MASS SPECTROMETRIC IDENTIFICATION OF FORMALDEHYDE-INDUCED MODIFICATIONS OF PEPTIDES AND PROTEINS UNDER IN VIVO PROTEIN CROSS-LINKING CONDITIONS

by

Judy Toews

B.Sc., University of British Columbia, 1999

A THESIS SUBMITTED IN PARTIAL FULFULLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Chemistry)

THE UNIVERSITY OF BRITISH COLUMBIA

FEBRUARY 2008

ABSTRACT

Formaldehyde cross-linking has been used to study protein-protein interactions in cells. Its short spacer arm, ability to permeate through cell membrane and the reversibility of the cross-linking reaction makes this a desirable cross-linker for *in vivo* studies. Although it has been widely used as a cross-linking reagent, the detailed chemistry behind protein cross-linking is not well understood. *In vitro* studies conducted under extended incubation periods (2 days) have shown that a multitude of amino acids are reactive to formaldehyde and that residue accessibility appears to play a role in reactivity. How applicable these findings are to formaldehyde cross-linking studies done under *in vivo* conditions (10-20 min incubations) is unclear. The chemistry of formaldehyde cross-linking was therefore investigated in model peptides under conditions similar to those used in *in vivo* studies. It was observed that only a subset of amino acids (amino termini and side chains of lysine and tryptophan) that were found reactive under extended incubation times was reactive in the much shorter incubation period. No cross-linking was detected between peptides, and elevating the peptide and formaldehyde concentrations resulted in only a minimal amount of cross-linked peptides. The relationship between residue accessibility and formaldehyde reactivity was assessed in model proteins that contain a more complex tertiary structure. It was shown that the extent of formaldehyde reactivity was dependent on the state of protein unfolding, i.e., solvent accessibility of reactive residues, and that an unfolded protein showed a significantly higher number of formaldehyde-induced modifications than a folded form, with lysine being the predominant reactive site. Formaldehyde treatment of proteins in

their native form resulted in a low number of modifications even under an increased incubation time, suggesting that the protein remains folded during the course of the reaction. This is important for *in vivo* cross-linking studies where specificity and stability of protein-protein interactions is dictated by protein tertiary structure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank foremost my supervisor Dr. Juergen Kast for accepting and training me as a student in his laboratory, and allowing me to work on a project that is both unique and exciting. I would also like to acknowledge Jason Rogalski for his endless patience, and Shujun Lin and the rest of the current members of the Kast Lab for their support and guidance during my studies.

On a person note, I would like to thank my family for their unwavering support throughout this entire process.

To Ryan and our baby,

CO-AUTHORSHIP STATEMENT

In the first manuscript: Mass Spectrometric Identification of Formaldehyde-Induced Peptide Modifications under *In Vivo* Protein Cross-linking Conditions, I helped my supervisor in the design of the research program, performed all the research and did all the data analysis. For the manuscript, I wrote the Materials and Methods and part (<50%) of the results and discussion sections, prepared all the tables and all but one of the figures (Figure 2-9).

In the second manuscript: Solvent Accessibility Governs the Relative Reactivity of Basic Residues in Formaldehyde-induced Modifications, I also helped in the design of the research program. I conducted all the research and data analysis except for the ESI-MS analysis of myoglobin (Figure 3-4a), for that experiment, I prepared the samples. For the second manuscript, I wrote the Materials and Methods section and help co-write (~50%) the abstract, results, discussion and conclusions sections. I prepared all the tables and figures except for Figure 3-4a.

1 INTRODUCTION

1.1 Protein Cross-linking

Proteins are vital to living organisms and are involved in many functions in a cell. Their functions can include structural maintenance, communication and transportation of materials throughout the body amongst many others. The study of these proteins is therefore understandably very important, explaining the emergence and rapid expansion of proteomics.

Proteomics is defined as the large-scale study of the structure, function and regulation of proteins in a cell under varying conditions. The unknown function of a protein of interest can be investigated by identifying proteins it interacts with. Through database searches for protein function, the known function of the interacting proteins can be classified into cellular pathways. Based on the identified pathways, the function of the protein of interest may be elucidated.

When a group of two or more associating proteins come together through protein-protein interactions they can form protein complexes which can be isolated using techniques such as immunoaffinity chromatography (Figure 1-1). Immunoaffinity chromatography is a separation technique that can isolate a protein of interest from a biological sample through the affinity between the protein and specific antibodies immobilized on beads in a column *(1-4)*. In this technique, the biological sample is introduced onto the column where binding between the protein and immobilized antibodies occur. The column is

washed several times to remove unwanted cellular material and proteins that do not bind or bind loosely or non-specifically to the antibodies or the beads. After the wash steps, the protein of interest is eluted and collected from the column for further processing/analysis.

With the immunoaffinity chromatographic technique, it is possible to bind not only the protein of interest but also proteins it may interact with (or bind to) (Figure 1-1a). This purification method can therefore be used to isolate protein complexes for protein interaction studies. Unfortunately, the stringent wash conditions typically used to remove unwanted cellular material can cause protein-protein interactions with high dissociation constants, i.e., weak or transient interactions, to be lost during this process (5-7) (Figure 1-1b). To preserve these protein interactions, protein cross-linking can be applied prior to purification (Figure 1-1c). The protein components in the complex, if all concurrently eluted, can be separated and each identified by mass spectrometry. Using this technique, the interacting partners of a protein of interest can be determined and the function of the protein clarified.

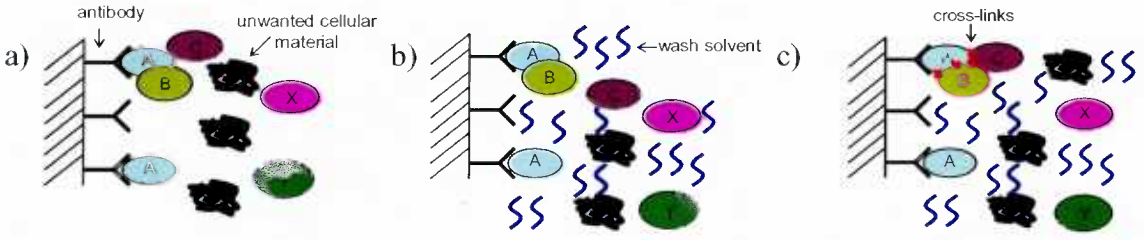


Figure 1-1. A schematic illustrating a) the binding of the protein of interest (protein A) and its protein complex (composed of proteins A, B & C) onto an immunoaffinity column. Without cross-linking b) protein interactions of high dissociation constants (between proteins A and C) are often lost; while with cross-linking c) these interactions are preserved. Note: Proteins X and Y are non-interacting proteins of protein A.

With protein cross-linking, two proteins that are in close proximity are linked together through the formation of common bonds via a small bifunctional molecule known as a chemical cross-linker. The cross-linker can be thought of as a bridge that joins two proteins. It typically has two functional groups (reactive sites), that react with functional group on each protein to form a covalent bond (cross-link).[1] The number of cross-links between two proteins is dependent on the number of reactive functional groups on the interacting proteins' surfaces. Cross-links should be present between neighboring proteins making up the complex (Figure 1-1c). The presence of these multiple cross-links within the protein complex holds the complex together during the stringent washes used in the immunoaffinity chromatography process.

Proteins are simply long peptides (polypeptides) that fold into a 3-dimensional structure (tertiary structure) based on their amino acid sequence and bonding interactions between side chains of amino acids. Cross-linking between proteins (inter-protein cross-linking) and within a protein (intra-protein cross-linking) can produce inter-peptide bonds upon proteolysis (Figure 1-2a). Reactive functional groups within a single peptide stretch can also form cross-links, and those bonds are referred to as intra-peptide bonds since they occur with the same peptide (Figure 1-2b). The chemical cross-linker may also react with only one reactive functional group and not form cross-links at all, this type of modification has been referred to in the literature as dead-end, end capped, hanging or decorated cross-links (Figure 1-2c) *(8)*.

[1] The functional groups originate from the amino acids that make up the protein.

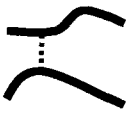
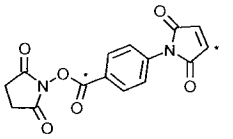
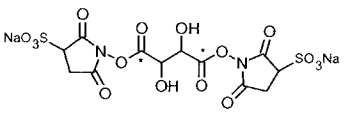
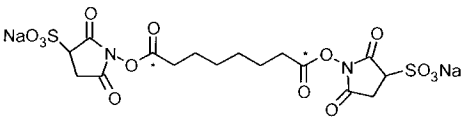
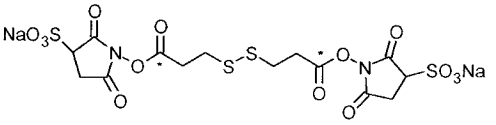
a)  b)  c) 

Figure 1-2. The types of bonds that can be formed in the reactions between proteins and a chemical cross-linker: a) inter-peptide, b) intra-peptide and c) dead-end cross-links. Peptides and cross-linking bond are indicated by the solid and dotted lines, respectively.

The length of a cross-linking bond (spacer arm) depends on the distance between the two reactive ends of the chemical cross-linker. Table 1-1 shows a few examples of chemical cross-linkers with varying spacer arm lengths. The two reactive sites on each are highlighted with asterisks; they can be the same (homobifunctional) or different (heterobifunctional).

The reactive sites in the chemical cross-linkers shown in Table 1-1 are all electrophiles (electron poor). The reaction between functional groups on proteins and these cross-linkers involve an electron-rich functional group (nucleophile) on proteins, such as the ε-amino group on lysine side chains, reacting with these electrophiles. Figure 1-3 shows the reaction between two hypothetical proteins and the chemical cross-linker BS^3.

Table 1-1. Examples of chemical cross-linkers used in protein-protein interactions studies. Reactive sites are indicated with asterisks.

Name	Structure	Spacer Arm Length (Å)
Formaldehyde		2 [5, 9, 10]
m-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS)		9.9 [11]
Disulfosuccinimidyl tartarate (sulfo-DST)		6.4 [8]
bis(sulfosuccinimidyl) suberate (BS^3)		11.4 [8]
3, 3'-dithiobis[sulfosuccinimidylproprionate] (DTSSP)		12 [9]

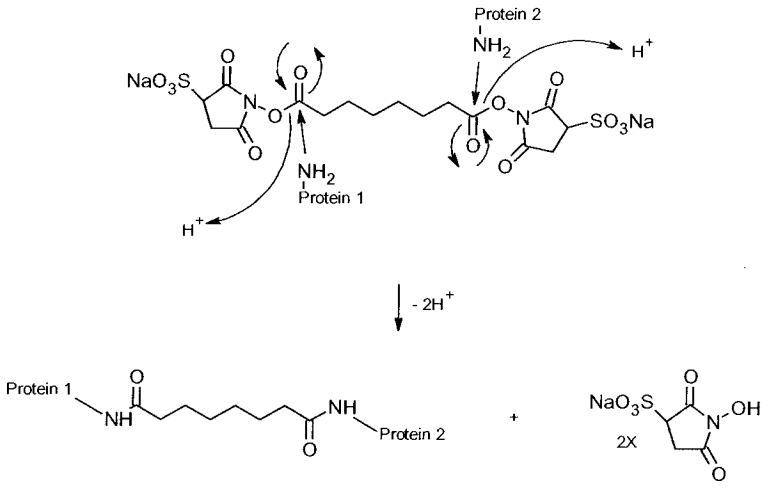


Figure 1-3. The reaction between the chemical cross-linker bis(sulfosuccinimidyl) suberate and two proteins.

1.2 Formaldehyde as a Cross-linker

Formaldehyde is a small molecule (30 Da) with a short spacer arm (2 Å). A short spacer arm is desirable in protein-protein interaction studies in order to minimize the chances of cross-linking non-interacting proteins that do not typically associate in such close proximity. With this advantage as well as the others listed below, formaldehyde shows much promise as a selective tool for the stabilization of protein complexes.

Advantages of using formaldehyde to study protein-protein interactions are:

1) able to permeate through cell membrane and bacterial cell wall,

2) reacts with multiple amino acids,

3) the reaction is reversible, and

4) formaldehyde is water-soluble and inexpensive.

Table 1 indicates that the two reactive sites on formaldehyde are located on the same atom. The explanation for this can be found by looking at the mechanism of formaldehyde cross-linking (Figure 1-4). In the first step of the reaction, a nucleophile on the protein attacks the carbonyl carbon of formaldehyde to form the methylol intermediate (R-X-CH_2-OH, where X = nucleophile) with a mass shift of +30 Da which may then lose water, to produce the imine in the case of primary amines, with a mass increase of +12 Da. The imine formed between a carbonyl compound and a primary amine (free amino terminus or ε amino group of lysine) is called a Schiff base. In the second step, a nucleophile on another protein attacks the carbon of the imine bond to produce the cross-linked product. The carbon in the cross-linking bond that remains from

formaldehyde forms a methylene bridge between proteins. The methylene bridge produces an overall mass increase of +12 Da to the added mass of the two cross-linked proteins. Step 2 of the reaction has been shown to be partially reversible by heating at high temperature (95°C) *(5, 10, 12)*.

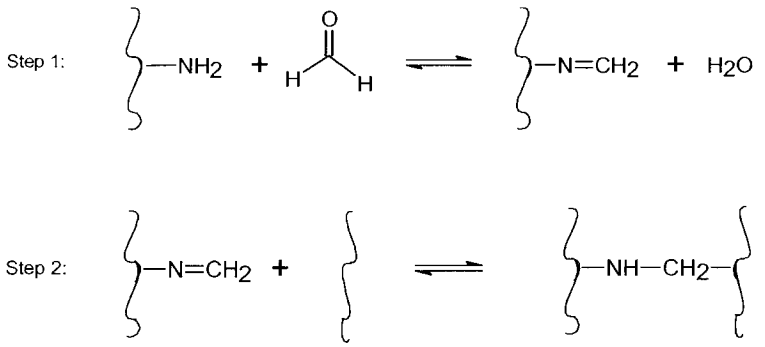


Figure 1-4. Two-step reaction of formaldehyde cross-linking. Step one involves the formation of the imine (mass shift of +12 Da) and water while step 2 involves the reaction of the imine with a nucleophile on another protein to form the cross-linking bond. Proteins are represented by wavy lines.

1.3 Mass Spectrometry

Mass spectrometry (MS) has emerged as a powerful and valuable tool in the field of proteomics and has been used in protein identification for *in vivo* protein-protein interaction studies using formaldehyde as the cross-linking reagent *(5-7, 10, 12)*. This is all possible due to the invention of two "soft ionization" techniques in the late 1980s: matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) *(13, 14)*. Previously impossible, thermally labile and involatile biomolecules such as peptides and proteins could now be vaporized and ionized, and their mass/charge (m/z) detected with high accuracy using various mass analyzers. The ability to detect low

analyte concentrations (sub-micromolar) at minimal sample consumption (microliter amounts) makes this technique compatible to proteomic studies where only small amounts of biomolecules at low concentrations can be isolated, i.e., proteins obtained from a cell lysate or an *in vivo* animal model.

1.3.1 MALDI

For MALDI MS, samples are prepared by mixing low volumes (0.1-1 µL each) of solution of a matrix (a small UV-light absorbing acid) and analyte solution. An aliquot of the mixture is deposited onto a metallic plate and the droplet is dried. Upon being dried, analyte molecules become incorporated into the crystal lattice of the matrix; with this technique low concentrations of salts and buffers can be tolerated *(8, 15)*. A laser pulse is focused onto the surface of a dried sample spot and the energy from the laser beam is absorbed by the abundant matrix molecules which causes the sample surface to undergo localized vaporization leading to cluster (of matrix and analyte-matrix) emission (Figure 1-5). Ionization of the biomolecules is achieved by charge transfer with the matrix. During ionization single proton transfer to the analyte typically occurs (except for proteins, where a few protons can be attached) and the analytes are detected at a charge state of 1. MALDI can be found coupled to a wide variety of mass analyzers (time-of-flight (TOF), Fourier transform (FT), magnetic sector, ion trap) and their hybrid/tandem forms. However, the most common are MALDI-TOF instruments. The short pulsed ion supply provided by the laser source of this ionization technique is ideally suited for TOF instruments since there is a defined time for ion generation.

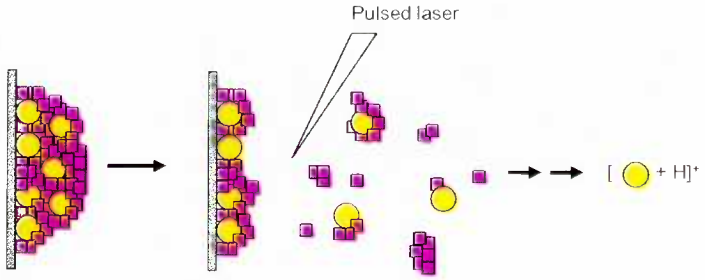


Figure 1-5. Schematic illustrating the MALDI process. The spheres and squares represent analyte and matrix, respectively. Energy from the laser beam is absorbed by the matrix which causes the sample surface to undergo localized vaporization. Analyte ions are produced from charge transfer with the matrix.

1.3.2 ESI

For ESI MS samples, analyte ions are pre-formed in solution from solvent protonation of basic functional groups on the analyte. The sample solution flows through a narrow capillary where a voltage is applied immediately before or at the capillary tip to generate a high electric field. If the capillary is biased positively, positively charged droplets leave the capillary and are accelerated towards the negative electrode located at the interface of the mass spectrometer while shrinking in size due to evaporation of solvent and Rayleigh fission (Figure 1-6). For efficient analysis, solvents are volatile and devoid of non-volatile salts for easy evaporation. Since ESI MS samples are solution based, front-end separation techniques such as liquid chromatography (LC) can be easily used to separate components of a complex mixture on line (such as peptides from a protein digestion) prior to MS analysis. Similar to MALDI, ionization for this technique occurs through proton attachment, but it is much more extensive and heterogeneous for proteins, leading to mass spectra with multiple signals representing different charge states of the protein. The different charge states of a protein are collectively referred to as the protein charge

state envelope. Deconvolution software is used to convert the data in the protein charge state envelope to protein mass. ESI allows protein samples to be analyzed under neutral pH, allowing study of biochemical processes such as protein folding. This ionization technique is used in combination with a wide array of mass analyzers (quadrupoles (Q), ion traps, TOF and FT instruments) and their hybrid/tandem forms, such as the ESI-Q-TOF instrument.

Two benefits of having multiply charged species produced in ESI present themselves when studying intact proteins. The presence of multiple charge states for proteins in ESI makes it compatible with quadrupole mass spectrometers that have a limited m/z range. The other benefit is in the study of the folding state of a protein. The distribution of charge states in the spectrum is proportional to the number of accessible basic residues (Lys, Arg, His) on the solvent exposed surface of the protein, which is therefore related to the extent of protein folding. An unfolded protein will have more solvent exposed residues than its folded counterpart, therefore more residues available for protonation resulting in a more highly charged ion. Thus the m/z shift of the protein charge state envelope can be used to study major changes in protein folding.

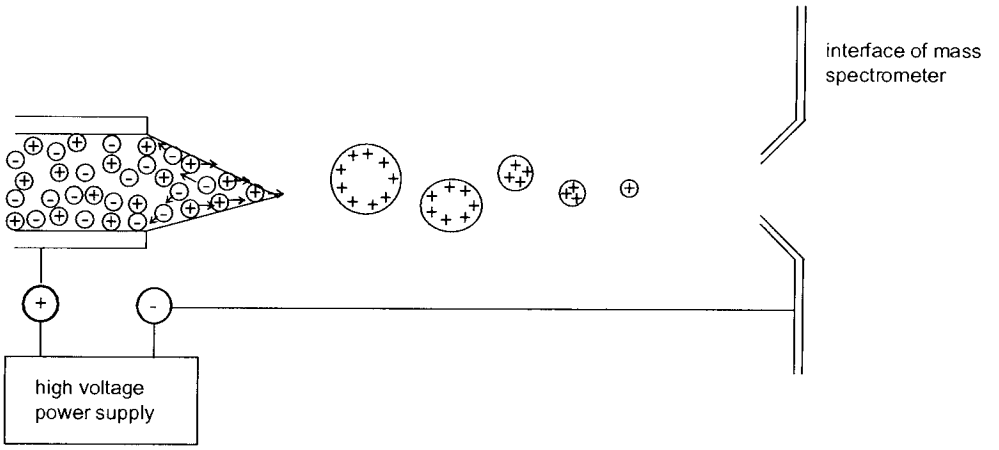


Figure 1-6. Schematic illustrating the ESI process. Charged droplets leave the capillary tip and shrink in size through evaporation and Rayleigh fission before entering the interface of the mass spectrometer.

1.3.3 MS and MS/MS

In mass spectrometric analyses there are two main analysis types: MS and MS/MS. In MS mode, the m/z of the biomolecule (peptide/protein) is measured, which can then be converted to the mass when the charge state is known. To determine the identity of a biomolecule, mass may be used, but the confidence level in the assignment would not be high, since different sequence combinations, thus different biomolecules can produce the same mass. To obtain a higher confidence in the identification of a biomolecule, primary structure information (amino acid sequence information) can be collected in MS/MS mode and be used to confirm the identity of the biomolecule. Peptides can be analyzed as is, however commonly proteins need to be enzymatically digested to yield peptides for analysis (bottom-up sequencing). Proteins can also be sequenced directly (top-down sequencing) but the need for mass spectrometers with high resolution, accuracy and ability to fragment large ions have limited the use of this approach.

MS/MS analysis is performed on hybrid instruments by 1) selecting for the ion of interest (precursor ion) by selecting for its m/z in the first mass analyzer, 2) introducing the peptides into a collision cell for fragmentation and 3) MS analysis of the fragment ions for m/z information in the last mass analyzer (Figure 1-7). For protein identification, the m/z of fragment ions for each peptide obtained from enzymatic digestion are compared to predicted fragment ions of sequences stored in protein databases to determine the identity of the protein. For peptides, aside from sequence confirmation, this mode can identify and localize chemical and post-translational sites of modification.

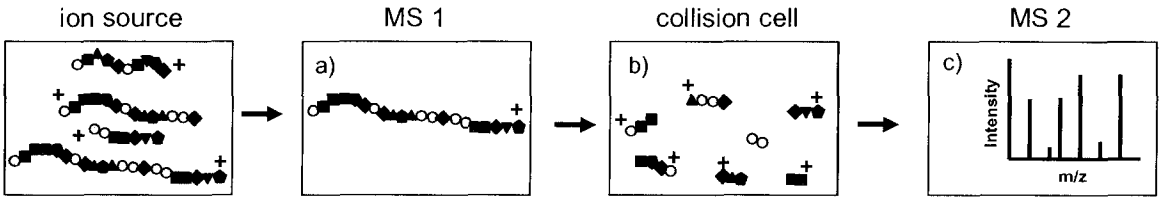


Figure 1-7. A schematic illustrating the process of peptide fragmentation in MS/MS mode. a) The precursor ion is selected from a mixture of ions and b) is sent to the collision cell for fragmentation. c) The resulting fragment ions are sent to the last mass analyzer for m/z determination.

1.3.4 Collision-Induced Dissociation

In the fragmentation process, the precursor ion is collided with an inert gas. The collision converts the kinetic energy of the precursor ion to internal energy, which if great enough induces fragmentation. For peptides, fragmentation occurs for the most part along the peptide backbone, which contains three types of bonds: C_α-C, C-N and N-C_α (Figure 1-8). When a bond breaks, two complementary pieces are produced, one containing the N-terminus and the other containing the C-terminus of the original peptide. The

complementary ions produced by the three types of bond breakages are: a and x, C_α-C bond; b and y, C-N bond; and c and z, N-C_α bond (Figure 1-8). The most common type of ions produced when fragmenting tryptic peptides with collision induced dissociation are b and y ions.

Two bonds can also break along the peptide backbone to produce internal fragment ions that are usually produced by a combination of y-type and b-type fragmentations *(16)*. In the low mass region of the mass spectrum, fragment ions representing individual amino acid residues may also be observed and are in the form of their respective immonium ion (NH=CH-R). Chemical or post-translational modifications on peptides would produce fragment ions with mass increases corresponding to the modification(s). By identifying fragment ions that have modification-related mass shifts, the site of modification on the peptide sequence can be easily determined.

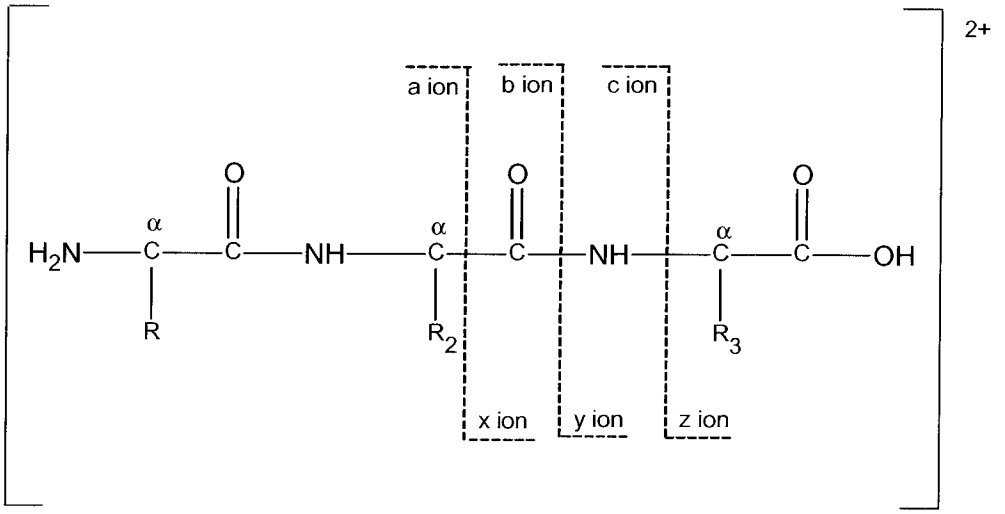


Figure 1-8. A simple schematic illustrating the complementary ions produced from fragmentation along the peptide backbone.

1.3.5 TOF Mass Analyzers

TOF mass analyzers separate ions of different mass by applying an electric potential to accelerate the ions out of the ion source and into the field-free drift region to the same kinetic energy (few keV). The ions then drift through a field-free path (1-2 m) where heavier ions will travel at a lower velocity than lighter ions, causing them to take longer to travel down the flight tube. This difference in "time of flight" separates ions of different mass in space prior to arriving at the detector for detection (Figure 1-9).

TOF mass analyzers can operate in two modes: linear and reflectron. Modern linear mode instruments use delayed ion extraction and space focusing to obtain a more uniform flight time for ions of the same mass by equating their initial kinetic energy and starting positions relative to one another *(17, 18)*. A more unified flight time translates to sharper peaks and hence better mass resolution. In MALDI TOF analyzers, desorption of analyte and matrix from the sample plate produces a plume of analyte ions and neutral molecules. When the analyte ions are accelerated, they collide with the neutral molecules and a kinetic energy spread is created which decreases mass resolution. To counteract this problem, there is a time delay before the ions are accelerated to allow neutral molecules to expand to a low density in the source. This method is called delayed ion extraction *(17, 18)*. Space focusing, however, corrects for ions of the same mass starting at different positions in the source. The focusing is accomplished by applying a small electric potential to the ions. Ions situated further away from the exit of the ion source experience a larger potential difference than ions closer to the exit and move more

quickly. The two electric potentials (applied for space focusing and to accelerate ions out of the ion source) are set up so that the position where the ions meet is at the detector.

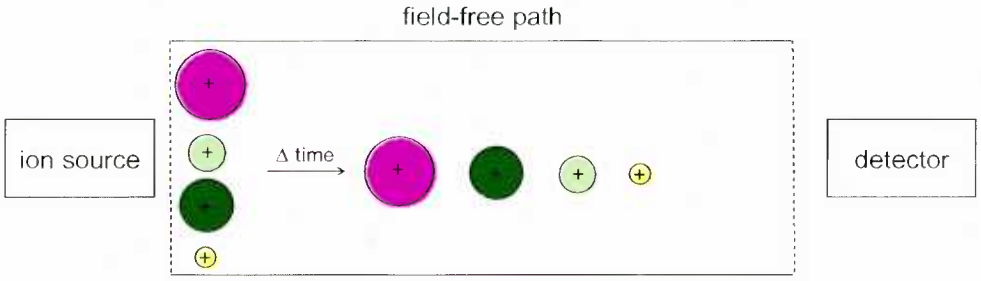


Figure 1-9. Schematic illustrating the separation of ions of varying mass in a time of flight mass analyzers. Ions are accelerated out of the ion source with the same kinetic energy and drift in a field-free path where they are separated in space due to their varying velocities (kinetic energy = ½ mass*velocity^2).

Reflectron mode uses the above two techniques as well as energy focusing to increase mass resolution. Energy focusing corrects for ions of the same mass, starting from a similar position, traveling down the flight path with different kinetic energy resulting in different flight times. The focusing is accomplished by ion mirrors (reflectron field) located at the end of the flight tube. Ions with slightly higher kinetic energy will travel deeper into the reflectron field than ions with slightly lower kinetic energy before they are both reflected upstream to the detector. This results in slightly different path lengths and compensates for the difference in kinetic energy such that the ions arrive at the detector at a more unified time. There is a mass limit for the reflectron field due to limitations in the applied voltage at the reflectron so this focusing technique can only be applied to lower molecular weight biomolecules such as peptides and small proteins. For larger proteins (>8-10 kDa), linear mode is used.

1.4 In Vitro Studies Using Formaldehyde and Mass Spectrometry

The ease of mass spectrometry to accurately detect small mass changes and its ability to sequence biomolecules and therefore identify and isolate sites of chemical modifications makes this analysis technique well suited to monitor and investigate products of the reaction between peptides with formaldehyde. In a study by Metz *et al. (19)*, model peptides were incubated for 48 h with formaldehyde in order to determine amino acid residues that were reactive with the cross-linker in the first step of the reaction. Products were separated by reversed-phase LC and analyzed by ESI-MS (quadrupole ion trap). Aliphatic and acidic residues were found unreactive with formaldehyde while electron-rich residues such as cysteine, arginine, tryptophan, histidine, lysine and the amino termini were.

Mass spectrometric analysis of the products revealed dead-end modifications on peptides with a characteristic mass increase of +30 or 12 Da, which corresponded to decoration with the methylol intermediate and its dehydrated form, respectively (see Section 1-2). Intra-peptide cross-linking of peptides (via methylene bridge(s), see Figure 1-4) were detected with mass increase of +12 or 24 Da. Formaldehyde cross-linking between peptides was not reported.

In a follow-up study, the same research group *(20)* wanted to determine if location of reactive residues in a structurally more complex system will affect their reactivity toward formaldehyde, or formaldehyde and glycine. They chose the structurally simple protein,

insulin[2] as their model and incubated with formaldehyde or formaldehyde and glycine for 1 week – an even longer incubation time than what was used in their peptide study *(19)*. Protein was digested using endoproteinase Glu-C and the resulting peptides were analyzed by ESI-LC-MS (Q-TOF).

In the reactions between insulin and formaldehyde, MS/MS analysis revealed that the amino termini and selective Cys and Tyr residues were modified with a mass increase of 12 Da. The authors do not explain why other reactive residues, identified in their peptide studies *(19)* that are present in insulin, as well as Tyr residues at other positions, were unreactive. Only in the reaction between insulin with formaldehyde and glycine was the reactivity of selective Tyr residues explained by solvent accessibility which was position dependent. The main product from the reaction between the protein with formaldehyde and glycine was the cross-linking of the amino acid (via formaldehyde) to selective residues on the protein. In either reactions, formaldehyde cross-linking between the two insulin chains or proteins were not reported. Together, results from these studies confirm the general consensus that formaldehyde reacts with multiple amino acids. However, the applicability of these findings to *in vivo* cross-linking studies has not been addressed since the difference in incubation time between the two (48 h and 1 week versus $\leq$ 1h) is so vast. *In vivo* cross-linking studies done under these extended incubation conditions would result in extensive cross-linking which would result in significant protein loss due to insolubility and precipitation.

[2] Insulin has a molecular weight of 5700 Da and contains two chains that are connected via two disulphide bridges.

Solvent accessibility and residue reactivity towards formaldehyde was used by Peterson *et al. (9)* to study the 3-dimensional structure of a protein by observing sites of dead-end modifications, and intra- and inter-peptide cross-linking. For comparison and confirmational purposes the protein was also treated separately with the cross-linker DTSSP (Table 1-1), which reacts exclusively with primary amines. Their protein of interest was α-crystallin[3], found at high levels in optical lens. The protein was treated with formaldehyde or DTSSP for 1 h, an incubation time more similar to those used in *in vivo* cross-linking studies *(5, 6, 12, 21)*. Tryptic peptides were analyzed using a quadrupole ion trap (QIT) mass spectrometer equipped with ESI source.

Consistent with the earlier work of Metz *et al. (19)*, dead-end modifications were detected with peptide mass increases of +12 Da and +30 Da and cross-links were formed via the methylene bridge (+12 Da as well). From the extent of formaldehyde and DTSSP dead-end modification and/or cross-linking, a list of criteria was created to help with the structural assignments for α-crystallin:

1) Lack of both dead-end modifications and cross-linking to certain regions of the protein suggested that they were buried deep within the protein's 3-dimensional structure.

2) Dead-end modifications or cross-linking by only the smaller cross-linking reagent, formaldehyde, indicated intermediate solvent accessibility.

3) Dead-end modifications but no cross-linking of a region suggested that the area was highly exposed and solvent accessible.

[3] α-crystallin has a molecular weight of 700-800 kDa and is composed of 3:1 mixture of αA- and αB-crystallin subunits.

4) Cross-linking between subunits indicated interaction between subunits. If the cross-linking was accomplished by formaldehyde it suggested that the subunits were no greater than 2 Å apart.

A limitation to this study was the use of software which only considered lysine residues capable of being modified or cross-linked. However, since currently there is no study that have identified residues that are reactive to formaldehyde under short incubation conditions (≤ 1 h), this may be a valid assumption to make.

1.5 In Vivo Cross-linking of Proteins

Studying the chemistry of formaldehyde cross-linking to as near *in vivo* cross-linking conditions as possible can only approximate the conditions of formaldehyde cross-linking between proteins in cells. Due to the high protein concentration in cells, where greater than 50% of cellular material is protein based *(22)*, and sub-cellular distribution and concentrations of these proteins, as well as the presence of salts and cofactors that may influence the reaction, an *in vitro* system may come close to but never truly mimic the cellular environment. Therefore the absence of cross-linking between peptides and proteins as observed in some *in vitro* experiments *(19, 20)* is not indicative of the ability of formaldehyde to cross-link proteins *in vivo*. This is demonstrated below, in the discussion of several *in vivo* formaldehyde protein cross-linking studies that have successfully identified interacting protein partners of a protein of interest.

For studying protein-protein interactions *in vivo*, the "nearest neighbor" approach is taken which assumes protein that are in close proximity interact. The ability of formaldehyde to cross-link proteins that only associates in close proximity (2 Å) and the ease at which it can permeate cellular membrane without external manipulation has made this cross-linking reagent desirable in several *in vivo* protein cross-linking studies *(5, 6, 12, 21)*. The cellular system may vary considerably from yeast cells *(6, 12)* to mammalian cells *(5)* to brain cells of live mice *(21)*, as well as vary in the location and function of the protein of interest. However, as discussed below similar experimental approaches were used in each study.

In vivo protein cross-linking conditions use low concentration of formaldehyde (≤ 4%) and a short incubation time (≤ 1 h). This is to prevent significant protein loss from excessive cross-linking which lead to protein insolubility and precipitation. The cross-linked protein complexes of interest were typically isolated using an immunoaffinity column that has a high affinity for the protein of interest compared to other cellular proteins *(5, 6, 12)*. The cross-links in the complexes may be reversed by heating (95°C, ≤20 min) *(5, 12)* or left intact *(6, 21)*. Separation of the proteins using SDS-PAGE[4] may be employed prior to enzymatic digestion with trypsin to reduce the complexity of the sample *(5, 12)*. Tryptic peptides were either analyzed directly by mass spectrometry (12) or separated by reverse phase LC prior to mass spectrometric analysis *(5, 6, 21)*. To identify the proteins in the tryptic digest, the mass of the tryptic peptides of each protein (MS analysis) *(12)* or mass and its fragments (MS and MS/MS analysis) *(5, 6, 21)* were

[4] SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis, a technique that separates proteins according to their relative size.

compared to sequences in databases for a protein match. Higher quality protein identifications are obtained using the latter of the two methods, since not only peptide mass but also its fragments are used to confirm the identity of a protein.

In the above mentioned *in vivo* formaldehyde cross-linking studies *(5, 6, 12, 21)*, unfortunately only the identity of the interacting protein partners of the protein of interest was sought. The composition and number of isolated protein complexes or the geometry of such complexes were not studied. The geometry of protein complexes provides detailed information regarding which proteins interact directly and come into contract with the protein of interest, or indirectly and are separated from the protein of interest by one or more proteins. To study the geometry of protein complexes, protein surfaces (inter-cross-linked peptides) involved in cross-linking of the protein complex need to be preserved, i.e. the cross-links can not be reversed by heating so they can be identified and characterized. Of the studies mentioned above *(5, 6, 12, 21)*, two did not have their cross-links reversed *(6, 21)* but the tryptic digestion mixture, that may have been rich in modified peptides (dead-end modifications, intra- and inter- cross-linked peptides) were ignored in their mass spectrometric analysis since these structures and their MS patterns are largely undefined.

1.6 Thesis Theme and Hypothesis

Formaldehyde cross-linking, in conjunction with mass spectrometric analysis, can be used to study the detailed 3-dimensional structure of a protein, and to identify interacting partners of a protein of interest. It would also be interesting to study the protein

interactions in detail by identifying the interacting surfaces and locating the reactive residues involved in the cross-linking. This would require that the protein complexes isolated from purification not have their cross-links reversed and instead be directly subjected to proteolytic digestion. The peptide mixture produced from protein digestion would be composed of peptides from the complex that are unmodified, dead-end modified and cross-linked.

Before these modified and cross-linked peptides can be studied, there needs to be a better understanding of the chemistry of formaldehyde cross-linking under *in vivo* conditions (short incubation times with formaldehyde, < 1 h). So far, only one other research group has studied the chemistry of formaldehyde reactions with model peptides and proteins in detail by mass spectrometry *(19, 20)*, although under conditions not readily amenable to *in vivo* protein cross-linking. Therefore, one of the aims of the work described in this thesis was to gain a more detailed understanding of the chemistry of formaldehyde cross-linking in model peptides using experimental conditions similar to those used *in vivo* (i.e., short incubation times). It is hypothesized that a select number of amino acid residues will be reactive to formaldehyde and based on a previous *in vivo* study *(5)*, the reaction yield will be dependent on formaldehyde concentration and incubation time. Therefore the products of the formaldehyde reaction under varying conditions were analyzed by mass spectrometry so that characteristic mass shifts of the products could be defined and characteristic fragments could be identified that will aid in the development of a detection method that can highlight these products in a complex mixture.

To be more aligned to a cellular context, the second aim of my thesis was to study the chemistry of formaldehyde cross-linking on a structurally more complex system, i.e. on model proteins. Mass spectrometry was used to monitor the products of the reaction before and after proteolysis.

1.7 References

1. Hage, D. S. (1999) *Clin Chem* **45,** 593-615.
2. Weller, M. G. (2000) *Fresenius J Anal Chem* **366,** 635-45.
3. Van Emon, J. M. (2001) *J AOAC Int* **84,** 125-33.
4. Subramanian, A. (2002) *Mol Biotechnol* **20,** 41-7.
5. Vasilescu, J., Guo, X., and Kast, J. (2004) *Proteomics* **4,** 3845-54.
6. Guerrero, C., Tagwerker, C., Kaiser, P., and Huang, L. (2006) *Mol Cell Proteomics* **5,** 366-78.
7. Tagwerker, C., Flick, K., Cui, M., Guerrero, C., Dou, Y., Auer, B., Baldi, P., Huang, L., and Kaiser, P. (2006) *Mol Cell Proteomics* **5,** 737-48.
8. Sinz, A. (2003) *J Mass Spectrom* **38,** 1225-37.
9. Peterson, J. J., Young, M. M., and Takemoto, L. J. (2004) *Mol Vis* **10,** 857-66.
10. Hájek, P., Chomyn, A., and Attardi, G. (2007) *J Biol Chem* **282,** 5670-81.
11. Di Bernardo, S., and Yagi, T. (2001) *FEBS Lett* **508,** 385-8.
12. Layh-Schmitt, G., Podtelejnikov, A., and Mann, M. (2000) *Microbiology* **146,** 741-7.
13. Karas, M., and Hillenkamp, F. (1988) *Anal Chem* **60,** 2299-301.
14. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) *Science* **246,** 64-71.
15. Schermann, S. M., Simmons, D. A., and Konermann, L. (2005) *Expert Rev Proteomics* **2,** 475-85.
16. Papayannopoulos, I. A. (1995) *Mass Spectrom Rev* **14,** 49-73.

17. Brown, R. S., and Lennon, J. J. (1995) *Anal Chem* **67,** 1998-2003.

18. Whittal, R. M., and Li, L. (1995) *Anal Chem* **67,** 1950-4.

19. Metz, B., Kersten, G. F., Hoogerhout, P., Brugghe, H. F., Timmermans, H. A., de Jong, A., Meiring, H., ten Hove, J., Hennink, W. E., Crommelin, D. J., and Jiskoot, W. (2004) *J Biol Chem* **279,** 6235-43.

20. Metz, B., Kersten, G. F., Baart, G. J., de Jong, A., Meiring, H., ten Hove, J., van Steenbergen, M. J., Hennink, W. E., Crommelin, D. J., and Jiskoot, W. (2006) *Bioconjug Chem* **17,** 815-22.

21. Schmitt-Ulms, G., Hansen, K., Liu, J., Cowdrey, C., Yang, J., DeArmond, S. J., Cohen, F. E., Prusiner, S. B., and Baldwin, M. A. (2004) *Nat Biotechnol* **22,** 724-31.

22. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1994) Molecular Biology of the Cell, Garland Publishing Inc., New York.

2 MASS SPECTROMETRIC IDENTIFICATION OF FORMALDEHYDE-INDUCED PEPTIDE MODIFICATIONS UNDER IN VIVO PROTEIN CROSS-LINKING CONDITIONS[5]

2.1 Introduction

The study of protein-protein interactions in living cells is an important aspect of understanding the function of individual proteins and their role within the cellular context. Affinity-based enrichment of proteins and their interaction partners from cell or tissue lysate is a commonly applied technique to isolate protein complexes and to identify and characterize its components by subsequent mass spectrometric analysis. Moreover, it is at the heart of several large-scale approaches to map the interactome of model organisms such as yeast *(1-3)* and E. coli *(4)*, or signaling pathways in human cell lines *(5, 6)*.

A major limitation of this general strategy is the loss of cellular context prior to the affinity enrichment step. As cell lysis converts the highly ordered protein assemblies inside a cell into disordered and dilute protein solutions, it eliminates the spatial and temporal constraints that govern protein interactions in cells. This process thus exposes the proteins under investigation to abundant cellular proteins they would not necessarily be in contact with otherwise. Moreover, it disturbs the equilibria of individual protein interactions and introduces molecular diffusion as an additional factor. As a consequence, interactions that show higher dissociation constants are lost, i.e. those that

[5] A version of this chapter will be submitted for publication. Toews, J. and Kast, J. Mass Spectrometric Identification of Formaldehyde-Induced Peptide Modifications Under *In Vivo* Protein Cross-linking Conditions.

are of transient and weak nature and result in high turnover and/or low stoichiometry. Together, these two processes lead to an increase in false positive and a decrease in true positive interactions over time. Several ways of combating this phenomenon have been introduced, which include the use of tandem affinity tags *(7)* to eliminate non-specific binding and cryolysis *(8)* to slow down the protein association and dissociation kinetics.

Formaldehyde-mediated protein cross-linking has recently emerged as an additional means to preserve cellular protein interactions during affinity enrichment. It has been shown to be compatible with: affinity enrichment based on single *(9)* or tandem affinity tags *(10, 11)* as well as the endogenous protein *(12)*; incorporation of stable isotope labels *(10, 11)*; the use of non-denaturing *(9, 12)* as well as denaturing *(10, 11)* purification conditions; and the application in cultured cells *(9-11)* as well as animal tissue *(12)*. In all of the aforementioned cases, interacting proteins were identified by tandem mass spectrometry, whereas the actual cross-linking sites were not determined. This is in stark contrast to *in vitro* cross-linking strategies that utilize homo- or hetero-bifunctional cross-linking reagents with characteristic linker lengths for the analysis of protein interaction geometries *(13-15)*. These are based on well-known reactions that generate chemical structures with predictable changes in peptide masses, which facilitate their analysis and identification by mass spectrometry. Conversely, the chemistry of the formaldehyde cross-linking is presumed to be considerably more complex, and the resulting mixture of products to be much more heterogeneous and difficult to assess by mass spectrometry. Two recent studies involving model peptides *(16)* and proteins *(17)* have shed some light onto the reaction products after formaldehyde treatment. Extended length of

formaldehyde exposure over several days has been shown to result in the modification of a multitude of different amino acid residues (amino-termini, Lys, Arg, His, Cys, Tyr, Trp, Phe), seemingly confirming the complexity of this reaction. Whether the results of these studies are applicable to conditions typically used for the cellular protein interaction studies outlined above and in particular with respect to their much shorter reaction time of 10-20 minutes, has not been addressed.

Here we present a detailed analysis of the reaction of formaldehyde with model peptides, which demonstrates that only a subset of the previously reported amino acid residues show significant reactivity within the first 10 minutes. Moreover, consistent with previous studies we find that the relative position of the residue in the peptide sequence is an additional constraint of its reactivity. Peptide cross-linking was negligible under the conditions in this study, and previously reported peptide mass increases of 30 Da were rarely observed. Conversely, increases by multiples of 12 Da were associated with the major reaction products and mostly located on the peptides' amino-termini as well as lysine and tryptophan residues.

2.2 Experimental

2.2.1 Chemicals

Model peptides were either purchased from Sigma (St. Louis, MO) or synthesized in-house. α-cyano-4-hydroxycinnamic acid (CHCA) was also obtained from Sigma. Paraformaldehyde (PFA), formic acid (FA, 88%) and acetonitrile (ACN, HPLC grade) were purchased from Fisher (Fair Lawn, NJ). C18 extraction tips were either obtained

from Millipore (Bedford, MA) or Varian (Lake Forest, CA). 0.22 μm filters were purchased from Pall Corporation (Ann Arbor, MI). Deionized water (18 MΩ·cm) was prepared using a Nanopure Ultrapure Water System from Barnstead (Dubuque, IA).

2.2.2 Preparation of Paraformaldehyde Solution

A 1.3 M formaldehyde stock solution was prepared by heating (80°C) PFA in phosphobuffer saline (PBS, pH 7.5) for 30 min, cooling to room temperature and filtering through a 0.22 μm filter.

2.2.3 Reaction with Peptides

14 model peptides (100 μM) were incubated in PBS (37°C) in the presence of 83 mM formaldehyde for 10 min and 2 days. Control samples were prepared by replacing the formaldehyde volume with PBS. Reactions were quenched by decreasing the pH through the addition of 5% formic acid in water. Peptides were isolated using C18 extraction tips and eluted in 50:45:5 ACN:H_2O:FA.

2.2.4 Cross-linking of Model Peptides

A mixture of the peptide PLSRTLSVAAKK and one of the peptides pGlu-GLPPRPKIPP, pGlu-ADPNKFYGLM-NH_2, HDMNKVLDL, GAPVPYPDPLEPR, YGGFMRRVGRPE, and pGlu-HWSYGLRPG-NH_2 (750 μM each) was incubated in PBS (37°C) in the presence of 333 mM formaldehyde for 20 min. Reaction quenching and peptide isolation are as described in the previous section.

2.2.5 Varying Reaction Conditions

Incubation time was studied by incubating peptide PLSRTLSVAAKK (75 μM) in PBS (37°C) in the presence of 83 mM formaldehyde for up to 48 hours. Time 0 was prepared by the addition of 5% formic acid to the sample prior to introducing the formaldehyde. Formaldehyde concentration effects were studied in model peptides PLSRTLSVAAKK and pGlu-GLPPRPKIPP (75 μM) by incubating each in PBS (37°C) in the presence of 0, 42, 83, 167 or 333 mM formaldehyde for 10 min. Peptide concentration effects were studied by incubating peptide PLSRTLSVAAKK at 15, 75, 150 or 1500 μM in PBS (37°C) in the presence of 83 mM formaldehyde for 5 or 10 min. Reaction quenching and peptide clean-up are as described in the section "Reaction with Peptides".

2.2.6 Mass Spectrometric Analysis

Eluent from the C18 extraction was diluted 1:1 (v/v) with a saturated solution of CHCA (in 50:50 ACN:H_2O) and spotted onto a MALDI target using the dried-droplet method, air dried and analyzed by MALDI-TOF-TOF (4700 Proteomics Analyzer) controlled by 4700 Explorer version 2.0 software (Applied Biosystems, Foster City, CA). MS and MS^2 spectra were collected with a total of 3000 (75 shots/subspectrum, 40 subspectra) and 10 000 shots (100 shots/subspectrum, 100 subspectra), respectively. Tandem mass spectra were acquired with atmospheric gas as the collision gas at a pressure of 3-4 x 10^{-8} Torr and collision energy of 1000 eV. MS^3 analysis was performed on the eluent from for the singly modified peptide PGHDPPISYYETN-NH_2 using nanospray-ESI-Q-TRAP (2000 Qtrap) controlled by Analyst version 1.4.2 software (Applied Biosystems, Foster City, CA). Tandem mass spectra were acquired with nitrogen as the collision gas, collision

energy of 55 eV and excitation energy of 100 eV. Nanospray emitters were purchased from Proxeon Biosystems (Odense, Denmark).

2.2.7 Interpretation of Data

To compare the reaction product distribution between different peptides or reaction times the peak area of each unmodified and modified monomeric peptide species was divided by the total peak area of all monomeric species of the peptide. Note: a +30 Da mass shift was included as a modification since it was assumed to be an intermediate of the +12 Da product. To compare the overall reactivity of each peptide, the average number of formaldehyde-induced modifications for a given peptide, the "degree of modification" (D.O.M.) was calculated. The calculation took the peak area of a modified species, multiplied it by the number of modifications that it contains (i.e., singly, multiply by 1; doubly, multiply by 2; triply, multiply by 3), divided by the total peak area of all species of the peptide and summed the results for the different species of the peptide. A sample calculation is shown below. It is assumed for the D.O.M calculation that the relative response factors for all modified and unmodified species of a peptide are the same. This assumption is supported by the consistent similarity observed in the tandem mass spectra between the different modified states of peptides (for an example see Figure 2-5). It should be noted that our analysis uses acidic conditions, so all bases should still be protonated initially. A change in charge distribution due to formaldehyde-induced modifications would have to be visible in altered fragment ion distribution, both in the type of fragment ions seen as well as their intensity, which is not the case. An assessment of the inter-day reproducibility for the D.O.M. calculation was performed on

two peptides (PLSRTLSVAAKK and pGlu-GLPPRPKIPP) and good reproducibility was observed with coefficient of variance (CV) of 11% (n=7) and 3% (n=5), respectively.

Sample Calculation:

After 10 min incubation with formaldehyde, the mass spectrum for peptide X showed it to be singly, doubly and triply modified with formaldehyde with peak areas (PA) of 2000, 3000, 5000, respectively.

D.O.M. = (PA singly/total PA*1) + (PA doubly/total PA*2) + (PA triply/total PA*3)

D.O.M. = (2000/10000*1) + (3000/10000*2) + (5000/10000)*3) = 2.3

The average degree of modification for peptide X is approximately 2.

To determine the actual sites of formaldehyde modification, D.O.M. was calculated for each detectable b- and y-ion fragments using the peak heights from the MS^2 spectrum of the most highly modified peptide species. Peak height as opposed to peak area was recorded due to the possibility of fragment ions peaks of very similar m/z overlapping that would negatively impact peak area to a greater extent than peak height. For clarification, a sample calculation for the b-ion series is shown below. For easy interpretation, results are expressed in a bar graph with peptide sequence along the x-axis and D.O.M. along the y-axis.

Sample Calculation:

After 10 min incubation with formaldehyde, the highest number of modifications observed for peptide AKK was two. MS/MS was performed on the doubly modified

peptide and the fragments ions b_1, b_2+12, b_3+24 were observed with peak heights (PH) of 300, 500 and 200, respectively.

D.O.M. for fragment b_1 = (PH singly/total PH*1) + (PH doubly/total PH*2)

= (0/300*1) + (0/300*2) = 0

Note: total PH for fragment b1 = PH b_1 + PH b_1+12 + PH b_1+24 = 300+0+0=300

D.O.M. for fragment b_2 = (PH singly/total PH*1) + (PH doubly/total PH*2)

= (500/500*1) + (0/500*2) = 1

D.O.M. for fragment b_3 = (PH singly/total PH*1) + (PH doubly/total PH*2)

= (0/200*1) + (2/200*2) = 2

The degree of modification for the b1, b2 and b3 fragments ions are 0, 1 and 2, respectively. Each incremental increase in degree of modification indicates a site of modification, therefore K2 and K3 were modified in this example.

To locate the reactive residues in a peptide sequence, the peak heights recorded for the detectable b- and y-ion fragments for the singly modified specie were used. As the fragment ions could only contain 0 or 1 modification in this case, their degree of modification was identical with their percentage of modification. The calculation expressed the peak height for the modified fragment ion as a percentage of the sum of it and its corresponding unmodified fragment ion (i.e., peak height modified b_3/(peak height modified b_3+peak height b_3)*100%). Once again for easy interpretation, results were expressed in a bar graph with peptide sequence along the x-axis and % modified along the y-axis.

2.3 Results and Discussion

In order to determine whether the results of extended formaldehyde treatment over several days are suitable substitutes for the formaldehyde cross-linking reaction in living cells, 14 model peptides that differed in length and amino acid composition were treated with formaldehyde at reaction conditions (100 μM, 83 mM formaldehyde, 10-20 minutes, 37°C) that are similar to those typically used for *in vivo* cross-linking. The reaction was quenched after 10 minutes or 2 days of incubation and the reaction mixture was subjected to mass spectrometric analysis to determine the mass-to-charge (m/z) ratios of the reaction products. Based on the reaction mechanism for formaldehyde cross-linking of protein, two types of peptide products were expected: modified or cross-linked peptides (Figure 2-1). The modified peptides may be in the form of the methylol intermediate (+30 Da) or, following the loss of water, an imine in the case of primary amines (+12 Da) (Figure 2-1, step 1). The imine formed between formaldehyde and a primary amine (free amino terminus or ε amino group of lysine) is called a Schiff base. Cross-linked peptides (Figure 2-1, step 2) could also be produced with an added mass of the two cross-linked peptides and the methylene bridge(s) joining them (increments of +12 Da).

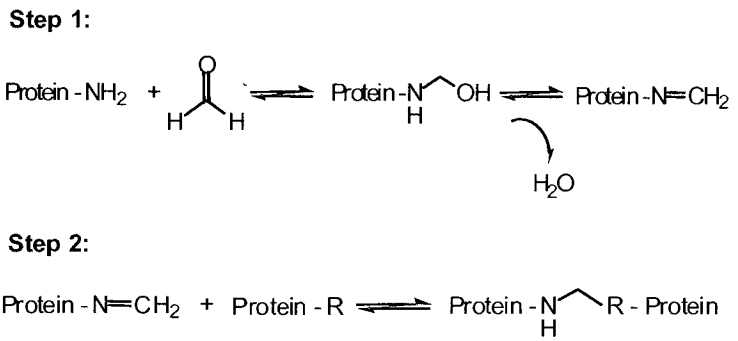


Figure 2-1. The two-step reaction for formaldehyde cross-linking.

The reaction products detected for the 14 model peptides under the two incubation conditions were entirely formaldehyde-induced modifications from the first step of the reaction, with mass shifts mainly in increments of +12 Da. No signals corresponding to cross-linked peptides were observed, which was not surprising since none was reported between peptides in previous studies *(16, 17)*. Conversely, the previously described modifications with mass shifts of +30 Da *(16, 17)* were not commonly detected. This seems to contradict previous reports that relied on electrospray as the ionization technique. It should be noted that the ionization technique used in this study (MALDI) dries the sample prior to analysis. This dehydration step may shift the equilibrium heavily towards the imine (Figure 2-1, step 1) thus converting the +30 Da intermediate into the +12 Da modification. Further investigations will be necessary to study the +30 Da modification in more detail.

As an example, Figure 2-2a and b shows the MALDI mass spectra of the reaction product distribution for peptide PLSRTLSVAAKK (m/z 1271) after 10 min and 2 days of formaldehyde treatment, respectively. The reaction products for peptide PLSRTLSVAAKK occurred solely in multiples of 12 Da of the unmodified peptide. After 10 min incubation with formaldehyde, a small amount of the unmodified peptide (0 Da) was still detected and the major products were the singly (+12 Da) and triply modified peptide species (+36 Da) (Figure 2-2a). With the extended incubation time (2 days), reaction products in multiples of 12 Da were still detected (Figure 2-2b, inset) but with a different product distribution. The unmodified peptide was absent and the major reaction products were the triply (+36 Da) and quintuply (+60 Da) modified species

(Figure 2-2b, inset). The abscissa for the mass spectrum obtained using a 2 days incubation time was extended to include the higher m/z region where cross-linked peptides would appear if present.

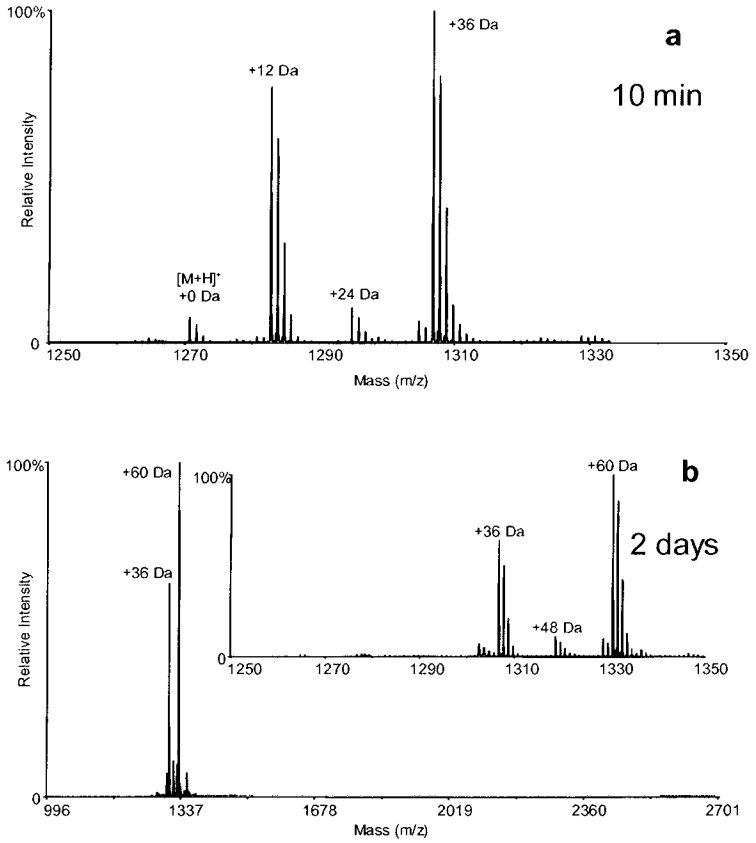


Figure 2-2. MALDI mass spectra showing the product distribution after formaldehyde incubation for peptide PLSRTLSVAAKK a) after 10 min and b) 2 days.

In an attempt to increase the cross-linking yield under an incubation time similar to what is used *in vivo*, the concentration of the peptide that showed the highest reactivity, PLSRTLSVAAKK, was increased several fold and mixed with equal amounts of one other peptide. The resulting equimolar mixtures were incubated with formaldehyde for 20 minutes, the reaction quenched, and mass spectra of the reaction products were acquired and inspected for the presence of cross-linking (Table 2-1). Three peptide

combinations (peptide PLSRTLSVAAKK with pGlu-GLPPRPKIPP, YGGFMRRVGRPE or pGlu-HWSYGLRPG-NH_2) yielded signals corresponding to cross-linked hetero-dimers with one peptide (pGlu-GLPPRPKIPP) yielding a cross-linked homo-dimer. Figure 2-3 shows the mass spectrum for the peptide combination PLSRTLSVAAKK (m/z 1271) and pGlu-GLPPRPKIPP (m/z 1182) that yielded both hetero- and homo-dimers. The predominant products from the reaction are non-cross-linked modifications of the two peptides, with up to two (+24 Da) and five modifications (+60 Da) detected separately for peptides pGlu-GLPPRPKIPP and PLSRTLSVAAKK, respectively. In comparison, very little of the cross-linked peptides was detected (Figure 2-3, inset), however based on the mass of the cross-linked peptides, the homo-dimer appears to be composed of two pGlu-GLPPRPKIPP peptides plus 4 modifications/methylene bridges and the three hetero-dimers composed of peptides PLSRTLSVAAKK with pGlu-GLPPRPKIPP plus 5-7 modifications/methylene bridges. As the low yield of cross-linked species proved insufficient for subsequent fragmentation analysis, no further investigations on the cross-linked peptides could be performed. For all of the combinations that yielded cross-linked products, the signal intensities of cross-linked peptides ranged from 0.01% to 0.4% of their respective monomer. Despite the lower resolution and signal response inherent in the MS analysis of larger peptides, non-specific cross-linking does not appear to be a major side reaction in these dilute solutions. Whether this is due to the lack of affinity between the peptides, their high diffusion rates in solution, or a combination of both remains to be seen. Nevertheless, it suggests that dilution of the reaction mixture upon cell lysis may contribute to efficient quenching of the formaldehyde cross-linking reaction of proteins in living cells.

Table 2-1. Overview of the cross-linked peptides produced after incubating peptide PLSRTLSVAAKK with another peptide (750 μM each) for 20 min in the presence of 333 mM formaldehyde.

Peptide Sequence	Cross-linked Peptides	Homo-dimer	Hetero-dimer
PGHDPPISYYETN-NH_2	No	-	-
pGlu-GLPPRPKIPP	Yes	√	√
pGlu-ADPNKFYGLM-NH_2	No	-	-
HDMNKVLDL	No	-	-
GAPVPYPDPLEPR	No	-	-
YGGFMRRVGRPE	Yes	-	√
pGlu-HWSYGLRPG-NH_2	Yes	-	√

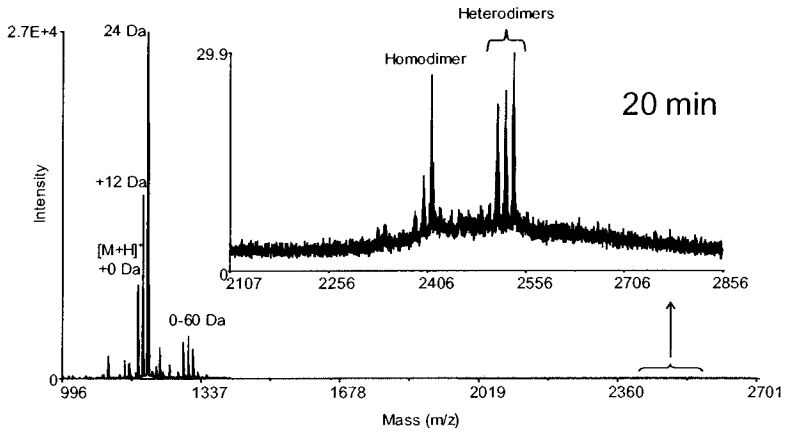


Figure 2-3. MALDI mass spectra showing the product distribution after formaldehyde incubation for peptides PLSRTLSVAAKK and pGlu-GLPPRPKIPP after 20 minutes.

With non-specific cross-linked peptides to be a minor product in the reaction, the focus was shifted back to the formaldehyde-induced modifications observed for the 14 model peptides after treatment with the cross-linking reagent for 10 min and 2 days. The distribution of the reaction products for each of the 14 peptides - listed as a percentage contribution of each species, and each peptide's overall reactivity – expressed in terms of average degree of modification, is shown in Table 2-2 for the 10 min incubation, and in Table 2-3 for the 2 day exposure to formaldehyde (see section *Interpretation of Data* for a detailed explanation of both calculations). Comparing the products distribution for each peptide under the two incubation times reveals an overall increase in the extent of

modification with time. This trend may be explained in each peptide by some residues are more reactive than others and are therefore modified quickly and detected in the 10 min sample. The less reactive residues that did not get significantly altered in 10 min, but eventually do become modified, are detected in sample after 2 days. For example, after a 10 min incubation with formaldehyde, peptide APGDRIYVHPF showed little modification with 98% of the peptide remaining unmodified, and only 2% singly modified (Table 2-2). This indicates that, in this peptide sequence, the amino acid residues are not very reactive towards formaldehyde. However, given enough time (2 days), these less reactive residues can still become modified, resulting in the conversion of 35% and 62% of the peptide to the singly (+12 Da) and doubly (+24 Da) modified species (Table 2-3), respectively. This difference in product distribution of each peptide for the two incubation periods suggests that using an extended incubation time (2 days) reduces the specificity of the amino acids that react with formaldehyde in the first step of the reaction (further discussed in a later section below).

Table 2-2. Overview of the products and calculated degree of modification (D.O.M.) for 14 peptides after 10 minutes of incubation with 83 mM formaldehyde.

Peptide Sequence	0, 0 Da	1, 12 Da	2, 24 Da	3, 36 Da	4, 48 Da	5, 60 Da	D.O.M
PLSRTLSVAAKK	3	40	4	51	0	2	2.11
MHRQETVDCLK-NH_2	33	33	20	14[a]	0	0	1.15
pGlu-GLPPRPKIPP	51	6	41	2[a]	0	0	0.94
pGlu-HWSYGLRPG-NH_2	27	69[a]	4[a]	0	0	0	0.77
PGHDPPISYYETN-NH_2	61	39	0	0	0	0	0.39
DYMGWMDF-NH_2	68	26	6	0	0	0	0.38
EQKLISEEDL-NH_2	81	13	6[b]	0	0	0	0.25
YGGFMRRVGRPE	88	8	3	0	0	0	0.14
YGGFMRGL	90	7	3	0	0	0	0.13
pGlu-ADPNKFYGLM-NH_2	91	9	0	0	0	0	0.09
RPPGFSPFR	96	1	3	0	0	0	0.07
GAPVPYPDPLEPR	97	3	0	0	0	0	0.03
HDMNKVLDL	97	3	0	0	0	0	0.03
APGDRIYVHPF	98	2	0	0	0	0	0.02

Note: fluctuations in the D.O.M. values are approximately ≤11% based on a 7-day inter-day reproducibility assessment.
[a] partially present as +30 Da modification, [b] partially oxidized

Table 2-3. Overview of the products and calculated degree of modification (D.O.M.) for 14 peptides after 2 days of incubation with 83 mM formaldehyde.

Peptide Sequence	0, 0 Da	1, 12 Da	2, 24 Da	3, 36 Da	4, 48 Da	5, 60 Da	D.O.M
PLSRTLSVAAKK	0	0	0	32	5	62	4.26
MHRQETVDCLK-NH_2	0	0	3	23	58[a]	17[a]	3.92
pGlu-GLPPRPKIPP	2	2	61	35[a]	0	0	2.29
HDMNKVLDL	4	4	87	5	0	0	1.93
RPPGFSPFR	9	3	87	1	0	0	1.80
APGDRIYVHPF	4	35	62	0	0	0	1.59
GAPVPYPDPLEPR	0	44	56[b]	0	0	0	1.56
EQKLISEEDL-NH_2	0	52	44[b]	5	0	0	1.55
YGGFMRRVGRPE	1	61	38	0	0	0	1.37
YGGFMRGL	1	62	36	0	0	0	1.34
pGlu-HWSYGLRPG-NH_2	13	54[a]	28[a]	4[a]	0	0	1.22
DYMGWMDF-NH_2	0	83	17	0	0	0	1.17
pGlu-ADPNKFYGLM-NH_2	26	47	28[a]	0	0	0	1.03
PGHDPPISYYETN-NH_2	0	98	2	0	0	0	1.02

Note: fluctuations in the D.O.M. values are approximately ≤11% based on a 7-day inter-day reproducibility assessment.
[a] partially present as +30 Da modification, [b] partially oxidized

In addition, this difference in reactivity between amino acid residues manifests itself in the degree of modification (D.O.M.) observed for the 14 peptides within each incubation period. After 2 days, the most reactive peptide was modified 4.26-fold, while the least reactive peptide showed an average degree of modification of 1.02 (Table 2-3), indicating that under these conditions, each peptide contained at least one site that was susceptible to formaldehyde-induced modification. Given that these peptides were chosen to have a

wide range of amino acid compositions, this seemed to support previous reports that identified a large number of amino acid residues as targets of this reaction *(16, 17)*. While incubation for 10 minutes also resulted in varying degrees of modification, the overall spread was lower, with 2.11-fold being the highest observed value (Table 2-2). A number of peptides showed very little modification, with seven peptides falling below 0.2-fold, and three peptides containing essentially no modification at all. Given the overall lower degree of modification with the shorter incubation time (10 min), it appears that although there are a number of residues highly reactive to PFA under these conditions, it is but a subset of these that would react if allowed to proceed for 2 days

A possible complication for the aforementioned comparison was that for these reaction conditions, the influence of kinetic effects on the relative degree of modification had not been determined. Consequently, additional studies were carried out to gain more insight into the factors that influence the relative rates of peptide modification. The most reactive peptide (PLSRTLSVAAKK), from Tables 2-2 and 2-3, was chosen for these studies due to its quick response to formaldehyde treatment. First, the peptide was reacted with formaldehyde, and aliquots of the reaction were taken at various time points to obtain a more detailed look at how the reaction proceeds over time (5 min – 2 days). The time points from Tables 2-2 and 2-3 were encompassed in this study in order to monitor the progress of the reaction between these two vastly different times. Following quenching of the reaction, the degree of modification was plotted against the incubation time (Figure 2-4a). A non-linear dependence of the degree of modification on the reaction time was observed, with the increase in degree of modification being higher at

early than at later time points. This suggests that the more reactive residues on the peptide were modified quickly, so it is desirable to keep the reaction fast. It can also be concluded that the subsequent decrease in the rate of modification at the later time points indicates that the reaction ultimately reaches saturation and the apparent increase after 2 days is sufficiently small to conclude that the reaction has reached completion. Even though the peptide concentration varied slightly between this study (75 μM) and the ones used to generate Tables 2-2 and 2-3 (100 μM), the degree of modification for the two overlapping time points (10 min, 2 days) was fairly similar. The same study was performed for peptide pGlu-GLPPRPKIPP and a similar trend was observed.

As this suggested an important role of the reaction kinetics for short reaction times, additional experiments were performed to test the effects of the relative reactant concentration on the degree of modification. Both of the aforementioned peptides were exposed to various formaldehyde concentrations for 10 minutes before quenching the reaction, and the resulting degrees of modification were plotted (Figure 2-4b). For both peptides, an increase in the degree of modification was observed with increasing formaldehyde concentrations, with a higher overall reactivity observed for peptide PLSRTLSVAAKK relative to peptide pGlu-GLPPRPKIPP. The profile of the graph is very similar to the graph generated by varying incubation time (Figure 2-4a) and suggests that PFA concentration is also key to specificity. Although both graphs (Figure 2-4a & b) show very similar profiles, it should be noted that at the highest formaldehyde concentration tested (333 mM) the degree of modification does not reach the same level as it did for the long incubation times in Figure 2-4a. Similar results were obtained when

the peptide concentration was varied. To this end, peptide PLSRTLSVAAKK was incubated with formaldehyde at various concentrations for 5 or 10 minutes, respectively, before the reaction was quenched and analyzed by MALDI-MS. The resulting plots show a non-linear increase in the degree of modification for both time points, with concentration increases having a more pronounced effect on the degree of modification for lower than for higher peptide concentrations indicating a diffusion controlled reaction rate (Figure 2-4c). A higher degree of modification was observed for all peptide concentrations after 10 minutes compared to 5 minutes, which is in agreement with the results obtained for the time-course study, while the overall trends were very similar.

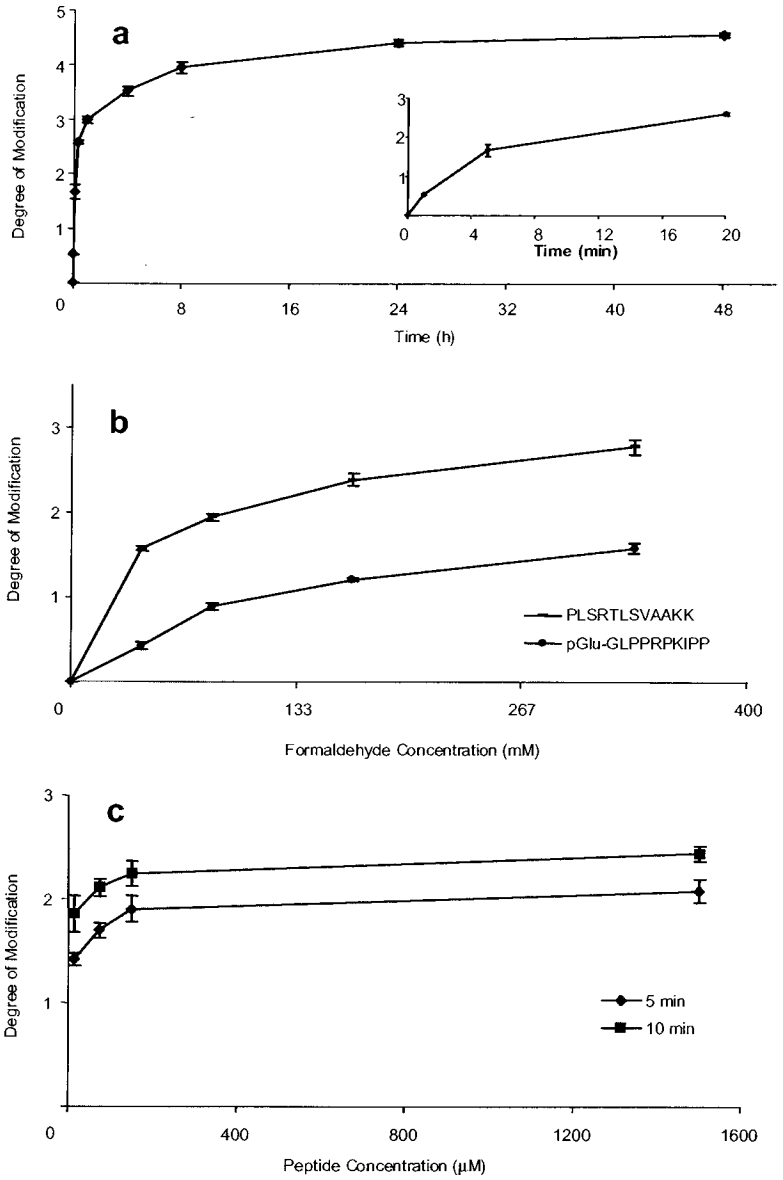


Figure 2-4. Degree of formaldehyde-induced modification on peptide a) PLSRTLSVAAKK versus incubation time, b) PLSRTLSVAAKK and pGlu-GLPPRPKIPP versus increasing formaldehyde concentration and c) PLSRTLSVAAKK versus increasing peptide concentration.

Although major differences in reactivity were observed between 10-minute and 2-day incubations, no clear link between amino acid composition and reactivity could be established from MS data alone. In order to gain more insight into the influence of the peptide sequence on its reactivity under a short incubation time (10 min), information on

the actual modified amino acid residues was sought. To this end, the reaction products of the seven peptides that showed degrees of modification higher than 0.2-fold after 10 minutes of incubation (Table 2-2) were subjected to tandem mass spectrometric analysis. Two separate MS/MS experiments were carried out, in which either the singly modified peptide or the peptide with the highest number of modifications was isolated and fragmented. As an example, the resulting tandem mass spectra of the triply modified and the singly modified peptide PLSRTLSVAAKK are shown in Figure 2-5a and 2-5b, respectively. Comparing the two tandem mass spectra, it became apparent that the fragment ion distributions were very similar, i.e. the same level of sequence information was retained in both, as well as the unmodified peptide (data not shown). Assignment of individual fragment ions revealed almost complete b- and y-ion series in both spectra, and made possible the determination of the modification state of each amino acid residue in the modified peptide. The similarity in the tandem mass spectra between the different modified states of the peptide (unmodified, singly & triply modified) also helps validate certain assumptions made in calculating degree modification – that the relative response factor of the different modified species are similar. Our analysis uses acidic conditions, so all bases should still be protonated initially, a change in charge distribution due to formaldehyde-induced modifications would have to be visible in altered fragment ion distribution, both in the type of fragment ions seen as well as their intensity, which is not the case.

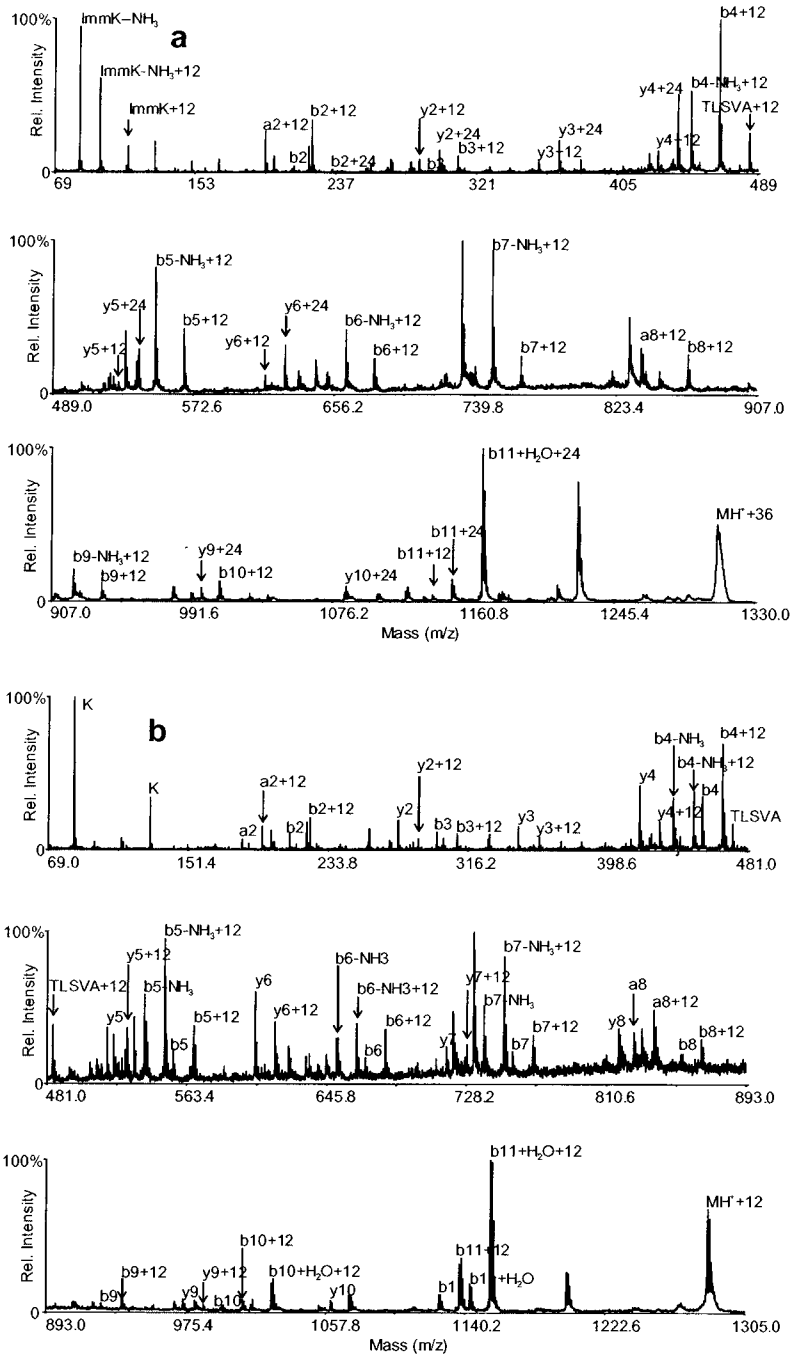


Figure 2-5. MALDI tandem mass spectra of a) triply modified and b) singly modified peptide PLSRTLSVAAKK. Key fragment ions are highlighted.

In order to facilitate the interpretation of these spectra, a different visualization was devised, highlighting the positions of the modified residues and their degrees of

modification. For each of the b- and y-ions, the degree of modification was determined by applying the same approach that was used to calculate the average degree of modification of each intact peptide. This value was then plotted against the peptide sequence, resulting in bar graphs for the b- and the y-ion series of the peptide that could then be inspected. Note that the y-ion which contains the N-terminus is in fact the precursor ion and therefore is fully modified. For the triply modified peptide PLSRTLSVAAKK, the mass spectrum shown in Figure 2-5a was thus converted into the bar graph shown in Figure 2-6a. The degree of modification plotted on the y-axis showed significant increases at the amino-terminal proline, and the two carboxy-terminal lysine residues, as well as almost identical values for the sequence stretch between these residues in both ion series. The small decrease in degree of modification for the b3 ion in comparison to fragment ions b4-b10, as well as the slight fluctuation in degree of modification for y ions 2-10 in Figure 2-6a is due to variability in the method. This identified the amino-terminus and the two carboxy-terminal lysine residues as the main modified residues for this peptide and demonstrates that by fragmenting the most highly modified peptide species, the sites of modification can be determined.

In a similar fashion, to determine the reactivity difference between the above mentioned residues, the mass spectrum of the singly modified peptide PLSRTLSVAAKK shown in Figure 2-5b gave rise to the bar graph shown in Figure 2-6b. Coming from a singly modified precursor the fragment ions could only contain 0 or 1 modification therefore their degree of modification was identical with their percentage of modification. By

plotting the percentage of modification on the y-axis, the contribution of individual reactive sites to the total reactivity of the peptide could be determined. A 60-70% contribution was determined for the amino-terminal proline (most reactive residue), while the remaining 30-40% were found on the C-terminal lysine. In contrast, the penultimate lysine showed very little change associated with its b-ion series, suggesting that this residue contributed only marginally to the modification of the singly modified peptide species.

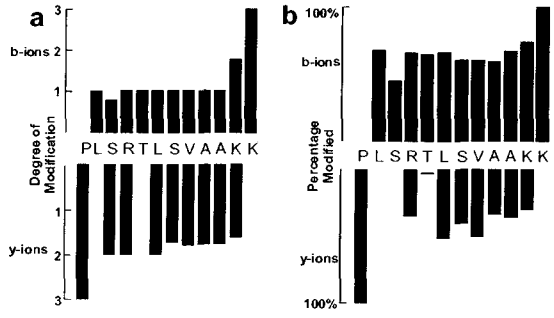


Figure 2-6. Degree of formaldehyde-induced modification for the b-ion (top) and y-ion (bottom) series of a) triply and b) singly modified peptide PLSRTLSVAAKK.

This example demonstrated that when peptides can be isolated in their singly and multiply modified forms, complementary information becomes accessible that can be used to gain additional insight into the reaction of the peptide with formaldehyde. Specifically, the peptide species with the highest number of modifications highlights the individual modification sites, while the singly modified peptide species provides information on the relative reactivity of those residues. However, information on the reactivity and the modification sites converges when the singly modified peptide species also represents the highest degree of modification, which was the case for several peptides after a 10-minute incubation.

Additional insight into reactive residues was gained by analyzing the bar graphs of the singly and multiply modified species of the other six most reactive peptides after 10 minutes of formaldehyde incubation. For peptide PGHDPPISYYETN-NH_2 (Figure 2-7a), all visible b-ions (b4 and higher) appeared in their modified form, while the y-ions were not modified until after y10 (as indicated by the presence of unmodified (0%) y8-y10 fragment ions). This information localized the modification to the amino-terminal end of the peptide (residues PGH), but only upon subsequent MS/MS/MS analysis of the b4+12 ion at 419.2 Th was it shown that the modification was exclusively present on the amino-terminal proline (Figure 2-8a). While the internal fragment ion HD indicated that neither of these residues was modified, the b2+12 and a2+12 fragment ions unequivocally localized the modification to the amino-terminal proline, as glycine does not contain any sites that are susceptible to formaldehyde modification. Likewise, tandem mass spectrometric analysis of the singly modified peptide pGlu-HWSYGLRPG-NH_2 showed tryptophan to be a reactive site based on the b- and y-ion series (Figure 2-7b). This could be confirmed by the presence of the corresponding modified immonium ion at 171.08 Th (Figure 2-8b). These targeted fragmentation studies identified the amino-termini and the side chains of lysine and tryptophan residues to be the main reaction products after 10 minutes of formaldehyde incubation. Combining this information with that presented in previous publications *(16, 17)* leads to the proposed chemical structures of the individual modification sites shown in Figure 2-9.

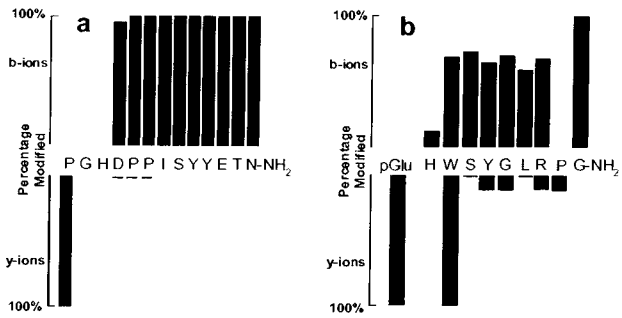


Figure 2-7. Degree of formaldehyde-induced modification for the b-ion (top) and y-ion (bottom) series of singly modified peptides a) PGHDPPISYYETN-NH_2 and b) pGlu-HWSYGLRPG-NH_2.

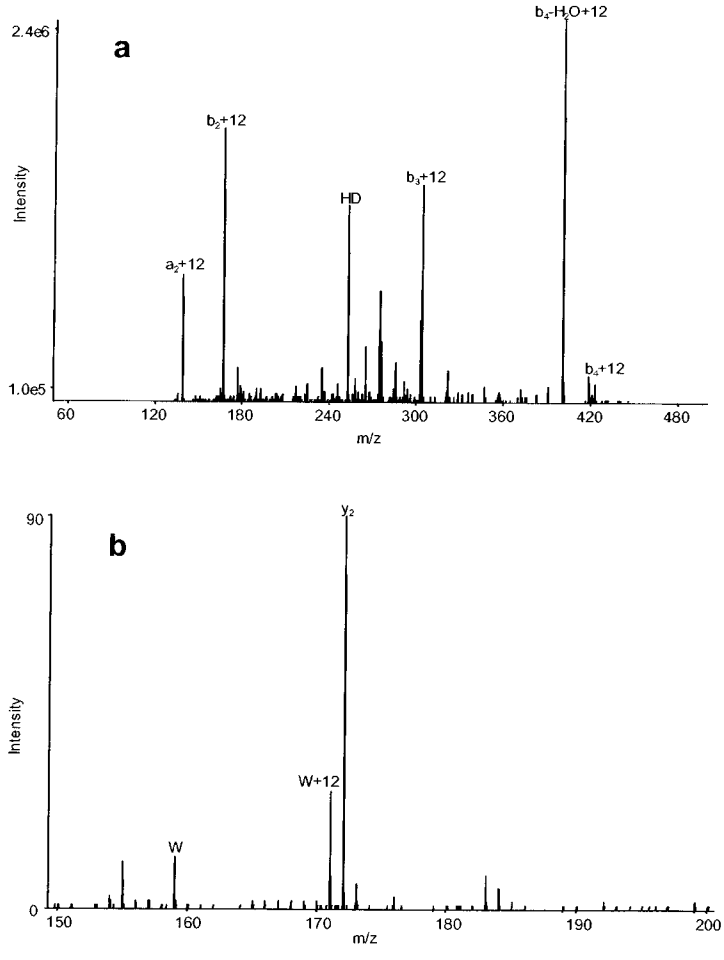


Figure 2-8. Identification of individual modification sites by their +12 Da mass shifts as a) the amino-terminal proline in the MS/MS/MS spectrum of the peptide

PGHDPPISYYETN-NH_2 and b) the tryptophan side chain in the immonium region of the MS/MS spectrum of the peptide pGlu-HWSYGLRPG-NH_2.

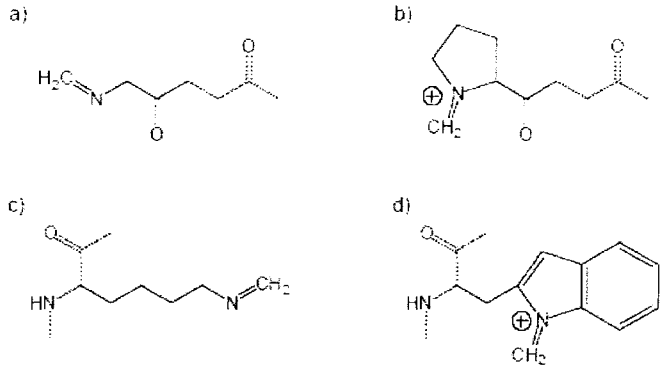


Figure 2-9. Proposed structures of the modification sites on a) primary amino-terminal groups, b) amino-terminal proline, c) lysine side chains, and d) tryptophan side chains.

Upon performing this type of analysis for all of the seven reactive peptides, the information on the reactive residues was summarized in Table 2-4. This list includes primary amino groups at the peptides' amino-termini and lysine side chains, which are functional groups known to be highly reactive nucleophiles. Interestingly, arginine as an additional basic residue showed significantly lower reactivity under these conditions, despite the fact that it was reported to be highly reactive in previous studies *(16, 17)*. Moreover, other amino acid residues that contain nucleophilic groups such as cysteine as well as serine, threonine, and tyrosine were not reactive at all under these conditions. As the secondary amino groups of tryptophan side chains and amino-terminal proline, usually not considered to be major reactive sites in nucleophilic substitutions, were also found to be reactive, it suggests that additional factors may make the formaldehyde reaction under these conditions more complex.

That such an additional level of complexity may exist is also supported by the findings of a comparison of the reactive residues in Table 2-4 with the modifiable peptides in Tables 2-2 and 2-3. While several peptides that contained lysine groups had significant reactivity associated with these residues at the 10-minute time point, several of the non-reactive peptides also contained lysine residues. The reactivity of the same type of residue appeared to be higher when it was located at or close to one of the peptide's termini than when it had a number of flanking residues on either side. These less reactive residues were found modified after 2 days (data not shown), which suggests that their reduced reactivity is caused by their surface accessibility, as observed in a previous study *(17)*. This effect may have major consequences when cross-linking proteins *in vivo*. If only readily accessible amino acid residues are highly reactive under the conditions outlined here, then highly folded proteins may be reactive at few locations other than highly accessible binding sites.

Table 2-4. Overview of the reactive residues for the seven most reactive peptides after 10 minutes of incubation with 83 mM formaldehyde.

Peptide Sequence	Modified Residues in Order of Decreasing Reactivity
PLSRTLSVAAKK	P1, K12, K11
MHRQETVDCLK-NH_2	K11, M1, H2
pGlu-GLPPRPK1PP	K8, R6
pGlu-HWSYGLRPG-NH_2	W3
PGHDPPISYYETN-NH_2	P1
DYMGWMDF-NH_2	D1
EQKLISEEDL-NH_2	E1, K3

2.4 Conclusions

This study was undertaken to determine whether previously identified amino acid residues that react with formaldehyde over the course of several days, i.e. amino-termini, Lys, Arg, His, Cys, Tyr, Trp and Phe are identical to those involved in formaldehyde cross-linking in living cells. From the results presented here, it is evident that these two conditions lead to overlapping, but clearly distinct outcomes. Although the residues that have reacted with formaldehyde to a significant extent after 10 minutes, i.e. amino-termini, Lys, and Trp, are contained in the list of those found modified after several days, they represent a much smaller subset. Targeted mass spectrometric experiments were used to detect and identify the individual modification sites. Key questions that remain are how significant the accessibility of residues is to their reactivity, and whether the low yield of non-specific cross-linking is due to the lack of affinity between the peptides, their high diffusion rates in solution, or a combination of both. Additional studies on folded proteins will be necessary to address these questions. These are currently being carried out and will undoubtedly shed more light onto the relative contribution of these factors.

2.5 Acknowledgments

The authors would like to thank Jason Rogalski for his support in the acquisition and interpretation of mass spectra, as well as Brent Sutherland and Jason Rogalski for critical reading of the manuscript. This work was funded by an operating grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), and infrastructure

grants by the Michael Smith Foundation for Health Research (MSFHR) and the Canada Foundation for Innovation (CFI).

2.6 References

1. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) *Nature* **415**, 141-7.

2. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002) *Nature* **415**, 180-3.

3. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari,

G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) *Nature* **440**, 631-6.

4. Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005) *Nature* **433**, 531-7.

5. Blagoev, B., Kratchmarova, I., Ong, S. E., Nielsen, M., Foster, L. J., and Mann, M. (2003) *Nat Biotechnol* **21**, 315-8.

6. Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P. O., Bergamini, G., Croughton, K., Cruciat, C., Eberhard, D., Gagneur, J., Ghidelli, S., Hopf, C., Huhse, B., Mangano, R., Michon, A. M., Schirle, M., Schlegl, J., Schwab, M., Stein, M. A., Bauer, A., Casari, G., Drewes, G., Gavin, A. C., Jackson, D. B., Joberty, G., Neubauer, G., Rick, J., Kuster, B., and Superti-Furga, G. (2004) *Nat Cell Biol* **6**, 97-105.

7. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999) *Nat Biotechnol* **17**, 1030-2.

8. Cristea, I. M., Williams, R., Chait, B. T., and Rout, M. P. (2005) *Mol Cell Proteomics* **4**, 1933-41.

9. Vasilescu, J., Guo, X., and Kast, J. (2004) *Proteomics* **4**, 3845-54.

10. Tagwerker, C., Flick, K., Cui, M., Guerrero, C., Dou, Y., Auer, B., Baldi, P., Huang, L., and Kaiser, P. (2006) *Mol Cell Proteomics* **5**, 737-48.

11. Guerrero, C., Tagwerker, C., Kaiser, P., and Huang, L. (2006) Mol Cell *Proteomics* **5**, 366-78.

12. Schmitt-Ulms, G., Hansen, K., Liu, J., Cowdrey, C., Yang, J., DeArmond, S. J., Cohen, F. E., Prusiner, S. B., and Baldwin, M. A. (2004) *Nat Biotechnol* **22**, 724-31.

13. Sinz, A. (2003) *J Mass Spectrom* **38**, 1225-37.

14. Trakselis, M. A., Alley, S. C., and Ishmael, F. T. (2005) *Bioconjug Chem* **16**, 741-50.

15. Sinz, A. (2006) *Mass Spectrom Rev* **25**, 663-82.

16. Metz, B., Kersten, G. F., Hoogerhout, P., Brugghe, H. F., Timmermans, H. A., de Jong, A., Meiring, H., ten Hove, J., Hennink, W. E., Crommelin, D. J., and Jiskoot, W. (2004) *J Biol Chem* **279**, 6235-43.

17. Metz, B., Kersten, G. F., Baart, G. J., de Jong, A., Meiring, H., ten Hove, J., van Steenbergen, M. J., Hennink, W. E., Crommelin, D. J., and Jiskoot, W. (2006) *Bioconjug Chem* **17**, 815-22.

3 SOLVENT ACCESSIBILITY AND THE RELATIVE REACTIVITY OF BASIC RESIDUES IN FORMALDEHYDE-INDUCED MODIFICATIONS[6]

3.1 Introduction

Chemical cross-linking, in combination with mass spectrometric product analysis, is emerging as a powerful tool to study the interactions and geometries of protein complexes. Commonly applied cross-linking strategies rely on well-defined conditions *in vitro* to react homo- or heterobifunctional cross-linkers of characteristic linker lengths with purified proteins or protein complexes. Enzymatic digestion of the cross-linked protein complexes and mass spectrometric analysis of the resultant peptides are then carried out to determine the interaction sites of the proteins. Distinguishing cross-linking products from the bulk of unchanged peptides has turned out to be challenging, thus several strategies to highlight the cross-linked dipeptides by incorporating stable isotopes or photo-induced reporter groups into the cross-linking reagent have recently been introduced *(1-6)*.

A major limitation of the *in vitro* cross-linking strategy is the lack of cellular context of the interactions. Three systematic issues may affect these types of studies in particular: the absence of additional proteins that directly or indirectly interact with one of the complex components or compete for binding; changes in the tertiary structure of the proteins due to considerably lower concentration of salts, co-factors, and other proteins; and altered protein diffusion rates in dilute solutions relative to those observed in high-

[6] A version of this chapter will be submitted for publication. Toews, J., Rogalski, J. C., Kast, J. Solvent Accessibility and the Relative Reactivity of Basic Residues in Formaldehyde-induced Modifications.

density cellular environments. Performing the cross-linking reaction inside a living cell is a promising way to circumvent these restrictions. A number of groups have reported the feasibility of this approach using formaldehyde as a cross-linking reagent *(7-11)*. Owing to its small size, it rapidly permeates membranes and quickly diffuses inside cells, thus is readily available for cross-linking of protein complexes without the need for prior cell lysis. Formaldehyde cross-linking in live cells has been utilized successfully to enrich for bait proteins and many cross-linked interaction partners that could then be identified by mass spectrometry, yet cross-linked dipeptides, indicative of the actual interaction site, have not been reported in any of these cases.

This lack of knowledge about the cross-linking sites on proteins upon *in vivo* formaldehyde treatment can be attributed to two challenges. The dipeptides containing this information are usually obscured by a complex matrix of more abundant peptides that are unaffected by formaldehyde treatment. This is compounded, however, by a limited understanding of what governs the reactivity of proteins under these conditions. Although two recent studies determined the products of the formaldehyde reaction with model peptides and proteins *(12, 13)*, the long formaldehyde exposure of several days used therein is not suitable for formaldehyde cross-linking inside cells, as it would lead to extensive protein modification, protein insolubility, and sample loss. We recently demonstrated that the side chains of lysine and tryptophan residues, as well as amino-termini, were the dominant reactive sites in model peptides under intracellular protein cross-linking reaction conditions *(14)*, however on proteins the dominant reactive site is

the ε-amino group of lysine. However this study also indicated that accessibility of reactive residues likely influenced their reactivity.

The current study was designed to investigate whether complex tertiary structural characteristic of proteins influences formaldehyde-induced reaction rates on modification. The data suggest that a folded protein having fewer solvent exposed reactive residues shows fewer formaldehyde-induced modifications than an unfolded protein. This indicates that the protein remains folded over the course of the reaction and that the extent of formaldehyde-induced modifications can be directly related to the extent of the protein unfolding. Moreover, a relationship between accessibility and reactivity is evident may enable for the creation of surface topology maps to help highlight residues with unexpected reactivity, e.g. those of active sites in enzymes, as well as aid in determination of the cross-linked interaction sites.

3.2 Experimental

3.2.1 Chemicals

Myoglobin from equine heart, carbonic anhydrase I (CA I) from human erythrocytes, lysozyme from chicken egg white, α-cyano-4-hydroxycinnamic acid (CHCA), dithiothreitol (DDT), trizma base, ammonium bicarbonate and sodium hydroxide were all purchased from Sigma (St. Louis, MO). Guanidine hydrochloride (Gdn HCl) was obtained from Schwarz/Mann Biotech (Cleveland, Ohio). Paraformaldehyde (PFA), formic acid (FA, 88%) and acetonitrile (ACN, HPLC grade) were purchased from Fisher (Fair Lawn, NJ). Endoproteinase Glu-C was obtained from Roche Applied Science

(Penzberg, Germany). C4 and C18 extraction tips were obtained from Varian (Lake Forest, CA). 3 kDa MW cut-off filters, 0.22 μm filters were purchased from Pall Corporation (Ann Arbor, MI). 3.5 kDa MW cut-off dialysis cassettes were obtained from Pierce (Rockford, IL). Deionized water (18 MΩ·cm) was prepared using a Nanopure Ultrapure Water System from Barnstead (Dubuque, IA).

3.2.2 Preparation of Paraformaldehyde Solution

A 1.3 M formaldehyde stock solution was prepared by heating (80°C) PFA in phosphobuffer saline (PBS, pH 7.5) for 30 min, cooling to room temperature and filtering through a 0.22 μm filter.

3.2.3 Reactivity of Folded versus Unfolded Myoglobin

Myoglobin (100 μM) was incubated in PBS (37°C) in the presence of 333 mM PFA for 20 min. To unfold the protein, 3 M Gdn HCl was included in the sample. Control samples were prepared by replacing the PFA volume with PBS. Reactions were quenched by the addition of 1 M tris buffer (pH 7.5). For this experiment a stock solution of 6 M Gdn HCl was prepared in PBS. For sample clean-up, see the appropriate section below.

3.2.4 Glu-C Digestion

Proteins were digested overnight at room temperature in 50 mM ammonium bicarbonate (pH 8) with endoproteinase Glu-C. The digestion was quenched using 5% formic acid in water, lyophilized and resuspended in 0.01% formic acid in water.

3.2.5 Reactivity of Folded versus Unfolded Protein

Carbonic anhydrase I and lysozyme (50-100 μM) were incubated in PBS (37°C) in the presence of 333 mM PFA for 20 min. To unfold the protein, 6 M Gdn HCl was added to the sample. To reduce the disulphide bonds in lysozyme, 100 mM DTT was also added to the sample and heated for 1 h at 56°C prior to reaction at room temperature with PFA. Control samples were prepared by replacing the PFA volume with PBS. Reactions were quenched by the addition of 1 M tris buffer (pH 7.5). Protein was isolated and concentrated using either C4/C18 extraction tips and eluted in 75:25 ACN:5% FA.

3.2.6 Change in Reactivity with Varying Concentrations of Gdn HCl

Myoglobin (100 μM) was incubated in PBS (37°C) in the presence of 333 mM PFA and 0, 0.5, 1, 2, 3, 4, 5, 6 or 7 M Gdn HCl for 5 min. Control samples were prepared by replacing PFA and Gdn HCl with PBS. Reactions were quenched by the addition of 1 M tris buffer (pH 7.5). For these experiments a stock solution of Gdn HCl was prepared at 8 M in PBS and adjusted with sodium hydroxide to maintain physiological pH. For sample clean-up, see the appropriate section below.

3.2.7 Comparison of the Electrospray Charge State Profile of Folded Untreated and Folded Formaldehyde-Treated Myoglobin

Myoglobin (100 μM) was incubated in PBS (37°C) in the presence or absence of 333 mM PFA for 50 min. The reactions were quenched by the addition of 1 M tris buffer (pH 7.5). Proteins were washed and resuspended from the filter membranes using either

water (pH 6) or 0.01% formic acid in water (pH 3). The protein resuspended in water (pH 6) was used to prepare the pH 4 and 5 samples by diluting with 0.001 and 1.75×10^{-4} % formic acid in water, respectively. All samples were diluted with ACN for analysis by ESI-MS (2000 Q Trap, Applied Biosystems, Foster City, CA). The average charge state for each spectrum was calculated using weighted peak areas for all charge states with signal to noise ratio greater than 5.

3.2.8 Unfolding Myoglobin During Reaction with Formaldehyde

Myoglobin (100 μM) was incubated in PBS (37°C) in the presence of 333 mM PFA for 30 min at which time the protein was unfolded by addition of 3 M Gdn HCl. The reaction was continued for another 20 min. Aliquots were taken throughout the experiment at times between 5 and 50 min. Two similar experiments were performed, one without the addition of 3 M Gdn HCl and one with the addition of Gdn HCl at t=0 min. All three experiments were compared to a control which lacked both Gdn HCl and treatment with PFA. The reactions were quenched by the addition of 1 M tris buffer (pH 7.5). For sample clean-up, see section below.

3.2.9 Myoglobin Sample Clean-up

To avoid precipitation of the Gdn HCl-containing protein samples, they were first subjected to dialysis prior to filtration. The Gdn HCl-containing samples were dialyzed (3.5 kDa cassettes) against 50 mM tris buffer for 2 h at room temperature. The dialysis buffer was changed and dialysis continued for another 2 h at room temperature. The

dialysis buffer was changed again and the sample was dialyzed overnight at 4°C. For the dialyzed samples and all other non Gdn HCl-containing samples, myoglobin was isolated and concentrated using 3.5 kDa MW-cut off filters. Proteins were washed and resuspended from the filter membranes using 0.01% formic acid in water.

3.2.10 Mass Spectrometric Analysis

Protein and peptide samples were mixed with saturated solution of CHCA (in 50:50 ACN:5% FA) and spotted onto a MALDI target using the dried-droplet method, air dried and analyzed by MALDI-TOF-TOF (4700 Proteomics Analyzer) controlled by 4700 Explorer version 2.0 software (Applied Biosystems, Foster City, CA). MS spectra for proteins and peptides were collected with a total of 5000 (100 shots/subspectrum, 50 subspectra) and 3000 (75 shots/subspectrum, 40 subspectra), respectively while MS^2 spectra were collected with a total of 10 000 shots (100 shots/subspectrum, 100 subspectra). Tandem mass spectra were acquired with atmospheric gas as the collision gas at a pressure of 3-4 x 10^{-8} Torr and collision energy of 1000 eV. For the analysis of whole protein (myoglobin, CA I and lysozyme), the centroid mass was recorded. For the analysis of the myoglobin protein digest, peak areas for each peptide and its related products were recorded. MS/MS spectra were obtained for unmodified and formaldehyde-modified peptides by MALDI-TOF-TOF-MS. Peak height for detectable b- and y- ion fragments were recorded.

3.2.11 Interpretation of Data

As previously described *(14)*, to compare the overall reactivity of each peptide, the average number of formaldehyde-induced modification for a given peptide, the "degree of modification" (D.O.M.) was calculated. The calculation took the peak area of a modified species, multiplying it by the number of modifications that it contains (i.e., singly, multiply by 1; doubly, multiply by 2; triply, multiply by 3), divided by the total peak area of all species of the peptide and summed the results for the different species of the peptide. It is assumed for the D.O.M calculation that the relative response factors for all modified and unmodified species of a peptide are similar. This assumption is supported by the consistent similarity observed in the tandem mass spectra between the different modified states of peptides observed in the peptides obtained from proteolysis (data not shown) as well as in model peptides from a previous study *(14)*. Our analysis uses acidic conditions, so all bases should still be protonated initially. A change in charge distribution due to formaldehyde-induced modifications would have to be visible in altered fragment ion distribution, both in the type of fragment ions seen as well as their intensity, which is not the case. An assessment of the inter-day reproducibility for the D.O.M. calculation was performed on two peptides (PLSRTLSVAAKK and pGlu-GLPPRPKIPP) and good reproducibility was observed with coefficient of variance (CV) of 11% (n=7) and 3% (n=5), respectively.

As previously described *(14)*, to determine the actual sites of formaldehyde modification, D.O.M. was calculated for each detectable b- and y-ion fragments using the peak heights from the MS^2 spectrum of the most highly modified peptide species. Peak height as opposed to peak area was recorded due to the possibility of fragment ions peaks of very similar m/z overlapping that would negatively impact peak area to a greater extent than peak height. For clarification, a sample calculation for the b-ion series is shown below. For easy interpretation, results are expressed in a bar graph with peptide sequence along the x-axis and D.O.M. along the y-axis.

3.3 Results and Discussion

A previous study suggested that relative position of reactive residues in a peptide sequence greatly affects the extent of formaldehyde modification under short incubation conditions. For instance, terminal lysine residues reacted much more readily than internal lysines. Hence, solvent accessibility appears to be a contributing factor in determining reactivity between amino acid residues and formaldehyde *(14)*. To further examine this association, we investigated how the formaldehyde reaction behaves using more complex protein systems, which contains tertiary structures. We speculated that differences in reactivity due to tertiary structural elements might manifest themselves as differences in the products of the reactions of PFA between folded and unfolded states of any given protein. Myoglobin was used in our initial study because it previously shown to be completely unfolded in 3 M Gdn HCl *(15, 16)*. Therefore we examined the protein's reactivity in a buffered solution containing 333 mM PFA in the presence or absence of 3 M Gdn HCl. As shown in Figure 3-1, the mass spectra of the two PFA-

treated samples (Figure 3-1c, d) show a mass shift over their non-PFA treated controls (Figure 3-1 a, b). The folded PFA-treated protein (Figure 3-1c) showed a very small mass shift from its unmodified form (Figure 3-1a), indicative of little modification under the conditions used. The unfolded PFA-treated protein (Figure 3-1d), however, produced a much higher mass shift from its unmodified form (Figure 3-1b), indicating an increased level of modifications compared to its folded counterpart.

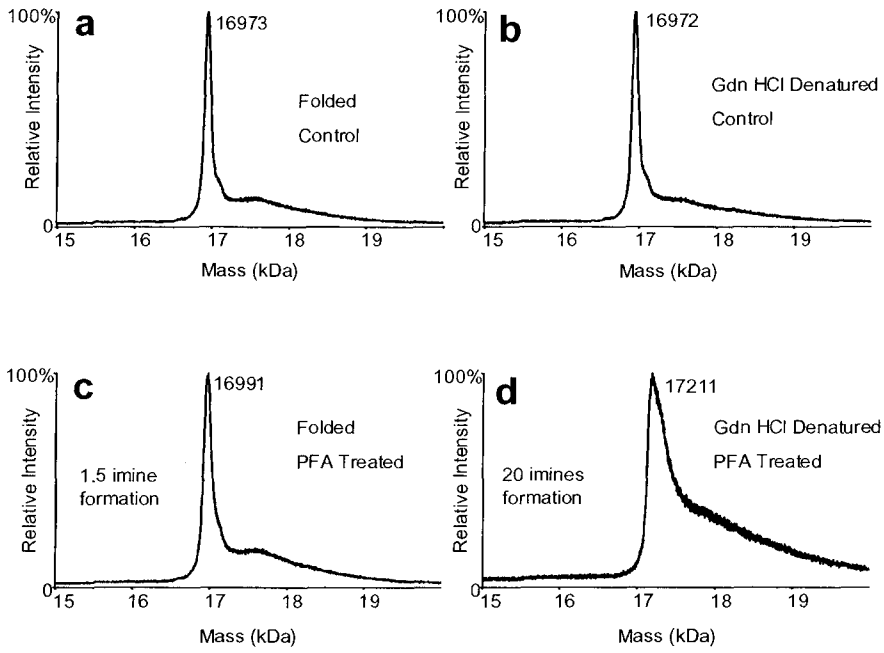


Figure 3-1. MALDI MS spectra showing the [M+H]+ of myoglobin a) untreated (control), and treated with b) guanidine hydrochloride, c) paraformaldehyde and d) guanidine hydrochloride and paraformaldehyde.

This suggests that denaturing the tertiary structure of myoglobin allows reactive residues that were otherwise solvent inaccessible to become accessible, and modified, thus leading to an overall higher reactivity, translating in a higher mass shift and a higher number of potential isoforms. Indeed, since formaldehyde-induced modifications appear as mass shifts in multiples of 12 Da *(14)*, we estimate that the folded and unfolded proteins

produced mass shifts upon treatment with PFA which correspond to the formation of an average of 1.5 and 20 modifications, respectively. The controls showed no difference in m/z for the folded and unfolded states, indicating that guanidine hydrochloride does not likely bind with the protein itself. To further rule out this possibility, that guanidine hydrochloride is cross-linked (via PFA) to myoglobin, the protein was digested to examine the peptide products.

The protein samples used to generate spectra seen in panels c and d of Figure 3-1 (folded and unfolded myoglobin that was treated with PFA) were treated to enzymatic digestion with Glu-C and their peptide products were compared to those of their respective controls. Glu-C was used as opposed to trypsin because Glu-C cleaves at glutamic acid residues which are not expected to be modified whereas the lysines targeted by trypsin are expected to be *(12-14)*. The mass spectra obtained from the digestion of PFA-treated myoglobin indeed showed mass shifts in multiples of 12 Da from the unmodified peptide, verifying that the mass shift observed in the intact protein analysis was due solely to protein reacting with PFA (data not shown). The average degree of modification for all detected peptides were calculated (Table 3-1) and further illustrate the difference in reactivity between the folded and unfolded states of the protein.

Table 3-1. Comparison of the degree of modification (D.O.M.) of the peptides obtained from the Glu-C digestion of PFA-treated (333 mM, 20 min) folded and guanidine hydrochloride denatured myoglobin.

Peptide m/z	Position	Folded D.O.M	Denatured D.O.M.	Number of Lysines
1282	28-38	0.13	0	0
1625	28-41	0.28	0	0
2044	1-18	0.15	0.78	1
3277	106-136	0	1.78	2
1468	137-148	0.46	1.85	2
1421	42-52	0.22	3.35	4
1621	42-54	0.16	3.26	4
2316	86-105	0.10	3.36	4
2580	60-83	0.05	4.25	5
2780	60-85	0.05	4.33	5

Generally, a higher degree of modification was observed for the peptides obtained from the digestion of the unfolded protein. This trend agrees with the higher mass shift observed in analysis of their respective intact proteins as seen in Figure 3-1 c and d. A strong correlation is evident between the degree of modification of the peptides from the denatured PFA-treated myoglobin and the number of lysines in each peptide sequence. Indeed these findings are consistent with previous work done with model peptides *(14)* where it was found that under the short incubation times used (10-20 min), the major reactive residues are the N-terminus, lysine and tryptophan. Since a protein only has one N-terminus and tryptophan is not a prevalent amino acid in proteins and is hydrophobic, lysine would likely account for the vast majority of reactive residues on proteins. MS/MS studies on these peptides indeed showed lysines to be the reactive residues.

To examine what sites were modified on intact proteins, MS/MS spectra were obtained for each modified peptide from myoglobin. Figure 3-2a shows an example of an MS/MS spectrum obtained from the fragmentation of the quadruply modified species of the

peptide with sequence 42-52, with m/z 1421, found in the peptide digest of the PFA-treated, unfolded myoglobin. There are several modified forms of many different fragment ions, making the spectrum relatively complex. As in a previous study *(14)* MS/MS spectra were interpreted and collapsed into bar graphs plots (Figure 3-2b). Figure 3-2b shows complementary data from the b-ion series (top half of the graph) and y-ion series (bottom half of the graph) from the MS/MS spectrum above, with peptide sequence on the x-axis and degree of modification on the y-axis. The D.O.M in this case measures the degree of modification at each individual peptide fragment. As indicated on the b-ion series (top) portion of the graph, proceeding from the N-terminus to C-terminus in the peptide sequence, it was determined that the D.O.M. of b-ions jumps incrementally at each lysine residue, indicating that the lysines in this peptide are the major reactive sites. Similarly, the y-ion series (bottom), despite lacking some ions from the data set entirely, show marked increases in the D.O.M. of the fragment ions for each successive lysine in the sequence. The slight fluctuation in D.O.M for fragment ions between lysine residues is due to variability in the method.

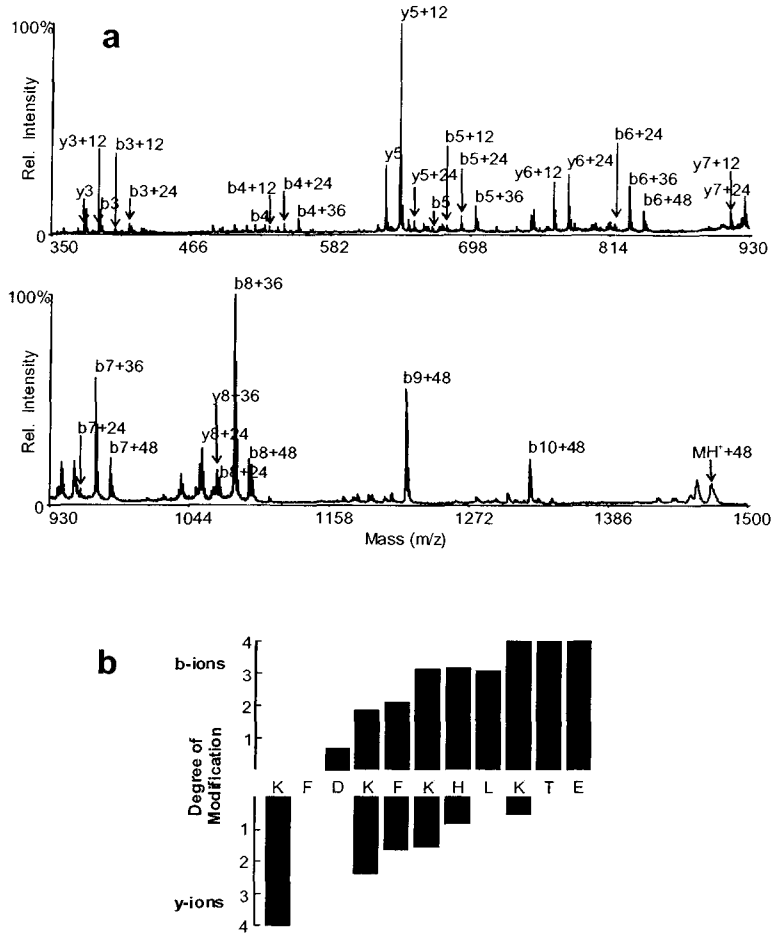


Figure 3-2. a) MALDI tandem mass spectrum of quadruply modified peptide KFDKFKHLKTE, with [M+H]+ = 1421 Th. b) Degree of formaldehyde-induced modification for each b-ion (top) and y-ion (bottom) from quadruply modified peptide KFDKFKHLKTE.

It is interesting that, for peptides with sequences 28-38 and 28-41 (Table 3-1), the reverse trend was observed and a slightly higher degree of modification was observed for the folded than for the unfolded protein. This reverse trend may be explained by the sparse reactive residues in these peptides. Peptides 28-38 and 28-41 have an overlapping peptide sequence (VLIRLFTGHPE and VLIRLFTGHPETLE, respectively), and from

MS/MS analysis of the modified peptide species (singly modified in each case), the reactive residue was identified as the arginine (R) in position 4. Interestingly, it was previously shown that arginine is not considered to be a very reactive residue *(14)*, but can react slowly under certain conditions. It is postulated that arginine can be modified when there is a lack of more prominent reactive residues. However, when the protein is unfolded many reactive lysine residues on other sites become accessible, and are available to be preferentially modified over arginine due to higher reactivity.

Two other model proteins, carbonic anhydrase I and lysozyme, were also tested to determine if their tertiary structures have similar affects on formaldehyde-induced modifications. Table 3-2 summarizes the effects of PFA treatment on these proteins. Similar to myoglobin, CA I showed a significant increase in PFA-induced modifications between the unfolded and folded states (6 M Gdn HCl) *(17)* of the protein (22 ± 1 and 3 ± 1, respectively). Interestingly, lysozyme did not show as large a difference in level of modification between the unfolded (6 M Gdn HCl) *(18)* and folded states (7.2 ± 0.4 and 4.1 ± 0.4, respectively). To determine if the discrepancy of lysozyme's behavior was a result of the 4 disulphide bonds in the protein holding the tertiary structure together and thereby preventing some potential modification sites from becoming accessible, samples were both reduced and denatured prior to PFA treatment (Table 3-2). However, addition of the reduction step in the denaturation process did not increase the level of modification of the unfolded protein. This inferred to us that guanidine hydrochloride alone was sufficient to unfold the protein to an extent that all reactive residues present in the protein were solvent exposed and able to react with PFA. Therefore we speculate that the low

number of modifications on lysozyme might be a reflection of the number of lysine residues present on this protein (six) compared to myoglobin and CA I (19 and 18, respectively), since lysine was previously shown to be the most reactive.

Table 3-2. Comparison of the reactivity with formaldehyde (333 mM, 20 min) for the folded and unfolded states of carbonic anhydrase I (CA I) and lysozyme.

Protein	Denaturant	No. Modifications in the Folded State	No. Modifications in the Unfolded State
Myoglobin	Gdn HCl	1.5	20
CA I	Gdn HCl	3 ± 1	22 ± 1
Lysozyme	Gdn HCl	4.1 ± 0.4	7.2 ± 0.4
Lysozyme	DTT & Gdn HCl	4.4 ± 0.4	6.8 ± 0.3

The concentration of guanidine hydrochloride in the previous experiments was chosen to be at a level where complete unfolding of the protein should occur *(15-18)*, however we also wanted to investigate how a partially unfolded protein reacts with formaldehyde. Guanidine hydrochloride has been known to partially unfold myoglobin in conjunction with its concentration *(15, 16)*, therefore a study was designed to monitor the changes in formaldehyde-induced modifications with varying concentrations of guanidine hydrochloride. Indeed, a positive correlation between concentration of guanidine hydrochloride and the level of modification was observed for the protein up to 3 M Gdn HCl, after which there was no change in the level of modification (Figure 3-3).

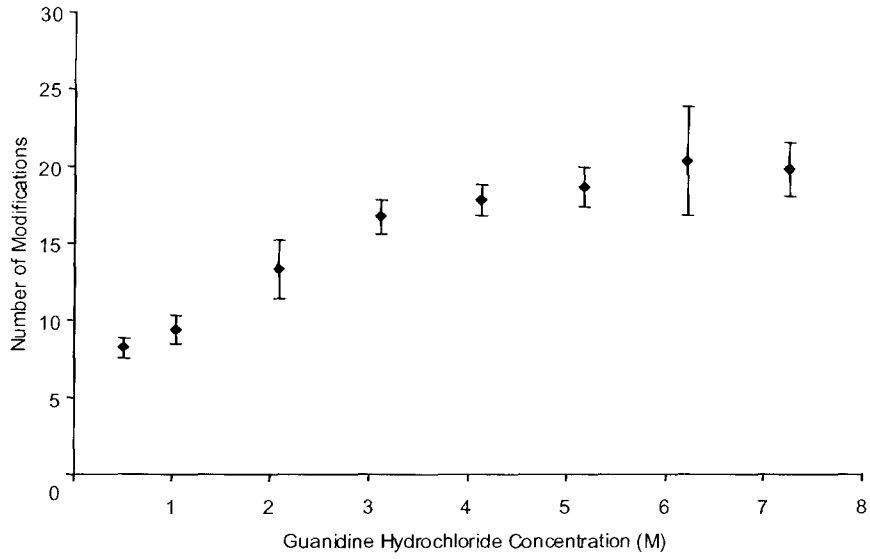


Figure 3-3. The effects of guanidine hydrochloride concentration on the number of formaldehyde-induced modifications on myoglobin.

There are sufficient differences in PFA reactivity between the folded and unfolded versions of a protein, indicating that intact proteins exposed to PFA treatment did not undergo spontaneous unfolding. This type of structural stability during the *in vivo* cross-linking process is significant because major changes in protein tertiary structure likely adversely influence specificity and stability of interactions in a cellular context. Although the tertiary structures remained intact, it was not known if there were any major changes in their stability upon PFA treatment. Since the use of formaldehyde as an *in vivo* cross-linker will modify proteins in their native states, enhancement or degradation of tertiary structural stability of proteins in the presence of PFA was examined. The stability of the tertiary structures can be studied by employing pH as a denaturant since it is known that a low pH can unfold a protein *(19)*. Major differences in stability can

manifest themselves in changes in the critical pH at which a protein unfolds. In an ESI mass spectrum, a folded protein will exhibit a lower average charge state profile than its unfolded form due principally to the accessibility of residues available for protonation. In this way the protein charge state envelope was used to compare the tertiary protein structure of native and PFA treated myoglobin samples at various pH levels.

Both PFA-treated and untreated folded myoglobin were measured by ESI-Q-TOF MS in solutions held at pH 6, 5, 4 and 3 (spectra not shown). The average charge state is plotted versus pH for the PFA-treated and untreated folded myoglobin (Figure 3-4a, ■ and ♦, respectively). The major unfolding event occurred at the same pH for both species (between pH 4 and 3), indicating that PFA does not influence structural integrity under these conditions. This is likely because, despite a large number of modifiable residues in myoglobin, keeping the protein in its folded state allows only a small number of these residues to be modified, and therefore affects on tertiary structural stability are minimized. These results show no large change in tertiary structural stability, suggesting that reacting a folded protein with PFA under the conditions used for *in vivo* cross-linking experiments does not significantly affect the overall structural behavior of a protein in the cell.

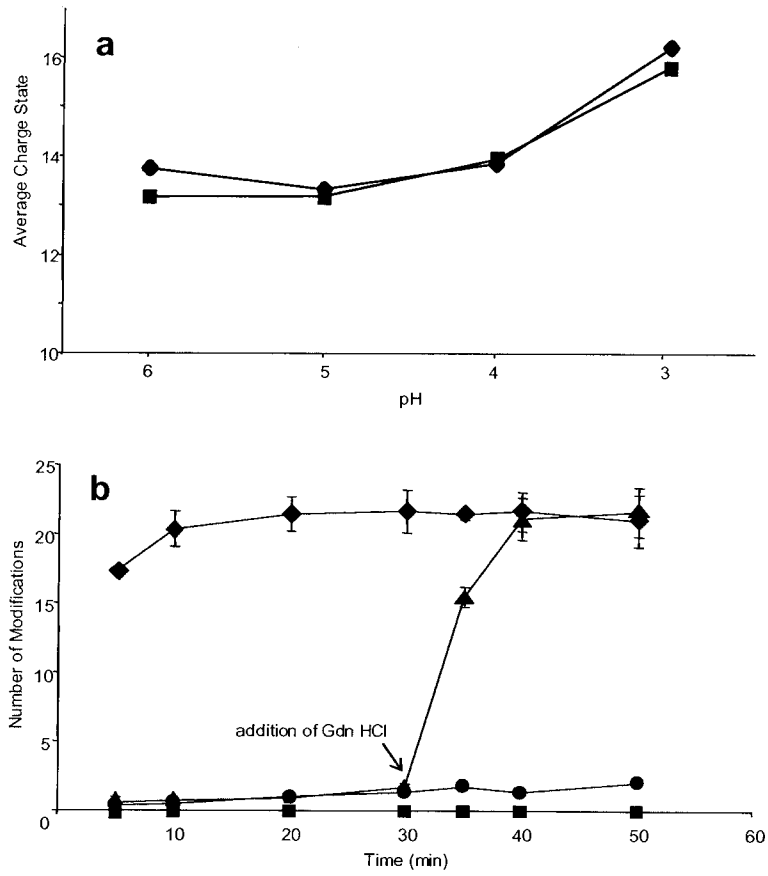


Figure 3-4. a) The average charge state is plotted for untreated (♦) and paraformaldehyde treated (■) myoglobin at pH 3, 4, 5 and 6. b) The number of formaldehyde-induced modifications on myoglobin (■) untreated and during treatments with (●) paraformaldehyde, (♦) paraformaldehyde and guanidine hydrochloride and (▲) paraformaldehyde with guanidine hydrochloride added at t=30 min.

Extended incubation conditions (up to 50 min) showed that myoglobin beginning in its native state showed only a slight increase in the number of modifications, with a maximum of two modifications, indicating no major unfolding occurred (Figure 3-4b, ●). Conversely, when pretreated with guanidine hydrochloride, a high level of modification was rapidly achieved with a negligible increase after t=10 min; indicating that all the reactive sites in the unfolded protein structure were quickly modified (Figure 3-4b, ♦).

The PFA modification on the reactive lysines is fast, specific, does not induce unfolding and is governed by solvent accessibility of reactive residues. In fact, when the folded form of myoglobin is allowed to react for 30 min with excess PFA, only a low level of modification is achieved (~ 2). When a denaturant is added after this point, the protein quickly becomes highly modified due to a sudden increase in residues available for modification (Figure 3-4b, ▲). All samples are compared to the control which was neither treated with guanidine hydrochloride or PFA (Figure 3-4b, ■).

3.4 Conclusions

Solvent accessibility of lysine side chains has a large affect on the number of formaldehyde-induced modifications observed on a protein upon incubation in *in vivo*-like protein cross-linking conditions. Lysine residues have been found to be the most reactive residues under these conditions, with the number of modifications on a protein being a function of the number of lysines in the sequence, and on the proteins state of unfolding. Interestingly we also demonstrate that extensive treatment of proteins with formaldehyde does not unfold a protein, nor does it have a large affect on tertiary structural stability. These findings are of particular significance since they suggest that proteins treated *in vivo* with formaldehyde will not likely exhibit drastic structural and functional change from their native state.

3.5 Acknowledgments

The authors would like to Brent Sutherland for critical reading of the manuscript. This work was funded by an operating grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), and infrastructure grants by the Michael Smith Foundation for Health Research (MSFHR) and the Canada Foundation for Innovation (CFI).

3.6 References

1. Müller, D. R., Schindler, P., Towbin, H., Wirth, U., Voshol, H., Hoving, S., and Steinmetz, M. O. (2001) *Anal Chem* **73**, 1927-34.
2. Lin, H.J., and Kodadek, T. (2005) *J Pept Res* **65**, 221-8.
3. Suchanek, M., Radzikowska, A., and Thiele, C. (2005) *Nature Meth* **2**, 261-67.
4. Seebacher, J., Mallick, P., Zhang, N., Eddes, J.S., Aebersold, R., and Gelb, M.H. (2006) *J Proteome Res* **5**, 2270-82.
5. Pimenova, T., Nazabal, A., Roschitzki, B., Seebacher, J., Rinner, O., and Zenobi, R. (2007) *J Mass Spectrom* **electronic publication**, ahead of print.
6. Li, H. T., Lin, X. J., Xie, Y. Y., and Hu, H. Y. (2006) *Protein Pept Lett* **13**, 385-90.
7. Layh-Schmitt, G., Podtelejnikov, A., and Mann, M. (2000) *Microbiology* **146**, 741-7.
8. Schmitt-Ulms, G., Hansen, K., Liu, J., Cowdrey, C., Yang, J., DeArmond, S. J., Cohen, F. E., Prusiner, S. B., and Baldwin, M. A. (2004) *Nat Biotechnol* **22**, 724-31.
9. Vasilescu, J., Guo, X., and Kast, J. (2004) *Proteomics* **4**, 3845-54.
10. Guerrero, C., Tagwerker, C., Kaiser, P., and Huang, L. (2006) *Mol Cell Proteomics* **5**, 366-78.
11. Tagwerker, C., Flick, K., Cui, M., Guerrero, C., Dou, Y., Auer, B., Baldi, P., Huang, L., and Kaiser, P. (2006) *Mol Cell Proteomics* **5**, 737-48.

12. Metz, B., Kersten, G. F., Hoogerhout, P., Brugghe, H. F., Timmermans, H. A., de Jong, A., Meiring, H., ten Hove, J., Hennink, W. E., Crommelin, D. J., and Jiskoot, W. (2004) *J Biol Chem* **279**, 6235-43.

13. Metz, B., Kersten, G. F., Baart, G. J., de Jong, A., Meiring, H., ten Hove, J., van Steenbergen, M. J., Hennink, W. E., Crommelin, D. J., and Jiskoot, W. (2006) *Bioconjug Chem* **17**, 815-22.

14. Toews, J., and Kast, J. (2007) *Anal Chim Acta* **to be submitted**.

15. Ahmad, F. (1985) *J Biol Chem* **260**, 10458-60.

16. Dai, S. Y., and Fitzgerald, M. C. (2006) *Biochemistry* **45**, 12890-97.

17. Kern, G., Kern, D., Schmid, F. X., and Fisher, G. (1995) *J Biol Chem* **270**, 740-5.

18. Vernaglia, B. A., Huang, J., and Clark, E. D. (2004) *Biomacromolecules* **5**, 1362-70.

19. Shosheva, A., Miteva, M., Christova, P., and Atanasov, B. (2003) *Eur Biophys J* **31**, 617-25.

4 CONCLUDING CHAPTER

4.1 Discussion and Conclusions

It has been shown under incubation conditions similar to those used in *in vivo* cross-linking studies (10-20 min incubations) that the amino termini, and side chains of Lys and Trp residues were reactive with formaldehyde. These residues are a subset of those found reactive (amino termini, Lys, Trp, Arg, His, Cys, Tyr and Phe) *(1, 2)* in studies conducted under an extended incubation period of 2 days. In either study, minimal or no cross-linking was observed between peptides. The lack of appearance of cross-linked peptides was attributed to either a lack of affinity between peptides and/or the diffusion rates of biomolecules *in vitro*. This suggests that for *in vivo* cross-linking studies, cell lysis may be sufficient to quench the cross-linking reaction by diluting the reaction partners.

It was also observed that the location of reactive residues within a peptide influences the overall peptide reactivity with formaldehyde. This was further pursued in a structurally more complex system, model proteins. It has been well established that protein tertiary structure affects the solvent accessibility and therefore reactivity of individual reactive residues in protein-chemical reactions *(3-5)*. The reaction between model proteins and formaldehyde was found to follow the same trend, with the extent of formaldehyde-induced modifications correlated to the extent of protein unfolding. Mass spectrometric analysis of peptides obtained from the digested protein showed that Lys was the predominant reactive residue. This result matches with findings in studies with model peptides performed under similar experimental conditions.

In *in vivo* formaldehyde cross-linking, interacting proteins are cross-linked in their natural environment and conformation. An underlining assumption in these studies has been that formaldehyde treatment does not affect the tertiary structure of proteins since changes in protein structure would change the specificity and stability of such interactions *(6-10)*. The stability of protein tertiary structure with formaldehyde treatment was therefore tested in myoglobin by extending the incubation time with formaldehyde up to 50 minutes rather than the 10-20 minutes used *in vivo (7-10)*. No increase in the number of formaldehyde-induced modifications was observed on the protein with the increased incubation time suggesting that under *in vivo* formaldehyde cross-linking conditions, the tertiary structure of a protein remains intact.

4.2 Future Directions

With a better understanding of the chemistry in the first step of formaldehyde cross-linking, the next goal should be to study, under similar conditions, the chemistry of the second step. To facilitate that, a sufficient amount of cross-linked peptides need to be generated and identified in a sample containing a complex mixture of peptides. To generate a sufficient amount of cross-linked peptides, *in vitro* systems of well known interacting proteins, e.g. pancreatic trypsin inhibitor and trypsin, should be used. The experiments can be conducted quickly, and produce a sample of controlled complexity. To search for these cross-linked peptides, analytical tools are needed to highlight them in a mass spectrum. A promising approach is to combine isotopically labeled formaldehyde *(2)* and computational approaches that can recognize mass spacing between isotopic pairs

(11). Possible candidates of cross-linked peptides can be further analyzed by MS/MS to confirm their identity, and to locate the reactive sites involved in the second step of formaldehyde cross-linking reaction. Once the details of the cross-linking reaction are established in *in vitro* model studies, the information can be used to aid more targeted analysis of the products of *in vivo* cross-linking experiments.

4.3 References

1. Metz, B., Kersten, G. F., Hoogerhout, P., Brugghe, H. F., Timmermans, H. A., de Jong, A., Meiring, H., ten Hove, J., Hennink, W. E., Crommelin, D. J., and Jiskoot, W. (2004) *J Biol Chem* **279**, 6235-43.
2. Metz, B., Kersten, G. F., Baart, G. J., de Jong, A., Meiring, H., ten Hove, J., van Steenbergen, M. J., Hennink, W. E., Crommelin, D. J., and Jiskoot, W. (2006) *Bioconjug Chem* **17**, 815-22.
3. Sharp, J. S., Guo, J. T., Uchiki, T., Xu, Y., Dealwis, C., and Hettich, R. L. (2005) *Anal Biochem* **340**, 201-12.
4. Truhlar, S. M., Croy, C. H., Torpey, J. W., Koeppe, J. R., and Komives, E. A. (2006) *J Am Soc Mass Spectrom* **17**, 1490-7.
5. Ladner, C. L., Turner, R. J., and Edwards, R. A. (2007) *Protein Sci* **16**, 1204-13.
6. Layh-Schmitt, G., Podtelejnikov, A., and Mann, M. (2000) *Microbiology* **146**, 741-7.
7. Schmitt-Ulms, G., Hansen, K., Liu, J., Cowdrey, C., Yang, J., DeArmond, S. J., Cohen, F. E., Prusiner, S. B., and Baldwin, M. A. (2004) *Nat Biotechnol* **22**, 724-31.
8. Vasilescu, J., Guo, X., and Kast, J. (2004) *Proteomics* **4**, 3845-54.
9. Guerrero, C., Tagwerker, C., Kaiser, P., and Huang, L. (2006) *Mol Cell Proteomics* **5**, 366-78.
10. Hájek, P., Chomyn, A., and Attardi, G. (2007) *J Biol Chem* **282**, 5670-81.
11. Sniatynski, M. J., Rogalski, J. C., Hoffman, M. D., and Kast, J. (2006) *Anal Chem* **78**, 2600-7.