Implementation of chemical protein-nucleic acid cross-linking into mass spectrometric workflows and mass spectrometric database searches

Dissertation
for the award of the degree
"Doctor of Philosophy"
Division of Mathematics and Natural Sciences

of the Georg-August-Universität Göttingen

within the doctoral program *Biomolecules: Structure - Function - Dynamics*
of the Georg-August University School of Science (GAUSS)

submitted by
Fanni Laura Bazsó

from Nagykanizsa, Hungary

Göttingen, 2021

Thesis Committee

Prof. Dr. Henning Urlaub, Bioanalytical Mass Spectrometry Group, Max-Planck-Institute for Biophysical Chemistry and Clinical Chemistry, University Medical Center, Göttingen

Prof. Dr. Vlad Pena, Mechanisms and Regulation of pre-mRNA Splicing, The Institute of Cancer Research (ICR), London

Dr. Juliane Liepe, Quantitative and Systems Biology group, Max-Planck-Institute for Biophysical Chemistry, Göttingen

Members of the Examination Board

Referee: Prof. Dr. Henning Urlaub, Bioanalytical Mass Spectrometry Group, Max-Planck-Institute for Biophysical Chemistry and Clinical Chemistry, University Medical Center Göttingen

2nd Referee: Prof. Dr. Vlad Pena, Mechanisms and Regulation of pre-mRNA Splicing, The Institute of Cancer Research (ICR), London

Further members of the Examination Board

Dr. Juliane Liepe, Quantitative and Systems Biology group, Max-Planck-Institute for Biophysical Chemistry, Göttingen

Alexis C. Faesen, Ph.D., Biochemistry of Signal Dynamics, Max-Planck-Institute for Biophysical Chemistry, Göttingen

Prof. Dr. Argyris Papantonis, Translational Epigenetics Laboratory, Institute of Pathology, University Medical Center, Göttingen

Prof. Dr. Jörg Stülke, Department of General Microbiology Institute for Microbiology and Genetics Georg-August-Universität, Göttingen

Date of oral examination: 27. 10. 2021

“No man ever steps in the same river twice,
for it’s not the same river and he’s not the same man”
Heraclitus

List of Abbreviations

Abbreviation	Full name
A	Adenosine monophosphate
aa	Amino acid
ACN	Acetonitrile
AGC	Automatic Gain Control
bp	Base pair
BRP	Basic reversed phase
C	Cytidine monophosphate
ChIP	Chromatin Immunoprecipitation
CID	Collision-induced dissociation
CSM	Crosslinked spectrum match
dA	Deoxyadenosine monophosphate
dC	Deoxycytidine monophosphate
DEB	1,2:3,4-diepoxybutane
dG	Deoxyguanosine monophosphate
DNA	Deoxyribonucleic acid
dsDNA	Double stranded DNA
E. coli	*Escherichia coli*
ECD	Electron capture dissociation
ESI	Electrospray Ionization
ETD	Electron transfer dissociation
EThCD	Electron-Transfer/Higher-Energy Collision Dissociation
FA	Formaldehyde
FDR	False Discovery Rate
G	Guanosine monophosphate
Gb	Guanine
HCD	Higher-energy C-trap dissociation
HPLC	High performance liquid chromatography
IT	Injection time
MS	Mass Spectrometry
ms	millisecond
Nb	Nucleobase
NCE	Normalized Collision Energy
Nt	Nucleotide
PSM	Peptide spectrum match

PTM	Post translational modification
RNA	Ribonucleic acid
Silica	Silicon dioxide
SP3	Single-pot, solid-phase-enhanced sample preparation
T	Thymidine monophosphate
TFA	Trifluoroacetic acid
TiO_2	Titanium-dioxide
U	Uridine monophosphate
XL	Crosslink
XL-MS	Crosslinking Mass Spectrometry

List of Figures

List of Tables

List of Supplementary Tables

Extended Supplementary Tables

List of Supplementary Figures

List of Supplementary Texts

Summary

Protein nucleic acid interactions play a pivotal role in cells, from transcription to translation. Functions of proteins in protein nucleic acid interactions are diverse: histones are responsible for the packaging of the genomic DNA; ribosomal proteins are part of the ribosome, which is responsible for protein synthesis.
Crosslinking mass spectrometry proved to be a useful tool to identify protein-nucleic acid interactions and their dynamics in the cell. A vast amount of effort has been spent to elucidate protein-nucleic acid interactions with UV crosslinking mass spectrometry, whereas chemical crosslinking techniques were left untouched. Therefore, the current study was aiming for the implementation of chemical crosslinking into available mass spectrometric workflows. Two chemical crosslinkers were investigated with crosslinking mass spectrometry for the elucidation of protein-nucleic acid complexes: formaldehyde and 1,2:3,4-diepoxybutane (DEB).
In the case of DEB-crosslinking, crosslinks to all four nucleobases were observed. In the case of formaldehyde crosslinking, crosslinks to three nucleobases were observed (adenine, guanine, and cytosine). Fragmentation behavior of the crosslink induced peptide-deoxynucleotide conjugates were investigated with both chemicals. Specific fragmentation of the DEB linker was observed and later used to establish a dedicated mass spectrometric measuring method, called MS2/MS3 method. Based on the formaldehyde linker fragmentation, predominantly leading to complete cleavage, a crosslinker specific mass spectrometric method was built as well, named as MS2/MS2 method.
DEB-crosslinking was applied in the study of DNA-containing complexes such as linker histone double stranded DNA complexes and nucleosome-based complexes. DEB crosslinking could provide information about the crosslinking site localization in the protein sequences, often to single amino acid resolution.
Formaldehyde-crosslinking was applied on *in vitro* reconstituted and native nucleosome complexes, the bacterial 70S ribosome and *in vivo* crosslinked *Escherichia coli* and HeLa cells.
The simple fragmentation pattern of formaldehyde-crosslinking and the developed sequential RNA digestion technique made it possible to identify hundreds of proteins crosslinked to RNA *in vivo*. The established MS based crosslinking strategy can give an extensive picture of the RNA interactome in well studied and less investigated organisms.

1 Introduction

1.1 Mass spectrometry

In mass spectrometry ions are measured based on their mass-to-charge ratio (*m/z)*. During a mass spectrometric measurement, molecules are ionized in the ion source, the resulting ions are focused and moved in the electrostatic field, *m/z* values and relative number of ions are measured. Thus, all mass spectrometers need to have an ion source, an analyzer, and a detector. Mass spectrometers can be divided into categories based on their mass analyzers. Quadruple, linear ion trap and orbitrap analyzers are described in detail; these analyzers were used for mass spectrometric measurements in this thesis.

1.1.1 Ionization techniques

Electrospray ionization (ESI) is one of the most widely used ionization technique, allowing the analysis of biomolecules such as peptides or proteins. Another reason for the popularity of the electrospray ionization, is that the ionization technique can be easily connected with liquid chromatograph: chromatography then can be used for separation of the biomolecules.
During electrospray ionization, compounds are dissolved in a solvent and the solvent is pressed through a capillary with high pressure, resulting aerosol formation. [1]. The capillary is placed opposite to a high voltage field, meanwhile, solvent evaporates causing charges to migrate to the surface of the droplets. When the Coulomb repulsion of same charges and the counteracting surface tension are in the same order of magnitude, a so called ‘Coulomb explosion’ happens. The droplet became instable, and the Coulomb explosion tears the droplet into smaller charged droplets. This mechanism continues until the size of the droplets reach a critical minimal size, enough for the electric field to desorb them, resulting ‘quasi molecular’ ions.

1.1.2 Mass analyzers

1.1.2.1 Quadrupole analyzer

A quadrupole analyzer consists of four cylindrical electrodes, which are positioned parallelly, in an imaginary cuboid longer sides. The opposite electrodes are connected electronically, and AC and DC potential is applied on them [2] Thus, two opposite rods have the same electric potential, rods next to each other have opposite electric potentials. When the ions are sent into the quadrupole analyzer, depending on their *m/z* values and application of corresponding electrical potential, they will move on a defined trajectory. With the constant change of the amplitude and the frequency of the field, ions only with a certain *m/z* can be selected and travel through the electrostatic field, the rest doesn’t have stable trajectory and hit the electrodes. Quadrupole analyzer is often called mass filter because it can filter out all ions except for the one with a stable trajectory.

1.1.2.2 Ion trap analyzers

1.1.2.2.1 Quadrupole ion trap analyzer

Quadrupole ion trap analyzer (3D ion trap) contains three, hyperbolic shaped electrodes, a ring electrode and two end cap electrodes [3]. Similarly, to quadrupole analyzer, AC and DC potential are applied on the electrodes, creating an alternating electrostatic field. In the analyzer ions, depending on their *m/z* values, are trapped and moved in the electric field. Their trajectory is shaped on the basis of the AC potential applied. The quadrupole ion trap has an alternative configuration - the linear ion trap [4]. It has a high ion storage capacity and high scan rate.

1.1.2.2.2 Orbitrap analyzer

Orbitrap analyzer is an ion trap analyzer, with a specific configuration. Quadrupole ion trap has low resolution and a so-called low mass cutoff: fragmented low *m/z* ions cannot be trapped. Orbitrap overcomes on these limitations - it can achieve high resolution and has no low *m/z* cutoff [5]. Orbitrap analyzers have two, differently shaped electrodes. The inner electrode is spindle shaped and the outer electrode is barrel shaped [6].

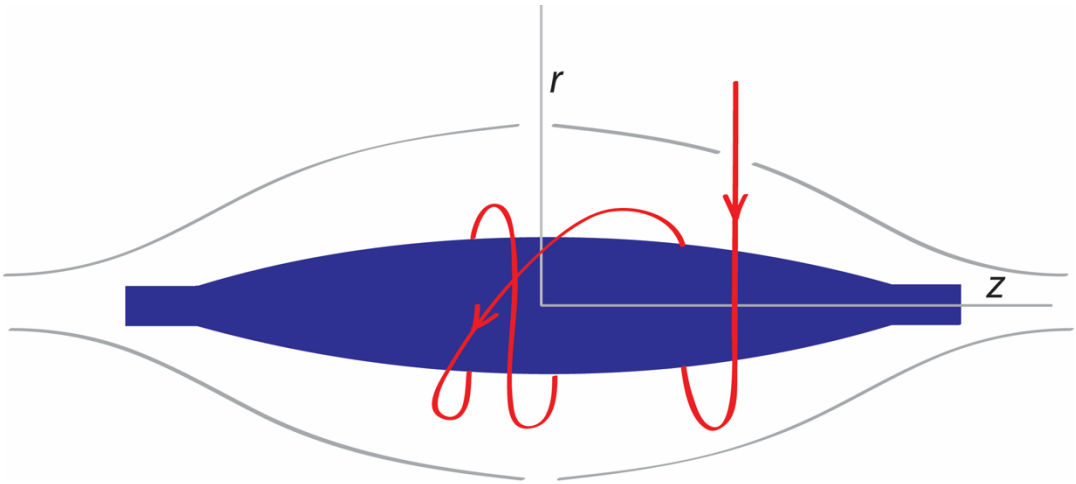


Figure 1.1-1: Schematic picture of the Orbitrap mass analyzer

Ions are injected in a tangential manner into the mass analyzer (Figure 1.1-1), where they are electrostatically trapped; the ions cycle around the inner electrode with an elliptical trajectory, moving back and forth between the two ends of the electrode. The ions' momentum creates an image current measured on the barrel shaped electrode. The signal is Fourier transformed and converted into a mass spectrum [7]. Orbitraps function as both a mass analyzer and a detector.

1.1.3 Mass spectrometers containing several analyzers

Mass spectrometers with single analyzer can be employed as stand-alone devices that can carry out a whole mass spectrometric experiment, for instance, ion trap analyzer can be used for recording survey scans and MS/MS scans in the same space. However, nowadays mass spectrometers often consist of more than one mass analyzers, combining their specific strengths. These mass spectrometers are called hybrid mass spectrometers. Typical hybrid mass spectrometers are the Q-Exactive instruments, where two mass analyzers are connected:

quadrupole mass analyzer and Orbitrap mass analyzer. Moreover, in other setups three mass analyzers can be used, for example in the Oribitrap Fuison (Lumos) instruments. The instrument consists of a quadruple analyzer which is used as the previous described mass filter. An ultra-high field Orbitrap mass analyzer, which allows the recording of high-resolution mass spectra and a high sensitivity, dual pressure, quadrupole linear ion trap analyzer, that allows successive fragmentation events.

1.2 High performance liquid chromatography

High performance liquid chromatography (HPLC) is a very widely used technique for the separation of compounds, based on their physicochemical properties [8].
In the normal phase chromatography, the solid phase is highly polar, and the liquid phase is non-polar, typically an organic solvent. When the polarities of the liquid and solid phase are inverted, the method is named reversed-phase (RP) chromatography. In RP-HPLC the stationary phase (also called column) is non-polar, and the liquid phase (also called mobile phase) is moderately polar, typically aqueous based. The most widely used stationary phase in peptide RP-HPLC is octadecyl carbon chain (C18) functionalized silica and the most commonly used mobile phase is water or water acetonitrile (ACN) mixture. During RP-HPLC, hydrophilic molecules elute from the column earlier, than the hydrophobic ones. When gradient elution is used, hydrophobic molecules elute from the column with the increasing concentration of organic solvents. RP-HPLC is often connected with mass spectrometry for the separation of biomolecules and improving the analysis depth of complex mixtures.

1.3 Mass spectrometry-based proteomics

1.3.1 Tandem mass spectrometry

During tandem mass spectrometry an analyte is subjected to fragmentation for acquiring structural information. Tandem mass spectrometry is a two-step process, where first, a mass spectrum is recorded. Then, an ion with a defined *m/z* value, often called precursor ion, is isolated and fragmented. During fragmentation the covalent bonds of the precursor ions are cleaved. Bond cleavage can be achieved in several ways, for instance by collision with an inert gas. Collision-induced dissociation (CID) [9] and higher energy c-trap dissociation (HCD) [10] are typically using collision for the fragmentation of the precursor ions. Bond cleave can be induced by other ways with electron capture dissociation (ECD) [11] or electron transfer dissociation (ETD) [12]. Moreover, fragmentation methods can be combined, such as Electron-Transfer/Higher-Energy Collision Dissociation (EThcD). The common feature of all fragmentation methods is that they produce fragments of the precursor ion that can provide meaningful sequencing or structural information about the biomolecule that is analyzed. These fragment ions can differ greatly based on the chosen fragmentation method.

1.3.1.1 Peptide fragmentation

Mass Spectrometry is often the method of choice for peptide analysis [13]. Peptides are often multiply charged during mass spectrometric measurements in positive ion mode. Typically, peptide fragmentation happens on the peptide backbone, possible backbone cleavage sites are

shown in Figure 1.3-1. N-terminal fragments are named as a, b and c ions. C-terminal fragments - as x, y, and z ions. Peptide fragmentation can result a, b, c and complementary x, y, z ions.

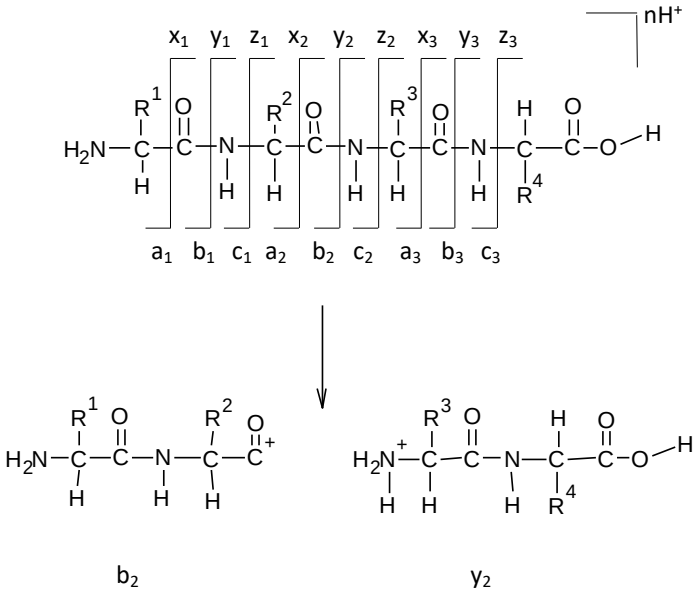


Figure 1.3-1: Possible cleavage sites of peptide fragments and their nomenclature during fragmentation adapted from [14]

One possible explanation of the peptide fragmentation is the mobile proton theory [15]. When peptides are protonated, protons are generally localized on the basic residues - such as lysine or arginine, or on the peptide N-terminal NH_2 group [14]. When a proton is localized on the latter, this proton is more mobile can move more freely on the peptide backbone amid bonds as well., thus localized at different positions. Due to the heterogeneity of peptides with different proton localizations, during fragmentation, different fragments are produced, as proton localized on different amid bonds have a great effect on which bonds are cleaved. The applied fragmentation technique greatly affects which type of peptide fragments are generated. During CID and HCD fragmentations, typically a, b and y ions are formed. Whereas, during ETD and ECD fragmentation typically c and z ions are formed. Combined fragmentation techniques can often give a higher degree of the peptide fragmentation and can offer more complementary peptide fragments.

1.3.1.1.1 Tandem mass spectrum of peptides

During fragmentation, peptide fragment ions are generated, and tandem mass spectrum is recorded. Tandem mass spectrum is the function of *m/z* values of the fragment ions and their respective intensities, intensity values are often normalized. Illustration of tandem mass spectra are shown on Figure 1.3-2.

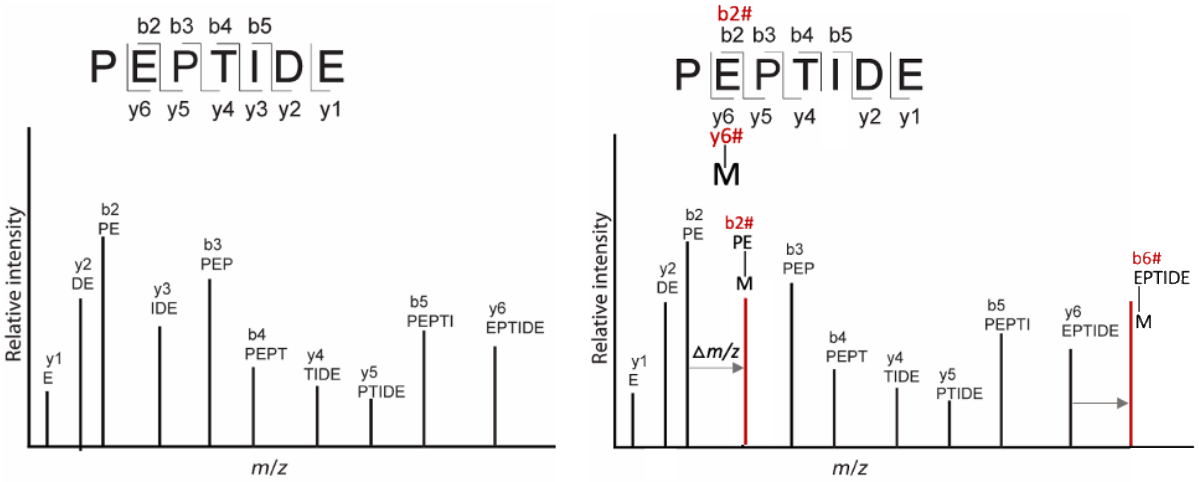


Figure 1.3-2: Illustration of a typical MS/MS spectrum of a peptide

Left panel: Illustration of a tandem mass spectrum of peptide where b and y ion series are present. Right panel: illustration of a tandem mass spectrum of modified peptide where b and y ions are present, as well as M mass shifted b and y ions. M mass shifted b and y ions are highlighted in red.

In the example tandem mass spectrum b and y fragment ions of the peptide are shown. When a peptide carries a modification (chemically induced or post-translational) a mass shift appears on the peptide fragments. In the tandem mass spectrum, the *m/z* values of the peptide fragment ions can be shifted with the *m/z* value of the modification as it is illustrated on Figure 1.3-2, right panel.

1.3.1.2 Protein identification with mass spectrometry

As the need for identification of proteins from a complex mixture has emerged, and peptide identification with mass spectrometry has advanced, mass spectrometry became one of the primary tools used for protein identification [16]. Different approaches are available for studying proteins with mass spectrometry, here the bottom-up approach is discussed. Bottom-up approach deemed most useful for the investigation of complex samples.
In the bottom-up approach the proteins are digested with an endoproteinase to peptides. Trypsin is the most commonly used endoproteinase, due to its amino acid specificity [17] and its stability under different biochemical conditions.
After the enzymatic digestion, the peptides are separated with liquid chromatography and measured with mass spectrometry. For the measurement of the peptides an MS-MS/MS strategy is used. First, intact masses of peptides are measured in a so-called MS (or MS1) run. During an MS run, intact masses of peptides are recorded in a defined *m/z* region. Then, the most abundant peptides from the MS1 scan are chosen for fragmentation, one by one. After the fragmentation of the peptide, *m/z* values of the peptide fragments are recorded in an MS/MS (or MS2) run. The produced, the mass spectrometric data is analyzed with proteomics database searches.

1.3.1.2.1 Peptide identification and validation in proteomics database strategy

The main goal of the database strategy is to identify which proteins were present in the measured sample. In the bottom-up strategy, the protein identification is based on the identification of the measured MS/MS spectra of the peptides. To identify the MS/MS spectra of the peptides, database search strategy used. The database strategy is as follows: the proteins present in the database are in-silico digested into peptides and theoretical peptide sequences are generated. Masses of the theoretical peptides are calculated based on their elemental composition. These theoretical peptide masses are used for comparison to the experimentally measured masses of peptides. The calculated and measured masses are matched within the defined error tolerance, and the candidate peptides' theoretical MS/MS spectra are generated. The measured and theoretical MS/MS spectra are compared, and similarity score is calculated. Different strategies for similarity calculation are used from cross correlation calculation [18] to E-value calculation [19] depending on which database search strategy is used. The database search results a list of proteins, their respective peptides as well as peptide spectrum matches (PSM).

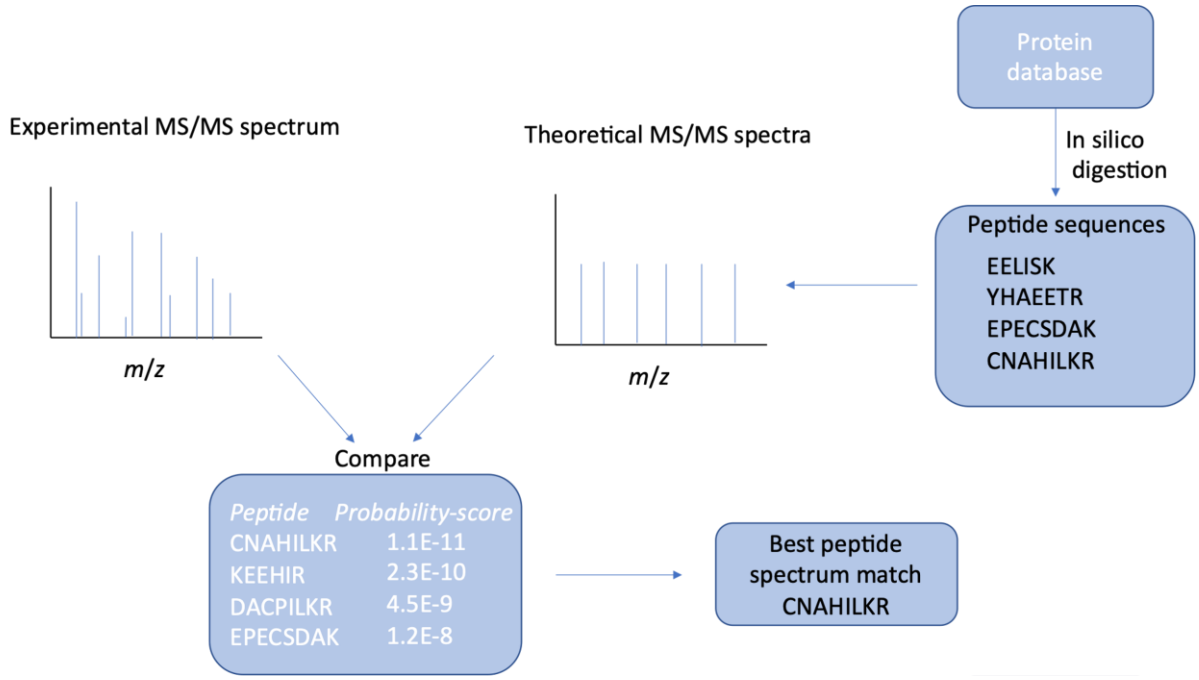


Figure 1.3-3: Illustration of the MS-based proteomics database search

Calculating any type of a similarity score for the best matching theoretical peptide is not enough for the data assignment. There is an overlap by score between correctly assigned peptides (true positive hits) and incorrectly assigned peptides (false positive hits). False positive matches can have as high similarity score as true positive matches. To overcome on this obstacle, statistical validation of the peptide spectrum matches is needed. The most widely used method for multiple testing correction is the so-called false discovery rate (FDR) [20]. One possible FDR calculation is as follows: a decoy database is generated from the targets database (original protein database). The FDR estimation assumes that false positive

hits in both the target and decoy database are equal. Then the FDR to a given score threshold is the following [21]:

$$FDR = \frac{\#decoys}{\#targets}$$

The number of decoys hits divided by the number of target hits.

1.3.1.2.2 Protein identification and validation in proteomics database strategy

In the bottom-up strategy, proteins are enzymatically digested into peptides, peptides are measured with HPLC-MS/MS, the generated peptide sequences are assigned by database search and the peptide identification is validated by spectral FDR. The assigned peptide sequences are mapped against onto the protein sequences in the protein database for protein identification. Protein identification based on peptide identification is a complicated process. Difficulty arises from the fact that, some proteins can be identified with several peptides, whereas others with only few. An even bigger problem is the "uniqueness" of the peptides, as proteins can share common sequences. One way to overcome this obstacle is to calculate protein parsimony [22]. Protein parsimony is used to estimate the needed minimal number of proteins which can describe the identified peptide sequences. Another way to overcome this obstacle is to calculate protein probability with mixture model EM [23]. For the validation of the identified proteins a protein FDR is calculated, similarly as spectral FDR, based on target-decoy approach [24] [25].

1.4 Protein nucleic acid interactions and their possible identification

1.4.1 Protein-RNA interactions

Protein-RNA interactions regulate various cellular processes. Recently mass spectrometry-based studies point towards that approximately 5-10% of the proteome in a cell could be interaction with RNA [26]. Some proteins have one or more RNA-binding domains such as double-stranded RNA-binding motif (DSRM motif), RNA-recognition motif (RRM) or hnRNPK homology domain (KH domain) [27]. Other proteins don't have distinct RNA binding domains, these proteins often have a dual function: they are known for their metabolic functions, but they also function as RNA-binding proteins [28]. Among others these proteins are kinases, heat and cold shock proteins, etc. [29].
There are several techniques to investigate proteome-wide interactions between proteins and RNA. Some techniques approach the interaction from the RNA site, these techniques are RNA-centric techniques. Other methods approach the protein-RNA interactions from the protein site, these are protein-centric techniques. RNA-centric techniques are typically CLIP techniques (crosslinking and immunoprecipitation) where cells are *in vivo* UV crosslinked, proteins, which have been crosslinked to RNA are immunoprecipitated with specific antibody, and RNA is fragmented and converted to cDNA for next generation sequencing analysis [30].
Protein-centric methods often use mass spectrometry as a tool. In the case of proteome-wide studies, proteins are *in vivo* UV crosslinked to RNA, RNA-protein complexes are isolated with solid phase extraction [30] or liquid-liquid extraction [31]–[33] and the isolated proteins are identified with HPLC-MS/MS and database searches.

1.4.2 Protein-DNA interactions

Protein-DNA interactions are critical for several cellular processes; proteins interact with DNA to regulate gene transcription or to organize chromatin. Protein-DNA interactions can be sequence specific, typically transcription factors bind to DNA in a sequence specific manner to DNA in the major groove [34]. Proteins can interact with DNA nonspecifically as well, furthermore it has been shown that proteins which bind in sequence specifical manner to DNA also can nonspecifically bind to DNA as well [35].
In the case of protein-DNA interaction studies, techniques can also be divided into DNA-centric and protein-centric methods as well. In the DNA-centric view, protein-DNA interactions can be for instance investigated with Chromatin Immunoprecipitation (ChIP) [36]. During ChIP, *in vivo* crosslinking is applied with formaldehyde, genomic DNA is fragmented to shorter stretches, approximately 200-1000 bp in size, then protein of interest is pulled down with specific antibody. Because DNA is chemically crosslinked to the proteins, interacting DNA is also isolated with the protein of interest. Crosslinking is heat reversed and DNA is sequenced. Formaldehyde crosslinking is used in another DNA centric method, in FAIR-Seq (Formaldehyde-Assisted Isolation of Regulatory Elements) [37]. In FAIR-Seq after formaldehyde crosslinking in *vivo*, and DNA fragmenting and a phenol-chloroform extraction is used to extract the free, non-crosslinked DNA. The DNA is isolated from the aqueous phase, DNA is fluorescently labeled followed by hybridization to a DNA microarray and Sequencing.
Protein-centric methods are commonly mass spectrometry-based methods. Chromatin enrichment for proteomics (ChEP) [38] is a chromatin isolation technique, where cells are formaldehyde crosslinked *in vivo*, the crosslinked cells are lysed, DNA and DNA-protein complexes are isolated with high-speed centrifugation under denaturing conditions. Formaldehyde crosslinking is heat reversed and the proteins are analyzed with HPLC-MS/MS.

1.5 MS based protein nucleic acid crosslink identification and its challenges

Although, the previously listed MS-based methods used crosslinking for the preservation of the contact sites between proteins and nucleic acids, crosslinking is often reversed, or non-crosslinked peptides are used for protein identifications. These methods give an indirect information about which proteins were in interactions with nucleic acids. Direct proof for the interaction between proteins and nucleic acids can only be acquired by the identification of the contact sites. Crosslinking Mass Spectrometry (XL-MS) is one of the possible methods for contact side identification. Generally used workflow for MS-based crosslinking site identification is shown on Figure 1.5-1.

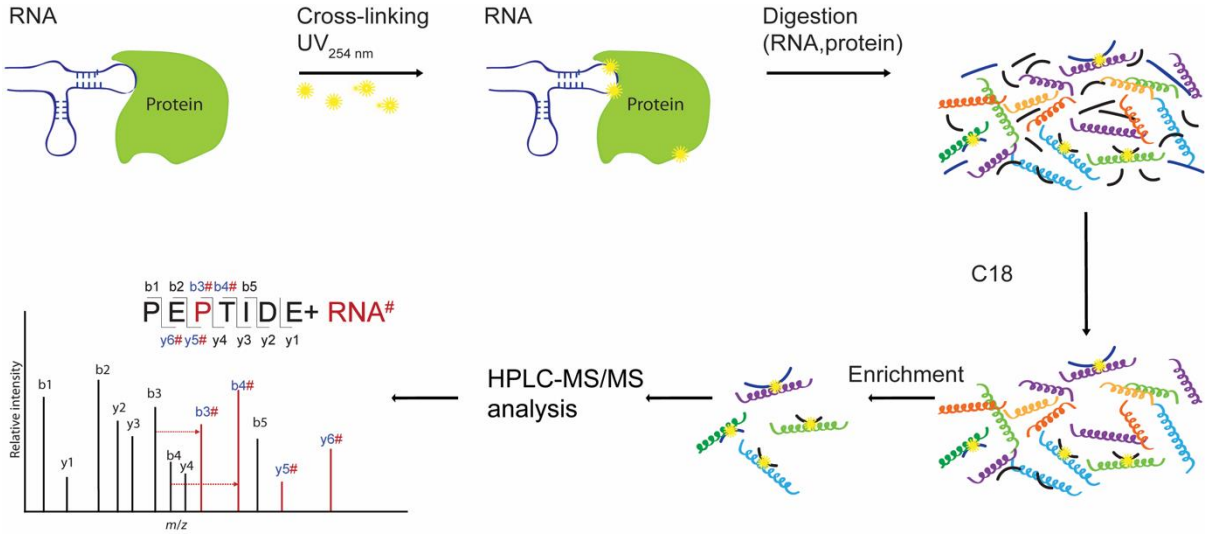


Figure 1.5-1: Schematic figure of crosslinking mass spectrometry-based contact site identification between proteins and nucleic acids

In a typical crosslinking MS experiment [39] protein-nucleic acid complexes are crosslinked, the nucleic acids and proteins are enzymatically digested into oligo(deoxy)nucleotides, peptides, and peptide-nucleic acid heteroconjugates. Free oligonucleotides and salts are removed with C18 reversed-phase chromatography. Selective enrichment strategy is used for the isolation of the peptide-nucleic acid heteroconjugates. The heteroconjugates are then separated by HPLC and mass spectrometric measurement is performed. The measured data is analyzed with specialized database search workflows [40].

The data analysis of peptide-nucleic heteroconjugates pose a challenge due to the complexity of the MS/MS fragmentation of the peptide-nucleic acid heteroconjugates.

As it was shown in the previous chapter, peptides can fragment through the peptide backbone. Depending on the fragmentation method, different peptide fragment ions are generated during fragmentation. During HCD dissociation b and y ions are generated. Nucleic acids also fragment during HCD, typically, the N-glycosylic bonds and the sugar-phosphate bonds cleave. When peptides and deoxynucleotides are covalently attached, both compounds fragment. In a peptide-centric view, the attached deoxynucleotide fragments are called mass shifts, referring to that the peptide fragment masses are shifted with the deoxynucleotide fragment masses. This fragmentation is more complex than the MS/MS fragmentation of a modified peptide (Figure 1.5-2 left panel). When a modified peptide is fragmented a fix mass shift appears on the peptide fragments.

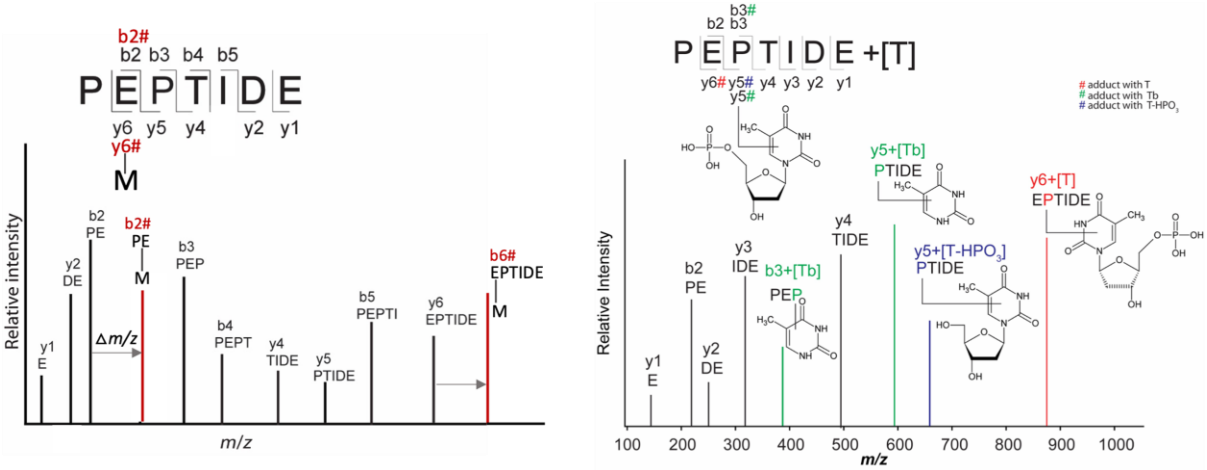


Figure 1.5-2: Illustration of a typical MS2 spectra of modified peptides and peptide-DNA heteroconjugates

Left panel: Illustration of a typical MS2 spectrum of a modified peptide. Right panel: illustration of a typical MS2 spectrum of a UV crosslinked peptide-thymidine monophosphate heteroconjugate

As it is shown on Figure 1.5-2, when a peptide is covalently attached to a nucleic acid - in this case to DNA- the peptide MS/MS fragments are shifted with several, different masses which are the masses of the DNA fragments. For instance, in the case of UV crosslinked thymidine monophosphate-peptide heteroconjugates, three different mass shifts appear on the peptide MS/MS fragments: the mass shift with the thymine base, Tb (Figure 1.5-2 right panel, peaks highlighted in green), the mass shift with thymidine, T-HPO_3 (Figure 1.5-2, right panel, peak highlighted in blue) and the mass shift of the thymidine monophosphate itself, T (Figure 1.5-2 right panel, peak highlighted in red). When UV crosslinking is used, the number of the mass shifted peaks are prominent in an MS/MS spectrum. Moreover, the higher the number of possible mass shifts are the more combinations are possible, making it extremely challenging to correctly identify the mass shifted peptide peaks. This phenomenon becomes more important in complex samples, where thousands of proteins are present, typically, in an *in vivo* crosslinking experiment.

1.5.1 Data analysis possibilities of the peptide-nucleic acid heteroconjugates

1.5.1.1 RNP^{xl}

As it was mentioned previously data analysis of peptide-nucleic acid heteroconjugates is complex and bioinformatically challenging task. Only one bioinformatics search engine is available for the analysis of peptide nucleic acid heteroconjugates, RNP^{xl} [40] which was developed in the framework of the OpenMS project [41]. For the analysis of peptide-nucleic acid heteroconjugates, a bioinformatics pipeline was established. The pipeline consists of different bioinformatic tools which are used for the reduction of the possible crosslinked candidates. Then the RNP^{xl} search can be used for the analysis of the peptide-nucleic acid heteroconjugates. RNP^{xl} search can consider all previously defined mass shifts on the peptide fragments, thus it allows for correct crosslink identification. However, additional manual validation of the crosslinked MS/MS spectra is needed to avoid false positives, which is laborious and time consuming.

1.5.1.2 Other approaches for the identification of peptide-nucleic acid heteroconjugates

To avoid manual validation of the peptide-nucleic acid heteroconjugates, a so-called open search strategy can be used. In an open search the precursor mass tolerance is opened with hundreds of Daltons to detect known or unknow modifications of the peptides. For instance, +/-500 Da mass tolerance is used for open searches, instead of the +/-10 ppm, which is typically used for narrow searches. In the open search strategy, for MS2 mass tolerance is set to the same value as in the narrow searches (typically 20 ppm). Thus, in open search approaches the peptide identification is based on the MS/MS fragmentation where high mass accuracy is used. The mass difference between the theoretical peptide mass and the measured precursor mass is the delta mass. Delta masses are then used for the identifications of the modification on the peptides. When open search is used for crosslinking identification, delta masses are the RNA adduct masses. Open search strategy was used by Peil *et al.* [42] for the identification of UV crosslinked protein-RNA heteroconjugates. Open searches cannot consider the mass shifts on the peptide mass fragments; thus, open searches can only confirm the crosslink identification on the MS1 level. Therefore, open search strategies as not widely used for UV crosslink identification.
Another approach was used by Panhale *et. al* [43], where RNA adducts were treated as post translational modifications (PTMs), they have called it RNA-PTMs. In their approach they have reduced the number of modifications to 34 when crosslinked RNA length is maximum 3 nucleotide long. They have used the RNA-PTMs setup together with a de novo peptide sequencing software, PEAKS [44]

1.6 Protein-nucleic acid chemical crosslinking

1.6.1 Protein-DNA crosslinking with 1,2:3,4-diepoxybutane

There are several chemicals which can be used for protein-DNA crosslinking, for instance aldehydes, transition metals and anti-cancer agents [45]. One possible crosslinker for protein-DNA crosslinking is 1,2:3,4-diepoxybutane (DEB). DEB is a bis-electrophile, a symmetric epoxy, with high reactivity and membrane permeable, introducing crosslink between the two strands of DNA [46].

1.6.1.1 DEB crosslinking reaction scheme

DEB induces crosslinks between proteins and DNA, more over DEB induces crosslinks between proteins too. DEB has two epoxy groups, both can react with nucleophiles, such as the NH_2 groups of proteins or respective groups of nucleic acids. The reaction mechanism between a guanine and an NH_2 group of a lysine molecules is shown in Figure 1.6-1.

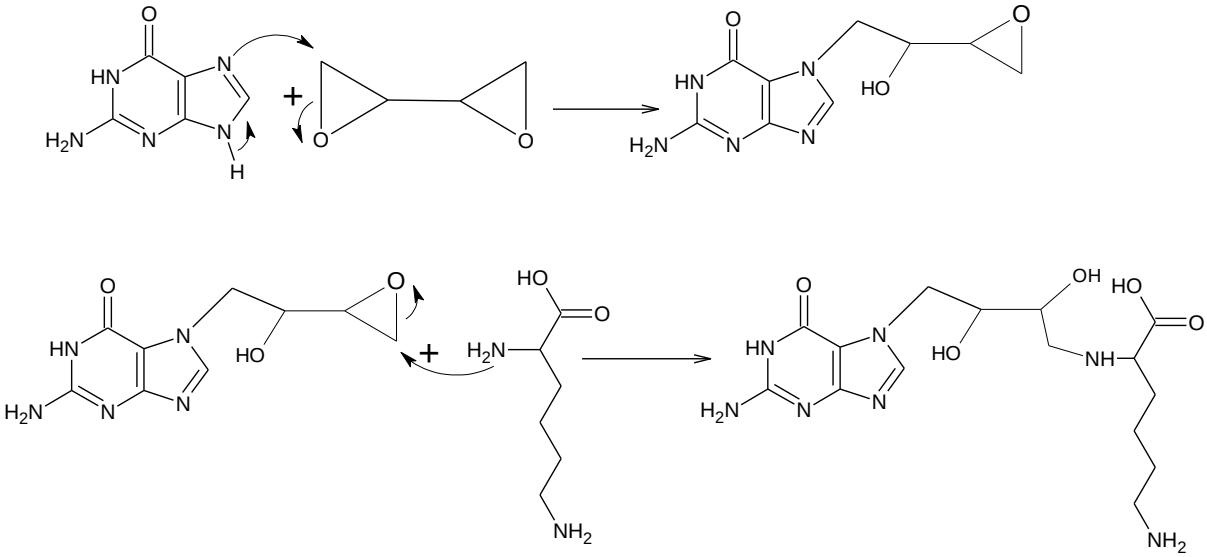


Figure 1.6-1: Proposed crosslinking reaction scheme between guanine and lysine with 1,2:3,4-diepoxybutane, adapted from [47]

In the first step one epoxy group opens and reacts with the guanine at the N7 position forming a 2-hydroxyl-3,4-epoxybut-1-yl intermediate, which through the other ring opening reacts with the lysine NH_2 group resulting a crosslink between the DNA and the protein [48]. In principle all nucleobases can react with DEB, not only guanine. Although, those reaction mechanisms are not investigated in a cellular context, adenine, cytosine, and thymine DEB alkylation have been described in [49]. In the case of adenine, Tretyakova *et al.* [50] have shown when deoxyadenosine and calf thymus DNA were treated with DEB, after acidic hydrolysis, a single product, N6-(2,3,4-trihydroxybut-1-yl)adenine have been detected (see numbering on Figure 1.6-2). Thus, the reaction took place on the 6-position in the deoxyadenosine. In the case of thymine and cytosine, the nitrogen in the N3 position has the highest likelihood to react with DEB [47]. Since this thesis is not aiming for the exact position identification, and crosslinking mass spectrometry is not suitable for structural investigation in such detail, further on, no crosslink position localization will be shown in the case of cytosine, adenine, and thymine.

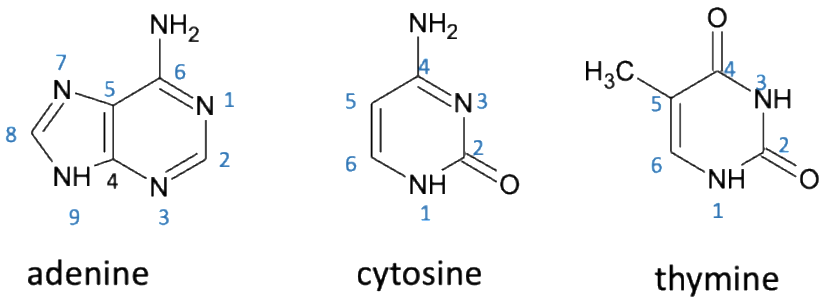


Figure 1.6-2: Atom numbering of adenine cytosine and thymine bases

1.6.1.2 Protein nucleic acid crosslinking with formaldehyde

Formaldehyde is a widely used crosslinking agent [51], it is used in several biochemical techniques such as tissue fixation [52], with the polymerized form of formaldehyde, paraformaldehyde it is also used for protein-DNA crosslinking in Chromatin immunoprecipitation [53]. The crosslink, formed by formaldehyde between proteins and nucleic acids is heat reversable [54]. This feature is often used in chromatin immunoprecipitation when proteins are released from their crosslink with DNA on a higher temperature and DNA is sequenced.

1.6.1.2.1 Formaldehyde crosslinking reaction scheme

Formaldehyde is known to form a methylene bridge between the close proximate proteins or protein nucleic acids **[51]**. The reaction mechanism is illustrated in Figure 1.6-3, the reaction is shown between lysine and guanine. In the first step formaldehyde reacts with one of the nucleophilic groups of lysine, forming a methylol group. This methylol group, through a water loss, transforms into a Schiff base (Figure 1.6-3 step 1). In the second step the Schiff base reacts with another nucleophilic group, forming the methylene bridge between the lysine and the guanine (Figure 1.6-3, step 2).

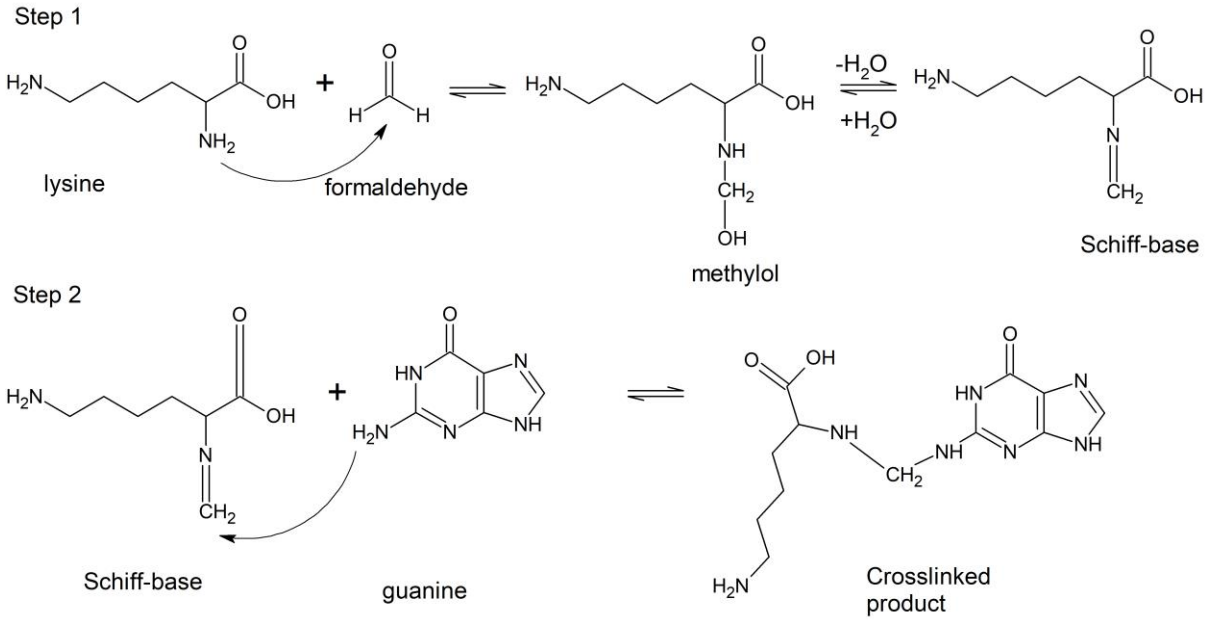


Figure 1.6-3: Proposed crosslinking reaction scheme of formaldehyde crosslinking between lysine and guanine adapted from [51]

The crosslinking reaction is an equilibrium reaction, that can be shifted towards the starting materials with elevated temperatures. Amino groups of adenine and cytosine can react with formaldehyde as well [51]; adenine reacts on 6-position NH_2 group [55] and cytosine reacts on 4-position NH_2 group. In theory, any amino acid can react with formaldehyde, containing reactive group such as -NH_2, -NH, -SH and -OH group, although only lysine, cytosine, tryptophane and histidine amino acids have been shown to crosslink to DNA with formaldehyde [56].

It has been reported that guanine can react with more than one formaldehyde [56]. Guanine has two possible reaction sites: the NH group at the 1-position and the NH_2 group at the 2-position. Lu *et. al*. has already shown [56] a multiple step formaldehyde incorporation between guanine and lysine, forming a ring like structure (Figure 1.6-4). The reaction mechanism is adapted from Lu *et al.* After the first linker incorporation (Step 2 crosslinked product Figure 1.6-3), the 1-position NH group reacts with another formaldehyde molecule, forming a methylol group, which after a water loss reacts with the lysine NH group forming a ring-like structure together with the guanine. In the last step another formaldehyde molecule reacts with the 2-postion NH group (Figure 1.6-4, crosslinked product 2) resulting an additional hydroxymethyl group on the guanine. This hydroxymethyl group does not convert into a Schiff base, thus there is no water loss in the last step (Figure 1.6-4, crosslinked product 3).

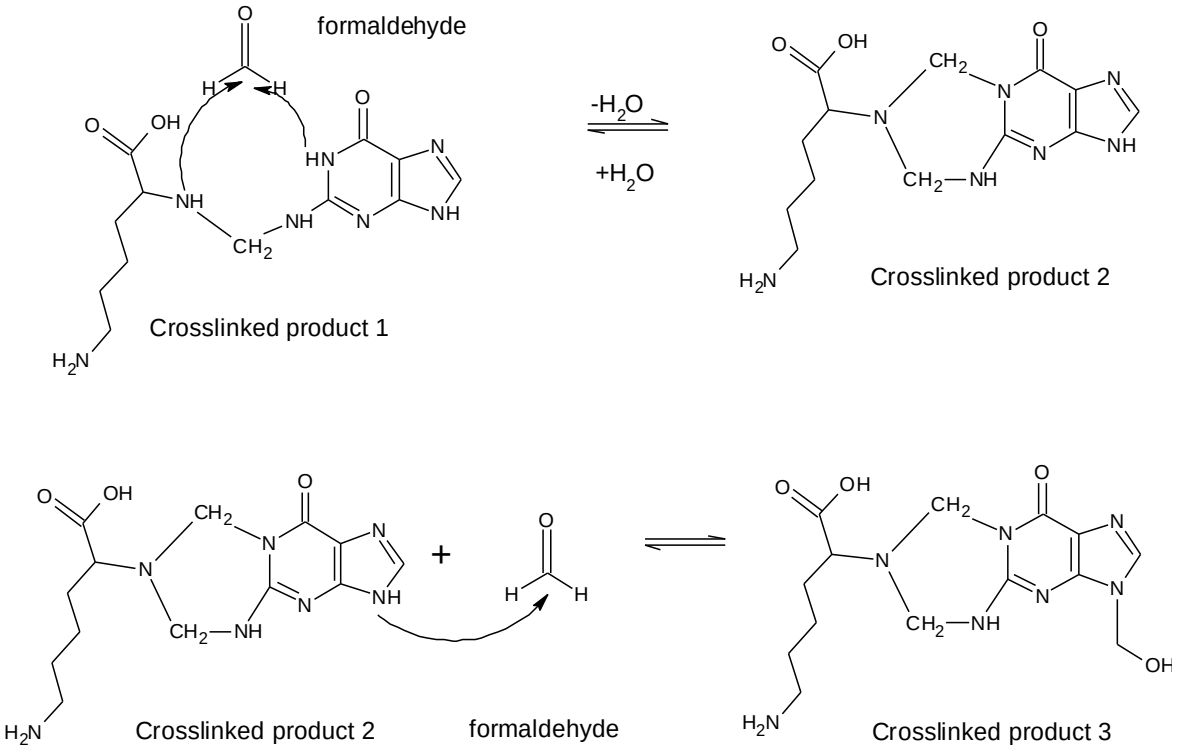


Figure 1.6-4: Proposed reaction scheme for addition of two formaldehyde linkers between formaldehyde crosslinked lysine and guanine heteroconjugate adapted from [56]

2 Objectives

In recent years, a vast amount of effort has been made for the development of mass spectrometric analysis of the UV crosslinked protein-RNA [40] and protein-DNA complexes [57,58] and protein-RNA UV crosslinking *in vivo* [33,34,43,59–62]. Meanwhile, chemical protein-nucleic acid crosslinking with mass spectrometric analysis have not been investigated thoroughly, only a very limited number of studies are available. Therefore, there are no standardized methods nor general protocols for mass spectrometric applications. The scope of the thesis is to investigate chemical crosslinking between proteins and nucleic acids in combination with mass spectrometric analysis. The following steps of method development are required to achieve that goal:

- Understand and describe the crosslinking chemistry between proteins and nucleic acids
    - Predict the chemical composition and length of the chemical crosslinker incorporated between proteins and nucleic acid.
    - Assess which chemical groups of the respective amino acids and nucleic acids are participating in the crosslinking reactions
- Investigation of the MS/MS fragmentation of peptide-(deoxy)oligonucleotide heteroconjugates
    - Examine which chemical bonds of the peptide-(deoxy)oligonucleotide heteroconjugates are prone to fragmentation and what are the chemical compositions of the possible MS/MS fragments
- Implementation of acquired MS/MS fragmentation observations into existing database searches
    - Establish a robust way to analyze the produced datasets for reliable crosslink identification.
- Apply crosslinking mass spectrometry on model systems with gradual increase of sample complexity
    - Establish specific HPLC-MS/MS measuring methods based on the observed MS/MS fragmentation patterns of the peptide-(deoxy)oligonucleotide heteroconjugates

When such a method is developed, sample complexity is gradually increased to assess the depth and sensitivity of the methodology. This study aims to investigate crosslinking chemistry and mass spectrometric analysis starting from a simple one-protein one-nucleic acid complex and adapting the method in proteome wide studies.

3 Materials and methods

3.1 Materials

3.1.1 Commonly used chemicals

Chemical	Supplier
4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)	Sigma-Aldrich
Acetic acid	Merck Millipore
Acetonitrile, LiChrosolv (ACN)	Merck Millipore
Ammonia solution, 25%	Merck Millipore
Ethanol, Analytical grade	Merck Millipore
Formaldehyde 37%	Sigma-Aldrich
Formic Acid	Sigma-Aldrich
Glycerol	Merck Millipore
Isopropanol	Merck Millipore
Lysozyme from chicken eggwhite	Sigma-Aldrich
Magnesium Chloride ($MgCl_2$)	Sigma-Aldrich
Methanol, LiChrosolv	Merck Millipore
Methanol, LiChrosolv (MeOH)	Merck Millipore
Potassium Chloride (KCl)	Sigma-Aldrich
Titansphere TiO_2 Bulk Material (10 µm)	GL Sciences
Trifluoric acid	Carl-Roth
Tris(hydroxymethyl)aminomethane (Tris)	Sigma-Aldrich
Urea	Sigma-Aldrich
Water, LiChrosolv	Merck Millipore
Zinc Chloride ($ZnCl_2$)	Sigma-Aldrich

3.1.2 Commonly used enzymes

Enzyme	Supplier
Antarctic phosphatase (5 U/µl)	New England Biolabs
Benzonase nuclease 25000 U	Merck Millipore
LysC (Mass Spec Grade)	Promega, Sigma-Aldrich
Nuclease P1 (100 U/µl)	New England Biolabs.
Pierce™ Universal Nuclease for Cell Lysis (250 U/µl)	Thermo Fisher Scientific
RNase A (1000 U/µl)	Thermo Fisher Scientific
RNase I (10 U/µL)	Thermo Fisher Scientific
RNase I (100 U/µl)	Invitrogen
RNase T1 (10 U/µl)	Thermo Fisher Scientific
Trypsin (sequencing grade)	Promega, Sigma-Aldrich

3.1.3 Commonly used buffers

Buffer	Composition
SP3 digestion Buffer	50 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 1.5 mM $ZnCl_2$.
Loading Buffer	2% (v/v) ACN, 0.05% (v/v) TFA
Phosphate buffered saline (1x PBS)	1.059 mM KH_2PO_4, 2.966 mM Na_2HPO_4, pH 7.4, 155.1724 mM NaCl
Native nucleosome buffer	20 mM Tris-HCl pH 7.5, 1 mM EDTA, 3 mM DTT, 10 mM Na-butyrate, 20% glycerol.
Oligonucleosome buffer	20 mM HEPES-KOH pH 7.5, 20 mM NaCl, 2 mM DTT, 1 mM EDTA
Mononucleosome and Mononucleosome+H1.4 buffer	10 mM Tris-HCl, pH 7.5, 25 mM NaCl, 2 mM DTT, 1 mM EDTA
70S ribosome buffer	20 mM HEPES-KOH pH 7.6, 10 mM Mg-acetate, 30 mM KCl, 7 mM β-mercaptoethanol
TiO_2- Buffer A	80% (v/v) ACN, 5% (v/v) TFA, 5% (v/v) glycerol
TiO_2- Buffer B	80% (v/v) ACN, 5% (v/v) TFA
TiO_2- Buffer C	0.3 M NH_4OH
RIPA Buffer	50 mM Tris-HCl, 150 mM NaCl, 1% Triton-X100, 0.1% Na-deoxycholate, 0.1% SDS

3.1.4 Consumables

Consumable	Manufacturer
C18 Micro SpinColumns	Harvard Apparatus
Diamond Tower Pack tips	Gilson Co.
Reprosil-Pur 120 C18-AQ, 1.9 µm	Dr. Maisch
Safe-Lock Tubes	Eppendorf
C18 extraction disks	3M
SpeedBeads magnetic carboxylate modified particles, hydrophobic 50mg/ml	Cytiva Lifesciences
SpeedBeads magnetic carboxylate modified particles, hydrophobic 50 mg/ml	Cytiva Lifesciences
Oasis HLB, 1cc Vac Cartridge 30 mg	Waters

3.1.5 Equipment

Equipment	Manufacturer
Bioruptor sonication apparatus UCW-201TM	Diagenode
Dionex Ultimate 3000 UHPLC	Thermo Fischer Scientific

Eppendorf Concentrator 5301	Eppendorf
Heraeus Fresco 17 Microcentrifuge	Thermo Fischer Scientific
Heraeus Multifuge X3R	Thermo Fischer Scientific
Heraeus Pico 17 Microcentrifuge	Thermo Fischer Scientific
Lab scale BP 211D	Sartorius
Lab scale CPA 423S	Sartorius
NanoDrop 1000 Spectrophotometer	Thermo Fischer Scientific
Orbitrap Fusion Lumos Tribrid	Thermo Fischer Scientific
Orbitrap Fusion Tribrid	Thermo Fischer Scientific
Q Exactive HF Plus	Thermo Fischer Scientific
Q Exactive HF-X	Thermo Fischer Scientific
Savant SPD121P Speed Vac	Thermo Fischer Scientific
Sonifier cell disrupter W 250	Branson Ultrasonics
Thermomixer C	Eppendorf
Thermomixer comfort	Eppendorf
Vortex-Genie 2	Scientific Industries

3.1.6 Software, programs, and online tools

ACD/ChemSketch https://www.acdlabs.com/resources/freeware/chemsketch/
Adobe Illustrator CS4 14.0.0
FragPipe 12.2, 16.0
GenBank https://www.ncbi.nlm.nih.gov/genbank/
MetaCyc https://metacyc.org
Microsoft Office 16.45
MSFragger 3.1.1 and 3.3
OpenMS 2.4.0 and 2.6.0
PANTHER 16.0 http://www.pantherdb.org
Philosopher 3.3.11, 4.0.0
ProteinProspector http://prospector.ucsf.edu
Proteome Discoverer 2.1
PyMOL 1.5.0.5
Python 3.7.4
R 4.0.0
R Studio 1.1.463
STRING 11.0
UniProt database https://www.uniprot.org/
Xcalibur 4.1.31.9

3.1.7 Samples

187 bp dsDNA, linker histone H1.4, mononucleosomes, oligonucleosomes and nucleosome H1.4 linker histone complexes (chromatosomes) have been prepared and kindly gifted by Dr. Alexandra Stützer [57]. Chicken linker histone H5 was purchased from Abcam (Cat No b81966). 70S ribosomes were purchased from New England Biolabs (Cat No P0763). HeLa native nucleosomes were purchased from BPS Bioscience (Cat No 52039). HeLa S3 cells were kindly gifted by Thomas Conrad (Research group Structural Biochemistry and Mechanisms, MPI bpc).

3.2 Methods

3.2.1 C18 reversed-phase chromatography

C18 reversed-phase chromatography was carried out with self-packed columns or prepacked C18 columns (Harvard Apparatus or Waters Oasis HLB 1cc Vac Cartridge 30 mg). Self-packed columns were prepared as follows: 5 layers of C18 material (C18 extraction disks, 3M) were layered on top of each other in a pipette tip. Equilibration, washing and elution steps were carried with 100 µl for self-packed tips, 200 µl for Harvard Apparatus and 1 ml for Waters columns. C18 reversed- phase chromatography was carried out as follows: columns were equilibrated twice with ACN, 80% (v/v) ACN, 0.1% (v/v) TFA, and with 60% (v/v) ACN, 0.1% (v/v) TFA and three times with 5% (v/v) ACN, 0.1% (v/v) TFA. The concentration of the sample was adjusted to 5% (v/v) ACN 0.1% (v/v) TFA and the sample was loaded onto the column. The column was washed three times with 5% (v/v) ACN, 0.1% (v/v) TFA. Elution was achieved with 80% (v/v) ACN, 0.1% (v/v) TFA.

3.2.2 TiO_2 enrichment

Samples were resuspended in TiO_2-Buffer A. Self-made columns were used for enrichment, prepared as described in Sharma *et al*. [63] The TiO_2 beads were equilibrated consecutively with TiO_2-Buffer B and TiO_2-Buffer A. The sample was loaded onto the column and washed three times with TiO_2-Buffer A and four times with TiO_2-Buffer B. An additional wash with 80% (v/v) ACN, 0.1% (v/v) TFA was performed to remove the excess TFA before elution. Samples were eluted with 0.3 M NH_4OH and dried under vacuum.

3.2.3 Single-pot, solid-phase-enhanced (SP3) sample preparation

3.2.3.1 SP3 bead preparation

20 µl 50 mg/ml hydrophobic carboxylate modified magnetic beads (Cytiva lifesciences) and 20 µl 50 mg/ml hydrophilic carboxylate modified magnetic beads (Cytiva lifesciences) were resuspended in 160 µl water. In-house made magnetic rack was used for bead separation from solution. Beads were washed twice with 200 µl water and resuspended in 100 µl water. Prepared SP3 beads were kept at 4°C for one week maximum.

3.2.3.2 SP3 protein clean-up

Respective amounts of beads, typically ranging from 1:1 to 1:10 (wt: wt) protein to beads ratio were used for SP3 cleanup. Prepared SP3 beads were added to the sample together in addition of ethanol to final concentration of 50% (v/v). The mixture was incubated in thermomixer at 1000 rpm to induce protein binding to the magnetic beads. The beads were separated on a magnetic rack and the supernatant was removed. Next, the beads were washed with 80% (v/v) ethanol, twice and were resuspended in digestion buffer, containing 50mM HEPES-NaOH pH 7.9, 1.5 mM $MgCl_2$, 1.5 mM $ZnCl_2$ or 50 mM Tris-HCl pH 7.9. Nucleases were added and DNA or RNA digestion was performed, followed by trypsin digestion overnight.

3.2.3.3 SP3 peptide cleanup

ACN was added to the digested peptide-oligonucleotide or peptide-deoxy(oligo)nucleotide mixture, including magnetic beads, to final concentration of 95% (v/v). The ACN containing solution was incubated for 5 minutes in a Thermomixer, at 1000 rpm. Magnetic rack was used for separation of the beads from the solution, the supernatant was removed, and the beads were washed three times with 100% ACN. Beads were quickly spun down and with the use of the magnetic rack, excess ACN was removed. Beads were resuspended in loading buffer to elute bound peptides and the samples were subjected to HPLC-MS/MS analysis.

3.2.4 DEB-crosslinking of linker histone-ds DNA systems

50 μg of linker histone H1.4 (2.29 nmol in 10 mM Tris-HCl pH 7.5, 20 mM NaCl, 1mM EDTA) and 187 bp ds DNA (in H_2O) were reconstituted in 2:1 molar ratio in 46 mM Na_2HPO_4, on ice for 30 minutes before crosslinking. The mixture was crosslinked with 46 mM DEB for 1 hour at 37 °C.
Chicken linker histone H5, 50 μg (2.41 nmol, in its storage buffer) and 187 bp (in H_2O) DNA were reconstituted in a 2:1 molar ratio in 50 mM Na_2HPO4, on ice for 30 minutes prior to crosslinking. 50 mM DEB was used for crosslinking for 1 hour at 37 °C.
DTT was added in a final concentration of 29 mM to quench the reactions. $MgCl_2$ was added in a final concentration of 1 mM and 375 U of benzonase nuclease was used for digestion. The samples were incubated for 1 hour at 37 °C. Four volumes of acetone were added, and the samples were kept at -20 °C until further processing.
The acetone precipitated complexes were centrifuged at 16.000xg for 30 minutes at 4 °C. The pellets were washed twice with ice-cold 80% (v/v) ethanol, followed by second centrifugation. The precipitates were dried on air and resuspended in 4M Urea, 50 mM Tris-HCl pH 7.9, 1mM $MgCl_2$. The solution was diluted to 1M Urea with 50 mM Tris-HCl pH 7.9, 1mM $MgCl_2$. Followed by nuclease digestion for 2 hours at 37 °C by the addition of 500 U benzonase nuclease. Trypsin was added in a 1:10 (wt/wt) ratio, and enzymatic digestion took place overnight at 37 °C. C18 reversed-phase chromatography was carried out in the following way: self-packed columns were activated with methanol, washed with 95% (v/v) ACN, 0.1% (v/v) formic acid; 80% (v/v) ACN, 0.1% (v/v) formic acid; 50% (v/v) ACN, 0.1% (v/v) formic acid and 0.1% (v/v) formic acid. The sample concentrations of ACN and formic acid were adjusted to 5% (v/v) and 0.1% (v/v), respectively. The sample were loaded onto the column followed by washing twice with 0.1% (v/v) formic acid and elution with

increasing concentration of ACN containing buffers: 10% (v/v) ACN, 0.1% (v/v) formic acid; twice 50% (v/v) ACN, 0.1% (v/v) formic acid and 80% (v/v) ACN, 0.1% (v/v) formic acid. The samples were dried under vacuum and subjected to TiO_2 enrichment, described as above. Samples were resuspended in 10 µl loading buffer and 4 µl were injected for HPLC-MS/MS analysis.

3.2.5 DEB-crosslinking of nucleosome containing complexes

60 µg of oligonucleosomes were crosslinked in their reconstitution buffer at final concertation of 200 mM DEB for 2 hours at 37 °C. The reaction was quenched with Tris-HCl pH 7.9 at a final concentration of 200 mM. 30 µg of mononucleosomes and 15 µg of mononucleosome +H1.4 complexes were crosslinked in their reconstitution buffer at final concentration of 200 mM DEB for 1 hour at 37 °C. Reactions were quenched with Tris-HCl pH 7.9 at final concentration of 220 mM. SP3 protein cleanup was used for buffer exchange in each case. Prepared SP3 beads were added to an approx. 1: 10 (wt: wt), complex to beads ratio. SP3 protein cleanup was performed. Beads were reconstituted with digestion buffer: 50 mM HEPES-NaOH, pH 7.9, 1.5 mM $MgCl_2$, 1.5 mM $ZnCl_2$. 500 U of universal nuclease and 200 U of Nuclease P1 were added to each sample and nuclease digestion was performed for 1 hour at 37 °C. Trypsin was added to each sample, to oligonucleosomes 1:20 (wt: wt) trypsin to protein ratio or 1:10 (wt: wt) in the case of mononucleosomes and mononucleosome+H.14 samples. Digestion was carried out overnight at 37 °C in the case of oligonucleosome sample and at 30°C in the case of mononucleosome and mononucleosome+H1.4 samples. SP3 peptide cleanup was performed. Beads were reconstituted with 50 µl of loading buffer in the case of oligonucleosome samples and 30 µl of loading buffer in the case of mononucleosome and mononucleosome+H1.4 samples. 8 µl of the oligonucleosome sample were kept for quality control. All samples were dried under vacuum and TiO_2 enrichment took place. Samples were eluted from the TiO_2 matrix and dried under vacuum. Samples were resuspended in 12 µl of loading buffer. 6 µl of the mononucleosome samples, 8 µl of the mononucleosome+ H1.4 sample, and 7 µl of the oligonucleosome sample were subjected to HPLC-MS/MS analysis.

3.2.6 Formaldehyde crosslinking

3.2.6.1 Formaldehyde-crosslinking of Mononucleosomes

Mononucleosomes were crosslinked with formaldehyde in total concentration of 1%, in their respective reconstitution buffer for 1 hour at 37 °C, 220 mM Tris-HCl pH 7.9 was added to stop the reaction. SP3 protein cleanup was used to remove the crosslinker, as described in 3.2.3.2. Beads were added at an approx. 1:10 (proteins to beads, wt: wt ratio). SP3 cleanup was performed as described in 3.2.3.2. Beads were reconstituted with 50 mM HEPES-NaOH, pH 7.9, 1.5 mM $MgCl_2$, 1.5 mM $ZnCl_2$. 500 U of universal nuclease and 200 U of nuclease P1 were added, and nuclease digestion was performed for 1 hour at 37 °C. Trypsin was added at a 1:10 (trypsin: protein wt: wt) ratio and proteolysis was performed overnight at 30 °C, followed by SP3 peptide cleanup. The sample was reconstituted with 30 µl loading buffer and dried under vacuum. The digest was subjected to TiO_2 affinity chromatography. The sample was dried under vacuum. The sample was resuspended in 12 µl loading buffer, 6 µl of which were subjected to HPLC-MS/MS analysis.

3.2.6.2 Control experiments for formaldehyde-crosslinking

Linker histone H1.4, 20 μg (0.916 nmol, in 10 mM Tris-HCl pH 7.5, 20 mM NaCl, 1mM EDTA) was mixed with equimolar 187 bp DNA (in H_2O) at a total volume of 131.5 μl in duplicates. In parallel only protein and only DNA samples were prepared by diluting the respective biomolecule to the same volume (131.5 μl)
All samples were incubated on ice for 30 minutes. One of the linker histone H1.4-187 bp DNA complex replicate was processed, without crosslinking. The rest were crosslinked with 2% formaldehyde, for 1 hour at 37 °C. Reactions were quenched with 600 mM Tris-HCl pH 7.9.
SP3 protein cleanup was used for crosslinker removal by adding 20 μg of beads to each sample. Beads were resuspended in 200 μl 50 mM HEPES-NAOH pH 7.9, 1.5 mM $MgCl_2$, 1.5 mM $ZnCl_2$. 2500 U universal nuclease and 600 U Nuclease P1 were added, and nuclease digestion was performed for 1 hour at 37 °C. 1 μg of trypsin was added to each sample and overnight digestion was carried out at 30 °C.
The SP3 beads were sonicated for 1 minute and quickly spun down. Magnetic rack was used for bead separation and the supernatant was subjected to C18 reversed-phase chromatography (Harvard Apparatus columns). Elution was performed in a two-step manner: with 40% (v/v) ACN, 0.1% (v/v) TFA and 60% (v/v) ACN, 0.1% (v/v) TFA. The samples were dried under vacuum and subjected to TiO_2 enrichment The samples were suspended in 12 μl loading buffer and 4 μl were subjected to LC-MS/MS analysis. 50 ng/μL HeLa Protein Digest Standard (Pierce, Thermo Fischer Scientific Cat. No:88328) supplemented with 0.25x iRT standard (Biognosys) was prepared in advance, 1 μl of the mixture was used for additional control, which was submitted to LC-MS/MS analysis.

3.2.6.3 Formaldehyde-crosslinking of HeLa native nucleosomes

105.6 μg HeLa nucleosomes were crosslinked in their storage buffer with formaldehyde in final concentration of 1% for 1 hour at 37 °C. The reaction was quenched with the addition of 350 mM Tris and the sample was subjected to SP3 cleanup with SP3 beads in approx. 1:12 ratio (sample to beads wt: wt). SP3 protein cleanup was carried out. The bound beads were reconstituted with 150 μl 50 mM HEPES-NaOH, pH 7.9, 1.5 mM $MgCl_2$, 1.5 mM $ZnCl_2$. 2000 U of universal nuclease and 500 U nuclease P1 were added to the sample. Nuclease digestion was carried out at 37 °C for 2.5 hours. 10 μg trypsin was added to the sample and digestion was carried out at 30 °C, overnight. The following day, 1250 U of universal nuclease were added, and nuclease digestion was carried at 37 °C for 1 hour. SP3 peptide cleanup was carried out. The sample was reconstituted with 100 μl loading buffer to elute the supernatant, which was dried under vacuum, followed by TiO_2 enrichment. Enriched analytes were resuspended in 12 μl of loading buffer, 5 μl of which were subjected to HPLC-MS/MS analysis.

3.2.6.4 Formaldehyde-crosslinking of the 70S ribosome

200 μg 70S ribosome was crosslinked in its storage buffer with formaldehyde at a total concentration of 1 (v/v) % at 37 °C. The reaction was quenched with Tris-HCl pH 7.9 in a total concentration of 350 mM. SP3 cleanup was used for buffer exchange, beads were added to a 1:2 (complex to beads ratio). The sample was reconstituted with 50 mM Tris-HCl pH 7.9. 50 U RNase 1, 2000 U RNase A were added, and nuclease digestion was performed at 37

°C for 1 hour. Trypsin was added to 1:20 ratio (trypsin to protein, wt: wt) and overnight digestion was carried out at 37 °C. Sample was subjected to C18 reversed-phase chromatography (Harvard apparatus). The sample was eluted from the C18 column with 60% (v/v) ACN, 0.1% (v/v) TFA, dried under vacuum and subjected to TiO_2 enrichment. A second nuclease digestion was carried out as follows: the sample was resuspended in 50 µl 50 mM sodium acetate, pH 5.5 and 30 U nuclease P1, 1.5 U Antarctic phosphatase were added. Nuclease digestion was carried out for 4 hours at 37 °C after which C18 reversed-phase chromatography was used for cleanup. After it was dried under vacuum, the sample was resuspended in 12 µl loading buffer and 2 µl sample was subjected to HPLC-MS/MS analysis.

3.2.7 Bacterial growth and crosslinking

Escherichia coli cells (Rosetta 2 DE3) were cultured in Lysogeny broth (LB) medium supplemented with chloramphenicol 25 µg/ml. 200 ml of the LB medium was inoculated with 100 µl of competent cells, and the culture was incubated at 37 °C, overnight. Five times of 400 ml of LB medium was inoculated with 500 µl of the preculture and incubated at 37 °C until a OD600 of 0.6 was reached. The culture was split in two portions and harvested by centrifugation. The cells were washed with ice-cold PBS and were resuspended in 40 ml PBS with 1% FA for crosslinking experiments or PBS only for control conditions. Crosslinking was achieved at 37 °C for 10 minutes followed by quenching with Tris, in total of 400 mM final concentration at room temperature for 10 minutes. Cells were spun down at 3000x g for 10 minutes and washed with ice-cold PBS, PBS was removed, and cells were aliquoted, flash frozen and aliquots were kept at -80 °C until further processing.

3.2.8 Silica-based enrichment of *E. coli* cells

Silica-based enrichment protocol was executed as previously described [64]. Approx. 1.2E11 of *E. coli* cells were resuspended in 300 µl PBS containing 9 mg of lysozyme and the cell pellet was incubated for 5 minutes at room temperature. 3ml 8M Urea containing 100 mM HEPES-NaOH pH 7.9 and 20mM EDTA was added. Sonifier cell disrupter was used for mechanical lysis (30% vibration amplitude, 0.5s on, 2s off, total 150 pulses). The urea concentration was reduced to 1M and 200 µg trypsin was added for overnight digestion. 5 ml of the cell lysate was used for silica enrichment with Qiagen RNeasy maxi kit. Sample was processed according to manufacturer's instructions. Each centrifugation step was performed at 3.000xg. Briefly, 19 ml RLT Buffer, supplemented with β-mercaptoethanol (10 µl/1ml RLT Buffer), and 14 ml ethanol were added to the cell lysate. The sample was loaded onto the column and centrifuged for five minutes. The column was washed with 7.5 ml RW1 Buffer, and twice with 10 ml RPE Buffer. The sample was eluted in a two-step manner with RNase-free water and 8 M Urea, 100 mM HEPES-NaOH pH 7.9 were added to final concentration of 1M Urea, 12.5 mM HEPES-NaOH pH 7.5. Protein digestion was carried out for 2 hours at 37 °C with the addition of 20 µg trypsin, followed by second RNA purification as previously described. The column eluate was further processed for RNA digestion. Concentration of Urea was adjusted to 1M and HEPES-NaOH pH 7.9 was adjusted to 12.5 mM, $MgCl_2$ and $ZnCl_2$ were added at a total concentration of 1 mM each. 1250 U of universal nuclease, 50 U RNase I, 300 U nuclease P1, 2 U RNase T1 and 200 U RNase A were added, and nuclease digestion was carried out, overnight at 30 °C. C18 reversed-phase chromatography was performed to remove free oligonucleotides from peptide-nucleotide heteroconjugates. C18 reversed-phase chromatography was carried out as it was described in section 2.7. The purified sample was eluted with 600 µl 60% (v/v) ACN, 0.1% (v/v) TFA.

The eluate was dried under vacuum and the sample was resuspended in 100 µl, 50 mM sodium acetate pH 5.5, 80 U nuclease P1 and 4 U Antarctic phosphatase were added, and digestion was carried out for 2 hours at 30 °C. Followed by C18 reversed-phase chromatography with self-made C18 columns. The sample was eluted with 200 µl of 60% (v/v) ACN, 0.1% TFA, dried under vacuum, resuspended in 12 µl loading buffer, 2µl and 3 µl were subjected to HPLC-MS/MS analysis.

3.2.9 HeLa cells formaldehyde-crosslinking

Approx. 3.5E9 number of HeLa S3 cells were washed with 800 ml cold phosphate-buffered saline (PBS) and pelleted at 300xg for 5 minutes at 4 °C. Cells were crosslinked with 1 (v/v) % formaldehyde in PBS (3.5E6 cells/ml) for 10 minutes at 37 °C at 100 rpm rotation. Crosslinking was quenched with Tris at a final concentration of 350 mM for 10 minutes at room temperature. Cells were pelleted after crosslinking with 300xg for 5 minutes, washed twice with 1x PBS and aliquots of approx. 2E8 number of cells were pellet. Cells were flash frozen and stored at -80 °C until further processing.

3.2.9.1 Silica enrichment of HeLa cells

Approx. 2E8 HeLa cells were lysed with 15 ml RIPA buffer (as described in [62]) and sonicated with ultrasonic sonifier for 30 pulses of 30% vibration amplitude, 0.5 on/2 off. 400 U of RNase inhibitor was added, followed by trypsin digestion overnight with the addition of 100 µg trypsin at 30 °C. The next morning additional sonication step was carried out with ultrasound sonifier, for 10 seconds at 30% amplitude on ice. The sample was spun down, and the supernatant was used. Silica based enrichment strategy was used with Qiagen RNeasy maxi kit. Sample was processed according to manufacturer's instructions; each centrifugation step was performed at 3.000xg. Briefly, 19 mL RLT Buffer, supplemented with β-mercaptoethanol (10 µl/1ml RLT Buffer), and 14 ml ethanol were added to the cell lysate. The sample was loaded onto the column and centrifuged for five minutes. Column was washed with 15 ml RW1 Buffer, and twice with 10 ml RPE Buffer, eluted in a two-step manner with 1600 µl and 600 µl RNAse free water and 8 M Urea and 100 mM HEPES-NaOH pH 7.9 were added to final concentration of 1M Urea, 12.5 mM HEPES-NaOH pH 7.9. Protein digestion was carried out for 2 hours at 37 °C with the addition of 20 µg trypsin. Second RNA purification was performed with Qiagen RNeasy maxi kit, as described above, except 10 ml RW1 Buffer was used instead of 15 ml RW1 Buffer. RNA was eluted twice with 1ml RNase free water. Eluates were pooled together and the concentration of Urea and HEPES-NaOH pH 7.9 were adjusted to 1 M and 13 mM, respectively. $MgCl_2$ and $ZnCl_2$ were added to a final concentration of 0.8 mM. Nuclease digestion was carried out at 30 °C overnight with the addition of 70 U RNase I, 700 U of universal nuclease 500 U nuclease P1, 12.5 U RNase A and 0.125 U RNase T1. The next morning, the sample was quickly spun down, and the supernatant was used for C18 reversed-phase chromatography, described in section 3.2.1 with 1cc (Oasis HLB, 1cc Vac Cartridge 30 mg) C18 columns. 600 µl of 40% ACN, 0.1% TFA were used for sample elution. The sample was dried under vacuum and subjected to a sequential RNA digestion, after which it was resuspended in 50 µl 50 mM sodium acetate supplemented with 200 U of nuclease P1 and 150 U Antarctic phosphatase. Digestion was carried out for 2 hours at 30 °C and the sample was subjected to C18 reversed-phase chromatography with self-packed columns. The sample was eluted with 100 µl 40 (v/v) % ACN, 0.1 (v/v) % TFA and sample was dried in SpeedVac. Sample was resuspended in 12 µl loading buffer and 5 µl samples were subjected as replicates to HPLC-MS/MS analysis.

3.2.9.2 HeLa Silica enrichment and basic reversed-phase fractionation

Approx. 2E8 HeLa cells were lysed with 15 ml RIPA buffer and sonicated with ultrasound sonifier with 30 pulses of 30% vibration amplitude, 0.5 on/ 2off. 480 U of RNase inhibitor was added to the sample. 100 µg of trypsin was used for protein digestion, which was performed overnight at 20 °C. In the next morning, the sample was spun down, and supernatant was used. Silica enrichment strategy was used for RNA enrichment with the use of Qiagen RNeasy maxi kit, according to the manufacturer's instructions. Each centrifugation step was performed at 3.000xg. Briefly, 19 mL RLT Buffer, supplemented with β-mercaptoethanol (10 µl/1mL RLT Buffer) and 14 ml ethanol was added to the cell lysate. The sample was loaded onto the column and centrifuged for five minutes. The column was washed with 15 ml RW1 Buffer, and twice with 10 ml RPE Buffer, followed by elution in a two-step manner with RNase-free water, each step with 1ml. Urea and HEPES-NaOH pH 7.9 were added to the final concentration of 1M Urea, 10 mM HEPES-NaOH pH 7.9. 20 µg trypsin was added and digestion was carried out for 2 hours at 30 °C. Second RNA purification was carried out as previously described, with Qiagen RNeasy maxi kit. RNA was eluted from the column with 1ml RNase free water, twice. Eluates were pooled together and the concentrations of Urea and HEPES-NaOH pH 7.9 were adjusted to 1 M and 13 mM, respectively. $ZnCl_2$ and $MgCl_2$ were added to the sample in a final concentration of 0.5 mM. 2500 U of universal nuclease, 500 U RNase I, 800 U of nuclease P1, 12.5 U RNase A and 0.125 U RNase T1 were added, and digestion was carried out at 30°C overnight. Followed by C18 reversed-phase chromatography, described in section 3.2.1. The sample was eluted with 600 µl 40% (v/v) ACN and 0.1% (v/v) TFA and dried under vacuum. For sequential RNA digestion, the sample was resuspended in 50 µl sodium acetate pH 5.5, and 7.5 U Antarctic phosphatase, 150 U nuclease P1, 12.5 U RNase A and 0.125 U RNase T1 were added for digestion at 30 °C for 1.5 hours. The sample was subjected to C18 reversed-phase chromatography, with self-made columns, was eluted from the column with 100 µl 40 (v/v) % ACN, 0.1(v/v) % TFA and dried under vacuum.
Enriched peptide-nucleotide heteroconjugates were fractionated with basic reversed-phase chromatography (Agilent Santa Clara, USA, 1100 series,), on a C18-X-Bridge column (length 150 mm, 1.0 inner diameter, particle size 3.5 µm; Waters, Milford, USA). Two buffers were used under high pH conditions (pH ~10), Buffer A: 10 mM NH_4OH and Buffer B: 10 mM NH_4OH, 80% (v/v) ACN. The chromatographic system was used with the flow rate of 60 µl/min. The column was equilibrated with the mixture of 5% Buffer B and 95% Buffer A. Linear gradient was used with increasing amount of Buffer B from 10% to 35% for 35 minutes, followed by up to 50% over 8 minutes. At the end of the chromatographic run, the column was washed with increasing amount of Buffer B from 90% to 95% for 5 minutes. Fractions were collected in one minute basis; fractions were concatenated into 24 fractions. The samples were dried in SpeedVac, samples were resuspended in 12 µl loading buffer, 8 µl was subjected to HPLC-MS analysis.

3.3 Data processing

3.3.1 Protein databases

Fasta files were downloaded from UniProt, except for H2A, H2B and H3 Luger histone sequences, int those cases Fasta files were downloaded from GenData with the following identifiers: H2A: CAD89676.1, H2B: CAD89678.1 and H3: CAD89679.1. Custom made databases were used in the case of linker histones, mono- and oligonucleosomes, chromatosomes (mononucleosomes+H1.4 linker histone) and 70S ribosomes (*E. coli* K12

strain 70S ribosomal protein sequences were used), containing individual protein sequences. Human nuclear database was downloaded from UniProt (31. 01. 2020) and was used in the case of HeLa native nucleosomes. Human database was downloaded from UniProt (20. 09. 2020) supplemented with common contaminant protein sequences, provided by MaxQuant [65]. In the case of *E. coli*, concatenated BL21 DE3 database and K12 strain database were used (version 03. 03. 2020 and 26. 09. 2019 respectively), supplemented with common contaminant protein sequences, provided by MaxQuant.

3.3.2 Data conversion

Raw files were converted with Proteome Discoverer 2.1 to the mzML format. The parameters in the spectrum selection were as follows: Min. Precursor mass 350 Da, Max. Precursor mass 20000 Da. Signal to noise ratio threshold was set to 1.2.

3.3.3 Database searches

mzML files were searched with the RNP^{xl} [40] processing node from OpenMS [41] (version: 2.4.0 and 2.6.0) and with FragPipe (version: 12.2 and 16.0) including MSFragger [66], Philosopher [67] and the PTM Shepard [68] pipelines. Modification searches were performed with the Proteome Discoverer (version: 2.1) environment.

3.3.3.1 RNP^{xl} search parameters of DEB-crosslinked protein-DNA complexes

MS^1 precursor tolerance was set to 8 ppm, precursor charge state from +2 to +5 were enabled. Isotope error was set to 1, allowing precursor mass correction. MS2 fragment tolerance was set to 20 ppm. Trypsin was set as endonuclease with maximum missed cleavages of 3. Minimal peptide length was set to 4 with no maximum peptide length.

Table 3.3-1: RNP^{xl} specific parameters for DEB-crosslinked peptide-DNA heteroconjugates identification

deoxynucleotide length	4
target deoxynucleotides	A=C10H14N5O6P
	G=C10H14N5O7P
	C=C9H14N3O7P
	T=C10H15N2O8P
can crosslink	AGCT
Fragment adducts	
	G:C4H4O;DEB-H2O
	G:C4H6O2;DEB
	G:C14H20N5O8P;DEB+G
	G:C9H9N5O2;DEB+Gb-H2O
	G:C9H11N5O3;DEB+Gb
	G:C10H9N5O2;G-H3PO4-H2O
	G:C5H5N5O;Gb

	A:C4H6O2;DEB
	A:C9H11N5O2;DEB+Ab
	A:C9H8N4O2;DEB+Ab-NH3
	A:C14H20N5O8P;DEB+A
	A:C10H9N5O;A-H3PO4-H2O
	A:C5H5N5;Ab
	T:C4H6O2;DEB
	T:C9H12N2O4; DEB+Tb
	T:C9H10N2O3; DEB+Tb-H2O
	T:C14H21N2O10P;DEB+T
	T:C14H18N2O6;DEB+T-H3PO4
	T:C14H20N2O7;DEB+T-HPO3
	T:C5H6N2O2;Tb
	C:C4H6O2;DEB
	C:C8H11N3O3;Cb+DEB
	C:C13H20N3O9P1;DEB+C
	C:C4H5N3O;Cb
	T:C14H22N2O13P2;DEB+T+HPO3
	T:C19H26N2O14P2;DEB+T+HPO3+C5H4O
Precursor adducts	
	G:C4H6O2
	A:C4H6O2
	T:C4H6O2
	C:C4H6O2
	G:C4H6O2-C5H9O6P
	A:C4H6O2-C5H9O6P
	C:C4H6O2-C5H9O6P
Scoring	Slow
Decoys	enabled

Manual validation of crosslinks was performed on the 10% FDR cut CSMs.

3.3.3.2 RNP^{xl} search parameters of the FA-crosslinked protein-DNA complexes

MS^1 precursor tolerance was set to 10 ppm, precursor charge state from +2 to +5 were enabled. Isotope error was set to 1, allowing precursor mass correction. MS2 fragment tolerance was set to 20 ppm. Trypsin was used as endonuclease with maximum missed cleavages of 2. Minimum peptide length was set to 4, no limit was set to maximum peptide length. RNP^{xl} specific parameters were as follows:

Table 3.3-2: RNP^{xl} specific parameters for formaldehyde-crosslinked peptide-DNA heteroconjugates identification

deoxynucleotide length	3
target deoxynucleotides	A=C10H14N5O6P
	G=C10H14N5O7P
	C=C9H14N3O7P
	T=C10H15N2O8P
can crosslink	AGC
fragment adducts	C:C;FA
	A:C;FA
	T:C;FA
	G:C;FA
	A:C6H5N5;Ab+FA
	C:C5H5N3O;Cb+FA
	G:C6H5N5O;Gb+FA
	C:C10H14N3O7P;C+FA
	A:C11H14N5O6P;A+FA
	G:C11H14N5O7P;G+FA
	C:C2;FA2
	A:C2;FA2
	G:C2;FA2
	A:C5H5N5;Ab
	G:C5H5N5O;Gb
	C:C4H5N3O;Cb
	A:C10H14N5O6P;A
	G:C10H14N5O7P;G
	T:C10H13N2O7P;T-H2O
	G:C11H9N5O2;G-H3PO4-H2O+FA
	C:C4H5N3O;Cb
	C:C9H14N3O7P;C
Precursor adducts	G:C
	A:C
	C:C
	C:C2
	G:C2
	A:C2
	G:C-HPO3
	A:C-HPO3
	C:C-HPO3

	G:C3H2O
	A:C3H2O
	G:C3H2O-HPO3
	A:C3H2O-HPO3
Scoring	Fast
Decoys	enabled

3.3.3.3 Search parameters for open searches

Decoy database for the MSFragger search was generated with the DecoyDatabase node in TOPPAS (OpenMS version 2.6.0), with the reversal of the protein sequences in the database.
FragPipe platform was only used for the database searches of the 70S ribosome, *E. coli* and HeLa datasets, therefore the mass offset values only contain RNA masses. MSFragger [66] [69], PeptideProphet [70], ProteinProphet [23], PTMShepard [68] were used in the FragPipe (version 12.2 and version 16.0) environment. MSFragger was used for the mass offset search with the following parameters:
Precursor mass tolerance was set to +/- 20 ppm, 1 Da isotope error was allowed. Mass calibration and optimization was enabled, fragment tolerance was set to 20 ppm, enzyme was set to trypsin with a maximum number of 3 missed cleavages. Peptide length was set from 5 to a maximum number of 65 amino acids. Peptide mass range was calculated between 300 and 5000 Da. Variable modifications were as follows: methionine oxidation, loss of ammonia from the peptide N-terminal and pyroglutamic acid. Mass offsets were as follows (in Da):
0 255.0855 279.0967 295.0917 341.0525 335.0519 357.0474 359.0631 375.0580 560.1268 561.1108 584.138 585.1221 600.1330 601.1169 624.1441 640.0931 640.1391 664.1044 665.0884 680.0993 688.1156 704.1105 720.1054.
MSFragger search results validation was performed with Philosopher pipeline [44] in the FragPipe environment. False discovery rate (FDR) was calculated based on the target decoy approach. PeptideProphet was used to calculate the peptide probabilities. PeptideProphet parameters were as follows: extended mass model (1000 Da), nonparametric modelling, using expectation scores only, cLevel 2. ProteinProphet was used to calculate the protein inference, ProteinProphet was used with the recommended settings for mass offset search: maxppmdiff flag was set to 2000000 ppm. Filtering was performed at 1% FDR on the PSM, peptide and protein level. Sequential and razor flags were used for the additional FDR calculation on the PSMs and on the identified protein list.

3.3.3.4 Marker ion search

3.3.3.4.1 Protein-DNA formaldehyde crosslinking datasets

CSM files of the RNP^{xl} search were imported in R. .mzML files were imported into R as well, with the use of the mzR package [71–74]. The following R packages were used for the downstream analysis: stringr package [75] and stringi package [76]. Based on the DNA adduct compositions, their respective marker ions and mass shifted marker ions were searched in the MS2 spectra, when the respective marker ions were not present CSMs were

removed. CSM tables with the assigned DNA adducts and crosslinked deoxynucleotides were imported.

3.3.3.4.2 Protein-RNA formaldehyde crosslinking datasets

PSM files of the open search results were imported in R and delta masses were extracted. Delta masses were compared with list of RNA adduct compositions. mzML files were read in R with the use of mzR package [71–74]. The following R packages were used for downstream analysis: stringr package [75], stringi package [76] and protr package [76,77]. Based on the RNA adduct composition, their respective marker ions masses were searched in the mzML files. Compositions of RNA adducts were conformed when their respective marker ion masses were present, when marker ions were not present, PSMs were removed. Crosslinked peptides' mass accuracy was calculated, if mass accuracy was below -10ppm or above +10 ppm (except for peptides with isotope errors), PSMs were removed. When peptide positions in their respective protein sequences were not given by the MSFragger search, protr package was used for the localization of the peptide sequences in the protein sequences. PSM tables with the assigned RNA adducts, crosslinked nucleotides, mass accuracy and peptide positions were imported.

3.3.3.5 Search parameters for modification searches

Proteome Discoverer 2.1 platform was used for modification search in the case of linker histones, 187 bp ds DNA experiments. Processing steps contained Spectrum files, Spectrum selector, Sequest HF, Target Decoy PSM validator and ptmRS nodes [78]. raw files were read in with the Spectrum Files node. Spectrum selector node was used with the following parameters: MS(n-1) Precursor was used, minimal precursor mass was set to 350 Da, maximal precursor mass was set to 5000 Da, maximum collision energy was set to 1000. Signal to noise ratio was set to 1.5. Sequest HT search parameters were as follows: trypsin was set as endonuclease with the maximal number of 4 missed cleavages. Precursor mass tolerance was set to 10 ppm, fragment mass tolerance was set to 0.02 Da. Neutral losses from a, b and y ions were allowed. Maximal number of equal modifications were set to 3. DEB (86.037 Da) and DEB+Gb (237.086) were used on the following amino acids (used one letter representations): (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T and Y). Target Decoy PSM validator node was used with maximum Delta Cn of 0.05. Strict Target FDR was set to 0.01 and Relaxed Target FDR was set to 0.05. ptmRS node [78] was used for modification site probability calculation. Diagnostic ions were allowed to rule out potential isoforms.
Consensus steps were used with the recommended values: results files of the Processing steps were sent to Consensus step with MSF Files node. PSM Grouper node was used of PSM grouping. Peptide validator node used the calculated PSM FDR values. Strict Target FDR was set to 0.01, Relaxed Target FDR was set to 0.05. Peptide validation was based on q-values, high confidence peptides were used with minimal peptide length of 6 amino acids. Strict Protein FDR was set to 0.01, Relaxed Protein FDR was set to 0.05. Strict parsimony principle was set to true in the Protein Grouping node.

3.3.4 R script for crosslink validation in MS2/MS3 measurement

R script was used for connecting MS levels and assigning crosslinked deoxynucleotides for MS3 level. MS/MS Spectrum Info table and PSM table were exported from the Proteome Discoverer platform as csv files. The two tables were read in R. Based on the scan numbers, precursor masses in MS2 and MS3 were assigned. Delta masses and theoretical DNA masses

were compared, and a match was considered, when mass differences between the delta masses and the theoretical masses were below 50 ppm. Result table, containing the identified crosslinks were exported as .csv format for the generation of supplementary tables.

3.3.5 GO term analysis

UniProt IDs of the identified crosslinked proteins were uploaded to the STRING [79] database and to the PANTHER [80] classification system, where functional enrichment analysis and functional classifications of the GO (Gene Ontology) terms were performed. In the case of proteins groups, first protein hit was chosen to represent the protein group. KEGG pathway analysis [81] was performed by STRING.

3.3.6 RNA-binding domain analysis

Identified crosslink peptide sequences were uploaded to the UniProt peptide search tool. UniProt peptide search gives information about the crosslinked peptide position in the protein sequence as well as the RNA binding domain position in the respective protein sequence. Manual comparison was performed between the position of the crosslinked peptide in the protein sequence and the RNA binding domain in the protein sequence. When an overlap between the crosslinked peptides' sequence positions and proteins domains' sequence positions was observed, it was considered a match. When peptide was shared between crosslinked proteins, it was considered as one match to the RNA binding domains.

3.3.7 Manual validation of DEB-crosslinked data

mzML files were searched with RNP^{xl}: mzML and idXML (RNP^{xl} results files) were loaded to the TOPPView [82] tool of the OpenMS pipeline, where crosslinked spectra were visualized. In the case of DEB crosslinked results, manual annotation of the crosslink identification was performed with the following considerations: good sequence coverage of the crosslinked peptide; presence of the DEB and DEB+Nb mass shifted ion series with at least 3 consecutive ions; in the case of low peptide backbone fragmentation - presence of DEB mass shifted peptide precursor ion and/or DEB+nucleobase mass shifted peptide precursor ions.

3.3.8 Data visualization

MS/MS spectra of crosslinked peptide-(oligo)deoxynucleotides were exported from the TOPPView tool of OpenMS as bmp. When MSFragger search was used, pepXML files were converted to idXML format with the IDFileConverter node in OpenMS. TOPPView tool was used for spectra visualization of the MSFragger results figures were exported as bmp files. Structures were visualized with PyMol and pictures of structures were exported as png files from PyMol. Biochemical workflows were hand drawn in Adobe Illustrator.

3.4 Parameters for HPLC-MS/MS measurements

3.4.1 HPLC parameters

Enriched crosslinked peptide-deoxy(oligo)nucleotide or peptide- (oligo)nucleotide heteroconjugates were resuspended in the loading buffer (2% ACN, 0.05% TFA) and were submitted to HPLC-MS/MS analysis. Dionex Ultimate 3000 RSLC nano system was coupled with a Q Exactive HF, Orbitrap Fusion Tribrid and Orbitrap Fusion Lumos Tribrid instruments. Peptide-deoxy(oligo)nucleotide or peptide-(oligo)nucleotide heteroconjugates were loaded on a Pepmap 300 C18 trap column (Thermo Fisher). The heteroconjugates were separated on an in-house packed (ReproSil- Pur 120Å, 1.9 µm, C18-AQ, 30 cm, 75 µm inner diameter) main column. During the chromatographic runs, buffer A (0.1% (v/v) formic acid) and Buffer B (80% (v/v) ACN, 0.08% (v/v) formic acid) were used. In every chromatographic run, the column was incubated with the respective Buffer B (ranging from 2-8%) for three minutes. During the measurements the main column was kept at 50 °C.

Table 3.4-1: Parameters for chromatographic separation during HPLC-MS/MS acquisition

Sample name/experiment name	**Gradient composition (Buffer B%)**	**Chromatographic run**
Mononucleosme (DEB, FA), Oligonucleosome (DEB)	(5-45)%	43 min
H1.4+dsDNA, H5+dsDNA (DEB)	(2-46)%	43 min
HeLa, *E. coli*	(8-44)%	104 min
HeLa BRP	(8-44)%	44 min
70S ribosome	(5-44)%	43 min
Native nucleosome	(5-48)%	43 min
FA control experiments	(5-44)%	43 min

3.4.2 MS acquisition parameters

Measurements were performed on a Q Exactive HF Hybrid Quadrupole-Orbitrap, Orbitrap Fusion Tribrid and Orbitrap Fusion Lumos Tribrid instruments. Electrospray ionization (ESI) was used with the nano-HPLC system, which was directly coupled to the ESI source. ESI was used with 2300-2400 V ionization voltage, ion transfer tube was heated to 275 °C. All measurements were recorded in a data dependent mode. MS2 spectra were recorded from 105 *m/z* or 110 *m/z*, HCD fragmentation was used in every MS/MS experiment.

Table 3.4-2: MS and MS/MS acquisition parameters for HPLC-MS/MS acquisition

The used abbreviations in the table are as follows: R: Resolution, ACG: Automatic Gain control, *z*: charge state, IT: injection time, NCE: normalized collision energy.

Sample name(s)	**Instrument type**	**Acquisition Method**	**MS1 scan (R/ACG/IT/mass range)**	**Precursor charge state selection**	**Isolation window/(*m/z*)**	**Dynamic exclusion**	**MS2 scan (R/ACG/IT/NCE)**
Mononucleosome (DEB, FA), Mononucleosome+H1.4 (DEB)	Q-Exactive HF	Top30	120.000/1E6/60 ms/200-2000 *m/z*	+(2-8)	1.6	9s	60.000/2E5/120ms/30%
Oligonucleosome (DEB)	Q-Exactive HF	Top30	120.000/1E6/60 ms/350-1600 *m/z*	+(2-8)	1.6	9s	60.000/2E5/120ms/30%
H1.4+dsDNA (DEB), H5+dsDNA (DEB)	Orbitrap Fusion Lumos	3s TopSpeed	120.000/1E6/50 ms/380-1580 *m/z*	+(2-6)	1.4	10s	30.000/1E5/128ms/25%
HeLa, *E. coli*	Orbitrap Fusion	3s TopSpeed	12.000/1E6/50 ms/350-1580 *m/z*	+(2-6)	1.4	20s	30.000/1E5/80ms/(z:+2-36%, z:+(3-6)-34%)
HeLa BRP	Orbitrap Fusion	3s TopSpeed	12.000/1E6/50 ms/350-1580 *m/z*	+(2-6)	1.2	10s	30.000/5E4/120ms/35%
70S ribosome	Orbitrap Fusion Lumos	3s TopSpeed	12.000/1E6/50 ms/350-1580 *m/z*	+(2-6)	1.4	20s	30.000/5E4/80ms/32%
Native nucleosome	Orbitrap Fusion	3s TopSpeed	12.000/1E6/50 ms/350-1580 *m/z*	+(2-6)	1.4	10s	30.000/5E4/100ms/28%
FA control experiments	Orbitrap Fusion Lumos	3s TopSpeed	12.000/1E6/50 ms/350-1580 m/z	+(2-6)	1.4	20s	

30.000/5E4/80ms/32%

3.4.2.1 Parameters used for MS2/MS3 measurements are as follows:

MS3 acquisition parameters: Top2 most abundant peaks in the MS2 spectra were used as precursor masses in the MS3 acquisition. Isolation window: 1.4 *m/z*, precursor charge state selection: +(2-4) Resolution: 30.000, ACG target 1E5, collision type: CID, Normalized collision energy: 32%, maximum injection time: 128 ms.

3.4.2.2 Parameters for MS2/MS2 measurements as follows:

3.4.2.2.1 HeLa native nucleosomes:

Targeted masses: 112.0511, 136.0623, 152.0572, 124.0511, 148.0623, 164.0572, 161.0324. The targeted masses were searched within the top 10 most abundant peaks in the MS2 spectra, when two of the target masses were present within +/- 25 ppm window, another MS2 spectra was triggered.
Triggered MS2 scan parameters: isolation window: 1.4 *m/z*, HCD collision energy: 34%, Resolution: 60.000, Normalized ACG target: 200%, maximum injection time: 160 ms.

3.4.2.2.2 Formaldehyde control experiments

Target masses of for formaldehyde DNA crosslinked experiments: 112.0511, 136.0623, 152.0572, 124.0511, 148.0623, 164.0572, 161.0324. The targeted masses were searched within the top 10 most abundant peaks in the MS2 spectra. When one of the target masses was present within +/- 8 ppm window, another MS2 spectra was triggered.
Triggered MS2 scan parameters:
Charge state +2: isolation window: 1.6 *m/z*, HCD collision energy: 38%, Resolution: 60.000, Normalized ACG target: 100%, maximum injection time: 160 ms.
Charge state +(3-6): isolation window: 1.4 *m/z*, HCD collision energy: 36%, Resolution: 60.000, Normalized ACG target: 100%, maximum injection time: 160 ms.

3.4.2.2.3 70S ribosome

Target masses of for formaldehyde RNA crosslinked experiments: 124.0511, 148.0623, 164.0572. The targeted masses were searched within the top 10 most abundant peaks in the MS2 spectra. When one of the target masses was present within +/- 8 ppm window, another MS2 spectra was triggered.
Triggered MS2 scan parameters:
Charge state +2: isolation window: 1.6 *m/z*, HCD collision energy: 38%, Resolution: 60.000, Normalized ACG target: 100%, maximum injection time: 160 ms.
Charge state +(3-6): isolation window: 1.4 *m/z*, HCD collision energy: 36%, Resolution: 60.000, Normalized ACG target: 100%, maximum injection time: 160 ms.

3.4.2.2.4 Silica enrichment measurements of *E. coli* and HeLa cells

Targeted masses for formaldehyde RNA crosslinked experiments: 124.0511, 148.0623, 164.0572. The targeted masses were searched within the top 10 most abundant peaks in the MS2 spectra. When one of the target masses was present, within +8/-5 ppm window, another MS2 spectra was triggered.
Triggered MS2 scan parameters:

Charge state +2: isolation window: 1.4 *m/z*, HCD collision energy: 39%, Resolution: 60.000, Normalized ACG target: 220%, maximum injection time: 160 ms.
Charge state +(3-6): isolation window: 1.4 *m/z*, HCD collision energy: 37%, Resolution: 60.000, Normalized ACG target: 220%, maximum injection time: 160 ms.

3.4.2.2.5 HeLa silica enrichment sample with BRP fractionation targeted parameters:

Targeted masses for formaldehyde RNA crosslinked experiments: 124.0511, 148.0623, 164.0572. The targeted masses were searched within the top 10 most abundant peaks in the MS2 spectra. When one of the target masses was present within +10/-5 ppm window, another MS2 spectra was triggered.
Triggered MS2 scan parameters:
Charge state +2: isolation window: 1.2 *m/z*, HCD collision energy: 39%, Resolution: 60.000, Normalized ACG target: 220%, maximum injection time: 160 ms
Charge state +(3-5): isolation window: 1.2 *m/z*, HCD collision energy: 37%, Resolution: 60.000, Normalized ACG target: 100%, maximum injection time: 160 ms.

4 Results

4.1 DEB crosslinking of protein-DNA complexes

4.1.1 Proposed MS/MS fragmentation pattern of the DEB crosslinked peptide- DNA heteroconjugates

In the case of protein-nucleic acid crosslinking, the complexity of MS/MS spectra arises from the fact that both peptides and deoxynucleotides can fragment at the same time. As a result, several combinations of fragments fragment ion masses need to be taken into consideration, which complicates spectral annotation. Fortunately, fragmentation of DEB crosslinked peptide-DNA heteroconjugates occurs in a simpler way. All four deoxynucleotides can participate in the DEB crosslinking reaction. Generally, dA, dG and dC crosslinks show similar fragmentation pattern, while dT heteroconjugates exhibit distinct fragmentation based on the observation of the current study.

The DEB crosslinking reaction between guanine and proteins has been investigated previously by Michaelson-Richie *et al.* [45], although detailed MS/MS fragmentation pattern was not described in their publication. Michaelson-Richie *et al*. have used an important feature of the DEB crosslinked DNA. At elevated temperatures, the N-glycosylic bond of purine bases is cleaved, releasing adenine and guanine, this reaction is called depurination. When DNA is crosslinked to a protein, depurination can be induced, which leads to release of the purine base together with the crosslinked protein from the DNA backbone. This results in a DEB-crosslinked guanine (or adenine) on the proteins. Probably, this thermal reaction also happens during mass spectrometric measurement on the peptide-DNA heteroconjugates, although this hypothesis was not proved. In any case, DEB crosslinks to guanine were very prominently observed in the current study in every DEB crosslinking dataset.

4.1.1.1 Crosslinks to guanine

MS/MS fragmentation of DEB-crosslinked peptide-DNA heteroconjugates was first investigated with guanine in the current study, as crosslinks to guanine were most frequently observed. The MS/MS fragmentation predominantly occurs at the DEB linker between deoxynucleotides and peptides. A possible common fragmentation is exemplified with a DEB crosslinked lysine-guanine heteroconjugate.

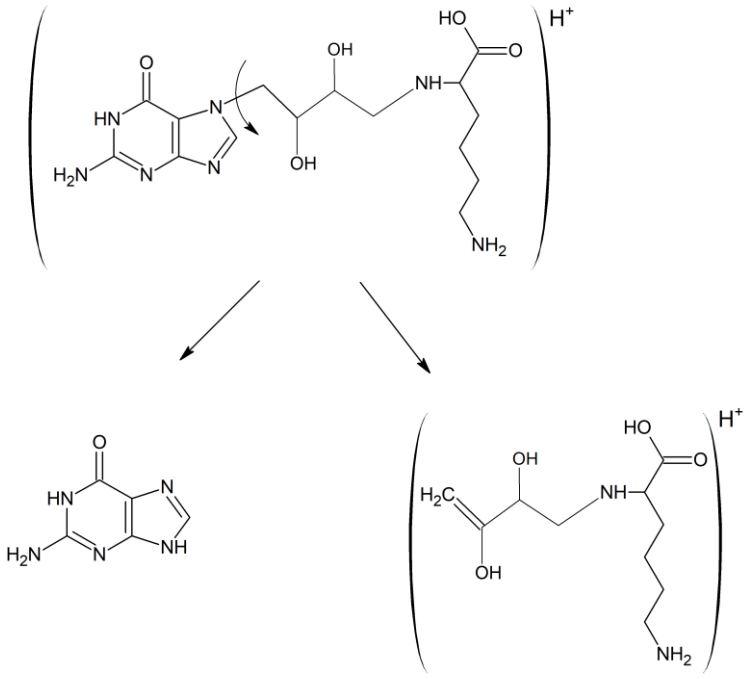


Figure 4.1-1: Proposed fragmentation of the DEB-linker in the DEB-crosslinked lysine guanine heteroconjugate

In the proposed DEB linker fragmentation of the lysine-guanine heteroconjugate, the DEB linker cleaves off from the guanine and stays on the lysine amino acid

One possible fragmentation event is the bond cleavage between N7 position of the guanine and the carbon atom of the DEB linker (Figure 4.1-1). The amino acid carrying the residue of the linker, for example lysine in Figure 4.1-1, this residual DEB linker is later referred as +DEB shift (86.0367 Da). When peptides are DEB crosslinked not only to a guanine base but to deoxynucleotides, another prominent MS/MS fragmentation feature of the peptide-DNA crosslinks is the N-glycosylic bond cleavage between the nucleobase and the deoxyribose. In those cases, there are two fragmentation possibilities. If the DEB linker doesn't fragment, the nucleobase will remain attached to the peptide, this will be referred as +DEB+Gb shift (237.0862 Da) further on. If the DEB linker cleaves between peptides and deoxynucleotides, the previously discussed +DEB modification appears on the peptides. Additionally, peptides usually also fragment through the peptide backbone, resulting b and y peptide fragment ions, as well as +DEB shifted and/or +DEB+Gb mass shifted peptide fragment ions.

In this result section, example MS2 spectra are shown to illustrate the above-described MS/MS fragmentation patterns. Example spectra were chosen from the DEB crosslinked H1.4 ds DNA complex and the DEB crosslinked mononucleosome complex (Material and Methods section 3.2.4 and section 3.2.5). In each case, annotation of the most abundant peaks is shown, to avoid peak annotation crowding.
Example spectrum was chosen to illustrate the DEB linker fragmentation between peptides and guanine. MS/MS spectrum of the DEB crosslinked peptide, EIAQDFK of the H3 histone protein was chosen (Figure 4.1-2). The peptide is crosslinked to guanine.

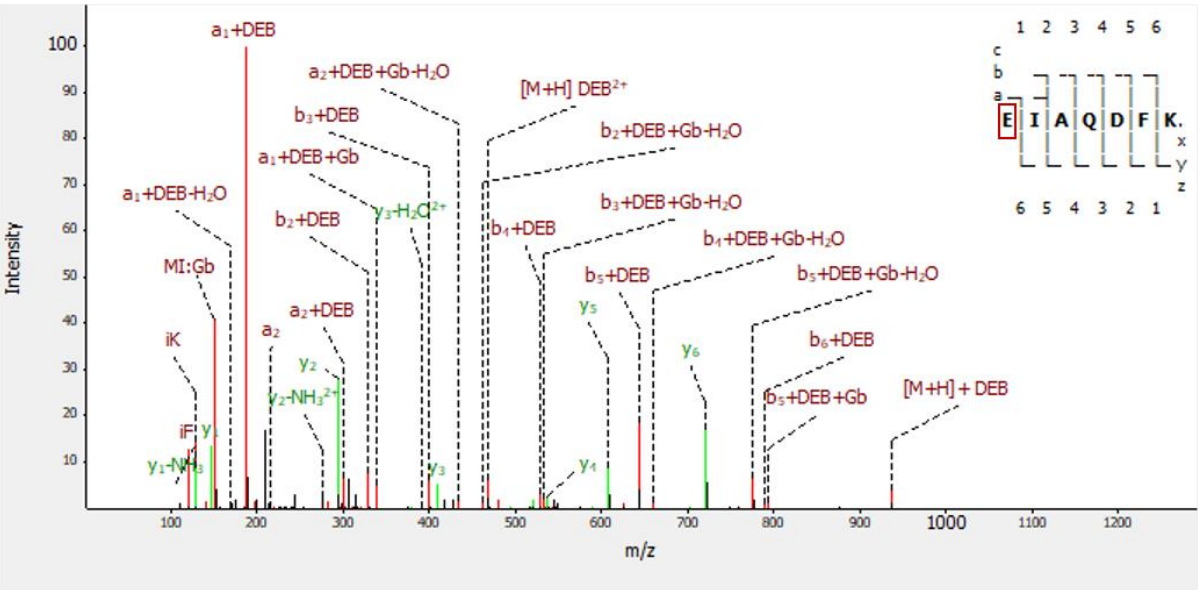


Figure 4.1-2: MS/MS spectrum of H3 peptide, EIAQDFK crosslinked to guanine

Crosslinked amino acid is shown in red rectangle, peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

In the lower *m/z* region in Figure 4.1-2, the guanine, Gb marker ion (152.0572 Da) is present, followed by the most abundant peak in the spectrum, DEB shifted a1 ion. This indicates that the crosslinked amino acid is glutamic acid in the a1 position. In the spectrum, DEB mass shifted series are present, containing a1, a2, b2, b3, b4, b5 and b6 ions, as well as the peptide precursor (M). DEB+Gb-H_2O shifted series is also present, containing a2, b2, b3, b4, b5 ions. All identified fragment ions support the crosslink site localization on the glutamic acid.

4.1.1.2 Crosslinks to adenine

In the case of adenine crosslinks in the DNA adduct, a very similar fragmentation was observed as the guanine crosslinks. DEB and DEB+Ab (221.0913 Da) mass shifted peaks are present in the spectrum (Figure 4.1-3). The first DEB mass shifted ion is y5, followed by extensive DEB mass shifted ion series from y5 to y9. DEB+Ab mass shifted ions can be identified as well - y8, y9 and y10. Some fragmentation ions are represented by both DEB and DEB+Ab mass shifted fragment ions (e.g.: y8 and y9). Due to the higher degree of fragmentation, resulting in DEB mass shifted fragment ion series, in comparison to DEB+Ab, considerably intense Ab marker ion is present at the low *m/z* region.

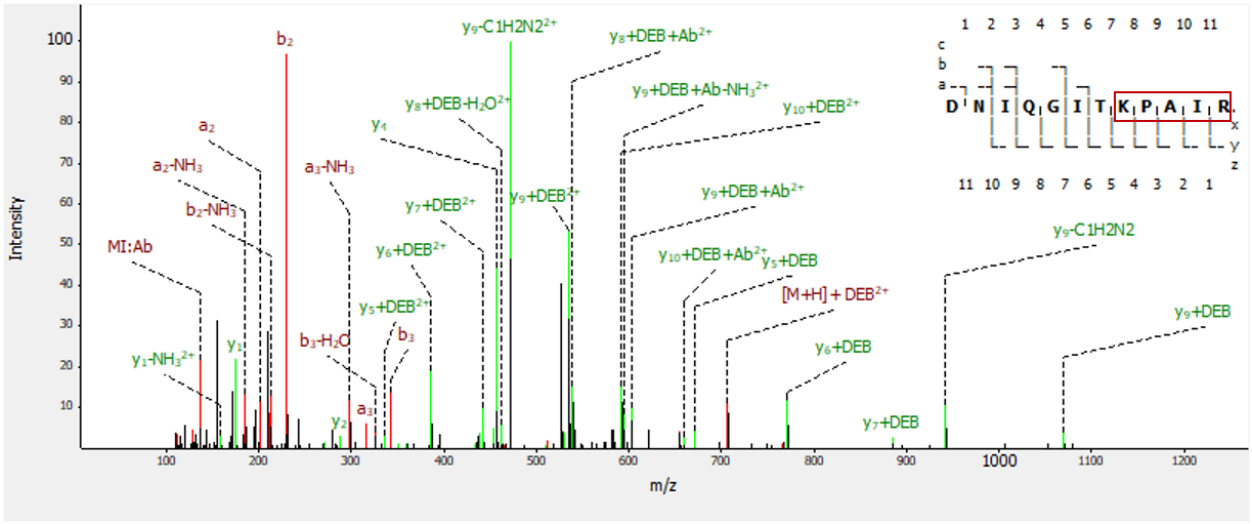


Figure 4.1-3:MS/MS spectrum of H4 peptide, DNIQGITKPAIR crosslinked to deoxyadenosine monophosphate

Crosslink site localization is shown in red rectangle, Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

Taking a closer look at the spectrum, there is an intense y4 fragment ion peak, this might indicate that crosslink is localized on this, particular lysine. Due to the lack of mass shifted b ion series, the crosslink site cannot be localized to a single amino acid with absolute certainty, therefore the crosslink window can be localized at KPAIR amino acids between y5 and y1 positions.

4.1.1.3 Crosslinks to cytosine

In the case of DEB crosslinking, cytosine crosslinks were observed to lesser extent. The observed MS/MS fragmentation pattern is similar to the previously described MS/MS fragmentation, when guanine and adenine crosslinks were present. MS2 spectra was chosen (Figure 4.1-4) for the illustration for the fragmentation of the DEB crosslinks to cytosine, crosslinked peptide of the H3 histone protein, EIAQDFKTDLR. The peptide is crosslinked to dC.

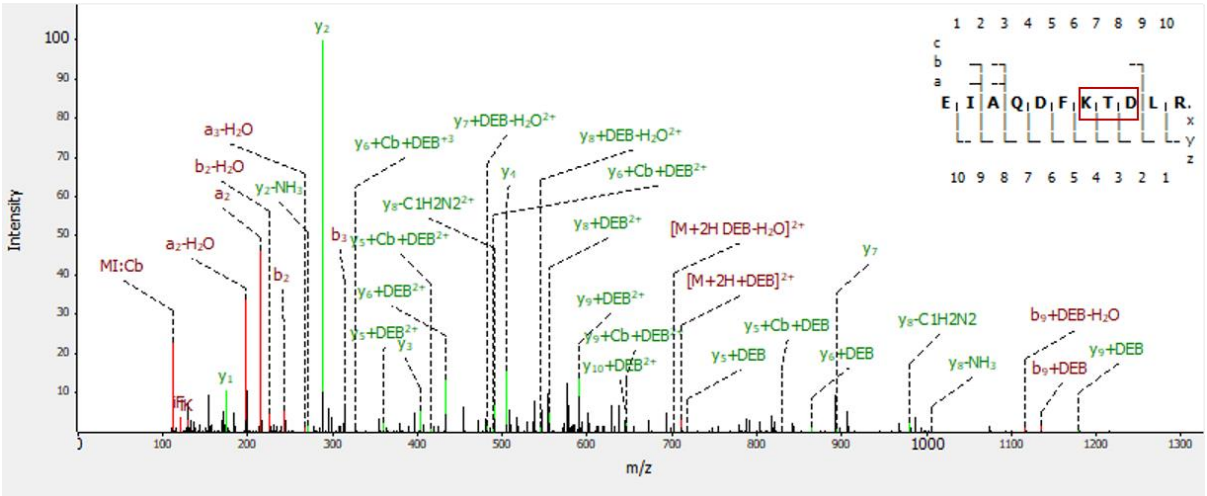


Figure 4.1-4: MS/MS spectrum of H3 peptide, EIAQDFKTDLR crosslinked to deoxycytidine monophosphate

Crosslinked site localization is shown in red rectangle, Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

DEB mass shifted ion series are present from y5 to y10. DEB+Cb (197.0800 Da) mass shifted ions are also present, such as y5 and y6. One DEB mass shifted b ion is present, b9. This narrows down the crosslinking site localization to three amino acids, marked in the red rectangle in Figure 4.1-4. Due to the fact that the majority of mass shifted ions are DEB adducts, the complementary Cb marker ion is present in the low *m/z* region.

4.1.1.4 Crosslinks to thymine

In the case of thymine crosslinks, not only DEB and DEB+Tb (212.0797 Da) mass shifts are present, but also DEB+T (408.0934 Da) – a mass shift of the whole deoxy thymidine monophosphate. This extra shift is different in its fragmentation pattern from the other three deoxynucleotide crosslinks. A shift of the mass of the deoxynucleotide was not observed for other deoxynucleotides, as cleavage of the N-glycosylic bond was commonly induced at all other nucleobases. An example spectrum for the fragmentation is shown in Figure 4.1-5, MS2 spectra of the histone H3 peptide, YQKSTELLIR crosslinked to dT.
Mass shifted fragment ion series are present, starting from b3 and containing a5, a6, b6, b7 and b8 ions. This implies crosslink localization on the lysine. The complementary mass shifted ions: y8 and y9 also support the crosslink localization on lysine.

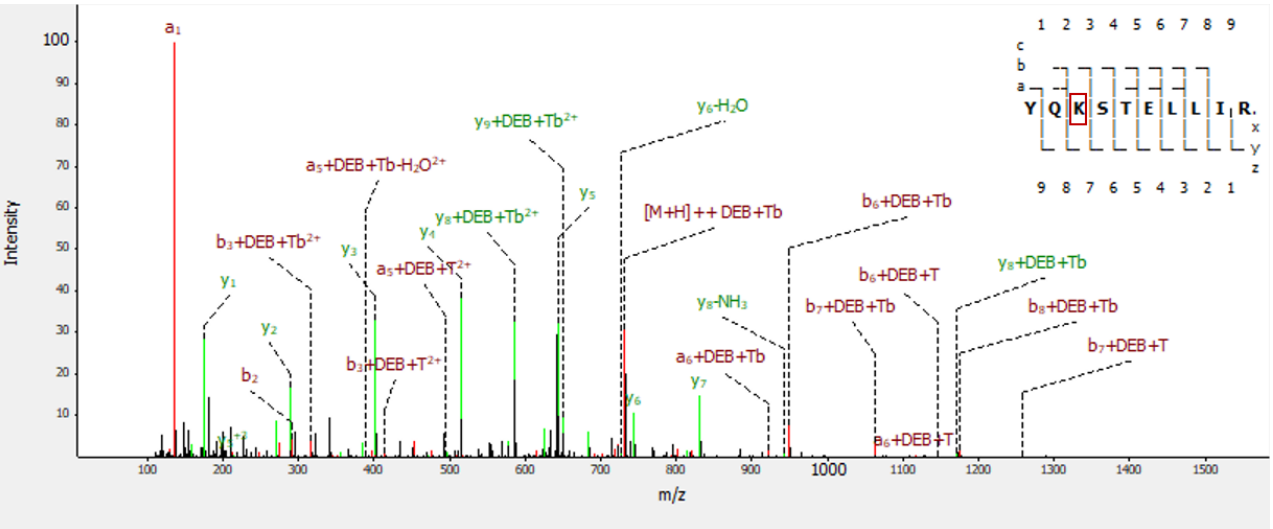


Figure 4.1-5: MS/MS spectrum of H3 peptide, YQKSTELLIR crosslinked to deoxythymidine monophospahte

Crosslinked amino acid is shown in red rectangle, peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

Due to the complementary mechanism of generation, when the majority of peaks are shifted with DEB+T or DEB+Tb, then no or very low intensity marker ions of the crosslinked deoxynucleotide can be expected in the spectrum, as it is shown in Figure 4.1-5.
To fully demonstrate the complexity of a seemingly less liable N-glycosylic bond in thymine cross-links, an additional example fragment ion spectrum was chosen to show the fragmentation of DEB crosslinked thymidine monophosphate.

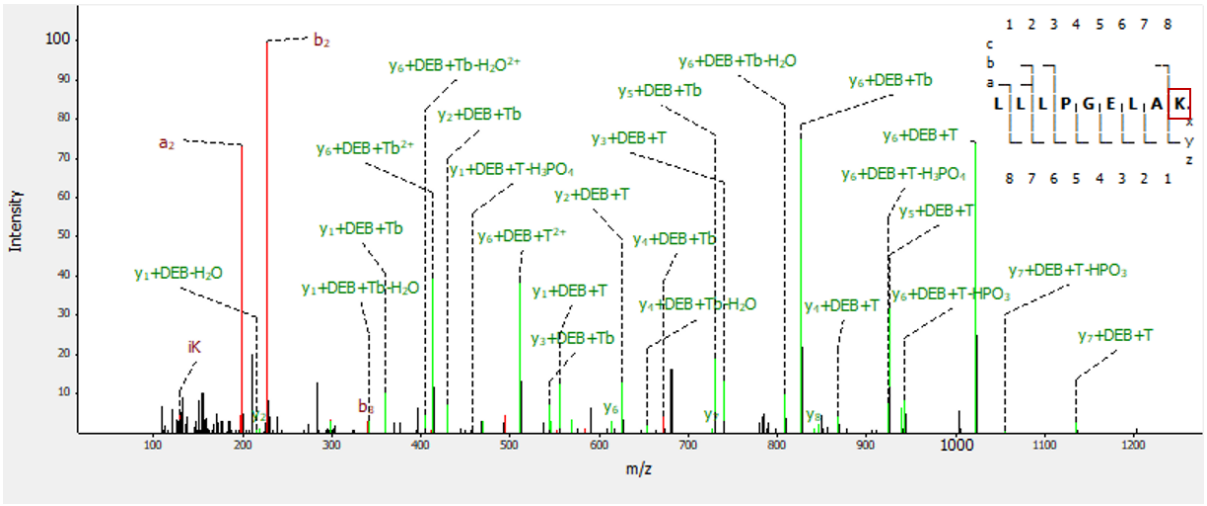


Figure 4.1-6: MS/MS spectrum of H2B peptide, LLLPGELAK crosslinked to deoxythymidine monophosphate

Crosslinked amino acid is shown in red rectangle**,** Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

As it is shown in Figure 4.1-6 the majority of the peptide peaks are mass shifted. Six different mass adducts are present: DEB, DEB+Tb, DEB+Tb-H_2O (194.0691 Da), DEB+T, DEB+T-HPO_3 (328.1271 Da), DEB+T-H_3PO_4 (310.1165 Da), shifting product ion series from y1 to y7. This fragmentation resembles to the UV crosslinked deoxythymidine monophosphate fragmentation [57].
Overall, the above outlined results show that DEB crosslinking can be used for all four-deoxynucleotide crosslink identification.

4.1.1.5 Localization of the crosslinking site

Localization of the crosslinking site in the peptide is possible in most of the cases with DEB crosslinking, owing to the low probability of the linker to be cleaved off from the peptide during fragmentation in the gas phase and therefore it marks the position of the crosslink site. As it was shown previously, localization to a single amino acid is only possible, if mass shifted b and y ion series are present.
A particular challenge when trying to pinpoint the site of the crosslink occurs when isobaric heteroconjugates are present. Isobaric crosslinks have the same precursor mass but different structure, often having similar physicochemical properties. Thus, it is not possible to separate them either by liquid chromatography or on MS1 level. They are analyzed together, which results in a mixed MS/MS spectrum, having fragments from both heteroconjugates. The crosslinking site localization is controversial.

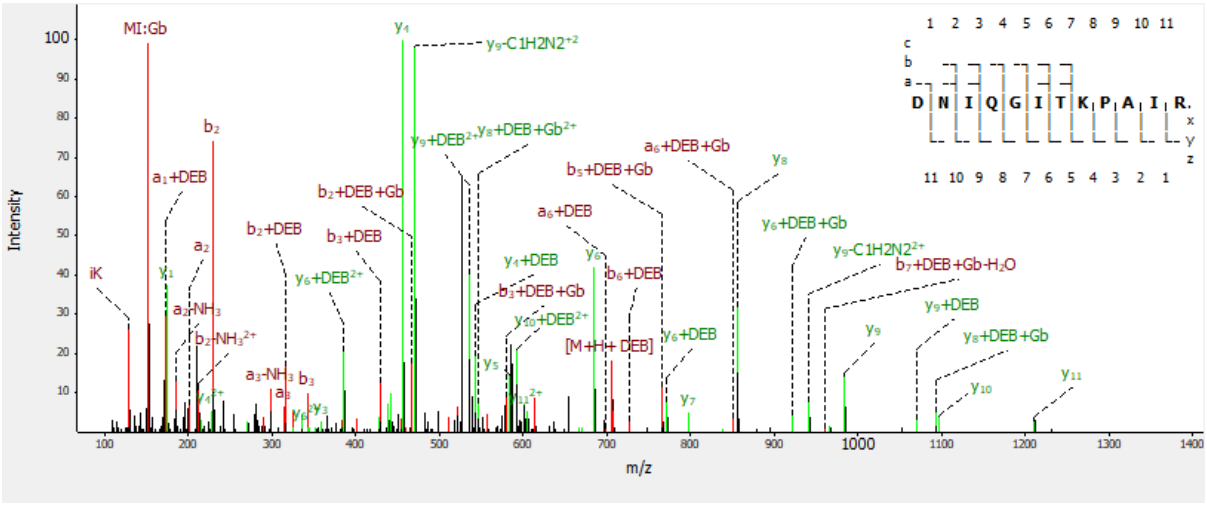


Figure 4.1-7: MS/MS spectrum of H4 peptide, DNIQGITKPAIR crosslinked to deoxyguanosine monophosphate

Due to the MS/MS spectrum is a mixture spectrum of isobaric peptides, no confident crosslink localization is possible**.** Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

A good example is the crosslinked peptide DNIQGITKPAIR of the H4 histone protein, linked to deoxyguanosine monophosphate (Figure 4.1-7). DEB mass shifted ion series are present from a1 to b3, which indicates the crosslinked site on glutamine at the b1 position in the sequence. At the same time, DEB mass shifted ions y4 and y6 are also present. This indicates that the crosslinking site is in the y1-y4 region of the sequence. These mass shifted ions can only be present at the same time, if they belong to two different crosslinking sites in the same peptide, resulting in a mixture MS/MS spectrum. This is due to coelution during liquid chromatography of the two crosslinked peptide variants, with different crosslinking sites.

4.1.1.6 Length of deoxynucleotides and charge state effects MS/MS fragmentation

4.1.1.6.1 The effect of the deoxy nucleotide length

The previously shown examples are all crosslinks that contain only a single crosslinked deoxynucleotide or nucleobase; however, DNA adducts can contain more deoxynucleotides y57. Two parameters have been identified to have major effect on the fragmentation: the length of the crosslinked deoxynucleotide chain and the charge state of the crosslink.
In this section, three exemplary spectra are shown. All three examples have the same crosslinked peptide but differ in the length of the crosslinked deoxynucleotide chain and the charge state of the crosslink.
For reference, a MS/MS spectrum of a crosslinked peptide (GTGASGSFKLNK peptide of H1.4 histone protein) is shown, where only a guanine base is crosslinked. Shifted y ion series are present from y5 to y10 and complementary shifted ion series is present from b9 to b11. This marks the crosslink site to the lysine at the y4 position Figure 4.1-8.

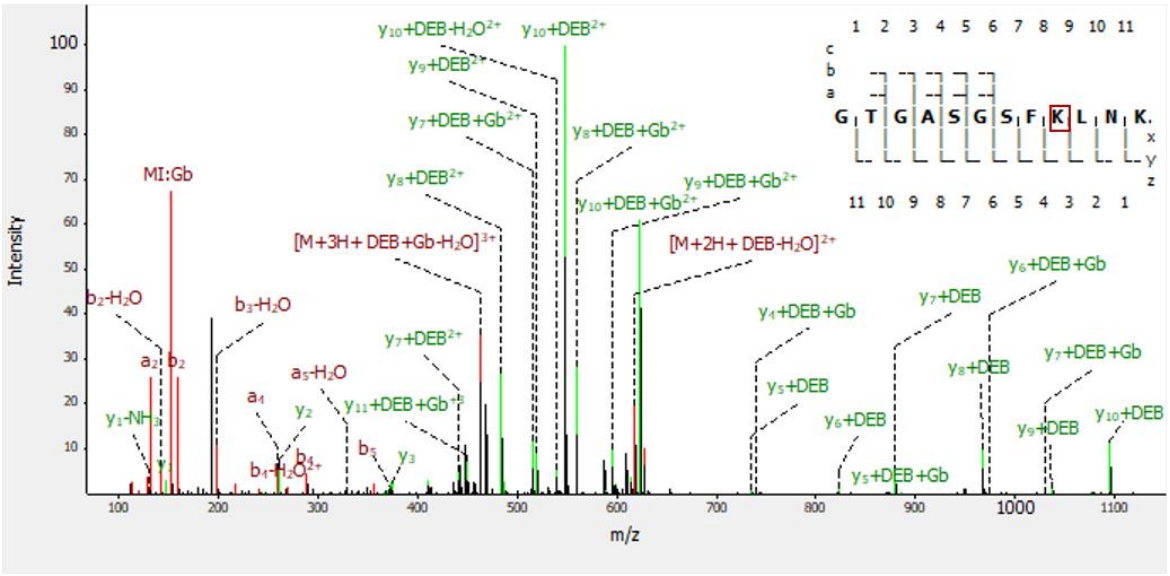


Figure 4.1-8: Figure: MS/MS spectrum of H1.4 peptide, GTGASGSFKLNK crosslinked to guanine

Crosslinked peptide is localized in the red rectangle, peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

The MS/MS fragmentation pattern is significantly different when the same peptide is crosslinked to the dinucleotide dGdG (Figure 4.1-9). When more than one deoxynucleotide is present, the internal energy is spent on bond cleavages between deoxynucleotides in the dinucleotide stretch as well as for the N-glycosylic bond between the nucleobase and the deoxyribose. Therefore, during fragmentation, one guanine stays attached to the intact peptide ($[M+2H+DEB+Gb]^{2+}$ peak on Figure 4.1-9), one guanine appears as marker ions in the low *m/z* region (Gb, G). Another prominent fragment ion is the intact peptide mass shifted with DEB ($[M+2H+DEB]^{2+}$). Fragment ions derived from peptide are of low intensity compared to the high intensity DEB and DEB+Gb mass shifted peptide peaks, thus, a confident crosslinking site localization is not possible.

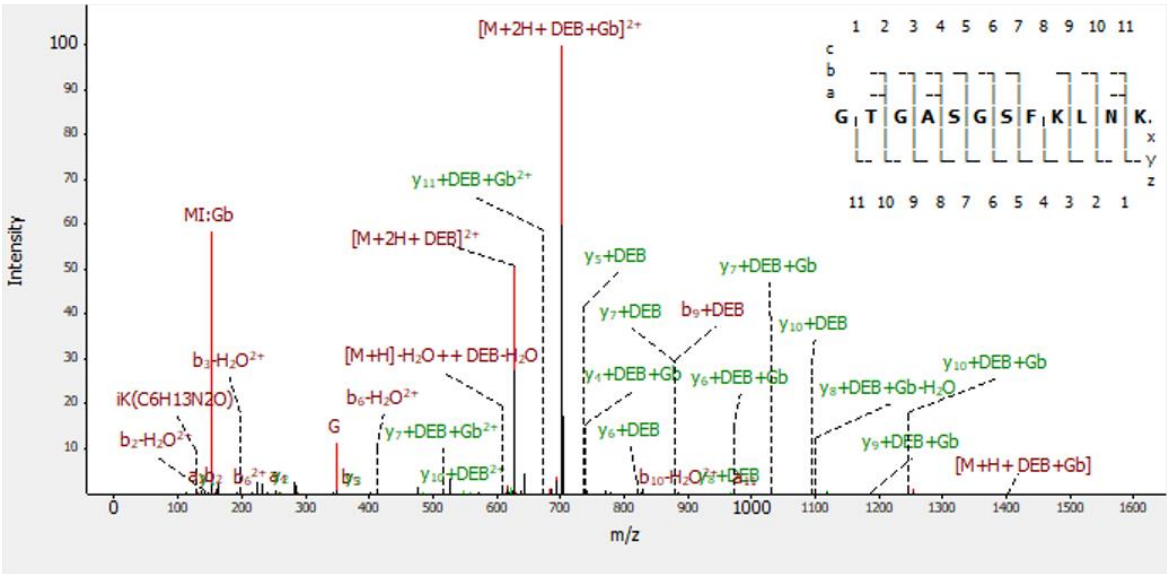


Figure 4.1-9: MS/MS spectrum of H1.4 peptide, GTGASGSFKLNK crosslinked to dGdG dinucleotide

Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green. Due to the low intensity of peptide fragments no confident localization was possible.

4.1.1.6.2 The effect of the charge state of the peptide-DNA heteroconjugates

The charge state of the crosslinked heteroconjugate has as significant effect on the fragmentation pattern. The previously shown crosslinked peptide, GTGASGSFKLNK, crosslinked to guanine has a charge state of +3 (Figure 4.1-8). The same crosslink with charge state of +2 has a different fragmentation pattern, as it is shown on Figure 4.1-10.

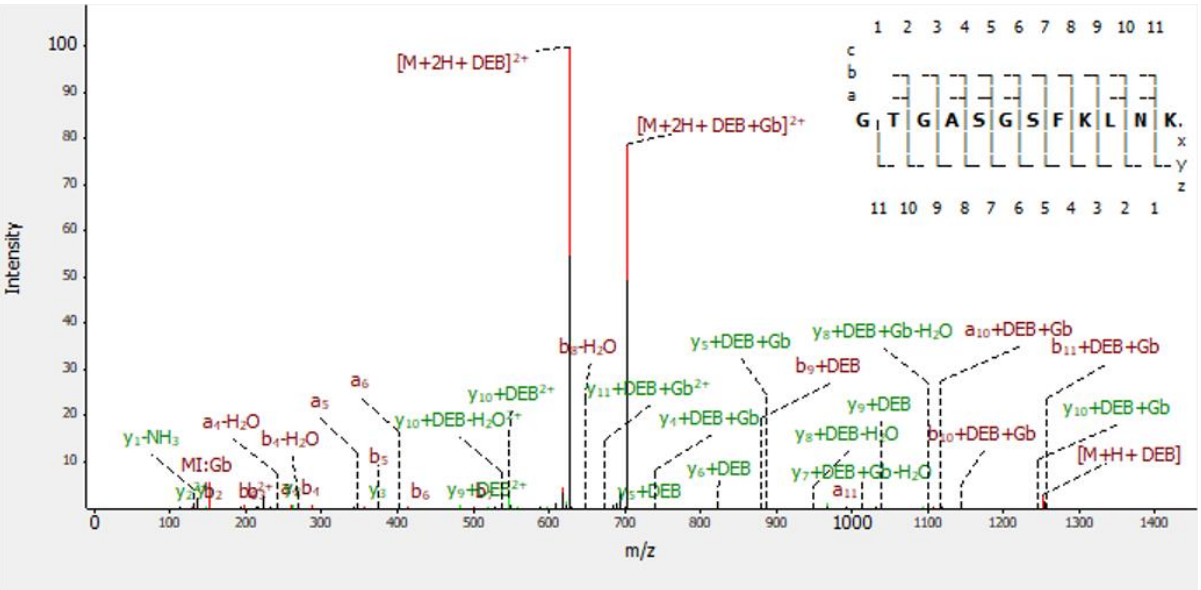


Figure 4.1-10: MS/MS spectrum of H1.4 peptide, GTGASGSFKLNK crosslinked to guanine, charge state +2

Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green. Due to the low intensity of peptide fragments no confident localization was possible.

The peptide is crosslinked to guanine, charge state of +2, Fragment ions derived from peptide backbone fragmentation are of low intensity compared to the high intensity DEB and DEB+Gb mass shifted peptide peaks, thus, a confident crosslinking site localization is not possible. The observed fragmentation patterns in Figure 4.1-9 and Figure 4.1-10 are very similar. Two characteristic peaks of intact peptide+DEB and peptide+DEB+Gb are present in both spectra. There are slight differences between the two spectra: when the peptide is crosslinked to guanine, the intensity of guanine marker ion is lower as compared with the marker ion present in the MS/MS fragmentation spectrum of the peptide crosslinked to dGdG. The most abundant peak in the guanine crosslinked spectrum is $[M+2H+DEB]^{2+}$, which is in good agreement with the low marker ion intensity. The most abundant peak of the dGdG crosslink is $[M+2H+DEB+Gb]^{2+}$.

4.1.2 MS2/MS3 acquisition method

The aim of applying a dedicated MS2/MS3 acquisition mass spectrometric workflow was not only to understand crosslinking fragmentation and to identify crosslinks, but also to establish a general method for convenient identification of peptide-DNA crosslinks. The crosslinked deoxynucleotide length varies from one up to five deoxynucleotides in length, which has a significant effect on the fragmentation.

As previously discussed, often an incomplete fragmentation is observed, where intense peptide+DEB and peptide+DEB+Gb fragment ions are generated but the intensity of other fragment ions is low. Therefore, this type of MS/MS fragmentation seemed advantageous to

reduce the size of the deoxynucleotide moiety via gas phase fragmentation. In line with that thought, it was hypothesized that if all spectra would have similar fragmentation as Figure 4.1-9 and Figure 4.1-10, the complexity of the MS/MS fragmentation of the heteroconjugates and the complexity of the data analysis could be reduced.
The occurrence of Peptide+DEB and peptide+DEB+Gb ions in MS2 can be used for another round of fragmentation in which Peptide+DEB and Peptide + DEB + Gb are selected. With an additional fragmentation (MS3) the sequence coverage of the crosslinked peptide by producing more intense product ions is improved and the crosslinking site localization can be achieved. The acquisition method is shown on the example of guanine crosslinks for which it was originally developed. The final goal is to create an automated workflow, where laborious manual spectra annotation could be avoided.

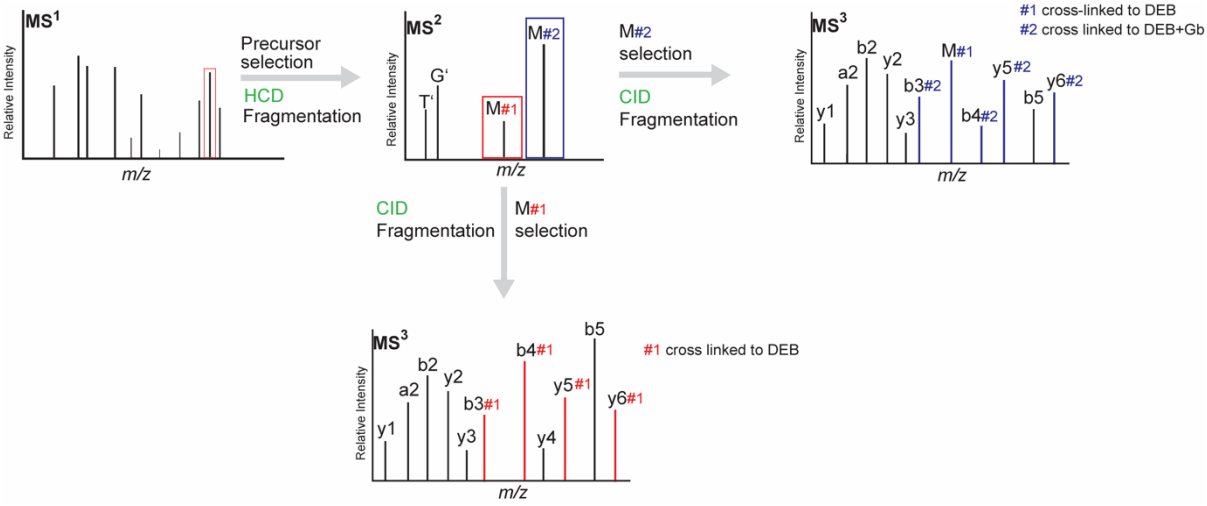


Figure 4.1-11: Illustration of the MS2/MS3 acquisition method, established in DEB crosslinking

HCD fragmentation is applied on the MS2 level, a lower collision energy is used to ensure that the preferred fragmentation of the crosslinked deoxynucleotide only. The intact peptide +DEB and the intact peptide+DEB+Gb (M#2 on Figure 4.1-11) are used as a precursor for another round of fragmentation, which are typically the two most abundant peaks in the MS2 spectra when lower collision energies are used and the heteroconjugate is guanine crosslink. CID fragmentation is used in this manner at higher values in the next stage to ensure the peptide backbone fragments. In the resulting MS3 spectra b and y ion series and b and y ion series shifted with DEB (#1 shift on Figure 4.1-11) or DEB+Gb (#2 shift on Figure 4.1-11) are present. Generally, fragmentation in the MS3 spectra is very simple. As a result, the peptide sequence and crosslinking site identifications happen on the MS3 level and the determination of the composition of the crosslinked deoxynucleotide(s) happens on the MS^1 level. MS2 level is used for the fragmentation for the crosslinked deoxynucleotides so that only the intact peptide plus DEB or the intact peptide plus DEB plus Gb remains, which can then be selected for an additional fragmentation (MS3). The only exception from this MS2/MS3 workflow, is when the peptides are crosslinked to guanine base only; then crosslink identification is already possible on the MS2 level).
Although the principle of this method is simple, the data analysis poses a challenge due to the fact that proteomics search engines are not tailored for special applications. Most proteomics search engines, as well as the RNP^{xl} pipeline, do not support MS3 level identification (with some exceptions such as MaxQuant or SEQUEST). Thus, it was needed to find a suitable way to analyze the data derived from MS3 fragmentation.

4.1.2.1 Data analysis of the MS3 level

SEQUEST search was performed for the MS3 fragmentation, with set modifications of DEB and DEB+Gb. Pairing up MS3 and MS2 was performed with a use of an R script, designed for this task (see section 3.3.4 and Supplementary Text 1). A schematic figure of the MS2/MS3 method data analysis strategy is shown on Figure 4.1-12.

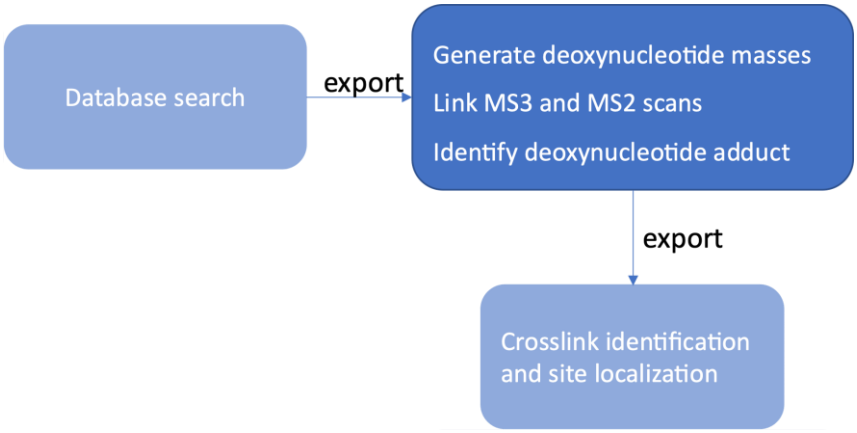


Figure 4.1-12: Modification-search based data analysis strategy for crosslink identification on the MS3 level

The strategy contains the following steps: first, a regular database search is performed on the MS3 level, where DEB, and DEB+Gb are set as variable modifications. A custom R script was used (Supplementary Text 1) to connect the MS levels and to assign the crosslinked deoxynucleotides for MS3 level, utilizing two result tables from the SEQUEST database search, the identification table and the MSMS Spectrum Info table. MSMS Info is a table containing scan numbers of the MS^1, MS2 and MS3 spectra, the *m/z* values and retention times. Based on that information the MS1-MS2-MS3 data can be logically connected. The masses of the deoxynucleotide moiety can be inferred from the mass difference between the precursors of MS2 and MS3 levels. For instance, if a peptide VKLLR is crosslinked with DEB to dG in a dinucleotide of dGdG (VKLLR+DEB+dGdG), and in the MS3 level the precursor was VKLLR+DEB+Gb the mass difference between the precursors of MS2 and MS3 is dGdG-Gb.
To that end, in the designed R script, a list of combination of theoretical (oligo) deoxynucleotide masses is generated with compositions up to five deoxynucleotides. Mass differences between the MS3 and MS2 precursors are calculated and assigned based on the theoretical deoxynucleotide mass list. MS2 and MS3 precursor masses are assigned, and original crosslink composition is linked to MS3 identification. A final results table, containing the identified crosslinks and crosslink site probabilities are exported.
After setting up the workflow for data analysis, the main advantage of the MS2/MS3 method was that the data analysis was automated and laborious manual annotation could be avoided.

4.1.3 General workflow for crosslink identification

In Figure 4.1-13 the general workflow used in the Urlaub laboratory for identification of the crosslinked sites between proteins and nucleic acids is shown. This strategy was originally used for UV-crosslinked protein-RNA complexes (Kramer *et al.* [40]) and later extended to other crosslink samples, including chemically crosslinked protein-DNA complexes.

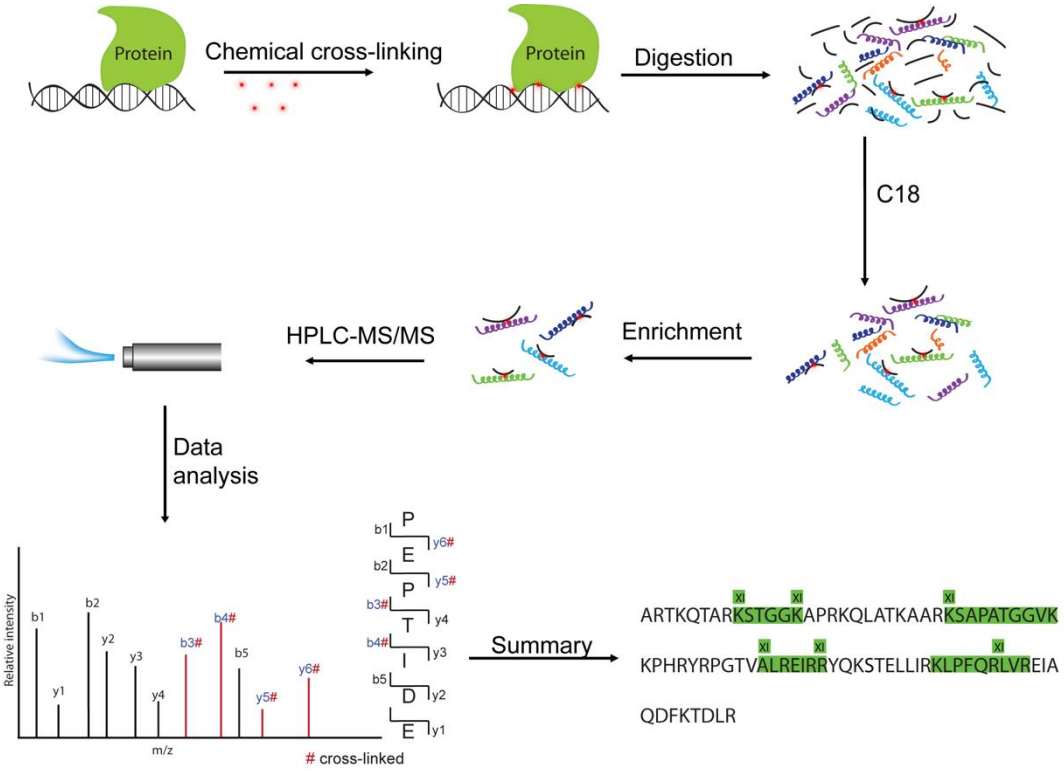


Figure 4.1-13: General workflow for protein-DNA crosslink identification

The first step of the workflow is chemical crosslinking, followed by acetone precipitation of the protein-DNA complexes. The pellet is resuspended in urea and enzymatic digestion takes place: proteins are digested with trypsin protease and nucleic acids are digested with endonucleases - resulting in a mixture of peptides, non-crosslinked (oligo)deoxynucleotides and crosslinked peptide-DNA heteroconjugates. Non-crosslinked (oligo)deoxynucleotides are removed by a C18 reversed-phase column. Crosslinked peptide-DNA heteroconjugates are enriched by TiO_2 affinity purification and then. subjected to LC-MS/MS analysis, followed by data analysis with the RNP^{xl} pipeline of the OpenMS framework [41].
Two complexes were used as model systems for investigation of DEB protein-DNA crosslinking: chicken linker histone H5 and human linker histone H1.4 reconstituted with a 187 bp dsDNA. Linker histones are known to bind to the linker DNA of the nucleosome and regulate the chromatin organization [83], therefore it was the natural choice for a model system. Moreover, a complex of one protein and one DNA is simple enough for thorough investigation of the resulting peptide-DNA crosslinks.

4.1.4 DEB crosslinking of the H1.4–dsDNA complex

H1.4 linker histone and 187 bp ds DNA were used as one of the model systems. After complex formation, DEB crosslinking took place, followed by the general sample preparation workflow. LC-MS/MS measurement took place on the enriched peptide-DNA heteroconjugates, followed by data analysis with two different strategies: modification search with MS2 and MS3 acquisition and RNP^{xl} search after MS2 acquisition.

4.1.4.1 MS2/MS3 strategy- modification search results

More than one third of the identified crosslinks were guanine crosslinks, followed by deoxynucleotide and deoxy(oligo)nucleotide crosslinks derived from dinucleotides with up 5 deoxynucleotide length were detected. Crosslinks only to single deoxynucleotides could not be detected. Crosslink localization was based on the ptmRS search node [78] output of

Proteome Discoverer 2.1, which gives a localization probability value for each site. Identifications based on MS3 fragment spectra were linked to the precursor mass with its crosslinked deoxynucleotide moiety as detected by MS2 with the use of the previously described R script (section 3.3.4). Crosslinking sites (Supplementary Table 1) with indicated site probabilities were mapped onto the protein sequence of linker histone H1.4 (Figure 4.1-14).

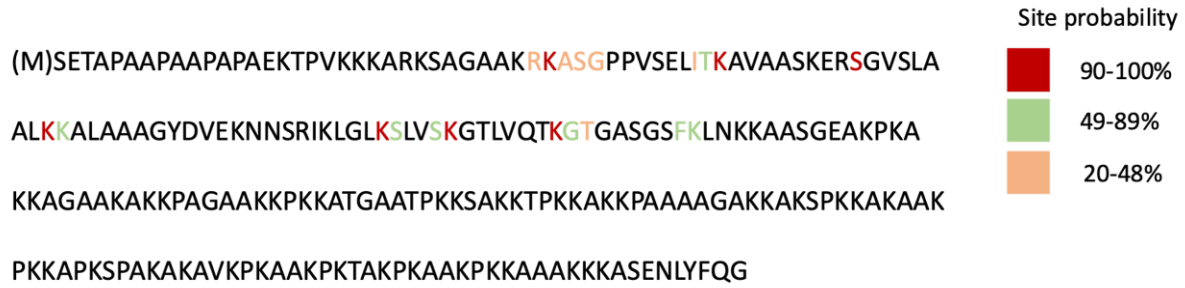


Figure 4.1-14: Crosslinking sites of H1.4 linker histone, based on Sequest HT modification search

Protein sequence is represented with a single letter amino acid code, Crosslinking sites are highlighted in orange green and red. Each color code is related to the crosslinking site probability, calculated in the ptmRS node in proteome discoverer.

4.1.4.2 RNP^{xl} search results

The previously described general method (section 4.1.3) was used for sample preparation and data analysis with the RNP^{xl} tool. As crosslink search is only possible on the MS2 level, the low peptide backbone fragmentation impeded crosslinking site identification (Supplementary Table 2). Thus, only guanine crosslinks were taken into consideration in order to localize the crosslinked amino acid. When thymine crosslinks were present, crosslinking site localization was not considered. In the case of guanine crosslinks complete peptide backbone fragmentation was present in the spectra. Those crosslink sites were mapped onto the H1.4 protein sequence with the use of the crosslink site localization from RNP^{xl}.

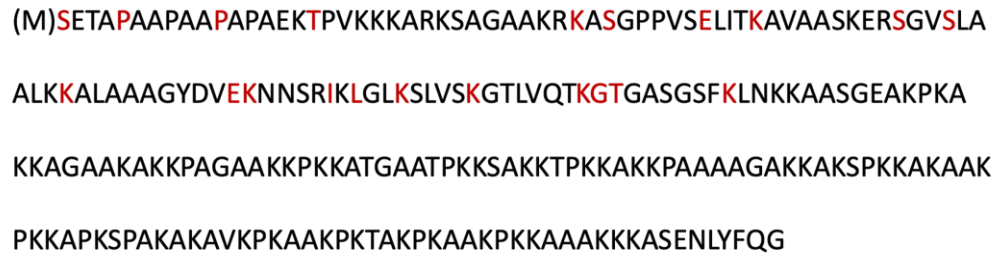


Figure 4.1-15: Crosslinking sites of H1.4 linker histone, based on RNP^{xl} search

Protein sequence is represented with a single letter amino acid code, Crosslinking sites are highlighted in red, based on the RNP^{xl} crosslinking site localization of guanosine crosslinks

Crosslinks to guanine and thymine were identified, while cytosine and adenine crosslinks were not present. Around half of the crosslinks were guanine crosslinks followed by mo, di, tri-and tetranucleotides of different composition. The great number of guanine crosslinks most probably occurs due to the heat liability of guanine crosslinks [84]. Guanine crosslink spectra have the typical DEB fragmentation pattern, where the majority of the peaks are mass shifted with DEB and DEB+Gb. Crosslinks were found in the short N-terminal tail of H1.4, as well as in the H1.4 globular domain. Interestingly, no crosslinks were found in the C-

terminal domain (CTD) of the proteins. One explanation could be that the CTD is rich in lysines and chemical crosslinking highly modifies these lysines, making identification not possible. Manually selected and validated crosslinks of the H1.4 globular domain were mapped onto the cryo-EM structure of H1.4 linker histone-nucleosome complex (pdb: 7K5Y, [85]).

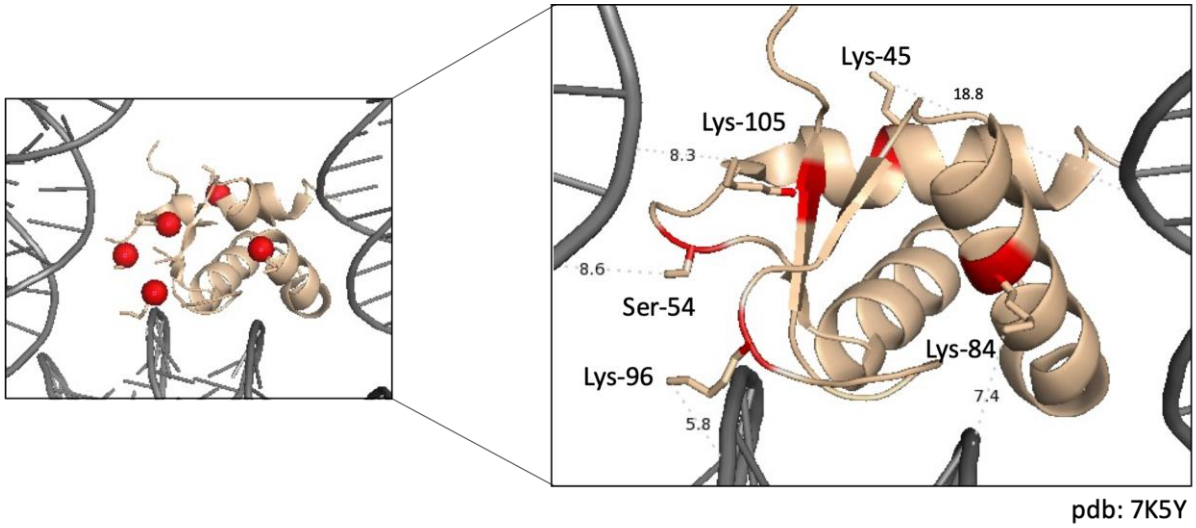


Figure 4.1-16: DEB Crosslinking sites of the H1.4 linker histone in complex with ds DNA in pdb 7K5Y **[**85]

Left side: C-alpha atoms of the crosslinked amino acids are highlighted as red spheres. Right side: measurements (in Ångström) between the crosslinked side chains of Lys-45, Ser-54, Lys-96 and lys-105 and the nearest deoxynucleotides of the DNA.

Five crosslinks of the globular domain were chosen to map the crosslinked amino acids into the structure: Lys-45, Ser-54, Lys84, Lys-96 and Lys-105. Four out of the five crosslinking sites Ser-54, Lys84, Lys-96 and Lys-105 are less than 10 Å from the closest DNA residues in the structure (amino acid numbering refers to the numbering in the structure). One crosslink was found above 10 Å, Lys-45. The higher degree of conformation changes in solution might explain why some of the crosslinks are considerably far from the DNA.

4.1.5 DEB crosslinking of the H5–dsDNA complex

Similarly, H5 linker histone (from *Gallus gallus*) was reconstituted with dsDNA and crosslinked with DEB, followed by general sample preparation for mass spectrometry measurement (Figure 4.1-13). Two data analyses strategies were used for crosslink identification, MS2/MS3 method in combination with modification search and RNP^{xl} search.

4.1.5.1 MS2/MS3 strategy- modification search results

Sequest HT modification search was used for crosslink site identification (section 3.3.3.5). Results are listed in Supplementary Table 3. Similarly, as it was described in the H1.4 section, crosslink site probabilities are shown on Figure 4.1-17.

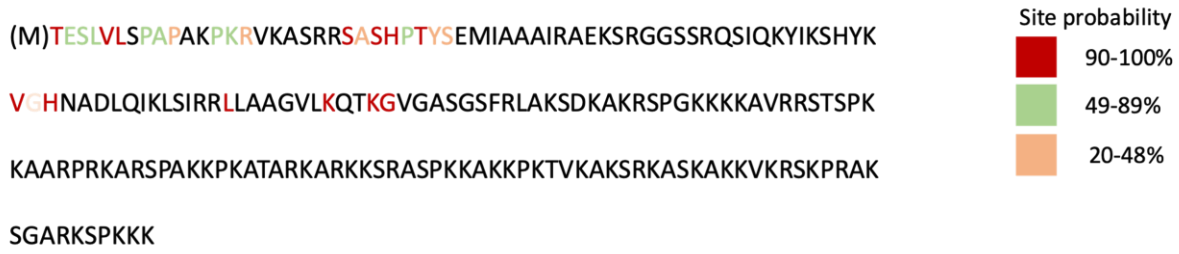


Figure 4.1-17: Crosslinked sites of H5 linker histone, based on Sequest HT modification search

Protein sequence is represented with a single letter amino acid code, crosslinking sites are highlighted in orange, green and red. Each color code is related to the crosslinking site probability, calculated in the ptmRS node in proteome discoverer.

4.1.6 RNP^{xl} search results

In the case of RNP^{xl} search of the MS2 fragmentation data crosslinks to guanine and thymine were found, adenine and cytosine crosslinks could not be identified in the dataset. The composition of (oligo)deoxynucleotides of the identified crosslinks is as follows: significant number of crosslinks are guanine crosslinks (similarly to H1.4), then mo, di, tri and tetranucleotides follow (Supplementary Table 4). Crosslink sites identified by RNP^{xl} were mapped onto the sequence of H5.

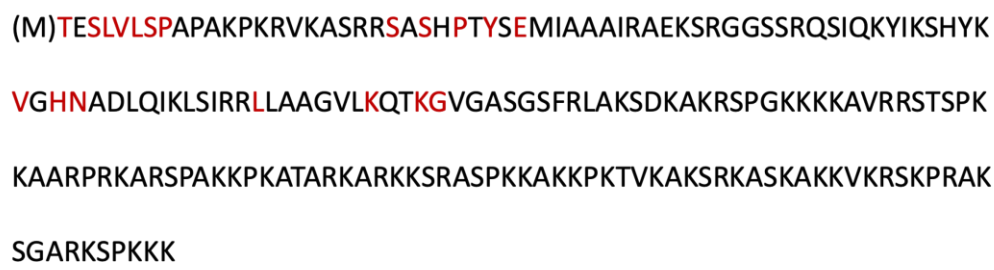


Figure 4.1-18: Crosslinked sites of H5 linker histone, based on RNP^{xl} search

Protein sequence is represented with a single letter amino acid code, Crosslinking sites are highlighted in red, based on the RNP^{xl} crosslinking site localization of guanosine crosslinks

Similarly, to H1.4, crosslinks were found in the N-terminal tail of H5 and the globular domain (Figure 4.1-18). Zhou *et al.* [85] have shown that the overlap between the conformation of a free H5 globular domain and the nucleosome bound GH5 is high, therefore crosslinks found in the GH5 domain were mapped onto the cryo-EM structure of *Xenopus laevis* nucleosome containing GH5 [86] (pdb: 4QLC).

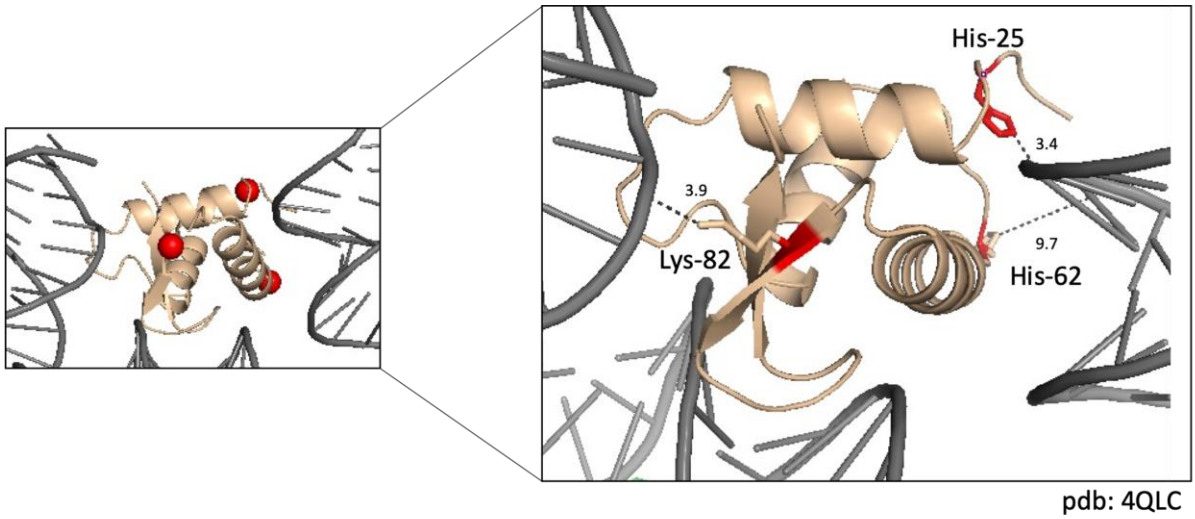


Figure 4.1-19: DEB Crosslinking sites of the H5 linker histone in complex with ds DNA in pdb 4QCL **[**86]

Left side: C-alpha sites of the crosslinked amino acids are highlighted as red spheres. Right side: measurement (in in Ångström) between the crosslinked side chains of His-25, His-62 and Lys-82 and the nearest deoxynucleotides.

Three crosslinking sites were mapped into the structure, His-25, Lys-82 and His-62. Measurements between the side chains of the crosslink sites and the closest deoxynucleotide of the DNA strands were performed - all crosslink sites were found to be in close proximity to the DNA (Figure 4.1-19 right side).

4.1.7 DEB crosslinking of the mononucleosome and mononucleosome containing complexes

4.1.7.1 Modified workflow for crosslink identification

The original workflow (Figure 4.1-13) was modified with the implementation of the SP3 (Single-pot, solid-phase-enhanced sample preparation) [87]; purification that was introduced into the workflow. SP3 allows all sample processing steps in one reaction tube avoiding sample loss during preparation. The modified workflow is executed as follows; steps included in the SP3 sample preparation are highlighted in red (Figure 4.1-20).

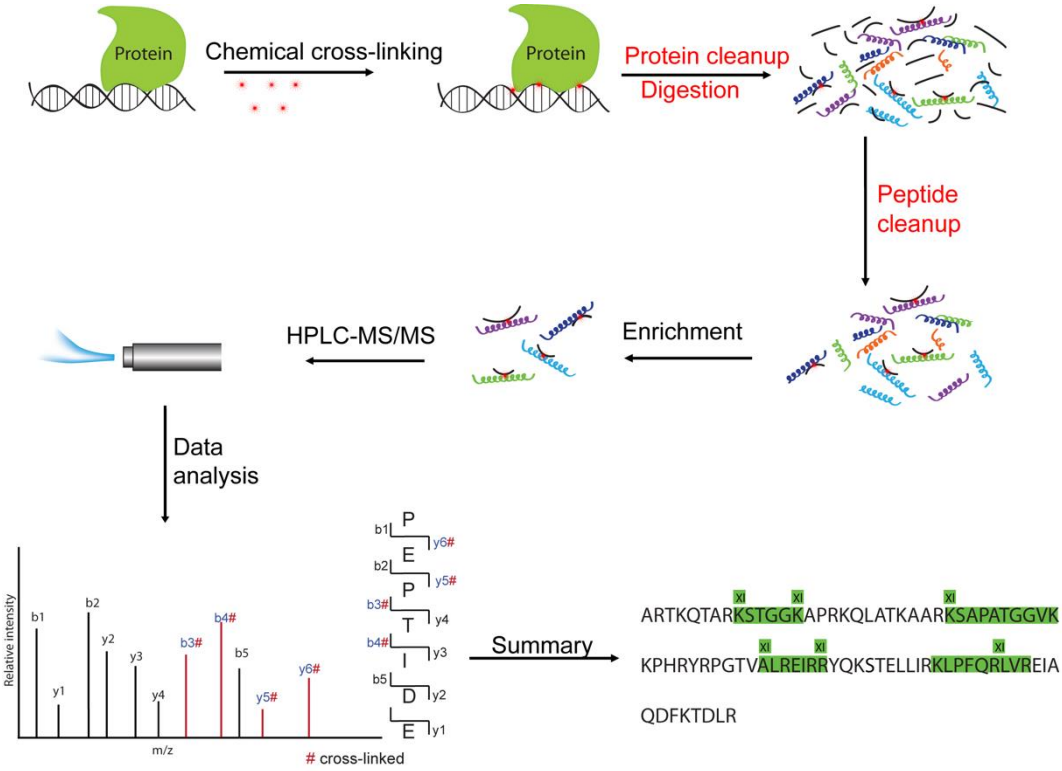


Figure 4.1-20: Single-pot, solid-phase-enhanced sample preparation [87] **integrated into the general crosslinking workflow**

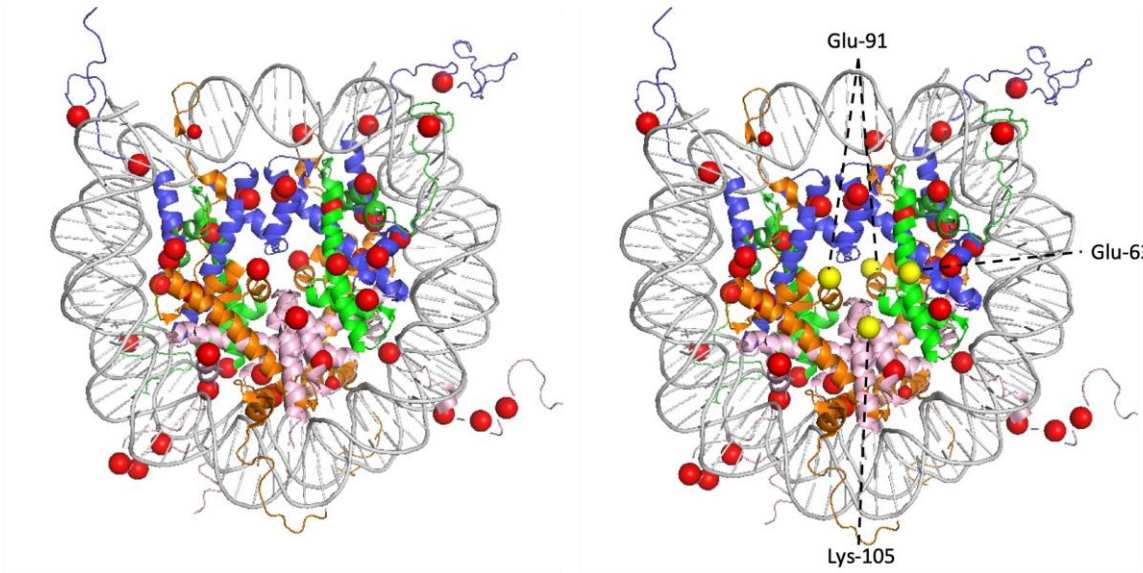
Three steps were performed on the SP3 beads: protein cleanup, protein and nucleic acid digestion and peptide cleanup. These steps are highlighted in red.

After crosslinking, SP3 magnetic beads were used for the removal of the crosslinker residues. SP3 beads are also used for buffer exchange, which is needed for the proteolytic and deoxynucleotide digestion. For more complete deoxynucleotide digestion an endonuclease enzyme mixture was used, containing universal nuclease and nuclease P1. After digestion, (oligo) deoxynucleotides are removed with the so-called peptide cleanup step, also performed on the SP3 beads. This substitutes the C18 reversed-phase chromatographic step from the original protocol. The rest of the protocol is the same as the general protocol - peptide-DNA heteroconjugates are enriched by TiO_2 affinity chromatography. The main advantage of SP3 implementation into the crosslinking workflow is that acetone or ethanol precipitation can be bypassed, which is a critical point of sample preparation, where significant sample loss may occur.

The modified protocol was used for further investigation of the DEB crosslinking method, in which mononucleosome-based complexes were now used in a feasibility study. The mononucleosome, also called nucleosome core particle (NCP), is the basic unit of chromatin, and therefore it is a classical model system for testing protein-DNA crosslinking methods. Mononucleosome consist of 2 copies of the four histone proteins, H2A, H2B, H3 and H4 which are wrapped around with double stranded DNA [88]. Mononucleosome based complexes are also more complex systems in terms of their protein content, than the previously shown linker histone-dsDNA complexes. Three different nucleosome-based complexes were investigated with DEB crosslinking: mononucleosomes, mononucleosome-H1.4 linker histone complex (chromatosomes) and 12mer oligonucleosomes.

4.1.7.2 DEB crosslinking results of the mononucleosome complex

Mononucleosomes were crosslinked with 200 mM DEB and the modified workflow with integrated SP3 cleanup was applied. After mass spectrometric measurement, RNP^{xl} search was used for data analysis.
Manual validation of the CSMs was performed on the identified crosslinking spectra (section 3.3.7), 37 crosslinked peptides were identified (Supplementary Table 5). Crosslinks to DNA of all four core histones were found. Based on the number of identified CSMs, the number of crosslinks found in each protein is decreasing in the manner: H3>H2B>H4>H2A. This order does not necessarily reflect the crosslinking effectivity, because other factors such as ionization efficiency of a peptide or peptide length are important factors in crosslink identification. In some cases, the crosslinking site localization was not possible due to the presence of the unfragmented precursor and low peptide derived fragment ions coverage.
Based on the RNP^{xl} crosslink localization outputs, among others, lysine, glutamic acid, tyrosine, serine, histidine, cysteine, and glutamine crosslinking sites were found. In some cases, proline, leucine, and alanine were identified by RNP^{xl} as crosslink sites. Manual validation of the crosslinking sites took place in the critical cases, including all crosslinking sites, which were mapped on the crystal structure of the *Xenopus laevis* nucleosome core particle (pdb 1KX5 [88]). Crosslinking sites with single amino acid precision were used. When more than one crosslinking site was identified within a peptide, and crosslinking sites were neighboring, only one site was chosen to avoid congestion. Cases with no possible localization or ambiguous localization were not taken into consideration.



pdb: 1KX5

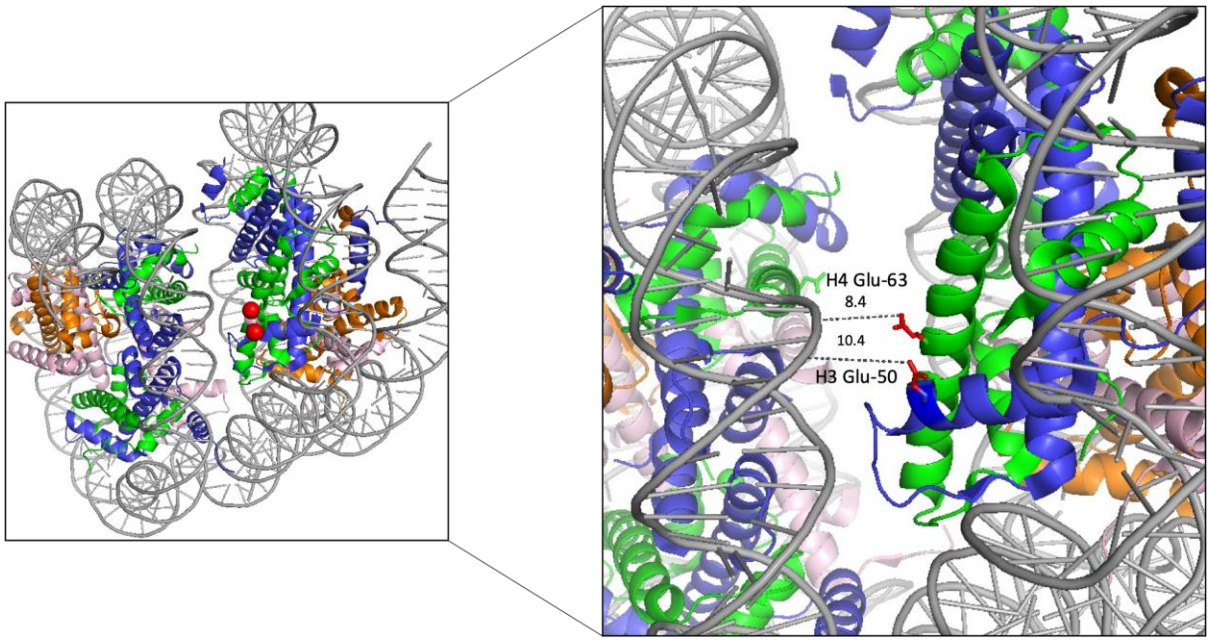
Figure 4.1-21: DEB crosslinking sites in nucleosome core particle (pdb 1KX5 [88**]) of *Xenopus laevis***

H2A colored in orange, H2B colored in pink, H3 colored in blue, H4 colored in green, DNA colored in gray. Left side: C-alpha atoms of crosslinked residues are shown as spheres and highlighted in red. Right side: Crosslink sites more than 20 Å from the closest deoxynucleotide are highlighted in yellow: H2A Glu-91, H2B Lys-105 and H4 Glu-63.

The majority of the crosslinks within the histone proteins are in close proximity to the DNA, as it is shown on Figure 4.1-21 (left side). Several crosslinking sites were found on the N-terminal tail of the H3 and H2B and H2A histones. These tails are flexible, therefore interaction with the DNA is highly plausible. Three crosslinking sites were found to be more than 25 Å away from the DNA strands, localized in the middle part of the nucleosome: H2A Glu-91, H2B Lys-105 and H4 Glu-63 (Figure 4.1-21 right side). Even considering the length of the linker and the length of the side chains, this distance is still far from the DNA. Similar results were found in the UV crosslinked mononucleosomes by Stützer *et al*. [57]. The identified UV crosslinking sites are localized in the same peptide sequences as the DEB crosslinking sites. For instance, in the case of UV crosslinking, H2A Lys-95 was found to be crosslinked in the peptide sequence of NDEELNK, whereas with DEB Glu-91 was found to be crosslinked in the same sequence. Similarly, H4 Val-61 was found to be UV crosslinked in the sequence of VFLENIR, whereas Glu-63 was found to be DEB crosslinked in the same sequence. H2B Lys-105 was commonly found to be crosslinked to DNA in both UV and DEB crosslinking experiments. Stützer *et al.* have suggested that these amino acids are not in interaction with the DNA from the same nucleosome, but in interaction with DNA of another nucleosome in an adjacent manner. Reim *et al.*, have also investigated protein-DNA UV crosslinking on human mononucleosomes [58]. Histones are conserved proteins, H2A type 2B from *Homo sapiens* has 100% sequence identity to histone H2A type 1 from *Xenopus laevis*, therefore the crosslink localization results are comparable. Reim *et al.* have identified crosslink sites which violate the distance constrain as well: H2A Asn-89 was found to be crosslinked to DNA in the NDEELNK peptide. This peptide is an overlap between all three studies (the current study, Stützer *et al.* and Reim *et al.*). Reim *et al*. have hypothesized that this crosslink site is due to a structural rearrangement of the nucleosome, during DNA unwrapping [89]. DNA unwrapping could also explain other crosslinks found in the alpha helixes of the H2B, which violate the crosslink distance constraint in the current DEB crosslinking dataset.

4.1.7.3 DEB crosslinking results of the 12-mer oligonucleosome complex

12-mer oligonucleosomes are mononucleosome-based complexes, in which 12 mononuclosomes are connected with each other through a linker DNA. 12mer oligonucleosomes were crosslinked with 200 mM DEB and the modified workflow was (see section 4.1.7.1) used for protein-DNA crosslink identification. Manual data validation was performed on the dataset (see section 3.3.7). Identified crosslinked peptides are listed in Supplementary Table 6. Crosslinks of all histone proteins were identified. Based on the number of CSMs, most crosslinks were found in H4 then a decreasing order of H3>H2B>H2A. An exact quantitative comparison is not possible between DEB crosslinking results of mononucleosomes and 12-mer oligonucleosomal arrays, since the samples were not measured at the same time, nor processed together. However, all crosslinked peptides identified in the oligonucleosomal arrays were also identified in the mononucleosomal dataset, with the exception of two crosslinked peptides- the H4 KVLR and TLYGFGG peptides. The H4 peptide VFLENIR was also identified with the same crosslink site (Glu-63) in the oligonucleosomal dataset. This crosslink site was identified more than 20 Å away from the closest deoxynucleotide, when considering the mononucleosome structure [90]. Since the nucleosome units in the 12mer nucleosomal array are packed in close proximity to each other, it is very likely that the introduced crosslinks are actually formed between the proteins of one nucleosome and the DNA of another nucleosome. This crosslink site was mapped onto the X-ray structure of an overlapping dinucleosome (pdb:5GSE, [90]) from *Homo sapiens*.



pdb: 5GSE

Figure 4.1-22: DEB crosslinking sites of H4 histone protein Glu-63 and H3 histone protein Glu-50 in the dinucleosome model (in pdb 5GSE)

H2A colored in orange, H2B colored in pink, H3 colored in blue, H4 colored in green, DNA colored in gray. Left side: C-alphas of Glu-63 and Glu-50 are shown as spheres and highlighted in red. Right side: Side chains of the respective H4 and H3 crosslink sites are highlighted in red. Distances (in Ångström) were measured between the side chains and the closest deoxynucleotides of the neighboring nucleosomes in the structure.

H4 Glu-63, which was not fitting into the mononucleosome model (pdb 1KX5 [88]). H4 Glue-63 crosslinking site is considerably close to the neighboring DNA in the dinucleosome model (Figure 4.1-22, right panel), the distance is below 10 Å. Interestingly, H3 Glu-50, which is in close proximity to the DNA in the mononucleosome model is also in close proximity to the neighboring DNA from another nucleosome (Figure 4.1-22, right panel).
The above illustrated results show that peptide-DNA crosslink information can be used to obtain structural information and investigate the spatial conformation of chromatin in solution.

4.1.7.4 DEB crosslinking results of the mononucleosome H1.4 linker histone complex (chromatosome)

The H1.4 human linker histone was reconstituted with nucleosome core particles from *Xenopus laevis.* The complex was DEB crosslinked, followed by the sample preparation for mass spectrometric measurement with the modified workflow (see section 4.1.7.1). Database search was performed with RNP^{xl} [40]. CSMs were manually validated (see section 3.3.7). Results are listed in Supplementary Table 7. More than 30% of the identified CSMs are H1.4 linker histone crosslinks, followed by crosslinks of the core histones in the decreasing order: H2B>H3>H4>H2A. Core histone crosslinks show a very similar picture as in the case of mononucleosome histone crosslinks. In the case of H1.4 crosslinks of both the N-terminal tail and of the globular domain were identified, but predominantly from the N-terminal tail. The same crosslink peptides were found, as in case of the H1.4 and dsDNA system. In the chromatosome, CTD domain crosslinks were identified as well, while these crosslinks were not identified int the H1.4 ds DNA complex (see section 4.1.4). Manually validated

crosslinks sites of H1.4 were mapped onto the cryo-EM structure of the chromatosome (pdb: 7K5Y, [85]

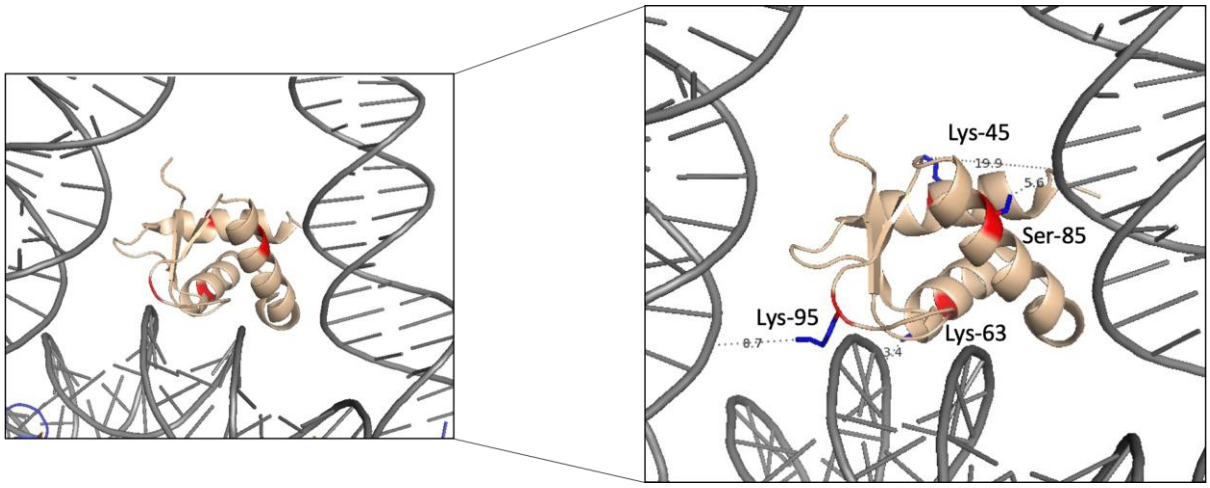


Figure 4.1-23: DEB Crosslinking sites of the H1.4 linker histone (form chromatosome) in complex with ds DNA in pdb 7K5Y **[**85]

Left side: C-alpha sites of the crosslinked amino acids are highlighted in red. Right side: measurement (in Ångström) between the crosslinked side chains of Lys-45 and Lys-63, Lys-95 and Ser-85 the nearest deoxynucleotides.

The cryo-EM structure the chromatosome only contains the globular domain of H1.4, therefore only the globular domain crosslink sites were mapped: Lys-45, Lys-63, Ser-85 and Lys-96. The general picture of the mapped crosslinking sites in the chromatosome is very similar to the H1.4 ds DNA system, as it is shown on Figure 4.1-23. Interestingly, Lys-45 was also identified in the chromatosome model, which is more than 18 Å distance from the closest deoxynucleotide.
According to the data obtained in the chromatosome model system, crosslinks of H1.4-dsDNA and chromatosome are in good agreement.

4.2 Formaldehyde crosslinking

4.2.1 Formaldehyde crosslinking of Protein-DNA complexes

In this section, MS/MS fragmentation behavior of the formaldehyde crosslinked peptide-DNA heteroconjugates are shown. Examples spectrums have been chosen from the formaldehyde crosslinked mononucleosome complex. Biochemical sample preparation was used as described in section 3.2.6.1.

4.2.1.1 MS/MS fragmentation behavior of the formaldehyde crosslinked peptide-(oligo)deoxynucleotide heteroconjugates

During formaldehyde crosslinking, a methylene bridge is formed between proteins and nucleic acids (see section 1.6.1.2.1). Formaldehyde crosslinking between amino acids (or peptides) and DNA was investigated with MS by Lu *et al.* [56]. Lu *et al.* have shown that the methylene bridge is the most liable part of the heteroconjugates during MS/MS fragmentation. The same assumption was used in the current study, it was hypnotized that the methylene bridge would fragment between peptides and DNA. This hypothesis was also based on the previously observed fragmentation of the DEB crosslinked peptide-DNA heteroconjugates, where often the DEB link fragmentation was observed between peptides and DNA (section 4.1.1).
After manual evaluation of formaldehyde induced peptide-DNA heteroconjugates' MS/MS spectra, reoccurring patterns of MS/MS fragmentation became noticeable. As it was expected, the linker cleaves between the peptides and DNA during MS/MS fragmentation. For the illustration of the formaldehyde liker cleavage during MS/MS fragmentation, an amino acid-deoxynucleotide heteroconjugate was chosen; the formaldehyde crosslinked lysine deoxyguanosine monophosphate heteroconjugate.

4.2.1.1.1 Proposed MS/MS fragmentation of the formaldehyde linker

One possible fragmentation of the linker between the deoxyguanosine monophosphate and the lysine is when the linker cleaves from the lysine and stays on the dG. After the formaldehyde linker cleavage, dG fragments as well, through the of the N-glycosylic bond dissociation, which results the loss of the deoxyribose and phosphate groups (Figure 4.2-1 left panel).

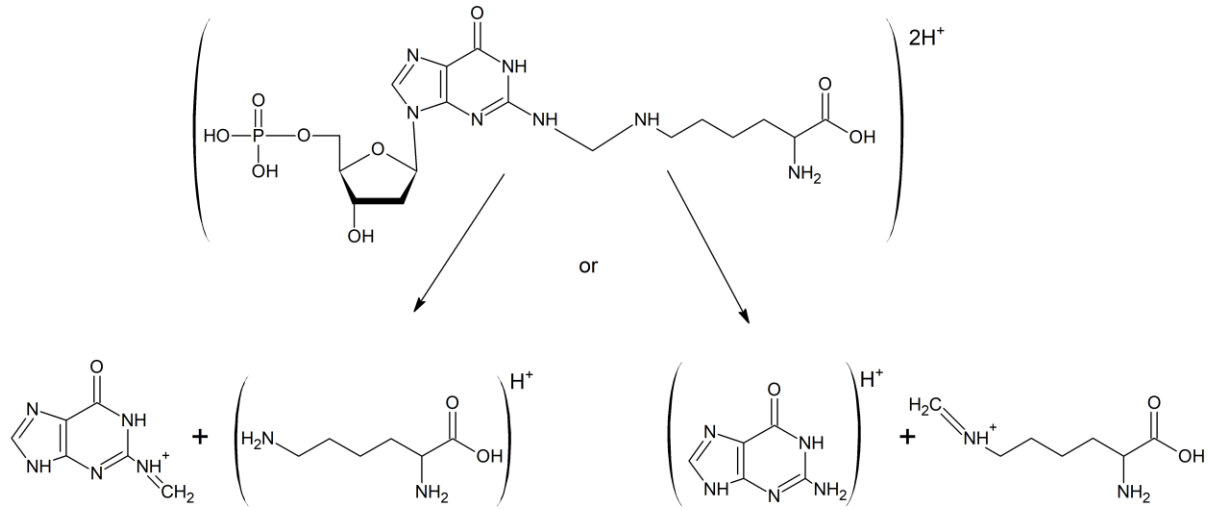


Figure 4.2-1: Proposed MS/MS fragmentation pathways of the formaldehyde crosslinked deoxyguanosine monophosphate-lysine heteroconjugate, adapted from [56]

Left panel: In the proposed MS/MS fragmentation scheme the linker cleaves between the deoxyguanosine monophosphate and the lysine amino acid and stays on the deoxyguanosine monophosphate. Deoxyguanosine monophosphate further fragments through the cleavage of the N-glycosylic bond. Right panel: In the proposed MS/MS fragmentation scheme the linker cleaves between the deoxyguanosine monophosphate and the lysine amino acid, the linker stays on the lysine amino acid. Deoxyguanosine monophosphate further fragments through the cleavage of the N-glycosylic bond.

The linker can cleave in the opposite way too: the formaldehyde linker cleaves off the deoxyguanosine monophosphate and stays on the lysine amino acid (Figure 4.2-1 right panel). Deoxyguanosine monophosphate fragments through the N-glycosylic bond.
In the case of the fragmentation of a peptide-DNA heteroconjugate, the linker cleaves between the crosslinked amino acid and the crosslinked deoxynucleotide, the peptide fragments through its backbone and deoxynucleotide fragments through the N-glycosylic bond. Generally, depending on which fragmentation pathway is more prominent in the MS/MS spectrum of the peptide-DNA heteroconjugates mass shifted marker ions are present or mass shifted peptide fragments are present. When linker fragmentation happens on both fragmentation pathways, both, mass shifted marker ions and mass shifted peptide fragments are present in the MS2 spectra.

One typical MS/MS spectrum of the fragmentation of the peptide-DNA heteroconjugates is shown on Figure 4.2-2. In the MS/MS spectrum an abundant mass shifted adenine marker ion is present (Ab+FA, 148.0623 Da). Which indicates that one of the main fragmentation events was the linker cleavage between the peptide and deoxyadenosine monophosphate. This spectrum is typical for that type of linker cleavage when the nucleobase carries the formaldehyde linker.

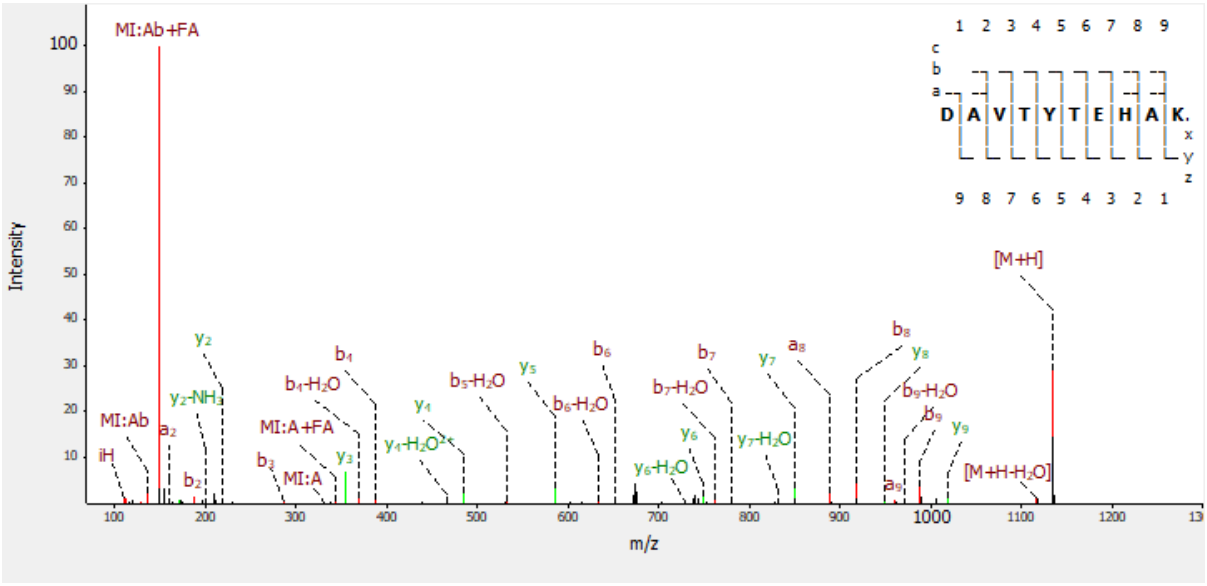


Figure 4.2-2: MS/MS spectrum of FA crosslinked peptide of the H4 histone protein, DAVTYTEHAK. The peptide is crosslinked to deoxyadenosine monophosphate

Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

Another example MS/MS spectrum was chosen to illustrate that both linker fragmentation pathways can be observed in one MS/MS spectrum (Figure 4.2-3).

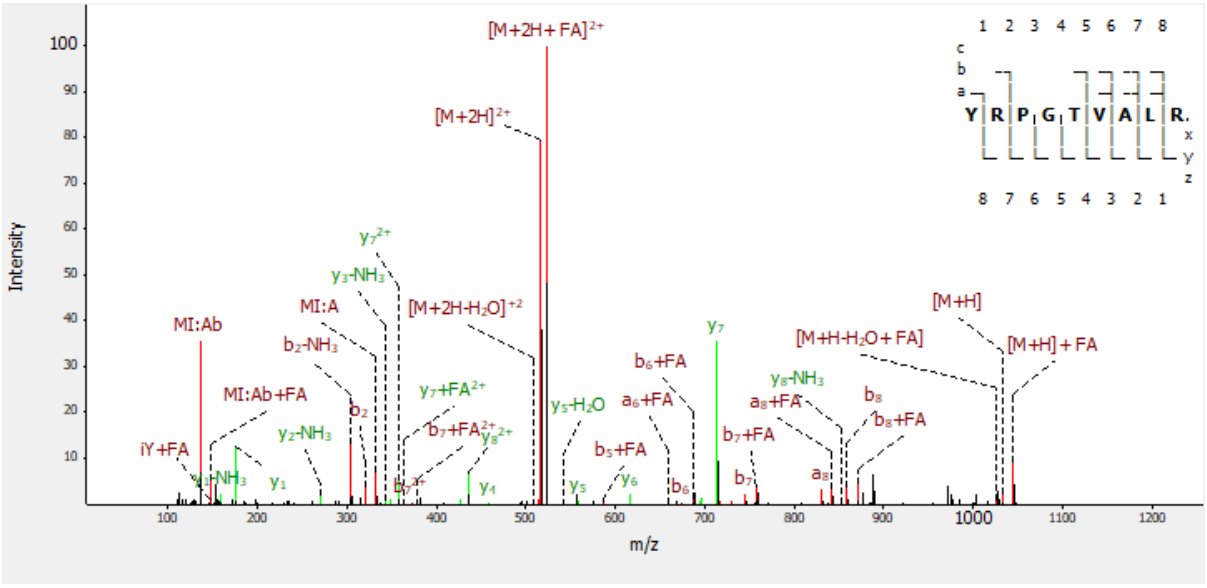


Figure 4.2-3: MS/MS spectrum of FA crosslinked peptide of H3 histone protein, YRPGTVALR. The peptide is crosslinked to deoxyadenosine monophosphate

Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

In the MS/MS spectrum mass shifted (+FA, 12 Da) and unshifted ion pairs are present, indicating that both linker fragmentation events happened on the peptide-DNA heteroconjugates. Figure 4.2-3 shows that marker ions can have a lower intensity, when peptide precursor related peaks have high intensities.
The two example spectra shows that the most liable bond between peptides and DNA is the formaldehyde linker and the MS/MS fragmentation of the peptide-DNA heteroconjugates follow a simple pattern.

4.2.1.1.2 Effect of the deoxynucleotide length on the MS/MS fragmentation of the peptide-DNA heteroconjugates

The DNA adduct of the peptide-DNA heteroconjugates can consists of one deoxynucleotide, two or three deoxynucleotides. When the DNA adduct is a single deoxynucleotide, the above shown fragmentation possibilities apply. When the DNA adduct composition consists of more than one deoxynucleotide, the phosphodiester bond cleaves between the deoxynucleotides during the fragmentation of the DNA-peptide heteroconjugates. Therefore, more than one marker ion appears in the lower *m/z* region of the MS/MS spectra. The presence of the marker ions and the mass shifted marker ions indicate the original deoxynucleotide composition of the peptide-DNA heteroconjugates. An example spectrum was chosen to illustrate the MS/MS fragmentation of the peptide-DNA heteroconjugates when DNA adduct consist of more than one deoxynucleotide. In the example MS/MS spectrum (Figure 4.2-4), RSTITSR peptide of the H2B histone is crosslinked to dinucleotide dCdG. In the spectrum, cytosine marker ion (Cb, 112.0511 Da) and mass shifted guanine marker ion (Gb+FA, 164.0572 Da) are present. As well, as deoxycytidine monophosphate and fragment of mass shifted deoxyguanosine monophosphate (dG-H_3PO_4-H_2O+FA, 244.0834 Da)

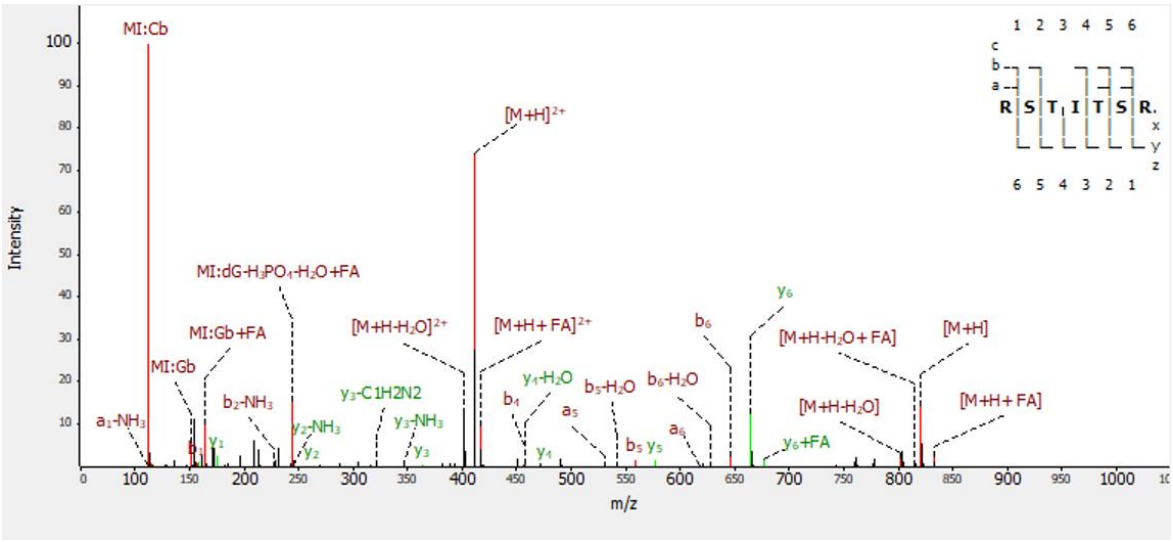


Figure 4.2-4: MS/MS spectrum of FA crosslinked peptide of H2B histone proteins, RSTITSR. The peptide is crosslinked to dCdG DNA adduct

Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

The presence of the Cb (112.0511 Da) and Gb (152.0572 Da) marker ions as well as the mass shifted Gb marker ion (Gb+FA, 164.0572 Da) indicate that the composition of the DNA adduct is dCdG and the peptide is crosslinked to dG.

Another example spectrum (Figure 4.2-5) was chosen for the description of the MS/MS fragmentation of the peptide-DNA heteroconjugates, when DNA adduct composition consists of more than one deoxynucleotide. LAHYNK peptide of the H2B histone protein, crosslinked to thymidine monophosphate-deoxyadenosine monophosphate dinucleotide (dTdA). Adenine marker (Ab, 136.0623 Da) ion and mass shifted adenine marker ion (Ab+FA, 148.0623 Da) are present in the low *m/z* region in the spectrum. An intense peak at 161.0596 Da was observed. This, 161.0596 Da characteristic peak appeared in all CSMs when DNA adduct contains deoxythymidine monophosphate in the MS2 spectra of the formaldehyde crosslinked mononucleosome dataset (186 CSMs out of 186 CSMs when DNA adduct contains dT). Elemental composition analysis of the characteristic peak was carried out, but

the chemical structure of the peak could not be confirmed. Further experiments are needed to confirm that the 161.0596 Da peak is a fragment ion of the deoxythymidine monophosphate.

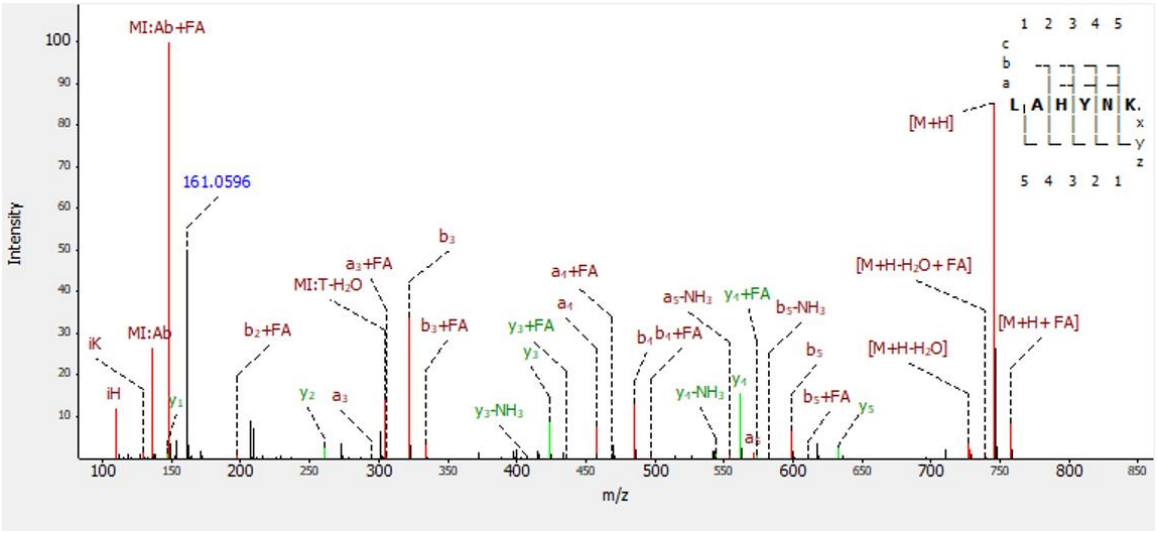


Figure 4.2-5: MS/MS spectrum of FA crosslinked peptide of H2B histone protein, LAHYNK. The peptide is crosslinked to dTdA DNA adduct

Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

4.2.1.1.3 Localization of the crosslinking site in the crosslinked peptide sequence

After thorough investigation of the crosslinking site localization, the general observation is, no confident crosslinking site localization is possible on the peptide sequence. This is due to two factors. One factor is the feature of the peptide-DNA heteroconjugates' MS/MS fragmentation. When the deoxynucleotide carries the linker, insufficient number of the mass shifted peptide fragments are available for confident crosslinking site localization. The second factor is the isobaric peptide-DNA crosslinks. It was already discussed at the DEB crosslink localization section (section 4.1.1.5), in some cases mixed spectra of the isobaric peptides were observed. When formaldehyde crosslinking is used, this phenomenon is more prominent. Two types of isobaric peptides were observed: when the peptide is crosslinked to DNA at different amino acid positions and when the peptide is crosslinked to different deoxynucleotides in the DNA adduct.

4.2.1.1.3.1 Peptide is crosslinked to DNA adduct at different amino acid positions

An example MS2 spectrum was chosen for the illustration of the MS/MS fragmentation of isobaric peptides, the MS2 spectrum of HLQLAVRNDEELNK peptide of H2A histone, peptide is crosslinked to deoxyadenosine monophosphate.

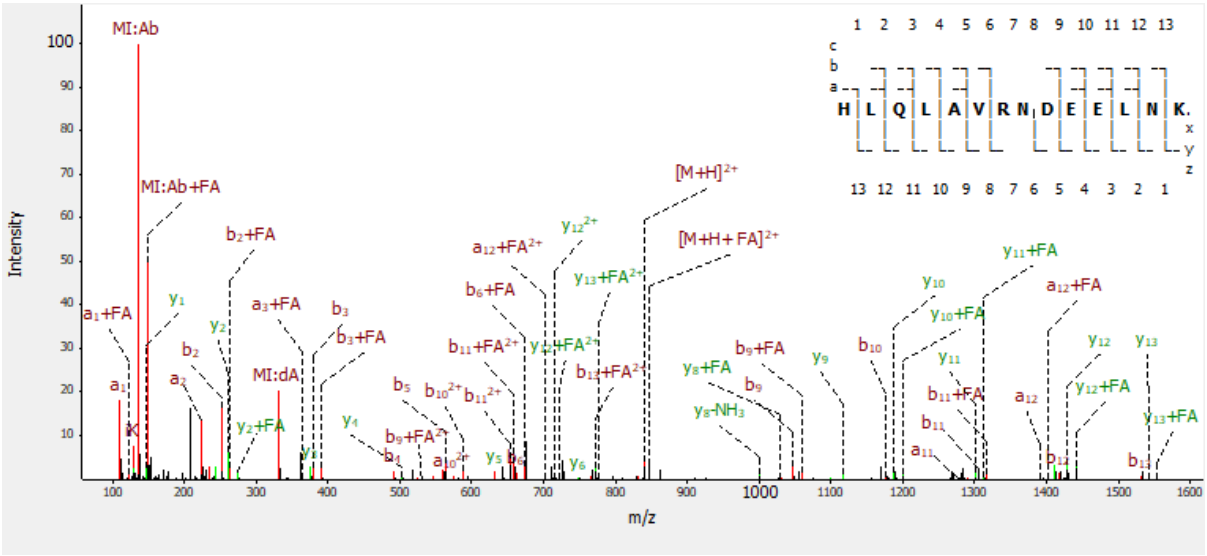


Figure 4.2-6: MS/MS spectrum of FA crosslinked peptide of H2A histone protein, HLQLAVRNDEELNK. The peptide is crosslinked to dA DNA adduct

Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

As it is shown on the Figure 4.2-6, a1 and mass shifted a1 fragments were observed, indicating that the crosslinking site is localized on the histidine in the peptide sequence. Mass shifted ion series are present on the y ion series as well, starting from the y2 ion. This observation is contradictory to the histidine crosslinking site because a1 and y2 ions are complementary ions. The presence of isobaric peptides can explain the incompatible mass shifted ion series in the spectrum. Based on the observed mass shifted ion series, crosslinking sites were originally localized on the histidine and on the asparagine amino acids. Often, isobaric peptides' crosslinking site localization is not evident due to the insufficient number of mass shifted ions. Proteomics search engines cannot correctly indicate the site localizations when isobaric peptides are present. Therefore, in the following experiments no crosslinking site localization was considered.

4.2.1.1.3.2 Peptide is crosslinked to different deoxynucleotides in a DNA adduct

Another example spectrum was chosen to illustrate the observed isobaric peptides' MS/MS fragmentation: crosslinked peptide DAVTYTEHAK, of the H4 histone protein. The peptide is crosslinked to dAdG dinucleotide. As it is shown on Figure 4.2-7 mass shifted adenine and mass shifted guanine peaks (dA+FA, 148.0623 Da and dG+FA 164.0572 Da) are present and mass shifted dG fragment: dG-H_3PO_4-H_2O (244.0834 Da). Moreover, dA (332.0760 Da) intact deoxynucleotide and mass shifted dA (dA+FA 344.0760 Da) intact deoxynucleotide peaks are present as well.

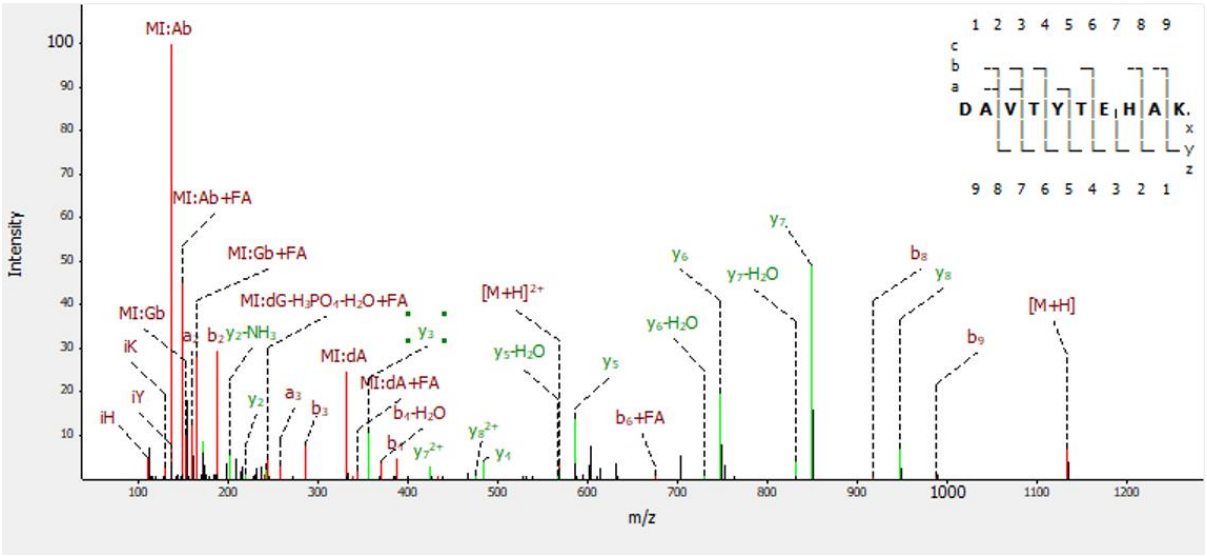


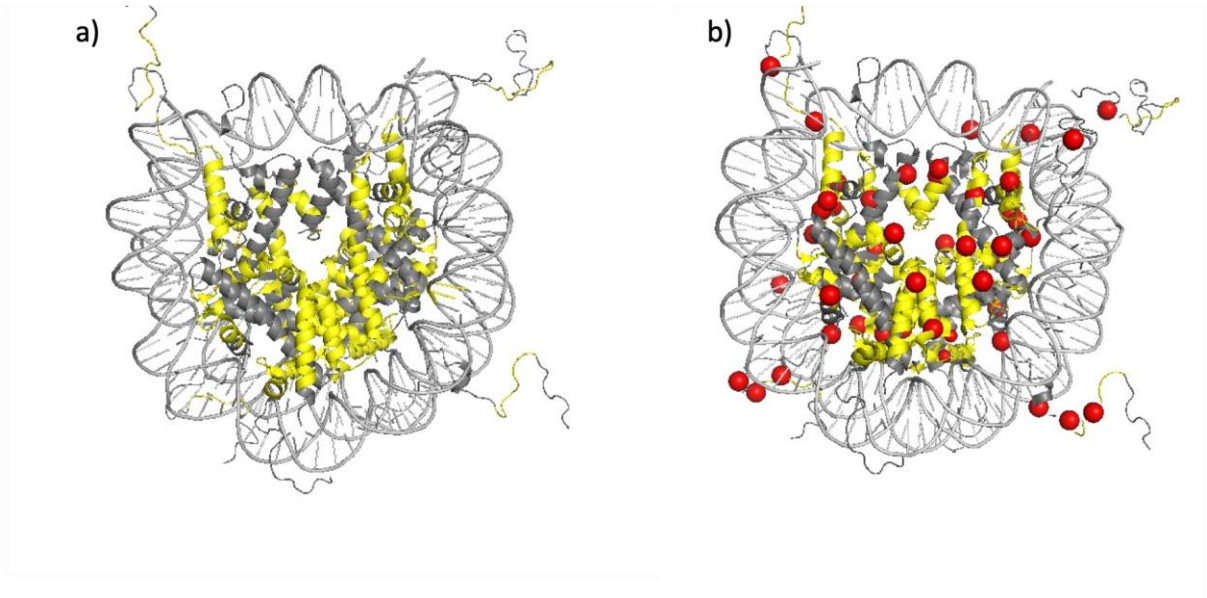
Figure 4.2-7: MS/MS spectrum of FA crosslinked peptide of the H4 histone protein, DAVTYTEHAK. The peptide is crosslinked to dAdG DNA adduct

Peptide precursor ions, a and b fragment ions are colored in red; y fragment ions are colored in green.

As it is shown on Figure 4.2-7, only one mass shifted peptide fragment is present in the MS/MS spectrum, b6+FA. This, one mass shifted ion cannot provide enough information for confident crosslinking site localization.

4.2.1.2 Formaldehyde crosslinking of the *in vitro* reconstituted mononucleosomes

Formaldehyde crosslinking of the *in vitro* reconstituted mononucleosomes was performed. Biochemical sample preparation was as follows: mononucleosomes were crosslinked with 1(v/v) % formaldehyde, the reaction was quenched with primary amines. The sample was subjected to SP3 protein cleanup, followed by protein digestion with trypsin and DNA digestion with nucleases. The sample was subjected to SP3 peptide cleanup and TiO_2 affinity chromatography. The peptide-DNA heteroconjugates were analyzed with HPLC-MS/MS. RNP^{xl} search was used for the MS data analysis, 1% FDR cut RNP^{xl} identifications were manually checked for correct peptide assignment. Identifications were further checked with an R script (section 3.3.3.4.1) for the presence of the respective marker ions of the DNA adduct in the MS/MS spectra and crosslinked deoxynucleotides and DNA adducts were confirmed. Crosslinks to all four core histones were identified (Supplementary Table 8). The number of CSMs in each protein as follows: H2B>H4>H2A>H3. Confident crosslink localization on a single amino acid was not possible, therefore whole crosslinked peptide sequences were used as crosslinking sites. Crosslinked peptides were mapped into the X-ray structure of *Xenopus laevis* nucleosome (Figure 4.2-8a), crosslinked peptide sequences are highlighted in yellow. As it is shown on Figure 4.2-8a, due to the lack of the confident crosslinking site localization, majority of the protein sequences are highlighted in yellow. DEB crosslinking sites of the mononucleosome were mapped into the structure as well, to show the overlap between the FA and the DEB crosslinking sites (Figure 4.2-8 b).



pdb: 1KX5

Figure 4.2-8: FA crosslinked peptide sequences mapped onto the crystal structure of *Xenopus laevis* nucleosome core particle.

Histone proteins and dsDNA are colored in gray. a): FA crosslinked peptide sequences are mapped into the nucleosome structure; crosslinked peptide sequences are highlighted in yellow. b): FA crosslinked peptides and DEB crosslinking sites mapped into the nucleosome crystal structure; FA crosslinked sequences are highlighted in yellow; DEB crosslinked sites are highlighted as red spheres.

As it is shown on Figure 4.2-8b the DEB crosslinking sites are localized on the FA crosslinked peptide sequences. Therefore, results of the DEB crosslinking and formaldehyde crosslinking are in good agreement with each other. When precise contact site identification is needed between DNA and proteins, formaldehyde crosslinking is not the best fitting tool, the advantageous and simple MS/MS fragmentation of the formaldehyde crosslinked peptide-DNA heteroconjugates could be better used for *in vivo* crosslinking.

4.2.1.3 The establishment of the MS2/MS2 acquisition method

As shown above, the MS/MS fragmentation of the formaldehyde crosslinked DNA-peptide heteroconjugates is not dependent the deoxynucleotide, but on the cleavage of the formaldehyde linker between peptides and deoxynucleotides. An MS acquisition method was developed (Figure 4.2-9) based on the observed MS/MS fragmentation patterns.

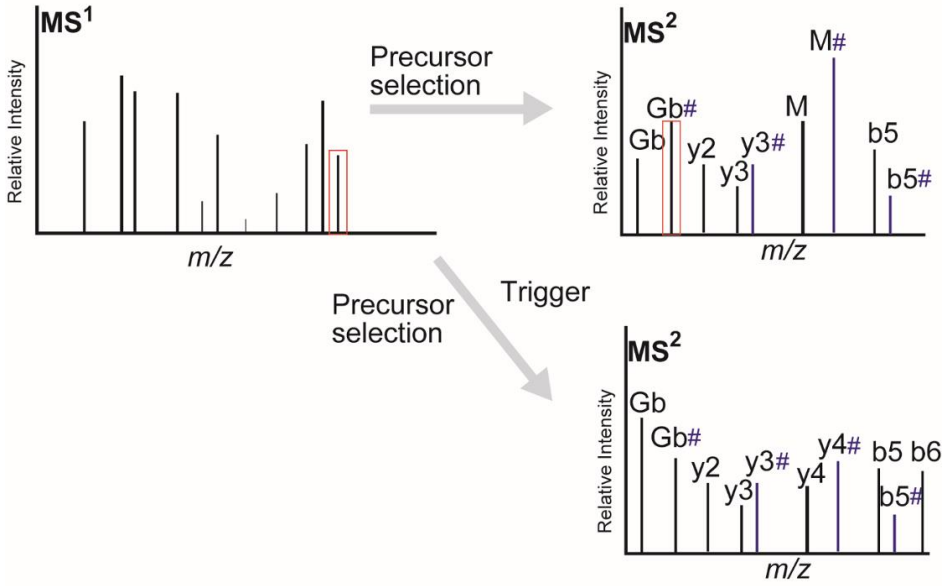


Figure 4.2-9: Illustration of the MS2-MS2 acquisition method

Marker ions and mass shifted marker ions were targeted during recording in each MS2 spectra, when one of target masses were found within the ten most abundant peaks in the MS2 spectrum, another MS2 spectrum was triggered and recorded of the same precursor ion.

Marker ions (Cb 112.0511 Da, Ab 136.0623 Da, Gb 152.0572 Da) and mass shifted marker ions (Cb+FA 124.0511, Ab+FA 148.0623, Gb+FA 164.0572 Da) are targeted during the MS2 acquisition. When the target masses are present within the ten most abundant peaks in the MS2 spectrum, MS2 spectrum is recorded again with longer injection times and a higher collision energy value (The whole MS acquisition method is available in Supplementary Text 2). The targeted method allows to spend more measuring time on the low abundant crosslinked peptides' MS/MS spectra and the higher injection times results better quality MS/MS spectra.
This acquisition method can be used to gain better crosslink identification, because higher quality and the higher number of crosslinked MS/MS spectra can boost the crosslink identification.

4.2.1.3.1 Control experiments for the validation of the MS2/MS2 acquisition method

MS2/MS2 acquisition method was established, based on the observed MS/MS fragmentation of the formaldehyde crosslinked peptide-DNA heteroconjugates. For the validation of the MS2/MS2 acquisition method, control experiments were performed. The following controls were used: not crosslinked, reconstituted linker histone H1.4 187 bp dsDNA system (referred as NXL), formaldehyde crosslinked linker histone H1.4 (referred as XL H1.4), formaldehyde crosslinked 187 bp dsDNA (referred as XL DNA), tryptic peptides of HeLa protein digest (referred as HeLa). The latter was used to investigate, if any masses, present in the MS/MS spectrum of peptides would interfere with the acquisition method. Formaldehyde crosslinked linker histone H1.4- 187 bp ds DNA sample was compared to all control samples. Biochemical sample preparation was performed as described in section 3.2.6.2.
During the MS2/MS2 acquisition method masses of marker ions (Cb 112.0511 Da, Ab 136.0623 Da, Gb 152.0572 Da) and FA mass shifted marker ions (Cb+FA 124.0511, Ab+FA 148.0623, Gb+FA 164.0572 Da) were targeted. The characteristic peak at 161.0595 Da (presumably derived from dT, see above) was targeted alongside the marker ions and mass shifted marker ions. When one of the target masses was found within the top 10 most abundant peaks in the MS2 spectrum another MS2 spectrum was recorded. Triggered MS2 spectra was only recorded in the crosslinked H1.4 ds DNA complex (183 MS2 spectra), no MS/MS spectra were triggered in the control samples, only in the crosslink sample, therefore

no interfering peaks were found in the control samples. Based on the control experiments the MS2/MS2 acquisition is crosslink specific and can be used for acquiring more crosslink MS2 spectra to support better crosslink identification.

4.2.1.4 Formaldehyde crosslinking of the native nucleosomes isolated from HeLa cells

Nucleosomes isolated from HeLa cells were crosslinked with 1 (v/v) % formaldehyde. Detailed description of the biochemical sample preparation in available in section 3.2.6.3 Briefly as follows: after crosslinking sample was subjected to SP3 protein cleanup. Followed by nuclease digestion with universal nuclease and nuclease P1 and protease digestion with trypsin, overnight. Sample was subjected to C18 reversed-phase chromatography followed by TiO_2 affinity chromatography for the enrichment of peptide-(oligo)deoxynucleotide heteroconjugates. Sample was subjected to HPLC-MS/MS analysis. Raw data analysis was performed by RNP^{xl}. RNP^{xl} identifications were manually checked for correct peptide assignment. Identifications were further checked with an R script (section 3.3.3.4.1 and Supplementary Text 3) for the presence of the respective marker ions of the DNA adduct in the MS/MS spectra and crosslinked deoxynucleotides and DNA adducts were confirmed. Quality control was performed by SDS PAGE separation of the proteins from the isolated HeLa nucleosome sample, where several other proteins were identified alongside of the nucleosomal proteins. MS identification of the sample's protein content was preformed, alongside the nucleosomal protein, several nuclear proteins were identified (data not shown). Thus, this sample was used as "semi" complex sample to investigate crosslinks in the nucleosome and in other nuclear proteins.
In total, 13 proteins were identified (Supplementary Table 9) to be crosslinked to DNA with formaldehyde, when a peptide sequence was shared between different proteins, the first protein hit was considered to avoid protein overrepresentation. Among others, all four core histones (H2A1B, H2B1B, H3.1, H4) and histone variants (H2AY, H3.3) were identified as DNA crosslinks, as well as nuclear proteins such as heterogeneous nuclear ribonucleoproteins C1/C2 (HNRNPC), prelamin A/C, filamin A and actin, cytoplasmic 1.
All peptides of the H2B1B, two peptides of the H3.1, three peptide of the H4 and peptide of the H2A1B have been identified in the *in vitro* reconstituted mononucleosome sample as well. Alongside to the DNA crosslinks, RNA crosslinks were observed as well, for instance H4 peptide VFLENIR was identified as adenosine monophosphate crosslink. Classical RNA binding proteins were identified as DNA crosslinks in this study, such as Heterogeneous nuclear ribonucleoproteins C1/C2 (HNRNPC), 60S ribosomal protein L12 and 60S ribosomal protein L30. 60S acidic ribosomal protein L0 have been identified to be crosslinked to RNA in this dataset. One of the classical RNA binding proteins have known DNA binding function: HNRNPC is known to bind the nucleosomal DNA [91].
Other, not classical DNA binding proteins were identified as DNA crosslinks, such as prelamin A/C, filamin-A and actin, cytoplasmic 1. Prelamin A/C plays an important role in regulation of heterochromatin [92]. Actin proteins are mostly known about their functions in the cytoplasm, although they are localized in the nucleus as well and regulate the gene transcription [93]. Identified proteins based on their molecular functions were visualized on a sunburst diagram (Figure 4.2-10).

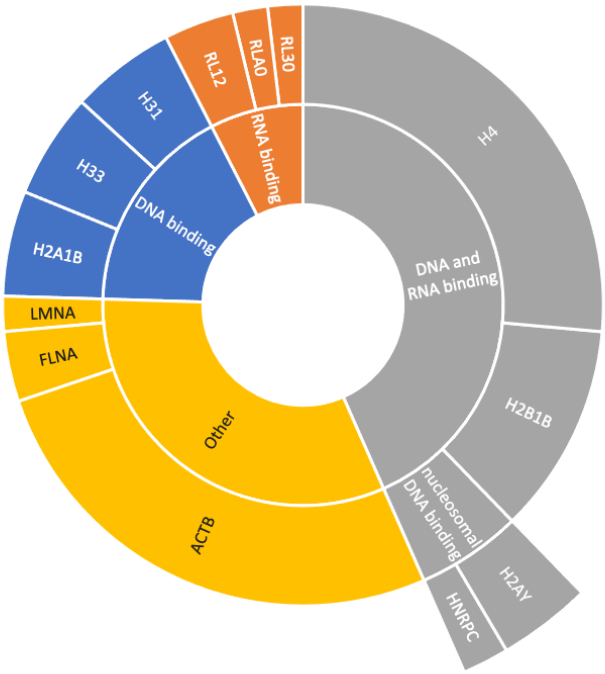


Figure 4.2-10: Sunburst diagram of molecular functions of the formaldehyde crosslinked, semi purified nucleosomes and their interacting proteins isolated from HeLa cells

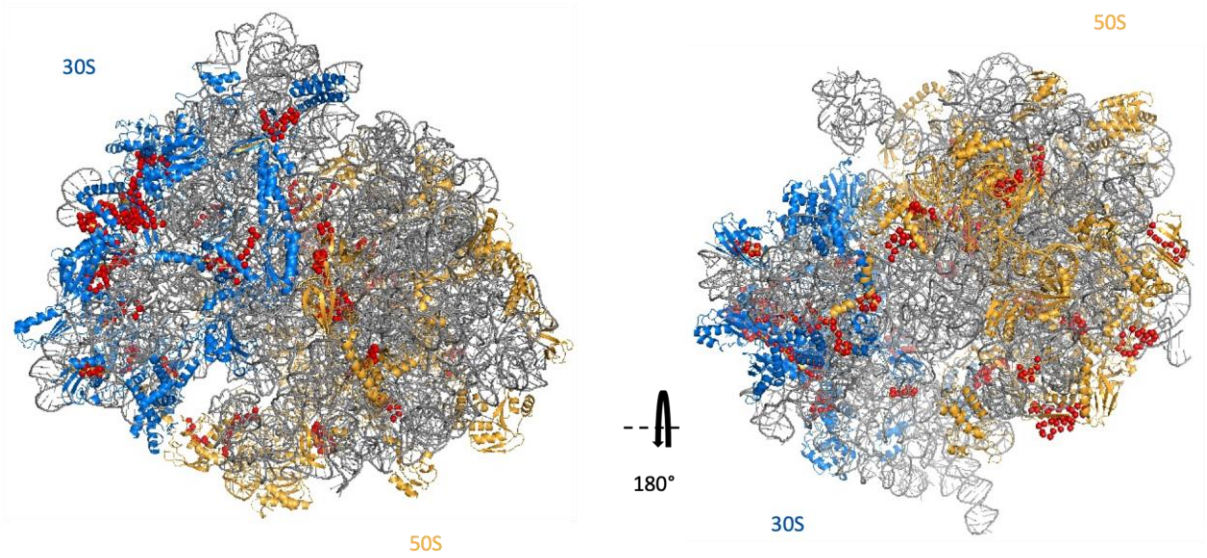
Based on the number of CSMs, histone protein H4 and actin 1 were scored the highest. Both proteins were identified to be crosslinked to RNA and DNA as well in the current study (Supplementary Table 9)
Results of this experiment have supported that RNA and DNA formaldehyde crosslinks can be measured together, and crosslinks to deoxynucleotides and nucleotides can be distinguished.

4.2.2 Formaldehyde crosslinking of protein-RNA complexes

As described for DNA (section 4.2.1), RNA can also crosslink with formaldehyde following the same chemical reaction. The crosslinking reaction involves the nucleobases of RNA and the respective chemical groups of the amino acids. The previously described biochemical sample preparation and MS2/MS2 acquisition method is applicable to protein RNA crosslinking because RNA shares three nucleobases with DNA.
First, the formaldehyde crosslinking workflow was tested on a well characterized protein-RNA containing complex, the 70S ribosome isolated from *Escherichia coli*.

4.2.2.1 Formaldehyde crosslinking of the 70S ribosome from *Escherichia coli*

The 70S ribosome isolated from *E. coli* (B strain) was *ex vivo* crosslinked with 1% formaldehyde. Detailed biochemical sample preparation is described in section 3.2.6.4. Sequential RNA digestion was implemented into the biochemical sample preparation workflow aiming for complete nucleoside digestion. For the MS data analysis an open search strategy was used, a so-called mass offset search. In a mass offset search precursor mass tolerances are opened at distinct delta masses. In the case of crosslink identification, mass offsets are the mass of the crosslinked RNA adducts. MSFragger search engine [66] was used for crosslink identification with mass offset setup. An additional marker ion search was performed for further crosslink validation, described in section 3.3.3.4 and Supplementary Text 4. During marker ion search, based on the heteroconjugates' assigned RNA adduct composition, the respective nucleotides' marker ions (mass shifted and unshifted) were searched in the MS2 spectra of the crosslinked heteroconjugates. When the respective marker ions were present in the MS2 spectrum, MS2 spectrum was accepted as crosslinked spectrum. This, marker ion search was used as an additional validation of the crosslinked spectra.
32 out of the 52 70S ribosomal proteins were identified as crosslinked to RNA, including 14 proteins of the small subunit and 18 proteins of the large subunit (Supplementary Table 10) In consideration of the crosslink peptides: 66% of the CSMs are adenine crosslinks, followed by guanine (20%) and cytosine (14%) crosslinks, therefore, the tendency of crosslinking sites is A>G>C. Based on the CSMs 50S L2 protein has the highest number of crosslinked spectrum matches (43 CSMs), then 30S protein S12 (31 CSMs) and 50L protein L31 follows (26 CSMs). 50S L2 protein in the 70S ribosome is known to be one of the main ribosomal RNA binding proteins [94]. in fact, it is essential for the association of the small (30S) and large (50S) ribosomal subunits [95].
Crosslinked proteins, crosslinked peptides and their corresponding crosslinked nucleotides are listed in Supplementary Table 10.
The identified crosslinks were mapped onto the cryo-EM structure of the ribosomal elongation complex in its classical state (PDB: 5LZE) [96] as shown in Figure 4.2-11. Crosslinked amino acid could not be identified due to the low number of formaldehyde mass shifted peaks; thus, the whole peptide sequences were considered as crosslinked protein regions.



pdb: 5LZE

Figure 4.2-11: Formaldehyde-induced crosslinks in the 70S ribosome from *E. Coli* in pdb 5LZE

Proteins of the 30S subunit are colored in blue, proteins of the 50S subunit are colored in yellow, C-alpha atoms in the amino acids of the crosslinked peptides are highlighted as red spheres.

The relative positions of crosslinked peptides in the small and large subunit differ. Crosslinks of the small subunit are closer to the surface of the structure, while crosslinks of the large subunit are closer to the 23S RNA, located at the center of the structure.
Several crosslinked peptides have been identified at the interface of the two subunits, like 30S protein S13 peptide LMDLGCYR. S13 is part of the small subunit, however, the identified crosslinked peptide is in closer proximity to the 23 S RNA of the large subunit, than to the 16S RNA. A similar observation was made for crosslinked peptide FDGNACVLLNNNSEQPIGTR of the 50S protein L14 which is located in closer proximity to the 16S RNA, even though the protein is part of the large subunit.
Four of the crosslinked proteins show known RNA binding domains. 30S S1 protein has 6 S1 motifs, 30S S3 protein has a K homolog RNA binding domain, 30S S4 protein has an S4 RNA binding domain and 30S S5 protein has a double stranded RNA binding motif (DRBM). Two proteins of the total four proteins have an overlap between the RNA binding domains and their crosslinked peptides: 30S protein S1 and 30S protein S5. Crosslinked peptides (aa 14-25 and aa 142-150) of 30S S1 protein overlap with the S1motif one (aa 21-87) and S1 motif two (aa 105-171). Crosslinked peptide of the 30S S5 protein (aa 69-86) overlaps with the DRBM domain (aa 11-74).

4.2.2.2 Sequential RNA digestion strategy results in formation of single nucleotides and nucleosides

Apart from a selective crosslink enrichment strategy, decreasing the complexity of the data is a bottleneck of the presented method. Data analysis can be facilitated by the reduction of crosslinked nucleotide lengths. In the case of the DEB crosslinking method (section 4.1.2) to shorten nucleotides' length, gas phase fragmentation was used. That approach did not meet all expectations, because did not result in a completely uniform fragmentation (section 4.1.1.4).
In this section, a sequential RNA digestion strategy is described, which was tested on the 70S ribosome, to see if RNA adducts can be digested to single nucleosides and whether data complexity is reduced upon reduction of nucleotide adducts. Reduction of the complexity of the RNA adducts can be advantageous for several reasons. MS1 signal of the peptide-RNA heteroconjugates are split between several RNA adducts of the same crosslinked peptide. When number of RNA adducts reduced, MS1 signal of crosslinked peptide increases, there is a higher chance to identify the low abundant peptide-RNA heteroconjugates. When limited number of RNA adducts are present there is a lower chance for false identification of the crosslinked peptide. In the first step of the sequential digestion RNase 1 and RNase A were used in the second step of the sequential digestion Nuclease P1 enzyme in combination with an alkaline phosphatase were used. The phosphatase was used to uniformly remove phosphate groups from all nucleotides. The identified RNA adducts on peptide precursors and the respective CSM counts are shown in Figure 4.2-12.

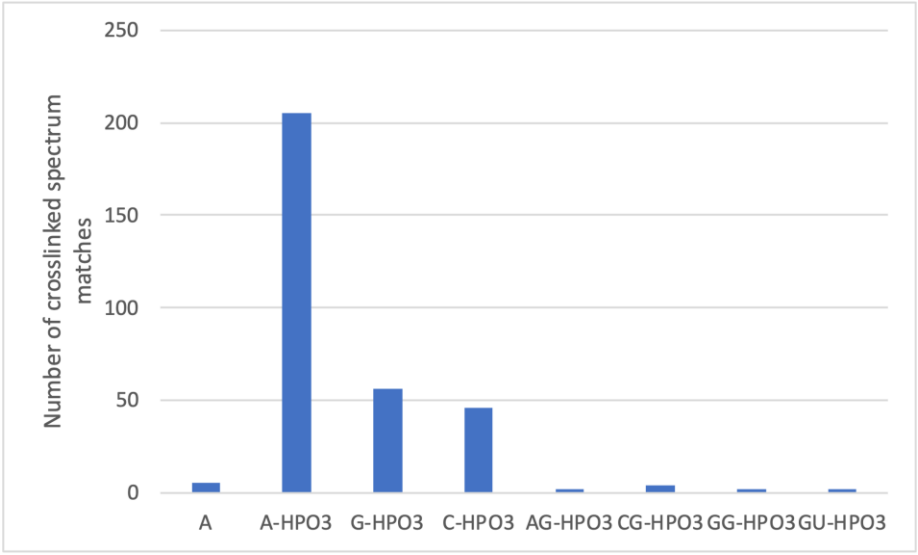


Figure 4.2-12: Number of identified crosslinked spectrum matches and their respective crosslinked nucleotide adducts

As it is shown in Figure 4.2-12 the majority of the RNA adducts (97% of the CSMs) are mononucleotide and nucleoside crosslinks. Ten CSMs were identified with dinucleotide RNA adducts: 2 CSMs of AG-HPO_3, 2 CSMs of UG-HPO_3, 4 CSMs CG-HPO_3, 2 CSMs GG-HPO_3.
Sequential digestion is beneficial for nucleoside digestion and can dramatically reduce the complexity of the RNA adducts.

4.2.3 *In vivo* crosslinking in *Escherichia coli*

4.2.3.1 Biochemical and mass spectrometric workflow for the identification of protein-RNA crosslinks in vivo

Formaldehyde crosslinking in *Escherichia coli* was used to establish *in vivo* crosslinking in cells. Silica enrichment (section 3.2.8). were used for the enrichment of formaldehyde crosslinked RNA-peptide heteroconjugates. General workflow for silica enrichment is shown in Figure 4.2-13. Enrichment method was adapted from Dr. Aleksandar Chernev's PhD thesis [64], with minor changes.

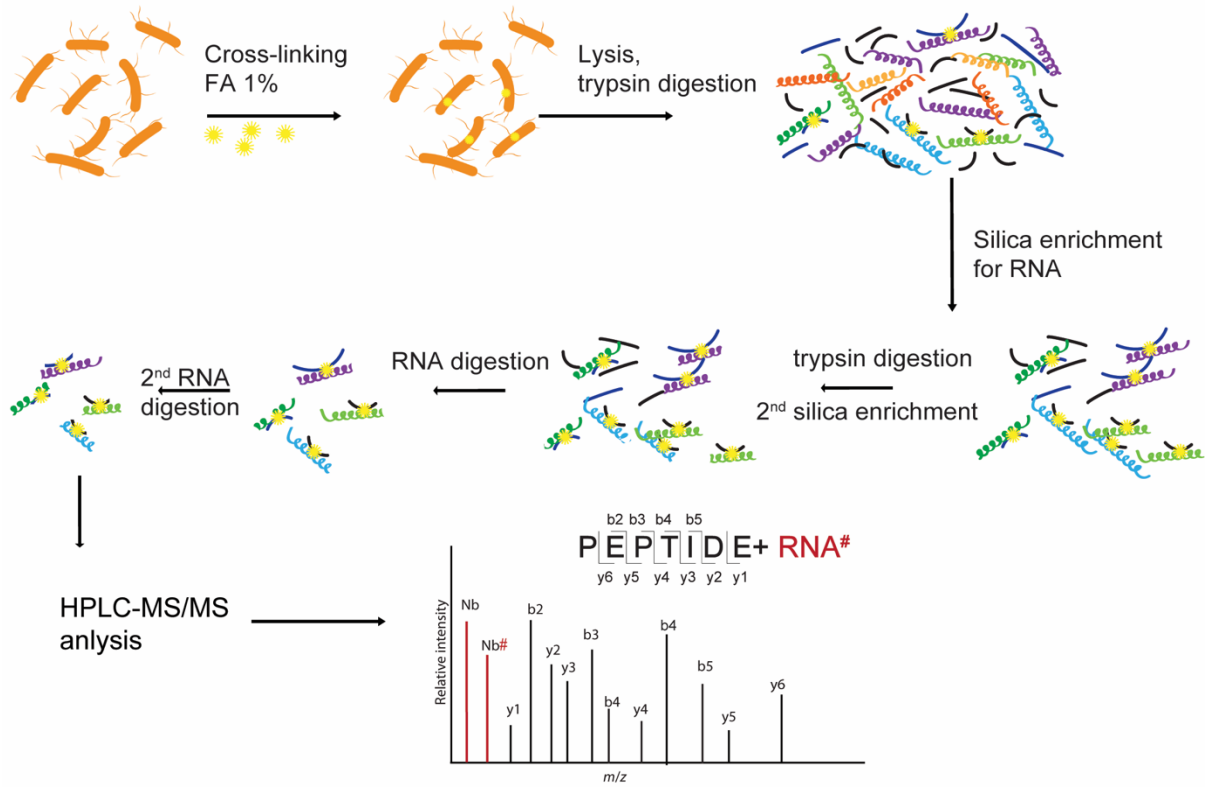


Figure 4.2-13: General biochemical workflow of in vivo formaldehyde crosslinking in cells combined with HPLC-MS/MS analysis and crosslink data analysis

Detailed biochemical workflow is shown is section 3.2.8 Briefly: in the first step of the workflow, *Escherichia coli* cells were crosslinked *in vivo* with 1% FA. After quenching the reaction with Tris, cell lysis was performed by addition of 30 mg/ml lysozyme and ultrasonic disruption, followed by overnight trypsin digestion of the proteins. A two-step silica-based enrichment workflow was performed, with additional protein digestion step between the enrichment steps. Subsequently, RNA was digested in two steps. After each RNA digestion step, C18 reversed-phase chromatography was performed to remove free nucleotides. The purpose of the two-step digestion was to remove free nucleotides, which can otherwise interfere with the downstream analysis, and to maximally digest RNA adducts, lowering the complexity of the data analysis. Enriched peptide-RNA heteroconjugates were subjected to HPLC-MS/MS analysis with MS2/MS2 trigger method. Followed by MSFragger database search (section 3.3.3.3) for the identification of RNA crosslinked proteins and marker ion search (3.3.3.4.2, Supplementary Text 4).

4.2.3.2 Functional enrichment analysis of the formaldehyde crosslinked proteins in Escherichia coli

The *in vivo* crosslinking experiment resulted 243 crosslinked proteins with 1% peptide FDR and 1% protein FDR and marker ion search (Results are available in Extended Supplementary Table 1). Constituent proteins of the 70S ribosome were found to be crosslinked to RNA. Several proteins were identified from different cellular processes, such as translation initiation, elongation and ribosome recycling. Translation initiation factors IF1, IF2, IF3, as well as elongation factor G, Ts, and Tu1/Tu2 were identified. tRNA ligase proteins, such as aspartate t-RNA ligase or phenylalanine t-RNA ligase alpha and beta were also found. DNA directed RNA polymerase: subunit β', β and α were identified.

Crosslink proteins' UniProt identifiers were uploaded into several databases for further analysis. STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) [79] was used for functional enrichment analysis (242 out of 243 UniProt IDs were annotated in STRING). UniProt retrieve function was used to identify the molecular functions of the proteins, as well as the RNA binding domains.
First, in UniProt the proteins' GO molecular functions section was checked to see how many proteins are known RNA binding proteins. Less than half of the proteins (93) were marked as RNA binding or structural constituents of the ribosome (Figure 4.2-14 a).

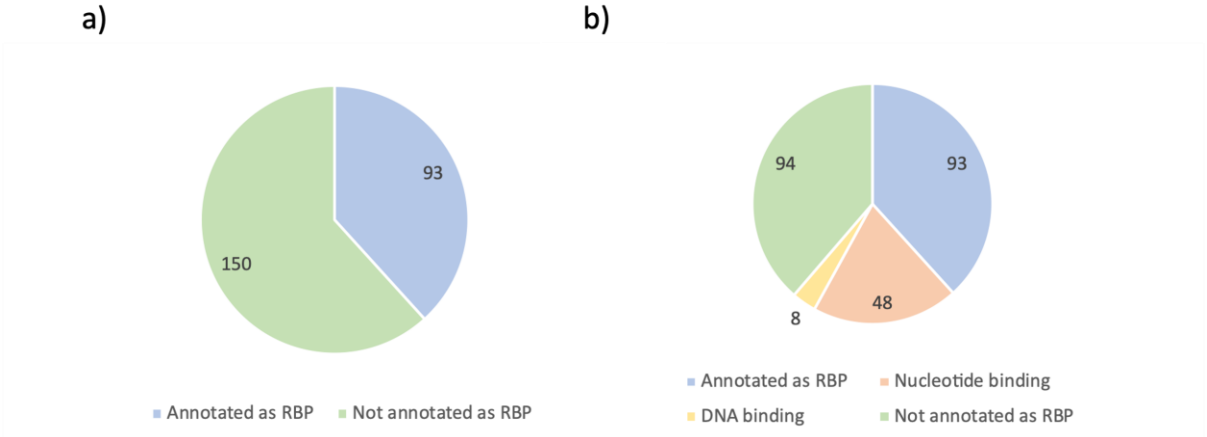


Figure 4.2-14: Pie charts of GO molecular functions of crosslinked proteins in *Escherichia coli*

a): Ratio between annotated RNA binding proteins in the UniProt database vs. Proteins with not known RNA binding molecular functions the UniProt database. b) Additional two molecular binding categories were added manually to the 'Not annotated as RNA binding protein' category: nucleotide binding and DNA binding.

Further two molecular function categories were added manually to the "not annotated as RNA binding" category (Figure 4.2-14 b): 'DNA binding' and 'nucleotide binding' proteins". Several proteins which are known to bind to RNA are binding to nucleotides as well, for instance metabolic enzyme glyceraldehyde-3-phosphate dehydrogenase, GAPDH [97]. A considerable number of proteins (48) were found to be nucleotide binding in the 'not annotated as RNA binding protein' category. It is likely that these proteins have shared functions between RNA binding and nucleotide binding.

To lesser extent, eight proteins with known DNA binding functions but not annotated RNA binding functions were identified, such as integration host factor subunit beta [98] or manganese superoxide dismutase [99].

To understand why there is a significant number of proteins (94) with not known RNA binding functions, all, identified protein IDs were uploaded into the STRING [79] database. STRING uses functional classification systems such GO (Gene Ontology) [100] or KEGG (Kyoto Encyclopedia of Genes and Genomes) [81] and performs functional enrichment analysis. Classification of the proteins, their molecular function and their participation in biological processes can help to have an overview of a dataset. Functional enrichment analysis results of the molecular function terms were sorted based on the FDR values, with a cutoff of 1%. Top 10 hits are shown on Figure 4.2-15a.

a)

#term ID	term description	false discovery rate
GO:0003723	RNA binding	3.50E-41
GO:0003735	structural constituent of ribosome	1.45E-36
GO:0097159	organic cyclic compound binding	2.33E-28
GO:1901363	heterocyclic compound binding	2.33E-28
GO:0005488	binding	4.92E-28
GO:0019843	rRNA binding	1.38E-25
GO:0003676	nucleic acid binding	7.92E-19
GO:0036094	small molecule binding	3.02E-08
GO:0140098	catalytic activity, acting on RNA	7.90E-08
GO:0000166	nucleotide binding	9.60E-08

b)

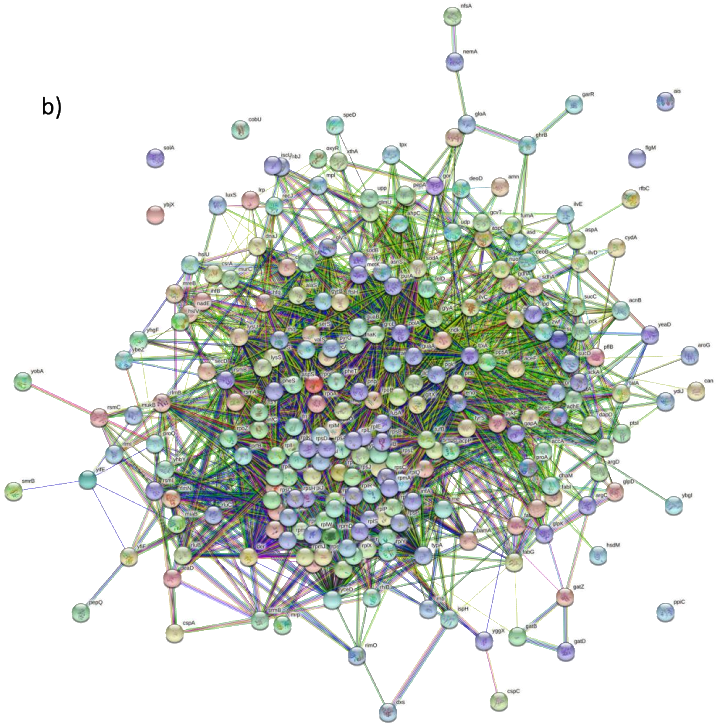


Figure 4.2-15: Molecular function analysis results and protein network of the formaldehyde crosslinked proteins

a) Top 10 hits of GO molecular function enrichment analysis with 1% FDR cutoff of the silica enriched formaldehyde crosslinked proteins in *Escherichia coli*. b) Protein interaction network of the identified crosslinked proteins in *Escherichia coli*.

Most significantly enriched proteins are RNA binding proteins, including ribosomal proteins, therefore rRNA binding term is significantly enriched as well. As it is shown in Figure 4.2-15b, local cluster of the ribosomal proteins in the protein interaction map is visible.
To see if these pathways are significantly enriched, KEGG pathway analysis was performed (in the framework of STRING). 21 pathways were found to be enriched in the KEGG pathway analysis, whereas 15 of them have less than 1% FDR (Table 4.2-1)

Table 4.2-1: Results of KEGG pathway analysis in the silica enrichment strategy

Names of each column are as follows: term ID: KEGG pathway ID; term description: name of the KEGG pathway; count in network: number of proteins annotated with the associated term out of all proteins associated with the term in the pathway; strength: log10(observed/expected), shows how large is the enrichment in the network; false discovery rate: shows how significant is the enrichment in the network (p-values corrected with the Benjamini–Hochberg procedure.)

#term ID	term description	count in network	strength	false discovery rate
eco03010	Ribosome	53/56	1.21	2.38E-37
eco01130	Biosynthesis of antibiotics	43/209	0.55	3.18E-10
eco01100	Metabolic pathways	82/708	0.3	2.53E-08
eco01200	Carbon metabolism	27/108	0.63	5.21E-08
eco01110	Biosynthesis of secondary metabolites	45/301	0.41	3.97E-07
eco00970	Aminoacyl-tRNA biosynthesis	13/25	0.95	7.09E-07
eco01230	Biosynthesis of amino acids	23/116	0.53	1.77E-05
eco03018	RNA degradation	9/16	0.98	4.01E-05
eco01120	Microbial metabolism in diverse environments	37/278	0.36	5.83E-05
eco00620	Pyruvate metabolism	14/52	0.66	8.28E-05
eco00020	Citrate cycle (TCA cycle)	10/27	0.8	0.00015
eco00010	Glycolysis / Gluconeogenesis	12/43	0.68	0.00022
eco00230	Purine metabolism	17/91	0.5	0.00045
eco01210	2-Oxocarboxylic acid metabolism	8/26	0.72	0.0024
eco03020	RNA polymerase	4/4	1.23	0.0029

Some result of the KEGG pathway does not come as a surprise, such as the term ‘ribosome’. In agreement with the molecular function enrichment analysis, several metabolic pathways were enriched in the analysis including glycolysis (12/43) or the citrate cycle (10/27) (shown in the Table 4.2-1)

In the last one decade, known proteins with catabolic functions in metabolic pathways have been shown to function as RNA binding proteins. Those enzymes, which carry out functions apart from their activity in metabolic processes are called moonlighting enzymes [97]. Typical example is the previously mentioned protein GAPDH. GAPDH is known for its function in glycolysis, where it catalyzes the conversion of glyceraldehyde-3-phosphate into d-glycerate-1,3-bisphosphate. GAPDH has also been shown to bind to RNA [101]. GAPDH participates in several processes from transcription to protection telomeric DNA [102]. Other example of moonlighting enzyme is the Aconitate hydratase B from the citrate cycle [103], where it catalyzes the conversion of citrate to isocitrate. Aconitate hydratase B moonlights as its apo-protein and binds to its own mRNA under lower iron levels [104]. Another example for moonlighting enzymes is carbon storage regulatory protein CsrA [105], or Enolase. Enolase was identified in the RNA degradosome in bacteria [97].

4.2.3.2.1 Glycolysis

Further investigation was performed to discover how many, and which proteins participate in the glycolysis and are RNA binding as well.
Five enzymes participating in the Embden-Meyerhof-Parnas pathway [106] have been identified as RNA crosslinked proteins (Figure 4.2-16). Fructose-diphosphate aldolase (fbaA) was identified as RNA crosslink, which catalysis the fructose 1,6-disphosphate conversion into dihydroxyacetone phosphate and glyceraldehyde 3-phosphate. Enzymes of consecutive steps in the glycolysis were identified, triosephosphate isomerase (tpiA) which catalysis the dihydroxyacetone phosphate to glyceraldehyde-3-phosphate conversion. As well as the glyceraldehyde 3-phosphate dehydrogenase (gapA) which catalysis the glyceraldehyde-3-phosphate, 1,3-disphosphoglycerate conversion. Phosphoglycerate kinase (pgk) was identified as well as crosslinked protein. Pgk is responsible for the conversion of 1,3-disphosphoglycerate into 3-phosphoglycerate. Phosphoglycerate mutase (pgmA) was not identified as RNA binding protein in the current study. The enzymes catalyze the steps of the conversion of 3-phosphoglycerate into 2-phosphoglycerate. Two more enzymes were identified as RNA crosslinked proteins, which are participating in two consecutive steps of the glycolysis: enolase (eno), which catalysis of the 2-phosphoglycerate 2-phosphoenolpyruvate conversion and pyruvate kinase (pykF) which catalysis the 2-phosphoenolpyruvate, pyruvate conversion.
Pyruvate then can be converted into Acetyl-CoA and being used in the TCA cycle.

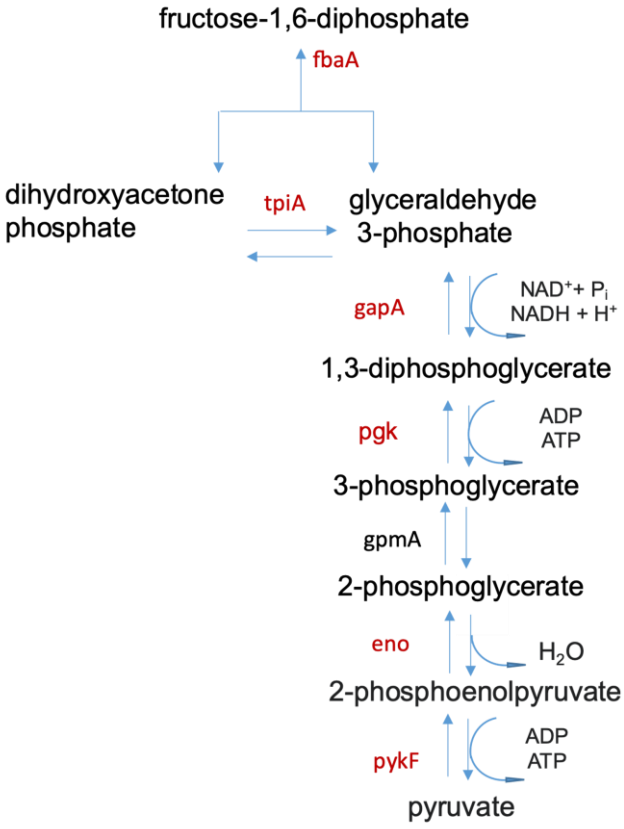


Figure 4.2-16: Consecutive steps of glycolysis between dihydroxyacetone phosphate to pyruvate conversion adapted from [106]

Steps, where at least one of the enzymes was found to be crosslinked to RNA are highlighted in red.

4.2.3.2.2 RNA binding domains

A small subset of the identified proteins have known RNA binding domains. Nevertheless, a subset of the identified proteins have known RNA binding domains, that fall under three categories: K homolog (KH) RNA binding domain, S1 like RNA binding domain and S4 RNA binding domain containing proteins. Crosslink peptides of the respective proteins were mapped onto their protein sequences to see if the crosslink peptides overlap with the RNA binding domains.

KH domain

Three proteins were identified with K homolog (KH) RNA binding domain: transcription termination/antitermination protein NusA, polyribonucleotide nucleotidyltransferase and 30S protein S3.

The crosslink peptide (aa 245-255) of transcription termination/antitermination protein NusA is located in the KH1 binding domain (aa 230-293). Polyribonucleotide nucleotidyltransferase has also one KH domain (aa 553-612) and one S1 domain (aa 622-690). The protein was identified with two crosslink peptides, one crosslink peptide was found at aa positions 427-458 and the other crosslink peptide is located at aa positions 337-345 in the sequence, which are not located in the known RNA binding domains.

In the case of 30S protein S3, 10 crosslink peptides were found. Only two crosslinked peptides were identified in the KH domain (aa 37-107) aa 90-114 and aa 66-79.

S1 like RNA binding domain

Seven proteins were identified with S1 like RNA binding domain (also named as S1 motif): transcription termination/antitermination protein NusA, ribonuclease E, translation initiation factor IF-1, polyribonucleotide nucleotidyltransferase, exoribonuclease 2, 30S ribosomal protein S1and protein YhgF. Transcription termination/antitermination protein NusA and polyribonucleotide nucleotidyltransferase do not have crosslinked peptides in the S1 like RNA binding domains.

RNAse E was identified with four crosslinked peptides, one crosslinked peptide at aa positions 76-91 was identified in the S1 motif (39-119). Translation initiation factor IF-1 was identified with two crosslinked peptides, both (aa 1-22 and aa 24-39) overlap with the S1 motif (aa 2-72). Exoribonuclease 2 was identified with two crosslink peptides aa 339-346 and aa 352-359, none of them overlap with the S1 motif (aa 561-643). Similarly, protein YhgF has an S1 motif (aa 651-720) and two crosslinked peptides of protein (aa 367-377 and aa 546-555) that do not overlap with the motif. 30S protein S1 was identified with 13 crosslinked peptides. 30S protein S1 has six S1 motifs, four crosslinked peptides overlap with the S1 motifs. Crosslinked peptides at aa 129-142 and aa143-150 positions overlap with the S1 motif 2 (aa 105-171), crosslinked peptide aa 228-244 overlaps with the S1 motif 3 (aa 192-260), crosslinked peptide aa 411-428 overlaps with S1 motif 5 (aa 364-434) and crosslinked peptide aa 465-485 overlaps with S1 motif 6 (aa 451-520).

S4 RNA binding domain

Four proteins with S4 RNA binding domains were identified: Tyrosine--tRNA ligase, ribosomal large subunit pseudouridine synthase B and C and 30S ribosomal protein S4. Tyrosine -tRNA ligase was identified with two crosslinked peptides at aa positions 61-86 and 215-231, none of which overlaps with the protein's S1 like domain (aa 357-414).

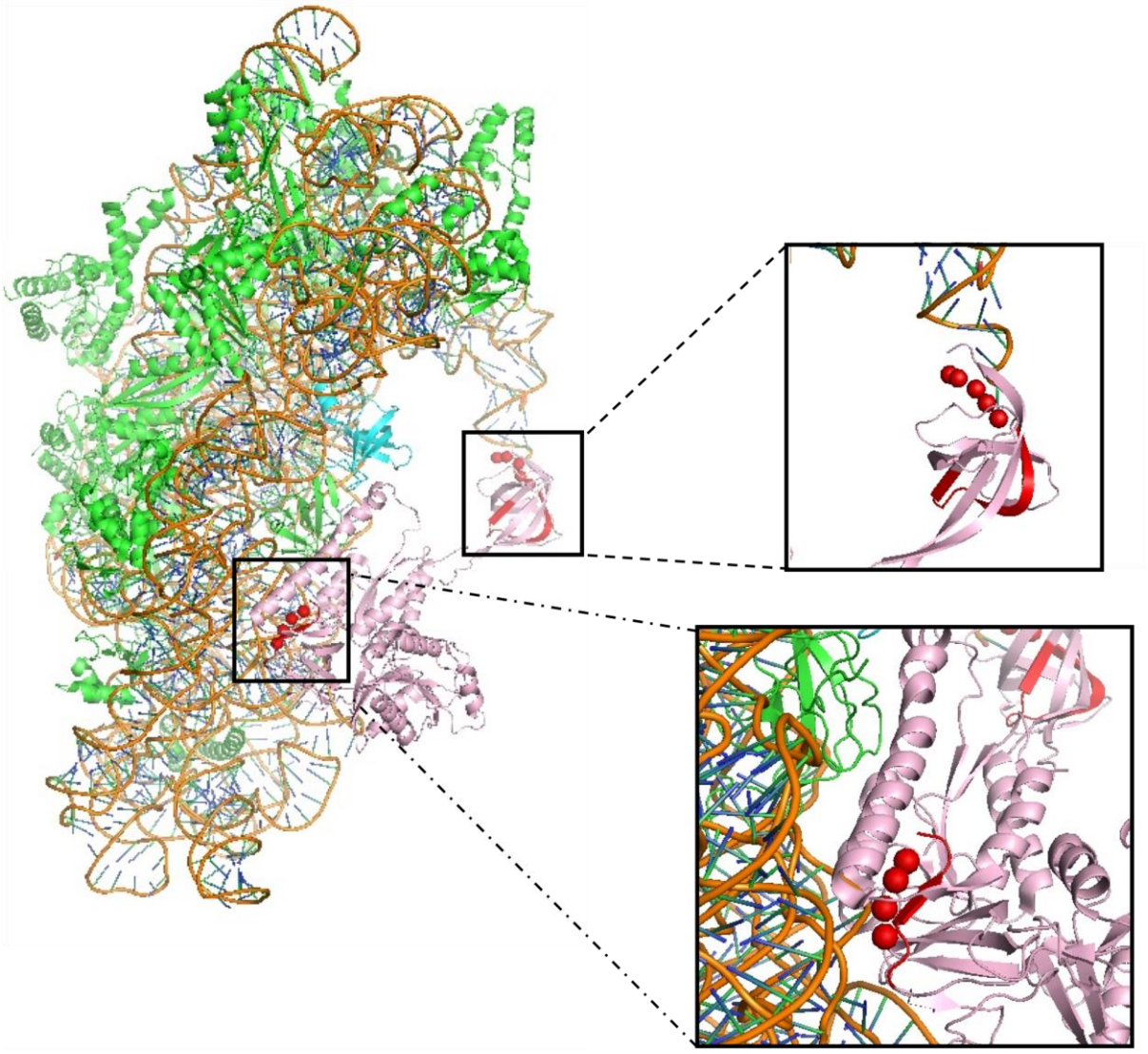
Pseudouridine synthase B was identified with two crosslink peptides (aa 69-82 and aa 109-128) the second peptide overlaps with the RNA binding motif (aa 124-152). Pseudouridine synthase C was identified with one crosslinked peptide at aa positions 246-259 - the peptide sequence does not overlap with the S4 motif (aa 28-80). 30S protein S4 was identified with five crosslinked peptides. One crosslinked peptide overlaps with the S4 motif (aa 96-156), crosslinked peptide aa 116-128.

Nucleotide binding domains
Crosslinking sites of the nucleotide-binding proteins and the nucleotide binding sites were mapped onto the protein sequences to investigate if the known nucleotide binding domains overlap with the identified crosslinking sites. Minority, 5 out of 60 proteins were identified with overlapping crosslink and nucleotide binding sites. Four out of five of them were found to be crosslinked to the same nucleotide as the nucleotide they bind to. For instance, serine--tRNA ligase was found to be crosslinked to adenosine (aa 355-368) at the same domain (aa 355-358) where serine--tRNA ligase interacts with ATP. *Castello et al.* [96] have proposed that metabolic enzymes act as RNA binding proteins might utilize their nucleotide binding domains to interact with RNA, as it has been demonstrated with the metabolic enzyme GAPDH. Interaction to RNA is not only transduced through known RNA binding domains, therefore, experimental mass spectrometric detection can be beneficial for the identification of proteins with not know RNA binding function.

4.2.3.3 Identified crosslinked peptides are in good agreement with available 3D structures

Two example structures were chosen to see if the identified crosslinks are in good agreement with the published X-ray/cryo-EM structures. The identified crosslinked peptides of elongation factor Tu2 were mapped on the structure of 70S ribosomal bound elongation factor Tu2 in *Escherichia coli*. The other chosen example is the 30S initiation complex, containing the 30S small subunit and translation initiation factor IF1 and IF2. The found crosslink sites were mapped onto the cryo-EM structure of the 30 S initiation complex in *Escherichia coli* (Figure 4.2-17).

Crosslinks identified in the translation initiation factor IF2 were mapped onto the structure of 30S initiation complex (pdb: 6O7K) [107]. Two crosslink peptides were identified in the translation initiation factor IF2: FGAIAGCMVTEGVVK and GDIVLCGFEYGR, both crosslinked to adenosine monophosphate. In the case of the first peptide, crosslink site could not be narrowed down to a single amino acid, but two shifted ions were identified, b4 and b5, which narrow down the crosslink site to the FGAIA amino acid stretch. Similarly, in the case of the GDIVLCGFEYGR peptide, shifted ion series of y8, y9 and y10 as well as y4 were present. Based on these fragments, the crosslink site could localize on the EYGR amino acid stretch.



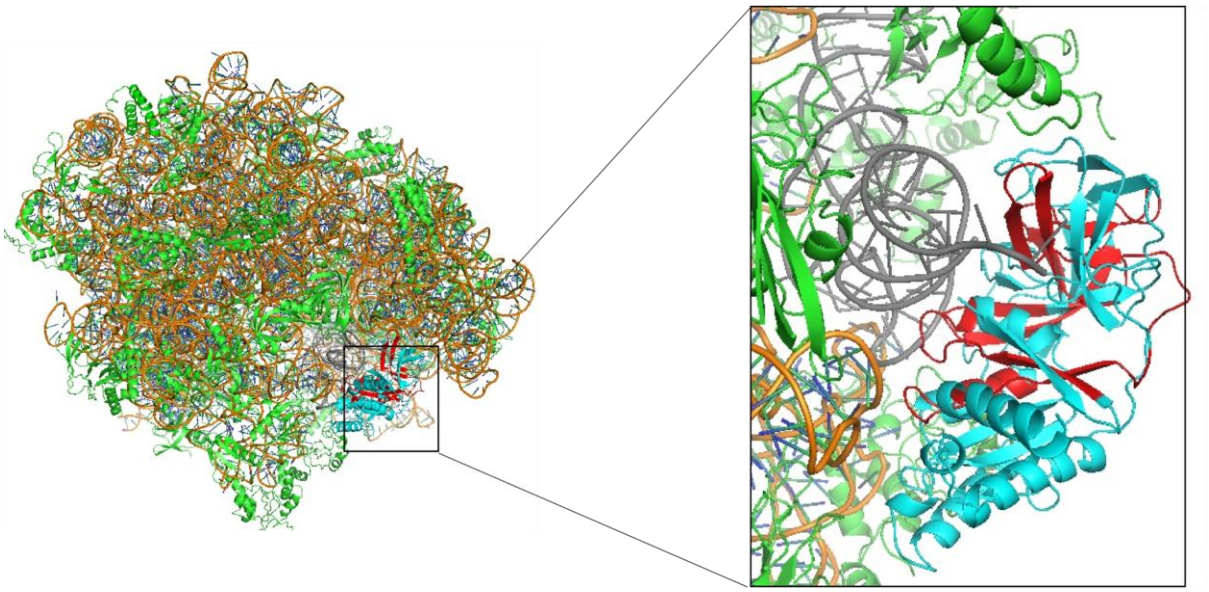
pdb: 6O7K

Figure 4.2-17: Crosslinked peptides of the Translation initiation factor IF2, in the 30S initiation complex in pdb 6O7K [107]

Translation initiation factor IF1 is colored in cyan, Translation initiation factor IF2 is colored in light pink, ribosomal proteins are colored in green, RNA is colored in golden brown. Crosslinked peptides are highlighted in red, crosslinked amino acids are shown as red spheres.

As it is shown in Figure 4.2-17, both crosslinked peptides of the Translation initiation factor IF2 are in close proximity to the RNA. Both crosslink peptides are localized on a beta sheet. The whole tryptic peptides are highlighted in red, crosslink sites were localized on a shorter stretch of amino acids, the C- alpha atoms of the crosslinked amino acids are shown as red spheres.

Crosslink peptides of elongation factor Tu2 are not unique, the peptides belong to elongation factor Tu1 too, as the two proteins differ in a single amino acid. The two proteins are often mentioned under one name and not distinguished. Therefore, these crosslinks were taken into consideration and mapped onto the cryo-EM structure of 70S ribosome-elongation factor Tu2 complex (pdb:5afi [108]).



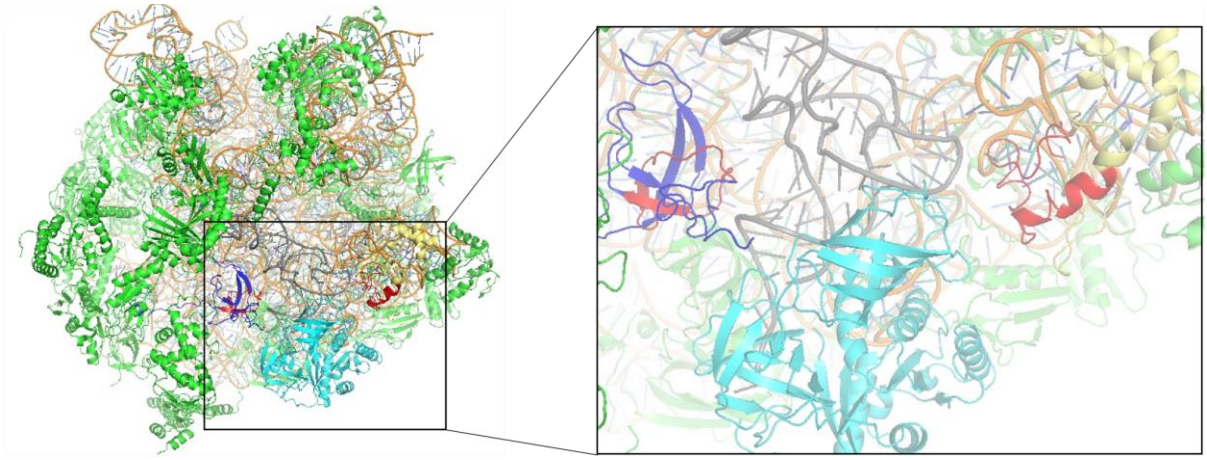
pdb: 5afi

Figure 4.2-18: Crosslink sites between the 70S ribosome bound elongation factor Tu2 and Phe-tRNA^{Phe} in pdb 5afi [108]

Elongation factor Tu2 is colored in cyan, crosslinked peptides are highlighted in red. Aminoacyl-tRNA is colored in gray, RNA is colored in golden brown.

Eight crosslinked peptides were mapped on the structure (Figure 4.2-18), crosslinked peptide sequence are as follows: AGENVGVLLRGIK (aa 270-282), AIDKPFLLPIEDVFS-ISGRGTVVTGR (aa 205-230), EHILLGR (aa 117-123), ELLSQYDFPGDDTPIVRGSALK (aa 155-176), GITINTSHVEYDTPTR (aa 59-74, HYAHVDCPGHADYVK (aa 75-89), MVVTLIHPIAMDDGLR (aa 358-373) and TKPHVNVGTIGHVDHGK (aa 8-24).
The tRNA bound to the 70S ribosome (colored in gray in Figure 4.2-18) is in close proximity to the crosslinked peptide AIDKPFLLPIEDVFSISGRGTVVTGR which is localized on one of the beta sheets of the elongation factor Tu2. Another crosslinked peptide of elongation factor Tu2, TKPHVNVGTIGHVDHGK is close proximity to the 23S RNA (colored in golden brown Figure 4.2-18).

Additional two crosslink sites were mapped onto the same cryo-EM structure of the 70S bound elongation factor Tu2, the crosslinked peptides of 50S ribosomal protein L11 and 30S ribosomal protein S12. These two proteins are in close proximity to the Phe-tRNA^{Phe} (Figure 4.2-19).



pdb: 5afi

Figure 4.2-19: Crosslink sites in 70S ribosome bound elongation factor Tu2 and aminoacyl-tRNA

Elongation factor Tu2 is colored in cyan, 50S ribosomal protein L11 is colored yellow, 30S ribosomal protein S12 is colored in blue, Phe-tRNA^{Phe} is colored in gray. Respective crosslinked peptides are highlighted in red.

50S protein L11 (colored yellow in Figure 4.2-19) was identified with one crosslink peptide: LQVAAGMANPSPPVGPALGQQGVNIMEFCK (aa 11-40). This peptide is located close to the aminoacyl-tRNA. Crosslinked peptide LTNGFEVTSYIGGEGHNLQEHSVILIR (aa 57-83) of the 30S protein S12 (colored in blue in Figure 4.2-19) is similarly close to the aminoacyl tRNA.

The above shown examples support that formaldehyde crosslinking captures interactions between proteins and RNA which have been close proximity to each other.

4.2.4 *In vivo* crosslinking of HeLa cells

Mass spectrometric based investigations of *in vivo* formaldehyde crosslinking of eukaryotic cells was performed. Detailed biochemical sample preparation is in section 3.2.9, briefly as follows: HeLa cells were crosslinked with 1% (v/v) formaldehyde, cells were lysed (lysis condition was adapted from Bae *et al.* [62] Cells were ultrasound disrupted and proteins were digested with trypsin overnight. A two-step silica enrichment protocol was used for the enrichment of RNA-peptide heteroconjugates, in between the silica enrichment step, trypsin digestion was performed, followed sequential RNA digestion. Two experiments were performed on the same number of cells. In the second experiment, lower temperatures were used for protein digestion as well as additional enzymes were used in the second RNA digestion, in the sequential RNA digestion step (see section 3.2.9.2) Another difference between the two samples was that the first sample was analyzed by HPLC-MS/MS after the second RNA digestion step, whereas the second sample was fractionated by basic reversed-phase chromatography (BRP) and the BRP fractions were analyzed by HPLC-MS/MS. The first sample will be referred as "Unfractionated Sample", the second sample will be referred as "Fractionated Sample" in the following sections. Both samples were analyzed with MSFragger with mass offset setup and marker ion search was used for database search (see section 3.3.3.3, section 3.3.3.4.2 and Supplementary Text 4) of formaldehyde crosslinked peptide-RNA heteroconjugates. In the case of "Unfractionated Sample" two injection replicates of LC-MS/MS were performed and both respective datasets were used for data analysis with MSFragger. In the case of "BRP Sample", all, 24 fractions derived from BRP were analyzed by LC-MS and 24 raw MS files were submitted to data analysis using MSFragger.
In both experiments a considerable number of formaldehyde crosslinked proteins (357 for Unfractionated and 355 for fractionated) were identified (Extended Supplementary Table 2 and Extended Supplementary Table 3). As expected, predominantly classical RNA binding proteins were identified, including the constituent proteins of the 80S human ribosome, translation initiation factors, RNA helicases, elongation factors, heterogenous ribonucleoproteins, as well as few constituent members of the splicesosomal complexes.

4.2.4.1 Unfractionated sample

Unfractionated sample identified proteins' molecular function were visualized on a sunburst diagram. PANTHER Classification system (Protein Analysis Through Evolutionary Relationships) [80] was used to assign for molecular function identification and classification. Since 357 proteins were identified, it is not possible to visualize all proteins on the same diagram, therefore only subcategories of the proteins' molecular functions are shown without listing the proteins' UniProt identifiers.

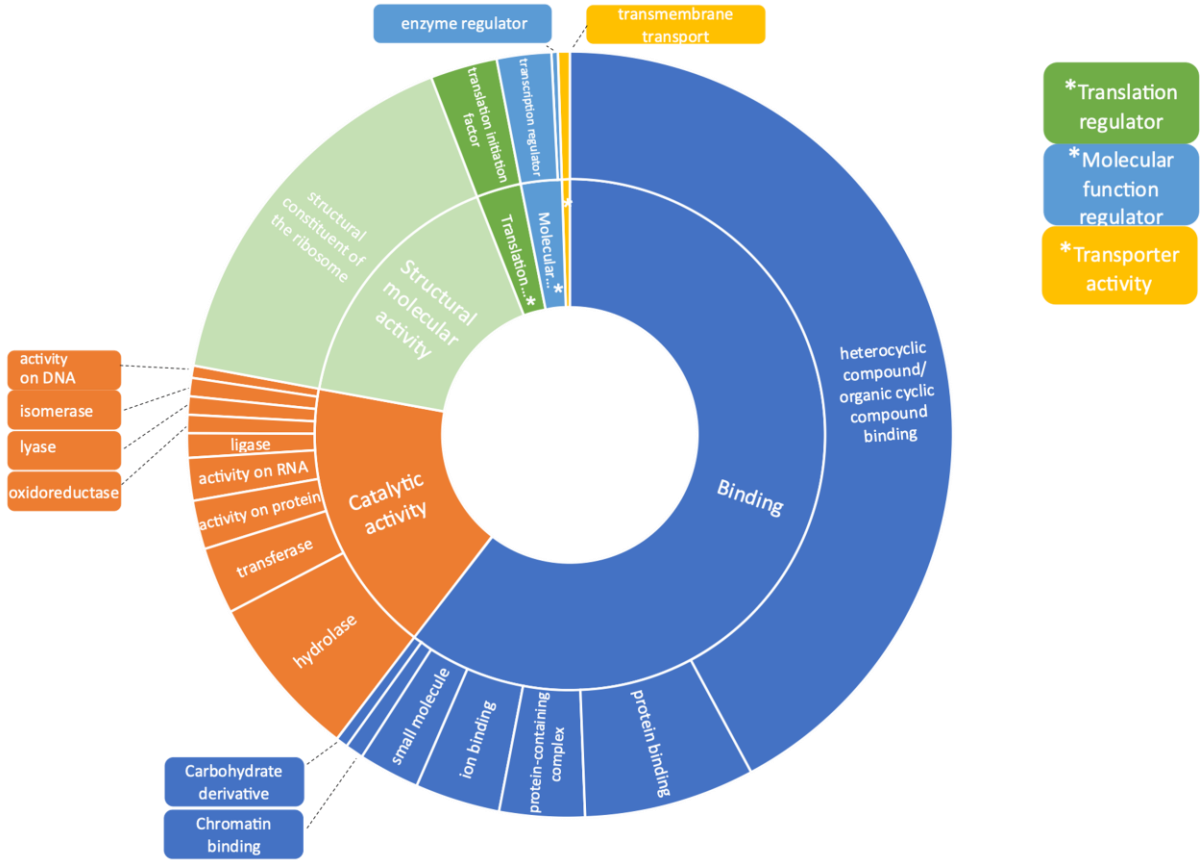


Figure 4.2-20: Sunburst diagram of the molecular functions of crosslinked proteins from formaldehyde crosslinked HeLa cells

Molecular function annotation was performed with the PANTHER classification system [80]. HeLa cells were formaldehyde crosslinked, *in vivo*. Cells were lysed, and proteolytic digestion took place, overnight. Two-step silica enrichment strategy was performed, in between the enrichment steps, proteolytic digestion was used. After silica enrichment, a two-step RNA digestion was performed. Sample was subjected to C18 reversed-phase chromatography and measured by HPLC-MS/MS analysis.

As it is shown in Figure 4.2-20, the majority of the protein identifications reveal the molecular function of ‘binding’. ‘heterocyclic compound/organic compound binding’. This molecular function category contains all RNA, DNA, and nucleotide binding proteins. The second biggest category is ‘structural molecular activity’ where all ribosomal proteins are located, including ribosomal proteins from the mitochondrion. The third biggest category is ‘catalytic activity’, containing RNA editing enzymes such as double-stranded RNA-specific adenosine deaminase, hydrolases such as Ubiquitin carboxyl-terminal hydrolase 10. Lesser number of proteins are in the ‘translation regulator activity’ category, where all initiation factors belong, such as eukaryotic translation initiation factor subunit A, B and D. The least expected category is ‘transporter activity’, proteins (chloride intracellular channel protein 1 and sodium channel protein type 1) under this category that do not have reported RNA binding functions.
The identified crosslinked proteins were uploaded into STRING [79] and functional enrichment analysis was performed. Protein interaction map was created with STRING. (Figure 4.2-21).

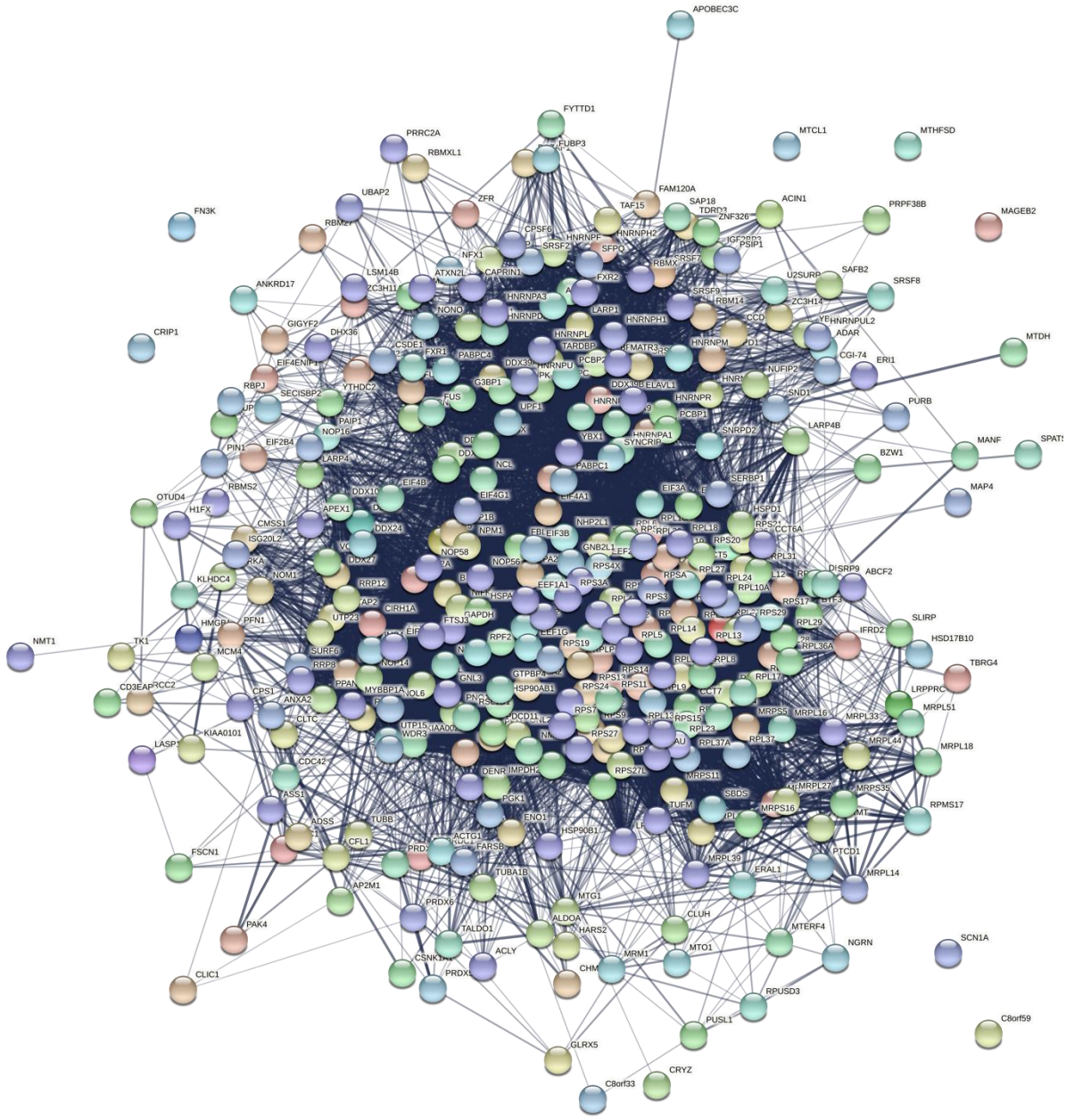


Figure 4.2-21: Protein interaction network of formaldehyde crosslinked proteins to RNA in HeLa cells, protein network was created with the STRING database [79]

As it is shown on Figure 4.2-21, one main cluster is visible in the protein interaction network, containing 80S ribosomal proteins, elongation factors, etc. Local cluster analysis of the STRING database shows that proteins participate in ‘peptide chain elongation’ and ‘Eukaryotic translation elongation’.
KEGG [79], [81] pathway analysis (also performed by STRING) showed that, splicing is the second most significantly enriched pathway in the dataset. 26 proteins related to this term were identified in the dataset. Among others, Serine/arginine-rich splicing factor 1, 2, 7 and 8 were identified, these proteins are participating in the pre-mRNA splicing. Small nuclear ribonucleoprotein Sm D2 was identified as well, this protein is part of the U1, U2, U4 and U5 snRNPs.

4.2.4.2 RNA binding domains

Out of the 357 number of proteins, 147 proteins were identified with known domain. Only a small fraction of them has RNA binding domains, such as RRM, KH, DEAD/DEAH box helicase domains. These proteins were taken under closer investigation; crosslinked peptide sequences were mapped onto their respective protein sequences. It is challenging to map a crosslinked peptide sequence on the respective binding domain, because crosslinked amino acid could not be localized, only crosslinked peptide sequences are available. In other words, the resolution of the crosslinking sites very low. The following procedure was used for crosslink peptide mapping: the crosslinked peptide's first amino acid position and last amino acid position were compared to the first and last amino acid position of the RNA binding domain, if any overlap was found it was considered a match.

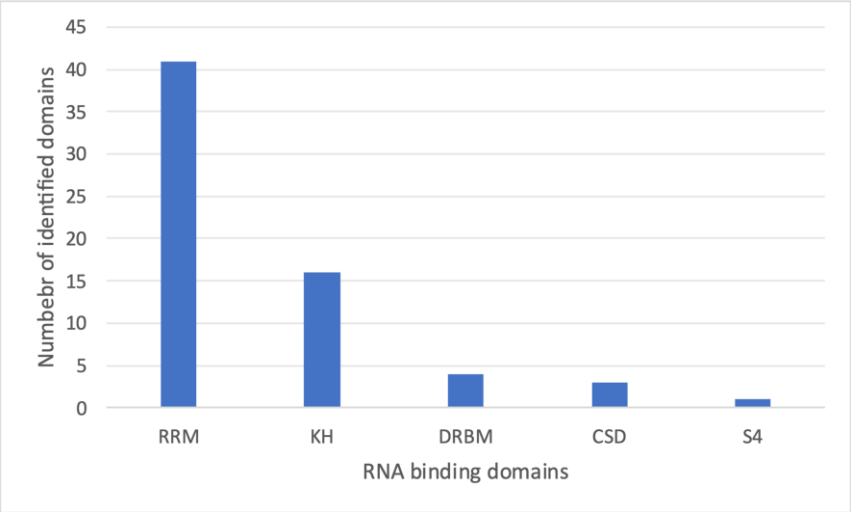


Figure 4.2-22: Number of identified RNA binding domains of crosslinked proteins

Crosslinked peptides in the various proteins are located manually

With this strategy, 30 proteins were identified with one or more RRM domain matching to their respective crosslink peptides. In total 41 RRM domain matches were found. 9 proteins were identified with KH domains, containing 16 crosslinked KH domains (Figure 4.2-22). A single S4 binding domain match was identified. 4 proteins were identified with overlap between their DRBM domain and the identified crosslinked proteins, three proteins with cold shock domain (CSD). The number of crosslinks, which fall into a known protein domain are considerably low in comparison to the total number of crosslinked peptides identified; this might be due to the fact the crosslink sites fall into not classical RNA binding domains.

4.2.4.3 Basic severed phase fractioned sample

Basic reversed-phase fractionation (BRP) was performed on another set of silica enriched formaldehyde crosslinked peptide-RNA heteroconjugates. Fractionation was performed to increase the LC-MS identification of very low abundant RNA binding proteins. When identified proteins of fractionated and unfractionated proteins were compared (Figure 4.2-23), only two third of the proteins have been found to be common (252).

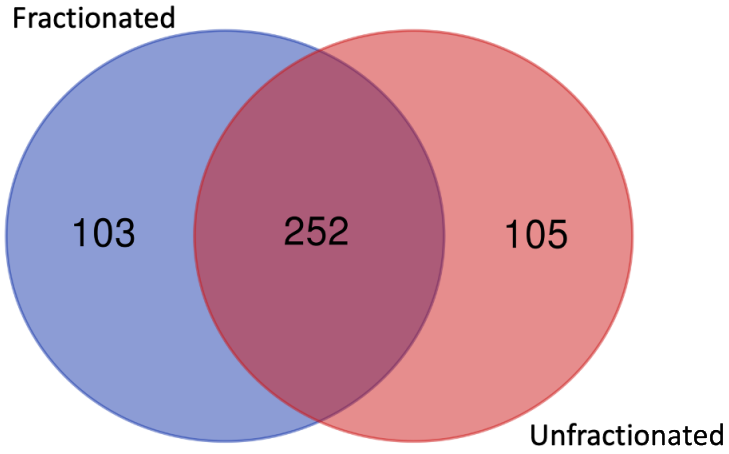


Figure 4.2-23: Venn diagram of number of proteins identified in the formaldehyde crosslinked HeLa cells between the Unfractionated and Fractionated samples

Unfractionated sample was processed with the silica enrichment strategy, fractionated sample was also processed with the silica enrichment strategy and basic reversed-phase chromatography took place for sample complexity reduction.

It was expected that the basic reversed-phase fractionation would result in larger number of crosslinked proteins, but the LC-MS and subsequent MSFragger data analysis did not support this assumption. First, crosslink identification of the basic reversed-phase fractionated sample was uploaded into PANTHER classification system [80], as well as into the STRING database [79]. Functional enrichment analysis was performed, to conform that the identified proteins fall under the same molecular function categories as in the case of the unfractionated sample.

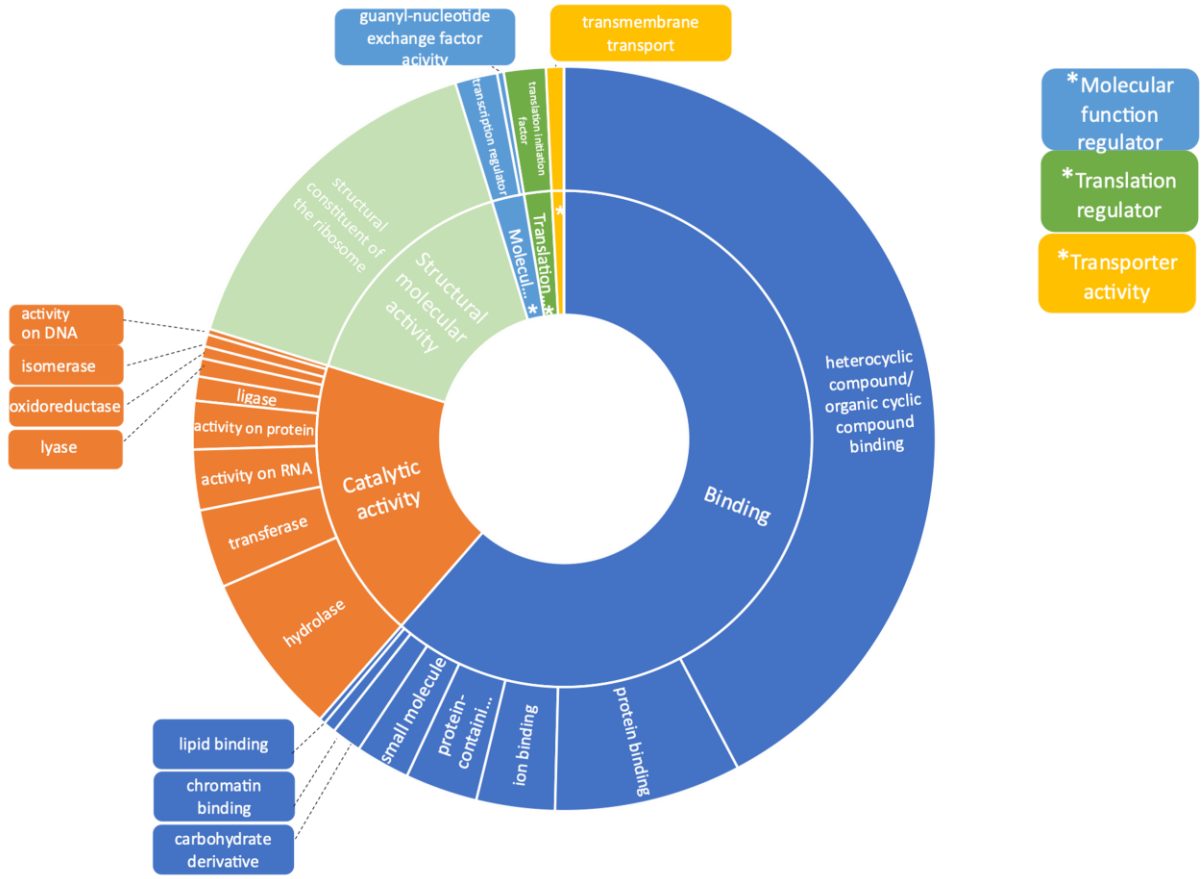


Figure 4.2-24: Sunburst diagram of the crosslinked proteins' molecular functions in formaldehyde crosslinked HeLa cells, fractionated sample

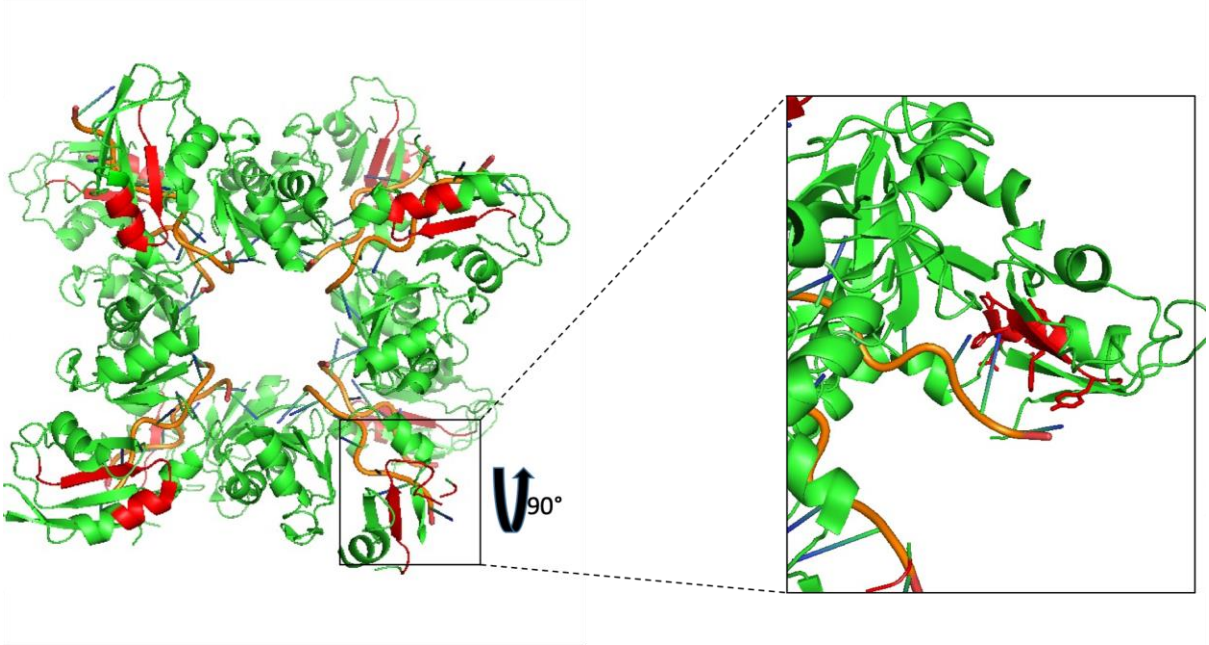
Molecular function annotation was performed with the PANTHER classification system [80]. HeLa cells were formaldehyde crosslinked, *in vivo*. Cells were lysed, and proteolytic digestion took place, overnight. Two-step silica enrichment strategy was performed, in between the enrichment steps, proteolytic digestion was used. After silica enrichment, a two-step RNA digestion was performed. Sample was subjected to C18 reversed-phase chromatography, followed by basic reversed-phase chromatography fractionation of the peptide-RNA heteroconjugates. The resulted twenty-four fractions were measured by HPLC-MS/MS analysis.

Figure 4.2-24 shows that molecular functions of the fractionated sample are very similar to the unfractionated sample (Figure 4.2-20). Similar inferences are drawn from the molecular function classification of the identified proteins (Figure 4.2-24). The three main categories of the molecular functions are 'binding', 'structural molecule activity' and 'catalytic activity'. This aligns with the molecular function categories of unfractionated sample. To identify the differences between the fractionated and unfractionated sample, further investigation took place, with more focus on the proteins identified in one sample and not identified in the other sample. Unfractionated sample contains more ribosomal proteins, including but not limited to: 80S ribosomal proteins (S13, S26, S30, L34, L28, L37, L37a, L39) as well as ribosomal proteins from mitochondrion, which have not been found in the fractionated sample (S11, S16, S17, L39, L44). Unfractionated sample contains different heterogeneous nuclear ribonucleoproteins than fractionated sample.
Fractionated sample contains more RNA processing proteins such as ribosome production factor 1 and RNA-specific editase 1. BRP sample also contains nine mitochondrial ribosomal proteins (S9, S14, S25, S34, L13, L35, L42, L50, L51) which have not been identified in the unfractionated sample. It is important to mention, two histone proteins were identified in the fractionated sample: histone protein H2.A type 1 and H2A variant z, whereas in the unfractionated sample linker histone H1.10 was identified.

The differences of the results between the two HeLa samples might come from the lower reproducibility of the enrichment method, because RNA degrades easily, or the difference might come from sample loss during fractionation or the lower reproducibility of the MS data dependent measurements or superposition of all mentioned above. Regardless, the total number of identified proteins from the crosslinked HeLa cells is 460.
Proteins have been identified from several cellular compartments of the cell, majority of the proteins are localized int the cell nucleus (heterogeneous nuclear ribonucleoproteins) and/or in the cytoplasm (ribosomal proteins), proteins from mitochondrion (constituent proteins of the mitochondrial ribosome, Elongation factor Tu, mitochondrial), even in endoplasmic reticulum (endoplasmin) were also identified. These two experiments show that formaldehyde crosslinking is a useful tool for RNA binding protein identification in eukaryotic cells.

4.2.5 Crosslinked peptides were mapped into complex structures

Crosslinked peptides were mapped into existing X-ray and cryo-EM structures of RNA-protein complexes to evaluate if crosslinked peptides are in agreement with the existing structures, i.e. whether the crosslinked peptides are adjacent. Polyadenylate-binding protein 1 (PAB1) was identified as one of the crosslinked proteins. PAB1 has four RRM domains: RRM1, RRM2, RRM3 and RRM4. Several crosslink peptides were identified from PAP1, such as GYGFVHFETQEAAER peptide, which is part of the RRM2 domain or the GFGFVCFSSPEEATK peptide localized in the RRM4 domain. Although, the latter peptide is not unique; it is a shared peptide between the poly(A)-binding protein 1, 3 and 4.
PAB1 RRM1 and RRM2 domains bind to the poly(A) tails of mRNA, RRM3 and RRM4 are nonspecifically bind to RNA. Crosslinked peptide, GYGFVHFETQEAAER of the RRM2 domain was identified as adenosine crosslink. The peptide sequence was mapped onto the crystal structure of poly(A) binding protein1 bound to polyadenylate RNA [109] (pdb: 1CVJ). The structure does not contain the full length of poly(A) binding protein, it contains only the two N-terminal RRM binding domains (1 and 2) linked with a short linker and poly(A) RNA. Poly(A) binding protein forms a homo-octamer structure together with four strands of polyadenine (Figure 4.2-25).



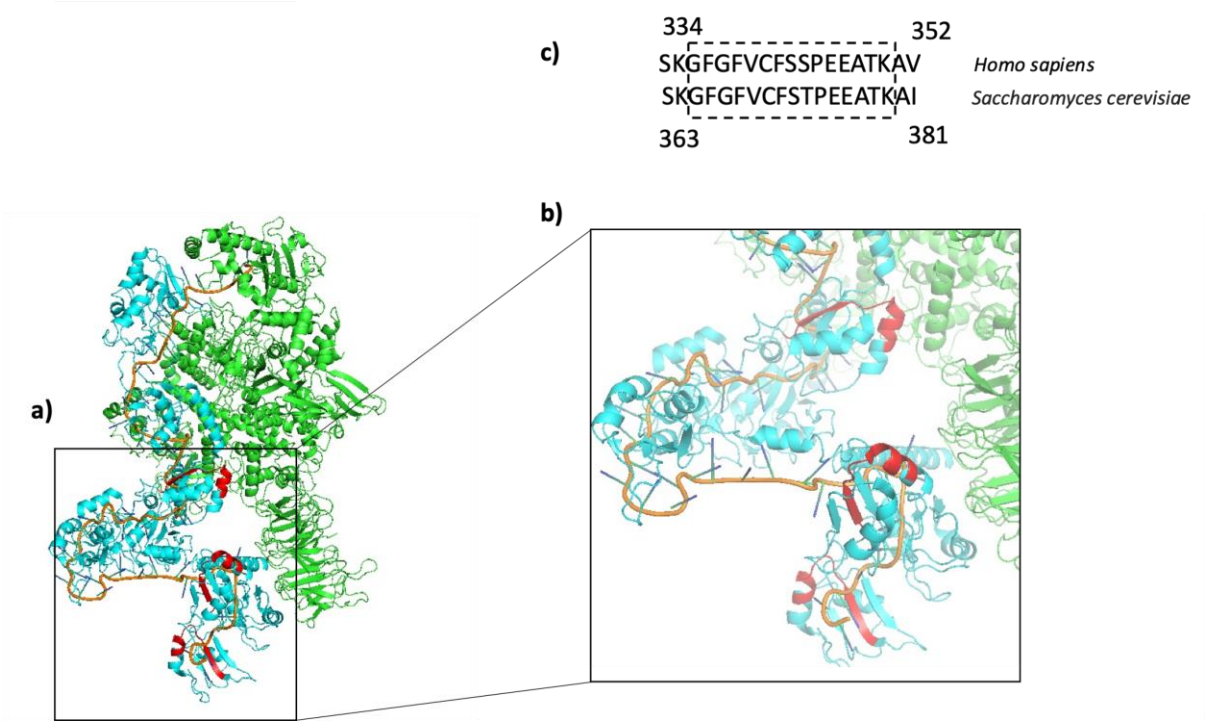
pdb: 1CVJ

Figure 4.2-25: Crosslinking sites of the Poly(A) binding protein, bound to poly(A) in pdb ICVJ

[109] a) PAB1 protein is colored in green, crosslinked peptides are highlighted in red, RNA is colored in golden brown b) 90 degrees rotation of the structure, for better visualization of the

crosslinking sites, side chains of crosslinks peptides are shown to emphasize that crosslink peptide is close proximity to RNA.

As it is shown in Figure 4.2-25, the peptide sequence in the RRM2 domain is localized on one beta sheet and partially on an alpha helix (highlighted in red). Amino acid side chains were shown to illustrate that, amino acids of the crosslinked peptide are close proximity to RNA.
Although, there is no available structure of the RRM3 and RRM4 domains of PAB1, structure of the homolog protein, Polyadenylate-binding protein (PAB1), cytoplasmic and nuclear from *Saccharomyces cerevisiae* is available. Crosslinked peptide of RRM4 was blasted against the yeast RRM4 domain, the two domains share high percentage of amino acids (Figure 4.2-26). Homologue sequence peptide was mapped onto the cryo-EM structure of poly(A) bound PAB1 together with Pan2-Pan3 deadenylase [110](pdb: 6R5K).



pdb: 6R5K

Figure 4.2-26: Crosslinking site of Poly(A) bound PAB1 protein in complex with Pan2-Pan3 deadenylase from *Saccharomyces cerevisiae* in pdb 6R5K [110]

a) Pan2-Pan3 complex is colored in green, three copies of PAB1 are colored in cyan. Crosslink peptides are highlighter in red. b) red highlighted sequences are close proximity to RNA. c) sequence alignment between crosslinked peptide of PAB1 from *Homo sapiens* and PAB1 from *Saccharomyces cerevisiae***,** crosslinked peptide sequences are shown in the dashed lined rectangular.

Pan2-Pan3 complex is colored in green, three copies of PAB1 are colored in cyan, crosslinked peptides are highlighted in red. As it is shown in Figure 4.2-26, red highlighted peptide sequences are localized close to the phosphodiester backbone of the RNA.

5 Discussion

5.1 MS/MS fragmentation of DEB-crosslinked peptide-(oligo)deoxynucleotide heteroconjugates

As it was shown in the results section, DEB crosslinks to guanine, dG, dA and dC show a similar MS/MS fragmentation pattern. Dominantly, DEB and DEB+ Nucleobase mass shifts were observed in the MS/MS spectra, although in some cases only DEB mass shift was present. The MS/MS fragmentation of the peptide-DNA heteroconjugates show a strong charge state and deoxynucleotide adduct size dependency. In the case of dT crosslinks, not only DEB and DEB+Tb mass shifts but also, other dT related mass shifts were observed: DEB+dT-HPO_3, DEB+dT-H_3PO_4, etc., These, observed mass shifts makes the DEB crosslinked dT's MS/MS fragmentation differ from the general MS/MS fragmentation. Although the DEB crosslinked dT show a more complex fragmentation pattern then the other three deoxynucleotides, the most prominent mass shifts are still DEB and DEB+Tb. DEB and DEB+Nucleobase mass shifts are simple mass adducts in the MS2 spectrum, thus advantageous for reliable crosslink identification.

5.2 Different data analysis possibilities for the DEB crosslinked datasets

In the results section, different approaches were shown for the analyses of the mass spectrometric raw data of the DEB-crosslinked protein-DNA complexes. When the MS2/MS3 method was developed, exclusively manual spectral validation was used on the RNP^{xl} results. The main goal of the MS2/MS3 method was to establish an automatic workflow and to exclude manual annotation. This approach was successful on guanine crosslinks, but due to the more complex fragmentation of the thymine crosslinks, the method was not applicable on dT-containing adducts. When the MS2/MS3 acquisition method was developed, no adenine and cytosine crosslinks were identified. This observation was revisited when higher number of DEB crosslinks could be identified in the nucleosome experiments, where adenine and cytosine crosslinks were identified alongside the guanine and thymine crosslinks. It became clear that enzymatic DNA digestion decreases the DNA adducts complexity, therefore gas phase fragmentation of the DNA adducts were not needed anymore. Additionally, it became clear that when the adducts is dT, M+DEB and M+DEB+Tb were not always the highest intense peaks, thus MS2/MS3 acquisition method could not be always used. These two observations put the MS2/MS3 method on the side.
The advised applications of the measurement methods and data analysis strategies are as follows. When longer DNA crosslinks are investigated, MS2/MS3acqusition method is a good approach for crosslinking site localization in combination with modification search. When DEB crosslinking is used for structural biology applications, DNA digestion to deoxynucleotides is advised in combination with RNP^{xl} data analysis. When DEB *in vivo* crosslinking is applied, sequential nucleic acid digestion is advised in combination with the RNP^{xl} data analysis.

5.3 Comparison of different chemical crosslinking techniques

In this work two chemical crosslinking techniques were used for protein-nucleic acid crosslinking - DEB, and formaldehyde crosslinking. DEB reacts with all four nucleobases (see section 4.1.1), whereas formaldehyde reacts with three nucleobases adenosine, guanosine, and cytosine (see 4.2.1.1 section). The crosslinking methods can be compared between the DEB and the FA crosslinked mononucleosome results. Mononucleosomes were crosslinked with DEB and formaldehyde in parallel, samples were processed and measured together. Results of the two crosslinking experiments were already shown in the results section (4.1.7.2 and 4.2.1.2 sections, respectively), here the results are used for the comparison of the two crosslinking methods. Based on the number of CSMs, in the case of DEB crosslinking the preferred identified nucleobase is guanine (49%), followed by thymine (33%), adenine (15%) and cytosine (3%). Whereas, in the case of formaldehyde crosslinking (with the exclusion of the isobaric peptides), no identification preference was observed between adenine (46%) and guanine (48%), whereas cytosine was identified in substantially lower numbers (6%). MS/MS fragmentation behavior is different when DEB or formaldehyde crosslinking is used. In the case of DEB crosslinking, DEB, and DEB+Nb mass shifts are the prominent mass shifts, except when dT crosslinks are present. Formaldehyde crosslinks have a very simple MS/MS fragmentation, which is dependent on the linker fragmentations, thus, data analysis is simpler than the data analysis of the DEB crosslinks.

Another major difference between the two crosslinking methods is the crosslinking site localization. DEB crosslinking site can be localized to a single amino acid, whereas formaldehyde crosslink localization is often ambiguous, or the crosslinking site localization is not possible. Therefore, the resolution of formaldehyde crosslinking site is much lower than the resolution of the DEB crosslinking site.

Isobaric peptides were present in both crosslinking experiments, more prominently in the FA crosslinking experiments.

High overlap was observed between the crosslinked peptides of DEB crosslinking and formaldehyde crosslinking (in the case of mononucleosome sample), which indicates that the two crosslinking methods capture the same interactions (see section 4.2.1.2).

Based on the observed features of the two crosslinking methods, DEB-mediated crosslinking is more beneficial for structural investigation of purified molecular complexes whereas formaldehyde crosslinking deemed more useful for *in vivo* crosslinking studies. Crosslinking site identification is crucial for the structural investigation of a purified protein-DNA complex; therefore, DEB crosslinking is preferred due to its precise crosslinking site localization. Whereas *in vivo* crosslinking aims for the identification of new RNA/DNA binding proteins and their respective RNA/DNA binding domains, crosslinking site identification to a single amino acid resolution, becomes less of importance.

5.4 Formaldehyde crosslinking, general observations

5.4.1 Formaldehyde linker incorporation between proteins and nucleic acids

One formaldehyde incorporation between proteins and nucleic acids is the most widely accepted view on the chemical reaction [51]. Tayri-Wilk *et al.* [111] have already shown that when formaldehyde is used for protein-protein crosslinking, often, between proteins not one formaldehyde delta mass but two formaldehyde delta masses are observed. Lu *et. al*. [54] have shown that three formaldehyde molecules can incorporate between deoxyguanosine monophosphate and peptides when they are formaldehyde treated, their finding were

supported with NMR measurements, which allowed structural investigation of the crosslinked product. Lu *et. al* have proposed a structure, where two formaldehyde molecules form a ring like structure between lysine and guanine and the third formaldehyde modifies the guanine, forming a mono-link on it (**Figure 1.6-4**). In the current study, the same, three formaldehyde-based linker composition was observed as proposed by Lu *et al*. The three-formaldehyde mass shift was observed in the datasets of formaldehyde crosslinked, *in vitro* reconstituted mononucleosomes (3% of the CSMs) and HeLa native nucleosomes (24% of the CSMs in Supplementary Table 9**,** 3FA crosslinks).

5.4.2 Formaldehyde crosslinked peptide-(oligo)nucleotide MS/MS fragmentation

The MS/MS fragmentation of the peptide(deoxy)oligonucleotide heteroconjugates are dependent on the linker cleavage between peptides and (deoxy)nucleotides. Generally, mass shifted marker ions and marker are present in the spectrum reflecting the DNA adduct composition, these marker ions and mass shifted marker ions have high intensity, they are often the base peak. In lower number of cases, marker ions or mass shifted marker ions were present with lower or low intensity. When the respective marker ion or mass shifted marker ion are not the most abundant peak in the MS/MS spectra, other phenomena are present, for instance, incomplete fragmentation of the precursor or proline effect.
Isobaric behavior of the formaldehyde crosslinked peptide-DNA heteroconjugates were described in Section 4.2.1.1.3. Two types of isobaric peptides were observed in the DNA crosslinked datasets: when crosslinked peptide was crosslinked to DNA at different amino acid positions and when peptide was crosslinked to different nucleotides in the DNA adducts. When peptides are crosslinked at different amino acid positions, non-overlapping mass shifted b and y-ions are present. When peptide was crosslinked to different nucleotides in the DNA adduct mass shifted nucleobases were present of the respective deoxynucleotides.
The same MS/MS fragmentation pattern was observed in the formaldehyde crosslinked peptide-RNA heteroconjugates, where mass shifted and unshifted marker ions were present in the MS2 spectra. The sequential RNA digestion resulted simpler RNA adducts and thus, as a tendency, more often mass shifted marker ions were present with high intensity and not pairs of mass shifted and unshifted marker ions with high intensity. In the case of guanine, adenine, and cytosine crosslinks in the MS2 spectra typically mass shifted marker ions appeared with high intensity, in the case of cytosine crosslinks in some cases unshifted marker ion were presented with high intensity in the MS2 spectra.

5.4.3 MS2/MS2 method development

Several versions of the MS2/ MS2 acquisition method were used in this study (see section 3.4.2.2). The acquisition method was originally established on the observation of the MS/MS fragmentation of the peptide-(oligo) heteroconjugates. It was observed that in crosslinked spectra, marker ions and mass shifted marker ions are present. Moreover, pairs of mass shifted and unshifted marker ions were consistently present, reflecting the composition of the deoxynucleotide adduct. In the formaldehyde crosslinked datasets, the length of the crosslinked deoxynucleotide chain was on average 2, thus often, two, high intensity peaks appeared in the low *m/z* region of the MS/MS spectrum, both related to the crosslinked deoxynucleotide adduct. For instance, if the deoxy(oligo)nucleotide adduct was dGdC and the crosslinked deoxynucleotide was dG, then mass shifted Gb marker ion and Cb maker ion appeared in the MS/MS spectrum. In order to target these peaks all marker ions and mass shifted marker ions were monitored and, if two of the target masses were found in the 10 most abundant peaks in the spectrum, another MS/MS spectrum was triggered. Later, it was observed that MS/MS fragmentation of the formaldehyde crosslinked peptide-

deoxynucleotide heteroconjugates has a charge dependence, Peptide precursor related peaks were usually present in the MS/MS spectra when crosslinks have charge state of +2. Therefore, different collision energies were used, dependent on the charge state of the crosslink. Precursor masses with +2 charge state were fragmented with higher collision energies to promote further dissociation and precursor masses of charge state between +3 and +6 were fragmented with lower collision energies. It was also observed that often not both marker ions had high intensity in the MS2 spectra but only one of them, therefore the targeting rule was changed to: if one of the target masses were found in the 10 most abundant peaks in the spectrum, another MS/MS spectrum was triggered. In the rest of the study, the method was only used for targeting peptide-RNA heteroconjugates and, due to the sequential RNA digestion, the crosslinked RNA adducts became simpler and more frequently the mass shifted marker ions were observed in the MS/MS spectrum with higher intensity not pair of mass shifted and unshifted marker ions. Thus, the number of targeted masses were reduced from six to three, only the mass shifted marker ions were targeted.

5.4.4 Validation of crosslinks: DNA or RNA?

The formaldehyde crosslinking reaction involves the nucleobases, three out of the four nucleobases were identified in the formaldehyde crosslinking experiments: adenine, guanine, and cytosine. Thymine and uracil crosslinks were not observed under the used reaction conditions. Because RNA and DNA share these nucleobases, the question is if it is possible to distinguish RNA crosslinks from DNA? As it is shown on Figure 5.3-1 ribose of the adenosine monophosphate has one more oxygen atom from the deoxyribose of the deoxyguanosine monophosphate. Whereas guanine has one more oxygen atom than adenine (Figure 5.3-1). Composition of adenosine monophosphate and deoxyguanosine monophosphate is the same, thus, their masses are the same. The two adduct masses are identical in MS1, but in the MS/MS spectra they are distinguishable, because of the presence of different nucleobase marker ions.

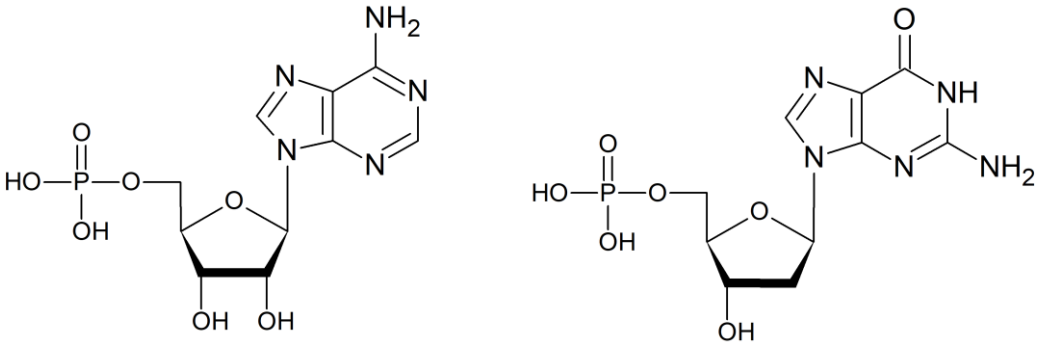


Figure 5.3-1: Molecular structure and molecular mass of adenosine monophosphate and deoxy-guanosine monophosphate nucleotides

In the case of a deoxyguanosine monophosphate crosslink a guanine marker ion, and in the case of an adenosine monophosphate crosslink an adenine marker ion appears. In the native nucleosome dataset DNA and RNA formaldehyde crosslinks were measured together. Example spectra of correct nucleotide/deoxynucleotide is shown in Supplementary Figure 1. Guanine DNA adduct was assigned by the RNP^{xl} search engine, although no guanine marker ion or mass shifted guanine marker ion was present in the MS/MS spectrum. After a closer

look on the MS/MS spectra it became clear a mass shifted adenine marker ion was present indicating that the peptide was crosslinked to adenosine monophosphate.
This example shows that RNA and DNA crosslinks are distinguishable based on the consistently observed marker ion identification.

5.4.5 Methionine oxidation

The mass difference between the respective RNA and DNA molecules is the mass of one oxygen atom (dA-A, dG-G, dC-C), therefore methionine oxidation is quite an interfering phenomenon. During sample preparation, an oxidative environment is present, where methionine can get oxidized on the SH side chain, forming a methionine sulfoxide, with the addition of an oxygen atom. This one oxygen atom is the exact difference between DNA and RNA nucleotides. Misassignment of the methionine oxidation can lead to false crosslink identification. When peptide fragment ions are shifted with the characteristic H_4COS loss [112] from methionine, it can indicate that that methionine amino acid is oxidized. (Supplementary Figure 2). When methionine oxidation is not correctly assigned, only the intact (deoxy)-nucleotide peak can help to distinguish the DNA and RNA crosslinks from each other, although, those peaks are not always present in the spectrum.

5.4.6 PTM assignments together with crosslink identification

Formaldehyde crosslinking has a very simple MS/MS fragmentation, which opens the possibility of crosslink and post translational modification assignment together. Especially that crosslink fragmentation pattern, where the nucleotide entirely cleaves off only appearing as the marker ion, is advantageous for PTM assignment (section 4.2.1.1). During data analysis, when MSFragger was used, datasets were searched for different post-translational modifications. It was important to see whether it is possible to identify PTMs on crosslinks. Although the *in vivo* crosslinking datasets were searched for PTMs such as lysine acetylation, lysine methylation, protein N-terminal acetylation - only protein N-terminal acetylation was identified. Example spectrum of N-terminal acetylation of 60S ribosomal protein L4 is shown in Supplementary Figure 3. Protein N-terminal acetylation of the 60S ribosomal protein L4 has already been shown [113]
Phosphorylation is another common PTM of proteins. Phosphorylation was not searched during database search, but because each sample was phosphatase treated (see 3.2.8 and 3.2.9 sections), it is likely that the used alkaline phosphatase removes the phosphate groups of peptides, not only the 5'-and 3' phosphates of the nucleotides.
The limited number of identified PTMs on crosslinked peptides is probably due to the fact that PTMs are generally low abundant and specific enrichment methods are needed for their identification. N-terminal acetylation, on the other hand, is one of the most abundant PTMs [114] therefore it can be identified without any special enrichment method.
Overall, it was possible to show that PTMs and crosslinks are identifiable on the same peptide when formaldehyde crosslinking is used.

5.5 General comments for *in vivo* formaldehyde crosslinking

In vivo formaldehyde crosslinking was performed for the identification of the RNA binding proteins in bacterial and human cells. Formaldehyde is cell permeable and highly reactive, therefore it can capture the protein-RNA (as well the protein-DNA and protein-protein) interactions. The two, *in vivo* experiments, *E. coli* and HeLa formaldehyde crosslinking

(section 4.2.3 and 4.2.4) are proof of principle experiments to show that formaldehyde crosslinking can be used in proteome-wide studies for confident crosslink identification.
In bacteria, there are approximately 180 known RNA binding [115]. This number in human cells is approximately 1300 [116]. Evidently, it is not possible to capture, all RNA binding proteins in a single, crosslinking experiment in human cells. There are two main factors why this the case. First, the RNA binding proteins are represented with different abundance in the cell, there is a higher chance to identify high abundant RNA binding proteins than the lower abundant ones. Second, the crosslinking reaction has a value of efficiency, some sources report that crosslinking efficiency is approx. 1% [117], although according to other sources this number is still unclear [118]. Regardless of the actual number of the crosslinking efficiency, crosslinked proteins have a lower population than non-crosslinked ones, thus there is a lower chance to measure them with mass spectrometry.
These two features of the proteins abundances and crosslinked protein abundancies are additive, therefore higher number of cells are needed as starting material in a proteome wide experiment. Generally, in crosslinking experiments two folds higher number of cells are used than in sequencing studies, where polymerase chain reaction is used for signal amplification [112].
Overall, a total number of 243 crosslinked proteins were identified in *E. coli* and 460 total number of proteins in HeLa cells. As expected, in both cases classical RNA binding proteins were identified such as ribosomal proteins, transferases, initiation factors etc.

5.5.1 Criteria for high throughput method establishment for *in vivo* crosslinking

For a high throughput, crosslinking method establishment, the following criteria need to be met.

- Simple MS/MS fragmentation of the peptide-RNA heteroconjugates for confident crosslink spectrum identification
- Suitable database search for crosslink identification, FDR control
- Selective biochemical enrichment strategy for peptide-RNA heteroconjugates
- Reduced heterogeneity of the crosslinked RNA adducts

In the case of formaldehyde crosslinking, all criteria have been matched. Formaldehyde crosslinked peptides' MS/MS fragmentation is simple, as it discussed in 4.2.1.1 section. The MS/MS fragmentation of the crosslinked peptides is uniform, it depends on the fragmentation of the linker between peptides and RNA. Due to that MS/MS fragmentation, open search strategy can be used for crosslink identification, where not only spectral FDR but also protein FDR can be calculated. Silica enrichment strategy is selective towards RNA and depletes the high intensity, linear peptides, which would interfere with the downstream analysis [64]. The sequential RNA digestion in the biochemical sample preparation results shorter RNA adducts, in average mono to dinucleotides lengths.
The previously listed features of formaldehyde crosslinking proves that formaldehyde is a useful technique for protein-RNA crosslink identification from high complexity samples.

5.5.2 Sequential RNA digestion

Two-step RNA digestion was performed in the biochemical sample preparation to reduce the heterogeneity of the RNA adducts. It was applied in semi complex (70S ribosome) and complex datasets (*E. coli* and HeLa datasets). The goal of the sequential digestion was to reduce the RNA adducts to three variants: A-HPO_3, G-HPO_3 and C-HPO_3. To achieve a nucleoside digestion nuclease P1 was used and to uniformly remove the 5´- and 3´-phosphates, Antarctic phosphatase was employed.

In the case of HeLa BRP dataset in the second RNA digestion - nuclease P1, Antarctic phosphatase, and additionally RNase T1 and RNase A were used. The addition of the two latter enzymes resulted a higher degree of nucleoside digestion compared to the HeLa unfractionated dataset. In the case of unfractionated HeLa experiment the nucleoside and mononucleotide crosslinks are 55% of the total CSMs, whereas in the case of the BRP fractionated HeLa dataset this number increases to 95% percent.
Although, it was aimed for reduction of the number of RNA adducts in the total of three, other RNA adducts were observed as well. 13% of the identified CSMs were crosslinks to G-H_2O in the HeLa BRP dataset. Water loss from guanosine monophosphate can happen during enzymatic digestion or during electrospray ionization. One possible explanation is that during the RNA digestion steps, when RNase T1 was used, intermediate product of the digestion, 2′,3′-cyclic phosphate [120] was formed. The hydrolysis of the intermediate did not occur, resulting a stable cyclic guanosine monophosphate. Another possible explanation is a water loss occurs during ESI ionization from the ribose, the phosphate group or from the base. Although, no water loss was observed from the marker ions and mass shifted marker ions, narrowing down the possibility of the water loss from to the ribose and phosphate groups.
37% of the total CSMs in the HeLa BRP dataset were G crosslinks, whereas 4% of the CSMs were crosslinks to G-HPO_3. Thus, complete removal of the phosphate groups was not successful. In the case of adenine crosslinks, the opposite tendency was observed, 1% of the CSMs were crosslinks to A, whereas 37% of the CSMs were A-HPO_3 %. Crosslinks to C-HPO_3 are significantly lower than adenine and guanine crosslinks, in total, 3% of the CSMs.
Overall, based on the number of CSMs in the HeLa BRP dataset, 95% of CSMs are mononucleotide and nucleoside RNA adducts: A, G, A-HPO_3, G-HPO_3, G-H_2O and C-HPO_3. Although, the results are promising, sequential nucleotide digestion still needs to be finetuned to achieve only 3 RNA adducts.

5.6 Crosslink results comparison between *ex vivo* crosslinked 70 S ribosome and 70S ribosome results of *in vivo* crosslinked *E. coli* cells

Ribosomal proteins are one of the most abundant RNA binding proteins in the cytosol [121]. Isolated 70S ribosome from *E. coli* was *ex vivo* formaldehyde crosslinked and peptide-RNA heteroconjugates were measured with HPLC-MS/MS and crosslinked proteins were identified with database search (see section 4.2.2.1)
E. coli was formaldehyde crosslinked *in vivo* and peptide-RNA heteroconjugates were identified as well (see section 4.2.3). Crosslinked peptides were compared between, the two systems, namely, *ex vivo* crosslinked and *in vivo* crosslinked 70S ribosomal proteins.
Significant number of CSMs are ribosomal protein crosslinks in the *E. coli* dataset (56% of the total CSMs). This covers all proteins of the 70S ribosomal proteins, 31 proteins of the 50S large ribosomal subunit and 21 proteins of the small ribosomal subunit. Whereas, in the case of *ex vivo* crosslinked 70S ribosomal proteins, 14 proteins of the small ribosomal subunit and 18 proteins of the large ribosomal subunit were identified as RNA crosslinked proteins. In the case of *ex vivo* crosslinked 70S ribosome, the most prominent crosslinks were compared with the *in vivo* crosslinked *E. coli* 70S ribosomal proteins. The highest number of crosslinked spectrum matches are related to crosslinked peptides of the 50S ribosomal protein L2, in the *ex vivo* crosslinked 70S ribosome. The 50S ribosomal protein L2 is needed for the association of the two subunits of the 70S ribosome as well as for the peptide bond formation [122]. In the case of the *in vivo* crosslinked 70S ribosome, 50S ribosomal protein L32 was identified with the highest number of CSMs. The majority of the crosslinked peptides (45) of the *ex vivo* crosslinked 70S ribosome have been identified in the *in vivo* crosslinking experiment (Figure 5.6-1)

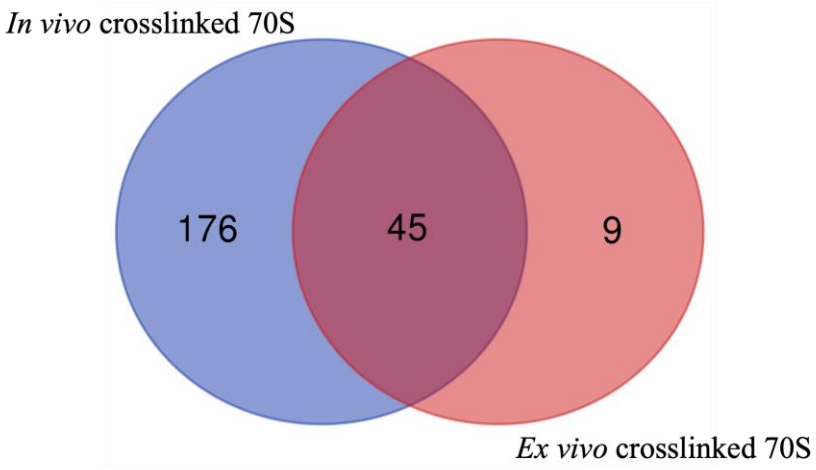


Figure 5.6-1: Venn diagram of the identified crosslinked peptides in the 70S ribosome of *in vivo* crosslinked *E. coli* and the *ex vivo* crosslinked 70S ribosome

Only minority (9) of the crosslinked peptides could not be identified in the *E. coli* dataset. Another 176 crosslinking sites were only identified in *E. coli*; these crosslinking sites are possibly related to different conformations of the ribosome as it was shown in the example of the Tu2 bound 70S ribosome in section 4.2.3.3.

5.7 Comparison of formaldehyde crosslinking results and UV crosslinking results from the literature in *Escherichia coli*

UV crosslinking is the classical method of choice for crosslinking protein-RNA complexes, therefore predominantly UV crosslinking studies are available in the literature. Moreover, no *in vivo* crosslinking study is available, where crosslinking site identification was performed in *Escherichia coli*. Thus, no study could be chosen for comparison of the formaldehyde crosslinking dataset in a one-to-one basis. On the other hand, there are studies available, where *in vivo* UV crosslinking was performed and linear peptides (or non-crosslinked peptides) of crosslinked proteins were identified. Identification of non-crosslinked peptides of a crosslinked proteins is a much simpler indirect approach, where crosslinking site identification is lost.

The other difference – apart from the lack of identified crosslinking sites in the UV crosslinking studies – is the crosslinking method itself. The main target of UV crosslinking is uracil; formaldehyde crosslinking mainly happens on guanine and adenine and – less prominently – on cytosine. Formaldehyde not only crosslinks proteins to nucleic acids but also proteins with each other, whereas UV crosslinking forms protein-nucleic acid bonds. These points all need to be taken into consideration when comparing different crosslinking methods with each other.

Two studies were chosen for comparison of the formaldehyde crosslinking results in *E. coli*, Queiroz *et al.* [33] and Shchepachev *et al.* [61].

Queiroz *et al.* have established an **O**rthogonal **O**rganic **P**hase **S**eparation of UV crosslinked protein-RNA complexes from *E. coli* cells and named the method OOPS. After the enrichment of protein-RNA complexes they have performed a sequential digestion of crosslinked proteins. First, LsyC digestion was performed, followed by a silica enrichment of the RNA or EtOH precipitation of RNA, then trypsin and RNAse digestions were performed. Followed by TiO_2 enrichment of the nucleotide-peptide heteroconjugates. The purpose of this step was not the enrichment for the peptide-nucleotide heteroconjugates, but rather the

separation of the peptide-nucleotide heteroconjugates from their adjacent peptides. These adjacent peptides are measured by HPLC-MS/MS. Using OOPS they have aimed to determine the RNA interactome in bacteria. They have identified 364 proteins in *Escherichia coli* K-12 strain.
Another UV crosslinking experiment in *Escherichia coli* was done by Shchepachev *et al.* They have performed silica enrichment of UV crosslinked and non-crosslinked protein-RNA complexes from SILAC (stable isotope labelling with amino acids in cell culture) media grown *E. coli*. The isolated RNA was digested, proteins were in-gel digested and peptides were measured with HPLC-MS/MS. They named the method TRAPP (**T**otal **R**NA-**A**ssociated **P**rotein **P**urification). Significantly enriched proteins in the crosslink sample (compare to the non-crosslinked sample) are the RNA binding proteins. They have identified 1168 significantly enriched proteins when 1360 mJ/cm^2 UVC irradiation was used.
In both studies linear peptides were used for the identification of the crosslinked proteins, whereas in this study crosslinked peptides were used for the identification of the crosslinked proteins. Peptide-RNA crosslinks are the direct proof of the interaction between proteins and nucleic acids. Due to the low intensity of the peptide-RNA heteroconjugates, lower number of proteins can be identified, than other studies, where crosslinked proteins are identified based on their linear peptides.
The most abundant RNA binding proteins, the ribosomal constituents, mRNA binding proteins have a big overlap between all the studies. TRAPP has identified the highest number of proteins (1168 proteins), followed by OOPS (364 proteins) and then the current study (243 proteins). Overlaps between the identified proteins in each study are shown in Figure 5.7-1.

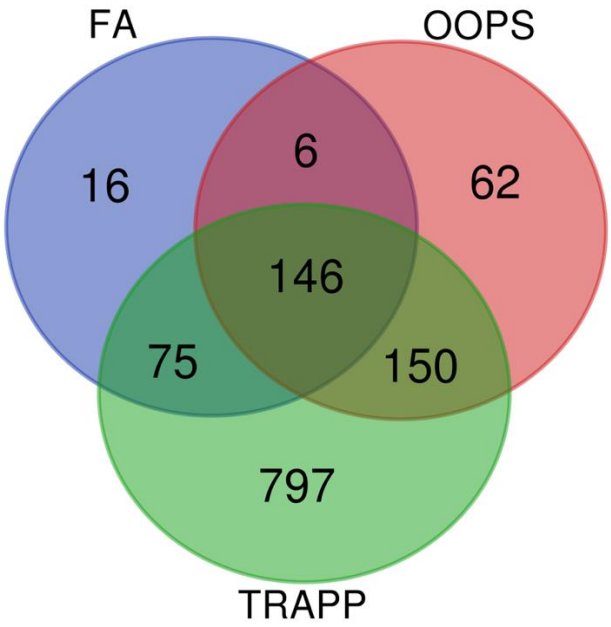


Figure 5.7-1: Venn diagram of identified proteins in OOPS, TRAPP and FA methods

The majority of the proteins (221) identified in the FA crosslinking experiment are identified with the TRAPP method as well, except for twenty-two proteins. Those proteins are being unique to the formaldehyde crosslinking dataset. Similarly, the majority of the proteins (296) identified in the OOPS experiment have been identified in the TRAPP experiment as well. The lowest overlap is between the FA and the OOPS experiment, with 152 proteins.
In the OOPS workflow has enriched GO terms such as ‘rRNA binding”, ‘tRNA binding’ showing that they have identified classical RNA binding proteins. Although, in that study, 234/364 RNA binding proteins were not associated with RNA related GO terms. The TRAPP workflow has identified proteins with RNA related GO terms such as ‘RNA binding’, ‘nucleotide binding’, ‘ribonucleotide binding’.

One of the most interesting findings of the current study is, high number of metabolic enzymes were of the identified as RNA crosslinks. Some of them are already known as RNA binding proteins such as GAPDH, isocitrate dehydrogenase or malate dehydrogenase [116]. The same observation was made in the TRAPP and the OOPS experiments by Shchepachev *et. al.* and Queiroz *et. al*. In their UV crosslinking studies they have identified several metabolic enzymes as RNA crosslinks as well. Queiroz *et. al*. have shown that, among others, Pyruvate kinase, GAPDH, Enolase and Phosphoglycerate kinase are identified as RNA binding proteins. Shchepachev *et al.* also discusses that those enzymes which participate in the glycolysis and gluconeogenesis are RNA binding proteins.
Several other interesting findings have been made by Queiroz *et al.* For instance, they have shown that SecA (protein translocase subunit SecA) is an RNA binding protein, which is a peripheral protein of the plasma membrane [123]. They hypothesized that SecA is a good candidate to shown that RNA is localized close to the plasma membrane.
In the current study, SecA was not found; but another subunit of the protein translocase complex was identified, SecD. Several proteins localized at the bacterial outer membrane were identified as RNA binding proteins as well. For instance, Outer membrane porin F protein (OmpF). OmpF is localized at the outer membrane, forming pores for small molecular diffusion [124]. Lipopolysaccharide core heptose(II)-phosphate phosphatase was identified in the current study as RNA binding protein. Lipopolysaccharide core heptose(II)-phosphate phosphatase is localized at the bacterial outer membrane and responsible of the dephosphorylation of heptose(II) in the liposaccharide core.
TRAPP also identified proteins related to 'membrane' GO term, such as Inner membrane protein YbaL and Outer membrane lipoprotein SlyB.
All three studies support the same conclusion, which is protein-RNA crosslinking is a useful to discover proteins with previously not known RNA binding function.

5.8 General comments about the *in vivo* crosslinking in human cells

Results, obtained in the *in vivo* formaldehyde crosslinked HeLa cells are not as striking as the *in vivo* formaldehyde crosslinked *E. coli* cells. Proteins identified in the formaldehyde crosslinked HeLa cells are known as RNA binding proteins, with a few exceptions. Not classical RNA binding proteins such as metabolic enzymes were identified as RNA binding proteins in the HeLa dataset (GAPDH, enolase, fructose-bisphosphate aldolase, etc.). Although, these metabolic enzymes are represented in lower number than observed in the *E. coli* dataset.
Formaldehyde crosslinking *in vivo*, in combination with mass spectrometric analysis was investigated in the current study. Similarly, as in the case of the *E. coli* dataset there is no available publication, where formaldehyde crosslinking together with mass spectrometry was used to identify protein-RNA crosslinking sites. Other UV crosslinking studies are available where UV was performed *in vivo* for the identification of protein-RNA crosslinks in cells [60].
Bae *et al*. [62] performed UV crosslinking in HeLaT cells in combination with a two-step silica enrichment for total RNA purification. They have performed a hydrogen fluoride (HF) digestion of the RNA-peptide heteroconjugates. HF cleaves RNA into nucleosides, leaving only a nucleoside on the crosslinked peptide. This approach highly reduces the complexity of the data analysis; therefore, they were able to identify crosslinking site to a single amino acid resolution.
Queiroz *et al.* applied OOPS workflow on three human cell lines: U2OS, MCF10A and HEK293 cells. Same workflow was used, as the previously discussed, adjacent peptide identification in case of *Escherichia coli* samples. Queiroz *et al*. has shown that 759 proteins are common the three cells lines which are RNA binding when OOPS workflow is used.

Trendel *et al*. [34] have performed UV crosslinking on MCF7 cells, they have used a similar approach as OOPS to extract protein-RNA complexes from cells with the use of Trizol.
Panhale *et al.* performed UV crosslinking in HEK293 cells [43]. Their approach combines the crosslink peptide identification with the adjacent peptide identification (similar to the OOPS study). In more details: poly(A) containing RNAs were isolated with oligo(dT) pull down. Lys-C digestion and subsequent trypsin digestion was performed on the crosslinked protein-RNA complexes, resulting peptides adjacent to crosslink-peptide heteroconjugates. These adjacent peptides were analyzed by HPLC-MS/MS. They have processed further the RNA-peptide conjugates. Nuclease treatment was performed on the RNA, the resulted crosslinked peptides-nucleotide heteroconjugates were identified with mass spectrometry. They have combined these two methods and named it CAPRI (**C**rosslinked and **A**djacent **P**eptides-based **R**NA-binding domain **I**dentification.
In the CAPRI study 135 proteins were identified based on only crosslink peptides and 543 proteins based on the adjacent peptides.

5.8.1 Comparison of formaldehyde crosslinking results with UV crosslinking results from literature in human cells

Crosslinked protein identification based on crosslinked peptides were performed in two studies, by Panhale *et al* [43]. and Bae *et al.* [62].The latter study identified higher number of proteins; thus, this study was used for comparison of the formaldehyde crosslinking results.
Bae *et al.* have identified 594 RNA binding proteins from a two-step silica enrichment of three dishes (150 mm) of UV-crosslinked HeLaT cells, they have identified crosslinking sites to single amino acid resolution. In the current study, in total, 460 crosslinked proteins were identified, in each experiment approximately ~ 2E8 number of cells were used. HF digestion resulted in complete RNA digestion to nucleoside. In the current study enzymatic RNA digestion was used; in the case of BRP fractionated sample, 95% of the identified crosslinked spectrum matches are mononucleotide and nucleoside crosslinks, which a similar result as the HF digestion of the RNA. Enzymatic digestion would be preferred over chemical digestion of the RNA, due to their selectivity and non-toxic properties. HF is a highly toxic chemical and the use of such highly reactive chemicals could introduce chemical artifacts during sample preparation.
The overlap between the crosslinked proteins in the two studies were calculated and visualized. In the case of formaldehyde crosslinking, all identified proteins were used in total 460 crosslinked proteins of the unfractionated and BRP fractionated samples. The 594 crosslinked proteins from UV crosslinking study were compared with the 460 crosslinked proteins from formaldehyde crosslinking. The Venn diagram of results is shown in Figure 5.8-1.

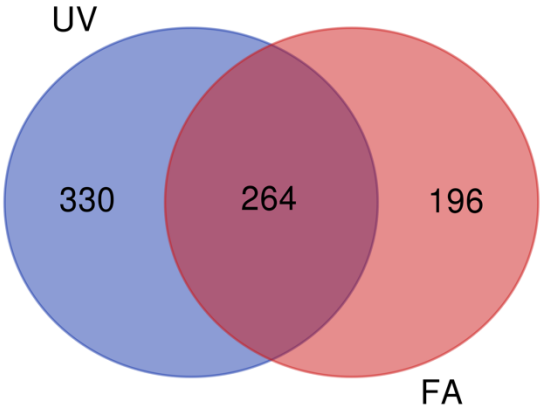


Figure 5.8-1: Venn diagram of crosslinked proteins identified in the UV crosslinking and formaldehyde crosslinking study

The two crosslinking methods commonly identified 264 proteins as RNA crosslinks. 196 crosslinked proteins were identified in the formaldehyde crosslinking experiment uniquely and 330 crosslinked proteins were identified in the UV crosslinking experiment uniquely.
STRING database [79] was used for the analysis of the not common proteins of each crosslinking experiment. Local network cluster analysis shows the difference between the proteins enriched in each experiment. In the case of formaldehyde crosslinking results local clusters are for instance, 'mitochondrial translation initiation' and 'peptide chain elongation' (Table 5.8-1 left table). In the case of UV crosslinking, 'ribosome biogenesis and helicase activity', 'preribosome and ribosome biogenesis' terms are part of the local clusters (Table 5.8-1 right table).

#term ID	term description	false discovery rate
CL:429	Mitochondrial translation initiation	4.80E-10
CL:146	Formation of a pool of free 40S subunits, and Elongation factor	1.81E-08
CL:153	Eukaryotic Translation Elongation	2.35E-08
CL:159	Peptide chain elongation	8.28E-07
CL:168	Cytoplasmic ribosomal proteins	2.02E-06
CL:170	Cytoplasmic ribosomal proteins	1.26E-05
CL:435	organellar large ribosomal subunit	4.81E-05
CL:171	Cytoplasmic ribosomal proteins	0.00038
CL:741	preribosome, and Ribosome biogenesis	0.0025
CL:438	organellar large ribosomal subunit	0.0043

#term ID	term description	false discovery rate
CL:735	ribosome biogenesis, and RNA helicase activity	5.68E-18
CL:736	ribosome biogenesis, and Helicase associated domain (HA2)	3.18E-16
CL:738	preribosome, and ribosome biogenesis	8.96E-16
CL:739	mixed, incl. preribosome, and Ribosome biogenesis	2.05E-15
CL:743	preribosome, and Ribosome biogenesis	7.49E-15
CL:748	preribosome, and Ribosome biogenesis	1.27E-14
CL:2271	mixed, incl. P-body, and endoribonuclease complex	6.68E-14
CL:750	preribosome, and Ribosome biogenesis	1.51E-13
CL:2272	mixed, incl. P-body, and endoribonuclease complex	7.96E-13
CL:751	preribosome, and Ribosome biogenesis	6.16E-12

Table 5.8-1:Local network cluster analysis results of the not common proteins between UV and FA crosslinking experiments in HeLa cells

Left table: First ten hits of the local network cluster analysis results (STRING) of formaldehyde crosslinked proteins, sorted by FDR. Right table: first ten hits of the local network cluster analysis results (STRING) of UV crosslinked proteins, sorted by FDR.

The preliminary results of the comparison of the two crosslinking methods show that the two crosslinking methods capture proteins in interactions of all types of RNAs (small, ribosomal, mRNA etc.). Based on the results, formaldehyde crosslinking captured more protein RNA interactions in the mitochondrion, whereas UV crosslinking captured more proteins related to ribosome biogenesis. This might be due to that formaldehyde as a chemical can penetrate easier into the cell, than UV light. Although, it is important to emphasize that quantitative comparison cannot be made on the two datasets. For better comparison, addressing the nature of crosslinking, the crosslinking efficiency, and the preferences towards RNAs and proteins need to be further investigated.

5.9 *Escherichia coli* and *Homo sapiens* orthologous proteins

RNA binding proteins identified from *Escherichia coli* and *Homo sapiens* were compared to see how many proteins are orthologous. InParanoid 8 [125] a database was used to compare the *E. coli* and human proteomes. In total, 530 clusters are present between the two proteomes. In the case of *E. coli*, in the silica enrichment experiment, 243 proteins were identified as RNA crosslinks. In the HeLa experiments, in total of the BRP and unfractionated samples 460 proteins were identified. RNA crosslinked proteins from each organism were mapped onto the respective clusters, if proteins of both organisms were present in one cluster, it was considered as a match.
Several metabolic enzymes were identified as RNA binding proteins in *E. coli*, if, these interactions have a specific function, they need to be evolutionary conserved and orthologues proteins in human should show similar RNA binding properties.
Same line of thought was used by Beckmann *et al.* [126], where they have isolated mRNA binding proteins from yeast and human and they have a compared protein functions from

human and yeast. They have found that several metabolic proteins are RNA binding in both human and in yeast. They have observed that in yeast 17% and in human 9% of the conserved RNA binding proteins function as metabolic enzymes.
In this study,108 proteins from *E. coli* have been found as a match to the 530 clusters and 56 proteins of the human proteins were identified in the clusters. In total 30 proteins were found to be orthologous between the identified proteins of *E. coli* and human. The majority of the matches are ribosomal proteins. 70S ribosomal proteins from *E. coli* and their orthologous proteins of the 80S ribosome and mitochondrial ribosome were found.
Among others, elongation factor Tu2 from *E. coli* and its orthologues protein from *Homo sapiens*, elongation factor Tu, mitochondrial were found. Two, previously discussed metabolic enzymes from *E. coli*, GAPDH and Enolase and their orthologous proteins: human GAPDH and alpha enolase were found as well.

6 Conclusions and outlook

In this study two chemical crosslinking methods, DEB and FA crosslinking coupled with mass spectrometry were used to investigate protein-nucleic acid interactions in low complexity and high complexity samples. MS/MS fragmentation behavior of each chemical crosslinking method was described. It was concluded that MS/MS fragmentation of the DEB crosslinking is adventurous when crosslinking site information is needed, because crosslinking side can be narrowed down to a single amino acid. This feature of the DEB crosslinking is beneficial for the elucidation of the structure of protein-DNA complexes. *In vivo* crosslinking of protein-DNA complexes with DEB was not possible due to the lack of selective enrichment strategy of the protein-DNA complexes from cells. If selective enrichment strategy is established, DEB crosslinking with MS2/MS3 method can be applicable to reduce the complexity of the crosslinking site identification and this would allow to produce a cell wide protein-DNA interaction map.

Formaldehyde *in vivo* crosslinking followed by mass spectrometry deemed to be a useful tool to it captures hundreds of RNA interacting proteins at the same time. The established method can be applied in several ways. One way to use the technique is to identify proteins with not known RNA-binding functions, the technique can be beneficial for the identification of the RNA interactome in less investigated organisms.

The formaldehyde crosslinking method can be used for targeted approaches, such as identification of proteins with certain post-translational modifications or RNA with post transcriptional modifications. Another application of the method is to monitor quantitative changes in the cell. The selective RNA enrichment method, the sequential RNA digestion and automatized data analysis would allow the method to be applied for quantitative crosslinking studies. For instance, differences between resting and excited states of the cell. Quantitative changes of the RNA interactome could be captured by formaldehyde crosslinking mass spectrometry. The method can be used in combination with chromatin immunoprecipitation. It is plausible that in the future, from one sample which have been formaldehyde crosslinked, mass spectrometric analysis as well as DNA sequencing can be carried out. Although the two techniques do not have the same sensitivity, hopefully, with higher sensitivity enrichment techniques the initial number of cells used for mass spectrometric analysis can be reduced.

Formaldehyde crosslinking is already a commonly used crosslinking technique, therefore the currently developed mass spectrometric method could be integrated into a platform where several techniques which aim for the investigation of protein-nucleic interactions can be used together.

Appendix

Supplementary Tables

Supplementary Table 1: Sequest modification search results of the DEB crosslinked H1.4 linker histone-187 bp ds DNA complex

Crosslinking Site Probabilities are taken from the ptmRS node in Proteome discoverer.

Protein Name (UniProt ID)	Peptide Sequence	Crosslinking Site Probability	DNA adduct(s)
Histone H1.4 (P10412)	ASGPPVSELITKAVAASK	K12(DEB): 99.03	dTdG
		K12(DEB_Gb): 98.88	dCdG
		T11(DEB): 49.66; K12(DEB): 49.66	Gb
		T11(DEB): 49.72; K12(DEB): 49.72	dGdG
		T11(DEB): 49.73; K12(DEB): 49.73	dCdTdG
		T11(DEB): 50; K12(DEB): 50	Gb
	GTGASGSFKLNK	F8(DEB): 49.46; K9(DEB): 49.46	dTdGdG
		F8(DEB): 49.66; K9(DEB): 49.66	dCdG
		F8(DEB): 49.83; K9(DEB): 49.83	dGdG
		F8(DEB): 49.84; K9(DEB): 49.84	Gb
	GTGASGSFKLNKK	F8(DEB): 49.78; K9(DEB): 49.78	Gb
	GTLVQTKGTGASGSFK	K7(DEB): 33.25; G8(DEB): 33.25; T9(DEB): 33.25	dCdTdGdA
		K7(DEB): 49.71; G8(DEB): 49.71	dCdG
			dCdTdGdG
		K7(DEB): 49.72; G8(DEB): 49.72	dGdG
			dTdG
		K7(DEB): 49.75; G8(DEB): 49.75	dTdG
		K7(DEB): 49.84; G8(DEB): 49.84	dTdGdG
		K7(DEB): 97.95	dCdTdG
		K7(DEB_Gb): 98.15	dTdGdA
		K7(DEB_Gb): 98.59	dGdGdA
	KASGPPVSELITK	A2(DEB): 33.17; S3(DEB): 33.17; G4(DEB): 33.17	dCdTdGdGdA
		A2(DEB): 99.31	dCdCdGdG
		K1(DEB): 100	dGdA
			Gb
			dTdG
		K1(DEB): 49.63; A2(DEB): 49.63	dTdTdG
		K1(DEB): 49.88; A2(DEB): 49.88	dCdGdA
		K1(DEB): 99.27	dCdTdGdA
		K1(DEB): 99.37	dTdTdGdA
		K1(DEB): 99.4	dCdCdG
		K1(DEB): 99.41	dTdGdG
		K1(DEB): 99.44	dTdGdG

		K1(DEB): 99.46 K1(DEB): 99.48 K1(DEB): 99.5	dTdGdA dGdGdG dCdG dTdG dTdGdGdA
		K1(DEB): 99.51 K1(DEB): 99.52	dGdGdA dCdGdG dCdTdGdG
		K1(DEB): 99.55 K1(DEB): 99.59 K1(DEB): 99.61 K1(DEB): 99.62 K1(DEB_Gb): 50; A2(DEB_Gb): 50	dCdTdGdG dGdG Gb Gb dCdTdG Gb
		K1(DEB_Gb): 99.99 Too many isoforms	Gb Gb
	KASGPPVSELITKAVAASK	I11(DEB): 33.33; T12(DEB): 33.33; K13(DEB): 33.33	Gb
	LGLKSLVSK	K4(DEB): 100 K4(DEB): 98.78 K4(DEB): 99.41 K4(DEB): 99.5 K4(DEB): 99.99	Gb Gb Gb Gb Gb
	RKASGPPVSELITK	R1(DEB): 20; K2(DEB): 20; A3(DEB): 20; S4(DEB): 20; G5(DEB): 20	Gb
	SGVSLAALK	S1(DEB): 99.62	Gb
	SGVSLAALKK	K9(DEB): 50; K10(DEB): 50 K9(DEB): 98.46 K9(DEB): 99.32 K9(DEB): 99.99	Gb dCdTdGdG Gb Gb dTdG
	SLVSKGTLVQTK	K5(DEB): 99.33 K5(DEB): 99.34 K5(DEB): 99.35 K5(DEB): 99.39 K5(DEB): 99.44 K5(DEB): 99.99 S4(DEB): 50; K5(DEB): 50	dCdG Gb dGdA dTdG Gb dTdG dCdGdG dCdTdGdG

Supplementary Table 2: RNP^{xl} search results of the DEB crosslinked H1.4 linker histone- 187 bp ds DNA complex

Protein Name (UniProt ID)	Peptide Sequence	DNA adduct(s)	Crosslinking site
Histone H1.4 (P10412)	ALAAAGYDVEK	dT	No localization
	ALAAAGYDVEKNNSR	Gb	E11
	ASGPPVSELITK	Gb	K12
			No localization
		dT	No localization
			No localization
	ASGPPVSELITKAVAASK	dAdCdCdG	No localization
		dAdCdGdT	No localization
		dAdGdT	No localization
		dAdGdTdT	No localization
		dCdG	No localization
		dCdGdT	No localization
		Gb	E8
			K12
		dGdG	No localization
		dGdT	No localization
		dT	No localization
			No localization
	ERSGVSLAALKK	dAdCdTdG-deoxyribose	No localization
	GTGASGSFKLNK	dAdGdT	No localization
		dCdG	No localization
		dCdGdT	No localization
		Gb	K9
			No localization
		dGdG	No localization
		dGdGdT	No localization
		dGdT	No localization
	GTGASGSFKLNKK	Gb	S7
	GTLVQTKGTGASGSFK	dAdCdT	No localization
		dAdGdG	No localization
		dAdGdT	No localization
		dAdGdTdT	No localization
		dAdT	No localization
		dCdG	No localization
		dCdGdT	No localization
			No localization
		dCdT	No localization
		Gb	G8
			K7
			T6

		dGdG	No localization
		dGdGdT	No localization
		dGdT	No localization
	IKLGLK	dAdCdGdT	No localization
			No localization
		dAdG	No localization
		dAdGdG	No localization
		dAdGdGdT	No localization
		dAdGdT	No localization
		dAdGdTdT	No localization
		dCdG	No localization
		dCdGdT	No localization
		Gb	I1 L3 K6
		dGdG	No localization
		dGdT	No localization
	IKLGLKSLVSK	dAdTdG-deoxyribose	No localization
	KASGPPVSELITK	dAdAdCdT	No localization
		dAdAdTdT	No localization
		dAdCdC	No localization
		dAdCdG	No localization
		dAdCdGdT	No localization
		dAdCdT	No localization
		dAdCdTdT	No localization
		dAdG	No localization
		dAdGdG	No localization
		dAdGdGdT	No localization
		dAdGdT	No localization
		dAdGdTdT	No localization
		dAdT	No localization
		dCdCdG	No localization
		dCdCdTdT	No localization
		dCdG	No localization
		dCdGdT	No localization
		dCdGdTdT	No localization
		dCdT	No localization
		dCdTdT	No localization
		Gb	K1
		dGdG	No localization No localization

		dGdGdG	No localization
		dGdGdT	No localization
			No localization
		dGdT	No localization
			No localization
		dGdTdT	No localization

Supplementary Table 3: Sequest modification search results of the DEB crosslinked H5 linker histone-187 bp ds DNA complex

Crosslinking Site Probabilities are taken from the ptmRS node in Proteome discoverer.

Protein name (UniProt ID)	Peptide Sequence	Crosslinking Site Probability	DNA adduct(s)
Histone H5 (P02259)	GVGASGSFR	G1(DEB): 97.67 G1(DEB): 99.14 G1(DEB): 99.21 G1(DEB): 99.24 G1(DEB): 99.35	dGdG Gb Gb Gb dGdG
	LLAAGVLK	L1(DEB): 99.57	Gb
	LLAAGVLKQTK	K8(DEB): 100	Gb dTdG dTdGdG
		K8(DEB): 99.17 K8(DEB): 99.47	dTdG dGdGdG
		K8(DEB): 99.49	dCdCdGdA Gb
		K8(DEB): 99.5 K8(DEB): 99.51	Gb dCdGdGdG dCdTdG
		K8(DEB): 99.52 K8(DEB): 99.53	dCdGdG Gb dTdG dTdGdGdA
		K8(DEB): 99.54	dCdCdGdG dCdTdGdG Gb dTdGdG
		K8(DEB): 99.56	Gb dTdGdA
		K8(DEB): 99.57 K8(DEB): 99.58 K8(DEB): 99.59	Gb Gb Gb dGdG
		K8(DEB): 99.6 K8(DEB): 99.61 K8(DEB): 99.99 K8(DEB_Gb): 99.1 K8(DEB_Gb): 99.11 K8(DEB_Gb): 99.26 K8(DEB_Gb): 99.28 K8(DEB_Gb): 99.29 K8(DEB_Gb): 99.3 K8(DEB_Gb): 99.33 K8(DEB_Gb): 99.37 K8(DEB_Gb): 99.39	Gb Gb Gb dGdG dCdG Gb dTdG Gb dGdA Gb dCdG dGdGdA

		K8(DEB_Gb): 99.92	dCdTdGdG
		K8(DEB_Gb): 99.96	dCdGdA
		K8(DEB_Gb): 99.99	Gb
	QTKGVGASGSFR	K3(DEB): 86.3	dCdG
		K3(DEB): 94.75	Gb
		K3(DEB): 98.28	Gb
		K3(DEB): 98.62	dTdG
		K3(DEB): 98.9	dGdG
		K3(DEB): 99.95	dGdGdA
		K3(DEB): 99.96	Gb
		K3(DEB): 99.97	Gb
		K3(DEB): 99.99	Gb
	SASHPTYSEMIAAAIR	H4(DEB_Gb): 49.66; P5(DEB_Gb): 49.66	Gb
		H4(DEB_Gb): 49.81; P5(DEB_Gb): 49.81	Gb
		H4(DEB_Gb): 97.25	Gb
		H4(DEB_Gb): 97.28	dCdCdGdGdA
		H4(DEB_Gb): 98.09	dGdG
		H4(DEB_Gb): 98.1	dCdGdG
		H4(DEB_Gb): 98.14	dCdCdGdG
		H4(DEB_Gb): 98.25	dCdTdGdG
		H4(DEB_Gb): 98.43	dCdTdGdA
		H4(DEB_Gb): 98.44	dCdTdG
		H4(DEB_Gb): 98.45	dCdGdG
		H4(DEB_Gb): 98.46	dTdG
		H4(DEB_Gb): 98.48	Gb
		H4(DEB_Gb): 98.53	dGdG
		H4(DEB_Gb): 98.55	dCdTdGdG
		H4(DEB_Gb): 98.64	Gb
		H4(DEB_Gb): 98.69	dTdG
		H4(DEB_Gb): 98.72	dGdG
		H4(DEB_Gb): 98.77	dTdG
		H4(DEB_Gb): 98.97	dTdG
		H4(DEB_Gb): 99	dCdG
		H4(DEB_Gb): 99.06	dCdG
		H4(DEB_Gb): 99.07	dCdCdGdGdA
		H4(DEB_Gb): 99.18	dTdG
		H4(DEB_Gb): 99.26	Gb
		H4(DEB_Gb): 99.28	dCdTdGdG
		H4(DEB_Gb): 99.37	dCdG
			dCdTdG
		P5(DEB_Gb): 19.8; T6(DEB_Gb): 19.8; Y7(DEB_Gb): 19.8; S8(DEB_Gb): 19.8; E9(DEB_Gb): 19.8	Gb
		P5(DEB_Gb): 33.19; T6(DEB_Gb): 33.19; Y7(DEB_Gb): 33.19	Gb
		P5(DEB_Gb): 33.33; T6(DEB_Gb): 33.33; Y7(DEB_Gb): 33.33	Gb
		P5(DEB_Gb): 49.76; T6(DEB_Gb): 49.76	Gb
		P5(DEB_Gb): 49.79; T6(DEB_Gb): 49.79	Gb

		P5(DEB_Gb): 49.82; T6(DEB_Gb): 49.82	Gb
		P5(DEB_Gb): 49.84; T6(DEB_Gb): 49.84	Gb
		P5(DEB_Gb): 50; T6(DEB_Gb): 50	Gb
		P5(DEB_Gb): 96.48	Gb
		S1(DEB): 99.14	Gb
		S1(DEB_Gb): 24.95; A2(DEB_Gb): 24.95; S3(DEB_Gb): 24.95; H4(DEB_Gb): 24.95	Gb
		S1(DEB_Gb): 99.99	dGdG
		S3(DEB_Gb): 48.43; H4(DEB_Gb): 48.43	dCdCdTdGdG
		S3(DEB_Gb): 48.85; H4(DEB_Gb): 48.85	dCdCdGdA
		S3(DEB_Gb): 49.47; H4(DEB_Gb): 49.47	dCdTdGdG
		S3(DEB_Gb): 97.13	Gb
		T6(DEB_Gb): 95.3	dCdCdGdG
	TESLVLSPAPAKPK	L6(DEB): 99.18	dTdG
		S3(DEB_Gb): 32.88; L4(DEB_Gb): 32.88; V5(DEB_Gb): 32.88	Gb
		S3(DEB_Gb): 32.94; L4(DEB_Gb): 32.94; V5(DEB_Gb): 32.94	Gb
		S3(DEB_Gb): 33.07; L4(DEB_Gb): 33.07; V5(DEB_Gb): 33.07	Gb
		S3(DEB_Gb): 33.15; L4(DEB_Gb): 33.15; V5(DEB_Gb): 33.15	Gb
		S3(DEB_Gb): 33.17; L4(DEB_Gb): 33.17; V5(DEB_Gb): 33.17	Gb
		S3(DEB_Gb): 49.17; L4(DEB_Gb): 49.17	Gb
		S3(DEB_Gb): 49.29; L4(DEB_Gb): 49.29	Gb
		S3(DEB_Gb): 49.37; L4(DEB_Gb): 49.37	Gb
		S3(DEB_Gb): 49.57; L4(DEB_Gb): 49.57	Gb
		S3(DEB_Gb): 49.69; L4(DEB_Gb): 49.69	Gb
		T1(DEB): 19.84; E2(DEB): 19.84; S3(DEB): 19.84; L4(DEB): 19.84; V5(DEB): 19.84	Gb
		T1(DEB): 19.93; E2(DEB): 19.93; S3(DEB): 19.93; L4(DEB): 19.93; V5(DEB): 19.93	Gb
		T1(DEB): 24.87; E2(DEB): 24.87; S3(DEB): 24.87; L4(DEB): 24.87	Gb
		T1(DEB): 32.69; E2(DEB): 32.69; V5(DEB): 32.69	Gb
		T1(DEB): 48.57; E2(DEB): 48.57	Gb
		T1(DEB): 48.61; E2(DEB): 48.61	Gb
		T1(DEB): 49.12; E2(DEB): 49.12	dTdGdGdA
		T1(DEB): 49.15; E2(DEB): 49.15	Gb
		T1(DEB): 49.33; E2(DEB): 49.33	Gb
		T1(DEB): 49.45; E2(DEB): 49.45	Gb
		T1(DEB): 49.49; E2(DEB): 49.49	Gb
		T1(DEB): 49.5; E2(DEB): 49.5	Gb
		T1(DEB): 49.51; E2(DEB): 49.51	Gb
		T1(DEB): 49.99; E2(DEB): 49.99	dTdGdG
		T1(DEB): 96.89	Gb
		T1(DEB): 98.77	dGdG
		T1(DEB): 98.96	Gb
		T1(DEB): 99.04	dCdCdGdA

		T1(DEB): 99.07 T1(DEB): 99.08 T1(DEB): 99.1 T1(DEB): 99.15 T1(DEB): 99.19 T1(DEB): 99.23 T1(DEB): 99.25	dCdGdA Gb dCdGdG Gb dTdA dTdG dCdTdGdG dCdTdGdGdA
		T1(DEB): 99.27	dCdCdGdG dGdA
		T1(DEB): 99.28 T1(DEB): 99.31 T1(DEB): 99.32 T1(DEB): 99.34 T1(DEB): 99.35 T1(DEB): 99.36	dGdG Gb Gb dCdGdG Gb Gb dGdG dGdGdG
		T1(DEB): 99.38	dCdTdGdG Gb dTdGdG
		T1(DEB): 99.39 T1(DEB): 99.4	Gb Gb dTdG
		T1(DEB): 99.41 T1(DEB): 99.43 T1(DEB): 99.44 T1(DEB): 99.46 T1(DEB): 99.49 T1(DEB_Gb): 100	Gb Gb Gb Gb dGdG-deoxyribose dGdG dTdG
		T1(DEB_Gb): 19.94; E2(DEB_Gb): 19.94; S3(DEB_Gb): 19.94; L4(DEB_Gb): 19.94; V5(DEB_Gb): 19.94 T1(DEB_Gb): 47.56; E2(DEB_Gb): 47.56 T1(DEB_Gb): 49.18; E2(DEB_Gb): 49.18 T1(DEB_Gb): 49.24; E2(DEB_Gb): 49.24 T1(DEB_Gb): 49.35; E2(DEB_Gb): 49.35 T1(DEB_Gb): 49.45; E2(DEB_Gb): 49.45 T1(DEB_Gb): 49.68; E2(DEB_Gb): 49.68 T1(DEB_Gb): 49.99; E2(DEB_Gb): 49.99	Gb dCdTdGdGdG dCdTdGdAdA dCdCdGdA dCdCdGdGdG Gb dTdTdGdAdA dCdCdG dCdTdTdGdG dTdG dTdGdA
		T1(DEB_Gb): 50; E2(DEB_Gb): 50	dCdCdG dCdCdGdA dCdCdGdG dCdGdAdA

			dCdTdG dCdTdGdG dTdG dTdGdA dTdGdG dTdTdGdA dTdTdTdA
		T1(DEB_Gb): 98.77 T1(DEB_Gb): 98.79 T1(DEB_Gb): 98.92	dCdCdGdGdA dTdGdGdA dCdCdTdGdG dCdGdGdGdG
		T1(DEB_Gb): 98.97	dCdCdTdG dGdA dGdGdGdA
		T1(DEB_Gb): 98.99 T1(DEB_Gb): 99 T1(DEB_Gb): 99.03 T1(DEB_Gb): 99.08	dCdGdGdG dCdTdGdGdG dCdCdGdGdA dCdCdGdG dTdTdGdA
		T1(DEB_Gb): 99.12 T1(DEB_Gb): 99.14 T1(DEB_Gb): 99.15 T1(DEB_Gb): 99.23 T1(DEB_Gb): 99.24	dCdTdG dTdGdG dGdGdG Gb dTdG dTdGdA
		T1(DEB_Gb): 99.99	dCdG Gb
	TESLVLSPAPAKPKR	Too many isoforms V5(DEB_Gb): 98.59 K12(DEB_Gb): 24.62; P13(DEB_Gb): 24.62; K14(DEB_Gb): 24.62; R15(DEB_Gb): 24.62 P13(DEB): 32.92; K14(DEB): 32.92; R15(DEB): 32.92 P13(DEB): 33.22; K14(DEB): 33.22; R15(DEB): 33.22 P13(DEB): 49.58; K14(DEB): 49.58 P13(DEB): 49.85; K14(DEB): 49.85 P13(DEB_Gb): 49.68; K14(DEB_Gb): 49.68 P13(DEB_Gb): 49.74; K14(DEB_Gb): 49.74	Gb Gb Gb dCdTdG dCdTdGdA dGdG Gb dCdTdGdG dTdG dTdGdA
		P13(DEB_Gb): 49.75; K14(DEB_Gb): 49.75	dCdTdG dCdTdGdG
		P8(DEB_Gb): 12.47; A9(DEB_Gb): 12.47; P10(DEB_Gb): 12.47; A11(DEB_Gb): 12.47; K12(DEB_Gb): 12.47; P13(DEB_Gb): 12.47; K14(DEB_Gb): 12.47; R15(DEB_Gb): 12.47	dGdGdA
	VGHNADLQIK	G2(DEB_Gb): 49.8; H3(DEB_Gb): 49.8 H3(DEB_Gb): 100 H3(DEB_Gb): 98.28	Gb Gb dGdG

Supplementary Table 4: RNP^{xl} search results of the DEB crosslinked H5 linker histone- 187 bp ds DNA complex

Protein Name (UniProt ID)	Peptide Sequence	DNA adduct(s)	Crosslinking site in peptide
Histone H5 (P02259)	AKSGARKSPK	dCdG-deoxyribose	No localization
	ASPKKAKKPK	dAdCdGdG-deoxyribose	No localization
	ASRRSASHPTYSEMIAAAIR	dAdAdG	No localization
		dAdGdT	No localization
	GVGASGSFR	Gb	G1
	KAKKPK	dAdAdG-deoxyribose	No localization
	KSRASPK	dAdCdG-deoxyribose	No localization
	LLAAGVLK	Gb	L1
	LLAAGVLKQTK	dAdCdCdG	No localization
		dAdCdG	No localization
		dAdCdGdT	No localization
		dAdG	Q9
		dAdGdG	Q9
		dAdGdGdT	No localization
		dAdGdT	No localization
		dCdG	No localization
		dCdGdT	No localization
			No localization
		Gb	K8
			No localization
		dGdG	No localization
			No localization
		dGdGdG	No localization
		dGdGdT	No localization
		dGdT	No localization
	QSIQKYIK	Gb	No localization
	QTKGVGASGSFR	dAdGdGdT	No localization
		dAdGdTdT	No localization
		dCdGdT	No localization
		Gb	G4
			K3
			No localization
		dGdG	K3
		dGdT	K3
	SASHPTYSEMIAAAIR	dAdAdTdT	No localization
		dAdCdCdG	No localization
		dAdCdGdT	A12
			No localization
			S1
		dAdCdT	No localization
			Y7

		dAdCdTdT	No localization
		dAdGdT	No localization
		dAdGdTdT	No localization
		dAdT	No localization S1
		dAdTdT	No localization
		dAdTdTdT	No localization
		dCdG	No localization
		dCdGdT	No localization
		dCdT	No localization
		Gb	E9 No localization P5 S1 S3 Y7
		dGdG	No localization S1
		dGdGdT	No localization
		dGdT	No localization Y7
	SPAKKPKATAR	dAdGdTdT	No localization
	TESLVLSPAPAKPK	dAdAdCdG	P8
		dAdAdCdT	T1 V5
		dAdAdT	V5
		dAdAdTdT	T1
		dAdCdCdG	T1
		dAdCdCdT	No localization
		dAdCdG	T1
		dAdCdGdT	A9 P8 T1 V5
		dAdCdT	S7 T1
		dAdCdTdT	T1 V5
		dAdG	P8 T1
		dAdGdGdG	T1
		dAdGdGdT	T1
		dAdGdT	T1
		dAdGdTdT	T1 V5
		dAdT	P8

			T1
		dAdTdT	T1
		dAdTdTdT	T1
		dCdCdG	T1
		dCdCdGdT	L4
			V5
		dCdCdT	P8
		dCdCdTdT	V5
		dCdG	P8
		dCdGdGdG	S7
		dCdGdT	T1
			V5
		dCdGdTdT	T1
		dCdT	T1
		dCdTdT	T1
		Gb	L4
			P8
			S3
			S7
			T1
			V5
		dGdG	P8
			S7
			T1
			V5
		dGdGdG	T1
		dGdGdGdG	S7
		dGdGdT	A11
			L4
			T1
		dGdT	L4
			L6
			T1
		dGdTdT	T1
			V5
	TESLVLSPAPAKPKR	dAdCdGdT	No localization
		dAdGdG	T1
		dAdGdT	T1
		dAdGdTdT	No localization
			V5
		dAdTdG-deoxyribose	No localization
		dCdGdT	A9
			No localization
		Gb	L6
			No localization
			S7
		dGdG	L4

		dGdT	No localization
	VGHNADLQIK	dAdAdTdT	No localization
		dAdCdGdT	No localization
		dAdCdTdT	No localization
		dAdG	V1
		dAdGdG	H3 No localization
		dAdGdT	No localization
		dAdGdTdT	No localization
		dAdT	H3
		dCdGdT	H3 No localization
		Gb	H3 N4 V1
		dGdG	H3 No localization
		dGdGdG	No localization
		dGdGdT	No localization
		dGdT	H3

Supplementary Table 5: Identified crosslinked proteins, their respective crosslinked peptide sequences and DNA compositions of the DEB crosslinked, *in vitro* reconstituted mononucleosomes

Protein (Identifier)	Peptide Sequence	DNA Adduct(s)	Crosslinking site in peptide
Histone H2A (GenBank CAD89676.1)	AGLQFPVGR	dG	A1
	HLQLAVR	dG	H1
	KGNYAER	dT	K1
	NDEELNK	dG	E3
			N1
	NDEELNKLLGR	dA	K7
			No localization
		dC	K7
		dCdG	No localization
		dCdT	No localization
		dG	L8
			No localization
		Gb	K9
			No localization
		dT	K7
			No localization
		dTdT	No localization
	VTIAQGGVLPNIQSVLLPKK	dG	K20
		Gb	K20
		dT	Q5
Histone H2B (GenBank CAD89678.1)	AVTKYTSAK	dG	K4
			No localization
		Gb	K4
		dT	K9
			No localization
	EIQTAVR	dG	E1
	ESYAIYVYKVLK	dG	K9
	HAVSEGTK	dG	H1
		dT	H1
		dTdT	H1
	HAVSEGTKAVTK	dCdT	K12
		dTdT	No localization
	KAVTK	dTdT	K1
	KESYAIYVYK	dT	K1
	LAHYNK	dA	H3
		dAb	H3
		dG	H3
		dT	H3

		dTdT	H3
	LLLPGELAK	dA	K9
		dG	K9
			P4
		Gb	L1
		dT	E6
			K9
	LLLPGELAKHAVSEGTK	dA	K9
		dAdT	No localization
		dG	G5
			H10
			K17
			K9
			L7
			No localization
		Gb	K9
		dT	K17
			No localization
		dTdT	K17
	QVHPDTGISSK	dA	Q1
		Gb	Q1
		dT	H3
	VLKQVHPDTGISSK	dT	D8
			V1
		dTdT	No localization
Histone H3 (GenBank CAD89679.1)	DIQLAR	dG	D1
	EIAQDFK	dA	E1
		dG	E1
	EIAQDFKTDLR	dA	K7
			Q4
		dAdC	No localization
		dAdT	K7
			Q4
		dC	F6
		dCb	F6
		dCdT	K7
			No localization
		dG	F6
			K7
			No localization
		Gb	No localization

			Q4
		dT	F6
			K7
		dTdT	K7
			No localization
	FQSSAVMALQEASEAYLVALFEDTNLCAIHAK	dG	C27
	KLPFQR	dCdT	No localization
		dG	K1
		Gb	K1
		dT	K1
		dTdT	K1
	KQLATK	dT	K1
		dTdT	K1
	QLATKAAR	dG	A3
			No localization
			Q1
		dT	No localization
	SAPATGGVK	dAdT	S1
		dG	S1
		dT	S1
	STELLIR	dA	S1
		dG	S1
	VTIMPKDIQLAR	dCdT	K9
		dG	K6
			P5
		Gb	K6
		dT	A11
			K6
		dTdT	No localization
	YQKSTELLIR	dA	K3
		dAdC	No localization
		dAdT	K3
			No localization
		dCdC	No localization
		dCdG	No localization
		dCdT	K3
			No localization
			S4
		dG	K3
			No localization
		Gb	K3

			No localization
		dGdG	K3
		dT	K3
			No localization
			T5
		dTdT	K3
			No localization
	YRPGTVALR	dA	Y1
		dG	Y1
		dT	Y1
Histone H4 (UniProt P62799)	DAVTYTEHAK	dG	A9
			E7
		dT	H8
	DAVTYTEHAKR	dCdT	Y5
	DNIQGITKPAIR	dAdC	No localization
		dCdCdG	No localization
		dCdG	No localization
		dCdT	No localization
		dG	D1
			I11
			I6
			K8
			No localization
			T7
		Gb	K8
			No localization
			Q4
			T7
		dT	K8
			P9
		dTdT	K8
	GGKGLGK	dT	No localization
	GLGKGGAK	dT	K8
	ISGLIYEETR	dA	No localization
			Y6
		dC	E7
		dG	E7
			E8
			I1
			No localization
		Gb	E8

			I1
		dT	E7
			No localization
			Y6
	VFLENVIR	dA	No localization
		dG	E4
			L3
			V1
		dT	E4
			L3

Supplementary Table 6: Identified crosslinked proteins, their respective crosslinked peptide sequences and DNA compositions of the DEB crosslinked, *in vitro* reconstituted 12mer oligonucleosomes

Protein (Identifier)	Peptide Sequence	DNA adduct(s)	Crosslinking site in peptide
Histone H2A (GenBank CAD89676.1)	AGLQFPVGR	dG	A1
		Gb	A1
	HLQLAVR	dG	H1
	NDEELNK	dG	E3
Histone H2B (GenBank CAD89678.1)	AVTKYTSAK	dG	K9
			No localization
	EIQTAVR	dG	E1
	ESYAIYVYK	dG	E1
	HAVSEGTK	dAdT	H1
		Gb	H1
	KESYAIYVYK	dG	K1
	LAHYNK	dG	H3
			No localization
		Gb	H3
		dT	H3
	LLLPGELAK	dG	L1
			L4
		Gb	L1
		dT	L1
	QVHPDTGISSK	dA	D5
			H3
			No localization
		dG	H3
			No localization
			Q1
		Gb	H3
			Q1
		dT	No localization
Histone H3 (GenBank CAD89679.1)	DIQLAR	dG	D1
		Gb	D1
		dT	D1
	EIAQDFK	dG	E1
		Gb	E1
	EIAQDFKTDLR	dCb	No localization
		dG	K7
			No localization
		Gb	K7
	KLPFQR	dG	K1

	KQLATK	dG	K1
	KSAPSTGGVK	dA	K1
		dAb	No localization
	STELLIR	dA	S1
		dG	S1
		Gb	E3
			S1
		dT	No localization
			S1
	YQKSTELLIR	dG	K3
	YRPGTVALR	dG	Y1
Histone H4 (UniProt P62799)	DAVTYTEHAK	dA	D1
			No localization
		dC	Y5
		dG	A9
			D1
			E7
		Gb	D1
			E7
			E8
			No localization
			Y5
	DNIQGITKPAIR	dCdG	No localization
		dG	D1
			No localization
			T7
		Gb	D1
			K8
			Q4
			T7
	ISGLIYEETR	dA	E8
			No localization
			Y6
		dC	E7
		dG	E7
			E8
			G3
			I1
			No localization

			Y6
		Gb	E7
			G3
			I1
			No localization
		dT	E7
			No localization
	KVLR	Gb	K1
	TLYGFGG	dG	T1
	VFLENVIR	dG	E4
			V1
		Gb	E4
			L3
			V1
		dT	E4
			L3

Supplementary Table 7: Identified crosslinked proteins, their respective crosslinked peptide sequences and DNA compositions of the DEB crosslinked, *in vitro* reconstituted mononucleosome H1.4 linker histone complex (chromatosomes)

Proteins (Identifier)	Peptide sequence	DNA adduct(s)	Crosslinking site in peptide
Histone H1.4 (UniProt ID P10412)	AGAAKAK	dAdCdG	K5
	AKKPAGAAK	dAdGdT	A1
		dCdGdT	A1
	ALAAAGYDVEK	dAdG	E10
		dCdG	No localization
		dG	E1
			E10
		Gb	E10
	ALAAAGYDVEKNNSR	dA	K11
		dCdG	No localization
		dG	No localization
			V9
		Gb	K11
	ASGPPVSELITK	dA	No localization
		dCdG	A1
		dG	A1
			E8
		Gb	A1
		dGdG	A1
	ASGPPVSELITKAVAASK	dG	No localization
		Gb	K12
		dGdT	No localization
	AVKPK	dAdCdG	A1
			No localization
		dAdCdGdT	No localization
		dAdG	A1
		dAdGdG	A1
		dAdGdT	No localization
		dAdT	A1
		dCdGdT	No localization
		Gb	A1
		dGdT	No localization
		dTdT	A1
	GTGASGSFK	dCdG	G1
	GTGASGSFKLNK	dAdCdT	No localization
		dAdG	No localization
		dCdG	L10
			No localization

		dCdGdT	No localization
		dG	K9
		Gb	K9
		dGdT	G1
	GTLVQTK	dG	G1
	GTLVQTKGTGASGSFK	dCdG	V4
		dG	K7
		Gb	K7
		dGdT	K7
	IKLGLK	dCdG	K6 L3
		Gb	L3
		dGdG	No localization
		dGdT	I1
	KALAAAGYDVEK	dAdG	K1
		dCdG	K1
		dG	K1
		Gb	K1
		dGdG	K1
		dGdT	K1
	KAPKSPAK	dAdGdG	No localization
		dGdT	K8
	KASGPPVSELITK	dAdA	No localization
		dAdC	No localization
		dAdCdG	No localization
		dAdCdT	No localization
		dAdG	K1 No localization
		dAdGdT	E9 K1 No localization
		dAdT	K1
		dCdG	K1
		dCdGdT	K1
		dG	K1
		Gb	K1
		dGdG	K1 No localization

		dGdT	K1
	KATGAATPK	dAdG	K1
		dGdT	K1
	LGLKSLVSK	dAdG	No localization
		dCdG	No localization
		dGdG	No localization
	RKASGPPVSELITK	dAdGdG	No localization
		dCdG	No localization
		dGdT	No localization
	SETAPAAPAAPAPAEK	dA	S1
		dAdA	S1
		dAdAdC	S1
		dAdAdG	S1
		dAb	S1
		dAdC	No localization S1
		dAdCdC	S1
		dAdCdG	S1
		dAdCdT	S1
		dAdG	No localization S1
		dAdGdG	No localization S1
		dAdGdT	S1
		dAdT	No localization S1
		dC	S1
		dCdC	No localization S1
		dCdCdG	S1
		dCdCdT	S1
		dCdG	No localization S1 T3
		dCdGdT	S1
		dCdT	S1
		dG	P8

			S1
		Gb	S1
		dGdG	S1
		dGdGdG	S1
		dGdGdT	S1
		dGdT	No localization
			S1
		dGdTdT	S1
		dT	S1
		dTdT	S1
	SETAPAAPAAPAPAEKTPVK	dA	A4
		dCdG	No localization
			S1
		dG	S1
		Gb	S1
		dGdT	S1
	SETAPAAPAAPAPAEKTPVKK	dAdG	S1
		dG	S1
		Gb	S1
		dGdG	S1
		dGdT	S1
	SETAPAAPAAPAPAEKTPVKKK	dCdG	S1
		dG	S1
		dGdG	S1
		dGdT	S1
	SGVSLAALK	dG	S1
	SGVSLAALKK	dG	A6
		Gb	A6
	SLVSK	dG	S1
	SLVSKGTLVQTK	dAdG	No localization
		dAdT	L2
		Gb	K5
		dGdT	No localization
Histone H2A (GenBank CAD89676.1)	AGLQFPVGR	dAdG	A1
		dG	A1
	HLQLAVR	dAdG	H1
		dG	H1
		dGdG	H1
	KGNYAER	dGdT	K1
	NDEELNK	dAdG	No localization
		dAdT	No localization

	NDEELNKLLGR	dG	E3
		dA	No localization
		dAdC	No localization
		dAdG	A5 No localization
		dAdT	No localization
		dCdG	No localization
		dG	K7
		Gb	No localization
		dGdT	No localization
Histone H2B (GenBank CAD89678.1)	AKSAPAPK	dAdAdCdG	A1
		dAdAdG	A1
		dAdAdT	A1
		dAdCdCdG	A1 No localization
		dAdCdG	A1 No localization
		dAdCdGdG	A1
		dAdCdGdT	A6 No localization
		dAdG	A1
		dAdGdG	A1
		dAdGdGdT	No localization
		dAdGdT	A1 A6 No localization
		dAdGdTdT	A6
		dAdT	A1 A6
		dCdCdCdG	A1
		dCdCdG	A1
		dCdCdGdT	A1
		dCdG	No localization
		dCdGdT	A1 A6 No localization
		dCdGdTdT	No localization

		dCdT	A1
		Gb	A1
		dGdG	A1
		dGdGdG	A1
		dGdGdT	A6
		dGdT	A1
			A6
			No localization
		dGdTdT	A6
			No localization
		dTdT	A1
	AVTKYTSAK	dA	K4
		dAdCdG	K4
		dAdG	K4
		dAdGdT	No localization
		dAdT	No localization
		dCdG	K4
		dCdGdT	K4
		dG	K4
		dGdT	No localization
		dT	K4
	EIQTAVR	dA	E1
		dG	E1
	ESYAIYVYK	dG	1
	HAVSEGTK	dAdGdG	H1
		dAdT	H1
		dCdT	H1
		dGdT	H1
		dT	H1
		dTdT	H1
	HAVSEGTKAVTK	dAdT	S4
		dCdG	S4
		dG	No localization
		Gb	K12
		dGdG	No localization
	KESYAIYVYK	dCdG	K1
		Gb	K1
	LAHYNK	dA	H3
		dAdA	H3
		dAdC	H3
		dAdCdG	H3

		dAdG	H3
		dAdT	H3
		dCdC	H3
		dCdCdG	H3
		dCdT	No localization
		dG	H3
		dT	H3
		dTdT	K6 No localization
	LAHYNKR	dAdG	No localization
		dAdT	No localization
		dCdT	No localization
		Gb	H3
	LLLPGELAK	dG	L1 P4
		Gb	L1
	LLLPGELAKHAVSEGTK	dAdGdG	No localization
		dCdG	K17
		dG	K9
		Gb	K9 No localization
	QVHPDTGISSK	dA	H3 Q1
		dAdG	H3 No localization Q1
		dAdGdG	H3
		dC	H3
		dCdC	No localization
		dCdG	D5 H3 Q1
		dG	H3 Q1
		Gb	Q1
		dGdG	H3 Q1
		dGdT	Q1
	SAPAPK	dAdAdG	S1
		dAdG	S1
		dAdGdG	S1
		dAdT	S1

	VLKQVHPDTGISSK	dGdG	S1
		dAdCdG	No localization
		dAdG	L3
		dAdGdG	V1
		dAdGdT	No localization
		dAdT	L3
		dG	L3
		Gb	L3
Histone H3 (GenBank CAD89679.1)	DIQLAR	dG	D1
	EIAQDFK	dA	E1
		dG	E1 K7
		Gb	E1
	EIAQDFKTDLR	dA	K7
		dAdC	E1
		dAdG	No localization
		dCdG	No localization
		dG	E1
		Gb	K7
		dGdT	No localization
	KLPFQR	dAdCdT	K1
		dAdGdG	K1
		Gb	K1
		dGdT	K1
	KQLATK	dCdG	K1
		dGdT	K1
	KSAPATGGVK	dAdCdG	No localization
		dAdCdT	No localization
		dAdG	K1 No localization
		dAdGdG	No localization
		dAdGdT	G7 No localization
		dAdT	No localization
		dCdCdG-deoxyribose	K10
		dCdG	K1
		dCdGdT	K1 No localization
		dG	K1
		Gb	K1
		dGdG	K1 No localization

		dGdT	No localization
	QLATKAAR	dCdG	No localization
	SAPATGGVK	dAdA	S1
		dAdAdG	S1
		dAdC	S1
		dAdCdG	S1
		dAdCdT	S1
		dAdG	S1
		dAdGdT	S1
		dAdT	S1
		dCdG	S1
		dCdT	S1
		dG	S1
		Gb	S1
		dGdG	S1
		dGdT	S1
		dT	S1
		dTdT	S1
	STELLIR	dG	E3 S1
	YQKSTELLIR	dAdGdT	No localization
		dCdG	K3 Y1
		dG	K3
		Gb	K3
		dGdG	Y1
		dGdT	Y1
	YRPGTVALR	dG	Y1
Histone H4 (UniProt P62799)	DAVTYTEHAK	dA	A9
		dAdG	A9 No localization
		dAdGdG	No localization
		dCdC	E7
		dCdG	E7 K10 No localization Y5
		dG	A9 D1 E7

			No localization
		Gb	A9
			E7
			H8
			K10
	DAVTYTEHAKR	dCdG	No localization
		dGdT	No localization
	DNIQGITKPAIR	dA	D1
		dAdA	No localization
		dAdC	No localization
		dAdG	D1
			I6
			No localization
		dCdG	D1
			No localization
		dG	D1
			K8
		Gb	D1
			T7
		dGdG	Q4
		dGdT	D1
			I3
			Q4
	GGKGLGK	dAdG	No localization
	GLGKGGAK	dAdCdT	K8
		dAdGdG	G3
		dAdGdT	K8
		dGdG	No localization
	ISGLIYEETR	dG	I1
			No localization
		Gb	E7
	KVLR	dAdA	No localization
		dAdCdT	K1
		dAdG	K1
			No localization
		dAdGdG	K1
		dAdT	K1
		dCdG	K1
		Gb	K1

Supplementary Table 8: Identified crosslinked proteins, their respective crosslinked peptide sequences, crosslinked DNA compositions and crosslinked deoxynucleotides of the formaldehyde crosslinked, *in vitro* reconstituted mononucleosomes

Protein name (Identifier)	Peptide Sequence	DNA adduct(s)	Crosslinked deoxy nucleotide
Histone H2A (GenBank CAD89676.1)	AGLQFPVGR	dA	dA
		dAdC	dA
		dCdG	dG
		dTdG	dG
	AGLQFPVGRVHR	dA	dA
		dAdC	dA
		dAdG	dA,dG
		dCdG	dG
		dTdA	dA
		dTdG	dG
	HLQLAVR	dA	dA
		dAdA	dA
		dAdC	dA
			dA,dC
		dAdG	dA,dG
		dCdC	dC
		dCdG	dG
			dG,dC
		dG	dG
		dGdG	dG
		dTdA	dA
		dTdG	dG
	HLQLAVRNDEELNK	dA	dA
		dAdA	dA
		dAdC	dA
		dAdG	dA,dG
		dCdG	dG
		dG	dG
		dGdG	dG
		dTdA	dA
		dTdG	dG

	IIPRHLQLAVR	dA dAdC dAdG dCdG dTdA dTdG	dA dA dA,dG dG dA dG
	KGNYAER	dA dTdG	dA dG
	NDEELNK	dA dAdG dCdG dTdG	dA dA,dG dG dG
	SSRAGLQFPVGR	dAdC dCdG dTdG	dA dG dG
	TRIIPR	dCdG	dG
	VHRLLR	dAdG dCdG dTdA dTdG	dA,dG dG dG,dC dA dG
	VTIAQGGVLPNIQSVLLPK	dA dAdC dAdG dCdC dCdG	dA dA dA,dC dA dA,dG dC dG
Histone H2B (GenBank CAD89678.1)	AKSAPAPK	dAdG (3FA) dAdGdG dCdG (3FA)	dA,dG dA,dG dG dG,dC
	EIQTAVR	dA	dA
	EIQTAVRLLLPGELAK	dA dAdA dAdC dCdG	dA dA dA dG

	HAVSEGTK	dA	dA
		dAdA	dA
		dAdC	dA
			dA,dC
		dAdG	dA
			dA,dG
		dCdC	dC
		dCdG	dG
			dG,dC
		dG	dG
		dGdG	dG
		dTdA	dA
		dTdC	dC
		dTdG	dG
	HAVSEGTKAVTK	dAdG (3FA)	dA,dG
		dAdGdG	dA,dG
		dCdG (3FA)	dG,dC
		dGdGdG	dG
		dTdG (3FA)	dG
	IAGEASR	dCdC	dC
	IAGEASRLAHYNK	dA	dA
		dAdA	dA
		dAdC	dA
		dAdG	dA
			dA,dG
		dCdG	dG
		dTdA	dA
		dTdG	dG
	LAHYNK	dA	dA
		dAdA	dA
		dAdC	dA
			dA,dC
		dAdG	dA,dG
		dCdC	dC
		dCdG	dG
			dG,dC
		dG	dG
		dGdG	dG
		dTdA	dA
		dTdC	dC

		dTdG	dG
	LAHYNKR	dAdC dAdG dTdA dTdG	dA dA,dG dA dG
	LLLPGELAKHAVSEGTK	dAdAdG(3FA) dAdG(3FA) dCdG(3FA) dGdGdG	dA,dG dA,dG dG,dC dG
	QVHPDTGISSK	dA dAdC dAdG dC dCdC dCdG dG dGdG dTdA dTdG	dA dA,dC dA,dG dC dC dG dG,dC dG dG dA dG
	RSTITSR	dA dAdC dAdG dCdG dTdA	dA dA dA,dC dA,dG dG dA
	STITSREIQTAVR	dA dAdA dAdC dAdG dCdG dGdG dTdA dTdG	dA dA dA dA,dG dG dG dA dG
	VLKQVHPDTGISSK	dAdG(3FA) dAdGdG dCdG(3FA)	dA,dG dA,dG dG dG,dC

Histone H3 (GenBank CAD89679.1)	DIQLAR	dA dCdG dTdA	dA dG dA
	DIQLARR	dA dAdC dTdA	dA dA dA
	EIAQDFK	dA dAdC dAdG dCdG dG dTdA dTdG	dA dA dA,dG dG dG dA dG
	EIAQDFKTDLR	dAdG(3FA) dAdGdG	dA,dG dG
	IRGERA	dA dCdG dTdA	dA dG dA
	KPHR	dTdA	dA
	KQLATK	dA dAdG dCdG dGdG dTdG	dA dA,dG dG dG dG
	KSAPATGGVK	dAdC dAdGdG dCdC dTdC	dA,dC dA,dG dC dC
	LVREIAQDFK	dA	dA
	QLATK	dCdG	dG
	SAPATGGVK	dAdC dCdC dCdG dTdC	dA,dC dC dG,dC dC
	STELLIR	dAdC dCdC dTdC	dA,dC dC dC dC

	YRPGTVALR	dA dAdA dAdC dAdG dCdG dTdA dTdG	dA dA dA dA,dC dA,dG dG dA dG
	YRPGTVALREIR	dAdC dAdG dCdG dTdA dTdG	dA dA,dG dG dA dG
Histone H4 (UniProt P62799)	DAVTYTEHAK	dA dAdA dAdC dAdG dC dCdC dCdG dG dGdG dTdA dTdC dTdG	dA dA dA dA,dC dA,dG dC dC dG dG,dC dG dG dA dC dG
	DAVTYTEHAKR	dA(3FA)	dA
	DNIQGITK	dA	dA
	DNIQGITKPAIR	dA dAdC dCdG dCdG(2FA) dTdA	dA dA dG dG dA
	QGRTLYGFGG	dA	dA

	RISGLIYEETR	dAdG dCdA(3FA) dCdG dCdG(2FA) dG dGdG dTdG dTdG(2FA)	dA,dG dA dG dG dG dG dG dG
	RQGR	dTdA	dA
	TVTAMDVVYALKR	dGdG(3FA)	dG
	VFLENVIR	dAdC dTdA	dA dA
	VFLENVIRDAVTYTEHAK	dAdA dAdC dAdG dGdG dTdA dTdG	dA dA dA dA,dG dG dA dG
	VLRDNIQGITK	dA dAdA dAdC dAdG dCdG dGdG dTdA dTdG	dA dA dA dA,dG dG dG dA dG
	VLRDNIQGITKPAIR	dA dAdC dAdG dCdG dGdG dTdA dTdG	dA dA dA,dG dG dG dA dG

Supplementary Table 9: Identified crosslinked proteins, their respective crosslinked peptide sequences, crosslinked DNA/RNA compositions and crosslinked (deoxy)nucleotides of the formaldehyde crosslinked HeLa native nucleosomes and nuclear proteins

Protein name (UniProt ID)	Peptide sequence	DNA/RNA adduct(s)	Crosslinked (deoxy) nucleotide
60S acidic ribosomal protein P0 (P05388)	AGAIAPCEVTVPAQ NTGLGPEK	A	A
60S ribosomal protein L12 (P30050)	EILGTAQSVGCNVDGR	dA	dA
	HPHDIIDDINSGAVECPAS	dA	dA
60S ribosomal protein L30 (P83731)	VCTLAIIDPGDSDIIR	dA	dA
Actin, cytoplasmic 1 (P60709)	LCYVALDFEQEMATA ASSSSLEK	A	A
		dA	dA
		dAdA	dA
		dAdT	dA
		dCdG	dG
Core histone macro-H2A.1 (O75367)	FVIHCNSPVWGADK	dA	dA
		dAdC	dA
Filamin-A (P21333)	ATCAPQHGAPGPGPADASK	dA	dA
Heterogeneous nuclear ribonucleoproteins C1/C2 (P07910)	IVGCSVHK	dAdA	dA
Histone H2A type 1-B/E (P04908)	VTIAQGGVLPNIQAVLLPK	dA	dA
Histone H2B type 1-B (P33778)	HAVSEGTKAVTK	dAdG(3FA)	dA,dG
	LLLPGELAKHAVSEGTK	dAdG(3FA)	dA,dG
		dCdG(3FA)	dG,dC

	QVHPDTGISSK	dA	dA
Histone H3.1 (P68431)	EIAQDFKTDLR	dAdG(3FA)	dA,dG
	FQSSAVMALQEACEAYLVGL FEDTNLCAIHAK	dAdC	dA
	VTIMPKDIQLAR	dAdG(3FA)	dA,dG
Histone H3.3 (P84243)	FQSAAIGALQEASEAYLVGLF EDTNLCAIHAK	dA	dA
		dAdC	dA
Histone H4 (P62805)	DAVTYTEHAK	dAdA	dA
	GVLKVFLENVIR	dAdG(3FA)	dA,dG
		dCdG(3FA)	dG,dC
	TVTAMDVVYALKR	dAdG(3FA)	dA,dG
	VFLENVIRDAVTYTEHAK	A	A
		dA	dA
		dAdA	dA
		dAdT	dA
		dGdT	dG
Prelamin-A/C (P02545)	AQNTWGCGNSLR	dA	dA

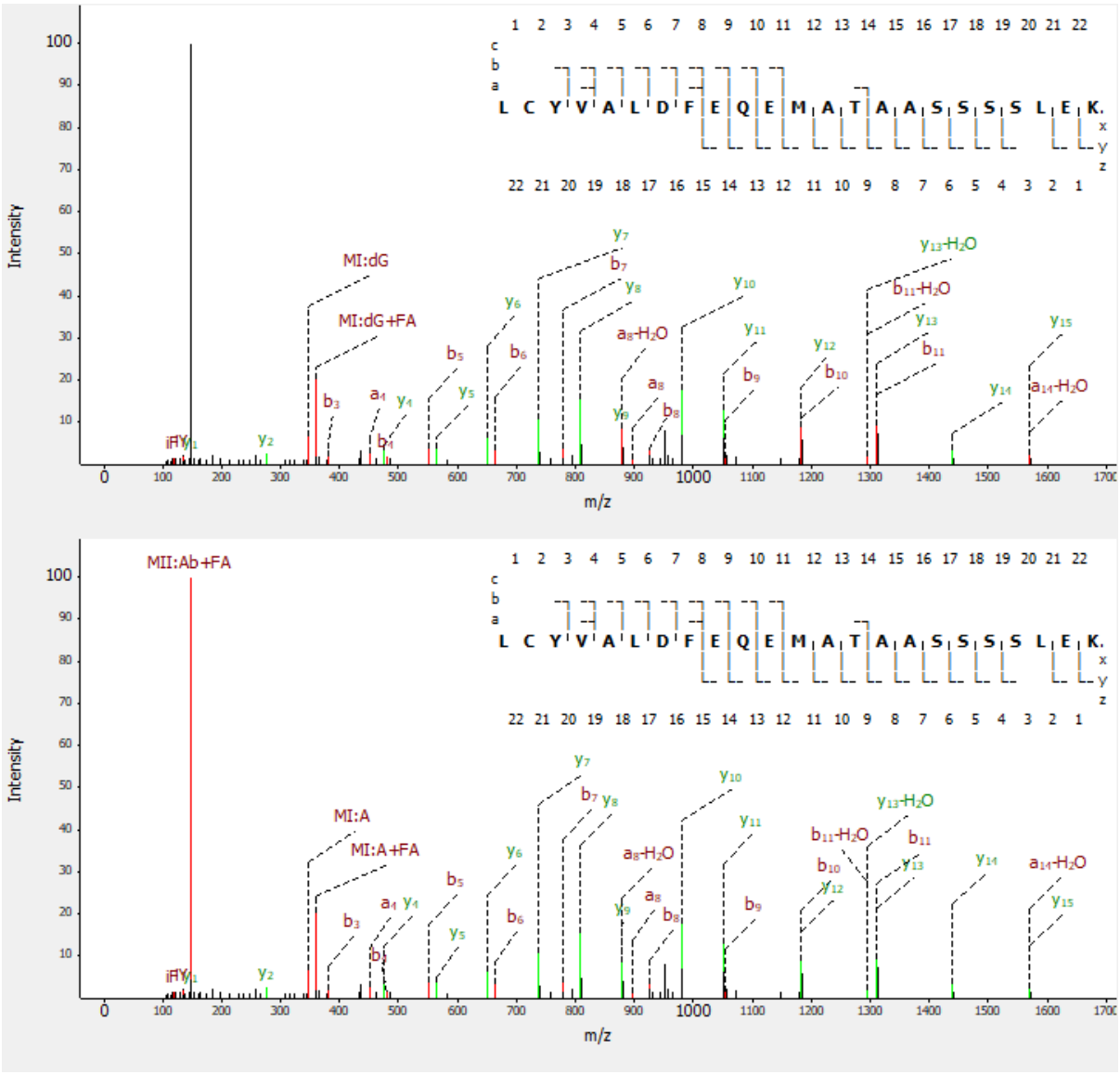
Supplementary Table 10: Identified crosslinked proteins, their respective crosslinked peptide sequences, crosslinked RNA compositions and crosslinked nucleotides of formaldehyde crosslinked 70S ribosome from *Escherichia coli*

Protein Name (UniProt ID)	Crosslinked Peptide	RNA Adduct(s)	Crosslinked Nucleotide
30S ribosomal protein S1 (P0AG67)	DTLHLEGK	A-HPO3	A
	EIETRPGSIVR	A-HPO3	A
30S ribosomal protein S10 (P0A7R5)	FTVLISPHVNK	A-HPO3	A
		C-HPO3	C
30S ribosomal protein S11 (P0A7R9)	ITNITDVTPIPHNGCRPPK	A	A
		A-HPO3	A
		C-HPO3	C
		G-HPO3	G
30S ribosomal protein S12 (P0A7S3)	GALDCSGVK	A-HPO3	A
		C-HPO3	C
		G-HPO3	G
	SNVPALEACPQK	A-HPO3	A
		AG-HPO3	A,G
		C-HPO3	C
		CG-HPO3	G
			G,C
		G-HPO3	G
		GG-HPO3	G
30S ribosomal protein S13 (P0A7S9)	IAGINIPDHK	A-HPO3	A
	LMDLGCYR	A-HPO3	A
		C-HPO3	C
		G-HPO3	G
30S ribosomal protein S17 (P0AG63)	ECRPLSK	A-HPO3	A
	LHVHDENNECGIGDVVEIR	A-HPO3	A
		G-HPO3	G
30S ribosomal protein S19 (P0A7U3)	GPFIDLHLLK	A-HPO3	A
30S ribosomal protein S2 (P0A7V0)	AGVHFGHQTR	A-HPO3	A
	DAALSCDQFFVNHR	A-HPO3	A
		G-HPO3	G
	DMGGLPDALFVIDADHEHIAIK	A-HPO3	A
		C-HPO3	C
	VHIINLEK	A-HPO3	A

30S ribosomal protein S21 (P68679)	EFYEKPTTERK	A-HPO3	A
		GU-HPO3	G
30S ribosomal protein S3 (P0A7V3)	ADIDYNTSEAHTTYGVIGVK	C-HPO3	C
	VHPNGIR	A-HPO3	A
	VPLHTLR	A-HPO3	A
		C-HPO3	C
30S ribosomal protein S4 (P0A7V8)	IEQAPGQHGAR	A-HPO3	A
30S ribosomal protein S5 (P0A7W1)	NMINVALNNGTLQHPVK	A-HPO3	A
30S ribosomal protein S6 (P02358)	HAVTEASPMVK	A-HPO3	A
		C-HPO3	C
30S ribosomal protein S8 (P0A7W7)	QAGLGGEIICYVA	A-HPO3	A
		G-HPO3	G
50S ribosomal protein L1 (P0A7L0)	GATVLPHGTGR	A-HPO3	A
	NGIIHTTIGK	A-HPO3	A
50S ribosomal protein L10 (P0A7J3)	AVEGTPFECLK	A-HPO3	A
		G-HPO3	G
50S ribosomal protein L13 (P0AA10)	VYAGNEHNHAAQQPQVLDI	A-HPO3	A
		C-HPO3	C
	VYYHHTGHIGGIK	A-HPO3	A
		C-HPO3	C
50S ribosomal protein L14 (P0ADY3)	FDGNACVLLNNNSEQPIGTR	A-HPO3	A
		C-HPO3	C
		G-HPO3	G
50S ribosomal protein L19 (P0A7K6)	VFQTHSPVVDSISVK	G-HPO3	G
50S ribosomal protein L2 (P60422)	ATLGEVGNAEHMLR	A-HPO3	A
	CKPTSPGR	A-HPO3	A
	GTAMNPVDHPHGGGEGR	A-HPO3	A
		C-HPO3	C
		CG-HPO3	G
	HPVTPWGVQTK	A-HPO3	A
		C-HPO3	C
	VVNPELHK	A-HPO3	A

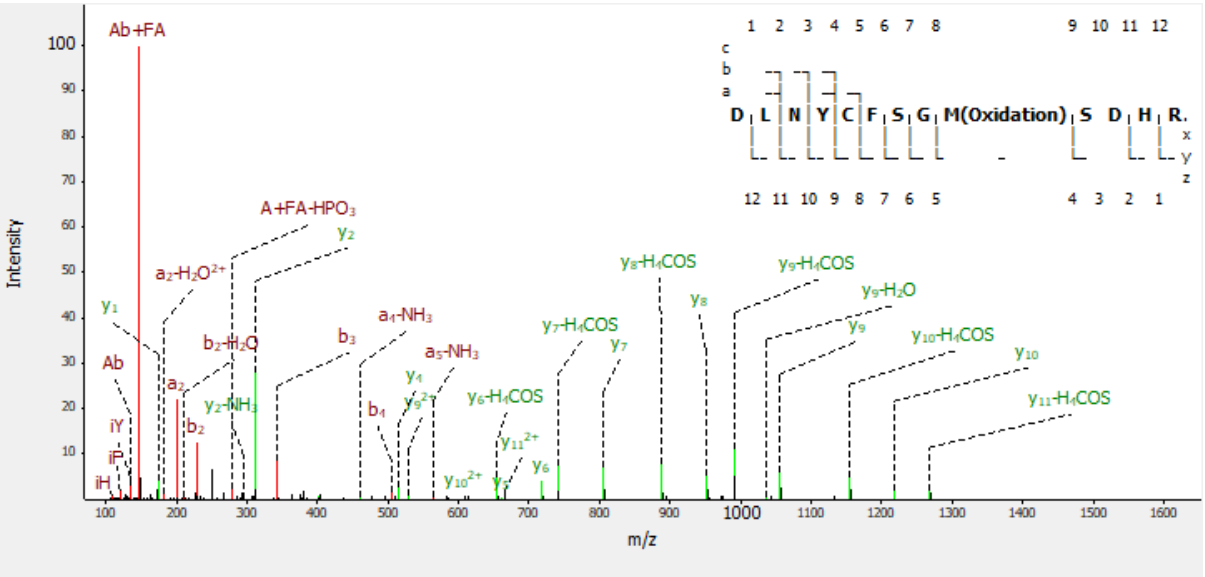
		C-HPO3	C
50S ribosomal protein L21 (P0AG48)	VSEGQTVRLEK	A-HPO3	A
50S ribosomal protein L27 (P0A7L8)	DHTLFAK	A-HPO3	A
	FHAGANVGCGR	A-HPO3	A
		C-HPO3	C
		CG-HPO3	G
		G-HPO3	G
50S ribosomal protein L28 (P0A7M2)	FLPNLHSHR	A-HPO3	A
	SHALNATK	A-HPO3	A
	VCQVTGK	A-HPO3	A
50S ribosomal protein L3 (P60438)	TQDATHGNSLSHR	A-HPO3	A
50S ribosomal protein L31 (P0A7M9)	CHPFFTGK	A-HPO3	A
		C-HPO3	C
	STVGHDLNLDVCSK	A-HPO3	A
		C-HPO3	C
		G-HPO3	G
	YEEITASCSCGNVMK	A	A
50S ribosomal protein L32 (P0A7N4)	HHITADGYYR	A-HPO3	A
	SHDALTAVTSLSVDK	A-HPO3	A
50S ribosomal protein L33 (P0A7N9)	LVSSAGTGHFYTTTK	A-HPO3	A
50S ribosomal protein L35 (P0A7Q1)	GDLGLVIACLPYA	A-HPO3	A
		C-HPO3	C
50S ribosomal protein L36 (P0A7Q6)	VICSAEPK	A-HPO3	A
		C-HPO3	C
		G-HPO3	G
50S ribosomal protein L5 (P62399)	ALLAAFDFPFRK	C-HPO3	C
	QGYPIGCK	A-HPO3	A
		C-HPO3	C
		G-HPO3	G
50S ribosomal protein L6 (P0AG55)	GNVINLSLGFSHPVDH QLPAGITAECPTQTEIVLK	A-HPO3	A
50S ribosomal protein L7/L12 (P0A7K2)	ALEEAGAEVEVK	A	A

Supplementary Figures



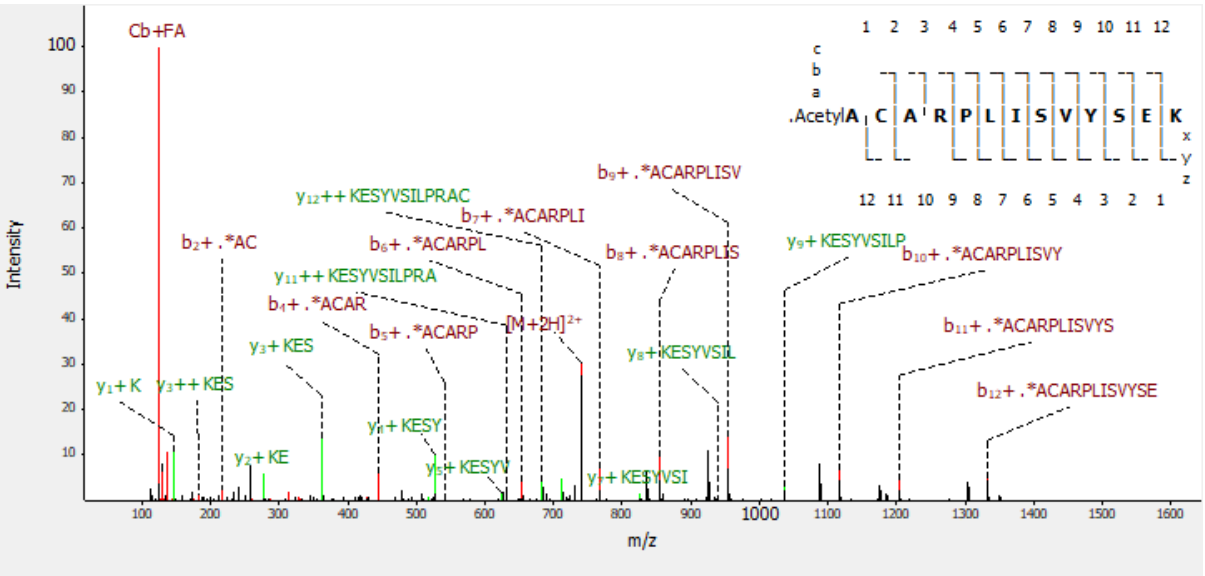
Supplementary Figure 1: MS/MS spectra of the formaldehyde crosslinked peptide, LCYVALDFEQEMATAASSSSLEK of the cytoplasmic actin 1 protein

The upper spectrum is assigned as deoxyguanosine monophosphate crosslink. The lower spectrum is assigned as adenosine monophosphate crosslink. In the lower spectrum, presence of mass shifted marker ion (Ab+FA) indicates that peptide was crosslinked to adenosine monophosphate. Peptide a and b fragment ions are colored in red; y fragment ions are colored in green



Supplementary Figure 2: MS/MS spectrum of the Heterogeneous nuclear ribonucleoprotein H peptide, DLNYCFSGM(Ox)SDHR formaldehyde crosslinked to adenosine

Peptide a and b fragment ions are colored in red; y fragment ions are colored in green. Peptide y ion series with characteristic H_4COS loss indicates that methionine amino acid is oxidized.



Supplementary Figure 3: MS/MS spectrum of 60S ribosomal protein L4 peptide, (Acetyl)ACARPLISVYSEK crosslinked to cytidine

Peptide a and b fragment ions are colored in red, y fragment ions are colored in green. Acetyl mass shifted ion series is highlighted with an asterisk. The acetyl mass shift on the peptide fragments indicates that acetylation was present on the peptide N-terminal.

Supplementary Texts

Supplementary Text 1: R script for data analysis for the MS2/MS3 method

```
#Set directory
setwd("E:/Data analysis/2017/November/PD_worksheets")
#Check
getwd()
# Read PSM info from table exported from PD

#zero everything to reset the values
psmInfo <- c(0)
MSMSInfo <- c(0)
PsmInfo_MSOrder <- c(0)
Reduced_table <- c(0)
Final_table <- c(0)
MassDiff <- c(0)
i= 0
j=0


psmInfo<-
read.table("F_Bazso_300517_140617_H1_4_187bp_DEB_MS3_NCE32_topp2_human_prot
eome_PSMs.txt", dec=".",sep="\t", header = TRUE, stringsAsFactors = FALSE)

# Read MSMS info form table xported from PD
MSMSInfo<-
read.table("F_Bazso_300517_140617_H1_4_187bp_DEB_MS3_NCE32_topp2_human_prot
eome_MSMSSpectrumInfo.txt", dec=".",sep="\t", header = TRUE, stringsAsFactors =
FALSE)

# Extract the rows which are in MS order 3
which(psmInfo$MS.Order!="MS2")
PsmInfo_MSOrder<-psmInfo[which(psmInfo$MS.Order!="MS2"),]

# Extract the rows which only have modifications
which(PsmInfo_MSOrder$Modifications!= "")
Reduced_table<-PsmInfo_MSOrder[which(PsmInfo_MSOrder$Modifications!= ""),]

i=0
j=0

#Transforming DEB+Gb into a DEB modification


for (i in 1:nrow(Reduced_table))
  {

        if(nchar(Reduced_table$Modifications[i])>8)
          {
            Reduced_table$MHplus.in.Da [i]=Reduced_table$MHplus.in.Da[i]-151.0494
          }
}
```

```
Final_table<- Reduced_table

#Connecting MS3 and MS2 levels with each other and calculating delta mass between
precursors of each level

i=0
j=0

MassDiff=matrix(nrow= nrow(Final_table), ncol=8)
for (i in 1:nrow(Final_table))
{
  for (j in 1:nrow(MSMSInfo))
  {
    if(Final_table$Master.Scans[i]== MSMSInfo$First.Scan[j])
    {
      MassDiff[i,1]=Final_table$Annotated.Sequence[i]
      MassDiff[i,2]=Final_table$Modifications[i]
      MassDiff[i,3]=MSMSInfo$Precursor.MHplus.in.Da[j]- Final_table$MHplus.in.Da[i]
      MassDiff[i,4]=Final_table$First.Scan[i]
      MassDiff[i,5]=MSMSInfo$First.Scan[j]
      MassDiff[i,6]=Final_table$ptmRS.Best.Site.Probabilities[i]
    }

  }
}

# Creating a header for MassDiff
headerforMassdiff <- c("Sequence", "Modifications", "Mass difference", "First Scan",
"Master Scan", "ptmRS", "Composition", "DeltamassAccuracy" )
MassDiff <- as.data.frame(MassDiff, stringsAsfactors= FALSE)
names(MassDiff)<- headerforMassdiff

#Creating vector for nucleotide and nucleobases masses and names

nucleotideMasses = c(307.0569, 322.0566, 347.0631, 331.0682)
nucleotideNames <- c("C","T", "G", "A")
nucleoBasesMasses <- c(111.0433, 126.0429, 151.0494, 135.0545)
nucleoBasesNames <- c("Cb", "Tb", "Gb", "Ab")

#creating a list of possible combination for masses

Possiblelosses <- c(18.01056, 17.02655, 97.9769, 79.9663)
PossiblelossesNames <- c("H2O", "NH3", "H3PO4","HPO3")
PossibleDinucleotideMasses <- matrix(nrow=4, ncol=4)
PossibleDinucleotidesNames <- matrix(nrow=4, ncol=4)
 for (i in 1:4)
 {
   if (i==1)
  {
   for (j in 1:4)
    {
```

```
    PossibleDinucleotideMasses[i,j]= nucleotideMasses[i]+nucleotideMasses[j]-18.01056
    PossibleDinucleotidesNames[i,j]= paste(nucleotideNames[i], nucleotideNames[j])
   }
  }

 else if (i==2)
  {
   for (j in 2:4)
   {

    PossibleDinucleotideMasses[i,j]= nucleotideMasses[i]+nucleotideMasses[j]-18.01056
    PossibleDinucleotidesNames[i,j]= paste(nucleotideNames[i], nucleotideNames[j])
   }
  }
  else if (i==3)
  {
   for (j in 3:4)
   {

    PossibleDinucleotideMasses[i,j]= nucleotideMasses[i]+nucleotideMasses[j]-18.01056
    PossibleDinucleotidesNames[i,j]= paste(nucleotideNames[i], nucleotideNames[j])
   }
  }
  else if (i ==4)
  {
   for (j in 4:4)
   {

    PossibleDinucleotideMasses[i,j]= nucleotideMasses[i]+nucleotideMasses[j]-18.01056
    PossibleDinucleotidesNames[i,j]= paste(nucleotideNames[i], nucleotideNames[j])

   }
  }
 }


v<-which(is.na(PossibleDinucleotideMasses[,])=="FALSE")
MassesofDinucleotides= c()
NamesofDiNucleotides = c()
 for (i in 1:10)
  {
  MassesofDinucleotides[i]= PossibleDinucleotideMasses[v[i]]
  NamesofDiNucleotides[i]= PossibleDinucleotidesNames[v[i]]
 }

NamesofDiNucleotides
MassesofDinucleotides
# The same just with the trinucleotides

PossibleTrinucleotidesNames <- matrix(nrow=4, ncol=10)
PossibleTrinucleotidesMasses <- matrix(nrow=4, ncol=10)

for (i in 1:4)
```

```
{
 if (i==1)
 {
  for (j in 1:10)
  {
   PossibleTrinucleotidesMasses[i,j]= nucleotideMasses[i]+MassesofDinucleotides[j]-
18.01056
   PossibleTrinucleotidesNames[i,j]= paste(nucleotideNames[i], NamesofDiNucleotides[j])
  }
 }

 else if (i==2)
 {
  for (j in 3:10)
  {
   PossibleTrinucleotidesMasses[i,j]= nucleotideMasses[i]+MassesofDinucleotides[j]-
18.01056
   PossibleTrinucleotidesNames[i,j]= paste(nucleotideNames[i], NamesofDiNucleotides[j])
  }
 }
 else if (i==3)
 {
  for (j in 6:10)
  {
   PossibleTrinucleotidesMasses[i,j]= nucleotideMasses[i]+MassesofDinucleotides[j]-
18.01056
   PossibleTrinucleotidesNames[i,j]= paste(nucleotideNames[i], NamesofDiNucleotides[j])
  }
 }
 else if (i ==4)
 {
  for (j in 10:10)
  {
   PossibleTrinucleotidesMasses[i,j]= nucleotideMasses[i]+MassesofDinucleotides[j]-
18.01056
   PossibleTrinucleotidesNames[i,j]= paste(nucleotideNames[i], NamesofDiNucleotides[j])
  }
 }
}

u<-which(is.na(PossibleTrinucleotidesNames[,])=="FALSE")
u
NamesofTrinucleotides= c()
MassesofTrinucleotides = c()
for (i in 1:24)
{
 NamesofTrinucleotides[i]= PossibleTrinucleotidesNames[u[i]]
 MassesofTrinucleotides[i]= PossibleTrinucleotidesMasses[u[i]]
}
NamesofTrinucleotides
MassesofTrinucleotides

# and the same for tetranucleotides
```

```
PossibleTetranucleotidesNames <- matrix(nrow=4, ncol=90)
PossibleTetranucleotidesMasses <- matrix(nrow=4, ncol=90)
Possible5nucleotidesNames <- matrix(nrow=4, ncol=90)
Possible5nucleotidesMasses <- matrix(nrow=4, ncol=90)
for (i in 1:4)
{
 if (i==1)
 {
  for (j in 1:90)
  {
   PossibleTetranucleotidesMasses[i,j]= nucleotideMasses[i]+MassesofTrinucleotides[j]-
18.01056
   PossibleTetranucleotidesNames[i,j]= paste(nucleotideNames[i],
NamesofTrinucleotides[j])
   Possible5nucleotidesMasses[i,j]= MassesofDinucleotides[i]+MassesofTrinucleotides[j]-
18.01056
   Possible5nucleotidesNames[i,j]= paste(NamesofDiNucleotides[i],
NamesofTrinucleotides[j])
  }
 }

 else if (i==2)
 {
  for (j in 2:90)
  {
   PossibleTetranucleotidesMasses[i,j]= nucleotideMasses[i]+MassesofTrinucleotides[j]-
18.01056
   PossibleTetranucleotidesNames[i,j]= paste(nucleotideNames[i],
NamesofTrinucleotides[j])
   Possible5nucleotidesMasses[i,j]= MassesofDinucleotides[i]+MassesofTrinucleotides[j]-
18.01056
   Possible5nucleotidesNames[i,j]= paste(NamesofDiNucleotides[i],
NamesofTrinucleotides[j])
  }
 }
 else if (i==3)
 {
  for (j in 3:90)
  {
   PossibleTetranucleotidesMasses[i,j]= nucleotideMasses[i]+MassesofTrinucleotides[j]-
18.01056
   PossibleTetranucleotidesNames[i,j]= paste(nucleotideNames[i],
NamesofTrinucleotides[j])
   Possible5nucleotidesMasses[i,j]= MassesofDinucleotides[i]+MassesofTrinucleotides[j]-
18.01056
   Possible5nucleotidesNames[i,j]= paste(NamesofDiNucleotides[i],
NamesofTrinucleotides[j])
  }
 }
 else if (i ==4)
 {
  for (j in 4:90)
```

```
   {
     PossibleTetranucleotidesMasses[i,j]= nucleotideMasses[i]+MassesofTrinucleotides[j]-
18.01056
     PossibleTetranucleotidesNames[i,j]= paste(nucleotideNames[i],
NamesofTrinucleotides[j])
     Possible5nucleotidesMasses[i,j]= MassesofDinucleotides[i]+MassesofTrinucleotides[j]-
18.01056
     Possible5nucleotidesNames[i,j]= paste(NamesofDiNucleotides[i],
NamesofTrinucleotides[j])
   }
  }
}

u<-which(is.na(PossibleTetranucleotidesNames[,])=="FALSE")
u
NamesofTetranucleotides= c()
MassesofTetranucleotides = c()
Namesof5nucleotides= c()
Massesof5nucleotides = c()
for (i in 1:90)
{
  NamesofTetranucleotides[i]= PossibleTetranucleotidesNames[u[i]]
  MassesofTetranucleotides[i]= PossibleTetranucleotidesMasses[u[i]]
  Namesof5nucleotides[i]= Possible5nucleotidesNames[u[i]]
  Massesof5nucleotides[i]= Possible5nucleotidesMasses[u[i]]

}
NamesofTetranucleotides
MassesofTetranucleotides


# Creating a vector with all possible losses and nucleotide combinations up to 3 nucleotides


AllCombinaionMasses1 <- c(nucleotideMasses, MassesofDinucleotides,
MassesofTrinucleotides, MassesofTetranucleotides, Massesof5nucleotides)
AllCombinaionMasses2<- c(MassesofDinucleotides- 98.0367, MassesofDinucleotides-
196.0137)
AllCombinaionMasses <- c(86.0367, Possiblelosses,98.0367, 196.0137, nucleoBasesMasses,
AllCombinaionMasses1, AllCombinaionMasses2)

#Creating the same with the nucleotide names
AllCombinaionNames1 <- c(nucleotideNames, NamesofDiNucleotides,
NamesofTrinucleotides, NamesofTetranucleotides, Namesof5nucleotides)
AllCombinaionNames2<- c(paste(NamesofDiNucleotides, "-sugar"),
paste(NamesofDiNucleotides, "-sugar+phosphate"))
AllCombinaionNames <- c("DEB", PossiblelossesNames,"sugar", "sugar+phoshate",
nucleoBasesNames, AllCombinaionNames1, AllCombinaionNames2)
AllCombinaionNames
#Searching against the allcombinationmasses the MS2 and MS3 level's mass difference

MassDiff$`Mass difference`<- as.numeric(as.character(MassDiff$`Mass difference`))
MassDiff$Composition <- as.character(MassDiff$Composition)
```

```
MassDiff$Composition <- "Unknown"
MassDiff$DeltamassAccuracy <- as.numeric(as.character(MassDiff$DeltamassAccuracy))


for (i in 1:nrow(MassDiff))
{
  for (j in 1: length (AllCombinaionMasses))
  {
    if (abs(((MassDiff$`Mass difference`[i]-
AllCombinaionMasses[j])/AllCombinaionMasses[j])*1000000) < 50)
    {
      MassDiff$Composition[i]= AllCombinaionNames[j]
      MassDiff$DeltamassAccuracy[i]= abs(((MassDiff$`Mass difference`[i]-
AllCombinaionMasses[j])/AllCombinaionMasses[j])*1000000)
    }

  }
}

AllCombinaionMasses

AllCombinaionNames
#if (j ==length(AllCombinaionMasses) & valami ==T)
#{
#  MassDiff$Composition[i]= "Unknown"
#}


MassDiff2<- MassDiff
MassDiff2<-MassDiff2[which(MassDiff2$Composition!= "Unknown"),]
MassDiff2<-MassDiff2[which(MassDiff2$Composition!= "H3PO4"),]
write.table(MassDiff2,
"F_Bazso_230817_240817_H4_XL_MS3_topp2_results_nucleotides_see.csv",col.names=T
RUE,sep = ";",row.names = FALSE,dec=".")
```

Supplementary Text 2: MS acquisition parameters for MS2/MS2 trigger method

Orbitrap Fusion Lumos Method Summary

Creator: P1608-71\Orbitrap_Lumos Last Modified: 07/02/2020 21:30:48 by P160871\Orbitrap_Lumos

Global Settings
Use Ion Source Settings from Tune = False
Method Duration (min)= 58
Ion Source Type = NSI
Spray Voltage: Positive Ion (V) = 2300
Spray Voltage: Negative Ion (V) = 600
Infusion Mode (LC)= False
Sweep Gas (Arb) = 0
Ion Transfer Tube Temp (°C) = 275
APPI Lamp = Not in use
FAIMS Mode = Not Installed
Contact Closure

Time (min)|	State|
0|	Open
0.2|	Closed

Internal Mass Calibration= User Defined Lock Mass
Application Mode = Peptide
Pressure Mode = Standard
Default Charge State = 2
Advanced Peak Determination = False
Xcalibur AcquireX enabled for method modifications = False Internal

Cal Positive

m/z
445.12003

Experiment 1
Experiment Name = MS
Start Time (min) = 0
End Time (min) = 58
Cycle Time (sec) = 3

Scan MasterScan
Desired minimum points across the peak = 6
MSn Level = 1
Use Wide Quad Isolation = True
Detector Type = Orbitrap
Orbitrap Resolution = 120K
Mass Range = Normal
Scan Range (m/z) = 350-1580
Maximum Injection Time (ms) = 50
AGC Target = 1000000
Normalized AGC Target = 250%

Microscans = 1
Maximum Injection Time Type = Custom
RF Lens (%) = 60
Use ETD Internal Calibration = False
DataType = Profile
Polarity = Positive
Source Fragmentation = False
Scan Description =
Enhanced Resolution Mode = Off

Filter ChargeState
Include charge state(s) = 2-6

Include undetermined charge states = False

Filter DynamicExclusion
Exclude after n times = 1
Exclusion duration (s) = 20
Mass Tolerance = ppm
Mass tolerance low = 10
Mass tolerance high = 10
Use Common Settings = False
Exclude isotopes = True
Perform dependent scan on single charge state per precursor only = False

Filter IntensityThreshold
Maximum Intensity = 1E+20
Minimum Intensity = 50000
Relative Intensity Threshold = 0
Intensity Filter Type = IntensityThreshold

Filter MIPS
Relax Restrictions when too few Precursors are Found = True
MIPS Mode = Peptide

Data Dependent Properties
Data Dependent Mode= Cycle Time
Scan Event 1

Scan ddMSnScan
MSn Level = 2
Desired minimum points across the peak = 6
Isolation Mode = Quadrupole
Isolation Window = 1.4
Isolation Offset = Off
Enable Auto PTR Windows = False
Reported Mass = Original Mass

Multi-notch Isolation = False
Scan Range Mode = Define First Mass
FirstMass = 105
Scan Priority= 1
Collision Energy Mode = Fixed
ActivationType = HCD
Collision Energy (%) = 32
Detector Type = Orbitrap
Orbitrap Resolution = 30K
Maximum Injection Time (ms) = 80
AGC Target = 50000
Normalized AGC Target = 100%
Inject ions for all available parallelizable time = False
Microscans = 1
Maximum Injection Time Type = Custom
Use ETD Internal Calibration = False
DataType = Profile
Polarity = Positive
Source Fragmentation = False
Scan Description =
Enhanced Resolution Mode = Off

Filter ProductIonTrigger
Use Groups = False
trigger only when at least n product ions from list are detected = True
Group ID must equal Isolation (m/z) = False
n = : 1
product ion(s) must be wihin top n = True

Top N Product Ions = 10
Mass tolerance low = 8
Mass Tolerance = ppm
Mass list type = m/z Mass tolerance high = 8 product ions(s) must

be above threshold (relative intensity, %) = False

Ignore charge state requirement for unassigned ions = False
Trigger Only When Correct Charge State of Product Ion is Detected = True
Product Ions Threshold = 0
Fail trigger if conditions are met = False

>>>>>>>>>>>>> Mass List Table <<<<<<<<<<<<<

CompoundName	m/z
Cb	112.0511
Cb+FA	124.0511
Ab	136.0623
Ab+FA	148.0623
Gb	152.0572
Gb+FA	164.0572
T	161.0324

>>>>>>>>>>>>> End Mass List Table <<<<<<<<<<<<<

Data Dependent Properties
Data Dependent Mode= Number of Scans
Number of Dependent Scans= 1

Filter ChargeState
Include charge state(s) = 2
Include undetermined charge states = False

Scan ddMSnScan
Desired minimum points across the peak = 6
MSn Level = 2
Isolation Mode = Quadrupole
Isolation Offset = Off
Isolation Window = 1.6

Enable Auto PTR Windows = False

Reported Mass = Original Mass

Multi-notch Isolation = False

Scan Range Mode = Define First Mass
FirstMass = 105

Scan Priority= 1

Collision Energy Mode = Fixed
ActivationType = HCD
Collision Energy (%) = 38
Detector Type = Orbitrap
Orbitrap Resolution = 60K
Maximum Injection Time (ms) = 160
AGC Target = 50000
Inject ions for all available parallelizable time = False
Normalized AGC Target = 100%
Microscans = 1
Maximum Injection Time Type = Custom
Use ETD Internal Calibration = False
DataType = Profile
Polarity = Positive
Source Fragmentation = False
Scan Description =
Enhanced Resolution Mode = Off

Filter ChargeState
Include charge state(s) = 3-6
Include undetermined charge states = False

Scan ddMSnScan
Desired minimum points across the peak = 9

MSn Level = 2
Isolation Mode = Quadrupole

Isolation Offset = Off
Isolation Window = 1.4

Enable Auto PTR Windows = False
Reported Mass = Original Mass
Multi-notch Isolation = False
Scan Range Mode = Define First Mass
FirstMass = 105

Scan Priority= 1

Collision Energy Mode = Fixed
ActivationType = HCD
Collision Energy (%) = 36
Detector Type = Orbitrap
Orbitrap Resolution = 60K
Maximum Injection Time (ms) = 160
AGC Target = 50000
Inject ions for all available parallelizable time = False
Normalized AGC Target = 100%
Microscans = 1
Maximum Injection Time Type = Custom
Use ETD Internal Calibration = False
DataType = Profile
Polarity = Positive
Source Fragmentation = False
Scan Description =
Enhanced Resolution Mode = Off

Supplementary Text 3: R script for the validation of the formaldehyde crosslinked peptide-DNA heteroconjugates

```
# The RNPxl.RNA colum name cannot contain FA as “+C “, please remove previously or split the colum by “+” into two columns
#Set directory
setwd("/Users/Fanni/Documents/Work/monos")
#Check
getwd()
#install packages
install.packages("stringr")
BiocManager::install(version = "3.12")
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager")
BiocManager::install("mzR")
install.packages("stringi")
install.packages("protr")

#load packages
library("protr")
library("stringi")
library("stringr")
library("protr")
library("mzR")

#read in mzml file
ms<-openMSfile("F_Bazso_290120_300120_HeLa_nucl_TiO2_lu_cc.mzml")
hd <- header(ms)
hd2 <- hd[hd$msLevel == 2, ]
 #zero everything to reset the values
RNPxl <- c(0)
i= 0
j=0

# read in RNPxl table
RNPxl<-
read.csv("/Users/Fanni/Documents/Work/open_FA_purchased/HeLa_nucl_forthesis.csv", dec=".", header = TRUE, stringsAsFactors = FALSE)

#adjusting index numbers for RNPxl files

Filtered_RNPxl<-RNPxl[which(RNPxl$RNPxl.RNA!="none"), ]
 for (i in 1:nrow(RNPxl))
   {
  RNPxl$index[i]=RNPxl$index[i]+1

 }

Filtered_RNPxl<-c(0)
Filtered_RNPxl<-RNPxl[which(RNPxl$RNPxl.RNA!="none"), ]
```

```
Filtered_RNPxl$index<-unlist(Filtered_RNPxl$index)

 #Looking for respecting marker ions based on the DNA adduct

Filtered_RNPxl$A<-c(0)
Filtered_RNPxl$A_FA<-c(0)
Filtered_RNPxl$G<-c(0)
Filtered_RNPxl$G_FA<-c(0)
Filtered_RNPxl$T<-c(0)
Filtered_RNPxl$C<-c(0)
Filtered_RNPxl$C_FA<-c(0)


for (i in 1:nrow(Filtered_RNPxl))

{j=1
    while(peaks(ms,Filtered_RNPxl$index[[i]])[j]<165)
      {
         if ((grepl( "A",Filtered_RNPxl$RNPxl.RNA[i], fixed = TRUE)))
           {   if(abs(((peaks(ms,Filtered_RNPxl$index[[i]])[j]-
136.0623)/136.0623)*1000000)<10)
             {

Filtered_RNPxl$A[i]=(peaks(ms,Filtered_RNPxl$index[[i]])[,2][j]/max(peaks(ms,Filtered_R
NPxl$index[[i]])[,2]))*100
             }
               else if(abs(((peaks(ms,Filtered_RNPxl$index[[i]])[j]-
148.0623)/148.0623)*1000000)<10)
                 {

Filtered_RNPxl$A_FA[i]=(peaks(ms,Filtered_RNPxl$index[[i]])[,2][j]/max(peaks(ms,Filtere
d_RNPxl$index[[i]])[,2]))*100
                 }

           }

           if ((grepl( "G",Filtered_RNPxl$RNPxl.RNA[i], fixed = TRUE)))
             {   if(abs(((peaks(ms,Filtered_RNPxl$index[[i]])[j]-
152.0572)/152.0572)*1000000)<10)
               {

Filtered_RNPxl$G[i]=(peaks(ms,Filtered_RNPxl$index[[i]])[,2][j]/max(peaks(ms,Filtered_R
NPxl$index[[i]])[,2]))*100
               }
                 else if(abs(((peaks(ms,Filtered_RNPxl$index[[i]])[j]-
164.0572)/164.0572)*1000000)<10)
                   {

Filtered_RNPxl$G_FA[i]=(peaks(ms,Filtered_RNPxl$index[[i]])[,2][j]/max(peaks(ms,Filtere
d_RNPxl$index[[i]])[,2]))*100
                   }
               }
```

```
          if ((grepl( "C",Filtered_RNPxl$RNPxl.RNA[i], fixed = TRUE)))
            {   if(abs(((peaks(ms,Filtered_RNPxl$index[[i]])[j]-
112.0511)/112.0511)*1000000)<10)
              {

Filtered_RNPxl$C[i]=(peaks(ms,Filtered_RNPxl$index[[i]])[,2][j]/max(peaks(ms,Filtered_R
NPxl$index[[i]])[,2]))*100
              }
                else if(abs(((peaks(ms,Filtered_RNPxl$index[[i]])[j]-
124.0511)/124.0511)*1000000)<10)
                {

Filtered_RNPxl$C_FA[i]=(peaks(ms,Filtered_RNPxl$index[[i]])[,2][j]/max(peaks(ms,Filtere
d_RNPxl$index[[i]])[,2]))*100
                }
              }

          if ((grepl( "T",Filtered_RNPxl$RNPxl.RNA[i], fixed = TRUE)))
            {   if(abs(((peaks(ms,Filtered_RNPxl$index[[i]])[j]-
161.0595)/161.0595)*1000000)<10)
              {

Filtered_RNPxl$T[i]=(peaks(ms,Filtered_RNPxl$index[[i]])[,2][j]/max(peaks(ms,Filtered_R
NPxl$index[[i]])[,2]))*100
                }
              }

          j=j+1
        }
    }

Filtered_RNPxl$DNAAdduct<-c(0)



  #Check if the respective marker ions are present based on the crosslinked deoxynucleotide
adduct composition

  for (i in 1:nrow(Filtered_RNPxl))
    {  Filtered_RNPxl$DNAAdduct[i]=Filtered_RNPxl$RNPxl.RNA[i]

        if (grepl("A", Filtered_RNPxl$RNPxl.RNA[i], fixed=TRUE) &
Filtered_RNPxl$A[i]==0 & Filtered_RNPxl$A_FA[i]==0)

        {
           Filtered_RNPxl$DNAAdduct[i]= "Not XL"
         }
       if (grepl("G", Filtered_RNPxl$RNPxl.RNA[i], fixed= TRUE) &
Filtered_RNPxl$G[i]==0 & Filtered_RNPxl$G_FA[i]==0)
```

```
        {
          Filtered_RNPxl$DNAAdduct[i]= "Not XL"
        }
      if (grepl("C", Filtered_RNPxl$RNPxl.RNA[i], fixed=TRUE)&
Filtered_RNPxl$C[i]==0 & Filtered_RNPxl$C_FA[i]==0)

        {
          Filtered_RNPxl$DNAAdduct[i]= "Not XL"
      }

  }

Filtered_RNPxl$CrosslinkedNucleotide=c(0)

 #Find  the crosslinked deoxynucleotide(s)
 for (i in 1:nrow(Filtered_RNPxl))
    {

        if (Filtered_RNPxl$A_FA[i]!=0 &Filtered_RNPxl$G_FA[i]==0 &
Filtered_RNPxl$C_FA[i]==0)
          {
            Filtered_RNPxl$CrosslinkedNucleotide[i]="A"
          }

        else if (Filtered_RNPxl$G_FA[i]!=0 & Filtered_RNPxl$A_FA[i]==0 &
Filtered_RNPxl$C_FA[i]==0)
          {
            Filtered_RNPxl$CrosslinkedNucleotide[i]="G"
          }

        else if (Filtered_RNPxl$C_FA[i]!=0 &Filtered_RNPxl$G_FA[i]==0 &
Filtered_RNPxl$A_FA[i]==0)
          {
            Filtered_RNPxl$CrosslinkedNucleotide[i]="C"
          }

        else if (Filtered_RNPxl$C_FA[i]!=0 &Filtered_RNPxl$A_FA[i]!=0 &
Filtered_RNPxl$G_FA[i]==0)
          {
            Filtered_RNPxl$CrosslinkedNucleotide[i]="A,C"
          }
       else if (Filtered_RNPxl$A_FA[i]==0 &  Filtered_RNPxl$C_FA[i]!=0 &
Filtered_RNPxl$G_FA[i]!=0 )
        {
          Filtered_RNPxl$CrosslinkedNucleotide[i]="G,C"
        }
       else if (Filtered_RNPxl$C_FA[i]==0 &Filtered_RNPxl$G_FA[i]!=0
&Filtered_RNPxl$A_FA[i]!=0)
        {
          Filtered_RNPxl$CrosslinkedNucleotide[i]="A,G"
        }
```

```
        else if (Filtered_RNPxl$C_FA[i]==0 &Filtered_RNPxl$G_FA[i]==0 &
Filtered_RNPxl$A_FA[i]==0 & Filtered_RNPxl$T[i]==0 & Filtered_RNPxl$C[i]==0
&Filtered_RNPxl$G[i]==0 & Filtered_RNPxl$A[i]==0 )

        {
           Filtered_RNPxl$CrosslinkedNucleotide[i]="Not XL"
        }

      }


#Selecting the verified crosslinks and removing downstream columns

Filtered_RNPxl<-Filtered_RNPxl[which(Filtered_RNPxl$DNAAdduct!="Not XL" ), ]
 Filtered_RNPxl<-Filtered_RNPxl[which(Filtered_RNPxl$CrosslinkedNucleotide!=0), ]
 Filtered_RNPxl=subset(Filtered_RNPxl, select= -c(A,A_FA,G,G_FA,C,C_FA,T))

 #writing output table

write.table(Filtered_RNPxl,
file="/Users/Fanni/Documents/Work/monos/Validated_Monos_RNPxl.csv"
,col.names=TRUE,sep = ",",row.names = FALSE,dec=".")
```

Supplementary Text 4: R script for the validation of the formaldehyde crosslinked peptide-RNA heteroconjugates

```
#Set directory
setwd("/Users/Fanni/Documents/Work/EC_silica_final")
#Check
getwd()
#installing packages
install.packages("stringr")
install.packages("stringi")
install.packages("protr")
BiocManager::install(version = "3.12")
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager")
BiocManager::install("mzR")

#loading packages
library("stringi")
library("stringr")
library("protr")
library("mzR")

#read in the mzml files
ms<-openMSfile("F_Bazso_021120_071120_EC_XL_rneasy_mod_118_cc_3.mzml")
hd <- header(ms)
hd2 <- hd[hd$msLevel == 2, ]
#zero everything to reset the values
MSFragger <- c(0)
i= 0
j=0

# read in MSfragger table
MSFragger<- read.csv("/Users/Fanni/Documents/Work/EC_silica_final/Ecoli_psms-2.csv",
dec=".", header = TRUE, stringsAsFactors = FALSE)
#Setting up columns which will be used later
MSFragger$MassDiffComposition <- c("Unknown")
MSFragger$MassDiffTheoMass <- c(0)
MSFragger$Accuracy <- c("NaN")

#Search the delta masses for crosslinks under 200 ppm
AllCombinationMasses<- c(255.0855, 256.0884, 279.0968, 280.0993, 295.0918, 296.0942,
335.0519, 336.0547, 357.0474, 358.0500, 359.0631, 360.0657, 375.0580, 376.0606,
560.1268, 561.1296, 561.1108, 562.1137, 584.1380, 585.1407, 585.1221, 586.1248,
600.1330, 601.1356, 601.1170, 602.1197, 608.1493, 609.1518, 624.1442, 625.1467,
640.0931, 641.0960, 640.1391, 641.1417, 641.0772, 642.0801, 664.1044, 665.1071,
680.0993, 681.1020, 688.1156, 689.1182,704.1105,705.1131,665.0884, 666.0911,
681.0833,682.0861,720.1054, 721.1080)
AllCombinationNames<- c("C-HPO3","C-HPO3_+1","A-HPO3", "A-HPO3_+1", "G-
HPO3","G-HPO3_+1","C","C_+1","G-H2O","G-H2O_+1","A","A_+1", "G","G_+1","CC-
HPO3","CC-HPO3_+1","CU-HPO3","CU-HPO3_+1", "AC-HPO3", "AC-HPO3_+1","AU-
HPO3", "AU-HPO3_+1", "CG-HPO3", "CG-HPO3_+1","GU-HPO3", "GU-
HPO3_+1","AA-HPO3","AA-HPO3_+1", "AG-HPO3","AG-HPO3_+1","CC",
```

```
"CC_+1","GG-HPO3","GG-HPO3_+1", "CU", "CU_+1", "AC", "AC_+1", "CG","CG_+1"
,"AA","AA_+1","AG",'AG_+1',"AU", "AU_+1", "GU", "GU_+1","GG","GG_+1" )
for (i in 1:nrow(MSFragger))
  {
     for (j in 1: length (AllCombinationMasses))
       {
          if (abs(((MSFragger$Delta.Mass[i]-
AllCombinationMasses[j])/AllCombinationMasses[j])*1000000) < 200)
           {
               MSFragger$MassDiffComposition[i]= AllCombinationNames[j]
               MSFragger$MassDiffTheoMass[i]= AllCombinationMasses[j]
                MSFragger$Accuracy[i]= abs(((MSFragger$Delta.Mass[i]-
AllCombinationMasses[j])/AllCombinationMasses[j])*1000000)
             }

           }
    }


#extracting the index numbers for MSFragger files
  strsplit(MSFragger[1,1],".", fixed = TRUE)

for (i in 1:nrow(MSFragger))
  {
     MSFragger$indexnumber[i]=strsplit(MSFragger$Spectrum[i], split="[.]")
     MSFragger$indexnumber[i]=MSFragger$indexnumber[[i]][2]
     MSFragger$indexnumber[i]=as.numeric(MSFragger$indexnumber[i])
     #to remove the zero from the first character
       if (regmatches(MSFragger$indexnumber[i], regexpr("\\d",
MSFragger$indexnumber[i]))== 0)
         {
           MSFragger$indexnumber[i]=MSFragger$indexnumber[i][-1]
         }
   }

Filtered_MSFrag<-c(0)
Filtered_MSFrag$New_Comp<-c(0)
Filtered_MSFrag<-MSFragger[which(MSFragger$Accuracy!="NaN"), ]


Filtered_MSFrag$indexnumber<-unlist(Filtered_MSFrag$indexnumber)

#Looking for the respective marker ions based on the RNA adduct composition

Filtered_MSFrag$A<-c(0)
Filtered_MSFrag$A_FA<-c(0)
Filtered_MSFrag$G<-c(0)
Filtered_MSFrag$G_FA<-c(0)
Filtered_MSFrag$U<-c(0)
Filtered_MSFrag$C<-c(0)
Filtered_MSFrag$C_FA<-c(0)

  for (i in 1:nrow(Filtered_MSFrag))
```

```
    {j=1
    while(peaks(ms,Filtered_MSFrag$indexnumber[[i]])[j]<165)
      {
         if ((grepl( "A",Filtered_MSFrag$MassDiffComposition[i], fixed = TRUE)))
           {   if(abs(((peaks(ms,Filtered_MSFrag$indexnumber[[i]])[j]-
136.0623)/136.0623)*1000000)<10)
             {

Filtered_MSFrag$A[i]=(peaks(ms,Filtered_MSFrag$indexnumber[[i]])[,2][j]/max(peaks(ms,
Filtered_MSFrag$indexnumber[[i]])[,2]))*100
             }
               else if(abs(((peaks(ms,Filtered_MSFrag$indexnumber[[i]])[j]-
148.0623)/148.0623)*1000000)<10)
                 {

Filtered_MSFrag$A_FA[i]=(peaks(ms,Filtered_MSFrag$indexnumber[[i]])[,2][j]/max(peaks
(ms,Filtered_MSFrag$indexnumber[[i]])[,2]))*100
                 }

           }

           if ((grepl( "G",Filtered_MSFrag$MassDiffComposition[i], fixed = TRUE)))
             {   if(abs(((peaks(ms,Filtered_MSFrag$indexnumber[[i]])[j]-
152.0572)/152.0572)*1000000)<10)
               {

Filtered_MSFrag$G[i]=(peaks(ms,Filtered_MSFrag$indexnumber[[i]])[,2][j]/max(peaks(ms,
Filtered_MSFrag$indexnumber[[i]])[,2]))*100
               }
                 else if(abs(((peaks(ms,Filtered_MSFrag$indexnumber[[i]])[j]-
164.0572)/164.0572)*1000000)<10)
                   {

Filtered_MSFrag$G_FA[i]=(peaks(ms,Filtered_MSFrag$indexnumber[[i]])[,2][j]/max(peaks
(ms,Filtered_MSFrag$indexnumber[[i]])[,2]))*100
                   }
               }



           if ((grepl( "C",Filtered_MSFrag$MassDiffComposition[i], fixed = TRUE)))
             {   if(abs(((peaks(ms,Filtered_MSFrag$indexnumber[[i]])[j]-
112.0511)/112.0511)*1000000)<10)
               {

Filtered_MSFrag$C[i]=(peaks(ms,Filtered_MSFrag$indexnumber[[i]])[,2][j]/max(peaks(ms,
Filtered_MSFrag$indexnumber[[i]])[,2]))*100
               }
                 else if(abs(((peaks(ms,Filtered_MSFrag$indexnumber[[i]])[j]-
124.0511)/124.0511)*1000000)<10)
                   {
```

```
Filtered_MSFrag$C_FA[i]=(peaks(ms,Filtered_MSFrag$indexnumber[[i]])[,2][j]/max(peaks
(ms,Filtered_MSFrag$indexnumber[[i]])[,2]))*100
                  }
               }

           if ((grepl( "U",Filtered_MSFrag$MassDiffComposition[i], fixed = TRUE)))
             {    if(abs(((peaks(ms,Filtered_MSFrag$indexnumber[[i]])[j]-
113.0351)/113.0351)*1000000)<10)
               {

Filtered_MSFrag$U[i]=(peaks(ms,Filtered_MSFrag$indexnumber[[i]])[,2][j]/max(peaks(ms,
Filtered_MSFrag$indexnumber[[i]])[,2]))*100
                 }
               }

           j=j+1
         }
    }



#Calculate crosslink accuracy

Filtered_MSFrag$RNAAdduct=c(0)
Filtered_MSFrag$CrosslinkedNucleotide=c(0)

  for (i in 1:nrow(Filtered_MSFrag))
    {

        if (grepl( "_+1",Filtered_MSFrag$MassDiffComposition[i], fixed = TRUE))
          {
             Filtered_MSFrag$CrossLinkAccuracy[i]= ((Filtered_MSFrag$Observed.Mass[i]-
(Filtered_MSFrag$Calculated.Peptide.Mass[i]+(Filtered_MSFrag$MassDiffTheoMass[i]-
1.0028)))/(Filtered_MSFrag$Calculated.Peptide.Mass[i]+Filtered_MSFrag$MassDiffTheoM
ass[i]))*1000000

          }

         else (Filtered_MSFrag$CrossLinkAccuracy[i]= ((Filtered_MSFrag$Observed.Mass[i]-
(Filtered_MSFrag$Calculated.Peptide.Mass[i]+Filtered_MSFrag$MassDiffTheoMass[i]))/(Fi
ltered_MSFrag$Calculated.Peptide.Mass[i]+Filtered_MSFrag$MassDiffTheoMass[i]))*1000
000)
      }


  #Check if the respective marker ions are present based on the crosslinked nucleotide adduct
composition

  for (i in 1:nrow(Filtered_MSFrag))
    {
```

```
        if (grepl("A", Filtered_MSFrag$MassDiffComposition[i], fixed=TRUE) &
Filtered_MSFrag$A[i]==0 & Filtered_MSFrag$A_FA[i]==0)

        {
           Filtered_MSFrag$RNAAdduct[i]= "Not XL"
         }
       if (grepl("G", Filtered_MSFrag$MassDiffComposition[i], fixed= TRUE) &
Filtered_MSFrag$G[i]==0 & Filtered_MSFrag$G_FA[i]==0)

        {
          Filtered_MSFrag$RNAAdduct[i]= "Not XL"
        }
      if (grepl("C", Filtered_MSFrag$MassDiffComposition[i], fixed=TRUE)&
Filtered_MSFrag$C[i]==0 & Filtered_MSFrag$C_FA[i]==0)

        {
           Filtered_MSFrag$RNAAdduct[i]= "Not XL"
        }
       if (grepl("U", Filtered_MSFrag$MassDiffComposition[i], fixed=TRUE)&
Filtered_MSFrag$U[i]==0)

        {
           Filtered_MSFrag$RNAAdduct[i]= "Not XL"
        }
     }

 #Checking for crosslinked nucleotide removing +1 from the RNAAdduct names when
isotoperror is present
 for (i in 1:nrow(Filtered_MSFrag))
    {

        if (Filtered_MSFrag$A_FA[i]!=0 &Filtered_MSFrag$G_FA[i]==0 &
Filtered_MSFrag$C_FA[i]==0)
         {
           Filtered_MSFrag$CrosslinkedNucleotide[i]="A"
         }

        else if (Filtered_MSFrag$G_FA[i]!=0 & Filtered_MSFrag$A_FA[i]==0 &
Filtered_MSFrag$C_FA[i]==0)
         {
           Filtered_MSFrag$CrosslinkedNucleotide[i]="G"
         }

        else if ((Filtered_MSFrag$C_FA[i]!=0 &Filtered_MSFrag$G_FA[i]==0 &
Filtered_MSFrag$A_FA[i]==0)||(Filtered_MSFrag$C[i]!=0 &Filtered_MSFrag$G_FA[i]==0
& Filtered_MSFrag$A_FA[i]==0))
        {
           Filtered_MSFrag$CrosslinkedNucleotide[i]="C"
         }

        else if (Filtered_MSFrag$C_FA[i]!=0 &Filtered_MSFrag$A_FA[i]!=0
&Filtered_MSFrag$G_FA[i]==0)
         {
```

```
            Filtered_MSFrag$CrosslinkedNucleotide[i]="A,C"
          }
       else if (Filtered_MSFrag$A_FA[i]==0 &  Filtered_MSFrag$C_FA[i]!=0 &
Filtered_MSFrag$G_FA[i]!=0 )
         {
            Filtered_MSFrag$CrosslinkedNucleotide[i]="G,C"
         }
       else if (Filtered_MSFrag$C_FA[i]==0 &Filtered_MSFrag$G_FA[i]!=0
&Filtered_MSFrag$A_FA[i]!=0)
         {
            Filtered_MSFrag$CrosslinkedNucleotide[i]="A,G"
         }

         else if (Filtered_MSFrag$C_FA[i]==0 &Filtered_MSFrag$G_FA[i]==0 &
Filtered_MSFrag$A_FA[i]==0 & Filtered_MSFrag$U[i]==0 & Filtered_MSFrag$C[i]==0
&Filtered_MSFrag$G[i]==0 & Filtered_MSFrag$A[i]==0 )

         {
            Filtered_MSFrag$CrosslinkedNucleotide[i]="Not XL"
         }

         if (Filtered_MSFrag$RNAAdduct[i]!="Not XL")

         {
            if (grepl( "_+1",Filtered_MSFrag$MassDiffComposition[i], fixed = TRUE))
              {

Filtered_MSFrag$RNAAdduct[i]=str_split(Filtered_MSFrag$MassDiffComposition[i],
"_")[[1]][1]    }
            else (Filtered_MSFrag$RNAAdduct[i]=Filtered_MSFrag$MassDiffComposition[i])

            }

       }

  #removing contaminants if they are present

  Filtered_MSFrag<-Filtered_MSFrag[which(nchar(Filtered_MSFrag$Protein)>8), ]

  #Find crosslinked peptide position in protein sequences

  Filtered_MSFrag$Peptide.Position_start=c(0)
  Filtered_MSFrag$Peptide.Position_end=c(0)

  for (i in 1:nrow(Filtered_MSFrag))
  {
    Filtered_MSFrag$Peptide.Position_start[i]= stri_locate(pattern =
Filtered_MSFrag$Peptide[i], getUniProt(Filtered_MSFrag$Protein.ID[i]), fixed =
TRUE)[[1]][1]
    Filtered_MSFrag$Peptide.Position_end[i]=Filtered_MSFrag$Peptide.Position_start[i] +
Filtered_MSFrag$Peptide.Length[i] -1
  }
```

```
 #Selecting the verified crosslinks
 Filtered_MSFrag<-Filtered_MSFrag[which(Filtered_MSFrag$RNAAdduct!="Not XL"), ]
 Filtered_MSFrag<-Filtered_MSFrag[which(Filtered_MSFrag$CrosslinkedNucleotide!=0), ]
 Filtered_MSFrag_final<-Filtered_MSFrag[which(Filtered_MSFrag$CrossLinkAccuracy >-
10), ]
 Filtered_MSFrag_final<-
Filtered_MSFrag_final[which(Filtered_MSFrag_final$CrossLinkAccuracy<10 ), ]
 Filtered_MSFrag_final<-rbind(Filtered_MSFrag_final,
Filtered_MSFrag[which(Filtered_MSFrag$CrossLinkAccuracy >30 ), ] )


 #Removing downstream analysis colums

 Filtered_MSFrag_final=subset(Filtered_MSFrag_final, select= -
c(A,A_FA,G,G_FA,C,C_FA,U,Accuracy,MassDiffTheoMass,MassDiffComposition))

#Exporting results
write.table(Filtered_MSFrag_final,
file="/Users/Fanni/Documents/Work/Reprocessed/Ecoli_ver2_2.csv" ,col.names=TRUE,sep
= ",",row.names = FALSE,dec=".")
```

BIBLIOGRAPHY

1. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64 (1989).
2. Kero, F. A., Pedder, R. E. & Yost, R. A. Quadrupole mass analyzers: theoretical and practical considerations. in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* (American Cancer Society, 2005). doi:https://doi.org/10.1002/047001153X.g301319.
3. March, R. E. An Introduction to Quadrupole Ion Trap Mass Spectrometry. *Journal of Mass Spectrometry* **32**, 351–369 (1997).
4. Douglas, D. J., Frank, A. J. & Mao, D. Linear ion traps in mass spectrometry. *Mass Spectrometry Reviews* **24**, 1–29 (2005).
5. Senko, M. W. *et al.* Novel Parallelized Quadrupole/Linear Ion Trap/Orbitrap Tribrid Mass Spectrometer Improving Proteome Coverage and Peptide Identification Rates. *Analytical Chemistry* **85**, 11710–11714 (2013).
6. Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Analytical Chemistry* **72**, 1156–1162 (2000).
7. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry* **40**, 430–443 (2005).
8. Corradini, D., Eksteen, E., Eksteen, R., Schoenmakers, P. & Miller, N. *Handbook of HPLC*. (CRC Press, 2011).
9. Levsen, K. & Schwarz, H. Gas-phase chemistry of collisionally activated ions. *Mass Spectrometry Reviews* **2**, 77–148 (1983).
10. Olsen, J. v *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods* **4**, 709–712 (2007).
11. Silivra, O. A., Kjeldsen, F., Ivonin, I. A. & Zubarev, R. A. Electron capture dissociation of polypeptides in a three-dimensional quadrupole ion trap: Implementation and first results. *Journal of the American Society for Mass Spectrometry* **16**, 22–27 (2005).
12. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9528 (2004).
13. Zhang, G., Annan, R. S., Carr, S. A. & Neubert, T. A. Overview of Peptide and Protein Analysis by Mass Spectrometry. *Current Protocols in Protein Science* **62**, 16.1.1-16.1.30 (2010).
14. Chhabil Dass. *Fundamentals of Contemporary Mass Spectrometry*. vol. 16 (John Wiley & Sons, 2007).
15. Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry* **35**, 1399–1406 (2000).
16. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
17. Olsen, J. v, Ong, S.-E. & Mann, M. Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues*. *Molecular & Cellular Proteomics* **3**, 608–614 (2004).
18. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**, 976–989 (1994).
19. Geer, L. Y. *et al.* Open Mass Spectrometry Search Algorithm. *Journal of Proteome Research* **3**, 958–964 (2004).

20. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).
21. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research* **7**, 29–34 (2008).
22. Zhang, B., Chambers, M. C. & Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *Journal of proteome research* **6**, 3549–3557 (2007).
23. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Analytical Chemistry* **75**, 4646–4658 (2003).
24. Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & cellular proteomics : MCP* **8**, 2405–2417 (2009).
25. The, M., Tasnim, A. & Käll, L. How to talk about protein-level false discovery rates in shotgun proteomics. *Proteomics* **16**, 2461–2469 (2016).
26. Gerber, A. P. RNA-Centric Approaches to Profile the RNA-Protein Interaction Landscape on Selected RNAs. *Non-coding RNA* **7**, 11 (2021).
27. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular cell* **78**, 9–29 (2020).
28. Balcerak, A., Trebinska-Stryjewska, A., Konopinski, R., Wakula, M. & Grzybowska, E. A. RNA-protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity. *Open biology* **9**, 190096 (2019).
29. Albihlal, W. S. & Gerber, A. P. Unconventional RNA-binding proteins: an uncharted zone in RNA biology. *FEBS Letters* **592**, 2917–2931 (2018).
30. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).
31. Asencio, C., Chatterjee, A. & Hentze, M. W. Silica-based solid-phase extraction of cross-linked nucleic acid–bound proteins. *Life Science Alliance* **1**, e201800088 (2018).
32. Urdaneta, E. C. *et al.* Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. *Nature Communications* **10**, 990 (2019).
33. Queiroz, R. M. L. *et al.* Comprehensive identification of RNA–protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nature Biotechnology* **37**, 169–178 (2019).
34. Trendel, J. *et al.* The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell* **176**, 391-403.e19 (2019).
35. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biology* **1**, reviews001.1 (2000).
36. Kalodimos, C. G. *et al.* Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science* **305**, 386 (2004).
37. Orlando, V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends in Biochemical Sciences* **25**, 99–104 (2000).
38. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research* **17**, 877–885 (2007).
39. Kustatscher, G., Wills, K. L. H., Furlan, C. & Rappsilber, J. Chromatin enrichment for proteomics. *Nature protocols* **9**, 2090–2099 (2014).
40. Kramer, K. *et al.* Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature Methods* **11**, 1064–1070 (2014).
41. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods* **13**, 741–748 (2016).

42. Peil, L. *et al.* Identification of RNA-associated peptides, iRAP, defines precise sites of protein-RNA interaction. (2018) doi:10.1101/456111.
43. Panhale, A. *et al.* CAPRI enables comparison of evolutionarily conserved RNA interacting regions. *Nature communications* **10**, 2682 (2019).
44. Zhang, J. *et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & cellular proteomics : MCP* **11**, M111.010587-M111.010587 (2012).
45. Michaelson-Richie, E. D. *et al.* DNA−Protein Cross-Linking by 1,2,3,4-Diepoxybutane. *Journal of Proteome Research* **9**, 4356–4367 (2010).
46. Park, S. *et al.* Interstrand and Intrastrand DNA−DNA Cross-Linking by 1,2,3,4-Diepoxybutane: Role of Stereochemistry. *Journal of the American Chemical Society* **127**, 14355–14365 (2005).
47. Tretyakova, N. Y., Groehler, A. & Ji, S. DNA–Protein Cross-Links: Formation, Structural Identities, and Biological Outcomes. *Accounts of Chemical Research* **48**, 1631–1644 (2015).
48. Park, S., Hodge, J., Anderson, C. & Tretyakova, N. Guanine−Adenine DNA Cross-Linking by 1,2,3,4-Diepoxybutane: Potential Basis for Biological Activity. *Chemical Research in Toxicology* **17**, 1638–1651 (2004).
49. La, D. K., Upton, P. B. & Swenberg, J. A. 14.05 - Carcinogenic Alkylating Agents*. in *Comprehensive Toxicology (Second Edition)* (ed. McQueen, C. A.) 63–83 (Elsevier, 2010). doi:https://doi.org/10.1016/B978-0-08-046884-6.01405-6.
50. Tretyakova, N., Sangaiah, R., Yen, T.-Y., Gold, A. & Swenberg, J. A. Adenine Adducts with Diepoxybutane: Isolation and Analysis in Exposed Calf Thymus DNA. *Chemical Research in Toxicology* **10**, 1171–1179 (1997).
51. Hoffman, E. A., Frey, B. L., Smith, L. M. & Auble, D. T. Formaldehyde crosslinking: a tool for the study of chromatin complexes. *The Journal of biological chemistry* **290**, 26404–26411 (2015).
52. Werner, M., Chott, A., Fabiano, A. & Battifora, H. Effect of Formalin Tissue Fixation and Processing on Immunohistochemistry. *The American Journal of Surgical Pathology* **24**, (2000).
53. Collas, P. The Current State of Chromatin Immunoprecipitation. *Molecular Biotechnology* **45**, 87–100 (2010).
54. Kennedy-Darling, J. & Smith, L. M. Measuring the Formaldehyde Protein–DNA Cross-Link Reversal Rate. *Analytical Chemistry* **86**, 5678–5681 (2014).
55. Delrue, I., Verzele, D., Madder, A. & Nauwynck, H. J. Inactivated virus vaccines from chemistry to prophylaxis: merits, risks and challenges. *Expert Review of Vaccines* **11**, 695–719 (2012).
56. Lu, K. *et al.* Structural Characterization of Formaldehyde-Induced Cross-Links Between Amino Acids and Deoxynucleosides and Their Oligomers. *Journal of the American Chemical Society* **132**, 3388–3399 (2010).
57. Stützer, A. *et al.* Analysis of protein-DNA interactions in chromatin by UV induced cross-linking and mass spectrometry. *Nature Communications* **11**, 5250 (2020).
58. Reim, A. *et al.* Atomic-resolution mapping of transcription factor-DNA interactions by femtosecond laser crosslinking and mass spectrometry. *Nature Communications* **11**, 3019 (2020).
59. Urdaneta, E. C. & Beckmann, B. M. Fast and unbiased purification of RNA-protein complexes after UV cross-linking. *Methods* **178**, 72–82 (2020).
60. van Ende, R., Balzarini, S. & Geuten, K. Single and Combined Methods to Specifically or Bulk-Purify RNA-Protein Complexes. *Biomolecules* **10**, 1160 (2020).
61. Shchepachev, V. *et al.* Defining the RNA interactome by total RNA-associated protein purification. *Molecular Systems Biology* **15**, e8689 (2019).

62. Bae, J. W., Kwon, S. C., Na, Y., Kim, V. N. & Kim, J.-S. Chemical RNA digestion enables robust RNA-binding site mapping at single amino acid resolution. *Nature Structural & Molecular Biology* **27**, 678–682 (2020).
63. Sharma, K. *et al.* Analysis of protein–RNA interactions in CRISPR proteins and effector complexes by UV-induced cross-linking and mass spectrometry. *Methods* **89**, 138–148 (2015).
64. Aleksandar Chernev. Identification of peptide-RNA heteroconjugates by mass spectrometry. (2020).
65. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367–1372 (2008).
66. Kong, A. T., Leprevost, F. v, Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods* **14**, 513–520 (2017).
67. da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nature Methods* **17**, 869–870 (2020).
68. Geiszler, D. J. *et al.* PTM-Shepherd: analysis and summarization of post-translational and chemical modifications from open search results. *bioRxiv* 2020.07.08.192583 (2020) doi:10.1101/2020.07.08.192583.
69. Yu, F. *et al.* Identification of modified peptides using localization-aware open search. *Nature Communications* **11**, 4065 (2020).
70. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Analytical Chemistry* **74**, 5383–5392 (2002).
71. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).
72. Pedrioli, P. G. A. *et al.* A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology* **22**, 1459–1466 (2004).
73. Martens, L. *et al.* mzML—a Community Standard for Mass Spectrometry Data*. *Molecular & Cellular Proteomics* **10**, R110.000133 (2011).
74. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30**, 918–920 (2012).
75. Hadley Wickham. stringr: Simple, Consistent Wrappers for Common String Operations. *R package version 1.4.0 https://CRAN.R-project.org/package=stringr* (2019).
76. Marek Gagolewski. stringi: Fast and portable character string processing in R. *https://stringi.gagolewski.com/ R package version 1.7.3* (2021).
77. Xiao, N., Cao, D.-S., Zhu, M.-F. & Xu, Q.-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **31**, 1857–1859 (2015).
78. Taus, T. *et al.* Universal and Confident Phosphorylation Site Localization Using phosphoRS. *Journal of Proteome Research* **10**, 5354–5362 (2011).
79. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607–D613 (2019).
80. Mi, H. *et al.* PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research* **49**, D394–D403 (2021).
81. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research* **38**, D355–D360 (2010).

82. Sturm, M. & Kohlbacher, O. TOPPView: An Open-Source Viewer for Mass Spectrometry Data. *Journal of Proteome Research* **8**, 3760–3763 (2009).
83. Bednar, J. *et al.* Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proceedings of the National Academy of Sciences* **95**, 14173 (1998).
84. Loeber, R., Rajesh, M., Fang, Q., Pegg, A. E. & Tretyakova, N. Cross-Linking of the Human DNA Repair Protein O6-Alkylguanine DNA Alkyltransferase to DNA in the Presence of 1,2,3,4-Diepoxybutane. *Chemical Research in Toxicology* **19**, 645–654 (2006).
85. Zhou, B.-R. *et al.* Distinct Structures and Dynamics of Chromatosomes with Different Human Linker Histone Isoforms. *Molecular Cell* **81**, 166-182.e6 (2021).
86. Zhou, B.-R. *et al.* Structural Mechanisms of Nucleosome Recognition by Linker Histones. *Molecular Cell* **59**, 628–638 (2015).
87. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature Protocols* **14**, 68–85 (2019).
88. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution††We dedicate this paper to the memory of Max Perutz who was particularly inspirational and supportive to T.J.R. in the early stages of this study. *Journal of Molecular Biology* **319**, 1097–1113 (2002).
89. Bilokapic, S., Strauss, M. & Halic, M. Histone octamer rearranges to adapt to DNA unwrapping. *Nature structural & molecular biology* **25**, 101–108 (2018).
90. Kato, D. *et al.* Crystal structure of the overlapping dinucleosome composed of hexasome and octasome. *Science* **356**, 205 (2017).
91. Mahajan, M. C., Narlikar, G. J., Boyapaty, G., Kingston, R. E. & Weissman, S. M. Heterogeneous nuclear ribonucleoprotein C1/C2, MeCP1, and SWI/SNF form a chromatin remodeling complex at the beta-globin locus control region. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15012–15017 (2005).
92. Camozzi, D. *et al.* Diverse lamin-dependent mechanisms interact to control chromatin dynamics. Focus on laminopathies. *Nucleus (Austin, Tex.)* **5**, 427–440 (2014).
93. Olson, E. N. & Nordheim, A. Linking actin dynamics and gene transcription to drive cellular motile functions. *Nature reviews. Molecular cell biology* **11**, 353–365 (2010).
94. Herold, M. & Nierhaus, K. H. Incorporation of six additional proteins to complete the assembly map of the 50 S subunit from Escherichia coli ribosomes. *Journal of Biological Chemistry* **262**, 8826–8833 (1987).
95. Diedrich, G. *et al.* Ribosomal protein L2 is involved in the association of the ribosomal subunits, tRNA binding to A and P sites and peptidyl transfer. *The EMBO journal* **19**, 5241–5250 (2000).
96. Fischer, N. *et al.* The pathway to GTPase activation of elongation factor SelB on the ribosome. *Nature* **540**, 80–85 (2016).
97. Castello, A., Hentze, M. W. & Preiss, T. Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. *Trends in endocrinology and metabolism: TEM* **26**, 746–757 (2015).
98. Yang, C.-C. & Nash, H. A. The interaction of E. coli IHF protein with its specific binding sites. *Cell* **57**, 869–880 (1989).
99. Steinman, H. M., Weinstein, L. & Brenowitz, M. The manganese superoxide dismutase of Escherichia coli K-12 associates with DNA. *Journal of Biological Chemistry* **269**, 28629–28634 (1994).
100. Consortium, G. O. The Gene Ontology project in 2008. *Nucleic acids research* **36**, D440–D444 (2008).

101. Nagy, E. & Rigby, W. F. C. Glyceraldehyde-3-phosphate Dehydrogenase Selectively Binds AU-rich RNA in the NAD^+-binding Region (Rossmann Fold) (∗). *Journal of Biological Chemistry* **270**, 2755–2763 (1995).
102. Tristan, C., Shahani, N., Sedlak, T. W. & Sawa, A. The diverse functions of GAPDH: views from different subcellular compartments. *Cellular signalling* **23**, 317–323 (2011).
103. Volz, K. The functional duality of iron regulatory protein 1. *Current opinion in structural biology* **18**, 106–111 (2008).
104. Benjamin, J.-A. M. & Massé, E. The iron-sensing aconitase B binds its own mRNA to prevent sRNA-induced mRNA cleavage. *Nucleic acids research* **42**, 10023–10036 (2014).
105. Liu, M. Y. *et al.* The RNA Molecule CsrB Binds to the Global Regulatory Protein CsrA and Antagonizes Its Activity in Escherichia coli *. *Journal of Biological Chemistry* **272**, 17502–17510 (1997).
106. Kim, B. H. & Gadd, G. M. *Bacterial Physiology and Metabolism*. (Cambridge University Press, 2008).
107. Kaledhonkar, S. *et al.* Late steps in bacterial translation initiation visualized using time-resolved cryo-EM. *Nature* **570**, 400–404 (2019).
108. Fischer, N. *et al.* Structure of the E. coli ribosome–EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. *Nature* **520**, 567–570 (2015).
109. Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burley, S. K. Recognition of Polyadenylate RNA by the Poly(A)-Binding Protein. *Cell* **98**, 835–845 (1999).
110. Schäfer, I. B. *et al.* Molecular Basis for poly(A) RNP Architecture and Recognition by the Pan2-Pan3 Deadenylase. *Cell* **177**, 1619-1631.e21 (2019).
111. Tayri-Wilk, T. *et al.* Mass spectrometry reveals the chemistry of formaldehyde cross-linking in structured proteins. *Nature communications* **11**, 3128 (2020).
112. Guan, Z., Yates, N. A. & Bakhtiar, R. Detection and characterization of methionine oxidation in peptides by collision-induced dissociation and electron capture dissociation. *Journal of the American Society for Mass Spectrometry* **14**, 605–613 (2003).
113. Khoury, G. A., Baliban, R. C. & Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.* **1**, (2011).
114. Linster, E. & Wirtz, M. N-terminal acetylation: an essential protein modification emerges as an important regulator of stress responses. *Journal of Experimental Botany* **69**, 4555–4568 (2018).
115. Holmqvist, E. & Vogel, J. RNA-binding proteins in bacteria. *Nature Reviews Microbiology* **16**, 601–615 (2018).
116. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology* **19**, 327–341 (2018).
117. Kim, T. H. & Dekker, J. Formaldehyde Cross-Linking. *Cold Spring Harbor Protocols* **2018**, (2018).
118. Gavrilov, A., Razin, S. v & Cavalli, G. In vivo formaldehyde cross-linking: it is time for black box analysis. *Briefings in Functional Genomics* **14**, 163–165 (2015).
119. Innis, M. A., Myambo, K. B., Gelfand, D. H. & Brow, M. A. DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 9436–9440 (1988).
120. Yoshida, H. [2] - The Ribonuclease T1 Family. in *Methods in Enzymology* (ed. Nicholson, A. W.) vol. 341 28–41 (Academic Press, 2001).
121. Ishihama, Y. *et al.* Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics* **9**, 102 (2008).

122. Cooperman, B. S., Wooten, T., Traut, R. R. & Romero, D. P. Histidine 229 in protein L2 is apparently essential for 50S peptidyl transferase activity. *Biochemistry and Cell Biology* **73**, 1087–1094 (1995).
123. Cunningham, K. *et al.* SecA protein, a peripheral protein of the Escherichia coli plasma membrane, is essential for the functional binding and translocation of proOmpA. *The EMBO journal* **8**, 955–959 (1989).
124. Duval, V., Nicoloff, H. & Levy, S. B. Combined inactivation of lon and ycgE decreases multidrug susceptibility by reducing the amount of OmpF porin in Escherichia coli. *Antimicrobial agents and chemotherapy* **53**, 4944–4948 (2009).
125. Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research* **43**, D234–D239 (2015).
126. Beckmann, B. M. *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nature communications* **6**, 10127 (2015).

Acknowledgements

First and most of all I would like to say thank you for my supervisor, Prof. Dr. Henning Urlaub for giving me the possibility working in his lab, I have a learnt a lot scientifically and personally during these years. I would like to say thank you for my committee members for Prof. Dr. Vlad Pena and Dr. Juliane Liepe for their guidance during our meetings. I would like to further say thank you for Dr. Alexis C. Faesen, Prof. Dr. Argyris Papantonis and Prof. Dr. Jörg Stülke, for participating in my extended committee. I would like to say thank you for Dr. Alexandra Stützer for the nucleosome and histone samples, for Vivek Susvirkar for the *E. coli* cells and for Thomas Conrad for the HeLa cells.
I would like to say thank you for Prof. Dr. Henning Urlaub, Dr. Aleksandar Chernev and Luisa Welp for reviewing parts of my thesis.

I would like to say thank you for all members of the bioanalytical mass spectrometry group for all kinds of help I have gotten throughout the years. I would like to say thank you for the best office mates: Sofia, Ivan and Ralf. Thank you for the discussions regardless of if they were scientific or just simply fun talks. I am grateful for the discussions in the downstairs office too, thank you guys, it was always the highlight of my day.
My gratitude goes for Judit and Csaba. Without your help I could have never reached this point, and I am so grateful for your scientific guidance and your friendship.
I am grateful for Dr. Kuan-Ting Pan for his scientific advices in the beginning of my PhD. I am in debt to Aleks Chernev for his enormous help at end of my PhD, for reviewing my whole thesis and for all the support I have gotten from him.
My special gratitude goes for Sofia who started out as my lab rotation student and turned into one of my closest friends, thank you for supporting me in the whole PhD process.
I always cherished the walks and talks with Eszti, it was so nice to have a bond with someone far from home.
I am grateful for Tóni for turning up in my life again, just at the right moment, when needed.
I am so grateful for the not intentional pep talks and the constant faith that I can do it.
I would like to say thank you for Tomi, for holding my hand in the beginning of the process.
I could not have done the PhD without Steven. You were and always are the light in the darkness.
I cannot be grateful enough for my sister and for my parents. It is truly a blessing to have you.
I would like to say thank you for everyone who has helped me in any ways during my PhD.

Curriculum Vitae

Personal data

Name: Fanni Laura Bazsó

Date of birth: 06.01.1990

Place of birth: Nagykanizsa, Hungary

Nationality: Hungarian

Education

2017-present **PhD studies** in the University of Göttingen / GGNB program Biomolecules: Structure - Function – Dynamics.
Thesis title: Implementation of chemical protein-nucleic acid cross-linking into mass spectrometric workflows and mass spectrometric database searches

2013-2014 **BSc studies in Mathematics** (3 semesters) Specialization in applied mathematics. Eötvös Loránd University, Budapest, Hungary.

2011-2013 **MSc in Chemistry**, (good), Eötvös Loránd University, Budapest, Hungary.
Thesis title: Improving reproducibility of tandem mass spectra (in Hungarian)

2008-2011 **BSc in Chemistry**, (with distinction), Eötvös Loránd University, Budapest, Hungary.
Thesis title: Investigation of conductive polymers with combined electrochemical methods (in Hungarian)

2004-2008 Batthyány Lajos High School, Nagykanizsa, Hungary

Publications

S. Osman, E. Mohammad, M. Lidschreiber, A. Stützer, F. L. Bazsó, K. C. Maier, H. Urlaub, P. Cramer: The Cdk8 kinase module regulates interaction of the Mediator complex with RNA polymerase II J Biol Chem. (2021) DOI: 10.1016/j.jbc.2021.100734.

F. L. Bazsó, O. Ozohanics, G. Schlosser, K. Ludányi, K. Vékey, L. Drahos: Quantitative comparison of tandem mass spectra obtained on various instruments J. Am. Soc. Mas. Spec. 27(8), pp. 1357-1365 (2016)

M. Bojtár, Z. Szakács, D. Hessz, F. L. Bazsó, M. Kubinyi, I. Bitter: Supramolecular FRET modulation by pseudorotaxane formation of a ditopic stilbazolium dye and carboxylato-pillar[5]arene Dyes Pigm. 133, pp. 415-413. (2016)

G. G. Láng, M. Ujvári, F. Bazsó, S. Vesztergom, F. Ujhelyi: In situ monitoring of the electrochemical degradation of polymer films on metals using the bending beam method and impedance spectroscopy Electrochim. Acta 73: pp. 59-69. (2012)

M. Ujvári, M. Takács, S. Vesztergom, F. Bazsó, F. Újhelyi, G. G. Láng: Monitoring of the electrochemical degradation of PEDOT films on gold using the bending beam method J. Solid State Chem. 15: pp. 2341-2349. (2011)