. Different permutations of ribosomal subunit combinations differ also in their motilities through the gel matrix, and each spot was assessed for relative mobility, which is directly related to mass. This data was cleverly combined to make hypotheses about which subunit constituents were proximal, an ingenious approach in a world without access to the accurate mass spectrometry available today (Lambert, Boileau, Cover, and Traut 1983). In 1986, a paper by Bragg and Hou describes a different method: use of disulfide-containing bifunctional crosslinking reagents reducible by 2-mercaptoethanol post-crosslinking applied to *Escherichia coli*'s F_1 adenosine triphosphatase from *Escherichia coli*. This method differs in that the crosslinked pairs of subunits were subsequently dissociated by the application of the 2-mercaptoethanol, rendering them individually identifiable by mass. Note that this method still requires *a priori* knowledge of the subunits present in the sample to be crosslinked, and relies on two-dimensional gel electrophoresis for the mass-determination step (Bragg and Hou 1986).

A discussion of crosslinking as effected exclusively by symmetrical bifunctional reagents, such as SuDP, and asymmetrical bifunctional ones like 2-iminothiolane excludes a special class of crosslinking reaction: the one in which the reaction is effected by light. It was discovered in the early 1960s that ultraviolet photons could induce crosslinking of proteins to nucleic acid substrates (Alexander and Moroson 1962) without even the small intermediary bond represented by small linkers like formaldehyde, though the mechanism by which the ultraviolet light pulses induced this crosslinking was not clear at the time. Subsequent investigation showed that photochemical crosslinking has a low chemical specificity; most amino acids of proteins may link with any nucleic acid. Preference against purine binding was, however, noted (Hockensmith et al 1986). Notably, the crosslinks that are produced when using ultraviolet laser pulses have no length. As they totally lack a chemical presence or agent of linking, the links add no distance between the two crosslinked species, which carves this method a niche unoccupied by any chemical crosslinkers; though formaldehyde and similar reagents add little distance between crosslinked substrate, they do form a link of a non-zero length. Larger reagents designed specifically for crosslinking have much larger linker regions; for example, SuDP adds 11.2 angstroms between crosslinked sites. Also noteworthy is that, while formaldehyde crosslinking has a storied history as a fixative, preservative process, ultraviolet crosslinking can be destructive to its target. Degradation of the moieties to be crosslinked can occur, especially to nucleic acids. This can reduce reliability somewhat, though crosslinking relies on economies of scale to a degree that the reduction is very manageable; destruction of a fraction of the sample to be analyzed is acceptable if enough sample remains intact and crosslinked. Thus, while other methods of crosslinking are generally easier, crosslinking by laser pulse does possess a specialized application wherever zero-length crosslinks could benefit an investigation.



(Figure from reference Farmer and Caprioli 1991)

Figure 1-3 Mass Spectrum of a Protein Multimer: A mass spectrum from an investigation of avidin and yeast alcohol dehydrogenase polymerization. Each labelled peak on the spectra represents a mono- or polymer, with the peak with the smallest mass-to-charge ratio (leftmost peak) being the monomer. Higher mass-to-charge peaks represent multimers, and the listed mass-to-charge value for each can be clearly seen as clean multiples of the monomer mass.

Analysis of intact multimeric proteins by crosslinking using the technique of mass spectrometry is also possible. Gentle ionization methods are required, but this technique is more similar to mass analysis by gel electrophoresis-based approaches, being a closer analog in principle than harsher, more-dissociative ionization method-based techniques. A crosslinked protein solution subjected to matrix-assisted ultraviolet laser desorption ionization undergoes minimal fragmentation. The method allows for no sequence-determining polypeptide fragmentation analysis but provides clear, unified peaks representative of the constituent monomers or crosslinked monomers of the intact protein complex, as illustrated in Figure 1-3 (Farmer and Caprioli 1991).

This technique, and gel-based methods, still have merit today, but cannot yet be easily scaled up into a reliable methodology for determining structure of monomeric proteins, as unfolded polypeptides may assemble in many more ways than a relatively small collection of subunits and require more sensitive mass spectrometry measurements for better characterization. Conceptually, though, this is still the foundational generalized process. More complex analysis is required to interpret the spectrometry data into a reasonable model of protein structure and assembly.

With the scope of crosslinker variety having been explored more and more fully, new and improving methods of analysis began to occupy the place of greatest innovation. The pre-eminent molecular mass measurement tool, mass spectrometry is an eminently applicable tool in the enterprise of structure determination by crosslinking, but the instrumentation and methodology associated with it have reached usefulness and relevance over the last decade compared to much earlier studies associated with gel methods. As discussed, relatively coarse mass spectrometry saw use in an analogous analytical role to gel electrophoresis when studying

multi-subunit species, but use of mass spectrometry brings a unique advantage: facilitation of study of linking between unidentified moieties. Gel methods require *a priori* categorization of the constituent subunits to be linked. Mass spectrometry entails analyte fragmentation, but the resulting product ion spectrum can be interpreted to provide a sequence, which can in turn be queried against an external database to identify the host protein, protein subunit (Komolov et al. 2017), or in more sophisticated methods, peptide location (Rappsilber, Siniossoglou, Hurt, and Mann 2000). Obviously, this requires mass spectrometers of a certain level of accuracy, but using crosslinking data to identify polypeptide portions of the host molecule (and provide structural constraints between those portions) and a sequence of that molecule from a database are an analytical combination that is increasingly being explored to its full potential, which exceeds the potential for structure determination possessed by simpler, multimeric-focused methods.

But if finer determinations of structure are to be made, accurate protein sequences are also critical. For a fully-sequenced protein, an identified polypeptide fragment's position in the one-dimensional sequence can be determined. For such an identified polypeptide in such a sequenced protein, after crosslinking, software can provide a specific location on the protein of the identified (crosslinked) fragment, but the mass addition of the linking region of the crosslinker itself, even if confined to a certain residue like lysine, increases the complexity of interpretation beyond what the algorithms used by programs to identify peptide sequences (e.g Sequest and Mascot (Perkins, Pappin, Creasy, and Cottrell 1999), which have existed for some time longer) can interpret (Sinz 2003). Software capable of analyzing the mass spectrometric data of crosslinked peptide species in this way and others have been a primary locus of sophistication in crosslinking analysis. Currently, many software packages specifically produced

for the purpose of crosslinking analysis exist, with varying suites of features. As part of my thesis research, a number of these were compared and considered; see Chapter 3.II. It is clear that much of the advancement that is to come in the progression towards true fine structure determination through crosslinking analysis will come less from ever more subtle methods and instruments, but more from software capable of interpreting the more complex data furnished by fragmented crosslinked peptide species and its insights into protein structure and dynamics.

Another alternative approach to crosslinking is use of isotopically-coded crosslinking reagents. A methodology developed relatively recently, dedicated bifunctional crosslinking agents are synthesized with heavy isotopes of constituent elements. These specially-synthesized, heavier crosslinkers are used to link a portion of protein sample, and an analogous linking reaction is performed on a similar portion of sample using unmodified (light) linker. The samples are combined and analyzed via mass spectrometry. Probable crosslinker-containing polypeptide fragments are in this way marked by a characteristic doublet with separation equal to the difference in mass between the unmodified and heavy-isotope-marked variants of the crosslinking agent used and can be readily identified by software following mass spectrometry (Seebacher et al. 2006). Care should be taken to ensure that the respective masses of the unmodified, light crosslinking reagent and the isotopically-coded, heavy crosslinking reagent are significantly different, as this increases ease of doublet identification; very close doublets can be difficult for software to interpret as different from naturally-occurring isotopic peak patterns. To this end, many substitutions are preferred over single substitutions, rendering common hydrogen the obvious choice of target, in addition to comparative ease and low cost with which it can be exchanged with deuterium, in comparison with the challenges of using heavy carbon. One additional advantage of usage of isotopically-coded crosslinking analysis is that it is not a strict

replacement of existing assays; it can be used with extant crosslinkers. Rather than offer a fundamental difference in the way a crosslinking assay is conducted, it merely facilitates detection of crosslinked polypeptides, and thus complements existing workflows instead of offering an entirely different path of analysis.

Thus, crosslinking analysis is a family of methods with a record of effectiveness, historically, that is currently benefitting from a wave of advancement in software and measurement tools. Modern mass spectrometry and software analysis promise to help it deliver information with greater speed and flexibility than it has in the past, and the process of crosslinking itself is usefully simple. There is complexity associated with choosing mass spectrometry parameters, but mixing a sample with crosslinker is itself a fundamentally uncomplicated process. The onus of difficulty has shifted to the writers of the software which interprets the data. It may not surpass other methods for sheer suitability in discovering structure of analytes, but it has advantages over other methods to rapidly probe protein interactions and may be suited to confidently identify structural dynamic change as well.

III: Specific History of SuDP

Reliable analysis of three-dimensional protein structure, including their interactions, by using chemical crosslinking agents first requires reliable reagents to effect the crosslinking. A capacity to link amino acid side chains is a necessity, as the majority of crosslinks generated after proteolytic digestion of the crosslinked protein will be intrapeptide links, but the structurally informative ones will be those generated by interpeptide crosslinks, particularly when trying to determine the interactions between two different proteins or domains within a protein. Based on the design of the reactive moiety, it is possible that no nucleic acids participate

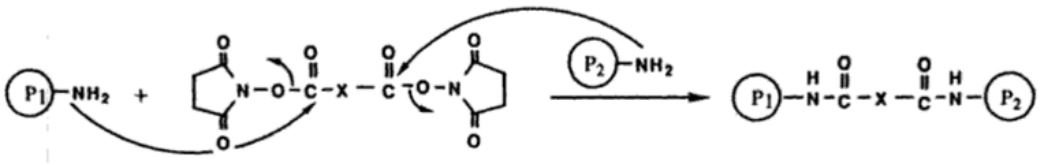
in the actual process of crosslinking, although study of conformational change in nucleic-acid binding proteins is one application of topological analysis by crosslinking. In this case, the binding of the chemical crosslinking agents would not be between the nucleic acids and the protein, but would provide evidence of shifting conformation of the protein between the nucleic acid-bound and unbound configurations.

Disuccinimidyl-succinamyl-aspartyl-proline (SuDP) is a representative of the protein-to-protein clade of chemical crosslinking agents. It possesses two functional succinimide groups connected by a short linker comprised of aspartate, proline, and succinate, and preferentially attacks aminoacyl R-chains, most typically lysine residues. The distance between its functional ends in extended conformation is a well-categorized 11.2 angstroms. Crosslinking with the substrate occurs at the terminal epsilon amino nitrogen of the side chain of lysine and the carbonyl carbon of the ester of the crosslinking agent. The succinimide rings flanking the functional esters on each end of the crosslinker (SuDP or a structural relative) serve as very ready leaving groups following attack by the amino group of the lysine side chain that is to participate in the link (Wong 1991). SuDP is readily synthesized using solid-phase peptide synthesis methods followed by esterification.

Very similar in functional application to bis-N-hydroxysuccinimidyl crosslinking reagents such as SuDP are di-isocyanates and di-isothiocyanates. These consist of a benzyl center portion linking two functional ends, in this case the cyanate or thiocyanate groups as opposed to esters and succinimidyl rings. In the mechanism for the reaction providing this group's functionality, which is largely analogous to the one providing bis-N-hydroxysuccinimidyl reagents', the nitrogen of the substrate amino group acts as the nucleophile in an attack on the electrophilic carbon of one of the functional ends of the crosslinking agent.

The crucial difference is that in the reaction of the di-isocyanates and di-isothiocyanates, there is no leaving group. The displaced electrons form a bond with hydrogen instead. The net result is the same, however: two linked peptides (Wold 1972).

Another cousin clade of crosslinking agents uses sulfonyl chlorides. This group effects crosslinking via a very similar mechanism to bis-N-hydroxysuccinimidyl and di-isocyanyl/ci-isothiocyanyl reagents; nucleophilic attack occurs at a functional carbon of one of a matched set of distal groups, but in the case of sulfonyl chlorides, the leaving group is chlorine (Dickson 1998). Still more esoteric crosslinking agent groups exist, such as dialdehyde- and bis-imidodiester-based linkers, but bis-N-hydroxysuccinimidyl reagents benefit from a well-established method of synthesis: the solid-phase peptide synthesis method (described in Chapter 2.II) is well-studied, reliable, and has established protocols for use in the application of crosslinker synthesis. This made it the method of crosslinker synthesis of choice for the experiments described.



(Figure from Wong 1991)

Figure 1-4 Generalized Homobifunctional Ester Crosslinking Reagent Mechanism of Action: P_1 and P_2 in this case represent different proteins. The NHS rings are the leaving groups.

IV: Comparison of Crosslinking with Traditional Methods of Structural Determination

The traditional tool for high-resolution protein three-dimensional structure determination is x-ray crystallography. Crystallography is capable of providing high-fidelity structural data for a wide variety of proteins, with additional functionality to be used even in cases where the protein in question possesses a characteristic bound substrate or co-acting factor, which can be co-crystallized. In contrast, crosslinking agents preferentially bind the surface of a protein in disproportion, and the structural data they do provide is lower-resolution, less authoritative, and more subject to chemistry of the crosslinker and the mass spectrometry acquisition and post-acquisitional analysis. The settings used in any given structural determination assay may be catalogued and transmitted to confirming experimenters, but software quality in mass spectral analysis is a factor to be considered always, whereas crystallographic data permits less variability of analysis.

In many cases, the main experimental determinant in whether crystallography can be used to characterize a protein's structure is the facility of that protein for forming suitable crystals. The principal advantage of protein structure analysis by crosslinking is the absence of such a restriction, rendering it fit for use with a greater diversity of proteins. The capacity of crosslinking agents to bind any protein is profound; SuDP, for example, preferentially binds lysine residues, and lysine is common across virtually all proteins, having structural usefulness. This is not to say that proteins with relatively low lysine representations could not exist; however, other crosslinking agents exist with different R-chain specificities, and a protein entirely devoid of R-chains that can be crosslinked would be highly improbable to encounter naturally, and this also disregards crosslinkers such as formaldehyde which do not act exclusively on the R-chain portions of the peptides they crosslink. Thus, virtually every protein

is susceptible to some kind of crosslinking, whereas entire clades of proteins are entirely intractable by crystallography. This is a great advantage to the crosslinking mass spectrometry approach.

The comparisons favorable to crosslinking over crystallography end here, however. Crystallography produces a diffraction pattern of "spots" where a crystal is struck by a beam of X-rays when held in a certain orientation and rotated. This pattern is highly precise and can be interpreted into a three-dimensional structure with similar precision. Though it is an arduous process, it is relatively objective and produces consistent results, and it yields a full three-dimensional model of the analyte. This degree of reliability (the X-ray scatter pattern is complex but relatively non-subjective) and richness of data (a full three-dimensional model with precision to the angstrom) is invaluable. Crosslinking can offer evidence that two residues are found within a certain distance of each other, but interpreting this into a three-dimensional structure is not possible with the mathematical precision with which a scatter pattern can be interpreted. Further complicating crosslinking analysis is the fact that detection of crosslinked polypeptides in mass spectra from crosslinked analytes comes with some degree of uncertainty, but a greater source of unreliability is the fact that crosslinking and fragmentation during subsequent mass spectrometry analysis tend to be stochastic processes. A specific crosslink between two loci on a protein may be detected with high confidence in one trial, and not at all on the next, due to the timing of the fragmentation event, which sets it apart from X-ray diffraction data collection.

Software performs the analysis of crosslinking data produced by mass spectrometry, and differences between software suites will be discussed later, but it is also important to note that scrupulous cataloguing of the software and algorithms used to perform the calculation of analysis is of greater import to crosslinking as it heavily influences reproducibility. Less analogous

cataloguing is required for crystallographic methods as the interpretation of the scatter pattern is more established, though crystal growing is an arduous process which can see benefits in terms of increased reproducibility if the process used is meticulously catalogued.

Crystals do, however, come with problems as they increase in size, setting them apart in another way; larger molecules often crystallize more poorly or with greater difficulty and offer lower-resolution images than do smaller molecules. Many biological molecules that would be the subject of study are complex macromolecules and relatively large, but the effect can still bias crystallography against larger proteins. Crosslinking is not subject to such biases. Further crystallization-based analysis requires sufficient protein to form a crystal, which is in many cases significantly more material than the femtomoles of analyte which is sufficient for crosslinking analysis. Multiple trials benefit crosslinking analytics and their stochastic nature in an outsized amount next to more deterministic X-ray crystallography, but the difference in required protein quantities can remain vast, providing another thin niche for crosslinking, though not irrelevant, given the rarity or difficulty of acquisition of some analyte species.

Another disadvantage to analysis by crystallography is that it captures the analyte protein in a static crystalline matrix and is therefore unsuited to analyzing active (mobile) conformations of these proteins. As crosslinking analysis takes place in solution, it is subject much less to this consideration.

Multidimensional nuclear magnetic resonance (NMR) spectroscopy is another common extant method of determining three-dimensional structure of biomolecules that is effective on large biologically-derived macromolecules such as proteins. NMR spectroscopy uses powerful magnetic fields to measure the resonance of analyte nuclei. The local magnetic environment of each constituent nucleus is subject to influence and shielding by nearby atoms, and the local

magnetic environment exerts influence, in turn, on the resonant frequency of the analyte nuclei. The effect of shielding is small, measured in parts per million, but can be measured. Two nuclei in physical proximity spatially, but separated by many covalent bonds (as tends to occur in large, convoluted structures such as proteins) also exert influence on each other via the nuclear Overhauser enhancement effect (Ganten and Ruckpaul 2006).

Nuclear magnetic resonance data, often in combination from multiple scans, can be deconvoluted to provide accurate three-dimensional structures of biomolecules. The process is, for proteins, somewhat labor-intensive, but capable of producing structures with a high degree of accuracy and certainty. The deconvolution process is more complicated than the one that makes sense of the scatter pattern produced in X-ray crystallography of protein crystals, and more ambiguous, and therefore more reliant on special software for rapid analyzing of the structure.

Complexity is cumulative in NMR spectroscopy determination of structures, and this precipitates one of its greatest weaknesses; the larger a protein analyzed, the more complicated its signature spectrum becomes. More complicated spectra are increasingly difficult to deconvolute with the required precision. Increased instrument precision, often facilitated by larger magnets, can ameliorate the problem to a degree, but never eliminate it, and magnets above a certain size become unwieldy, expensive, and difficult to operate and maintain. Even with excellent equipment, the upper limit at which NMR spectroscopy can typically provide consistently solvable spectra for a protein rests at approximately 30 kDa. Even for proteins with multiple subunits which can be analyzed and considered independently, this limit can be confining. One way to bypass this limit is to utilize nonuniform sampling, which exploits the tendency of samplings of data with shorter evolution decays to have higher sensitivities, to increase the effective sensitivity of the data. Existing methods of analysis via Fourier transform

algorithms require equally-spaced samplings, rendering this method incompatible with these traditional algorithms, but newer algorithms that can be used on unevenly-spaced sampling data sets can perform the deconvolution successfully, allowing full realization of the increase in sensitivity. In this way, the limit of usefulness can exceed 80 kDa (Cavanagh et al. 2010, 770) (Mobli and Hoch, 2014). Proteins, of course, commonly exceed this, with the largest having mass in excess of 3,800 kDa, but the majority are not so large, and NMR spectroscopy is useful and reliable enough that expanding its small window by any significant degree has value simply by dint of its effectiveness in analysis cases where it can be brought to bear.

Despite any shortcomings, NMR spectroscopy is consistently capable of much clearer deconvoluted structures of proteins than crosslinking analysis affords. The method is reliant on magnetic fields of very high strength, and the equipment capable of generating such large fields is cumbersome, and often requires the magnets be cooled by expensive and finicky liquid-helium systems. In comparison, crosslinking analysis requires less specialized equipment. Lending nuclear magnetic resonance spectroscopy an advantage over X-ray crystallography is the absence of the unpredictably-fulfillable requirement of crystallizability. Size of the analyte protein is the more-salient factor in tractability, and size is more easily assessed than crystallizability.

One more-recently described method of categorizing protein structures that stands as an increasingly viable alternative to NMR spectroscopy and X-ray crystallography for detailed structural analysis is cryo-electron microscopy. Cryo-electron microscopy uses flash-freezing of very small (sometimes as small as a single-particle) quantities of sample to render them suitable for imaging through electron microscopy. By cooling the samples rapidly enough that water ice forms amorphously as opposed to the usual crystalline form that occurs when the ice is permitted to nucleate, damage to biological samples is prevented and the analyte is trapped and can be

imaged with the microscope. Electron microscopy's maximum resolution is still limited by the wavelength of the electrons it makes use of, a limit which is harder to bypass than the less-strict one established in NMR spectroscopy, which is susceptible to trickery. Unfrozen single particles are typically not susceptible to visualization by electron microscopy due to Brownian motion (if in liquid solution) and flexibility of the molecule's structure (Jonge and Ross 2011), but, frozen in ice, they may be held still long enough to provide an image.

The image produced by cryo-EM is unique among outputs provided by NMR spectroscopy, X-ray crystallography, and crosslinking mass spectrometry analysis in the small amount of interpretation it requires to yield useful information. The method returns its data in the form of an electron micrograph, which, unlike any of the above methods with the possible exception of X-ray crystallography for experts, yields data that is available visually without interpretation. The micrograph is a picture of the structure. X-ray crystallography produces a scatter pattern, which, while related to the structure rather directly, is not usually interpretable as such by the unaided eye. NMR spectra and crosslinking mass spectrometry data both are almost entirely abstract and absolutely dependent on processing. But even this data, a two-dimensional picture, still does not reach its full informative usefulness with regards to the three-dimensional structure of the protein to be analyzed without some processing.

As with other methods, cryo-EM post-acquisitional analysis is generally performed by dedicated software and uses Fourier transform-based calculations. This software streamlines the task of assembling the two-dimensional picture data from the electron microscope into functional three-dimensional models, the kind that are the ideal end product of other structure determination methods. SSEHunter was the premiere software solution for this; it has since been folded into the Gorgon molecular modeling system (Jiang et al. 2008). The deconvolution is less intensive and

more deterministic than the one associated with NMR spectroscopy. The limitations of the cryo-EM technique lie largely in sample preparation. The proteins to be studied must be suspended in water such that single particles can be photographed individually, and the technique is most useful on larger molecules, often multi-subunit proteins, and complexes. The flash-freezing process is less arduous and unreliable than the process of precipitating crystal growth for X-ray crystallography, and typically faster as well. Thus, cryo-EM has significant advantages over both NMR spectroscopy and X-ray crystallography, but possessed of limitations of its own, the most restrictive of which is its ineffectiveness on smaller proteins, a weakness which is difficult to circumvent by improving instrument technology and technique.

It would be difficult to argue that crosslinking analysis can provide better or more or comparable structural information about proteins that can be studied by X-ray crystallography, cryo-EM, or NMR spectroscopy, but its advantages in other areas confer it usefulness. A method usable on structures intractable by crystallography (or, depending on the size of the protein, cryo-EM and/or NMR spectroscopy) is a niche unto itself, but, assuming previously-synthesized crosslinkers, the crosslinking process, including post-mass spectrometry data analysis, is relatively rapid and, importantly when considering the technique for routine analysis, less skilled-labor-intensive.

Some harmonization of the solvent conditions to the protein analyte under scrutiny is required, however, which can increase complexity. The chemical crosslinking agents used are soluble and borne to their substrates by the solvent, so the analyte protein substrate must be well solubilized, which for some proteins can be an additional challenge worthy of consideration. As usual, many factors contribute to a protein's solubility, and these characteristics can vary across multiple analyte proteins to be studied by crosslinking. Ambient pH plays a strong role in

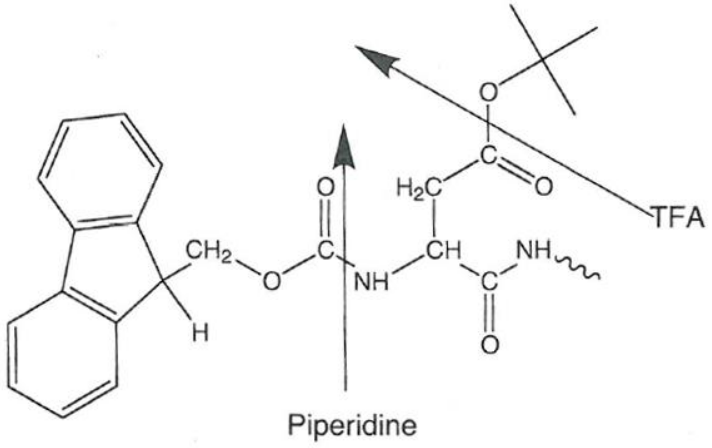
determining protein solubility as it affects protonation of amino acid residues containing dissociable groups, which in turn frequently play a critical role in maintaining solubility and protein structure in its native conformation (which is generally the form under study). Other solvent factors which affect proper folding should be taken into account as well, such as ionic strength and temperature. The vast majority of proteins which the method might in any likelihood be applied to do share biological origination, and cellular conditions vary within a relatively narrow range of pH and salt concentrations. Nonetheless, significant differences in solubility and native pH are exhibited by many proteins. Additionally, proteins with special conditions, such as transmembrane proteins, may be largely incompatible with solvation in water, due to their native forms requiring support from their surrounding lipid environment and possessing intrinsic crucial regions of hydrophobicity.

Crosslinking analysis is unlikely to surpass crystallography in fidelity and, therefore, overall usefulness, but it has potential as a secondary method or fallback method when crystallography fails, and should not be neglected.

CHAPTER 2 – Methods

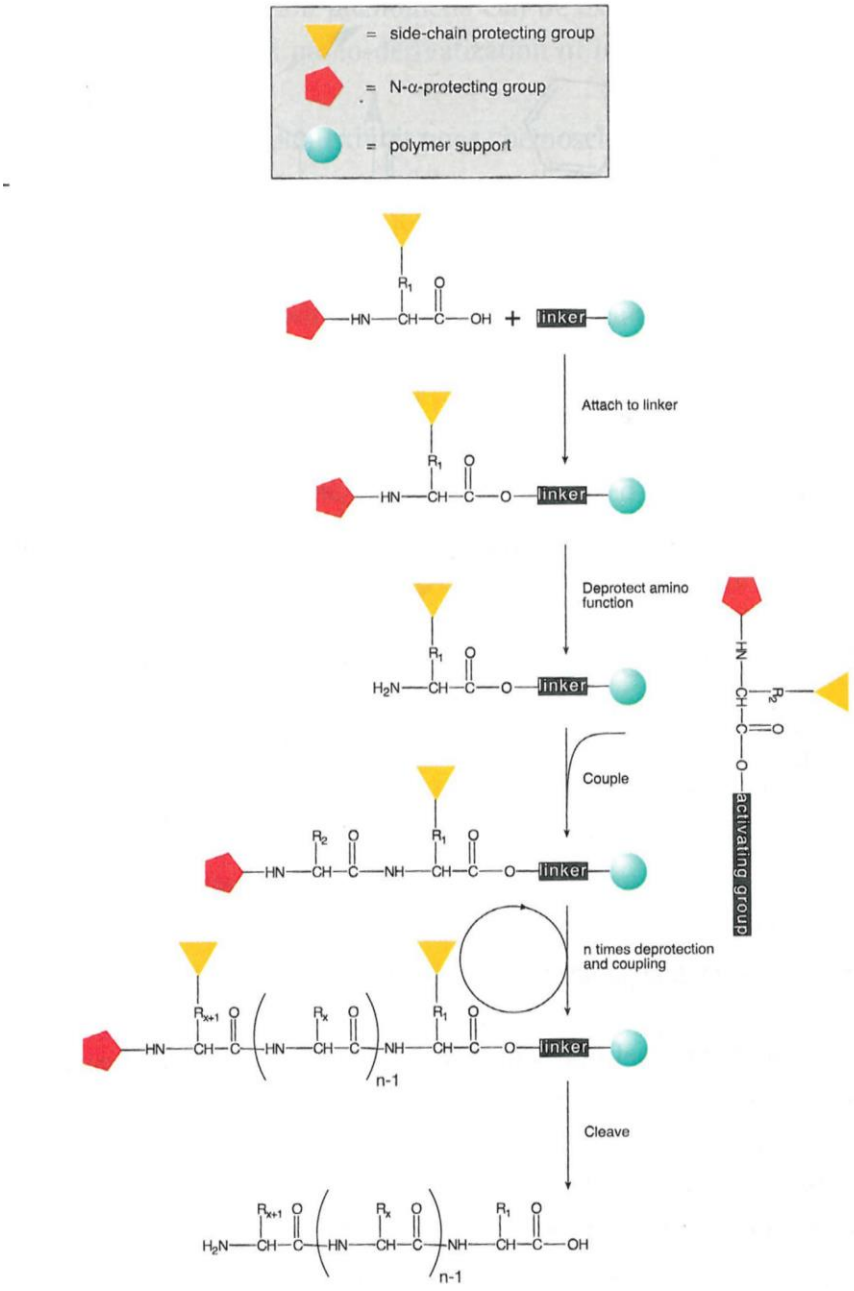
I: Synthesis of the SuDP Crosslinker

The crosslinker utilized was disuccinimidyl-succinamyl-aspartyl-proline (SuDP). This reagent cannot be purchased from conventional biochemical suppliers, and therefore was synthesized chemically using solid-phase peptide synthesis (SPPS). Synthesis proceeds C-terminal to N-terminal, in opposition to biochemical translation, with protecting groups shielding the N-termini and being removed when the next amino acid is to be added. Additionally, some R-chains were protected, and protected by different groups than were used to protect the N-termini, and had to be removed in a discrete process, using different reagents. The protecting groups used were selected for their ease of deprotection, and for the ease and safety of handling the reagents used to deprotect them. Variant, related disuccinimidyl crosslinking agents, such as disuccinimidyl-succinamyl-valyl-proline can be synthesized which do not have any R-chains which need protecting. SPPS is sufficient to produce the polypeptide backbone of the linkers used in these experiments, but these backbones are insufficient as crosslinking agents. The created backbones must be esterified, as it is the ester substituents that act as ready leaving groups and confer the agent its effectiveness. Post-synthesis purification and verification steps by high-performance liquid chromatography and mass spectrometry, respectively, were also performed to ensure production of the correct and complete crosslinker, as incomplete or partial esterification were potential challenges, given that two ends of the backbone had to be esterified.



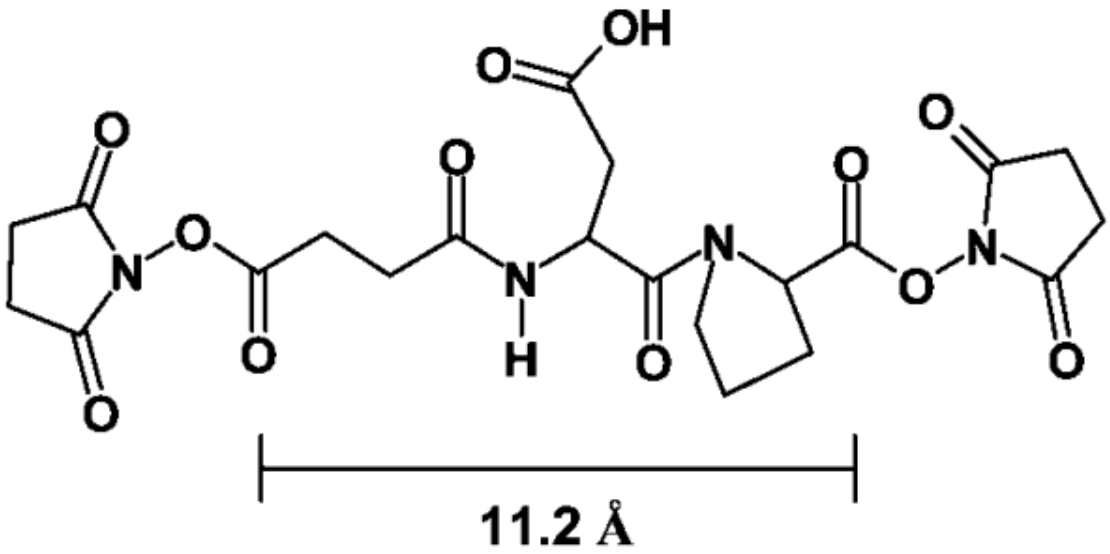
(Figure from Novabiochem 2008)

Figure 2.1 Piperidine Deprotection of Fmoc Amino Acid during Solid-Phase Synthesis: Solid-phase peptide synthesis as conducted on resin involves deprotection of Fmoc protecting groups from the nascent polypeptide's N-terminus, which was effected with a piperidine wash, which cleaves the Fmoc group as shown.



(Figure from Novabiochem 2008)

Figure 2.2 Solid-Phase Polypeptide Synthesis: Detailed diagram of the solid-phase synthesis process. The nascent polypeptide is bound by the linker region to immobilizing resin. R-chains of amino acids which must also be protected are capped with a protecting group distinct from the one which caps the N-terminus of the chain. In the work described in this thesis, that group is a tert-butyl group.



(Argo, Shi, Liu, and Goshe 2015)

Figure 2.3 SuDP Diagram: The SuDP molecule, like other homobifunctional chemical crosslinking agents, consists of a central linker region separating two NHS ring groups. It is the linker region which binds to the targets, and thus it is the length of the linker region which sets the distance constraint between the two bound residues.

A total of 500 mg of resin-bound proline (Fmoc-Pro-Novasyn) were mixed with 20 mL of dimethylformamide (DMF) in a peptide synthesis vessel with a sintered glass filter and stopcock. Periods of mixing were alternated with draining of the solvent through the filter and addition of 5 mL aliquots of 20% piperidine solution in DMF. Addition of 218 mg of non-resin bound aspartate with an N-terminal and R-chain protecting group (Fmoc-Asp (OtBu)-OH, Novabiochem) and 81 mg of hydroxylbenzotriazolemonohydrate (HOBT) in 2 mL of DMF and 90 µL of diisopropyo carbodiimide (DIC) was the next step. The glassware was gently rocked for 90 min, then rinsed through the filter three times with DMF. A volume of 10 mL of 20% piperidine solution was added and the glassware gently rocked for 20 min more before being rinsed twice with DMF, and these steps were repeated once. To add the third component to the growing peptide chain, 100 mg of succinic anhydride in 2 mL of DMF was added to the vessel, which was rotated for 1 h, then rinsed through with DMF twice, then thrice with dichloromethane (DCM). To detach the nascent crosslinker from the resin, five repetitions of the addition of 5 mL of DCM with 50 µL of trifluoroacetic acid (TFA) with mixing for 2 min followed by draining into a collecting vessel were performed, as were two rinses into the collection vessel with 5 mL of DCM. The volume in the collection vessel was reduced from 35 mL to 10 by blowing a stream of nitrogen over it, at which point 10 mL of water was added to the collecting vessel, which was agitated briefly, to ensure that the crosslinker was contained in the top aqueous layer.

The aqueous layer was collected and reduced through vacuum centrifugation only to be brought up in 1 mL of a solution of 0.1% formic acid (FA) in water, then purified via gradient high-performance liquid chromatography (HPLC) using mobile phase A: 0.1% FA in water, mobile phase B: 0.1% FA in acetonitrile (ACN) and a C_{18} column of 10 µM x 150 mM. A

gradient was used, starting from 98% mobile phase A and 2% B for 5 min, then transitioning to 5% A and 95% B over 31 minutes. Throughout, the flow was kept at 3 mL/min. The instrument used was an Agilent 1100 Series. The SuDP was reduced via vacuum centrifugation and brought up again, this time in 1 mL of dry ACN (filtered through aluminum oxide). Separately, 32 mg of di-succinimidyl bicarbonate was dissolved in 1 mL of dry ACN, but this solution was then mixed with the SuDP-ACN solution, 1 mL more of dry ACN, and 8 µL of pyridine, and then the mixture was shaken gently for 18 h. HPLC was performed again, with the same gradient and column as described above, to isolate the doubly-esterified SuDP. Another round of vacuum centrifugation later, deprotection was performed by bringing the dried SuDP up in only 5 µL of dimethylsulfoxide (DMSO), then simultaneously adding 5 µL of water and 190 µL of TFA and gently mixing via rotation for 15 min. One last round of vacuum centrifugation was performed, and then the SuDP was resuspended in 218.5 µL of mobile phase A (0.1% FA in water) and 11.5 µL mobile phase B (0.1% FA in ACN) for one equally-final round of HPLC, on the same gradient and machine, but utilizing a column of 4.6 µM x 100 mM and a flow rate of 1 mL/min, to collect the final SuDP.

Correct synthesis was verified by mass spectrometry using a Waters Premier quadrupole time-of-flight mass spectrometer via direct infusion of SuDP following purification by HPLC as described above.

II: Ester Quantification

Any statements that can be made from crosslinking analysis of a protein's structure are made from a position of pure comparison. A site that sees greater crosslinking activity than another site is likely more accessible to the solvent than the other; the number of crosslinks

identified at a site is useless except in concert with and relative to numbers from another site, or, more ideally, many other sites. Nonetheless, the concentration of crosslinking agent to be used, when applied in concert with a known concentration of protein to be crosslinked, ensures that a certain proportion of peptides will be crosslinked. Using not enough crosslinker can make it difficult to positively identify the signals of those residues that are crosslinked with confidence. However, the ultimate usefulness of the quantification lies in its facilitation of reproducibility. In order to observe similar relative levels of crosslinking of each lysine in an assayed protein, use of a consistent concentration of crosslinker is of primary value, and accurate determination of the quantity of crosslinking reagent synthesized and used is required to ensure the correct concentration, both for subsequent trials and reproducibility by others.

The final concentration of the SuDP after its synthesis was unknown, though an estimated concentration could be calculated from the reagents used in the synthesis. Quantification of the completed crosslinking agent was performed in accordance with Miron and Wilchek's 1982 spectrophotometric assay for N-hydroxysuccinimide (NHS) esters. The completed SuDP possesses N-hydroxysuccinimide esters at each functional end, and Miron and Wilchek's assay detects just this moiety, which absorbs light at the characteristic wavelength of 260 nm. The NHS groups must be released from their host molecule by weak base; strong base does separate them, but also tends to cause structural damage to the ring resulting in its failure to absorb light at 260 nm (Miron and Wilchek 1982).

To implement the assay, a standard curve was constructed. Acetic acid N-hydroxysuccinimide (AcNHS) was diluted from a 400 mM stock 100-fold using DMSO (to 4 mM), then serially diluted in DMSO to final concentrations of 2.0 mM, 1.0 mM, 0.5 mM, 0.25 mM, 0.125 mM, and 0.0625 mM, a 1:2 dilution each time. In turn, 100 µL of each AcNHS

solution was added to a methacrylate cuvette containing 900 µL of distilled water. Each solution's absorbance at 260 nm was measured, and these measurements served as a blank for the same solution with 10 µL of 1 M ammonium hydroxide (NH_4OH) added. These measurements from each solution were in turn used to form a linear curve model of absorbance as a function of NHS concentration, which was used to predict SuDP concentration for an absorbance value obtained using the same process with an initial solution of tenfold diluted SuDP solution (Figure 2.4). Note that each SuDP molecule possesses two dissociable NHS rings, whereas each AcNHS molecule has only one, doubling the SuDP's apparent concentration. This was corrected for, as was the tenfold dilution of the SuDP solution in distilled water in the cuvette. From the linear curve model was calculated the extinction coefficient for the esters, allowing the calculation. The calculated extinction coefficient was found to be within an acceptable margin of error to the one prescribed by Miron and Wilchek's experimentation (ibid.). Their paper cited a figure of $\varepsilon = 9700$ m^2/mol as the extinction coefficient. We consistently calculated extinction coefficients of between 8600 and 11000 m^2/mol using house-made standards.

III: Crosslinking Reaction and Filter-Aided Sample Preparation

A volume of 100 mL of phosphate buffer was prepared using mono- and dibasic phosphate salts to a final concentration of 100 mM and pH of 7.5. To this was added NaCl to a final concentration of 25 mM and disodium EDTA to a final concentration of 1.0 mM. To a 15 mL aliquot of this solution was added 0.0100 g of bovine serum albumin (BSA), producing a concentration of 10 µM.

Four 1 mL Eppendorf tubes were filled with 150 µL of the protein solution. To one was added 3 µL of DMSO. To the second was added 0.3 µmol of AcNHS in 3 µL of DMSO (a concentration of 2 mM). To the third was added 0.15 µmol (calculated) of SuDP crosslinker dried down and resuspended in 3 µL of DMSO (concentration of 1 mM). To the last was added nothing, as it was the control. After adding the labeling reagents, the tubes were incubated at ambient temperature for 45 minutes immediately after. Quenching of the reaction was achieved by adding 5 µL of a 1.0 M Tris solution to each tube (a final concentration of 33 mM) and incubating at ambient temperature for 20 min (Argo, Shi, Liu, and Goshe 2015).

To prepare the crosslinked samples for mass spectrometric analysis, the protein was reduced with 100 mM dithiothreitol (DTT) to a final calculated concentration of 5 mM, incubated for 30 minutes at 56 degrees C, mixed with 200 µL of a stock of 8M urea in 0.1 M Tris buffer at pH 8.5 in a filter centrifuge tube (10000 kDa filter used), and centrifuged for 15 minutes. Then 200 µL more urea solution was added and the tube centrifuged again for 15 minutes. The flow-through was discarded and 100 µL of an iodoacetamide (IAA) solution consisting of 50 mM IAA in the urea stock solution described above was added to the filter, which was incubated for 20 min in the dark at room temperature. The filter tube was then spun for 10 min. 100 µL of the urea solution was added and the tube spun for 10 minutes, then this

was repeated twice, then the flow-through discarded. A volume of 100 µL of 50 mM ammonium bicarbonate (ABC) in water was added in the same way, spun for 10 minutes, and this process repeated twice. The flow-through tube was discarded following this and 50 µL of trypsin in ABC (as above) was added to the filter, made to ensure a protein-trypsin ratio of 100:1, and then incubated at 37 degrees C for 18 h. A volume of 40 µL of the ABC solution were then added and the basket spun into a new collection tube for 19 minutes, at which time 40 more µL of the ABC solution were added and the centrifugation repeated.

The crosslinked protein solution was analyzed via liquid chromatography-tandem mass spectrometry (LC/MSMS) using an LTQ Orbitrap Elite mass spectrometer (ThermoFisher Scientific) coupled to a liquid chromatography column (C_{18} packing material, 3 µm particle size, ReproSil brand produced by Dr. Maisch GmbH, an Easy-nLC 1000 system by ThermoFisher Scientific) through a PepMap 100 (C_{18}, 5 µm) trapping column. The mobile phases used were: Mobile Phase A, 98% water, 2% acetonitrile (ACN), and 0.1% formic acid (FA), and Mobile Phase B, 5% water, 95% ACN, 0.1% FA. The gradient used was a 5% Mobile Phase B to 40% B gradient over 80 min, at a flow rate of 300 µL/min. Positive-mode detection was used, with a survey MS scan with an m/z range of 400-2000 and a resolution of 60,000, and MS/MS was performed at 15,000 resolution in a data-dependent manner based on on the three most intense peaks from the survey scan. Enrichment for crosslinked polypeptide pairs was accomplished by scanning for +4 or greater ions, using an isolation width of 7.0, HCD activation and dynamic exclusion at default settings. A tryptic polypeptide is usually found ionized in the +2 state (or higher), and crosslinked polypeptides are composed of two tryptic polypeptides linked by a neutral crosslinker, and typically have a charge of +4 or more.

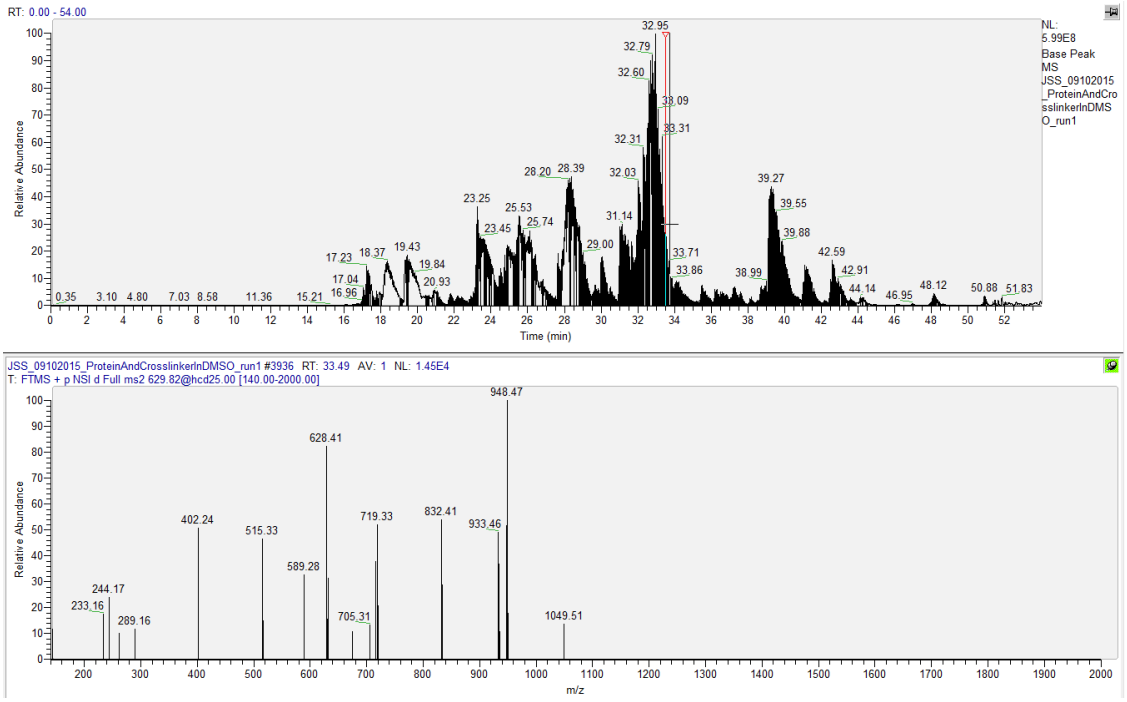


Figure 2.4 Example Mass Spectrum of Crosslinked Polypeptides: Above is pictured the base peak chromatogram and below is pictured the product ion spectrum from the scan selected in red. The product ion spectrum displays the fragmentation pattern of the selected precursor ion at m/z 629.82.

IV: Materials

Reagents for the synthesis of the crosslinking agent were generally HPLC-grade or higher where available. Piperidine, pyridine, dimethylformamide, dicholoromethane, diisopropyo carbodiimide, and dimethylsulfoxide were purchased from Sigma/Aldrich (St. Louis, MO) at 99% or higher purity. The Fmoc amino acids were purchased from Novabiochem/Millipore Sigma (Burlington, MA). Trifluoroacetate was purchased from Fisher (Waltham, MA) at HPLC grade. For the HPLC purification itself, American Chemical Society (ACS) reagent-grade acetonitrile (98%+) from Sigma/Aldrich and HPLC-grade formic acid from EMD-Millipore (Burlington, MA) were purchased. Ammonium hydroxide for the ester quantification was purchased from Sigma/Aldrich at ACS reagent purity. Lyophilized, low fatty-acid bovine serum albumin was purchased from Sigma/Aldrich at 99% purity, and the digest was performed with porcine trypsin from Thermo Scientific (Waltham, MA). Water was distilled and purified using a High-Q 103S water purification system (Wilmette, IL). For LC/MS/MS analysis acetonitrile and and water were optima grade for LC-MS from Thermo Fisher Scientific (Waltham, MA).

CHAPTER 3 – Result, Analysis, and Discussion

I: Overview

The ultimate product of the crosslinking process is mass spectral data. The pattern of m/z peaks resulting from protein crosslinking is nigh impossible to predict, and a simple confirmation of expected peaks is not viable. This data requires computer analysis to interpret usefully, and many programs are available to serve this function (See Chapter 3.II). Peptide recognition is very reliably carried out by modern computer software, but the nature of the crosslinking reaction dictates that crosslinked peptide-detecting software must be capable of detecting peptides linked (by a characteristic and predictable mass) to another peptide. The characteristic mass increase of bound crosslinker to a single peptide is easily detected by a number of different software packages due to its functional similarity to many kinds of post-translational R-chain modifications, which always increase the mass of the affected peptide by the characteristic mass and often possess specificity to one amino acid. In the case of SuDP, specific binding to lysine residues is very predictable, and detection of both halves of the cleavable reagent is accomplished with very standard software (using the LC/MS3 mode of analysis) (Soderblom 2008), but to predict which two peptides are bound by uncleaved crosslinking agents is fraught with more variables and requires more specialized software.

Prior to the crosslinking reaction itself and the subsequent mass spectrometric analysis of the crosslinked protein, the synthesized crosslinking agent was quantified. Accurate quantification of the crosslinking agent is vital, as the stoichiometric ratio of crosslinking agent to protein substrate can influence binding site preference. Saturation of the most accessible sites can increase disproportionate binding of less-favorable sites than would be observed in less-saturated conditions. More importantly, an accurate calculation of the amount of crosslinker

synthesized and used ensures repeatability and consistency in results. Simple stoichiometric calculations using the known masses of the reagents used in the synthesis of the crosslinking agent provide a maximum yield, but an assumption of some loss during the synthesis is very reasonable.

Quantification was accomplished using the ester quantification method devised by Erik Soderblom (Soderblom 2008). In this assay, an ester-containing compound of known concentration is serially diluted, tenfold per dilution, to create a standard for which the concentration of succinimide rings is known for each treatment. To each standard is added a small amount of base, to induce dissociation of the succinimide rings from their host molecules into solution. Addition of an excess of base can cause breakage of the rings, so care must be taken to avoid using an excess of base. Soderblom's original paper described use of sodium hydroxide, a strong base, as the base used to perform the dissociation, but ammonium hydroxide was found to work as well, and poses less risk to the dissociated succinimide rings. Each diluted solution was then assayed for light absorbance at the 260 nm wavelength with a spectrophotometer. This enabled creation of a standard in which absorbance was directly correlated to concentration.

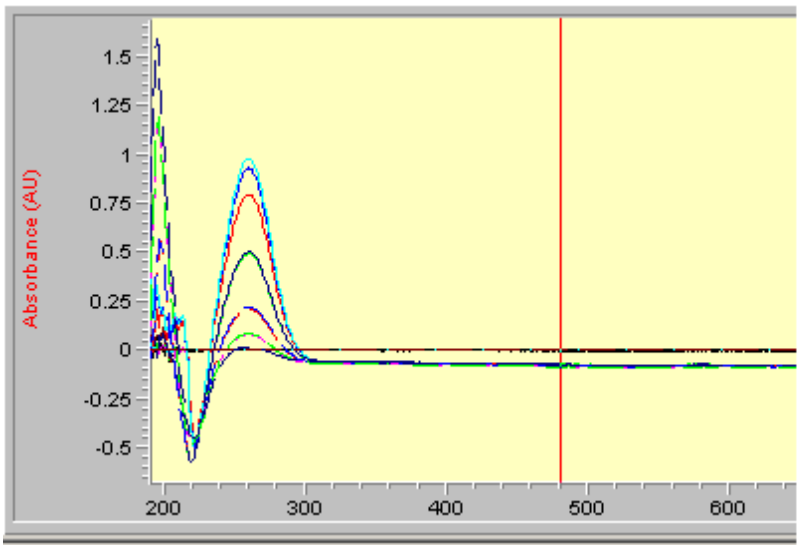


Figure 3.1 Ultraviolet Absorbance Spectroscopy Standard: The spectrophotometric data from the instrument. Peak absorbance occurs at 260 nm, by the base-dissociated NHS rings.

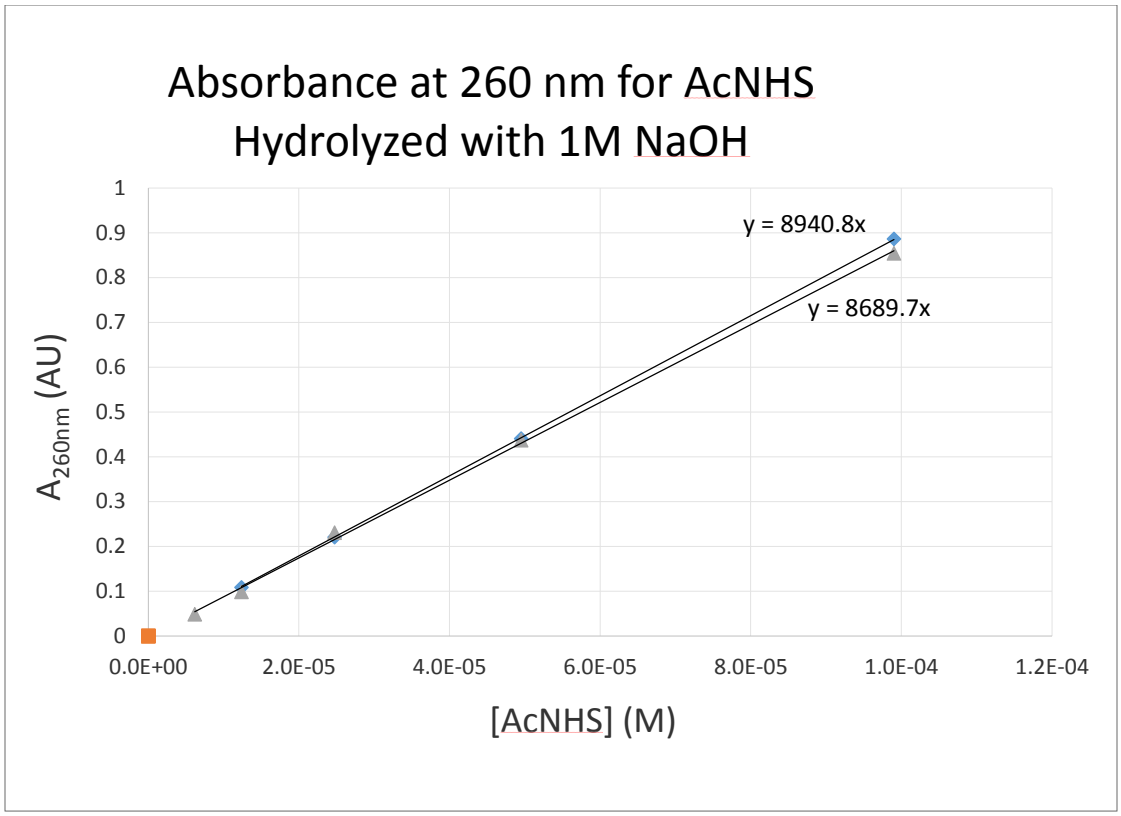


Figure 3.2 Absorbance Standard Curve: The plotted points from the AcNHS standards tested for absorbance at 260 nm. From these standards was determined the slope, from which is derived the extinction coefficient.

Using this standard allowed calculation of the concentration of SuDP that was synthesized. SuDP possesses dissociable succinimide rings just as acetic acid n-hydroxysuccinimide does, and it was treated in the exact same way, with ammonium hydroxide and spectrophotometric analysis, to yield an absorbance value which was converted to concentration via calculation using the standard developed as described.

II: Software Investigation

The crosslinked peptide analysis software used most often was StavroX, programs developed by Michael Götze of the Martin Luther University Of Halle-Wittenberg. StavroX was developed for the specific purpose of analyzing crosslinked peptide mass spectrometric data, whereas most mass spectrometry data analysis programs are more generalized. StavroX is used for interpreted detection of crosslinking by incleavable crosslinking agents. Interpretation of crosslinked peptide fragments (with the crosslinker still intact) is an intensive and specialized task, and StavroX was developed with that specific function in mind. Interpretation of peptide fragments bound by cleaved crosslinking agent fragments is less specialized, as the bound crosslinking agent fragments are analogous conceptually to posttranslational modifications of peptides, and software to interpret data of posttranslationally-modified proteins already exists in relative abundance, e.g. Mascot.

The crosslinked-peptide data was produced originally by the mass spectrometer in the form of several files in ".raw" format, a generic and bulky format. StavroX requires files input to be in the Mascot Generic Format (".mgf"). ThermoFisher's Proteome Discoverer software was used to simply convert the file into the proper format, which is proprietary to ThermoFisher and more streamlined. StavroX produces .csv output natively, a format conducive to script analysis.

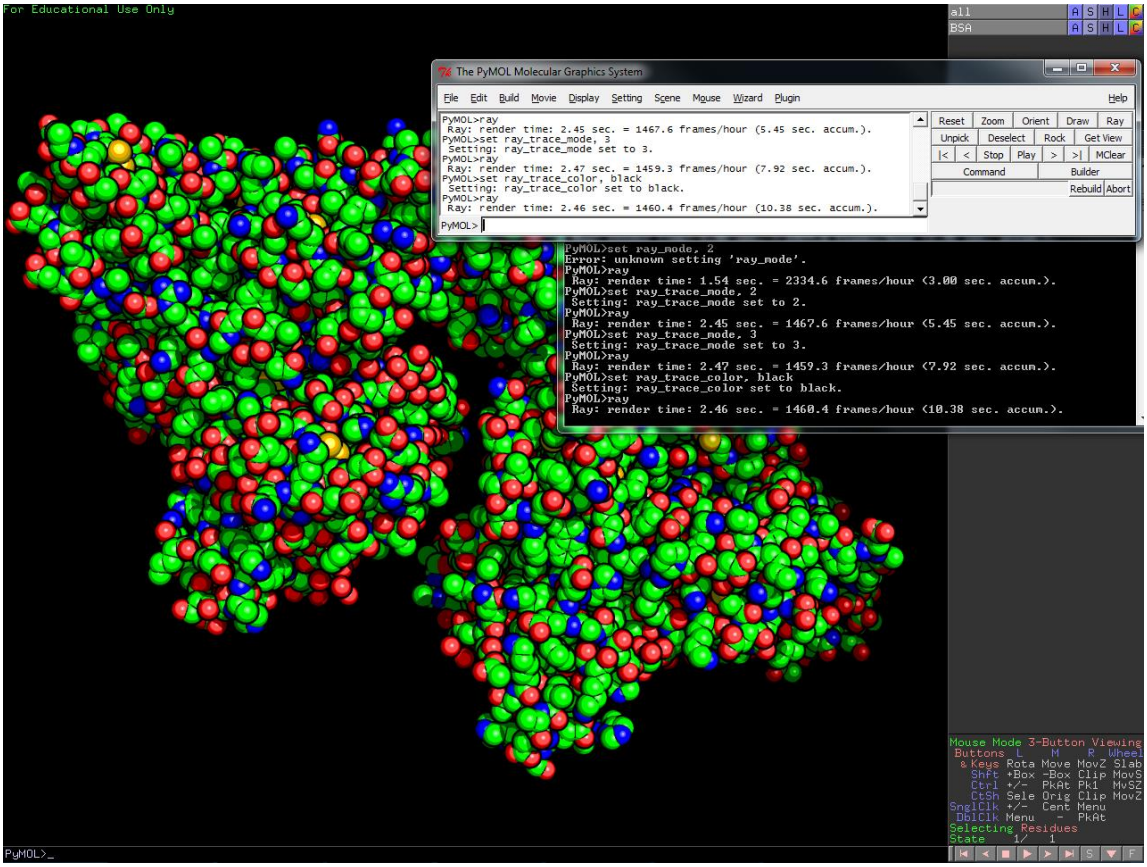
Scripts were primarily made in Python 3. Python is powerful and flexible, and can be used to interface directly with PyMOL, the modelling software, which accepts scripts in Python.

PyMOL itself, specifically version 1.3, was used primarily for visualization. Developed by DeLano Scientific, the program offers flexible viewing of three-dimensional models of molecules. The program is written in Python and accepts scripts written in Python, permitting robust manipulation of data obtained from the crosslinking experiments without excess prior sanitization. Example: the one-dimensional array produced by the crosslinking participation analysis script (see Figures 3.9, 3.10, and 3.11) can be represented as color on each constituent amino acid of the host protein despite the PyMOL software having no preprogrammed ability to process data specifically from crosslinking. This enables easy spatial visualization of such data as hydrogen-deuterium exchange data over time, as well as an intuitive way to portray crosslinking availability.

The University of California, San Francisco has developed a similar software suite to PyMOL called Chimera. Like PyMOL, it features the capability for flexible visualization of biomolecules in three dimensional views, and it also features extensibility; a wide variety of additional functionalities are available for download. One such module is XlinkAnalyzer (Kosinski et al. 2015). Developed by Jan Kosinski of the European Molecular Biology Laboratory, this module provides dedicated support for visualization of crosslinking directly. It accepts as input comma-separated value files and displays the putative crosslinks on the host protein model. The input files are of a style somewhat similar to those that can be generated from StavroX's standard output, but require some transformation, as they require the identified peptides and the index numbers of the linked residues, all of which are exported by StavroX, but in a differently-formatted form. The module does not perform any analysis of mass spectrometric

data itself and cannot identify crosslinkers from raw spectrometric data, and thus must be coupled to a program that serves this function, such as StavroX, but it does provide for robust visualization of the crosslinks detected by such a program with a graphical user interface, and offers some control in the form of selective display by such considerations as score and threshold distance between crosslinked loci.

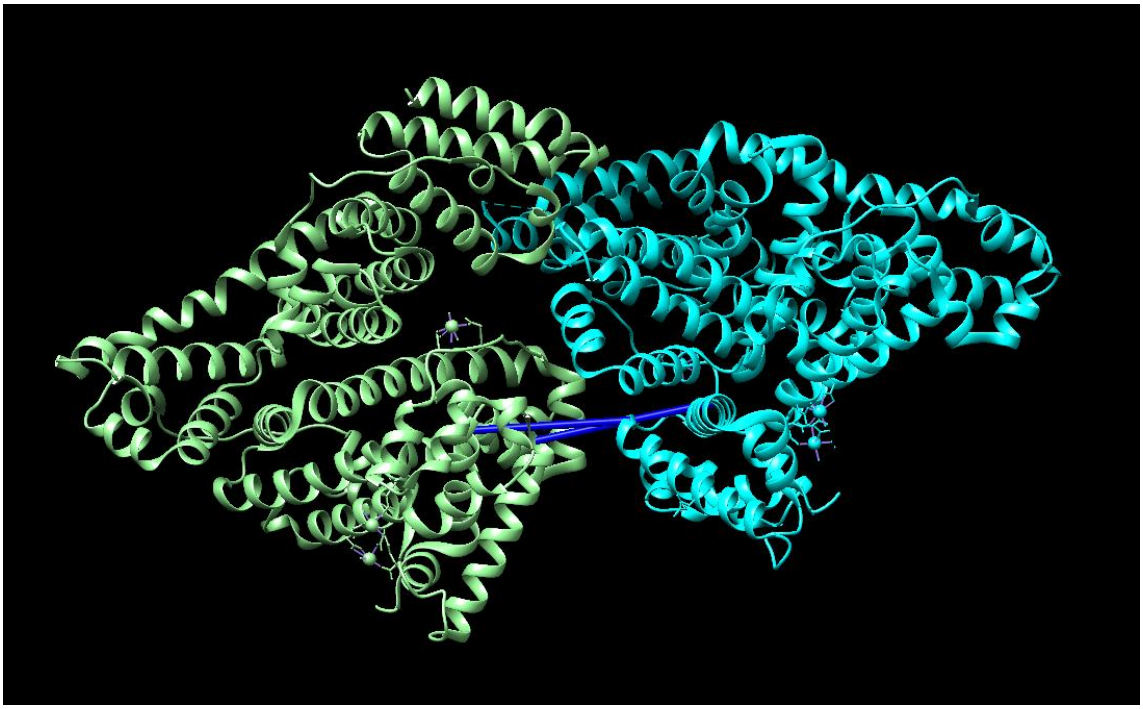
The software used is vital to the workflow as it affects the accuracy of the results, as might be expected, but also the reproducibility of the results. New software can ease the computer-assisted processing of crosslinking data but care must be taken to document the software being used, as it becomes obsolete quickly. Not only this, but defunct software is often unobtainable, as has been shown in Chapter 3.II, which can complicate reproducibility if it is no longer accessible to other, or indeed to the same, researchers. This problem is magnified by proprietary software and licensing issues. Care should be taken to document not only the software used and its origin, but also the algorithms the software employs to return its results, so that even if the access to the software is broken or impossible, an equivalent can be produced or found that uses the same algorithms as were used originally, since the same algorithm would return the same results regardless of the other individualized trappings of whatever program was originally used. This again is made more difficult by the use of proprietary software.



(Figure generated by PyMOL (Schrödinger, 2014))

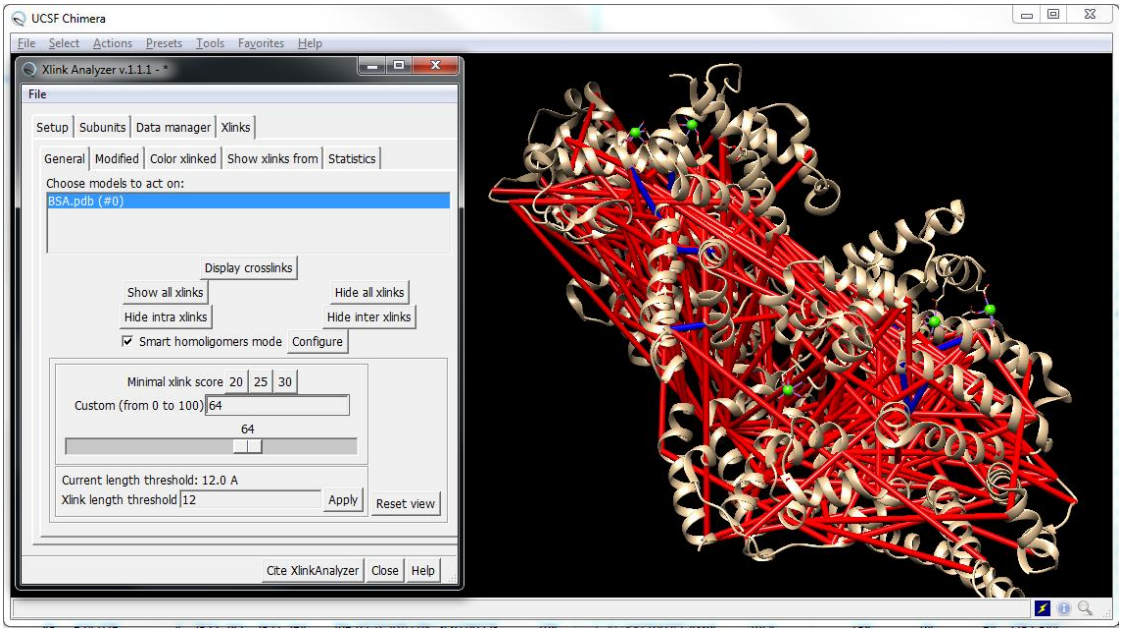
Figure 3.3 The PyMOL Interface and BSA: Input is given primarily through the command line in Python.

The number of crosslinking events identifiable with a certain level of confidence but involving two peptides of the same subunit and not separated by a greater distance than the length of the crosslinker used naturally narrows down the field of reliable crosslinking events that can be used, potentially, to make reliable statements about the three-dimensional structure or topology of a protein (subunit). The number of such crosslinking events which define reliable intramolecular parameters for a single trial of crosslinking can be quite low, on the order of a dozen or fewer. This paucity of results plus the stochastic nature of the crosslinking process (especially relative to more-absolute methods such as crystallography or nuclear magnetic resonance spectroscopy) makes, more than usual, a large number of trials advisable. More trials permits greater numbers of relatively reliable statements about the structure of an analyte moiety.



(Figure generated by Chimera (Pettersen et al. 2004))

Figure 3.4 The Chimera Interface and Crosslinked BSA: The two subunits have been tinted different colors and two example crosslinks have been mapped between the subunits (in thick blue for visibility).



(Figure generated by Chimera (Pettersen et al. 2004))

Figure 3.5 The Xlink Analyzer Plugin for Chimera: Pictured displaying a number of crosslinks from an externally generated list. Based on input parameters, the software can filter the list to only display crosslinks with certain characteristics, such as filtering by score.

However, many programs capable of crosslinking analysis exist and serve different niche functions. DX/DXDX/DXMSMS, produced by Creative Molecules, is specialized for the purpose of analyzing mass spectrometry data from crosslinking reactions involving isotopically-coded, cleavable crosslinkers only. Isotopically-coded crosslinkers are utilized for some crosslinking reactions due to the isotopically-coded crosslinking entity possessing a very specific (and thus detectable) mass. Use of isotopically-coded crosslinking agents as a mixture with the non-isotopically-coded analogues of the same agents engenders characteristic patterns in the MS spectrum that can be detected by the software. This raises confidence in the presence of certain crosslinking patterns (Petrotchenko, Olkhovik, and Borchers 2005). In crosslinking reactions rich in peptide diversity, relative signal strength of the sought crosslinked peptides may suffer, and a higher margin of confidence in detected crosslinks is a boon. However, DX/DXDX/DXMSMS is optimized for this specific sort of analysis, and holds no advantage if not applied to isotopically-coded crosslinked peptide data sets.

Another software solution, X-Linked Peptide Mapping Algorithm, exists, but is web-based, outputs via e-mail, and only works with a static list of crosslinkers. As SuDP is not on that static list and the list is not modifiable by the end user, this software was not considered.

Several other investigated programs have unresolvable dependencies that prevented their use. SIM-XL requires the MSFileReader software, which could not be obtained, to function. Crux, a generic mass spectrometry data analysis package containing the command “search-for-xlinks” which permits specific analysis of crosslinked peptide data, also relies on this MSFileReader software. Several other pieces of software seem to exist only as names: XLINK/Kojak, Crossfinder, pLink and MS2Assign were found alluded to but unsourceable for testing during the time of my thesis research.

III: Script Analysis

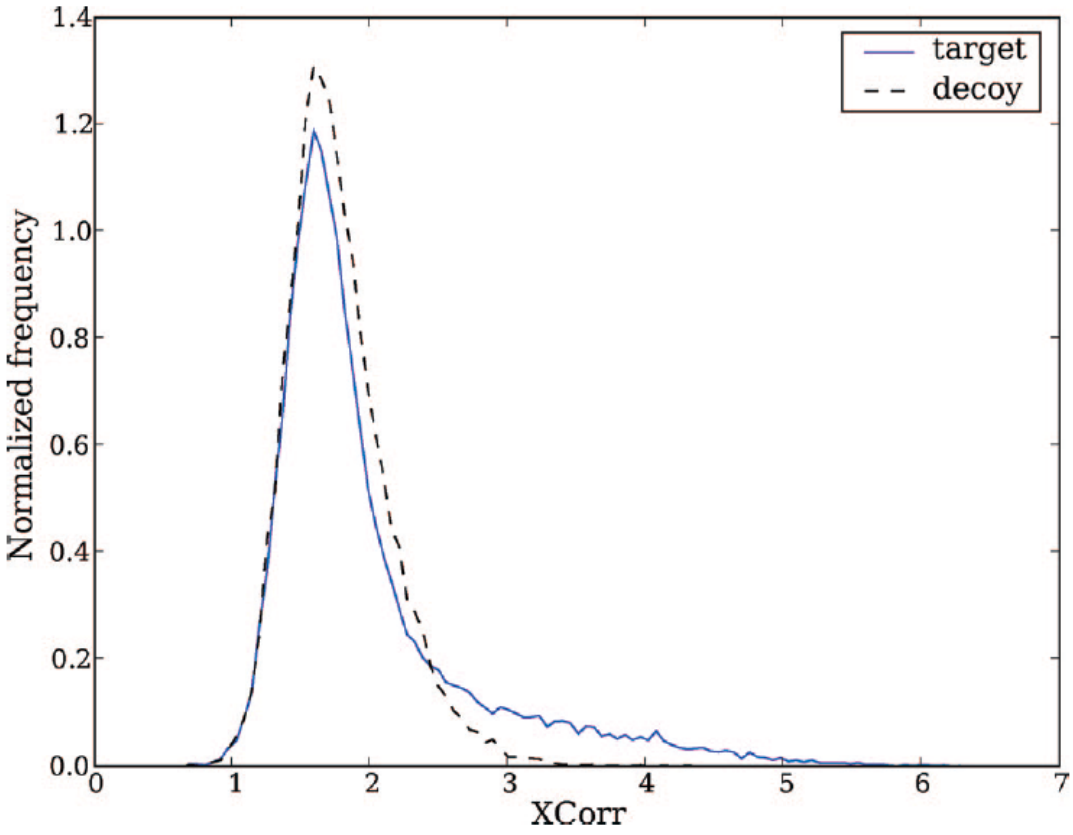
Amino acid residues located on the surface of the protein would be expected to be more accessible to crosslinking activity. The crosslinking agents are borne in the solvent, and the protein surface tends towards greater hydrophilicity and solvent-accessible surface area. To investigate this, a Python script was written which acted on a list of polypeptides. In this case, the list was the output of StavroX. It noted for an optional input threshold value how many times each constituent amino acid of the substrate protein was judged present in a crosslinked peptide by StavroX, with the input threshold value dictating the score of confidence required by StavroX for a crosslinked peptide to be included in the list of those to be tallied. The result is a simple integer count for each constituent residue, itself unsuitable for comparison with other data, but normalizing the counts with relation to the global maximum count gives relative participation in crosslinking for each amino acid residue of the substrate protein. This data can be compared with data on the solvent-accessible surface area of each amino acid residue. Higher solvent-accessible surface area would be expected to correlate well with crosslink participation frequency. For clear visualization, the one-dimensional count array can be plotted against the protein sequence index numbers of the constituent amino acids.

Determining the optimal threshold value to use for selection of the detected putative polypeptides was not straightforward. The confidence score output by StavroX is not without value, but its usefulness is obfuscated by several factors. The software is optimized to detect crosslinked polypeptides, and such is its determination that it will, for any data set of large enough size, incorrectly identify some crosslinked peptides not truly present simply by dint of chance. In order to correct for this tendency towards error, which would be extremely difficult to avoid without compromising the quality of the data for true crosslinked polypeptides as well,

estimation of the rate at which the software detects false positives is helpful. The StavroX software itself includes a tool for this estimation in the form of a decoy analysis routine. This analysis method works by modeling the null hypothesis, the assumption that there are no crosslinked polypeptides present in the data, and comparing the data to that model. Generally, the null hypothesis in peptide identification scenarios is modeled by directing the software at *in silico*-generated sequences which would not correspond to true polypeptides (Kall et al 2008). Several methods are traditionally used to achieve this, such as randomization by residue or several-residue chunk of an existing sequence, or inversion of an existing sequence. The latter option is deprecated due to biological conservation of motifs increasing the likelihood that the inverse of a sequence does in fact exist, though this phenomenon's applicability to the single-protein target of crosslinking analysis in the case of BSA in this work is of less, but not zero, concern. Once an acceptable model of the null hypothesis has been constructed, results from the target data can be compared to the results from the model. Some threshold of score is set such that, above this score of quality, a low percent (usually 5%) of results are those that would be expected to occur by chance if the null hypothesis is held to be correct.

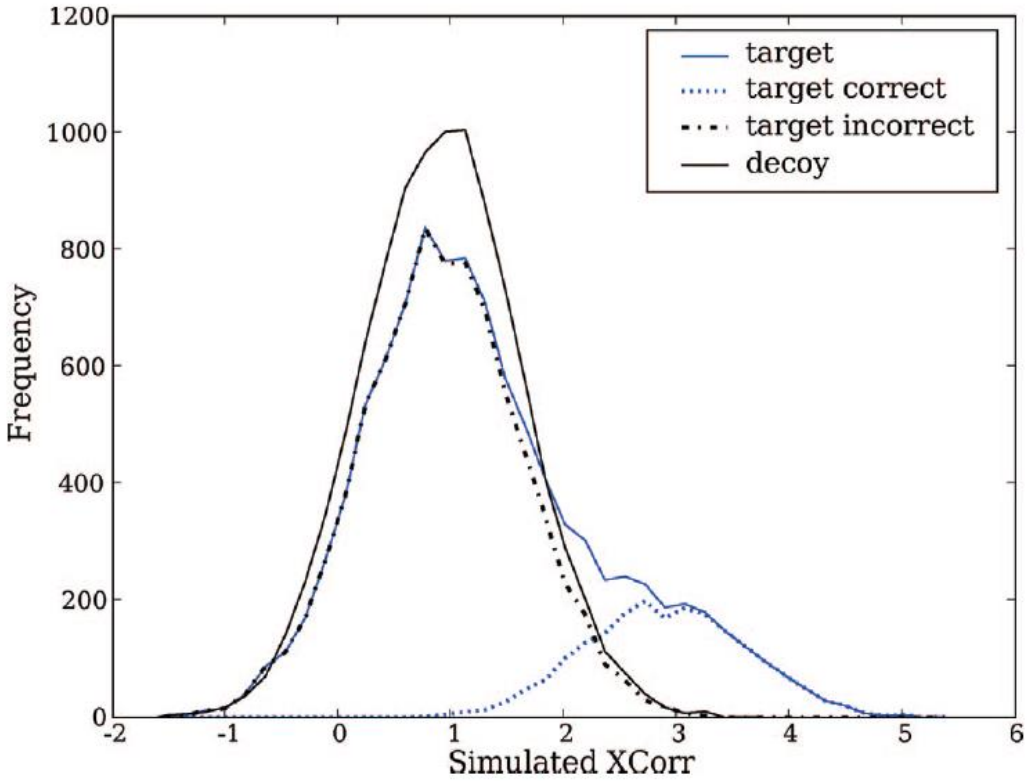
The relationship between the number of peptide identifications and score can be plotted, and this is the easiest way to determine which identified polypeptides should be considered. An analysis of experimental data would be expected to return a similar number of false identifications to the null hypothesis model data, simply because the software can draw incorrect conclusions from the experimental data, which contains the true sequences, just as it can from the null model, which ostensibly does not. However, in addition to the usual background incorrect identifications, the software would also return correct polypeptide identifications. Thus, as Figures 3.5 and 3.6 show, the graphs of null model vs. experimental data

appear similar at low scores of confidence, but the experimental data exhibits significant "tailing" towards higher scores, such that there exists a region where target curve area is maximized relative to null curve area, and target score should be chosen accordingly. As discussed above, the exact score above which the percentage of null model, decoy peptides relative to experimentally derived peptides can be determined for any desired percentage. A value of 5% is commonly used, and is the default for the crosslinked polypeptide analysis program StavroX referred to throughout this thesis. This number is dubbed the p-value, and describes, in summary, the probability if the null hypothesis is taken as correct that the observed result is due to chance. This is the fundamental principle behind decoy analysis, which was used to verify the crosslinking data obtained throughout the experiments described in this thesis.



(Figure from Kall et al. 2008)

Figure 3.6 Decoy Analysis 1: Distribution of values for target and decoy PSMs. Values for the target PSMs are represented by the solid line and for the decoys by the dashed line.



(Figure from Kall et al. 2008)

Figure 3.7 Decoy Analysis 2: Distribution of values for target and decoy PSMs. Values for the target PSMs are represented by the solid line and for the decoys by the dashed line.

An application of this concept has traditionally been used to verify the identification of linear polypeptides by searching an independently created database of theoretical spectral data for matches between the experimentally-derived spectrum and a created spectrum matching a given polypeptide sequence or fragment (Eng, Mccormack, and Yates 1994). Such matches are likely predictors that the matching sequence corresponds to a sequence in the sample. This correlation method has tremendous value in identifying the presence of polypeptide fragments of many proteins in a heterogenous sample, with its facility for matching a selection of targets limited only by the quality and availability of the spectral databases used. Proteins similar to known extant proteins may be identified (ibid) and by using large databases containing spectral data from many species, the species of origin of a sample may be identified quickly. The source of the approach's speed and flexibility is that the method works primarily at the software level. A high-quality input spectrum increases accuracy, but the fundamental approach to the mass spectrometric data acquisition changes very little across different analyses. Applying one experimental spectrum to a database containing protein sequences from many species can identify the species, but if this is known or irrelevant, as narrow a database of proteins to match as desired may be used, thereby reducing the pool of potential false matches. This is a choice which can greatly reduce the false detection rate in light of the fact that the larger the database being searched, the more false matches will occur simply due to chance. Higher mass measurement accuracy can reduce the incidence of this phenomenon, and the importance of using a high mass accuracy instrument to perform crosslinking analysis should be stressed. Today, several high-profile and actively maintained large-database search programs exist to automate this process and standardize it across the many potential usages and trials it might be

applied to. Mascot and SEQUEST are commonly used extant database-searching protein identification programs. (Keller, Nesvizhskii, Kolker, and Aebersold 2002)

Complicating this part of the analysis is the fact that there are extant programs for the analysis of existing peptides with user-customizable lists of post-translational modifications, but this software is not effective against crosslinked spectral data because the incleavable crosslinking agent is not a modification which can simply be added to target residues (likely lysine in this case) but also incorporates another bound polypeptide of variable length and sequence, making identification of crosslinks as static mass additions impossible. With this method and this software, it is, however, possible to identify crosslinker bindings in which the second functional end remains unbound. This is called a "dead-end" crosslinking event, and it still has informational value, as a bound crosslinking moiety at a locus still indicates solvent accessibility at that locus.

IV: Representation Analysis

A graph of the data produced from crosslinking of BSA when collated into representational counts has dramatic peaks and valleys, but typically shows a nonrandom pattern of regions of relatively higher crosslinking participation, lending credence to the hypothesis that crosslinking favors binding of regions more accessible to the solvent that bears them and is not random.

To test this hypothesis, in addition to the trials of crosslinked bovine serum albumin experimental preparations, similarly-prepared trials of bovine serum albumin mixed with "dummy" crosslinking agent acetic acid n-hydroxysuccinimide were also analyzed by mass spectrometry. This reagent has similar affinity for lysine binding to the true crosslinking agents,

but only one functional group. The mass spectrometry data from these trials was compared with the mass spectrometry data from the experimental preparations and likewise analyzed and subjected to decoy analysis by StavroX. The scores as reported by StavroX of the AcNHS mass spectrometry data indicated the presence of crosslinked peptides with a maximum score above 100, which compares well to the maximum scores attained by crosslinked peptides identified by StavroX in the experimental SuDP-crosslinked data. As AcNHS is the negative control, meant to emulate true crosslinking agents' behavior without linking the substrates, this high score may seem unexpected and indicative of a problem, but it can be explained using the principles already discussed. Specifically, any crosslinked polypeptide-identifying software is susceptible to some ratio of false detection of incorrect crosslinked peptides. That a sample of entirely uncrosslinked protein substrate could contain fragments which by chance match closely enough the mass of a true crosslinked pair of polypeptides is very reasonable. This is the purpose of the p-value accuracy assessment. An identification with a high score in a dataset from a sample which should not contain any crosslinked peptides is not conclusive evidence of crosslinking activity; the software makes errors with some regularity; over the size of such large datasets, these errors become nontrivial, and eliminating them is the purpose of the decoy analysis.

In Figures 3.7 and 3.8, the decoy analyses as produced and visualized by StavroX are shown. These figures were produced from StavroX decoy analysis of a crosslinked sample of BSA and a similar sample of BSA mixed with AcNHS (as a negative control). The concentration of AcNHS was twice that of the concentration of the crosslinking agent used, to account for the fact that AcNHS has only half the functional n-hydroxysuccinimide groups (one) that the bifunctional crosslinking reagent used possesses. It can be observed that the graph of decoy peptide scores for the AcNHS sample was much closer to the graph of the experimental data than

the graph of decoy peptide scores for the true crosslinked sample was to the graph of its own experimental data.

Correlation of solvent-accessible surface area with crosslinked peptide representation would be expected to be present. Solvent accessibility can be estimated to a reasonable degree of accuracy by molecular modeling programs such as PyMOL and the crosslinked peptide representational data correlated with this data and scrutinized for correlation.

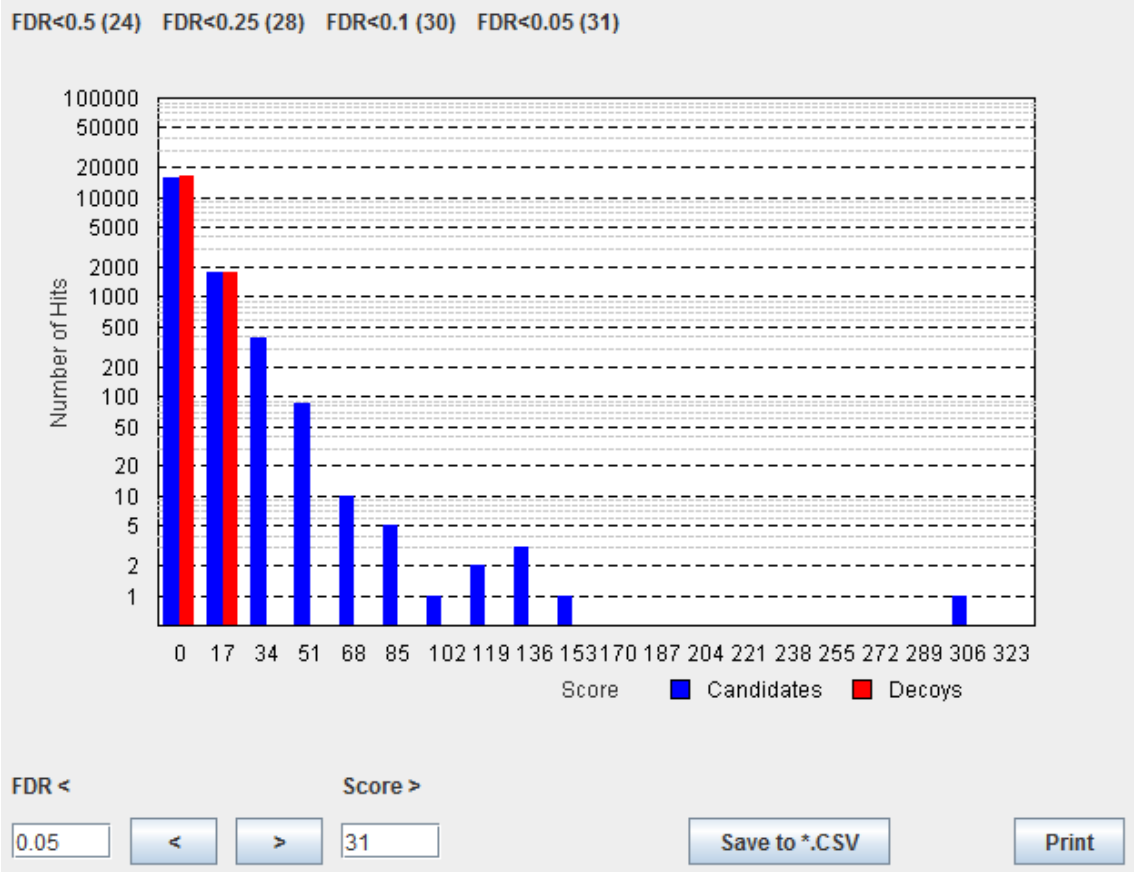


Figure 3.8 StavroX's Decoy Analysis 1: Analysis of an experimental crosslinked BSA sample. Candidates outscored the decoys to satisfaction above a score of 31.

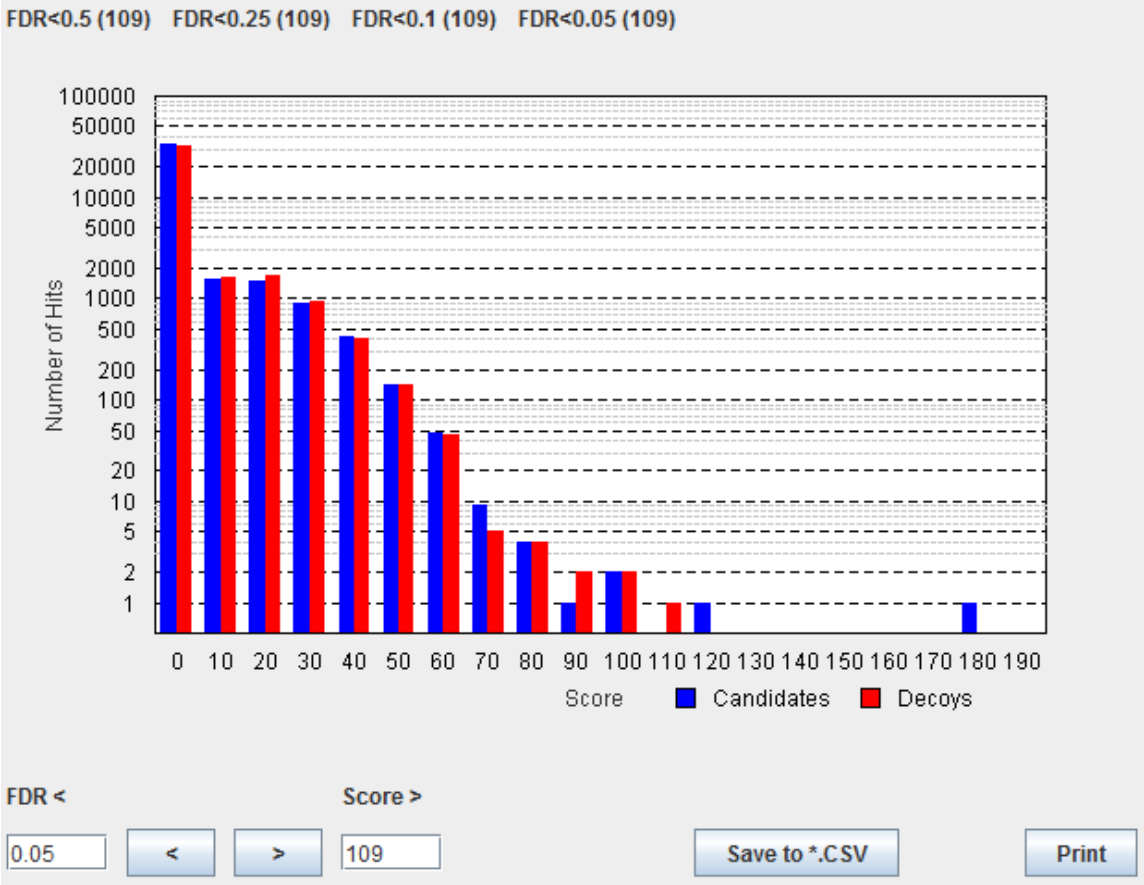


Figure 3.9 StavroX's Decoy Analysis 2: Analysis of a control sample of BSA labelled with AcNHS. Candidates outscored the decoys to satisfaction above a score of 109, leaving unusably low numbers of hits to consider.

First, analysis was performed on the data produced from a single crosslinking trial using SuDP. Shown in Figure 3.9 is the output of script analysis determining the frequency of crosslinking of each residue of the BSA analyte. Shown is data culled from only crosslinked peptides with a score of confidence of 0 or higher, as determined by StavroX. The data is presented by relative representation of each residue index of BSA by the number of times it appears in crosslinked polypeptides. Once tabulated, each count was divided by the maximum count for any one residue, yielding a scale of 0 to 1. Thus arranged, a pattern of peaks and valleys becomes apparent, suggesting that there are regions which tend to be more accessible to the crosslinker's reactivity than others. This is the predicted result. To test the effect of changing the set of considered polypeptides based on confidence, a secondary trial of the same data filtered to exclude all the identified crosslinked polypeptides with a StavroX-reported score of confidence of 50 or less. This sharply reduced the number of considered polypeptides. A total of 16,146 identified putative crosslinked polypeptides had a confidence score of 0 or greater. When the required confidence score was raised to 50, the number of eligible peptides dropped to 677, just 4.2 percent of the previous tally. Even with this dramatic reduction in the pool of considered polypeptides, the resultant graph of crosslinking participation, produced in the same way and also normalized to a 0-1 scale, bore resemblance to the less-discriminating 0-threshold result.

The graph produced by the polypeptides above 50 score is similar to the one produced by a 0-score filter, but the differences are not proportionate. The same regions along the analyzed host protein's length appear to be differently represented by both subsets. As an example, the maximum representation (1.0 on a 0 to 1 scale) on both graphs occurs approximately at the residue with the index number of 210. It is 1.0 in both cases. Moving along the protein's length, a region of increased crosslinking accessibility from a hypothetical baseline of approximately 0.1

is reported faithfully by both sets of data from index values of approximately 225 through 257. Though this region being higher than its immediate surroundings is agreed upon by consensus by the two sets of data, the data from only the higher-scoring peptides reports its average height as approximately 0.3, whereas the data with lower-scoring putative linking events included reports it as approximately 0.8. This difference is dramatic relative to the agreement a few dozen residues upstream on the residues at 1.0, nor is the rest of the data consistent with respect to the amount of agreement between the two graphs, though in almost every case, the data from the smaller pool shows lower values. It might be expected that such a limitation would scale the graph proportionately downwards save at an outlying maximum, but the observed shift is not proportionate. That there exists a difference in the susceptibility of different loci along the analyte protein's length to the activity of crosslinking is not surprising, but this indicates that the differences are reported differently by varying the score of the matches used.

These observations are meaningless in a vacuum. As has been discussed, the confidence score reported by StavroX does not give an objective measure of the quality of the assignments. For all datasets reported by the program, there exists no unified score number above which all putative links are considered valid. Only by applying decoy analysis (as explained in Chapter 3.III) to each dataset on an individual basis can it be determined which crosslinked peptides are not the result of a false positive caused by chance and by using a program which is designed to find links in very complex data to a given certainty (of 95%, usually). The differential patterns of reported crosslinked peptide representation at different scores indicate that changing the score of peptides considered does indeed change the outcome, confirming that assigning a cutoff score to the list of peptides used in any analysis is a vital step for accuracy. That cutoff point is automatically calculated by StavroX as part of its decoy analysis, and the program uses the

standard 95% certainty (p-value of 0.05, as in Käll et al). Including polypeptides below the assigned cutoff point means accepting data from links greater than 5% likely to be incorrect due to overzealous pattern-finding by StavroX, and therefore unworthy of consideration.

The score below which greater than 5% of identifications are liable to be false positives varies with each dataset, but it would be expected that the higher in quality, purity, and crosslinked polypeptide concentration the analyzed data is, the lower the required score to eliminate the less-likely polypeptides would be. This was borne out by the aforementioned AcNHS vs. SuDP crosslinking trials, in which the required score to reduce the p-value to 0.05 or less was typically approximately 30 for crosslinked samples and approximately 100 for AcNHS samples.

A graph of the relative representation of each residue of the host protein (BSA) in detected crosslinked peptides, limited to peptides only above a score corresponding to a p-value of 0.05, was prepared from a trial of 1:100 BSA:SuDP crosslinking, and can be seen in Figure 3.11, using the prescribed cutoff score of 31, as dictated by StavroX's decoy analysis to find the p-value. As predicted, the graph does resemble the graph of the same data limited to scores greater than 0 and the graph of the same data limited to scores greater than 50, but is a more accurate representation than the same data with a cutoff score of 0, which, as discussed, would include polypeptides unacceptably likely to be inaccurate, but still includes as many polypeptides as possible for maximum size of sample. The same data culled to a cutoff of 50 would contain even fewer potentially inaccurate polypeptides, and is thus as valid as a source of information as the set from 31, but setting the required score higher than necessary removes from consideration many polypeptides that are still likely valid sources of information. This analysis relies upon using a large volume of identified crosslinks to specify those regions which appear more

frequently on average. An increase in the score required for consideration would have large ramifications for the result accuracy given the tendency towards decreasing populations at higher scores. The set of all identified polypeptides with a score of 90-100 is much smaller than the set with scores of 0-10 for virtually all datasets. Thus, eliminating lower (but still acceptable) scores is an elimination of one of the richest sources of useable crosslinking representational data. Crosslinking being a stochastic process and the data processing an approach which demands as much useable data as possible to eliminate the effects of chance, a methodology which yields the maximum amount of usable identified polypeptides is most desirable.

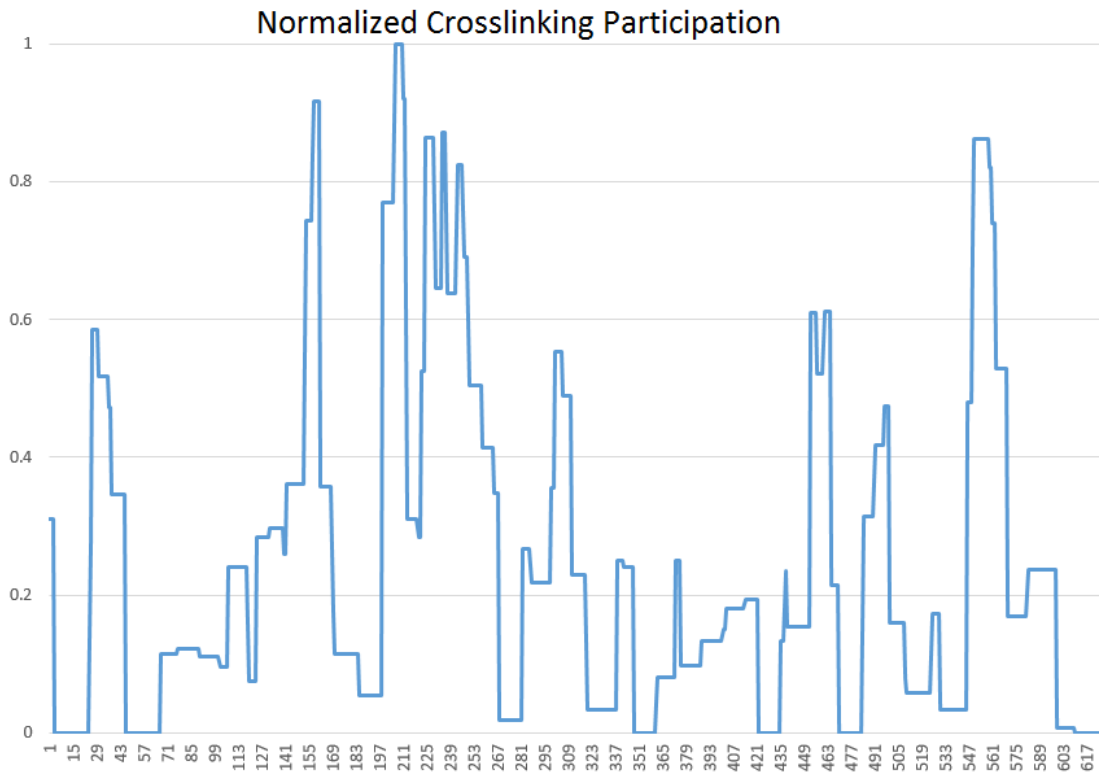


Figure 3.10 Representation Analysis of Crosslinking Participation: visualization of crosslinking participation (number of times any given residue appeared in crosslinked polypeptides) by representation of each residue of the analyte protein (index numbers along the X-axis) as a value normalized to a scale of 0-1.

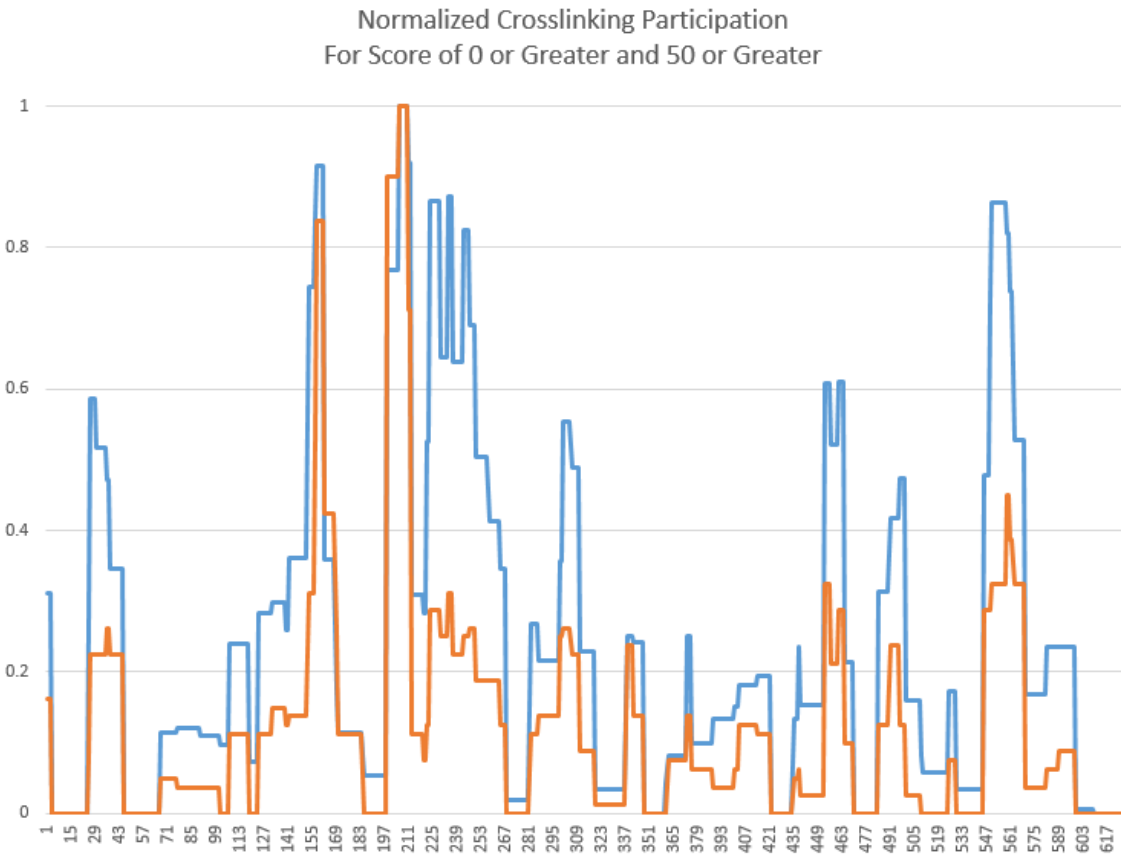


Figure 3.11 Representation Analysis of Crosslinking Participation – Score Filtered: visualization of crosslinking participation by representation of each residue of the analyte protein as a value normalized to a scale of 0-1, as in Figure 3.9, save that a second visualization composed from only polypeptides above a certain score of confidence is included as an overlay. It is normalized to a 0-1 scale as well, but exhibits differences from the indiscriminate data.

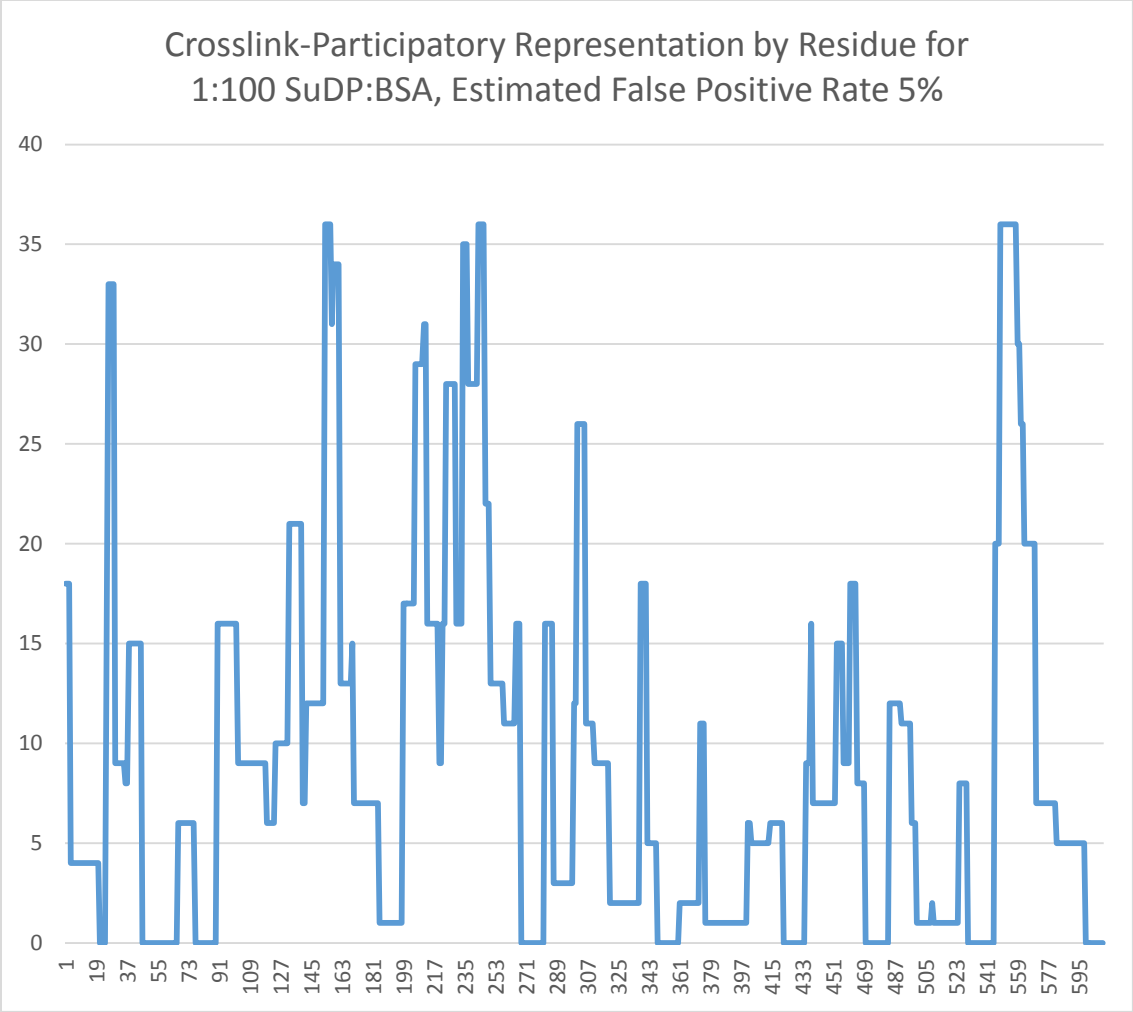


Figure 3.12 Representation Analysis of Crosslinking Participation – FDR Filtered: visualization of crosslinking participation by representation of each residue of the analyte protein as measured in the list of crosslinked polypeptides above the score at which the false detection rate was 5% or less.

V: Validity Verification

If tabulation of each residue's relative representation in crosslinked polypeptides yields proportional values to each residue's solvent accessibility, as would be expected from solvent-borne crosslinking reagents, this relationship would be verifiable by comparing each residue's representation as reported by StavroX to its solvent-accessible surface area, which can be returned by modeling programs such as PyMOL acting on a model of the geometry of the protein in question.

In order to test this, first an accurate determination of the solvent-accessible surface area of each residue of the analyte protein must be made. PyMOL has a native function, get_area, which performs this function on a per-residue basis. Dot solvent was set to "on" for the calculation, which was recorded in a one-dimensional array for ease of visualization. For comparison with the representational data, which was scaled from 0-1 and unitless, the area data was likewise scaled 0-1, with 1 being the maximum area of any one residue and each value divided by this value to yield relative data. The result is showed in Figure 3.12. Represented in this way, the data, unlike the graph of the data from the tabulated crosslinking participation analysis, does not show any patterns as clearly. The crosslinked polypeptide representation data shows several notable peaks of higher representation divided by valleys of relatively lower representation, a pattern which is clear to the eye, but the solvent accessibility graph is much harder to interpret without the aid of software.

In order to interpret the 0-1 scaled data from both sources (crosslinked polypeptide representation and solvent-accessible surface area), a simple comparison of differences was performed. For each residue along the protein's length, the absolute value of the difference between the relative representation and the relative solvent-accessible surface area was

calculated. The resultant numbers were meaningless. They were averaged across the entire set, however, and the number that this average yielded was compared to the average of the absolute values of the difference between the relative representation values and a randomly generated number between 0 and 1 for every residue along the protein's length. The average difference between the representational value and the surface area value was found to be approximately 0.2517. The average difference between the representational value and the random number was found to be approximately 0.35, though it varied somewhat around this centroid each time new random integers were chosen, by as much as 0.05 across ten recomputations. For comparison and control, a third comparison was made, this time between a random value between 0 and 1 and a second random value also between 0 and 1. A number of these differences equal to the total number of BSA residues used in the earlier computation were found, and the average of these calculated. This value was found to be approximately 0.33, variable by as much as 0.01 across ten recomputations. To remove the element of random chance, one more set of calculations was performed: the average of the absolute value of the difference between the crosslinked polypeptide representation and a forced value of 0.5, in place of the random number that was used before. For the value of 0.5, the average absolute value of the difference was found to be approximately 0.32, significantly below the true random value, but still more significantly above the experimental value.

The purpose of the representation-random comparison is to provide a control. The average value of the calculated difference between the absolute value of the representation value minus random value and absolute value of the random value minus a second random value are both similar enough to conclude that, based on the random-random values as a known negative control, the representation-random figure is an acceptable negative control.

A traditional correlation analysis using linearity of a graph with solvent-accessible surface area on one axis and relative crosslinking representation, relative AcNHS representation, or random numbers on the second, however, showed no closer relationship of crosslinking participation to solvent accessibility than to the negative control (random) numbers.

Based on this outcome, the validity of both crosslinking as a method for probing protein topology as well as collation of crosslinking representational data as means for expressing the resultant data is in doubt as executed and analyzed. Verification of this assumption is good practice, and this approach to the verification requires no additional procedures be run. The future applicability of crosslinking analysis methodologies to protein topological and/or structural investigations is in no small part a direct result of its reactivity towards the surface of substrate proteins, especially in the capacity of interprotein and intersubunit linking. Surface-preferential binding would increase the likelihood of observing such interactions with lessened interference from deep intraprotein or intrasubunit linking. It would be expected that the data from an investigation of surface area's correlation with crosslinking participation would be positive, but this was not conclusively shown by the traditional correlation analysis. There exist valleys of zero representation in crosslinked polypeptides, but the number of residues with zero such representation is assuredly small relative to the total number of residues located below the surface layer of the protein. A total of 27 residues of BSA were found to have zero representation in the data used to produce Figure 3.9, out of 610 residues considered. This is approximately 4.4% of the total protein's length, unlikely to represent all subsurface residues. More analysis is required to determine more conclusively if the assumption that solvent accessibility is positively related to crosslinking availability is valid.

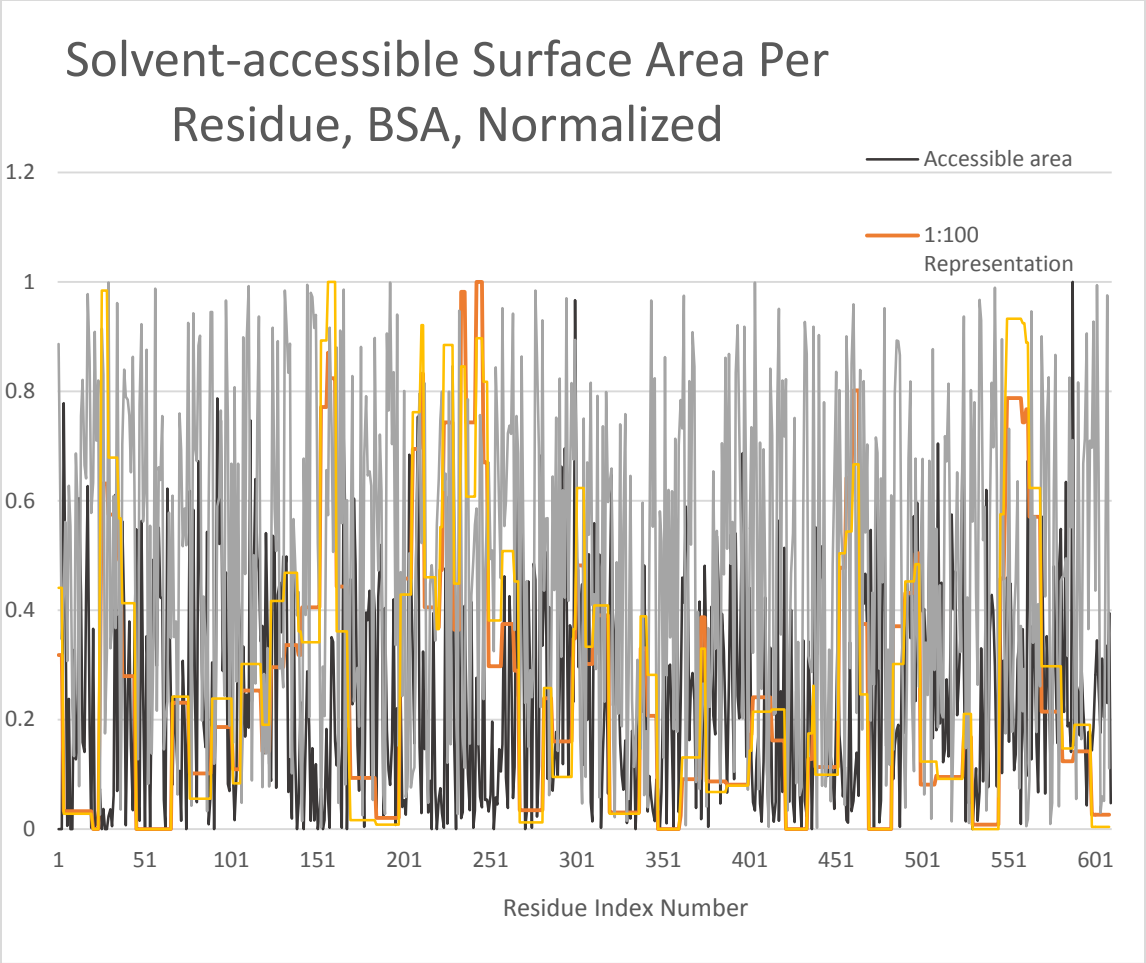


Figure 3.13 Representation and Solvent Accessibility: overlaid graphs by index number of crosslinked representation of both SuDP-linked and AcNHS control trials, solvent-accessible surface area, and randomly-generated values. Each set is normalized to a 0-1 scale.

VI: Script Analysis Revisited: Window Shifting

One other custom software method was used to analyze the crosslinked peptides' correlation with solvent-accessible surface area. One piece of code, written in Python, for a list of polypeptides (such as the one output by StavroX corresponding to crosslinked peptides) computed the average solvent-accessible surface area for each constituent amino acid residue of all polypeptides in the list. For each residue of each polypeptide, the solvent-accessible surface area was added to a master count, which was divided by the total number of amino acid residues considered this way from all polypeptides; note that this was not a computation of the average solvent-accessible surface area of each polypeptide in the list.

This was coupled to custom software which altered an input list of peptides originally from a specified protein by translating each amino acid residue by an input integer position from its original identity, with output as a new list of polypeptides, each the same length as the original input polypeptide it corresponds to, yet the entire polypeptide "shifted" a number of amino acid positions from its original identity. This shifted list was analyzed in the same way as the original list of polypeptides, using the same script. The average solvent-accessible surface area for the altered, shifted list of polypeptides was returned as output. Thus each input integer (the shift value) returns a determined change in the average (per amino acid residue) solvent-accessible surface area in a format suitable for plotting.

To this end, another function was written which, given an integer X, returns values of polypeptides "shifted" by each integer in the range of –X to X. This is an easily-graphed system of independent variable, the shift number, and its dependent result.

The purpose of this analysis method was to investigate if there existed a pattern in the solvent accessibility of the analyzed protein's structure and if this pattern could be held to be

predictive of that secondary structure. It would very reasonably be expected that the average solvent-accessible surface area of the amino acid residues of the set of residues constituting the more exposed portions of the protein's secondary structure would tend to be higher than the global average solvent-accessible surface area, but if modeled solvent accessibility can be assumed to correlate with the likelihood of being located in outer sections of the secondary structure is less clear. The analysis was conducted on data obtainable from an existing solved molecular structure of bovine serum albumin coupled with output from crosslinked-peptide identification software StavroX acting on in-house crosslinked bovine serum albumin mass spectrometry data. PyMOL's own analysis tools were used to produce the list of solvent-accessible surface areas that were used throughout the rest of the analysis from the structure file of bovine serum albumin, and the script which performed the analysis was written in-house.

The crosslinking reaction described in Chapter 2.III yields its results in the form of mass spectrometric raw data. Data in this format is all but impossible to interpret into useful information about the crosslinking events it may (or may not) represent, except perhaps to seasoned veterans, and the raw data files are quite large. The first step in their processing was utilizing Mascot to convert the raw mass spectrometric data to Mascot generic formal (file extension .mgf), the ThermoFisher proprietary format which StavroX requires for its input. Mascot is capable of identifying patterns indicative of peptides in the data which it acts on, but was in this use case employed only to convert the raw mass spectrometric data into a different format. Mascot can in fact be used to identify dead-end crosslinking events where a polypeptide is bound by the entire crosslinking agent unconnected from any second polypeptide, or cleaved-crosslinker bindings in which a portion of the crosslinking agent being used binds a polypeptide and is separated from the other portion of the agent, bound to another polypeptide or not, usually

by in-source fragmentation during the process of mass spectrometry, but this technique, while legitimate, was not employed.

Indeed, many software suites exist that are capable of enumerating post-translational modifications to polypeptide fragments in mass spectrometric datasets. Mascot certainly can (Creasy and Cottrell 2002), and in fact predates many of the alternative programs, which include PeaksPTM (Han et al. 2011), the open-source X! Tandem (Craig and Beavis 2003), and the similarly open-source SpectraST (Lam et al. 2008). Other programs exist specialized to detect polypeptides in mass spectrometric data analysis, but lack post-translational-modification detection that is customizable by the user, such as Sequest (Mann and Jensen 2003).

Regardless of the software used to preliminarily process the raw mass spectrometric data, the next step is to identify the polypeptides bound by crosslinking agents in the study. Other software purporting to fulfill its function does exist, and was examined (see Chapter 3.II for more details).

StavroX requires as input the protein's FASTA-formatted sequence file and the mostly-raw data in .mgf format, but it also must be tuned to the specific crosslinking agent being used for each set of data. This process is largely abstracted to the bare essentials of the crosslinking agent in question; the crosslinking agent's preferred substrate for each end (this is lysine, in the case of SuDP) and total molecular mass of the linker region as it will appear in the mass spectrometric data, i.e. without ester rings in the case of disuccinimidyl crosslinking agents such as those used here. The fine structure of the crosslinking agent used, including such factors as its R-chain(s), where applicable, while potentially influential of the reactivity and specificity of the agent in the assay, is not considered by StavroX. Indeed, the limited scope of the mass spectrometric data utilized by the program cannot display such influence in a vacuum.

Comparative studies can illuminate differences in binding patterns by crosslinkers possessed of differing sterics, structures, and R-chains; StavroX's abstraction of the crosslinking agent used in the study which produced the data fed to it is indeed a simplification, but a justifiable one with little negative impact on the software's results aside from the lack of consideration of non-specified crosslinker bindings. Disuccinimidyl agents preferentially bind lysine, it is true, but do have the capacity to bind other amino acid residues, and this possibility is not considered by StavroX.

Analytically perfect or not, once the software has acted, it returns data in a much more human-comprehensible format than the one used to encode the raw mass spectrometric data. This data is in the form of a spreadsheet containing information in tidy rows about each putative crosslinking event. This spreadsheet-like output format is relatively comprehensive; it includes the mass/charge ratio for each peak corresponding to a purported pair of linked polypeptides, charge and mass for same, metadata concerning which mass spectrometry scan yielded the crosslinking event, and a confidence score. As part of its polypeptide-identification duties, it also returns the sequence of each of the two crosslinked polypeptides, the index numbers of the crosslinked lysine residues within the crosslinked polypeptides, and, using the protein's FASTA-formatted file as information on the primary sequence of the protein being analyzed, the numerical indices of the polypeptides linked in each event, thereby eliminating the potential ambiguity that would be associated with returning exclusively the polypeptide sequences; any given polypeptide sequence could appear more than once in the primary structure of the protein in question, though the likelihood of such a duplication is low for polypeptides of larger length. The majority of polypeptides returned by the software are long enough that duplication is relatively unlikely, but polypeptides as short as three residues are included at not insignificant

frequency. Thus, these index numbers will prove useful in later steps where consideration of the proper instance of a potentially non-unique polypeptide by software is a concern. If a given polypeptide sequence appears once in the outermost, most solvent-accessible layer of a protein and once deep in its inner structure which is much less solvent-accessible, the ability to provide a distinction between the two to downstream analytical software becomes clear.

What has been discussed is StavroX's on-screen output. Complications arose when attempting to render it machine-readable. The program does support output to external file, in the comma-separated values (.csv) format. This format is a simplistic but effective format for data storage and natively read by popular spreadsheet program Microsoft Excel as well as easily read in by homebrewed scripts. Unfortunately, the inbuilt export writes out only the total mass of each identified polypeptide pair plus crosslinker conglomerate and the charge range of the peak used to identify it. This is not very useful data. It can be used in broad strokes to compare two crosslinking trials; across two attempts at crosslinking the same protein species, it could be expected that identical masses would appear in the output by StavroX of both trials at a greater rate than if they were different proteins, but for any analysis deeper than this, this scale of output is simply inadequate. But by copying and pasting the screen to an accommodating spreadsheet program, such as Microsoft Excel, which has the capability of exporting its contents to .csv format, the full scope of StavroX's potent output can be stored and passed on to further steps of analysis in the standard format. The need of using such an unconventional and unintuitive work-around is unfortunate.

Once the crosslinking data was successfully ensconced in .csv format, the real work of window-shifting analysis was initiated. This work begins with preparation and aggregation of all the required data from various sources. The in-house written window-shifting analysis scripts

require a list of polypeptides, a primary sequence for the protein in question, and a list of the solvent-accessible surface areas of each amino acid residue in the protein. This is a somewhat different list than StavroX requires. The first requirement, the list of polypeptides, is in fact supplied by StavroX itself; the .csv generated from its output is processed by a Python function which reads each line and adds the polypeptides to a list in memory. This function necessarily strips formatting characters from the .csv entry before the final addition to the list, yielding a battery of easily-dealt-with text strings. The required protein sequence is culled from the FASTA by another function which strips it of all formatting. The last requirement requires the most introduction; structural modeling software PyMOL can for any given portion of a protein return the solvent-accessible surface area in square Angstroms, and this capability was employed on a per-residue basis to yield a list of areas for any given residue in the host protein. This requires an existing structure for the host protein, as a simple one-dimensional polypeptide chain of the proper sequence would yield identical solvent-accessible surface areas for each residue of the same type. PyMOL is itself a Python program, and can perform the minor formatting the list of areas requires without an external helper function, and stores the list as a .csv.

Once the various helper functions and PyMOL have assembled the data into the proper format, the main function can be executed. This script iterates through each entry of the polypeptide list and locates its sequence in the primary protein sequence. This is where ambiguity could occur if not for StavroX's output of the index number for each polypeptide, if the sequence were not unique in the protein. Regardless, for each residue of each polypeptide in the list of all crosslinked peptides, the script adds its solvent-accessible surface area (drawing from the PyMOL's list of solvent-accessible surface areas) or, vitally, the solvent-accessible surface area of the residue located at the original residue's index number plus an integer skew

value, to a master count, and also increments a separate count to track how many residues have been considered. After all polypeptides have been considered, it divides the total solvent-accessible surface area of all considered residues by the number of residues considered, yielding an average solvent-accessible surface area value for the set of amino acid residues present in at least one crosslinked peptide.

This figure is not entirely without value by itself; a reported average solvent-accessible surface area value for crosslinking participant amino acids greater than the average solvent-accessible surface area value for all constituent amino acids would indicate that the software (both the house-made scripts and StavroX) is functioning correctly, as the crosslinker would be expected to bind at a greater rate to more solvent-accessible lysine residues simply because it is borne by the solvent. Window shifting makes use of an extension of this principle to beyond proving the obvious. The set of all polypeptides found participating in a crosslinking event would have a certain average solvent-accessible surface area, but that same set offset in the index numbers of the peptides by a number of positions would be expected to have a different average solvent-accessible surface area. Thus, if a given polypeptide identified by the program could be represented as spanning residues number 29 through 36, altering the index numbers by +1 would yield a slightly different peptide spanning residues 30 through 37. As this set would less well represent the preference of crosslinker binding (representing an alteration of the set which was experimentally found as preferable via actual crosslinker binding), the average value found would be expected to be lower.

This principle underlies the window-shifting method of analysis used. The direction of shift (represented by the positive or negative character of the integer used to skew the window of consideration for the input polypeptide list) should not significantly influence the tendency of

any offset to result in a reduction of the average solvent-accessible surface area, at least in theory. It would also be expected that the further from 0 the skew number used, the more the average solvent-accessible surface area would be expected to decrease, up to the point that the polypeptides formed by shifting the numbers of the original list has no more meaningful relationship to the originals and their likelihood of location in a relatively solvent-adjacent position in the host protein. It could be expected that the average solvent-accessible surface area would not invariably decrease with every step farther from 0. Once sufficiently removed from the original set of polypeptides, it is very possible that the new set of considered polypeptides could collectively emerge from the protein structure and increase their average exposed surface area to some degree relative to the set immediately prior, though the odds against the set stumbling into a configuration of greater average solvent accessibility than the naturally-occurring experimentally-determined one is highly unlikely. To summarize, it would be expected that the polypeptide set handed down by StavroX would represent one of the most favorable of all sets with respect to maximizing solvent-accessible surface, and altering this set by "shifting" the window of consideration for all constituent peptides would move the set from a state of maximized accessible area.

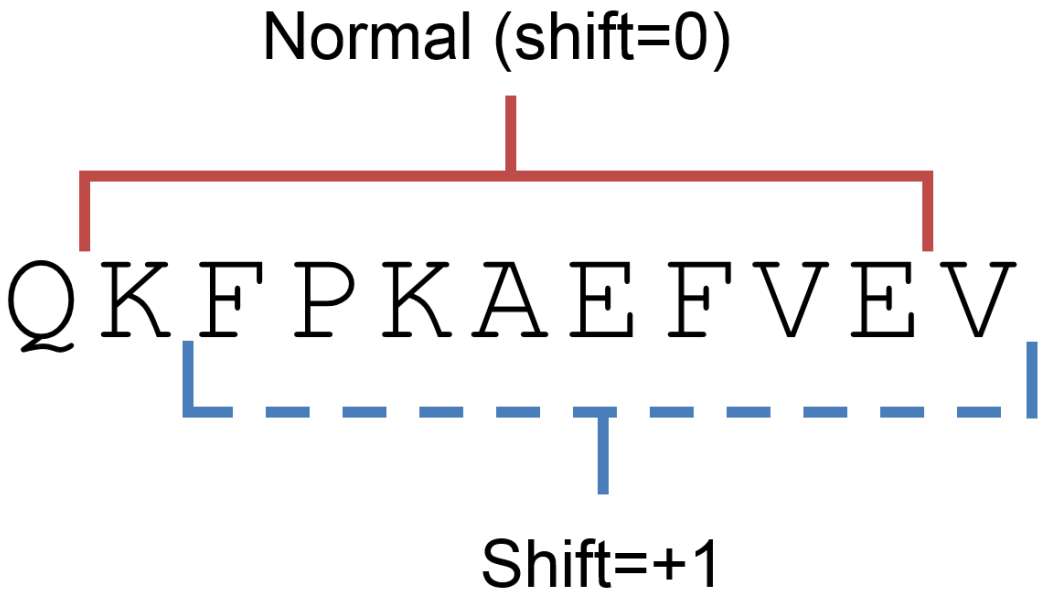
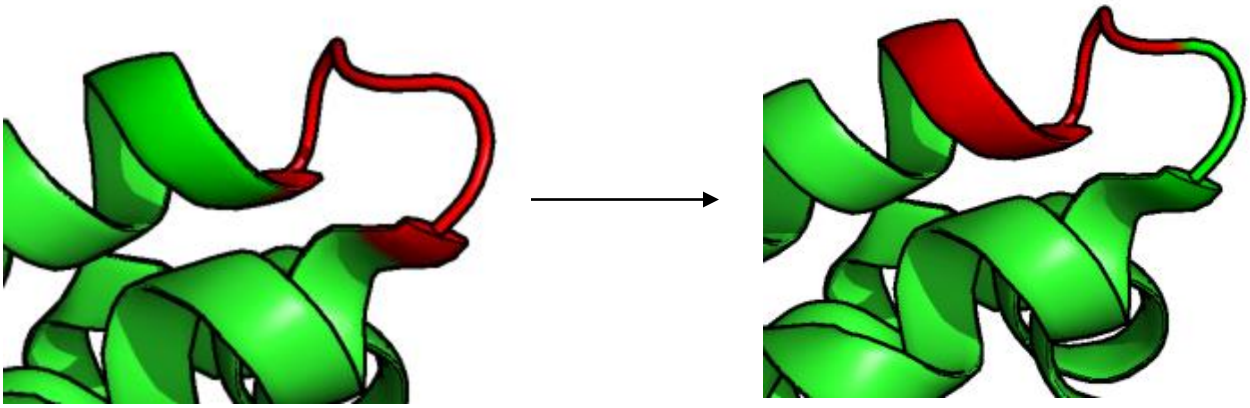


Figure 3.14 Window Shifting: the original "window" is the original listed polypeptide, shown here as a red-bracketed portion of a larger whole. The window is shifted by +1 to a blue-bracketed portion of the same length but one amino acid position further along the chain.



(Figure generated by PyMOL (Schrödinger, 2014))

Figure 3.15 Window Shifting on a Protein: the concept of window shifting illustrated on a three-dimensional protein model. The red region on the left protein represents a “window” of a polypeptide of some length in its original position. The red region on the right is the same number of polypeptides, but shifted by two positions further along the protein.

Is this borne out by the results of the crosslinking data analysis conducted on bovine serum albumin? The answer is inconclusive. To probe this line of investigation, more accessory functions were developed in Python. The master skew window calculation program accepts but one integer to use as the skew value, and returns the average solvent-accessible surface area for the set of peptides returned for that skew value. If the integer passed to the function is 0, it returns the average solvent-accessible area for the original set of peptides (as the shifted set is the same). For ease of interpretation, an accessory function was written to repeatedly call the original master skew calculation function to collect all average solvent-accessible areas for all integer skew values of a specified range window. -100 through +100 was used. At higher shift values, the number of positions shifted may cause many of the index numbers of the polypeptides of the original set to exceed the maximum or minimum index numbers of the host protein; this was accounted for by omitting any nonexistent residues from both the master count of total solvent-accessible surface area as well as the count of residues considered, effectively removing them from consideration without impacting the average as established by the other, existing polypeptides. The accessory function also represented each average area calculated by subtracting the original average solvent-accessible surface area (skewed 0 positions) from the new average. Thus each integer skew possibility was expressed as a difference in accessible area from the original configuration at a shift of zero.

Represented this way, it would be expected if solvent-accessible surface area correlates with crosslinking agent accessibility/reactivity, that the graphed relationship between skew number and its change to the average solvent-accessible surface area would exhibit a global maximum of 0 change at a skew value of ±0, the origin, as any alteration to the original peptide set would result in a net decrease of solvent accessibility in the new set, decreasing on either side

to a horizontal asymptote at the value corresponding to the average crosslinking event-participating solvent-accessible surface area minus the average total such surface area.

These hypotheses were mostly observed. A distinct global maximum was indeed observed very close to the origin as predicted. However, this peak occurred not at the origin precisely but located at a shift value of +11 amino acid residues, not at ±0, and at an average solvent-accessible surface area change not of ±0 but at approximately +7 squared Angstroms. However, as predicted, the average solvent-accessible surface area did indeed fall off on either side of this global maximum, somewhat dramatically. It also did have periods of concomitant increase of both average accessible area and distance from 0, and though local maxima some distance above the global minimum did occur, none of the resulting local maxima ever exceeded the roughly 0-centered high point.

The explanation for the experimentally-determined global maximum being so discrepant from the expected values at that particular point is cryptic. As it would be expected that the set of all polypeptides found participatory in crosslinking events would necessarily be a set possessed of maximum solvent accessibility due to the obvious and previously-discussed relationship between solvent-accessible surface area and crosslinker accessibility, a shifted window of consideration of this set of polypeptides would logically result in a lower average solvent or crosslinker accessibility regardless of the direction in which the window was shifted, yet the plotted results suggest an asymmetricality at the point of 0 window shift. Clearly at least one assumption was incorrect. Though it is not verifiable as the single cause of this effect, the interaction of trypsin's cleavage pattern and the disuccinimidyl-succinamyl-aspartyl-proline crosslinking agent used in the experiment may provide a partial explanation for this observed oddity. Following crosslinking using the SuDP, trypsin was added to hydrolyze the crosslinked

protein into smaller fragments pending mass spectrometric analysis and collection of data on the crosslinking patterns. Trypsin's cleavage pattern is predictable and specific; it cleaves its polypeptide substrate at the carboxylic termini of arginine and lysine residues. The binding preferences of SuDP (and similar disuccinimidyl linking agents) are likewise known and predictable; they preferentially attack lysine residues. Additionally, the software used to identify crosslinking events identifies only lysine-binding events. Thus it is an incorrect assumption that the set of all polypeptides associated with crosslinking events would provide equally-valid information on all of their constituent residues' solvent accessibility. For each such considered polypeptide, only one residue is the target of the crosslinking agent, and, moreover, that one residue's location and identity is not completely random. Far from chance selection, in fact, as trypsin's cleavage preference will very often leave a lysine residue at the C-terminus of any given polypeptide hewn from a host protein by its action. Lysine is not an uncommonly-represented amino acid residue in bovine serum albumin, or indeed in most proteins, but the small average length of identified crosslinked polypeptides limits the number of lysine residues present in most of them, and this greatly increases the odds that the crosslinking event which identified them to the software occurred at their trypsin-created terminal lysine residue. Accepting this as truth leads to the realization that the list of all crosslinking event-associated polypeptides is not strictly a list of the most solvent-accessible polypeptides but a list of the most solvent-accessible lysine residues and their closest associates, with a distinct preference for associated aminoacyl residues located in the N-terminal direction of the lysine which was bound by the crosslinking agent. This makes clear the possibility that shifting the window of consideration of polypeptides would produce differing results based on the direction in which it was shifted.

An additional functionality added to the window-shifting suite of programs was the capacity to filter considered polypeptides from the input master list based on an input desired minimum (and/or maximum) score of confidence as recorded and output by StavroX. For a given polypeptide from the list, if it was found in a crosslinking event identified with confidence score greater than an input minimum value (the default minimum being a confidence score of 0) or, less frequently used, less than an optional input maximum value, it is included in the new list of polypeptides to subject to window-shifting. The window-shifting analysis scripts account for polypeptides from the list which do not meet the minimum score of confidence or exceed the maximum score in much the same way they account for polypeptides shifted beyond the frame of the host protein; these polypeptides are not considered at all in the average solvent-accessible surface area tally or the master count, so they do not interfere with the results from the polypeptides which do meet the criteria.

When the list of polypeptides was filtered to exclude all crosslinking-participatory polypeptides below a confidence score of 50, the results changed remarkably little. Though the original graph accounted for polypeptides with confidence scores above 0 (up to and including the single polypeptide with the highest confidence score of 221) and restricting the scope of the considered polypeptides to only scores of 50 and above represents a total score range of 221 (221 to 0) decreasing to 171 (221 to 50), the number of considered polypeptides dropped by a more than proportionate amount; the total number of polypeptides with score greater than 0 was 16,146, versus a total number of polypeptides with score greater than 50 of only 677, representing only slightly more than four percent of all polypeptides reported by StavroX. This speaks to an agreement across a range of confidence scores that the observed results are

consistent without requiring extreme discrimination in selecting which crosslinked polypeptides to consider for the process of window-shifting.

Window-shifting analysis presents an alternative avenue of analysis of crosslinked protein datasets, but the true scope of its usefulness is unclear. It has been shown, in concert with StavroX, that for an existing crosslinked protein dataset, it can be used to provide evidence that the particular set of reported polypeptides in that set do represent some of the most solvent-accessible polypeptides, on average, in the protein, but this information is not in itself very useful. It is proof that the crosslinking agent is conveyed to its target by the solvent, but this did not require proving. Future usefulness might lie less in using the method on existing datasets and more in using it as a predictive tool of protein structure for proteins whose structure is not yet or not well determined by existing other methods such as X-ray crystallography or cryo-electron microscopy. To accomplish this, a list of polypeptides might be drawn from random samplings of the target protein's primary sequence and ensured to be of appropriate lengths within the range in which the lengths of actual identified polypeptides in an existing real dataset vary. This list could be input to the window-shifting algorithms without problem, no mechanism existing within the program to determine whether its data is reliable, and an average solvent accessibility diagram produced. Used in comparison with other randomly-selected sets and in concert with a list of the solvent accessibilities of each residue in an unfolded chain, it could provide information on which portions of the protein are likely to lie on the outside of the folded protein. This, however, adds complication when taken relative to simply estimating likely intraprotein location based on solvent accessibility. Furthermore, the computations associated with performing window-shift analysis on an amply-sized window of shift values is not trivial. For one structure, a thorough calculation of the skew map surrounding a value of 0 might extend 100

integers in either direction. This is a total of 201 iterations through the base function, which, while not problematically time-consuming in itself, rapidly sums to a significant time spent processing when run hundreds of times. Not even the time spent to create a single luxuriously extensive skew map is oppressive. In testing, the program required several minutes at most for the sizes tested, which never exceeded a few hundred skew values and, indeed, could not usefully be expanded beyond several hundred skew values in a single direction without reducing the comparable polypeptide set to vanishingly small sizes, so single skew maps would never pose a computational problem. A several-minute runtime would however pose problems for applications in which the skew map function had to be run many times. Analysis of randomly-generated polypeptide sets to assess the feasibility of many putative protein structures, a use case which could be plausible if the function were employed by an application aiming to determine structure of a folded protein via random structure assessment, would take a prohibitively long time if window-shift analysis were employed. Nonetheless, as a complement to existing methods, window shifting may have value as a tool for assisting in the determination of folded protein forms and could even more easily serve as an assessment tool for analyzing the plausibility of putative protein structures created in other structural determination programs.

In all, window-shifting analysis was a useful tool to assess the relationship between crosslinker accessibility and solvent accessibility, but its potential usefulness to those attempting fine structure prediction should not be overstated. Its application in determining viability of randomly-generated putative three-dimensional protein structures is somewhat limited by virtue of requiring a comprehensive list of the solvent-accessible surface areas of the substrate protein's constituent amino acid residues in the protein's natural folded confirmation as input, as well as the relatively large gestation time for a single skew map limiting its ability to be called as many

times in as short a time as such a random method would demand. The calculations associated with correcting the bias inherent in using that peculiar combination of a protein-digesting enzyme which produces C-terminal lysine residues and a crosslinker which acts, as far as the software used to interpret the mass spectrometric crosslinking data is concerned, exclusively on lysine are not necessarily entirely reliable. Still, it provides dramatic evidence that the list of crosslinked polypeptides produced by StavroX is a plausible one based on their positioning in the folded structure of the protein. This functionality cannot be parlayed efficiently into other domains of protein structural research, however, leaving this method of analysis better suited as a complementary tool of second opinion for existing methods of structural determination.

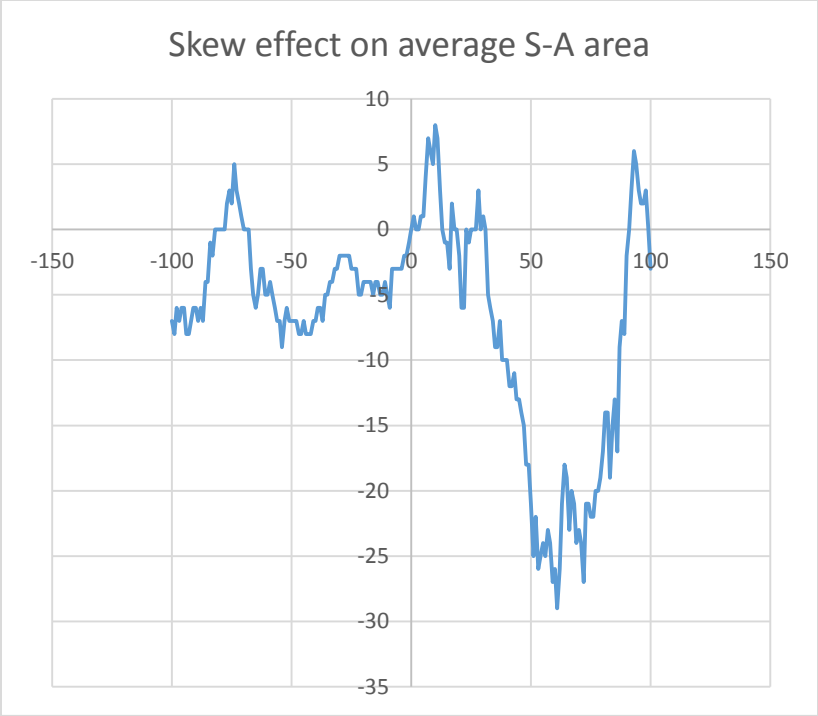


Figure 3.16 Skew Effect Mapping: map of the effect of positive and negative skew values on the change of average solvent-accessible surface area of the original set (at a skew value of ±0) from its original value. The origin represents no change from the original set and is thus 0,0. Skew values are shown for changes of up to 100 positions in either direction.

CHAPTER 4 – Conclusions

I: In Brief

Chemical crosslinking is a versatile tool of analysis with applications in several subfields of protein structural studies. It could be held to be comparable in scope of application to such other methods of structure determination as x-ray crystallography, nuclear mass resonance spectroscopy, and cryo-electron microscopy, though it has many differences from all of these methods that both make it less effective in many cases but also confer niche effectiveness for a small but potentially useful subset of potential analyte proteins or protein mixtures. Problems with accuracy exist, but the method's entire procedural breadth consists of mixing a variable ratio of the crosslinking agent with the protein to be analyzed, making it an appealingly simple method of analysis to pursue.

Its complexity is loaded in the post-acquisitional analysis, which does require specialized software and fine operation to interpret the mass spectrometric data, which is also required to be of a certain quality, but the ease of the initial step of actually effecting the crosslinking can hold great appeal relative to more laborious methods such as x-ray crystallography even though the data returned is of decidedly lower quality, as crystallography and, to lesser degrees, other methods of structural determination are beset by requirements as to the mandated state or properties of the protein to be analyzed by that method. Crosslinking does require solubility, which for some proteins, and especially membrane-bound proteins, can pose some degree of challenge. Adjusting solvent levels of salt, pH, and chelators can greatly assist in stabilizing a variety of proteins, though the resultant solvent environment must also continue to permit crosslinking activity to proceed, and, if the mixture to be crosslinked is heterogenous, must be

amenable to all proteins in the mixture. The postacquisitional step is largely contingent on available software, and the quality of software is steadily improving.

As reagents, the disuccininimidyl-succinamyl-aspartyl-proline used in these analyses also benefits from an ease of synthesis. It is produced by solid-phase peptide synthesis, which is very well-documented and straightforward, requiring little in the way of specialized equipment or esoteric or dangerous accessory reagents. This and the easy procedure to actually effect the crosslinking discussed above, coupled with greater tractability across a wide variety of analyte proteins, make the method a potential low-cost and -effort alternative to more-involved methods of structural determination, especially when analyzing a protein that has undergone some change in its environment (e.g. change in pH due to affinity chromatrography as part of protein purification as seen with biomanufacturing of monoclonal antibodies.)

The analysis phase of the application of crosslinking analysis to structural determination is, however, more complicated. The data produced is in mass spectrum format, which is largely unhelpful without the aid of computerized interpretation and transformation to more-comprehensible forms. This does mean that highly specialized software must be used, because the mass spectrometric data must be transformed not only to protein sequences with a mass addition (the crosslinking agent itself), but discontinuous protein sequences, due to the nature of bifunctional crosslinking agent binding. Indeed, the sequences may not lie near each other on the same subunit, or on adjacent subunits, or even within the same parent protein, if multiple proteins are present, for crosslinking agents which bind such nonspecific targets as lysine residues. Lysine is ubiquitous throughout the great majority of proteins. Specialized software is required to interpret this data, written specifically for the purpose. More generally available software can detect only monofunctionalized crosslinking agent binding, so-called “dead-end”

linkage events, and the most valuable structural data the crosslinking method can yield comes not from this but from confirmed linking of two disparate peptides, and the discontinuity cannot be processed by these conventional programs. Custom software must be written or brought to bear, and there are several solutions available for the identification of links, and several others for still more specialized tasks such as radiolabelled crosslinker analysis and visualization of a list of putative, independently-generated list of crosslinks.

Existing software packages for the analysis of crosslinking mass spectrometric data may be supplemented with more-specialized and limited techniques for faster analysis or verification of data. The window-shifting method described in this thesis attempts to assess whether an input list of polypeptides appear to have a maximum solvent accessibility, which is a valuable test to conduct given that a list of crosslink-participatory polypeptides would be expected to have such a maximum, and therefore can serve as a speedy gauge of the set's plausibility. Manual selection of obvious residues part of some set, such as interface or surface residues, and tabulating their representation in crosslinked peptides reported by the software can also be used to report on the relationship between that set and crosslinking participation in several ways.

These small-scale, computationally uncomplicated applications illustrate the strengths of analysis by crosslinking as well as the differences from traditional methods of structure determination. The results are not rich in information without significant processing, but limited and specialized processing can be utilized to provide data which trades comprehensiveness for relative ease of acquisition. It is easy to build a small battery of tests for study of various aspects of the structure or character of the analyte protein.

Verification of many of the assumptions associated with utilization of crosslinking agents to probe structure was an integral part of this work. It cannot, for example, be assumed without

verification that the crosslinking agents used would bind with any preferentiality to the outermost surfaces of the target protein(s). It would certainly be predicted, but verification is good practice and ensures validity of subsequent investigations which rest on the assumption. Much effort, particularly in the analysis phase of the procedures, was devoted to ensuring that any information which could apparently be drawn from the data was indeed based on valid premises and thus valid itself, to mixed results. Better verification of this core assumption is a primary target of future refinement of this work.

A complication which accompanies the use of crosslinking detection software, which is a requirement, is that it demands analysis of its output to verify that its results are more accurate than what would be obtained simply due to random chance. This is the reason for the decoy analysis; more verification. The necessity of this is a highly noteworthy drawback when assessing crosslinking analysis's usefulness relative to other methods of protein structure probing, such as nuclear magnetic resonance spectroscopy. Virtually every post-acquisition technique requires assessment to ensure that the conclusions which can be drawn from it are valid, whereas other methods of structural determination yield data which may provide more information more immediately or with less subjective processing.

In conclusion, the method of structure probing via chemical crosslinking has many flaws that hurt its comparison with other methods of structure determination. It offers low-detail results which must be passed through multiple layers of fallible analysis to yield up their information. However, the results which it does provide are easier and require less effort to obtain than are the ones produced by other methods. Thus it is reasonable to conclude that the method is a useful one most specifically as a supplement to other methodologies and a tool of verification of structures solved otherwise, not a primary instrument of *de novo* structure investigation.

II: Looking to the Future

This thesis has briefly outlined the current state of the art in chemical crosslinking analysis of protein structure, but it is also intended to guide the reader to a greater understanding of the direction the art may take in the future. The research discussed likewise has clear extensibility in a few directions not taken for a variety of reasons – time, sample availability, equipment availability, and expense chiefly. It can surely be surmised from reading this thesis that the software used to identify crosslinked polypeptides in the mass spectrometric data is of foundational importance to the quality of the conclusions that can be drawn from the method. A better or simply a different software package than StavroX with the same functionality would provide a second source of crosslink information to be fed to the scripts, and greater certainty in the identified links and/or a larger list of links would be a boon to using the data.

Alternative methods of gathering the data provide a second frontier for greater usefulness of the method. Isotopically-coded crosslinking agents (See Chapter 3.II) offer an option in the procedure which could result in easier and more confident identification of crosslinked polypeptides while not requiring any change in the interpretation workflow. This was not attempted due to the difficulty in obtaining isotopically-coded crosslinking agents and working with them, but it offers promise for later research.

Correlation of crosslinking accessibility data with hydrogen-deuterium exchange (HDX) data is a potential avenue to verify the validity of the results yielded by crosslinking analysis, HDX being itself a method to assess the solvent accessibility of a protein's constituent amino acids on a residue-by-residue basis. This was also not attempted due to the need for both HDX and crosslinking data on the same protein, and no protein which was crosslinked in the course of this study had relevant HDX data at hand, and no HDX data could be obtained for any protein

which was at hand for analysis by crosslinking. This analysis stands as a tempting undertaking for future researchers with access to both HDX and crosslinking data.

The possibility of use of crosslinking agents with different steric and polar qualities is also a route of research which involves little change in protocol or analysis methodology, yet may bear fruit. A substitution of one of the amino acid constituents of the crosslinking agents, as valine for aspartic acid, where the side chain of the replaced amino acid has different polar character, charge, or steric bulk, would yield a crosslinking agent with theoretically altered binding affinity which is no harder to synthesis and which may be analyzed with the same methods used for SuDP or the original crosslinking agent. A change in the binding affinity/reactivity which is reflected in the relative crosslinked polypeptide detection rates would recommend the method as a more versatile surface structure probe and expand its usefulness. This was attempted by the author of this thesis with disuccinimidyl-succinamyl-valyl-proline, but instrument troubles delayed results, precluding their inclusion in this thesis.

REFERENCES

Alexander, P., and H. Moroson. "Cross-linking of Deoxyribonucleic Acid to Protein following Ultra-Violet Irradiation of Different Cells." *Nature* 194.4831 (1962): 882-83.

Argo, Andrew S., Chunxiao Shi, Fan Liu, and Michael B. Goshe. "Performing Protein Crosslinking Using Gas-phase Cleavable Chemical Crosslinkers and Liquid Chromatography-tandem Mass Spectrometry." *Methods* 89 (2015): 64-73.

Bragg, Philip D., and Cynthia Hou. "Chemical Crosslinking of α Subunits in the F1 Adenosine Triphosphatase of Escherichia Coli." *Archives of Biochemistry and Biophysics* 244.1 (1986): 361-72.

Cavanagh, John, Fairbrother, Wayne J., and Palmer, III, Arthur G.. Protein NMR Spectroscopy: Principles and Practice (2). Burlington, US: Academic Press, 2010.

Craig, Robertson, and Ronald C. Beavis. "A Method for Reducing the Time Required to Match Protein Sequences with Tandem Mass Spectra." *Rapid Communications in Mass Spectrometry* 17.20 (2003): 2310-316.

Creasy, David M., and John S. Cottrell. "Error Tolerant Searching of Uninterpreted Tandem Mass Spectrometry Data." *Proteomics* 2.10 (2002): 1426-434.

Dickson, J.m., et al. “Development of a Coating Technique for the Internal Structure of Polypropylene Microfiltration Membranes.” *Journal of Membrane Science*, vol. 148, no. 1, 1998, pp. 25–36.

Eng, Jimmy K., Ashley L. Mccormack, and John R. Yates. "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database." *Journal of the American Society for Mass Spectrometry* 5.11 (1994): 976-89.

Farmer, Terry B., and Richard M. Caprioli. "Assessing the Multimeric States of Proteins: Studies Using Laser Desorption Mass Spectrometry." *Biol. Mass Spectrom. Biological Mass Spectrometry* 20.12 (1991): 796-800.

Fraenkel-Conrat, Heinz, Mitzi Cooper, and Harold S. Olcott. "The Reaction of Formaldehyde with Proteins." *Journal of the American Chemical Society* 67.6 (1945): 950-54.

Ganten, D., and Klaus Ruckpaul. "Multidimensional NMR Spectroscopy." *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*. Berlin: Springer, 2006. 1204-208.

Götze, Michael, Jens Pettelkau, Sabine Schaks, Konstanze Bosse, Christian H. Ihling, Fabian Krauth, Romy Fritzsche, Uwe Kühn, and Andrea Sinz. "StavroX—A Software for Analyzing Crosslinked Products in Protein Interaction Studies." *Journal of the American Society for Mass Spectrometry* 23.1 (2011): 76-87.

Han, Xi, Lin He, Lei Xin, Baozhen Shan, and Bin Ma. "PeaksPTM: Mass Spectrometry-Based Identification of Peptides with Unspecified Modifications." *J. Proteome Res. Journal of Proteome Research* 10.7 (2011): 2930-936.

Hershey, A. D. “Nucleic Acid Economy In Bacteria Infected With Bacteriophage T2: Ii. Phage Precursor Nucleic. Acid.” *The Journal of General Physiology* 37.1, (1953): 1–23.

Hockensmith, J. W., et al. "Laser Cross-Linking Of Nucleic Acids to Proteins. Methodology and First Applications To The Phage T4 DNA Replication System." *Journal of Biological Chemistry* 261.8 (1986): 3512-518.

Jiang, Wen, Matthew L. Baker, Joanita Jakana, Peter R. Weigele, Jonathan King, and Wah Chiu. "Backbone Structure of the Infectious Îµ15 Virus Capsid Revealed by Electron Cryomicroscopy." *Nature* 451.7182 (2008): 1130-134.

Jonge, Niels De, and Frances M. Ross. "Electron Microscopy of Specimens in Liquid." *Nature Nanotechnology* 6.11 (2011): 695-704.

Käll, Lukas, John D. Storey, Michael J. Maccoss, and William Stafford Noble. "Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases." *Journal of Proteome Research* 7.1 (2008): 29-34.

Keller, Andrew, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. "Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search." *Analytical Chemistry* 74.20 (2002): 5383-392.

Komolov, Konstantin E., Yang Du, Nguyen Minh Duc, Robin M. Betz, João P.G.L.M. Rodrigues, Ryan D. Leib, Dhabaleswar Patra, Georgios Skiniotis, Christopher M. Adams, Ron O. Dror, Ka Young Chung, Brian K. Kobilka, and Jeffrey L. Benovic. "Structural and Functional Analysis of a Î² 2 -Adrenergic Receptor Complex with GRK5." *Cell* 169.3 (2017).

Kosinski, Jan, Alexander Von Appen, Alessandro Ori, Kai Karius, Christoph W. Müller, and Martin Beck. "Xlink Analyzer: Software for Analysis and Visualization of Cross-linking Data in the Context of Three-dimensional Structures." *Journal of Structural Biology* 189.3 (2015): 177-83.

Lam, Henry, Eric W. Deutsch, James S. Eddes, Jimmy K. Eng, Stephen E. Stein, and Ruedi Aebersold. "Building Consensus Spectral Libraries for Peptide Identification in Proteomics." *Nature Methods* 5.10 (2008): 873-75.

Mann, Matthias, and Ole N. Jensen. "Proteomic Analysis of Post-translational Modifications." *Nature Biotechnology* 21.3 (2003): 255-61.

Miron, Talia, and Meir Wilchek. "A Spectrophotometric Assay for Soluble and Immobilized N-hydroxysuccinimide Esters." *Analytical Biochemistry* 126.2 (1982): 433-35.

Novabiochem. (2009). “2008-2009 Catalog”

Mobli, Mehdi, and Jeffrey C. Hoch. "Nonuniform Sampling and Non-Fourier Signal Processing Methods in Multidimensional NMR." *Progress in Nuclear Magnetic Resonance Spectroscopy* 83 (2014): 21-41.

Nygård, O, and H Nika. "Identification by RNA-Protein Cross-Linking of Ribosomal Proteins Located at the Interface between the Small and the Large Subunits of Mammalian Ribosomes." *The EMBO Journal* 1.3 (1982): 357–62.

Pettersen, EF, TD Goddard, CC Huang, GS Couch, DM Greenblatt, EC Meng, and TE Ferrin. "UCSF Chimera - A Visualization System for Exploratory Research and Analysis." *J Comput Chem* 25(13) (2004): 1605-12.

Perkins, DN, DJ Pappin, DM Creasy, and JS Cottrell. "Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data." *Electrophoresis* 20(18) (1999): 3551-67.

Petrotchenko, EV. "Isotopically Coded Cleavable Cross-linker for Studying Protein-Protein Interaction and Protein Complexes." *Molecular & Cellular Proteomics* 4.8 (2005): 1167-79.

PyMOL Molecular Graphics System, Version 2.0, Schrödinger, LLC, 2014.

Rappsilber, Juri, Symeon Siniossoglou, Eduard C. Hurt, and Matthias Mann. "A Generic Strategy to Analyze the Spatial Organization of Multi-Protein Complexes by Cross-Linking and Mass Spectrometry." *Analytical Chemistry* 72.2 (2000): 267-75.

Seebacher, Jan, Parag Mallick, Ning Zhang, James Eddes S., Ruedi Aebersold, and Michael Gelb H. "Protein Cross-Linking Analysis Using Mass Spectrometry, Isotope-Coded Cross-Linkers, and Integrated Computational Data Processing." *Journal of Proteome Research* 5.9 (2006): 2270-282.

Sinz, Andrea. "Chemical Cross-linking and Mass Spectrometry for Mapping Three-dimensional Structures of Proteins and Protein Complexes." *Journal of Mass Spectrometry* 38.12 (2003): 1225-237.

Soderblom, Eric J. "Collision-Induced Dissociative Crosslinking Reagents and Methodology for the Structural Analysis of Proteins and Protein-Protein Interactions Using Tandem Mass Spectrometry." (2008): 37-38.

Soderblom, Erik J., and Michael B. Goshe. "Collision-Induced Dissociative Chemical Cross-Linking Reagents and Methodology: Applications to Protein Structural Characterization Using Tandem Mass Spectrometry Analysis." *Analytical Chemistry* 78.23 (2006): 8059-068.

Sun, Haipeng, et al. "Co-Delivery and Controlled Release of Stromal Cell-Derived Factor-1α Chemically Conjugated on Collagen Scaffolds Enhances Bone Morphogenetic Protein-2-Driven Osteogenesis in Rats." *Molecular Medicine Reports* 14.1 (2016): 737-45.

Wold, Finn. "[57] Bifunctional Reagents." *Methods in Enzymology Enzyme Structure, Part B* (1972): 623-51.

Wong, Shan S. *Chemistry of Protein Conjugation and Cross-linking*. Boca Raton: CRC, 1991. 61.

Wu, Cheng-Hsien, et al. "Sequence-Specific Capture of Protein-DNA Complexes for Mass Spectrometric Protein Identification." *PLoS ONE* 6.10 (2011).

APPENDIX

Appendix A – Abbreviations Used

ABC: ammonium bicarbonate
ACN: acetonitrile
AcNHS: acetic acid n-hydroxysuccinimide
BSA: bovine serum albumin
DCM: dichloromethane
DIC: diisopropyocarbodiimide
DMF: dimethylformamide
DMSO: dimethylsulfoxide
DTT: dithiothreitol
EM: electron microscopy
FA: formic acid
HDX: hydrogen-deuterium exchange
HOBT: hydroxylbenzotriazolemonohydrate
HPLC: high-performance liquid chromatography
IAA: iodoacetamide
LC: liquid chromatography
MS/MS: tandem mass spectrometry
NHS: n-hydroxysuccinimide
NMR: nuclear magnetic resonance
SPPS: solid-phase peptide synthesis
SuDP: disuccinimidyl-succinamyl-aspartyl-proline
TFA: trifluoroacetic acid