Investigation of protein–RNA interactions by UV cross-linking and mass spectrometry: methodological improvements toward *in vivo* applications

Dissertation

for the award of the degree
"Doctor rerum naturalium" (Dr. rer. nat.)

Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

submitted by

Katharina Kramer
from Fulda, Germany

Göttingen 2013

Members of the Thesis Committee:

Prof. Dr. Henning Urlaub (Reviewer)	Bioanalytical Mass Spectrometry Group, Max Planck Institute for Biophysical Chemistry, Göttingen; Bioanalytics, Department of Clinical Chemistry, University Medical Center, Georg-August-Universität Göttingen
Prof. Dr. Jörg Stülke (Reviewer)	Department of General Microbiology, Georg-August-Universität Göttingen
Dr. Claudia Höbartner	Nucleic Acid Chemistry Group, Max Planck Institute for Biophysical Chemistry, Göttingen

Date of the oral examination:

30.5.2013

Affidavit

Hereby, I declare that the presented thesis entitled *Investigation of protein–RNA interactions by UV cross-linking and mass spectrometry: methodological improvements toward in vivo applications* was written entirely by myself and that I have only used the sources and materials cited.

Göttingen, 30.4.2013

Katharina Kramer

Summary

Protein–RNA complexes play key roles in a variety of cellular functions such as gene expression and its regulation. Detailed knowledge about the proteins and RNAs involved, as well as their three-dimensional arrangements, is required for complete functional understanding.

One frequently applied method for the investigation of direct protein–RNA interactions is UV induced cross-linking. Upon UV irradiation, covalent bonds are formed between nucleic acid bases and amino acid residues in close spatial proximity. This way, intermolecular interactions are fixed with high selectivity, which allows their exploration by various bioanalytical methods.

Mass spectrometry (MS) is increasingly utilized for the identification of proteins and peptides directly interacting with RNA. Following hydrolysis by RNases and endoproteinase, peptide–oligonucleotide heteroconjugates are enriched with suitable chromatographic methods such as size exclusion, C18 reversed phase, and titanium dioxide. Mass spectrometric analysis then identifies contact sites on a peptide or even amino acid level.

The major challenges for this application are the generally low yield of the UV induced cross-linking reaction and the lack of suitable tools for MS data analysis. In this work, both issues are addressed. The incorporation of the photoreactive base analogue 4-thio-uracil was investigated with a focus on the mass of the cross-linking products. The *E. coli* transcription antitermination complex NusB–S10 bound to a BoxA containing oligonucleotide served as a model system. A novel cross-linking pathway involving net loss of H_2S from 4-thio-uracil was identified.

In addition, a novel approach for automated identification of cross-linked peptides from mass spectrometry data was developed. It is based on the variation of experimentally determined masses by subtraction of the calculated masses of potentially cross-linked oligonucleotides. A subsequent database search of the mass variants by conventional algorithms identifies cross-linked peptides. After feasibility of the approach was established, it was further tested and optimized in the investigation of a model complex for *ASH1* mRNA transport in yeast and interactions of the yeast spliceosomal protein Cwc2 with U6 and U4 small nuclear RNAs. In both systems, several protein regions contacting RNA were identified; in many cases, the cross-linking site could be confined to a single amino acid.

Finally, it was demonstrated that the data analysis approach can be applied for unbiased searches against databases containing the entire yeast proteome. After isolation of capped RNAs bound by the protein Cbp20 under native conditions, ribosomal proteins, proteins with known RNA- or DNA-binding properties, and metabolic enzymes were found to directly interact with RNA. This illustrates the capability of UV induced cross-linking with MS analysis to identify novel RNA-binding proteins and domains. Importantly, the data analysis approach represents a key development toward the applicability of the method to *in vivo* cross-linking approaches.

For my parents...

Contents

List of Figures VI

List of Tables XII

1 Introduction 1

1.1 Protein–RNA complexes 1

1.2 Mass spectrometry in identification of biological macromolecules 4

1.2.1 Tandem mass spectrometry 5

1.2.1.1 Quadrupole time-of-flight (Q-ToF) mass spectrometers 6

1.2.1.2 Orbitrap mass spectrometers with linear ions traps (LTQ Orbitraps) 7

1.2.1.3 Fragmentation of peptides 8

1.2.1.4 Fragmentation of RNA 9

1.2.2 Data analysis in proteomics 10

1.3 UV induced protein–RNA cross-linking 12

1.3.1 Preparation of ribonucleoprotein complexes and UV cross-linking 13

1.3.1.1 Incorporation of substituted nucleotides 14

1.3.2 Sample preparation for mass spectrometry, enrichment and purification strategies 15

1.3.3 Mass spectrometry of peptide–RNA heteroconjugates 18

1.3.3.1 MALDI mass spectrometry 18

1.3.3.2 ESI mass spectrometry 18

1.3.3.3 RNA marker ions 19

1.3.3.4 Fragmentation modes: beam-type versus ion trap CID 19

1.3.4 Cross-link identification from mass spectrometry data 20

1.3.5 Application of UV induced cross-linking in combination with mass spectrometry 22

1.4 Objectives 25

2 Materials and Methods 27

2.1 Materials 27

2.1.1 Chemicals and solvents 27

2.1.2 Commercial buffers and solutions 28

2.1.3 Buffers 28

2.1.4 Enzymes and enzyme inhibitors 29

2.1.5 Proteins, peptides and (oligo)nucleotides 29

2.1.6 Antibodies 29

2.1.7 Other materials 30

2.1.8 Instruments and laboratory equipment 31

2.2 Methods . . . 32
  2.2.1 Media and plates for cell cultures . . . 32
    2.2.1.1 LB medium . . . 32
    2.2.1.2 YPD medium . . . 32
    2.2.1.3 YMM dropout medium . . . 32
    2.2.1.4 Auto-inducing medium . . . 32
  2.2.2 Standard molecular biology methods . . . 34
    2.2.2.1 Agarose gel electrophoresis of DNA fragments . . . 34
    2.2.2.2 PCI extraction . . . 34
    2.2.2.3 Ethanol precipitation . . . 34
    2.2.2.4 5’ labeling of RNA . . . 34
  2.2.3 Standard protein biochemical methods . . . 35
    2.2.3.1 Determination of protein concentration . . . 35
    2.2.3.2 Denaturing polyacrylamide gel electrophoresis (SDS-PAGE) . . . 35
    2.2.3.3 Colloidal Coomassie staining . . . 36
    2.2.3.4 Western blotting for immunodetection of proteins . . . 36
  2.2.4 Expression and isolation of the NusB–S10 protein complex . . . 37
  2.2.5 Introduction of a C-terminal TAP tag to the yeast protein Cbp20 by homologous recombination . . . 38
    2.2.5.1 Generation of DNA . . . 38
    2.2.5.2 Transformation . . . 39
    2.2.5.3 Yeast colony PCR . . . 39
    2.2.5.4 Confirmation of TAP tag inclusion by Western blot . . . 40
    2.2.5.5 Confirmation of TAP tag inclusion by sequencing . . . 40
  2.2.6 Yeast cell culture and extract preparation . . . 41
  2.2.7 TAP tag purification . . . 41
  2.2.8 UV induced protein–RNA cross-linking . . . 42
    2.2.8.1 Cross-linking of labeled RNA and visualization of cross-linking products by SDS-PAGE . . . 42
    2.2.8.2 Standard protocol for cross-linking and enrichment of cross-linked heteroconjugates for LC-ESI-MS/MS . . . 42
    2.2.8.3 Cross-linking of NusB–S10 to 4SU-substituted RNA . . . 43
    2.2.8.4 Cross-linking of the ASH1 complex . . . 43
    2.2.8.5 Cross-linking of Cwc2 . . . 44
    2.2.8.6 Cross-linking of protein–RNA complexes after TAP tag purification 44
  2.2.9 LC-ESI-MS/MS analysis . . . 45
    2.2.9.1 Nano-LC separation (Agilent) . . . 45
    2.2.9.2 Nano-LC separation (EASY-nLC II) . . . 45
    2.2.9.3 ESI-MS/MS analysis on the Q-ToF Ultima . . . 46
    2.2.9.4 ESI-MS/MS analysis on the LTQ Orbitrap Velos . . . 46
    2.2.9.5 ESI-MS/MS analysis on the Q Exactive . . . 46
  2.2.10 MS data analysis . . . 47
    2.2.10.1 Peptide identification with Mascot . . . 47
    2.2.10.2 Identification of cross-links with 94 Da adducts using Mascot . . . 47

2.2.10.3 Online tools for calculation of monoisotopic masses 48
2.2.10.4 Identification of cross-links by manual spectra interpretation 48
2.2.10.5 Identification of cross-linked peptides after precursor variant generation by a perl script 48
2.2.10.6 Identification of cross-linked peptides with OpenMS and OMSSA . . 49
2.2.10.7 Validation of cross-links 53

3 Results 55

3.1 Cross-linking products of 4-thio-uracil and a novel approach for automated data analysis 56
3.1.1 Influence of 4-thio-uracil on the cross-linking yield of the NusB–S10-complex . 56
3.1.2 Development of a novel approach for automated data analysis 58
3.1.2.1 Anticipated RNA combinations and modifications 58
3.1.2.2 Cross-linked RNA and standard database search 59
3.1.2.3 The precursor variant approach 60
3.1.2.4 Implementation of the precursor variant approach 61
3.1.3 Cross-linking products of 4-thio-uracil in the NusB–S10–BoxA RNA complex 63
3.1.3.1 Additive cross-linking product of 4-thio-uracil 63
3.1.3.2 Identification of a novel, 4-thio-uracil specific cross-linking product . 64
3.1.3.3 Observation of peptides with a 258 Da adduct 67
3.1.3.4 Feasibility of the precursor variant approach 69
3.1.4 Comparison of obtained cross-linking results to cross-linking with unsubstituted RNA 70
3.2 Instrumental and data analysis improvements and their implications for cross-link identification 73
3.2.1 LTQ Orbitrap Velos mass spectrometer 73
3.2.2 Integration of the precursor variant approach into the OpenMS environment . 73
3.2.3 Cross-linking of ASH1 76
3.2.4 ASH1 cross-links identified after LC-ESI-MS/MS measurement on the Q-TOF and data analysis with the perl script 77
3.2.5 ASH1 cross-links identified after LC-ESI-MS/MS measurement on the Orbitrap Velos and data analysis with OpenMS 80
3.2.5.1 Introduction of non-irradiated controls for validation 82
3.2.5.2 4-thio-uracil and native uracil form similar cross-linking products . . 84
3.2.5.3 Significant improvement of cross-link identification after measurement on the Orbitrap Velos and data analysis with OpenMS 85
3.2.6 Summary and functional implications of obtained cross-linking results 87
3.3 Application of the automated data analysis workflow to the spliceosomal protein Cwc2 in complex with RNA 89
3.3.1 Cross-linking of Cwc2 to U6 snRNA 89
3.3.2 Cross-linking of Cwc2 to U4 snRNA and U6 internal stem loop 92
3.4 Application of the automated data analysis workflow to complex systems 95
3.4.1 Isolation of protein–RNA complexes by TAP tag purification 95

3.4.2 Optimization of extract preparation, complex isolation, cross-linking, sample preparation, and LC-ESI-MS/MS . . . 96
3.4.2.1 UV cross-linking . . . 97
3.4.2.2 Preparation of yeast extract . . . 97
3.4.2.3 Sample preparation for LC-ESI-MS/MS analysis . . . 97
3.4.2.4 Gradient for LC-ESI-MS/MS analysis . . . 100
3.4.2.5 MS instrumentation . . . 100
3.4.3 Data analysis and integration of additional filters . . . 100
3.4.3.1 Validation of cross-link candidates: Extracted ion chromatogram and independent database search . . . 100
3.4.3.2 Validation of cross-link candidates: Mass spectra . . . 102
3.4.3.3 Integration of exclusion criteria into automated filtering tools . . . 103
3.4.4 Cross-links identified after TAP tag purification and isolation of cross-linked heteroconjugates . . . 105
3.4.4.1 Cross-links of ribosomal proteins . . . 106
3.4.4.2 Cross-links of RNA-/DNA-binding proteins . . . 109
3.4.4.3 Cross-links of proteins without annotated polynucleotide binding function . . . 110
3.4.5 Summary . . . 114

4 Discussion 115
4.1 Experimental insights and potentials of UV induced protein–RNA cross-linking . . . 116
4.1.1 Reactivity of nucleotides . . . 116
4.1.2 Reactivity of amino acids . . . 117
4.1.3 Influence of the three-dimensional structure . . . 118
4.1.4 Identification of the cross-linking site on the RNA level . . . 118
4.1.5 RNA-binding metabolic enzymes: Rossmann fold domains as RNA-binding motifs . . . 119
4.2 Mass spectrometry and data analysis . . . 120
4.2.1 Instrumentation . . . 120
4.2.2 Development and feasibility of the precursor variant approach . . . 122
4.3 Mass spectrometry and *in vivo* cross-linking . . . 124

Bibliography 127

Acknowledgements 136

Appendices 140

A Masses of identified cross-links 140

B MS/MS fragment spectra of identified cross-links 148
B.1 Annotation of MS/MS spectra of cross-linked peptides . . . 148
B.1.1 Peptide sequence ions . . . 148
B.1.2 RNA marker ions and fragments . . . 149
B.1.3 RNA adducts of peptides and their sequence ions . . . 150

B.2 MS/MS fragment spectra of NusB–S10 peptides observed as adducts with 258 Da . . 152
B.3 MS/MS fragment spectra of cross-links from the ASH1 complexes 153
B.4 MS/MS fragment spectra of cross-links from Cwc2 156
B.5 MS/MS fragment spectra of cross-links from yeast after TAP tag isolation 159
B.5.1 Cross-links of the 40S small ribosomal subunit 159
B.5.2 Cross-links of the 60S large ribosomal subunit 165
B.5.3 Cross-links of ribosome-related proteins 178
B.5.4 Cross-links of polynucleotide-binding proteins 179
B.5.5 Cross-links of proteins without any annotated polynucleotide-binding function 182

C Abbreviations **186**

D Curriculum Vitae **189**

List of Figures

1.1 Structure examples of the RRM, the KH domain, and the ZnF. 2
1.2 Schematic workflow of a typical large scale proteomics experiment. 5
1.3 Schematic representation of a quadrupole time-of-flight (Q-ToF) mass spectrometer. 6
1.4 Schematic representation of a LTQ Orbitrap mass spectrometer. 7
1.5 Nomenclature of peptide fragments resulting from backbone cleavage. 8
1.6 Nomenclature of RNA fragments. 9
1.7 Principles of sequence database searching. 10
1.8 UV induced reaction between uridine and tyrosine. 12
1.9 Strategies for isolation of protein–RNA complexes. 14
1.10 Isolation of cross-linked heteroconjugates from noncross-linked peptides by size exclusion chromatography. 16
1.11 Enrichment of cross-linked heteroconjugates with C18 and titanium dioxide chromatography. 17

2.1 Pipeline for retention time alignment of LC-MS data of control and UV irradiated sample. 50
2.2 ID filter pipeline for removal of MS/MS spectra with confident peptide identification. 51
2.3 Parameters of the RNP^{xl} tool. 52

3.1 Autoradiography of NusB–S10 cross-linked to ^{32}P-labeled BoxA containing RNA with and without 4-thio-uracil. 57
3.2 Schematic, simplified comparison of the MS and MS/MS spectra of the same peptide with and without cross-linked RNA. 59
3.3 Schematic representation of precursor variant generation. 60
3.4 Schematic description of data analysis with a perl script. 61
3.5 MS/MS fragment spectrum of carbamylated NusB peptide SFGAEDSHKFVNGVLDK (S113–K129) cross-linked to [(4SU)(4SU) –HPO$_3$]. 64
3.6 MS/MS fragment spectrum of NusB peptide SFGAEDSHKFVNGVLDK (S113–K129) cross-linked to [4SU –H$_2$S]. 65
3.7 MS/MS fragment spectrum of S10 peptide LIDQATAEIVETAKR (L17–R31) observed as adduct with [4SU –H$_2$S]. 66
3.8 Possible structures of 4SU cross-linking products. 67
3.9 MS/MS fragment spectrum of S10 peptide GPIPLPTR (G38–R44) cross-linked to [4SU –H$_2$S]. 68
3.10 MS/MS fragment spectrum of NusB peptide SDVPYKVAINEAIELAK (S96–K112) observed as adduct with 258 Da. 68

3.11 Comparison of NusB cross-links to unsubstituted and 4-thio-uracil substituted *rrn* BoxA containing RNA oligonucleotides. 70
3.12 Data analysis workflow with the RNP^{xl} tool. 75
3.13 MS/MS fragment spectrum of She2p peptide IGSNLLDLEVVQFAIK (I164–K179) cross-linked to [U]. 78
3.14 MS/MS fragment spectra of She3p-short peptide GPLGSMGNSSNNK (G334–K340) cross-linked to [U] and [AAU –HPO_3]. 79
3.15 MS/MS fragment spectrum of She2p peptide IGSNLLDLEVVQFAIK (I164–K179) cross-linked to [U]. 80
3.16 MS/MS fragment spectrum of She2p peptide YLSSYIHVLNK (Y27–K37) cross-linked to [AAU –HPO_3]. 81
3.17 MS/MS fragment spectrum of She2p peptide LSALDEEFDVVATKWHDK (L223–K240) cross-linked to [U +152]. 82
3.18 XICs and MS/MS spectrum of She2p peptide IGSNLLDLEVVQFAIK (I164–K179) cross-linked to [U –H_2O]. 83
3.19 Possible structures of 4SU and uracil cross-linking products. 84
3.20 Structure of Cwc2. 90
3.21 Protein gel and Western blot of different steps of TAP tag purification. 96
3.22 Size exclusion chromatogram and gel of corresponding fractions. 99
3.23 Effect on spectra numbers after filtering with ID and XIC filter across experiments. . 104
3.24 Effect on spectra numbers after filtering with ID and XIC filter in a single experiment.104
3.25 Schematic workflow for automated filtering of cross-linking data. 105
3.26 Representative examples for cross-links in the ribosome. 108
3.27 Structure of human S-adenosylhomocysteine hydrolyase. 111
3.28 Structure of yeast alcohol dehydrogenase 1. 112
3.29 Structure of yeast glyceraldehyde-3-phosphate dehydrogenases (GAPDH) 3. 113

4.1 Comparison of MS/MS fragment spectra of the unmodified S24-A/-B peptide DAVSVFGFR (D53–R61) and the same peptide cross-linked to uridine. 123

B.1 MS/MS fragment spectrum of NusB peptide IALYELSKR (I87–R95) observed as adduct with 258 Da. 152
B.2 MS/MS fragment spectrum of S10 peptide LVDIVEPTEKTVDALMR (L73–R89) observed as adduct with 258 Da. 152
B.3 MS/MS fragment spectrum of She2p peptide GPLGSMSK (M1–K3) cross-linked to [U –H_2O]. 153
B.4 MS/MS fragment spectrum of She2p peptide FYNDCVLSYNASEFINEGK (F64–K82) cross-linked to [U +152]. 153
B.5 MS/MS fragment spectrum of She2p peptide CVETFDLLNYYLTQSLQK (C106–K123) cross-linked to [U +152]. 154
B.6 MS/MS fragment spectrum of She3p peptide MDQLSKLAK (M130–K138) cross-linked to [U –H_2O]. 154
B.7 MS/MS fragment spectrum of She3p peptide GAVVQTLKK (G283–K291) cross-linked to [U –H_2O]. 154

B.8 MS/MS fragment spectrum of She3p peptide TNVTHNNDPSTSPTISVPPGVTR (T383–R405) cross-linked to [GU]. . . . 155
B.9 MS/MS fragment spectrum of She3p peptide NSSAIEQSCSEK (N139–K150) cross-linked to [U +152]. . . . 155
B.10 MS/MS fragment spectrum of Cwc2 peptide FVSPFALQPQLHSGK (F47–K61) cross-linked to [U –H_2O]. . . . 156
B.11 MS/MS fragment spectrum of Cwc2 peptide CEYLHHIPDEEDIGK (C87–K101) cross-linked to [AU +152]. . . . 156
B.12 MS/MS fragment spectrum of Cwc2 peptide FADYREDMGGIGSFR (F117–R131) cross-linked to [U]. . . . 157
B.13 MS/MS fragment spectrum of Cwc2 peptide TLYVGGIDGALNSK (T136–K149) cross-linked to [U]. . . . 157
B.14 MS/MS fragment spectrum of Cwc2 peptide HLKPAQIESR (H150–R159) cross-linked to [U –H_2O]. . . . 158
B.15 MS/MS fragment spectrum of Cwc2 peptide NCGFVK (N180–K185) cross-linked to [U +152]. . . . 158
B.16 MS/MS fragment spectrum of 40S ribosomal protein S1-A/-B peptide KWQTLIEANVTVK (K116–K128) cross-linked to [U –H_2O]. . . . 159
B.17 MS/MS fragment spectrum of 40S ribosomal protein S3 peptide GLSAVAQAESMKFK (G95–K108) cross-linked to [GU]. . . . 159
B.18 MS/MS fragment spectrum of 40S ribosomal protein S3 peptide GCEVVVSGK (G133–K141) cross-linked to [U +152 –H_2O]. . . . 160
B.19 MS/MS fragment spectrum of 40S ribosomal protein S5 peptide TIAETLAEELINAAK (T189–K203) cross-linked to [GU –H_2O]. . . . 160
B.20 MS/MS fragment spectrum of 40S ribosomal protein S11-A/-B peptide VQVGDIVTVGQCRPISK (V117–K133) cross-linked to [AU –H_2O]. . . . 160
B.21 MS/MS fragment spectrum of 40S ribosomal protein S14-A/-B peptide IYASFNDTFVHVTDLSGK (A: I19–K36, B: I20–K37) cross-linked to [UU]. . . . 161
B.22 MS/MS fragment spectrum of 40S ribosomal protein S14-A/-B peptide ADRDESSPYAAMLAAQDVAAK (A: A50–K70, B: A51–K71) cross-linked to [GU]. . . . 161
B.23 MS/MS fragment spectrum of 40S ribosomal protein S16-A/-B peptide VTGGGHVSQVYAIR (V69–R82) cross-linked to [U –H_2O]. . . . 161
B.24 MS/MS fragment spectrum of 40S ribosomal protein S17-A/-B peptide LCDEIATIQSK (L34–K44) cross-linked to [U +152]. . . . 162
B.25 MS/MS fragment spectrum of 40S ribosomal protein S17-A/-B peptide IAGYTTHLMK (I50–K59) cross-linked to [U –H_2O]. . . . 162
B.26 MS/MS fragment spectrum of 40S ribosomal protein S24-A/-B peptide DAVSVFGFR (D53–R61) cross-linked to [U]. . . . 162
B.27 MS/MS fragment spectrum of 40S ribosomal protein S24-A/-B peptide DKKIFGTGK (D115–K123) cross-linked to [CU]. . . . 163
B.28 MS/MS fragment spectrum of 40S ribosomal protein S29-A peptide VCSSHTGLIR (V23–R32) cross-linked to [U +152 –H_2O]. . . . 163
B.29 MS/MS fragment spectrum of 40S ribosomal protein S29-B peptide VCSSHTGLVR (V23–R32) cross-linked to [U +152 –H_2O]. . . . 163

B.30 MS/MS fragment spectrum of guanine nucleotide-binding protein subunit beta-like protein (Rack1) peptide GQCLATLLGHNDWVSQVR (G138–R155) cross-linked to [U +152 –H_2O]. 164
B.31 MS/MS fragment spectrum of 60S ribosomal protein L1-A/-B peptide SCGVDAMSVDDLKK (S79–K92) cross-linked to [U +152 –H_2O]. 165
B.32 MS/MS fragment spectrum of 60S ribosomal protein L2-A/-B peptide ASGNYVIIGHNPDENK (A129–K145) cross-linked to [U]. 165
B.33 MS/MS fragment spectrum of 60S ribosomal protein L2-A/-B peptide GVAMNPVDHPHGGGNHQHIGK (G201–K221) cross-linked to [AAGU –H_2O]. 166
B.34 MS/MS fragment spectrum of 60S ribosomal protein L3 peptide VACIGAWHPAHVMWSVAR (V249–R266) cross-linked to [U –H_2O]. 166
B.35 MS/MS fragment spectrum of 60S ribosomal protein L4-A/-B peptide SGQGAFGNMCR (S85–R95) cross-linked to [U –H_2O]. 167
B.36 MS/MS fragment spectrum of 60S ribosomal protein L4-A/-B peptide NVPGVETANVASLNLLQLAPGAHLGR (N221–R246) cross-linked to [AU –H_2O]. 167
B.37 MS/MS fragment spectrum of 60S ribosomal protein L4-A/-B peptide IINSSEIQSAIRPAGQATQK (I289–K308) cross-linked to [GU]. 168
B.38 MS/MS fragment spectrum of 60S ribosomal protein L4-A peptide TGTKPAAVFTETLK (T347–K360) cross-linked to [AU –H_2O]. 168
B.39 MS/MS fragment spectrum of 60S ribosomal protein L4-B peptide TGTKPAAVFAETLK (T347–K360) cross-linked to [AU –H_2O]. 168
B.40 MS/MS fragment spectrum of 60S ribosomal protein L5 peptide SYIFGGHVSQYMEELADDDEER (S197–R218) cross-linked to [U]. 169
B.41 MS/MS fragment spectrum of 60S ribosomal protein L6-A/-B peptide LRASLVPGTVLILLAGRFR (L30–R48) cross-linked to [GU –H_2O]. 169
B.42 MS/MS fragment spectrum of 60S ribosomal protein L6-A peptide WYPSEDVAALKK (W9–K20) cross-linked to [AU –H_2O]. 170
B.43 MS/MS fragment spectrum of 60S ribosomal protein L6-B peptide WYPSEDVAAPK (W9–K19) cross-linked to [AU –H_2O]. 170
B.44 MS/MS fragment spectrum of 60S ribosomal protein L6-A peptide HLEDNTLLISGPFK (H57–K70) cross-linked to [U –H_2O]. 171
B.45 MS/MS fragment spectrum of 60S ribosomal protein L6-B peptide HLEDNTLLVTGPFK (H57–K70) cross-linked to [U –H_2O]. 171
B.46 MS/MS fragment spectrum of 60S ribosomal protein L8-A peptide YGLNHVVALIENKK (Y134–K147) cross-linked to [GU –H_2O]. 172
B.47 MS/MS fragment spectrum of 60S ribosomal protein L8-B peptide YGLNHVVSLIENK (Y134–K146) cross-linked to [GU –H_2O]. 172
B.48 MS/MS fragment spectrum of 60S ribosomal protein L16-A/-B peptide LSTSVGWKYEDVVAK (A: L141–K155, B: L140–K154) cross-linked to [U –H_2O]. 172
B.49 MS/MS fragment spectrum of 60S ribosomal protein L16-A peptide AEELNISGEFFRNK (A38–K51) cross-linked to [ACU]. 173
B.50 MS/MS fragment spectrum of 60S ribosomal protein L16-B peptide AEALNISGEFFR (A37–R48) cross-linked to [CU]. 173

B.51 MS/MS fragment spectrum of 60S ribosomal protein L18-A/-B peptide ALFLSK (A51–K56) cross-linked to [AU]. 173
B.52 MS/MS fragment spectrum of 60S ribosomal protein L18-A/-B peptide AGGECITLDQLAVR (A117–R130) cross-linked to [U +152 –H_2O]. 174
B.53 MS/MS fragment spectrum of 60S ribosomal protein L23-A/-B peptide ECADLWPR (E121–R128) cross-linked to [U +152 –H_2O]. 174
B.54 MS/MS fragment spectrum of 60S ribosomal protein L26-B peptide KAYFTAPSSER (K17–R27) cross-linked to [GU]. 174
B.55 MS/MS fragment spectrum of 60S ribosomal protein L28 peptide INMDKYHPGYFGK (I43–K55) cross-linked to [U –H_2O]. 175
B.56 MS/MS fragment spectrum of 60S ribosomal protein L31-A/-B peptide LHGVSFK (L20–K26) cross-linked to [UU –H_2O]. 175
B.57 MS/MS fragment spectrum of 60S ribosomal protein L33-A/-B peptide IAYVYRASKEVR (I49–R60) cross-linked to [AU –H_2O]. 175
B.58 MS/MS fragment spectrum of 60S ribosomal protein L35-A/-B peptide SIACVLTVINEQQR (S50–R63) cross-linked to [U –H_2O]. 176
B.59 MS/MS fragment spectrum of 60S ribosomal protein L37-A peptide FKNGFQTGSASK (F74–K85) cross-linked to [ACU]. 176
B.60 MS/MS fragment spectrum of 60S ribosomal protein L37-B peptide FKNGFQTGSAK (F74–K84) cross-linked to [ACU]. 176
B.61 MS/MS fragment spectrum of Ubiquitin-60S ribosomal protein L40 peptide CGHTNQLRPK (C115–K124) cross-linked to [U +152 –H_2O]. 177
B.62 MS/MS fragment spectrum of 60S ribosomal protein L42-A/-B peptide CKHFELGGEK (C88–K97) cross-linked to [U –H_2O]. 177
B.63 MS/MS fragment spectrum of ribosome biogenesis protein RLP7 peptide GPLAVNIPNKAFK (G149–K161) cross-linked to [CUU]. 178
B.64 MS/MS fragment spectrum of cruciform DNA-recognizing protein 1 peptide IPEAGGLLCGKPPR (I105–R118) cross-linked to [U –H_2O]. 179
B.65 MS/MS fragment spectrum of elongation factor 1-alpha peptide FVPSKPMCVEAFSEYPPLGR (F402–R421) cross-linked to [U +152]. 179
B.66 MS/MS fragment spectrum of nucleolar protein 3 peptide ILNGFAFVEFEEAESAAK (I156–K173) cross-linked to [U –H_2O]. 180
B.67 MS/MS fragment spectrum of nucleolar protein 3 peptide ENSLETTFSSVNTR (E222–R235) cross-linked to [U]. 180
B.68 MS/MS fragment spectrum of nucleolar protein 13 peptide ILFVGNLSFDVTDDLLR (I240–R256) cross-linked to [GU –H_2O]. 180
B.69 MS/MS fragment spectrum of polyadenylate-binding protein peptide YQGVNLFVK (Y319–K327) cross-linked to [U]. 181
B.70 MS/MS fragment spectrum of single-stranded nucleic-acid binding protein peptide SKDTLYINNVPFK (S184–K196) cross-linked to [AC]. 181
B.71 MS/MS fragment spectrum of adenosylhomocysteinase peptide ECINIKPQVDR (E320–R330) cross-linked to [U +152]. 182

B.72 MS/MS fragment spectrum of alcohol dehydrogenase 1/3 peptide YSGVCHTDLHAWHGDWPLPVK (Adh1p: Y40–K60, Adh3p: Y67–K86) cross-linked to [U +152 –H_2O]. 182

B.73 MS/MS fragment spectrum of enolase 1/2 peptide IGLDCASSEFFK (I244–K255) cross-linked to [U +152 –H_2O]. 183

B.74 MS/MS fragment spectrum of glyceraldehyde-3-phosphate dehydrogenase 2/3 peptide ETTYDEIKK (E250–K258) cross-linked to [AU –HPO_3]. 183

B.75 MS/MS fragment spectrum of inorganic pyrophosphatase peptide NCFPHHGYIHNYGAFPQTWEDPNVSHPETK (N83–K112) cross-linked to [U +152 –H_2O]. 183

B.76 MS/MS fragment spectrum of peroxiredoxin TSA1 peptide NGTVLPCNWTPGAATIKPTVEDSK (N165–K188) cross-linked to [U +152]. 184

B.77 MS/MS fragment spectrum of phosphoglycerate kinase peptide YVLEHHPR (Y49–R56) cross-linked to [AU –HPO_3]. 184

B.78 MS/MS fragment spectrum of pyruvate kinase 1 peptide NCTPKPTSTTETVAASAVAAVFEQK (N370–K394) cross-linked to [U +152 –H_2O]. 185

B.79 MS/MS fragment spectrum of pyruvate kinase 1 peptide YRPNCPIILVTR (Y414–R425) cross-linked to [U +152 –H_2O]. 185

List of Tables

3.1 Cross-links of the NusB–S10 complex 69
3.2 Cross-links of the ASH1 complexes identified after analysis on the Q-ToF Ultima. . . 77
3.3 Cross-links of the ASH1 complexes identified after analysis on the Orbitrap Velos. . . 86
3.4 Cross-links of Cwc2 to U6 snRNA 91
3.5 Cross-links of Cwc2 to U6 snRNA, U4 snRNA and U6 internal stem loop 93
3.6 Overview on experiments in yeast 101
3.7 Cross-linked ribosomal proteins 107
3.8 Cross-linked RNA-/DNA-binding proteins 109
3.9 Cross-linked proteins without annotated RNA-binding function 111

A.1 Overview of cross-links of NusB and S10 and the corresponding mass values. 141
A.2 Overview of cross-links of She2p and She3p identified after measurement on the Q-ToF Ultima and the corresponding mass values. 142
A.3 Overview of cross-links of She2p and She3p identified after measurement on the Orbitrap Velos and the corresponding mass values. 142
A.4 Overview of cross-links of Cwc2 and the corresponding mass values. 143
A.5 Overview of cross-links from the 40S small ribosomal subunit and the corresponding mass values. 144
A.6 Overview of cross-links from the 60S large ribosomal subunit (proteins L1 to L8) and the corresponding mass values. 145
A.7 Overview of cross-links from the 60S large ribosomal subunit (proteins L16 to L42) and RPL7 and the corresponding mass values.. 146
A.8 Overview of cross-links of enzymatic and RNA-binding proteins (excluding ribosomal) and the corresponding mass values. 147

B.1 Calculated monoisotopic masses of RNA (marker) ions 149
B.2 Annotation of peptide–RNA adducts 150
B.3 RNA-adducts of single amino acids 151

1 Introduction

1.1 Protein–RNA complexes

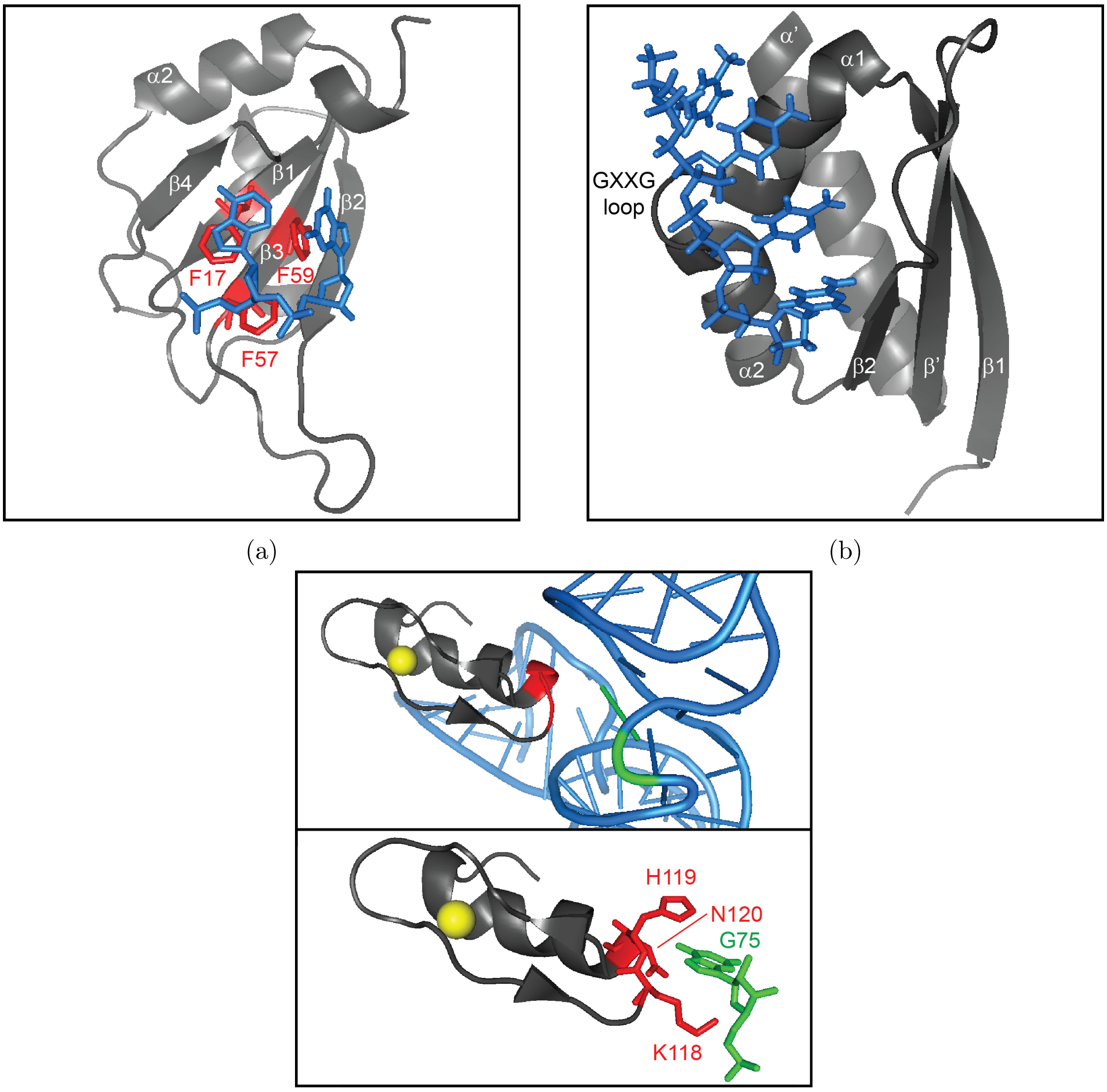
Protein–RNA complexes play a central role in many diverse cellular functions. The most prominent and well-studied example is the ribosome, containing and interacting with both proteins and RNA in translation.

RNA binding proteins (RBPs) can stabilize and protect RNAs, and mediate interactions to other proteins or RNAs in macromolecular assemblies. In addition, they can act as RNA modifying enzymes, e.g. nucleases that hydrolyze or degrade RNA, helicases that unwind double-strands, or transferases that directly modify nucleotides. For example, RBPs play key roles in the life of messenger RNA (mRNA), from translation through capping, polyadenylation, splicing, nuclear export, sub-cellular localization, translation and finally degradation.

RNAs are named according to their function and localization. The major classes are ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), messenger RNAs (mRNAs), small nuclear RNAs (snRNAs), microRNAs (miRNAs) and small interfering RNAs (siRNAs). rRNAs are the most abundant RNAs in the cell and the main components of the ribosome. tRNAs are the second most abundant group of RNAs and transport amino acids to the ribosome. mRNAs contain the genetic information after transcription and are the template for protein synthesis during translation. snRNAs, miRNAs and siRNAs are relatively small RNAs found in eukaryotic cells. snRNAs are part of the small nuclear ribonucleoprotein particles (snRNPs), the building blocks of the spliceosome. miRNAs and siRNAs bind to mRNAs and block translation or induce degradation, respectively.

Protein–RNA interactions are crucial for stability and function of ribonucleoprotein (RNP) particles. Therefore, knowledge of protein–RNA contact sites is required to understand the underlying molecular mechanisms. A single RNA binding region on a protein can be sufficient for RNA interactions. Similarly, a few nucleotides or short consensus sequences within the RNA are enough to mediate binding to RBPs.

In comparison to DNA, RNA secondary and tertiary structures exhibit a greater diversity. Therefore, interactions of RBPs with their target RNA molecules are more complex. Some RBPs contain conserved RNA-binding domains. Prominent examples are the RNA recognition motif (RRM [1,2]), the K homology domain (KH domain [3]), and zinc finger domains [4].



(a) (b)

(c)

Figure 1.1: Structure examples of the RNA recognition motif (RRM), the K homology (KH) domain, and the zinc finger (ZnF).

(a, b) Proteins are represented as cartoons in gray, RNA-binding residues as sticks in red. Nucleotides are shown in blue sticks.

(a) X-ray structure of hnRNP A1 bound to single stranded telomeric DNA. The figure shows details on the binding of RRM2 to deoxyadenosine and deoxyguanine involved in stacking interactions with F17 and F59, respectively. Residue F57 inserts between the deoxyriboses. Helix α1 lies behind the β-sheet. (pdb 2UP1 [5], representation adapted from [2])

(b) NMR structure of the third KH domain (type I) of hnRNP K bound to the DNA fragment TCCC. The DNA binding cleft is surrounded by the helices α1 and α2, the GXXG loop, the β2 strand and the variable loop connecting β1 and β'. (pdb 1J5K [6], representation adapted from [3])

(c) Crystal structure of ZnF 4 of transcription factor TFIIIA and loop E of 5S rRNA. The ZnF is shown in gray, the zinc ion in yellow, the RNA-binding residues in red, and the RNA in blue. The upper panel gives an overview while the lower panel shows binding details. The bulged G75 base (green) is contacted by H119 and N120, the ribose by K118 via hydrogen bonds. (pdb 1UN6 [7])

The RNA recognition motif, also referred to as RNA-binding domain (RBD) or ribonucleoprotein domain (RNP), is the most abundant RNA binding domain in higher eukaryotes. A typical RRM contains around 90 amino acids that form two α-helices that pack on a four-stranded β-sheet with a $\beta1$–$\alpha1$–$\beta2$–$\beta3$–$\alpha2$–$\beta4$ topology. The RRM binds between two and eight nucleotides, for example in the cap binding protein Cbp20 [8,9] or the spliceosomal U2B" protein [10], respectively. Primarily, interactions occur between the β-sheet and the RNA, more precisely through stacking of aromatic residues in the conserved RNP2 and RNP1 domains located in $\beta1$ and $\beta3$. This is illustrated in Figure 1.1a, where two phenylalanine residues stack with the two nucleotides bound by RRM2 of hnRNP A1. Additional contributions to binding affinity and selectivity can stem from loops connecting the secondary structure elements of the same RRM, other RRMs within the same protein, or different RNA-binding protein domains [2].

Another RNA-binding domain appearing in many RBPs of diverse functions is the K homology domain. The KH domain was first identified in and is named after the human heterogeneous nuclear RNP (hnRNP) K protein [11,12]. Typically, a protein contains multiple KH domains that can bind single-stranded RNA or DNA cooperatively or independently. The KH domain comprises around 70 amino acids that form a three-stranded β-sheet and three α-helices. $\alpha1$ and $\alpha2$ are commonly connected by a GXXG loop. The three-dimensional arrangement differs slightly between eukaryotic type I KH folds and prokaryotic type II KH folds. The KH domain can accommodate four bases in a cleft formed by the helices $\alpha1$ and $\alpha2$ together with the GXXG loop on one side and the $\beta2$ strand together with a variable loop on the other. As an example, the KH3 domain of hnRNP K is shown in Figure 1.1b. In contrast to stacking interactions in RRMs, protein–RNA contacts are primarily hydrogen bonds [3].

The third example for common RNA-binding domains are zinc fingers (ZnF). In these structural elements, zinc ions are coordinated by four histidine and cysteine residues within small, 20–30 amino acid long protein regions. Involvement of two or more zinc ions leads to the formation of more extended domains. The Cys_2His_2 zinc finger motif is the most abundant ZnF, appearing in about 3% of all human genes and primarily connected to DNA binding. It is composed of an α-helix and a β-sheet bound together by Zn^{2+} which is coordinated by two cysteines and two histidines at conserved positions. Cys_2His_2 ZnFs can bind double stranded DNA and RNA as well as a variety of single stranded RNA structural elements. Protein–RNA contacts can involve both hydrogen bonding and stacking interactions [4,7]. Figure 1.1c shows an example of a Cys_2His_2 ZnF of transcription factor TFIIIA binding to 5S ribosomal RNA.

Many studies have focused on RNA binding of specific RBPs *in vitro*. There has also been considerable effort in investigating protein–RNA interactions in the complex environment of the cell. Many approaches combine *in vivo* UV induced protein–RNA cross-linking with DNA sequencing techniques to identify binding sites on the RNA level [13]. UV cross-linking *in vitro*, together with mass spectrometry, has been applied in detailed structural and functional studies on RBPs. Very recently, it was demonstrated that UV cross-linking and mass spectrometry can be combined to identify RBPs after UV irradiation *in vivo* [14–16] (see below and 4.3).

1.2 Mass spectrometry in identification of biological macromolecules

Mass spectrometry has seen an exponential growth in applications in the field of biological macromolecules, especially proteomics, since the late 1990s. The expanding applications are tightly associated with technical and computational advances. A prerequisite for the success of mass spectrometry in the biological sciences was the development of soft ionization techniques. Electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI) had such a fundamental influence that John B. Fenn and Koichi Tanaka were awarded the Nobel Prize in Chemistry in 2002. They shared one half of the prize for their development of ESI (Fenn [17,18]) and MALDI (Tanaka [19]), respectively [20]. The German scientists Michael Karas and Franz Hillenkamp developed MALDI techniques almost simultaneously [21] to Tanaka and coworkers.

Over the years, ESI became the prevalent ionization method, particularly since it allows direct coupling of liquid chromatography to the mass spectrometer. As implied by the name, electrospray ionization is based on generating a fine spray of charged droplets. A solution containing the analytes continuously flows through a capillary. A potential difference is applied between the tip of the capillary and a counter-electrode at the interface of the mass spectrometer. The resulting electric field nebulizes the analyte solution into fine, charged droplets. The solvent evaporates while the droplets are transferred from atmospheric pressure into the vacuum of the instrument. When the Coulomb repulsion is greater than the surface tension, the droplet separates into smaller droplets (*Coulomb explosion*). This process could repeat itself until each droplet contains a single analyte ion (*charge-residue model*), further evaporation eventually leads to dissolved ions. Alternatively, free ions could be released from highly charged droplets (*ion evaporation model*). The exact mechanism might depend on the physical properties of the analyte and is still a subject of debate [22].

In the last decade, mass spectrometry (MS) has become one of the most important methods in proteomics, i.e. the investigation of proteins and their biological functions. Mass spectrometry based proteomics have proven extremely useful in qualitative and quantitative investigation of proteomes, including differentiation of sub-cellular or tissue dependent protein distributions. MS is also extremely valuable in the investigation of various post-translational modifications. Additionally, the composition of and interactions within macromolecular biomolecules can be investigated (MS based proteomics reviewed in [23,24]).

The mass determination and fragmentation of an isolated intact protein (*top-down mass spectrometry*) can yield valuable information, for example if it carries post-translational modifications. However, the *bottom-up* approach is more widely applied, where samples are analyzed by MS after proteolytic digestion. A typical large-scale proteomics experiment comprises the following steps, outlined in Figure 1.2: (1) The protein sample is prepared, e.g. by purifying proteins from a cell lysate or by isolation of a particular protein with its interaction partners by immunoprecipitation. (2) The protein mixture is separated by one-dimensional gel electrophoresis to decrease complexity. (3) Proteins are hydrolyzed by endoproteinases. (4) Peptides are further separated by reversed phase-high performance liquid chromatography (RP-HPLC) directly coupled to an ESI source of a mass spectrometer.

The peptide masses alone are not sufficient for unambiguous identification of the corresponding proteins in complex mixtures. Additional information is gained after dissociation of a single peptide

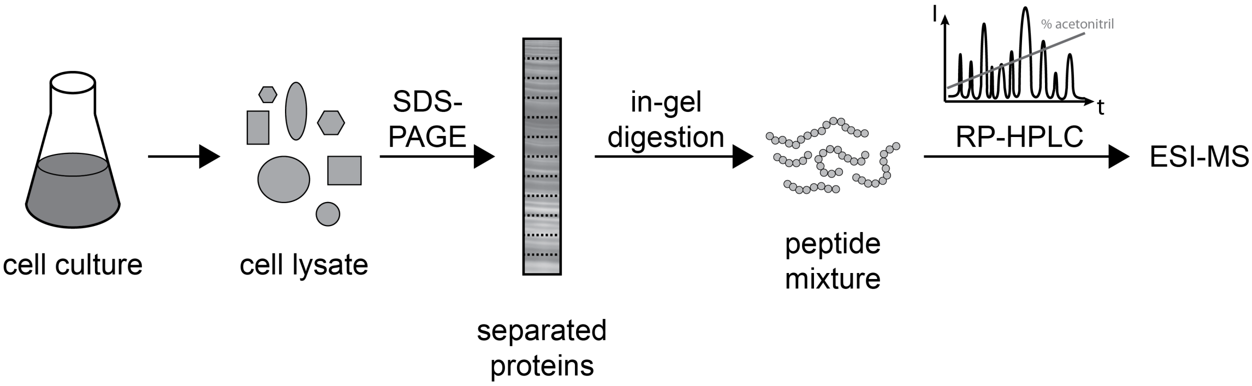


Figure 1.2: Schematic workflow of a typical large scale proteomics experiment. Cultured cells (or tissue samples) are lysed and the protein mixture is separated by SDS-PAGE. Next, proteins in individual gel slices are in-gel digested and peptides are eluted. Peptides are separated by RP-HPLC and subsequently analyzed by ESI-MS.

in the gas phase and monitoring the masses of the resulting fragments, a process termed *tandem mass spectrometry.* This approach, the necessary instrumentation and the analysis of the mass spectrometry data will be described in more detail below.

1.2.1 Tandem mass spectrometry

Tandem mass spectrometry or MS/MS combines two stages of MS: First, the mass of the intact ion is determined. Next, this *precursor ion* is isolated and fragmented. Low-energy collision induced dissociation (CID) is the most common fragmentation mode applied for large biomolecules. Therein, fragmentation is induced by collision with inert gas such as helium, argon or nitrogen. The measurement of the resulting product ions presents the second stage of the analysis.

Tandem MS can be carried out in three separate parts of the mass spectrometer (*tandem in-space*), i.e., selection of the desired ion, fragmentation and mass determination are performed by different components of the instruments. Since these instruments contain two mass analyzers, they are termed *hybrid mass spectrometers.* Quadrupole time-of-flight (Q-ToF) instruments (see below) are a prominent example of hybrid instruments that perform tandem in-space MS.

Tandem in-time instruments perform ion selection, fragmentation and mass determination in the same part of the instrument but sequentially in time. This applies to linear ion traps which can be found as stand-alone instruments. A linear ion trap is also part of most Orbitrap instruments, the second type of hybrid instrument which will be described in more detail below.

The majority of tandem mass spectrometry experiments in proteomics are carried out with *data dependent acquisition* (DDA) in the mass spectrometer. The instrument records a full scan (MS1), measuring the masses of all species eluting from the LC at that time point. Next, the species giving rise to the most intense signals are chosen for fragmentation. After the product ions scans (MS2, MS/MS) have been acquired, the instrument records the next set of MS1 and MS2 scans. This cycle is repeated over the entire duration of the chromatographic gradient. In DDA, low abundant species are less likely to be chosen for fragmentation, an effect that increases with sample complexity.

1.2.1.1 Quadrupole time-of-flight (Q-ToF) mass spectrometers

A Q-ToF mass spectrometer is a hybrid instrument combining a quadrupole and a time-of-flight mass analyzer (see Figure 1.3). A quadrupole is composed of four parallel metal rods that serve as electrodes. An electric field is generated by applying both direct current (DC) and radio frequency (RF) potentials to the metal rods. At a given combination of DC and RF, only ions within a narrow mass-to-charge (m/z) ratio window pass through the quadrupole. All other ions are not confined within the quadrupole and are removed by the vacuum system. If only a radio frequency is applied, ions over a wide m/z range can pass through the quadrupole. The quadrupole can scan though an m/z range by changing both DC and RF potentials while keeping their ratio constant. By detecting at which ratio ions reach a detector, a mass spectrum can be acquired.

The time-of-flight mass analyzer separates ions according to their m/z ratio in a field-free drift region. Ions with a small m/z ratio travel faster than those with a higher m/z ratio. Important for resolution and mass accuracy is that the ions enter the flight path at the same time with the same kinetic energy. The mass spectrum is recorded by detecting at which time ions reach the detector, the time is converted to the corresponding m/z ratio. The resolution of a ToF analyzer is increased with integration of a reflectron which serves as an electrostatic mirror. Ions with higher velocity penetrate deeper into the repelling electric field of the reflectron, compensating differences in kinetic energy of ions with the same m/z ratio. In addition, the flight path is increased, also leading to higher resolution.

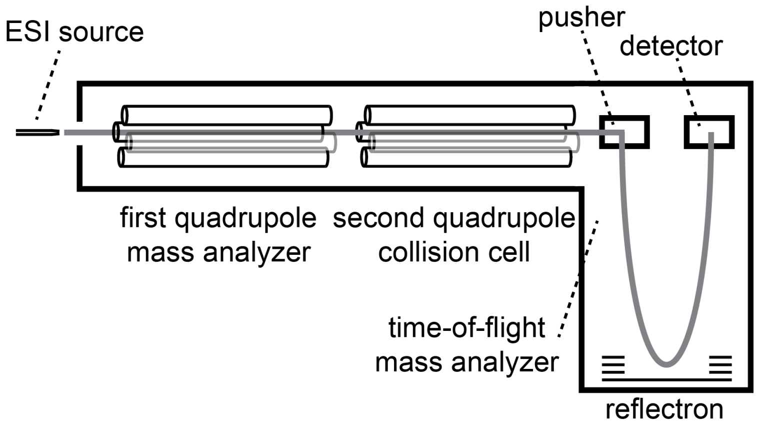

ESI source
pusher
detector
first quadrupole mass analyzer
second quadrupole collision cell
time-of-flight mass analyzer
reflectron

Figure 1.3: Schematic representation of a quadrupole time-of-flight (Q-ToF) mass spectrometer. It contains a regular quadrupole (mass analyzer) and an RF-only quadrupole (collision cell). Ions are directed into the time-of-flight mass analyzer by the pusher. The drift region is increased by the reflectron which guides the ions toward the detector.

A simplified Q-ToF mass spectrometer is depicted in Figure 1.3. In a first scan (*precursor ion scan*), all ions in a wide m/z range pass through both quadrupoles. The pusher applies a short pulse of an orthogonal accelerating field to the constant ion beam passing through the second quadrupole to direct a group of ions into the field-free drift region of the ToF. This way, the MS spectrum is recorded. For fragmentation experiments, the first quadrupole serves as a mass analyzer, selecting ions of the desired m/z ratio. These are subsequently fragmented in the second, RF only quadrupole by collision with inert gas (*beam-type collision induced dissociation*). The product ion scan is again recorded in the ToF mass analyzer [22].

1.2.1.2 Orbitrap mass spectrometers with linear ions traps (LTQ Orbitraps)

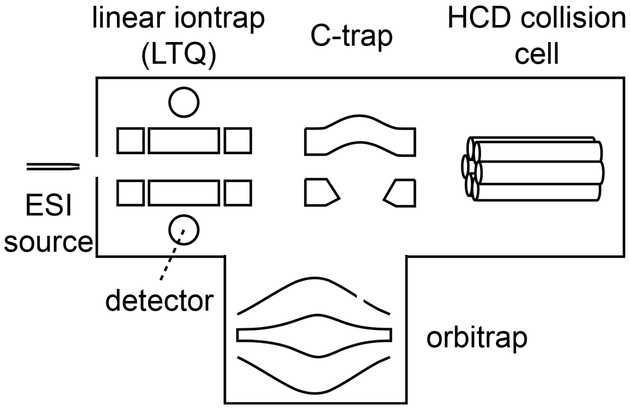


Figure 1.4: Schematic representation of a LTQ Orbitrap mass spectrometer. The first mass analyzer is a linear ion trap with adjacent detectors. The orbitrap serves as mass analyzer and detector. From the ion trap, ions can be passed to the HCD collision cell or injected into the orbitrap by the C-trap. Mass spectra can be recorded in the ion trap as well as the orbitrap. Fragmentation can be performed in the ion trap (CID) or in the HCD collision cell.

A more recently developed class of hybrid mass spectrometers are LTQ Orbitraps. They contain a linear ion trap (linear trap quadrupole, LTQ) and an orbitrap mass analyzer. A simplified scheme is shown in Figure 1.4.

As implied by the name, the LTQ shares similarities to quadrupoles. It is built of four hyperbolic rods that are typically separated into three axial sections. Ions are trapped in the axial direction by applying different DC voltages to the three sections, and in the radial direction by RF potentials between opposite rods within the same section. Two of the central rods have a small slit though which ions can be ejected towards the detectors. Alternating current (AC) voltages are applied to these rods for isolation, activation, and ejection of ions.

The ion trap is held under a low helium pressure. Ions gather kinetic energy during acceleration by the ion optics between ESI source and ion trap (omitted from the simplified representation in Figure 1.4). During trapping, slow collisions with the inert gas lead to decrease of kinetic energy (*cooling* of ions).

In order to record a mass spectrum, the RF amplitude is increased at a constant rate from low to high voltages. This leads to successive destabilization of ions with an increasing m/z ratio. The AC is kept at constant frequency but increasing amplitude. This way, instable ions are directed through the slits towards the detector.

In order to isolate ions in a narrow m/z window, all other ions are destabilized at a constant RF amplitude by changing the AC frequency, skipping the frequency at which the ions of interest would become instable. After isolation, this AC frequency is used to increase the kinetic energy of the ion of interest. However, the AC amplitude is considerably smaller than during isolation so that the ions are not ejected. Due to the increased kinetic energy of the ions, collisions with helium lead to fragmentation (*ion trap collision induced dissociation*). The fragment spectrum can then be recorded as described above.

The applied potentials cannot stabilize ions with a small m/z ratio. This *low mass cut-off* typically affects the lower third of the m/z range with respect to the uncharged precursor mass and thus prevents the detection of small fragmentation products.

The orbitrap consists of an axial central electrode and a co-axial outer electrode. The electrostatic field traps ions rotating around the central electrode and oscillating along its axis. Only the axial movement is independent of kinetic energy and spatial distribution of the ions, but it is related to the m/z ratio. The frequencies of the oscillating axial movement are detected by the current induced between the halves of the outer electrode and are converted into m/z ratios by Fourier transformation.

In a typical tandem MS experiment on LTQ Orbitraps, all ions entering the instrument are trapped in the LTQ, passed on to the C-trap and injected into the orbitrap where a high resolution precursor ion scan is recorded. Meanwhile, ions are isolated, fragmented and product ion scans are recorded in the linear ion trap (tandem in-time, see above). Since the sequencing speed of the LTQ by far exceeds that of the orbitrap, several product ion scans can be recorded in it while the precursor ion scan is acquired in the orbitrap. The CID spectra can also be recorded in the orbitrap, with the benefit of a considerably higher resolution and mass accuracy compared to the LTQ but at a significantly lower acquisition speed.

In addition to ion trap CID, LTQ Orbitraps offer a second fragmentation mode corresponding to beam-type CID on Q-ToF instruments, termed *higher-energy collision dissociation* (HCD). Ions are again collected in the linear ion trap, the desired ion is isolated and passed to the HCD collision cell (multipole). There, the ions are fragmented by collisions with nitrogen molecules. The product ions are ejected into the C-trap and transferred into the orbitrap where the fragment spectrum is recorded [22,25]. HCD is slower compared to MS/MS in the ion trap but does not exhibit a low mass cut-off. In addition, HCD fragmentation corresponds to beam-type CID and is beneficial for some applications in comparison to ion trap CID (see 1.3.3.4).

1.2.1.3 Fragmentation of peptides

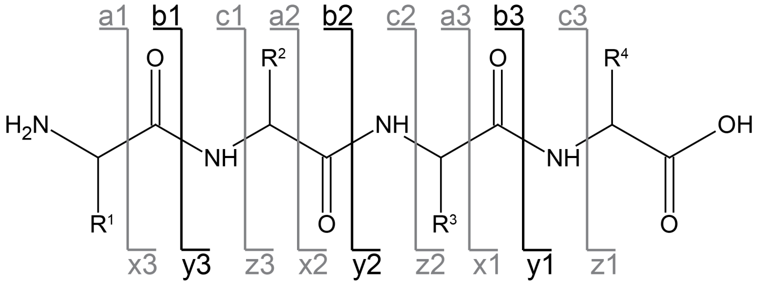


Figure 1.5: Nomenclature of peptide fragments resulting from backbone cleavage. Cleavage of the alkyl carbonyl bond produces a- and x-ions, cleavage of the amide bond b- and y-ions, and cleavage of the amino alkyl bond leads to c- and z-ions. a-, b-, and c-ions contain the peptide N-terminus while the corresponding C-terminal ions are called x, y, and z [26,27].

Mass spectrometric analysis of peptides is usually carried out from acidic solutions in positive ion mode. Fragmentation of peptides is mostly charge-directed, i.e., a proton at the cleavage site is required. On the protonated peptide ions, charges are preferentially located on basic residues (arginine, lysine, histidine) and at the peptide N-terminus. The energy transferred to the peptide upon collision with inert gas atoms or molecules can initiate redistribution of protons (*mobile proton model*) prior to fragmentation [28]. However, fragmentation of protonated peptides is highly complex and follows many different pathways [29]. Therefore, reliable prediction of observable fragments and especially their relative intensities is still not possible to the extent that this information could be the basis for automated peptide identification. Instead, all ions within a series are assumed to occur with the same probability and intensities are disregarded.

Figure 1.5 shows the three possible sites of fragmentation on the peptide backbone. Cleavage of the amide bond occurs most frequently, especially under CID conditions. The resulting spectra therefore contain b- and y-ion series as well as a-ions that are formed after loss of CO from b-ions. Within one series, the distance between two neighboring signals equals the mass of the corresponding amino acid (in its chain form without the water lost during amide bond formation). Therefore, amino acid sequences can be derived from calculated mass differences of fragment ions.

1.2.1.4 Fragmentation of RNA

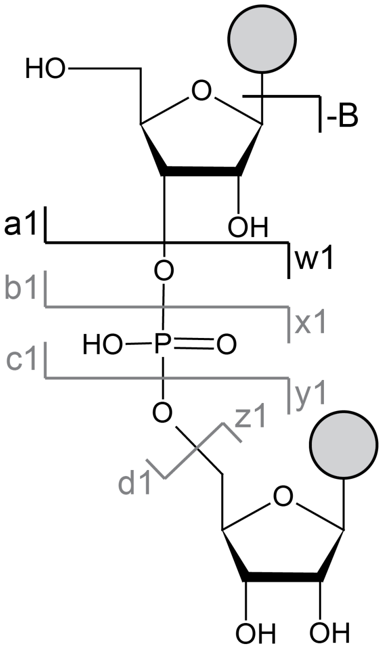


Figure 1.6: Nomenclature of RNA fragments. Bases are simplified as gray spheres. In analogy to the nomenclature of peptide sequence ions, fragments resulting from cleavage of the phosphate backbone are termed a, b, c, or d for fragments containing the 5' end or w, x, y, or z if the charge is retained on the 3' end. Loss of a base is denoted as -Bn(X), where n is the position of the base counting from the 5' end and X is the one letter code of the base [30].

In contrast to the widespread application of mass spectrometry techniques in proteomics, MS is much less frequently applied in investigation of DNA or RNA. For most questions, (DNA) sequencing techniques are preferred as they can handle longer oligonucleotide segments, are less expensive and provide greater multiplexing capabilities.

Mass spectrometric analysis of oligonucleotides is usually carried out from basic solutions in negative ion mode. The nomenclature follows rules similar to those of peptide fragments (see Figure

1.6). Upon CID fragmentation, the N-glycosidic bond is often cleaved to release the nucleic acid base, either as neutral loss or as a base anion. Additionally, backbone fragmentation predominantly leads to the formation of c- and y-ions. For DNA oligonucleotides, loss of the base is more dominant and backbone fragmentation leads to the formation of a- and (w-B)-ions. Fragmentation of protonated DNA in positive ion mode leads to similar product ion types. It has been proposed that the abundance of protonated bases after fragmentation correlates with the proton affinity, with $\mathrm{C} \sim \mathrm{G} > \mathrm{A} \gg \mathrm{T}$ [31].

1.2.2 Data analysis in proteomics

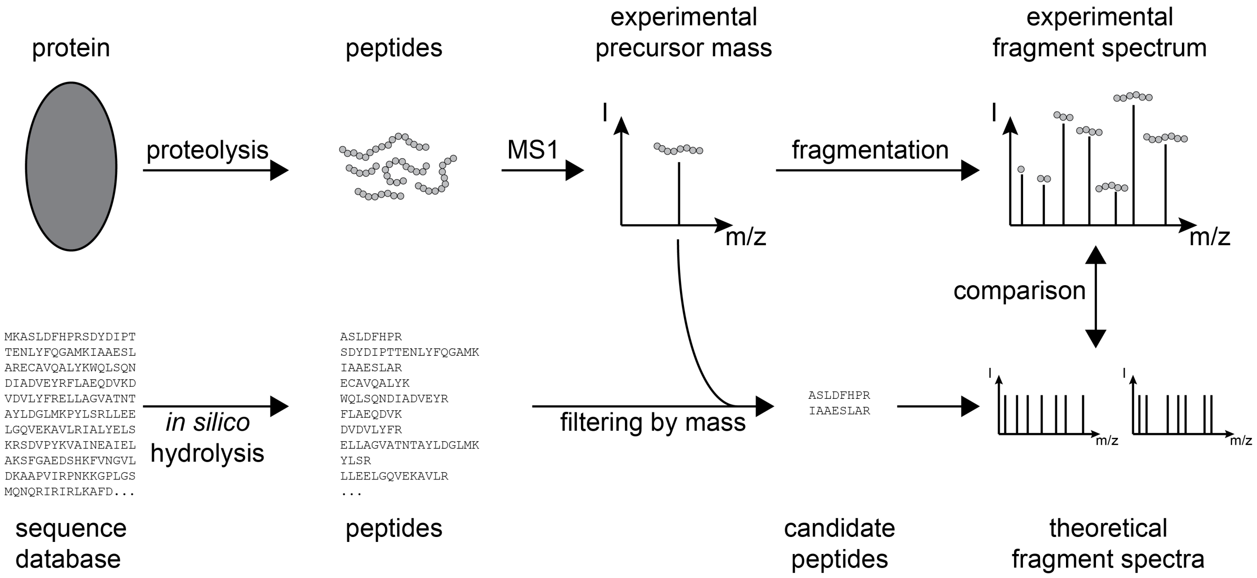


Figure 1.7: Principles of sequence database searching. Searches follow essentially the same steps as the experimental workflow. The sequence database is hydrolyzed *in silico*. Experimentally, the protein is hydrolyzed into peptides and their masses are determined. The experimental mass is used to filter for database peptides with the same calculated mass. Theoretical fragment spectra are generated and compared to the experimentally acquired spectrum.

Identification of a peptide from mass spectrometry data relies on matching both the mass of the intact peptide (*precursor mass*) and the masses of its fragmentation products. All algorithms for automated peptide identification could in principle allow for any combination of the 20 standard amino acids. However, to limit calculation times and decrease the number of false positive results, several constraints are introduced. The order in which these are applied depends on the exact algorithm. The most popular sequence database searching approach will be described in more detail (this and other algorithms are reviewed in [32]).

In sequence database searching (overview in Figure 1.7), MS data is searched against a protein sequence database. The search engine uses the precursor mass and the masses of the fragments to match a MS/MS spectrum to a database peptide. First, the protein database is digested *in silico*, thus protease specificity is taken into account. This way, the amino acid on the peptides' extreme N- or C-terminus is limited to one or a few candidates, e.g. to lysine or arginine as the C-terminal amino acid in the case of the endoproteinase trypsin. Next, the list of peptides is filtered for candidates with a mass corresponding to the experimental precursor mass determined in the MS analysis. A mass tolerance for deviation between experimental precursor and calculated peptide masses is set according to the mass accuracy of the MS instrument.

For all candidate peptides, a theoretical fragment spectrum is generated. These spectra are compared to the experimental product ion scan obtained in the MS analysis, taking the fragmentation mode into account. For example, comparisons are mainly based on b- and y-ions for collision induced dissociation. Agreement between expected and observed fragments is evaluated. Typically, only the peptide-to-spectrum match (PSM) with the highest overlap is considered as a possible correct match. Finally, the significance of the match is determined and expressed as a score. To this end, the probability is calculated for the PSM to be a random event, i.e. that the peptide was matched to the spectrum purely by chance. The exact scoring algorithms depend on the database search engine employed. The commercial search engine Mascot (Matrix Science [33]) and the open-source Open Mass Spectrometry Search Algorithm (OMSSA, National Center for Biotechnology Information [34]) report E-values. This value indicates the expected number of peptides randomly matching the spectrum with scores equal to or better than the score of the identified PSM. Consequently, low E-values correspond to a high significance, i.e. the match is less likely to be a random false positive hit.

Many post-translational modifications (PTMs) lead to a distinct mass increase of the modified amino acid. Therefore, mass spectrometry is ideally suited for identification and localization of many PTMs. In the database search, peptides containing a potentially modified amino acid are considered with and without the modification. This increases the search space exponentially with each PTM, consequently processing times are longer and false positive matches more likely. Therefore, only a limited number of PTMs can be considered in each search. The PTM does not only increase the mass of the protein and peptide, but also the mass of the peptide fragments containing the modified amino acid. This shift of product ion masses is used to localize the modification in the primary sequence.

1.3 UV induced protein–RNA cross-linking

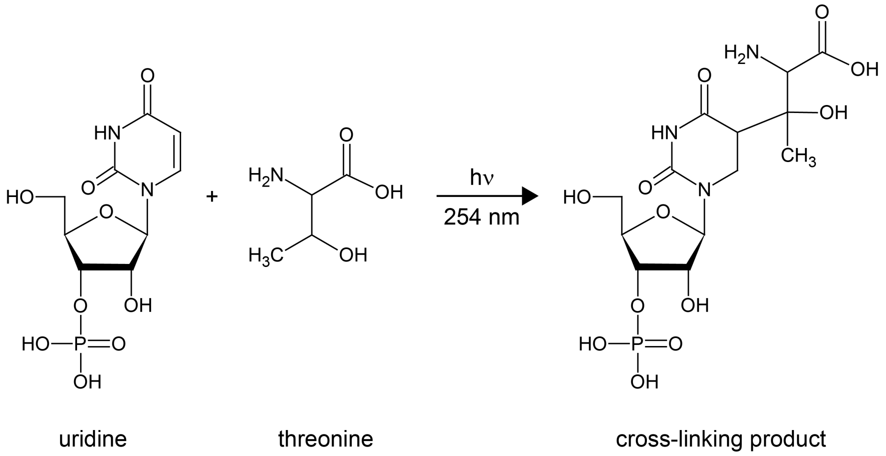


Figure 1.8: Possible reaction between uridine and threonine upon UV induced cross-linking. The uracil base is excited upon absorption of UV light. Subsequently, a hydrogen atom might be abstracted from threonine and the cross-linking product could be formed by radical combination.

This free radical based reaction presents the most common mechanism (according to [35] and references therein). In general, the mechanism of cross-link formation is not fully understood, the exact mechanisms for the presented example as well as other bases and amino acid residues might differ.

UV induced cross-linking of RNA (or DNA) to proteins is a frequently applied method for studying direct protein–nucleic acid interactions. The approach employs the natural reactivity of the nucleic acid bases: Upon absorption of UV light, the base is promoted into an excited electronic state. Subsequently, chemical reactions with amino acids can lead to formation of covalent bonds between nucleic acid bases and amino acid residues [35], so-called *zero-length cross-links*. The yield of UV induced protein–RNA cross-linking is very low, e.g. 0.2-0.5% for cross-linked peptides from prokaryotic ribosomal subunits after isolation by size exclusion and reversed phase chromatography [36]. Importantly, since the cross-link only consists of a single covalent bond, UV induced cross-linking only occurs between nucleic acid bases and amino acid residues that are in close spatial proximity.

While all nucleic acid bases and amino acid residues can form cross-links, significant differences are observed in their reactivity. In a series of studies, Shetlar *et al.* have systematically investigated cross-linking yields of nucleotides and amino acids. For cross-linking of single amino acids to DNA, they found cysteine, lysine, phenylalanine, tryptophan, and tyrosine to be the most reactive, while alanine, aspartic and glutamic acid, serine, and threonine were unreactive [37]. In addition, the reactivity of polynucleotides towards single amino acids (excluding proline) was tested. Polyuridylic acid reacted with all amino acids; polythymidylic, polyguanylic, polycytidylic, and polyadenylic acid reacted with a decreasing number of amino acids. Phenylalanine, tyrosine, and lysine were among the amino acids with the highest yields for all five polynucleotides [38,39].

The excited states of nucleic acid monomers have short lifetimes (picosecond range for the first singlet state S1, microseconds for the first excited triplet state T1), while interactions in polynucleotides can lead to longer-lived excited species [40]. The short period in which the cross-link can be formed has an important consequence: The ribonucleoprotein complex cannot undergo major distortions while the nucleic acid is excited. The short time for reaction initiation together with the formation of single covalent bonds are the reasons for the high specificity of UV induced protein–RNA cross-linking. Direct interactions of nucleotides and amino acids in the native ribonucleoprotein complex are fixed, enabling their investigation by various analytical methods. Cross-linking by chemical reagents does not provide this strict specificity. Maximum distances between reacting groups are determined by the structure of the reagent and the cross-linking chemistry. After the reagent has reacted with the first target group, it remains reactive until cross-linking is completed by the reaction with the second target group. In this prolonged time period, distortions of the investigated macromolecule are more likely to lead to artificial cross-links.

Bioanalytical methods that have been applied to identify RNA-binding proteins after UV induced cross-linking include SDS-PAGE (e.g. [41,42]), Western blotting (e.g. [43]) and immunoprecipitation (e.g. [41,44,45]). Recently, several studies reported the identification of RNA-binding proteins with mass spectrometry after cross-linking and purification of mRNA with oligo(dT) under stringent conditions, thus removing noncross-linked proteins almost completely [14–16]. In general, mass spectrometry based methods have the advantage that no prior knowledge about the sample protein content is required.

Detailed investigations of the cross-linking site on the protein level have initially been carried out by Edman sequencing after (semi-) preparative isolation of peptide–RNA oligonucleotide heteroconjugates (e.g. [36,46,47]). Identification of the cross-linking site by mass spectrometry has been a long-standing interest in our laboratory and will be described in more detail below. Knowledge of protein regions, peptides or even amino acids directly interacting with RNA can provide valuable information, for example about RNA-binding surfaces or novel RNA-binding domains.

1.3.1 Preparation of ribonucleoprotein complexes and UV cross-linking

Protein–RNA complexes for investigation with UV cross-linking and mass spectrometry can be prepared through either purification of native ribonucleoproteins or by *in vitro* reconstitution. For the latter, incomplete or nonspecific assembly should be excluded as it can result in artificial cross-links.

In general, protein–RNA complexes can be purified from cell extracts or after *in vitro* reconstitution with several strategies, e.g. as outlined in Figure 1.9. A protein and its interaction partners can be isolated after introduction of a tag suitable for affinity purification, e.g. histidine tags, glutathione *S*-transferase (GST) tags, or tandem affinity purification (TAP) tags. Similarly, aptamer tags can be introduced to an RNA sequence, e.g. stem loop structures specifically bound by the MS2 bacteriophage coat protein. Additionally, antibodies that specifically bind the protein of interest or certain RNA elements (e.g. the 5' cap structure of snRNAs) can be employed.

The conditions during UV irradiation, i.e. sample amount and concentration, buffer constituents, light source, and irradiation time, influence the obtained cross-links and should be carefully chosen. Mass spectrometric identification of cross-linked peptide–RNA heteroconjugates can be prevented by insufficient sample amounts. Cross-links can be below the detection limit or produce low intensity signals, resulting in poor quality spectra that do not permit unambiguous identification. For *in vitro* complex reconstitution, the sample volume should be kept small to avoid incomplete complex formation due to dilution effects.

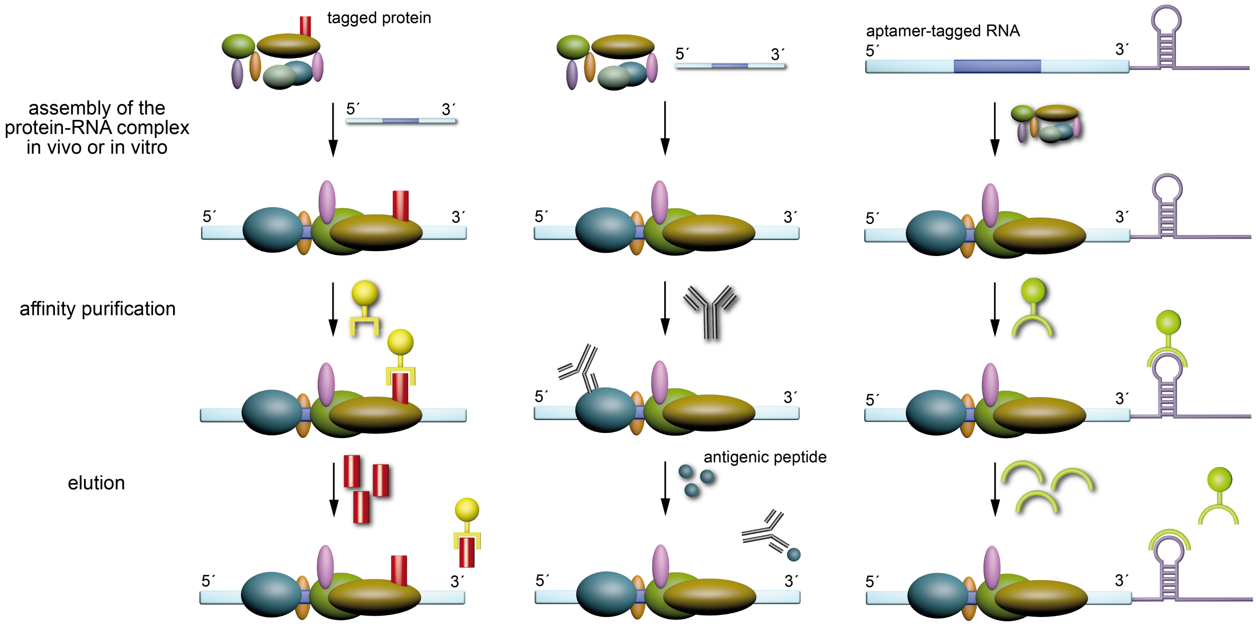


Figure 1.9: Strategies for isolation of protein–RNA complexes. Complexes can be prepared be affinity purification of a tagged protein (left), immunoprecipitation (middle), or through a tagged RNA (right).
Figure originally published in [48].

In principle, UV cross-linking tolerates a wide range of buffer conditions. However, higher concentrations of radical scavengers like glycerine should be avoided as they might prevent cross-link formation [49]. In addition, certain detergents like sodium dodecyl sulfate (SDS) are incompatible with LC-ESI-MS analysis and higher concentrations should be avoided [50]. Frequent contamination of e.g. Triton X-100, Tween and NP-40 with polyethyleneglycole (PEG) can cause substantial problems, high intensity PEG signals can completely suppress other signals in the mass spectrometer. These detergents should be avoided completely and in general the highest grade reagents and buffer constituents should be used.

Light sources for UV irradiation are UV lamps or monochromatic lasers [51]. The energies of the emitted light differ substantially. Laser light can induce two photon absorption and in consequence ionization. Ions have considerably longer life times than excited states and are more likely to lead to unspecific reactions upon structural perturbations. The optimal irradiation time depends on the light source and the complex under investigation. Longer irradiation may increase cross-linking yields but can cause photodamage to both proteins and RNA (e.g. [52]).

1.3.1.1 Incorporation of substituted nucleotides

Incorporation of photoreactive base-analogues such as 4-thio-uracil, 6-thio-guanine, 5-bromo-uracil, or 5-iodo-uracil is a strategy to increase the cross-linking yield. The absorption maxima of the base-analogues lie at higher wavelengths (4-thio-uridine 330 nm, 6-thio-guanine 342 nm, 5-bromo- and 5-iodo-uridine around 280 nm) compared to the native RNA bases (250-270 nm) [35]. Therefore, complexes containing photoreactive nucleotides are irradiated at longer wavelengths, typically 312 nm for halopyrimidines and 365 nm for 4-thio-uracil and 6-thio-guanine. Especially at 365 nm, no cross-linking of the native nucleotides occurs and undesired photocleavage and oxidation is reduced [52]. Halopyrimidines react via radical-based mechanisms and loss of the respective hydrohalogen [53]. To our knowledge, the reaction mechanisms of 4-thio-uracil and 6-thio-guanine have not been previously investigated in detail by mass spectrometry. The observation of frequent thymidine to cytidine transitions after cDNA sequencing of 4-thio-uracil substituted RNA suggests a product of the UV induced reaction that alters base pairing properties of the cross-linked nucleotide. The same might apply to 6-thio-guanine, where sequence reads were enriched in guanosine to adenosine transitions [54].

Photoreactive base-analogues can be incorporated site-specifically into synthetic RNAs. The approach is therefore limited to *in vitro* reconstituted complexes. Site-specific labeling of RNA with photoreactive nucleotides can be used to identify interaction sites on the RNA level. For example, binding sites of yeast spliceosomal proteins on U5 snRNA were investigated by site-specific labeling of U5 and identification of cross-linked proteins by immunoprecipitation [55]. MS-based cross-linking studies on protein–DNA complexes include, for example, incorporation of 5-bromo-deoxyuridine [56] or 4-thio-thymidine [57].

Alternatively, the photoreactive base-analogues can be incorporated randomly *in vivo*. The enhanced cross-linking yield improves results of different approaches that combine immunoprecipitation, isolation of the cross-linked RNA and its analysis by cDNA sequencing (e.g. cross-linking and immunoprecipitation (CLIP) [58,59], cross-linking and analysis of cDNAs (CRAC) [60,61], photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) [54]).

1.3.2 Sample preparation for mass spectrometry, enrichment and purification strategies

To ease interpretation of MS and MS/MS spectra of cross-linked heteroconjugates, both proteins and RNA need to be hydrolyzed thoroughly prior to and/or following enrichment or isolation strategies [62,63]. Disassembly and denaturation of protein–RNA complexes can be achieved in 1 M urea or guanidine hydrochloride, larger amounts of other detergents like SDS should be avoided (see above) [49]. As for most proteomic approaches, trypsin is favored as the enzyme for proteolysis as it cleaves after the basic residues lysine and arginine. This usually leads to peptide fragmentation series starting from the peptide C-terminus, aiding data interpretation. Chymotrypsin has also been applied successfully, especially in studies of snurportin 1 and U1 snRNA as well as reconstituted human [15.5K-61K-U4atac snRNA(-U6atac snRNA)] [47,64–66]. Use of different endoproteinases can lead to the identification of additional cross-linking sites within the same protein [45]. Since increasing length of the cross-linked RNA oligonucleotide leads to suppression of the peptide fragment signals, RNA hydrolysis to single or a low number of nucleotides is desirable, especially for ESI-MS [62]. This can also be achieved by complete hydrolysis of the oligonucleotide with HF [67].

Due to the low cross-linking yield and the usually limited amounts of starting material, the purification or enrichment of cross-linked heteroconjugates is a crucial step in the sample preparation for mass spectrometric analysis. The high excess of noncross-linked peptides and oligonucleotides would otherwise hinder cross-link detection and identification. Through enrichment, sample complexity is greatly reduced, which is beneficial in maximizing the number of potential cross-links chosen for MS/MS fragmentation and for data analysis. In addition, signal suppression by noncross-linked RNA oligonucleotides and peptides is decreased.

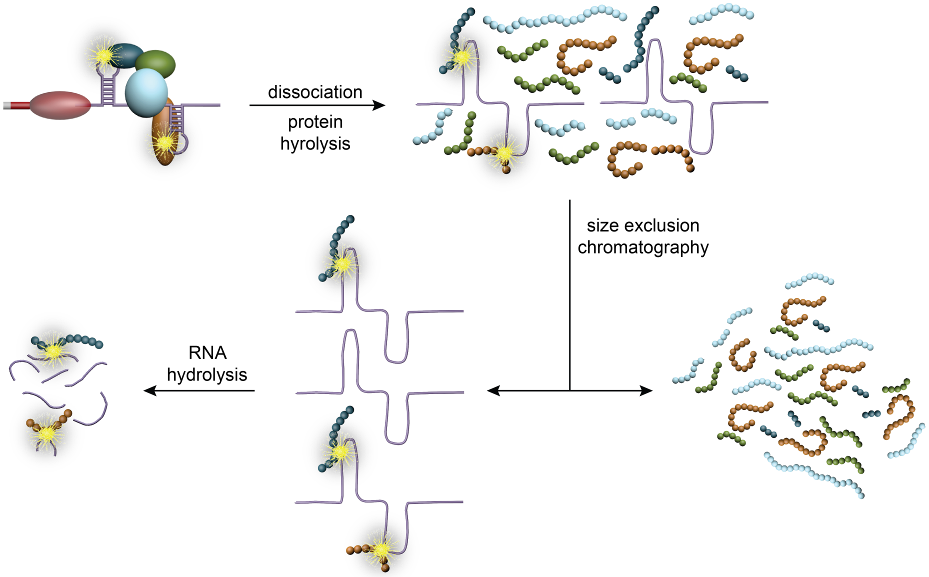


Figure 1.10: Isolation of cross-linked heteroconjugates from noncross-linked peptides by size exclusion chromatography. After proteolysis under denaturing conditions, full-length RNA together with cross-linked peptides can be isolated from peptides by size exclusion. After RNA hydrolysis, noncross-linked oligonucleotides need to be removed through suitable methods.

Purification by reversed phase high performance liquid chromatography (RP-HPLC) or size exclusion (SE) chromatography as well as enrichment via immobilized metal-ion affinity chromatography (IMAC) or titanium dioxide material have been applied successfully.

Size exclusion chromatography can be applied if proteins and RNA differ considerably in size or for isolation of RNA with and without cross-linked peptides following proteolysis. Both approaches were combined in cross-linking studies of the prokaryotic ribosome [36, 46, 68]. In a first SE, ribosomal RNA (rRNA) with cross-linked proteins was separated from noncross-linked ribosomal proteins. After proteolysis, a second SE step separated rRNA with cross-linked peptides from noncross-linked peptides. rRNA containing fractions were hydrolyzed by nucleases and cross-linked peptide–RNA heteroconjugates were separated by RP-HPLC. Isolated cross-links were subjected to Edman sequencing to identify the sequence of the cross-linked peptide. In many cases, the cross-linked amino acid led to a gap in the sequence analysis and could thus be identified.

In contrast to ribosomes, the size difference between uridine-rich small nuclear RNAs (U snRNAs) and their associated proteins is not sufficient to permit their separation by size exclusion chromatography. Therefore in studies of human small nuclear ribonucleoprotein particles (snRNPs), SE was only applied after proteolysis (as outlined in Figure 1.10). RNA containing fractions were subsequently hydrolyzed with nucleases and endoproteinases. The mixture was then separated by RP-HPLC. Monitoring absorption at both 220 nm (peptides) and 260 nm (RNA) allowed the detection of heteroconjugates (e.g. [64, 66]).

After size exclusion chromatography and hydrolysis of RNA-containing fractions, the mixture can also be directly subjected to on-line LC-ESI-MS/MS. This was demonstrated in a cross-linking study of the human U1 snRNP and the reconstituted [15.5K-61K-U4atac snRNA] complex. An extensive washing step was included to remove the noncross-linked RNA oligonucleotides, cross-linked heteroconjugates and residual peptides were retained on the trapping column [63].

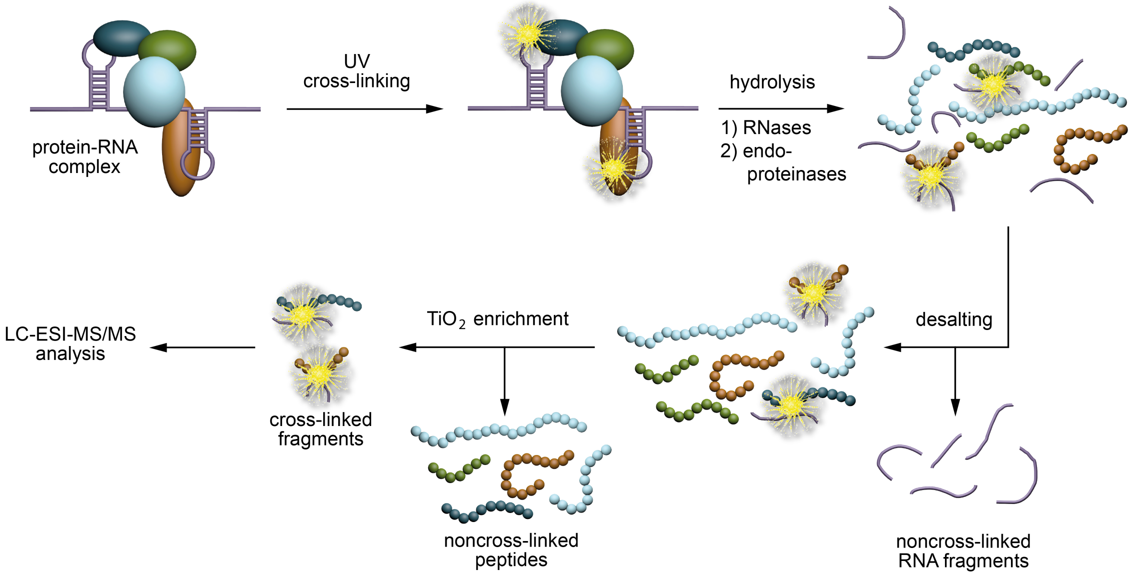


Figure 1.11: Enrichment of cross-linked heteroconjugates with C18 and titanium dioxide chromatography. The protein–RNA complex is UV irradiated, for native RNA at 254 nm. Next, the complex is hydrolyzed under denaturing conditions with RNases and endoproteinases. Desalting with C18 material removes the majority of noncross-linked RNA oligonucleotides. Finally, titanium dioxide chromatography separates noncross-linked peptides from the cross-linked heteroconjugates. These are then subjected to LC-ESI-MS/MS analysis.
Figure originally published in [48].

Several enrichment protocols for phosphopeptides are based on the interaction of the phosphate groups with metal ions. These can be adapted to enrich peptide–RNA oligonucleotide heteroconjugates via the phosphate groups in the RNA backbone. At first, enrichment protocols were based on immobilized metal-ion affinity chromatography (IMAC) with Fe(III) ions for cross-link to either RNA [62,66] or DNA [67,69].

More recently, enrichment based on titanium dioxide (TiO_2) chromatography was applied. Protocols initially established for phosphopeptides use competitive binding with 2,5-dihydroxy benzoic acid (DHB) to TiO_2 to reduce co-enrichment of acidic peptides [70]. This approach could be directly conveyed to cross-linked peptide–RNA oligonucleotide heteroconjugates [71]. After UV irradiation and ethanol precipitation, the protein–RNA complexes are hydrolyzed by RNases and trypsin. The sample is then desalted and the cross-links are subsequently enriched over TiO_2 columns in the presence of DHB [71] (see Figure 1.11). In contrast to IMAC agarose beads, titanium dioxide can be integrated into an HPLC setup. A two dimensional LC approach, combining C18 and TiO_2 columns, has been used for enrichment and subsequent automatic spotting for MALDI-MS analysis [72].

Overall, the most frequently applied method in our laboratory is TiO_2 enrichment with spin columns, as it requires less sample amounts compared to size exclusion or reversed phase isolation and is more selective than IMAC enrichment strategies. It also allows for the enrichment of several samples in parallel and is compatible with LC-ESI-MS/MS analysis, which is advantageous compared to the 2D LC setup with subsequent MALDI analysis in terms of sample processing time and quality of the obtained MS information.

1.3.3 Mass spectrometry of peptide–RNA heteroconjugates

The efficiency of mass spectrometric analysis of UV induced protein–RNA cross-linking experiments is closely connected to the performance of the mass spectrometer. Ionization technique, sensitivity, mass accuracy and resolution as well as sequencing speed greatly influence the obtained results. The MS methods that have been applied to the identification of cross-links often correlated to the development in MS instrumentation in general.

The divergent physico-chemical properties of proteins and RNA lead to their analysis in positive and negative ion mode, respectively. Peptide–RNA oligonucleotide heteroconjugates are usually analyzed in positive ion mode since unambiguous identification of the cross-linked peptide is desired. The ionization efficiency of heteroconjugates is decreased by the negatively charged phosphodiester backbone. In consequence, identification of cross-links requires higher sample amount compared to standard MS based proteomics experiments [52]. The negative ion mode is preferred only for mass determination of heteroconjugates containing longer RNA oligonucleotides as it increases signal intensities. Fragmentation of the peptide moiety in negative ion mode is poor, preventing its identification [65].

1.3.3.1 MALDI mass spectrometry

The first mass spectrometry based analyses of cross-linked heteroconjugates were performed with matrix assisted laser desorption/ionization (MALDI). Initially, only the mass of the intact heteroconjugate was determined, cross-links were identified by their mass and information derived from complementary experiments [46, 47, 64, 68]. Later, fragmentation by post source decay (PSD) was performed to acquire fragment information [65, 66]. The matrices dihydroxy benzoic acid (DHB) and 2,4,5-trihydroxyacetophenone (THAP) provided the best results in respect to signal intensity and spectrum quality [65]. Analysis was mainly done in positive ion mode, but in contrast to LC-ESI-MS/MS, switching between positive and negative mode within the same experiment was easier [62]. The loss of H_3PO_4 (98 Da) can be used as an initial indicator for phosphate containing species [65, 66, 72].

1.3.3.2 ESI mass spectrometry

Electrospray ionization (ESI) mass spectrometry directly couples liquid chromatography to the instrument (on-line LC-ESI-MS/MS). In comparison to MALDI, lower sample amounts are necessary and the obtained sequencing results are more informative. Therefore, ESI-MS gradually replaced MALDI-MS and is now the method of choice for the analysis of cross-linking experiments.

Various instrument types have been used for the analysis of heteroconjugates with ESI-MS: linear ion trap (linear trap quadrupole, LTQ, [73]), triple quadrupole/linear ion trap (Q-trap, [63, 66]), quadrupole time-of-flight (Q-ToF, [71, 74]), Fourier transform ion cyclotron resonance (FT-ICR, [73]), or orbitrap [75].

ESI-MS analysis of cross-linked samples can be done by targeted analysis on a Q-trap instrument. This approach requires two consecutive LC-MS runs. In the first run, the masses of the intact heteroconjugates (precursor masses) are recorded. RNA containing species are distinguished from residual peptides by monitoring the loss of 79 Da (PO_3^-) in negative ion mode. The gathered information is then used for the second run, in which phosphate-containing species are specifically fragmented in positive ion mode. The detection of RNA marker ions (see below), more specifically the nucleobases, triggers the acquisition of a high resolution fragments spectrum [63, 66].

More recently, LC-MS has been carried out with Q-ToF mass spectrometers [71, 74]. Data dependent acquisition (see 1.2.1) on Q-ToF instruments has a lower sensitivity compared to targeted experiments on a Q-trap. Low intensity cross-links might not trigger MS/MS fragmentation and thus cannot be identified. However, this type of analysis does not require two consecutive runs for the same experiment and MS/MS spectra are more informative. In addition, only cross-links producing RNA marker ions upon fragmentation are accessible with the targeted method; this bias does not apply to data dependent acquisition.

FT-ICR or orbitrap instruments acquire data with high mass accuracy and resolution. CID fragmentation is performed in the linear ion trap of the instrument and is disadvantageous for cross-link identification (see 1.2.1.2 and below). Nonetheless, characterization of an isolated microcin with MS/MS in positive and negative ion mode in the LTQ was demonstrated. This species contained a phosphoramidate group covalently linked to a peptide with seven amino acids, an adenosine, and a propylamine [75].

1.3.3.3 RNA marker ions

Upon CID fragmentation, cross-links can produce distinctive marker ions containing intact nucleotides (after neutral loss of water) and the nucleic acid bases (theoretical masses are listed in Table B.1). This is consistent with fragmentation of pure RNA, where cleavage of the N-glycosidic bond to produce nucleobases and of the 5'-P-O bond leading to shortened oligonucleotides is common (1.2.1.4, [76] and references within). However, this is not a reliable criterion to identify spectra of RNA-containing species. The marker ions of uracil are usually of very low intensity, consistent with the observation that uracil is rarely detected after fragmentation of RNA oligonucleotides due to its low proton affinity [76]. Also, intact guanosine has never been observed in previous cross-linking studies. Especially if only one nucleotide is cross-linked, usually no RNA signal is visible in the MS/MS spectrum. Thus in most cases, the MS/MS spectra are dominated by peptide fragments. Then, the cross-linked RNA moiety can only be derived from the difference between calculated peptide and experimental precursor mass. If less complete hydrolysis of the RNA is performed and in consequence the cross-linked RNA oligonucleotide is considerably longer, CID spectra are dominated by RNA fragments since the phosphodiester backbone is more readily cleaved [77].

1.3.3.4 Fragmentation modes: beam-type versus ion trap CID

Collision induced dissociation (CID) on orbitrap instruments is performed in the linear ion trap (see 1.2.1.2). There are substantial differences between beam-type CID as carried out in Q-ToF instruments and ion trap CID fragmentation. In the latter, several collisions with low energy induce fragmentation. In contrast, the kinetic energy of ions during beam-type CID is significantly higher and fragmentation is typically induced by single collisions with higher energy. In consequence, beam-type CID typically produces long y-ion series while only small N-terminal a- and b-ions are observed. Larger a-/b-ions are unstable and readily fragment further. In contrast, ion trap CID usually leads to longer y- and b-ion series.

The physical properties of the ion trap presents two major disadvantages. It cannot trap all fragment ions over the entire mass range, small ions are lost (*low mass cut-off*, see 1.2.1.2). In consequence, detection of important (diagnostic) ions like RNA marker ions, immonium ions and small peptide fragments is prevented. Secondly, the MS/MS fragment spectra are typically recorded with the detectors adjacent to the ion trap. While this process is faster compared to MS/MS acquisition in the orbitrap, it has the disadvantage of a considerably lower resolution and mass accuracy.

In general, ion trap CID was considered disadvantageous for protein–RNA cross-links compared to beam-type CID (Henning Urlaub, unpublished observation). Ion trap CID is likely to cleave off labile modifications (as for example cross-linked nucleotides). The resulting fragment, the intact peptide, is not activated to induce further collisions due to its lower mass. In addition, the low MS/MS mass accuracy and the low mass cut-off inhibit confident cross-link identification. The low mass accuracy prevents unambiguous assignment of fragments as y-, b-, or RNA containing ions (see below). RNA marker ions, which are important for confirmation of the cross-linked RNA, are lost due to the low mass cut-off, thus preventing an important step of cross-link validation.

Orbitrap instruments offer a second fragmentation mode corresponding to beam-type CID, namely higher-energy collision dissociation (HCD). Ions are fragmented in a multipole adjacent to the C-trap and the MS/MS fragment spectrum is recorded in the orbitrap. Therefore, the corresponding spectra do not exhibit the low mass cut-off and benefit from the high mass accuracy of the orbitrap. However, the number of observed fragments was slightly higher on the Q-ToF instrument employed in the comparison, making it the instrument of choice for analysis of protein-RNA cross-linking experiments [78].

In addition to fragmentation techniques based on collisions (CID, HCD), other methods can be applied. In a fragmentation study of a model peptide–RNA heteroconjugate, two alternative fragmentation modes were tested. Both electron capture dissociation (ECD) and electron transfer dissociation (ETD) of the model heteroconjugate led to a higher number of observed peptide fragments compared to ion trap CID. In addition, larger precursor charge states were shown to increase the peptide sequence information after fragmentation. In accordance with observations reported from our laboratory (see 1.3.1), it was concluded that increasing length of the cross-linked oligonucleotide is disadvantageous for fragmentation efficiency [73].

1.3.4 Cross-link identification from mass spectrometry data

Analysis of mass spectrometry data derived from cross-linking experiments is a limiting factor of the approach. The low yield of cross-linked heteroconjugates leads to spectra with low intensity signals, i.e. relatively poor quality in comparison to standard proteomics experiments. The large number of potentially cross-linked RNA oligonucleotides causes additional challenges. The observation of RNA marker ions is largely dependent on the composition of the cross-linked RNA (see 1.3.3.3). Therefore, they cannot be used as diagnostic ions to filter for spectra corresponding to heteroconjugates.

Helpful evidence for the presence of a cross-linked heteroconjugate can be derived from the comparison between the cross-linked sample and a non-irradiated control. Signals observed after UV irradiation but not in the corresponding control are very likely to originate from products of a UV-induced reaction. So far, such comparisons for the identification of cross-linked heteroconjugates have only been done manually (e.g. [71,79]). In contrast, several standard algorithms of quantitative mass spectrometry have been applied successfully to identify RNA binding proteins based on noncross-linked peptides [14,15]. This approach is only feasible after isolation of cross-linked proteins under stringent conditions.

Similarly, evidence for cross-links can be gained by treating one sample with calf intestinal alkaline phosphatase (CIP) while the other remains untreated. Comparison of the treated and untreated samples gives a characteristic shift of 80 Da in MS, corresponding to the loss of HPO_3 upon CIP treatment. Phosphate containing species can then be distinguished from acidic peptides that are co-enriched with IMAC or titanium dioxid. If noncross-linked peptides were separated via SE prior to enrichment, it is unlikely that the loss of HPO_3 is due to residual phosphopeptides. As an additional or alternative indicator, mass shifts due to nuclease treatment can be monitored [62].

In principle, the fractional mass (first digit of the molecular mass) could also be used to distinguish cross-linked heteroconjugates from peptides and oligonucleotides. Peptides contain a higher percentage of atoms with a mass excess, i.e. hydrogen with the mass of 1.0078 Da and nitrogen with 14.0031 Da. In comparison, oligonucleotides contain more oxygen (15.9949 Da) and phosphorus (30.9738 Da) that show a mass deficiency. In consequence, fractional masses of peptides are higher than those of oligonucleotides. Heteroconjugate fractional masses are intermediate as they contain both peptide and oligonucleotide moieties. These differences can be used to distinguish the three species, provided that masses are acquired on instruments with a high mass accuracy [80]. However, this approach has not been integrated in any algorithm for an automated recognition of heteroconjugates.

Previous studies have demonstrated that the mass of a cross-linked heteroconjugate is additive, i.e. the mass of the cross-links corresponds to the sum of the masses of its constituents, peptide and oligonucleotide (e.g. [65]). Exceptions were observed for halopyrimidines where the corresponding hydrohalogen is lost during the reaction. For example, the loss of HBr upon cross-linking of 5-Br-deoxiuridine substituted DNA has been confirmed by mass spectrometry [56].

Another important exception is the observation of an additional mass of 152 Da in cross-links between cysteine-containing peptides and RNA. It was first reported in a cross-linking study of snurportin 1 (SPN1) to U1 snRNA [66]. Various cross-links of the SPN1 C-terminus, differing in length of the cross-linked peptide as well as oligonucleotide, were observed with an additional mass of 152 Da. The same adduct has been observed in cross-links of Sm proteins to both U1 and U2 snRNAs [78] as well as a cross-link of the RNase H-like domain of Prp8 to an RNA resembling the U4/U6 snRNA duplex [81]. Despite various efforts, neither the origin nor the exact composition of the species leading to the observed mass adduct could be unambiguously determined (F. Richter, C. Endler, K.K., U. Zaman, H. Urlaub, Bioanalytical Mass Spectrometry group, unpublished).

Initial computer-aided data analysis strategies compared the experimental precursor masses with the calculated masses of all peptide–oligonucleotide combinations after *in silico* digest of the protein and RNA, respectively [62,65]. However, this might lead to a great number of putative hits, especially for large precursor masses. Therefore, laborious manual evaluation of the results is required. Secondly, the approach is only feasible if protein and RNA sequences of the sample constituents have been identified.

Interpretation of the mass spectrometry data obtained from cross-linking experiments remains challenging and laborious. So far, there has been no suitable software that can handle all effects observed in fragment spectra of cross-linked heteroconjugates and thus can be applied for automated identification.

1.3.5 Application of UV induced cross-linking in combination with mass spectrometry

The earliest cross-linking studies with analysis by mass spectrometry were focused on protein–DNA complexes which will be described briefly before focusing on protein–RNA cross-linking studies. The applied MS techniques mostly correlate to their developments in general. For example, the first studies investigating cross-linked heteroconjugates employed MALDI while the majority of later surveys was based on ESI-MS.

In the early 1990s, the first mass spectrometric measurements of peptide–mononulceoside heteroconjugates were published [82,83]. In the first feasibility study on phosphate-containing heteroconjugates, Jensen *et al.* used MALDI-MS to determine the molecular mass of protein–DNA adducts. They investigated cross-linking of phage T4 gene 32 protein to a (dT)20 oligonucleotide and *E. coli* transcription termination factor rho to the ATP-analogue 4-thio-uridine-diphosphate. For the first time ever, the mass of a cross-linked heteroconjugate was thus measured by mass spectrometry. Remarkably, the masses of the intact protein–DNA adducts were detected without any purification of the sample [84]. In a second study by Bennett *et al.*, *E. coli* uracil-DNA glycosylase was cross-linked to a (dT)20 oligonucleotide and isolated via DEAE and single-stranded DNA agarose chromatography to measure the intact mass. In addition, after irradiation and trypsin digestion, cross-linked peptides were isolated via DEAE chromatography. Subsequent MALDI-MS analysis identified four peptides that cross-linked to (dT)20 [85]. Finally, a chemically synthesized peptide–oligothymidylic acid conjugate was investigated comparing various MALDI matrices and employing ESI-MS/MS to obtain one of the first fragment spectra of a peptide–oligonucleotide heteroconjugate [77].

The first cross-linking studies on protein–RNA complexes employing MALDI-MS were done by Urlaub *et al.* [46,68]. Bacterial 30S ribosomal subunits were UV irradiated, either without or following 2-iminothiolane labeling of lysine residues. RNA with cross-linked proteins was isolated by size exclusion chromatography. Subsequently, the complexes were hydrolyzed by endoproteinases and RNase T1 and the cross-links were purified by RP-HPLC. N-terminal sequencing identified the peptide sequence, while the gap in the sequence corresponded to the cross-linked amino acid. The mass of the intact heteroconjugate was determined by MALDI-MS. The composition of the cross-linked oligonucleotide was derived from the difference between the heteroconjugate and peptide masses. The cross-linked RNA sequence was identified after partial alkaline hydrolysis and treatment with 3' → 5' phosphodiesterase. Analysis of the resulting hydrolysis products yielded a series of signals where the mass differences corresponded to the RNA nucleotides and the oligonucleotide sequence could be deducted by comparison to the primary sequence of 16S ribosomal RNA.

Another very early study on protein–RNA cross-linking focused on the interaction of Human Immunodeficiency Virus type 1 Tat protein with the trans-activation responsive region (TAR). Farrow *et al.* cross-linked a Tat peptide to a model TAR duplex in which a bulged uridine was substituted by 4-thio-uridine. After HPLC purification and proteolytic digestion, they identified several cross-linked arginine-containing peptides by MALDI-MS. As the peptide sequence contained a large number of arginines, they used peptides site-specifically labeled with ^{13}C and ^{15}N containing arginine to identify the actual cross-linked amino acid [86].

Human spliceosomal U1 small nuclear ribonucleoprotein particles (snRNPs) and the reconstituted [15.5K-61K-U4atac snRNA] complex of the minor spliceosome [47] have been studied extensively by cross-linking and mass spectrometry and served as model complexes for method development studies (see above, [62–66,72]).

The first study on cross-linking of spliceosomal complexes using mass spectrometry investigated the human U1 snRNP [64]. Identification of the cross-linked nucleotides was achieved by primer extension analysis. Cross-linking sites appear as discrete stops of the reverse transcriptase one nucleotide upstream to the cross-linked nucleotide, as a small peptide always remains covalently attached to the latter. To determine which protein is cross-linked, an immunoprecipitation step was integrated into the workflow prior to primer extension. The sequence of the cross-linked peptide was identified by N-terminal sequencing after preparative purification of the heteroconjugates. The cross-linked RNA was identified with MALDI-MS following the same workflow established for ribosomes (see above).

After UV irradiation of the reconstituted [15.5K-61K-U4atac snRNA] complex followed by size exclusion and microbore chromatography, the chymotrypsin specific 61K (hPrp31p) peptide SSTSVLPHTGY (S263–Y273) was found to be cross-linked to an oligonucleotide 5'-CAUAG-3' (C42–G46) of U4atac solely by the precursor mass measured by MALDI in positive reflectron mode. Fragmentation by post-source decay (PSD) confirmed the RNA sequence and identified U44 as the cross-linked nucleotide. The peptide sequence could only be obtained by fragmentation of a different precursor of the same peptide cross-linked to an AU dinucleotide. As mentioned above, increasing length of the cross-linked oligonucleotide leads to stronger interference with the detection of peptide fragments. Observation of a histidine-uridine heteroconjugate identified H270 as the cross-linked amino acid. This cross-linking site had been identified in previous studies by N-terminal sequencing and MALDI in positive linear mode [47]. Following the same experimental workflow for U1 snRNP, the 70K peptide RVLVDVER (R173–R180) was found to be cross-linked to the U1 snRNA RNase T1 specific fragment 5'-AUCACG-3' (A29–G34). MALDI-PSD proved U30 to be the cross-linked nucleotide. Furthermore, from the lack of the corresponding peptide fragment ion y6 and a signal corresponding to the next peptide fragment ion in the series, y7, additionally carrying a uracil base, it was concluded that L175 could be the cross-linked amino acid [65].

In a study of a partial complex of human spliceosomal U2 snRNP, Kühn-Hölsken *et al.* identified a contact site between the U2-specific protein p14/SF3b114a and the region of U2 snRNA interacting with the branch-site of the pre-mRNA [62]. This is in excellent agreement with previous studies that demonstrated direct contact of p14/SF3b14a with nucleotide G31, which is next to the branch-site interacting region of U2 snRNA (G33–A38) [87].

Further investigations of U1 snRNP bound to the nuclear import factor snurportin 1 (SPN1) combined UV induced protein–RNA cross-linking with IMAC enrichment. MS analysis was done by both MALDI-ToF(-ToF) and the targeted approach described in 1.3.3.2 after LC-ESI-MS. The C-terminal peptide of SPN1 was shown to directly interact with stem loop III of U1 snRNA [66].

Several recent studies have successfully combined titanium dioxide enrichment and ESI-LC-MS/MS analysis on a Q-ToF instrument: The structure of the NusB–S10 transcription antitermination complex of *E. coli* was studied by Luo *et al.* Combining their structure as determined by X-ray crystallography and the results of UV induced cross-linking of NusB–S10 to an *rrn* BoxA and a γNutR BoxA containing oligonucleotide enabled mapping of the RNA-binding surface of the NusB–S10 complex [71]. Ghalei *et al.* studied box C/D sRNPs of *Pyrococcus furiosus* which catalyze 2'-O-methylation of ribosomal RNA. Based on the results of cross-linking experiments analyzed by mass spectrometry, they identified the AFLR motif of the Nop5 protein directly binding to the RNA in and around stem II of the box C/D or C'/D' motifs. Electrophoretic gel mobility shift assays confirmed that deletion of the respective regions in Nop5 or RNA prevents complete formation of the sRNPs [74]. Mozaffari-Jovin *et al.* studied the interactions of the RNase H-like domain of the spliceosomal protein Prp8 with an RNA construct resembling a truncated U4/U6 snRNA duplex with cross-linking and mass spectrometry. They identified two amino acids of the RNase H-like domain that directly interact with RNA. Together with complementary experiments, including identification of the cross-linking site on the RNA by primer extension analysis, the results permitted modeling of the position of the U4/U6 duplex on the three-dimensional structure of the Prp8 RNase H-like domain [81].

Bley *et al.* studied the interaction of the telomerase RNA binding domain (TRBD) of the telomerase reverse transcriptase protein with an RNA fragment resembling the three-way helical junction CR4/5. The TRBD was expressed as a fusion protein with the maltose-binding protein and a His_6 tag. The RNA was randomly labeled with 5-iodo-uridine during *in vitro* transcription. After UV induced cross-linking at 302 nm, the fusion protein was isolated under denaturing conditions from noncross-linked RNA via the His-tag. After proteolysis, the peptide–RNA heteroconjugates were purified by gel electrophoresis. The RNA was hydrolyzed with RNases A or T1; the mixture was then subjected to MALDI-ToF-MS. Comparison of the mass spectrum to a non-irradiated control containing unlabeled RNA revealed signals corresponding to cross-linked heteroconjugates. The composition of the cross-link was derived from calculating the masses of all possible combinations of tryptic peptides and RNase specific oligonucleotides. The sample was then treated with ammonia and phosphatase to yield peptides covalently linked to single uridines. Subsequently, the peptide sequence was confirmed and the cross-linked residue identified by MALDI-ToF/ToF analysis after collision induced dissociation [79].

1.4 Objectives

Various experimental strategies have been developed for the identification of cross-linking sites on a peptide or amino acid level by mass spectrometric analysis. These have been applied to different purified native or *in vitro* reconstituted protein–RNA complexes. However, the complexes studied so far have been relatively simple, comprising few proteins and mostly single RNAs.

While studies of protein–RNA interactions *in vitro* yield valuable information of direct contact sites, they cannot account for all effects that influence interactions in the complex environment of the cell. DNA sequencing techniques allow investigation of the contact sites on the RNA level after UV induced cross-linking *in vivo*. No complementary methods allow the identification of contact sites on the protein level at a comparable resolution, i.e. peptides or amino acids.

To close the gap between relatively simple protein–RNA complexes and an unbiased identification of interaction sites on a peptide or even amino acid level by mass spectrometry after *in vivo* cross-linking, substantial improvements to the approach are necessary. The method faces two major challenges: the generally low yield of the UV induced reaction and a lack of data analysis tools that allow unbiased searches for cross-linked heteroconjugates.

The low cross-linking yield can be partially overcome by incorporation of photoreactive nucleotides *in vitro* or *in vivo*. However, the mass of the cross-linking product and its fragmentation behavior during MS analysis has to be known. Therefore, one aim of this project was the detailed investigation of one candidate for RNA labeling, 4-thio-uracil, in a simple test system.

Due to the rapid growth in the number of mass spectrometry applications, especially in proteomics, MS instrumentation has seen major improvements in the past decade. Sensitivity, resolution, mass accuracy, and sequencing speed have been increased significantly. Since the constant improvements might also affect the MS based identification of protein–RNA cross-links, another aim of this study was the evaluation of emerging instrumental advancements. One instrument in particular, the LTQ Orbitrap Velos, was to be tested extensively. It was reported to exhibit a significant increase in sensitivity in proteomics applications, especially for a particular type of fragmentation (HCD). The question was whether these improvements would also be beneficial for the mass spectrometric analysis of cross-linked heteroconjugates.

The second major challenge in investigation of UV induced protein–RNA cross-linking is data analysis. Existing manual spectra interpretation or computer-aided strategies were limited to a low number of protein sequences. Without a suitable data analysis approach, cross-links cannot be identified in extended ribonucleoprotein complexes or even entire cells. Therefore, the major aim of this PhD project was the development of data analysis workflows that would allow automated identification of cross-links in searches against larger protein databases and eventually entire proteomes.

These method development strategies had to be evaluated in biological systems, for which appropriate test systems had to be found. In addition, specific questions raised in two collaborations were to be resolved with UV cross-linking combined with MS. Experimental strategies for sample preparation, cross-linking, enrichment, and MS analysis had to be evaluated and, if necessary, optimized to allow identification of cross-links in the respective ribonucleoprotein complexes.

2 Materials and Methods

2.1 Materials

2.1.1 Chemicals and solvents

acetonitrile, LiChrosolv	Merck, Darmstadt, Germany
acetonitrile, Chromasolv	Sigma-Aldrich, Steinheim, Germany
agarose	Serva, Heidelberg, Germany
ampicillin	Roth, Karlsruhe, Germany
Bacto agar	Becton, Dickinson and Company, Sparks, MD, USA
Bacto peptone	Becton, Dickinson and Company, Sparks, MD, USA
Bacto yeast extract	Becton, Dickinson and Company, Sparks, MD, USA
bromophenol blue	Merck, Darmstadt, Germany
Coomassie Brilliant Blue G-250	Sigma-Aldrich, Steinheim, Germany
DHB	Aldrich, Steinheim, Germany
EGTA	Roth, Karlsruhe, Germany
formic acid	Sigma-Aldrich, Steinheim, Germany
glucose	Merck, Darmstadt, Germany
kanamycin	Roth, Karlsruhe, Germany
LB-medium capsules	MP Biomedicals, Illkirch, France
methanol, LiChrosolv	Merck, Darmstadt, Germany
α-lactose	Merck, Darmstadt, Germany
TFA	Roth, Karlsruhe, Germany
trypton	Merck, Darmstadt, Germany
urea	Sigma-Aldrich, Steinheim, Germany
water, LiChrosolv	Merck, Darmstadt, Germany
xylene cyanol FF	Sigma-Aldrich, Steinheim, Germany
YEP Broth	ForMedium, Norfolk, UK

All other chemicals/solvents were obtained from Fluka, Merck, Sigma-Aldrich or Roth in p.a. (pro analysis) grade. All buffers were prepared with water purified by deionization and filtration in a Milli-Q Biocel century system equipped with a Millipak 0.22 µm filter (both Millipore, Merck, Darmstadt, Germany). Buffers were sterilized with bottle top or syringe filters.

2.1.2 Commercial buffers and solutions

acrylamide for protein gels	Rotiphorese Gel 30 (30% acrylamide, 0.8% bis-acrylamide); Roth, Karlsruhe, Germany
Bradford staining solution	Bio-Rad Protein Assay; Bio-Rad, Munich, Germany
ECL solutions	Amersham ECL Western Blotting Detection Reagents; GE Healthcare, Munich, Germany
ethidium bromide	Roth, Karlsruhe, Germany
PCI solution	Roti phenol/chloroform/isoamyl alcohol (25:24:1); Roth, Karlsruhe, Germany
Pfu buffer (10x)	Promega, Mannheim, Germany
PNK buffer (10x)	New England Biolabs, Frankfurt, Germany
protein marker	Precision Plus Protein Standards (Unstained or All Blue); Bio-Rad, Munich, Germany

2.1.3 Buffers

AGK	10 mM Hepes pH 7.9, 1.5 mM $MgCl_2$, 50 mM KCl, 10% glycerol
CBB	25 mM Tris pH 7.9, 150 mM NaCl, 1 mM $Mg(OAc)_2$, 1 mM imidazole, 2 mM $CaCl_2$, 2 mM DTT
CEB	25 mM Tris pH 7.9, 150 mM NaCl, 1 mM $Mg(OAc)_2$, 1 mM imidazole, 25 mM EGTA, 0.02% NP40, 2 mM DTT
CE	10 mM cacodylic acid pH 7.0, 0.2 mM EDTA pH 8.0
PCR buffer, 10x	100 mM Tris pH 8.7, 500 mM KCl, 25 mM $MgCl_2$
SDS running buffer, 1x	25 mM Tris, 192 mM glycine, 0.1% SDS
SDS sample buffer	60 mM Tris, 1 mM EDTA, 16% glycerine, 2% SDS, 0.1% bromophenol blue, 50 mM DTT
TBE, 1x	0.1 M boric acid, 0.1 M Tris, 2 mM EDTA, pH 8.3

2.1.4 Enzymes and enzyme inhibitors

benzonase	benzonase nuclease; Novagen, Merck, Darmstadt, Germany
lyticase	Sigma-Aldrich, Steinheim, Germany
phusion DNA polymerase	15 ng/µl; Department of Cellular Biochemistry
PreScission protease	10 mg/ml; Department of Cellular Biochemistry
protease inhibitors	EDTA free Protease Inhibitor Cocktail tablets; Roche, Mannheim, Germany
RNase A	RP A grade; Ambion, Applied Biosystems, Darmstadt, Germany
RNase T1	biochemistry grade; Ambion, Applied Biosystems, Darmstadt, Germany
rRNasin RNase inhibitor	Promega, Mannheim, Germany
T4 polynucleotide kinase	New England Biolabs, Frankfurt, Germany
Taq DNA polymerase	4.8 mg/ml; Department of Cellular Biochemistry
trypsin	sequencing grade modified trypsin; Promega, Mannheim, Germany

2.1.5 Proteins, peptides and (oligo)nucleotides

ATP, [γ-^{32}P]-labeled	PerkinElmer, Waltham, MA, USA
BSA standards	Bradford: Albumin Standard; Thermo Fisher Scientific, Schwerte, Germany LC-MS: BSA Protein Digest Standard; Protea Biosciences, Morgantown, WV, USA
carrier DNA	DNA, MB-grade from fish sperm; Roche, Mannheim, Germany
DNA primers	Eurofins MWG Operon, Ebersberg, Germany
dNTPs	New England Biolabs, Frankfurt, Germany
GluFib	[Glu^{1}]-Fibrinopeptide B, human, synthetic; Aldrich, Steinheim, Germany
RNA oligonucleotides	Dharmacon, Thermo Fisher Scientific, Epsom, UK

2.1.6 Antibodies

anti-rabbit IgG	Peroxidase-conjugated AffiniPure Goat Anti-Rabbit IgG; Jackson ImmunoResearch, West Grove, PA, USA
Peroxidase Anti-Peroxidase antibody	P-2026; Sigma-Aldrich, Steinheim, Germany
TAP tag antibody	CAB1001; Pierce, Thermo Fisher Scientific, Schwerte, Germany

2.1.7 Other materials

bottle top filters	Filtropur BT50 0.2; Sarstedt, Nümbrecht, Germany
C18 column material	C18 AQ 120 Å 5 µm or 3 µm; Dr Maisch GmbH, Ammerbuch, Germany
Calmoduline beads	Calmodulin Affinity Resin; Agilent, Böblingen, Germany
chromatography columns (gravity flow)	Econo-Pac Chromatography Columns; Bio-Rad, Munich, Germany
dialysis cassettes	Slide-A-Lyzer 3.5K; Pierce, Thermo Fisher Scientific, Rockford, IL, USA
film (autoradiography)	Carestream Kodak BioMax MR Film; Sigma-Aldrich, Steinheim, Germany
film (ECL)	High Performance Chemiluminescence Film; GE Healthcare, Munich, Germany
glutathione sepharose	GE Healthcare, Munich, Germany
IgG beads	IgG Sepharose 6 Fast Flow; GE Healthcare, Munich, Germany
microtiter plates for cross-linking	black polypropylene 96 well microplates (# 655209); Greiner Bio-One, Frichenhausen, Germany
Ni-NTA agarose	Qiagen, Hilden, Germany
RNA isolation columns	MicroSpin G-25; GE Healthcare, Munich, Germany
syringe filters	Filtropur S 0.2; Sarstedt, Nümbrecht, Germany
titanium dioxide column material	titansphere 5 µm; GL Sciences, Tokyo, Japan
WB nitrocellulose membrane	Protran BA 83; GE Healthcare, Munich, Germany

2.1.8 Instruments and laboratory equipment

autoclaves	Varioklav steam sterilizer H + P; Thermo Fisher Scientific, Schwerte, Germany
centrifuges	benchtop centrifuges: eppendorf centrifuge 5415R; Eppendorf, Hamburg, Germany Heraeus Fresco 17 centrifuge Heraeus Biofuge pico Heraeus Megafuge 1.0R superspeed centrifuge: Sorvall Evolution RC ultracentrifuge: Sorvall Discovery 90SE large capacity centrifuge: Sorvall RC 12BP+ Heraeus/Sorvall: Thermo Fisher Scientific, Schwerte, Germany
clean bench	HeraSafe; Heraeus, Thermo Fisher Scientific, Schwerte, Germany
cross-linking apparatus	build in-house, operated with four 8 W lamps 254 nm: G8T5; Sankyo Denki, Japan 365 nm: F8T5BL; Sankyo Denki, Japan
film developer	Kodak X-OMAT 2000 Processor; Carestream, Stuttgart, Germany
gel documentation	Gel Doc 2000; Bio-Rad, Munich, Germany
gel electrophoresis	Mini-PROTEAN Tetra system Mini-SUB CELL GT both Bio-Rad, Munich, Germany
LC-MS	see 2.2.9
PCR thermo cycler	T3 Thermocycler; Biometra, Göttingen, Germany
phosphorimager	Typhoon 8600; GE Healthcare, Munich, Germany
rotors	Sorvall SS-34 T-865 H-12000 all Thermo Fisher Scientific, Schwerte, Germany
scintillation counter	Tri-Carb 2100TR; Beckmann, USA
spectrophotometer	Ultrospec 3000 pro; Amersham Pharmacia Biotech, Cambridge, UK Biophotometer; Eppendorf, Hamburg, Germany
thermomixers	Thermomixer comfort ThermoStat plus both Eppendorf, Hamburg, Germany
ultra centrifugal mill	ZM 200; Retsch, Haan, Germany
WB transfer cell	Mini Trans-Blot Electrophoretic Transfer Cell; Bio-Rad, Munich, Germany

2.2 Methods

If not noted otherwise, methods were according to standard protocols [88] with modifications as described.

2.2.1 Media and plates for cell cultures

All media and solutions for plates were prepared with deionized water and sterilized by autoclaving (121°C, 15 min, 15 psi).

2.2.1.1 LB medium

LB (lysogeny broth) medium contained 1% tryptone, 0.5% yeast extract, and 1% NaCl (all w/V).

2.2.1.2 YPD medium

YPD (yeast extract, peptone, dextrose) medium contained 1% yeast extract, 2% peptone, and 2% glucose (all w/V). For YPD agar, 2% (w/V) Bacto agar was added prior to autoclaving.

For large-scale yeast fermentation, YPD was prepared from YEP Broth by autoclaving and addition of glucose (filtered sterile) to a final concentration of 2% (w/V).

2.2.1.3 YMM dropout medium

YMM (yeast minimum media) dropout contained 0.67% yeast nitrogen base without amino acids, 2% glucose, and 0.2% drop-out powder (all w/V), 2% agar was added for YMM plates. The pH was adjusted by addition of NaOH. Drop-out powder for –URA selective plates contained 2 g each of the following: adenine, alanine, arginine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, histidine, isoleucine, lysine, methionine, phenylalanine, proline, serine, threonine, tyrosine, tryptophan, and valine. In addition, it contained 4 g leucine. The powder was ground for complete mixing.

2.2.1.4 Auto-inducing medium

Auto-inducing medium ZYM-5052 was prepared according to [89]. The 1000x trace metal solution was a kind gift of Dr. Sunbin Liu (Department of Cellular Biochemistry). Other stock solutions were prepared as listed below and sterilized by autoclaving. Medium was prepared as listed below under a clean bench.

stock solutions

solution	compound	concentration
ZY	trypton	1%
	yeast extract	0.5%
50x 5052	glycerol	25%
	glucose	25%
	α-lactose	10%
50x salt M	Na_2HPO_4	1.25 M
	KH_2PO_4	1.25 M
	NH_4Cl	2.5 M
	Na_2SO_4	0.25 M

for 1 l auto-inducing medium

solution	volume	compound	final concentration
ZY	956 ml	trypton	1%
		yeast extract	0.5%
1 M $MgSO_4$	2 ml	$MgSO_4$	2 mM
50x 5052	20 ml	glycerol	0.5%
		glucose	0.05%
		α-lactose	0.2%
50x salt M	20 ml	Na_2HPO_4	25 mM
		KH_2PO_4	25 mM
		NH_4Cl	50 mM
		Na_2SO_4	5 mM
1000x trace metals	200 µl		0.2x

2.2.2 Standard molecular biology methods

2.2.2.1 Agarose gel electrophoresis of DNA fragments

Agarose gel electrophoresis was carried out for separation and visualization of DNA fragments. Agarose (1.2% w/V) was dissolved in 60 ml 0.5x TBE by heating. For later visualization, 3 µl ethidium bromide (1% w/V) were added to the solution while it was cooling. DNA samples were mixed with 6x DNA gel-loading buffer and fractionated at a constant voltage of 150 V in 0.5x TBE as running buffer. DNA was visualized by UV illumination.

6x DNA gel-loading buffer

0.25%	bromophenol blue
0.25%	xylene cyanol FF
15%	Ficoll (Type 400; Pharmacia)
	H_2O

2.2.2.2 PCI extraction

Phenol-chloroform-isoamylalcohol extraction was used to separate DNA/RNA from proteins. The sample was mixed with one volume PCI solution by vigorous shaking. Phases were separated by centrifugation at 13 000 rpm for 5 min. The upper aqueous phase, containing RNA/DNA, was transferred into a clean microfuge tube. Optionally, the aqueous phase was again extracted by addition of one volume chloroform, vigorous shaking and phase separation by centrifugation as above. RNA/DNA was isolated from the aqueous phase by ethanol precipitation.

2.2.2.3 Ethanol precipitation

Proteins, RNA/DNA or protein–RNA complexes were precipitated by addition of 2.5 to 3 volumes ethanol and 1/10 volume 3 M NaOAc and incubation at –20°C for at least 30 min. Macromolecules were pelleted by centrifugation at 13 000 rpm and 4°C for 30 min. The pellet was washed with 80% ethanol and centrifuged as above. The supernatant was discarded and the pellet was dried in a centrifugal evaporator.

2.2.2.4 5' labeling of RNA

RNA oligonucleotides were 5' labeled with [γ-^{32}P]-ATP and T4 polynucleotide kinase (PNK) by incubation of the reaction mixture for 70 min at 37°C.

reaction mixture for 5' RNA labeling

1.5 µl	RNA oligonucleotide	5 pmol
1.5 µl	H_2O	
1 µl	10x PNK buffer	
5 µl	[γ-^{32}P]-ATP	8.3 pmol
1 µl	T4 polynucleotide kinase	

After incubation, 40 µl CE-buffer were added to the mixture. Free [γ-^{32}P]-ATP was removed by isolation with MicroSpin G-25 columns (GE Healthcare, Munich, Germany), used according to the manufacturer's protocol. The crude product was further purified by PCI extraction after adjusting the volume to 200 µl by addition of 150 µl CE buffer. The RNA pellet was redissolved in CE buffer.

2.2.3 Standard protein biochemical methods

2.2.3.1 Determination of protein concentration

Protein concentrations were determined with the method originally developed by Bradford [90]. It is based on the absorption maximum shift of Coomassie Brilliant Blue G-250 from 465 to 595 nm when the dye binds to protein in acidic solution. The commercially available Protein Assay (Bio-Rad) was used according to the manufacturer's protocol.

BSA protein standards were prepared, typically eight standards ranging from 0 to 15 µg/ml final concentration. The sample was diluted so that the final concentration was within the concentration range of the standards. The standard curve and the sample concentration were calculated by the spectrophotometer.

2.2.3.2 Denaturing polyacrylamide gel electrophoresis (SDS-PAGE)

Denaturing polyacrylamide gel electrophoresis was used for separation and visualization of proteins by Coomassie staining or Western blotting. Gels for SDS-PAGE were prepared and run in a Mini-PROTEAN Tetra system (Bio-Rad, Munich, Germany). Typically, gels with a 5.5% stacking gel and a 15% separating gel were prepared following the recipe below. Samples were mixed with SDS sample buffer (1:1 V/V) and heated to 95°C for 5 min prior to loading. Gels were run at 30 mA per gel with 1x SDS running buffer.

4x buffer for stacking and resolving gels

	Tris	SDS	pH
stacking gel	1.5 M	4%	8.8
resolving gel	0.5 M	4%	6.8

	5.5% stacking gel	15% separating gel
H_2O	2.95 ml	2.5 ml
30% acrylamide solution	950 µl	5 ml
4x buffer	1.25 ml	2.5 ml
10% APS	20 µl	35 µl
TEMED	20 µl	35 µl

2.2.3.3 Colloidal Coomassie staining

Proteins separated by SDS-PAGE were stained with colloidal Coomassie [91] over night and destained by several rinses with water. Colloidal Coomassie was prepared with water and methanol in LiChrosolv quality.

Colloidal Coomassie	
Coomassie Brilliant Blue G-250	0.08% (w/v)
phosphoric acid (conc.)	1.6% (v/v)
ammonium sulfate	8% (w/v)
methanol	20% (v/v)

2.2.3.4 Western blotting for immunodetection of proteins

For immunodetection of proteins by specific antibodies, the protein sample was first separated by SDS-PAGE. Proteins were then transferred onto nitrocellulose membranes by electrophoresis (1 h at 65 V and 4°C for 1 mm gels) in Western transfer buffer (20 mM Tris, 150 mM glycine). The membrane was washed with TBS-T (50 mM Tris pH 7.5, 150 mM NaCl, 0.1% Tween-20) and blocked with 5% (w/V) fat-free milk powder in TBS-T at 4°C over night. The membrane was then incubated with primary antibody (1:500) in 5% milk/TBS-T for 1 h at room temperature. The membrane was washed with TBS-T (5x 10 min) before incubation with the secondary antibody (typically 1:60 000) in 5% milk/TBS-T for 1 h at room temperature. The membrane was washed with TBS-T as mentioned above and immunodetected proteins were visualized by enhanced chemiluminescence according to the manufacturer's protocol (Amersham ECL Western Blotting Detection Reagents; high performance chemiluminescence film; both GE Healthcare, Munich, Germany).

2.2.4 Expression and isolation of the NusB–S10 protein complex

The NusB–S10 complex was purified following the published protocol [71] with slight modifications. A glycerol stock of an *E. coli* BL21(DE3) strain containing plasmids encoding for NusB and S10 was kindly provided by Xiao Luo (Strukturbiochemie, Prof. Markus C. Wahl, Freie Universität Berlin).

The *E. coli* strain was grown in autoinducing medium in the presence of 100 µg/ml ampicillin and 25 µg/ml kanamycin to an OD_{600} of 0.5 at 37°C and subsequently over night at 20°C. Cells were harvested by centrifugation at 5 000 rpm and 4°C for 30 min. Cell pellets were washed once with binding buffer (50 mM Tris pH 7.5, 150 mM NaCl), centrifuged as above, and resuspended in binding buffer supplemented with protease inhibitors. Cells were disrupted by sonication and cell debris were pelleted at 15 000 rpm and 4°C for 30 min.

Glutathione sepharose was equilibrated with binding buffer. The NusB–S10 complex was trapped on glutathione sepharose through the N-terminal GST tag of S10 by incubation at 4°C for 3 h. The protein complex was eluted with binding buffer supplemented with 15 mM reduced glutathione. The eluate was incubated with PreScission protease (1 mg/ml; 1:100 w/w) at 4°C over night to cleave off the GST tag. In the second purification step, the complex was trapped on Ni-NTA agarose via the N-terminal His_6 tag of NusB. Ni-NTA agarose was pre-equilibrated with binding buffer containing 20 mM imidazole prior to incubation with the sample for 90 min at 4°C. Beads were washed with binding buffer containing 50 mM imidazole and the protein complex was eluted with binding buffer supplemented with 500 mM imidazole. The eluate was concentrated by centrifugation.

In the final isolation step, the protein complex was injected onto a gel filtration column (Superdex 75 10/300 GL; GE Healthcare, Munich, Germany). Eluent was binding buffer with 2 mM DTT. Coomassie stained SDS-PAGE gels of the protein fractions were used to choose fractions where NusB and S10 were present in a 1:1 ratio.

2.2.5 Introduction of a C-terminal TAP tag to the yeast protein Cbp20 by homologous recombination

The applied protocol described in the sections below follows established procedures [92] with slight modifications unless noted otherwise. Dr. Kum-Loong Boon (Department of Cellular Biochemistry) gave technical support in the experiments.

2.2.5.1 Generation of DNA

DNA template was the pBS1539 plasmid which was constructed to introduce a C-terminal TAP tag and contains a *URA3* selective marker from *Kluyveromyces lactis* [93]. The pBS1539-psc plasmid used here, which contains a PreScission instead of the TEV cleavage site, was provided by Dr. Kum-Loong Boon.

The TAP cassette was amplified by polymerase chain reaction (PCR). Primers, reaction mix and PCR program are listed below. In the primer sequences, regions homologous to the pBS1539-psc plasmid are underlined. The 5' ends are homologous to the target gene *CBP20*.

forward primer:
5'-TCA GAC CAG GTT TCG ATG AAG AAA GAG AAG ATG ATA ACT ACG TAC CTC AGT CCA TGG AAA AGA-GAA GAT-3'

reverse primer:
5'-TAT ATA TAT ATC TGT GTG TAG AAT CTT TCT CAG ATA TAA ATT-GAT TGA TTT ACG ACT CAC TAT AGG GCG A-3'

250 µl PCR mix	
25 µl	10x Pfu buffer
20 µl	dNTPs (2.5 mM)
1.25 µl	forward primer (100 µM)
1.25 µl	reverse primer (100 µM)
195 µl	H_2O
5 µl	DNA (pBS1539-psc)
2.5 µl	phusion DNA polymerase

PCR program		
94°C	3 min	
94°C	30 s	
50°C	50 s	34 cycles
72°C	2.5 min	
72°C	3 min	

Formation of PCR product was confirmed by agarose gel electrophoresis. The PCR product was isolated with phenol-chloroform extraction. To this end, 240 µl PCI solution (1:1 V:V) were added and mixed by vortexing. Phases were separated by centrifugation at 13 000 rpm for 10 min. DNA was precipitated from the aqueous phase with 2.5 volumes of ethanol and 1/10 volume of 3 M NaOAc. The DNA was pelleted by centrifugation, air dried, and subsequently dissolved in 35 µl H_2O.

2.2.5.2 Transformation

To construct the yeast strain expressing TAP tagged Cbp20 (Cbc2p), the PCR product containing the C-terminal TAP tag cassette was transformed into yeast strain BJ2168 with the lithium acetate (LiOAc) method [94, 95].

Transformation mix was prepared by mixing 35 µl DNA solution, 36 µl 1 M LiOAc and 240 µl PEG$_{3350}$ solution (50% w/V), and 40 µl denatured fish sperm carrier DNA (2 mg/ml; DNA, MB-grade from fish sperm, Roche, Mannheim, Germany).

Competent yeast cells for transformation were prepared from a 50 ml overnight culture grown to an OD$_{600}$ of 0.6-1.0. Cells were spun down by centrifugation at 4 000 rpm and 4°C for 3 min, subsequently washed with 1 ml H_2O and centrifuged as above. Cells were resuspended in 400 µl 100 mM LiOAc.

50 µl cell suspension were incubated with the transformation mix on a rotating wheel at room temperature for 30 min, followed by a heat shock at 42°C for 20 min. Cells were pelleted by brief centrifugation, the transformation mix was removed and cells were resuspended in 125 µl H_2O. Cells were plated on –URA selective plates and incubated for 2-3 days at 30°C. The transformants were restreaked onto a fresh –URA selective plate for further validation.

2.2.5.3 Yeast colony PCR

Correct insertion of the TAP tag construct into the yeast strain was confirmed by yeast colony PCR. Forward primer was the same as above, homologous to the chromosomal sequence and the inserted TAP cassette (latter underlined in primer sequence). The reverse primer was homologous to the ProtA sequence (underlined). PCR products were verified by agarose gel electrophoresis.

forward primer:
5'-TCA GAC CAG GTT TCG ATG AAG AAA GAG AAG ATG ATA ACT ACG TAC CTC AGT CCA TGG AAA AGA-GAA GAT-3'

reverse primer:
5'-CCT TAA ATC AGG TTG ACT TCC CCG CGC A-3'

PCR mix (14 samples)	
36 µl	10x PCR buffer
57.6 µl	dNTPs (2.5 mM)
3.6 µl	forward primer (100 mM)
3.6 µl	reverse primer (100 mM)
247.2 µl	H_2O
3 µl	Taq
3 µl	Phusion

PCR program		
94°C	6 min	
94°C	20 s	
50°C	50 s	34 cycles
72°C	2.5 min	
72°C	3 min	

2.2.5.4 Confirmation of TAP tag inclusion by Western blot

Yeast clones confirmed by yeast colony PCR were further investigated by Western blotting. For sample preparation, 2 ml overnight cultures cultivated in YPD broth were harvested by centrifugation at 3 500 rpm and 4°C for 4 min. Cell pellets were resuspended in 500 µl 0.2 M NaOH and incubated on ice for 10 min. Subsequently, 27.5 µl TCA (100% w/V) were added and the cell lysate was further incubated on ice for 10 min. Proteins were spun down at 13 000 rpm for 30 s. Protein pellets were resuspended in 35 µl dissociation buffer (0.1 M Tris pH 6.8, 4 mM EDTA pH 8.0, 4% SDS, 20% glycerol, 20 mM DTT). After addition of 15 µl 1 M Tris, the sample was boiled at 95°C for 10 min. Cell debris were removed by centrifugation (10 s at 13 000 rpm). Proteins were separated by SDS-PAGE, and the TAP tagged protein was detected by Western blotting by peroxidase anti-peroxidase antibody and visualized by enhanced chemiluminescence (ECL).

2.2.5.5 Confirmation of TAP tag inclusion by sequencing

The yeast strain confirmed to express TAP tagged protein by Western blotting was further verified by DNA sequencing. Yeast DNA for PCR prior to sequencing was prepared from 50 ml yeast culture cultivated in YPD. Cells were harvested at 4 000 rpm and 4°C for 3 min. Cell pellets were washed once with 10 ml deionized water, centrifuged as above and resuspended in 10 ml SE buffer (0.9 M sorbitol, 0.1 M EDTA pH 8.0). 50 µl lyticase (20 mg/ml) were added and the cell suspension was incubated for 30–60 min at room temperature. Cells were spun down at 5 000 rpm for 5 min, resuspended in 500 µl lysis buffer (0.1 M Tris pH 8.0, 50 mM EDTA, 1% SDS) and 32 µl 4 M NaCl were added. In order to break the cells, glass beads were added to the cell suspension and the sample was vortexed for 1 min. Cell debris and glass beads were removed by centrifugation at 5 000 rpm for 5 min. DNA was isolated by PCI extraction and ethanol precipitation.

Two PCRs were prepared as described in 2.2.5.1 with the same forward primer. In the first PCR, the same reverse primer as in 2.2.5.3 was used. The reverse primer for the second PCR is listed below. It is homologous to the *URA3* sequence in the TAP cassette. PCR products were visualized by agarose gel electrophoresis and sequenced (SEQLAB Sequence Laboratories, Göttingen, Germany). The obtained sequencing results showed no mutations in the coding region.

reverse primer 2:
5'-AGA GAA TCA GCG CTC CCC AT-3'

2.2.6 Yeast cell culture and extract preparation

A cell culture of the yeast strain containing a C-terminal TAP tag on *CBP20* was grown in a 150 l fermenter (INFORS-HT, Bottmingen, Switzerland) by Thomas Conrad (Bioreactor Facility, Department of Cellular Biochemistry). Cells were inoculated in YPD in the presence of 50 mg/l ampicillin and 10 mg/l tetracycline to an OD_{600} of 5.7. Cells were washed once with water, collected in a nozzle separator (GEA Westfalia Separator Group, Oelde, Germany) and harvested by centrifugation at 4 500 rpm and 4°C for 10 min.

Cell pellets were resuspended in 0.7 volumes AGK buffer and cell droplets were flash frozen with liquid nitrogen. Cell beads were ground in an ultra centrifugal mill and cell debris were pelleted by centrifugation at 17 000 rpm and 4°C for 30 min in a SS-34 rotor. Optionally, polysomes were pelleted by ultracentrifugation at 37 000 rpm at 4°C for 60 min in a T-865 rotor. Cell extracts were flash frozen in liquid nitrogen and stored at –80°C.

2.2.7 TAP tag purification

The original protocol for TAP tag purification [93] was further optimized and simplified by Dr. Kum-Loong Boon to the TAP tag with PreScission cleavage site.

TAP tag purification for cross-linking and subsequent MS analysis was typically done with 10 ml yeast extract, corresponding to about 350 mg of protein. The first step of TAP tag affinity purification was performed with IgG beads and elution by PreScission protease cleavage of the ProteinA part of the TAP tag. In the second purification step, complexes were trapped on Calmoduline beads via the Calmoduline binding peptide part of the TAP tag.

300 µl IgG beads (600 µl bead suspension) were equilibrated with 5 ml AGK prior to addition of sample. Protein–RNA complexes were bound by incubation at 4°C for 2 h. The IgG column was washed with 20 ml CBB. Complexes were released from IgG by incubation with 12 µl PreScission protease in 2 ml CBB supplied with 1 µl RNasin at 4°C over night. The sample was eluted by gravity flow and addition of 1 ml CBB.

200 µl calmoduline beads (400 µl bead suspension) were equilibrated with 5 ml CBB. The sample was incubated with the beads at 4°C for 1 h. The beads were washed with 20 ml CBB. The sample was eluted by incubation with 1 ml CEB for 5 min, elution was repeated twice in total.

2.2.8 UV induced protein–RNA cross-linking

2.2.8.1 Cross-linking of labeled RNA and visualization of cross-linking products by SDS-PAGE

Cross-linking of proteins to RNA previously labeled with [γ-^{32}P]-ATP was carried out with a large excess of protein over RNA ($>$ 100fold) to ensure complete binding of RNA by protein. Experiments were typically carried out with 1-2 pmol labeled RNA. Protein(s) and RNA were mixed and incubated on ice for 30 min for complex formation. After cross-linking, samples were mixed with SDS sample buffer and directly subjected to SDS-PAGE. Cross-linking products were visualized by autoradiography.

2.2.8.2 Standard protocol for cross-linking and enrichment of cross-linked heteroconjugates for LC-ESI-MS/MS

The standard protocol for cross-linking and titanium dioxide enrichment was developed in our laboratory [71] and further optimized in the course of this thesis. Titanium dioxide enrichment follows procedures initially established for the enrichment of phosphopeptides [70]. Buffers for desalting and titanium dioxide enrichment were prepared with water, methanol and acetonitrile (ACN) in LiChrosolv/Chromasolv quality.

For reconstitution, RNA and protein were mixed in appropriate buffer, typically in a 1:1 molar ratio. The sample was incubated on ice for 30 min for complex formation. Reconstituted or isolated complexes were UV irradiated, typically for 10 min at 254 nm, in 100 µl aliquots in a microtiter plate placed on ice at a distance of 1 cm from the light source. Irradiated complexes were immediately ethanol precipitated.

Pelleted complexes were dissolved in 50 µl 4 M urea, 50 mM Tris pH 7.9 and diluted to 1 M urea, 50 mM Tris pH 7.9 with 150 µl 50 mM Tris pH 7.9. RNA hydrolysis was typically achieved with 1 µl each of RNases A (1 µg/µl) and T1 (1 U/µl) in a 2 h incubation at 52°C. In some cases, 1 µl benzonase (25 U/µl) was used instead of or in addition to RNases, for which $MgCl_2$ was added to the digestion buffer to a final concentration of 1 mM. Benzonase hydrolysis was typically carried out at 37°C for 1 h. Proteolysis was performed with trypsin, usually at an enzyme-to-protein ratio of 1:20 (w/w), in overnight incubation at 37°C. In general, incubations with enzymes were carried out in thermoshakers with mixing at 500 rpm.

Both C18 and TiO_2 spin columns were packed in-house. A pipette tip (epT.I.P.S. 0.5-10 µl; Eppendorf, Hamburg, Germany) was prepared with a piece of regular coffee filter around 2 mm^2 in size as a frit. C18 material, suspended in methanol, or TiO_2 material, suspended in 80% ACN, 0.1% TFA, was added to give a column of about 1.5 µl volume.

Desalting and removal of noncross-linked RNA fragments was carried out directly after hydrolysis. 10 µl ACN and 2 µl 10% FA were added to the sample to reach a final concentration of 5% ACN and 0.1% FA. All washing, loading and elution steps were performed by centrifugation at 5 000 rpm for 5 min. The C18 column was washed and equilibrated by passing 60 µl each of the following four solutions: 95% ACN, 0.1% FA; 80% ACN, 0.1% FA; 50% ACN, 0.1% FA; 0.1% FA. The sample was loaded on the column in 60 µl aliquots, washed twice with 60 µl 0.1% FA and eluted stepwise with two times 60 µl 50% ACN, 0.1% FA and 60 µl 80%ACN, 0.1% FA. The eluate was dried in a centrifugal evaporator.

Next, titanium dioxide enrichment was performed to remove noncross-linked peptides. All washing, loading and elution steps were performed by centrifugation at 3 000 rpm for 5 min. The TiO_2 spin column was equilibrated with 60 µl buffer B (80% ACN, 5% TFA). The sample was dissolved in 60 µl buffer A (200 mg/ml DHB in 80% ACN, 5% TFA) and loaded onto the spin column. Washing with buffer A (3x 60 µl) removed residual noncross-linked peptides, followed by extensive washing with buffer B (5x 60 µl) to remove DHB which is not compatible with LC-ESI-MS/MS analysis. Enriched peptide–RNA heteroconjugates were eluted with ammonia (0.3 M, 3x 40µl) and the sample was dried in a centrifugal evaporator.

For LC-ESI-MS/MS analysis, sample pellets were dissolved in the presence of 2 µl 50% ACN, 0.1% FA and diluted to a final concentration of 10% ACN, 0.1% FA by addition of 10 µl 0.1% FA. Of the 12 µl sample volume, 5 µl were injected for a single LC-MS/MS run.

2.2.8.3 Cross-linking of NusB–S10 to 4SU-substituted RNA

Cross-linking experiments with the NusB–S10 complex and the synthetic, 4SU substituted oligonucleotide 5'-CAC UGC UC(4SU) (4SU)(4SU)A ACA AUU A-3' were carried out with 2 nmol each of the protein complex and the RNA oligonucleotide in binding buffer. After incubation on ice for 30 min, the mixture was irradiated at 365 nm for 5 min. Hydrolysis and enrichment were carried out according to the standard protocol with 2 µl each of RNases A and T1 and trypsin at a ratio of 1:20 (w/w). LC-ESI-MS/MS was carried out on the Q-ToF Ultima.

2.2.8.4 Cross-linking of the ASH1 complex

ASH1 complexes were prepared by Roland Heym (Prof. Dierk Niessing, Institute of Structural Biology, Helmholtz Zentrum München) and isolated by gel filtration according to the published protocol [96]. Both ASH1 complexes contained a 51 nucleotide section of zip-code element E3 of the *ASH1* mRNA (5'-AUG GAU AAC UGA AUC UCU UUC AAC UAA UAA GAG ACA UUA UCA CGA AAC AAU-3'). The ASH1-FL complex consisted of full-length She3p and She2p, while the ASH1-short complex contained the C-terminus of She3p (92 amino acids; termed She3p-short from here onwards) and She2p. In reference to protein, 120 µg ASH1-FL and 100 µg ASH1-short were used for initial experiments. Protein buffer was 20 mM Hepes pH 7.8, 200 mM NaCl, 2 mM $MgCl_2$, 2 mM DTT. Complexes were UV irradiated at 254 nm for 10 min. Hydrolysis and enrichment were carried out according to the standard protocol with 2 µl each of RNases A and T1 and trypsin at a ratio of 1:20 (w/w). LC-ESI-MS/MS was carried out on the Q-ToF Ultima.

Experiments with LC-ESI-MS/MS analysis on the LTQ Orbitrap Velos were carried out with 20 µg ASH1-FL and 25 µg ASH1-short in each control and UV irradiated sample. Cross-linking, hydrolysis and enrichment were carried out according to the standard protocol with 10 min irradiation at 254 nm, 1 µl each of RNases A and T1 and trypsin at a ratio of 1:20 (w/w).

2.2.8.5 Cross-linking of Cwc2

Cwc2 protein and *in vitro* transcribed U4 and U6 snRNAs were prepared by Dr. Jana Schmitzová (Macromolecular Crystallography Group, Department of Cellular Biochemistry) according to published protocols [97].

Cwc2–RNA complexes were reconstituted *in vitro* by incubation of 100 µg Cwc2 with 1.5 µg RNA (U6 snRNA, U4 snRNA, or a synthetic oligonucleotide resembling internal stem-loop of U6) for 30 min on ice. The protein buffer contained 20 mM Hepes pH 7.5, 100 mM NaCl and 1 mM DTT. UV irradiation time was typically 10 min. Hydrolysis and enrichment were carried out according to the standard protocol with 1 µl each of RNases A and T1 and trypsin at a ratio of 1:50 (w/w). LC-ESI-MS/MS was carried out on the LTQ Orbitrap Velos.

2.2.8.6 Cross-linking of protein–RNA complexes after TAP tag purification

For cross-linking of protein–RNA complexes isolated from yeast extract by TAP tag purification, different variations were performed (see 3.4.2).

Experiments were typically started with 10 ml yeast extract ($\sim$ 350 mg protein) for both UV irradiated and control samples. For cross-linking of extract, samples were dialyzed against AGK buffer without glycerin. Cell extract, IgG or Calmoduline eluate was cross-linked for 2 min in petri dishes placed on ice with a liquid depth of around 1 mm.

Sample preparation for LC-ESI-MS/MS with C18 and titanium dioxide chromatography was essentially done according to the standard protocol. RNA hydrolysis was performed with 1 µl benzonase for 30 min at 37°C and 2 µl each of RNases A and T1 for 60 min at 52°C. Proteolysis was achieved by incubation with trypsin (1:50 w/w) at 37°C over night. Samples were typically split on two C18 columns to prevent overloading. All other steps followed the standard protocol.

Alternatively, samples were prepared for MS analysis with size exclusion, C18 and optionally TiO_2 chromatography. To this end, samples were ethanol precipitated after cross-linking of the IgG eluate. Pelleted complexes were dissolved in the presence of 20 mM Tris pH 7.5 and 1% SDS. The sample was diluted 1:10 with 20 mM Tris pH 7.5 to a final concentration of 0.1% SDS and hydrolyzed with trypsin (1:50 w/w) at 37°C over night. Size exclusion chromatography was performed with a Superdex 200 column (PC 3.2/30, 2.4 ml, Amersham Biosciences) in a SMART system (Pharmacia Biotech). Samples were injected in 50 µl aliquots. Running buffer was 20 mM Tris pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$ and flow rate 40 µl/min, absorption at both 254 and 280 nm were monitored. Fractions of 100 µl were automatically collected. Fractions showing high absorption at both 254 and 280 nm were pooled and ethanol precipitated. Samples were dissolved, hydrolyzed and desalted as described above. Optionally, they were further enriched with titanium dioxide according to the standard protocol.

2.2.9 LC-ESI-MS/MS analysis

Mass spectrometric analysis was carried out by sample injection into a nano-liquid chromatography (nano-LC) system directly coupled to the electrospray (ESI) source of a mass spectrometer. Three different mass spectrometers were used in this thesis:

- Q-ToF Ultima (Waters, Manchester, UK)
- LTQ Orbitrap Velos (Thermo Fisher Scientific, Schwerte, Germany)
- Q Exactive (Thermo Fisher Scientific, Schwerte, Germany)

Both Q-ToF and Velos were coupled to an Agilent LC-system (Agilent 1100 series, Agilent Technologies, Böblingen, Germany), the Q Exactive was coupled to an EASY-nLC II (Thermo Fisher Scientific). Details for LC separation and MS analysis are described below.

All columns used in nano-LC separation were packed in-house by Uwe Pleßmann (Bioanalytical Mass Spectrometry Group) with C18 AQ 120 Å material (Dr. Maisch GmbH, Ammerbuch, Germany; particle size 5 µm except for analytical column in EASY-nLC II, there 3 µm).

LC solvents were prepared with water and acetonitrile in LiChrosolv or Chromasolv quality.

2.2.9.1 Nano-LC separation (Agilent)

Samples were loaded onto a C18 trapping column (length $\sim$ 2 cm, inner diameter 150 µm) with a flow rate of 10 µl/min in 3% buffer B (buffer A: 0.1% FA; buffer B: 95% ACN, 0.1% FA) and washed for 5 min under the same conditions. Subsequently, a linear gradient of 3 to 36% buffer B was started with a flow rate of 300 nl/min. The gradient eluted the analytes from the trapping column onto the analytical column (length $\sim$ 15 cm, inner diameter 75 µm). On the analytical column, analytes were separated and eluted into the ESI source of the mass spectrometer. Elution time was 37 min (60 min gradient) or 97 min (120 min gradient). Finally, buffer B was raised to 95% for 7.5 min to elute any residual species and then lowered back to 3% to equilibrate the column for the next run.

2.2.9.2 Nano-LC separation (EASY-nLC II)

Washing and elution followed the same principles as described above. After loading onto the trapping column (length 4 cm, inner diameter 100 µm), samples were washed with a total volume of 25 µl buffer A at a maximum pressure of 280 bar. The linear gradient was from 4 to 36% buffer B within 92 min at a flow rate of 250 nl/min. The analytical column was 10 cm long with an inner diameter of 50 µm. Final elution was carried out at 95% buffer B for 8 min, column equilibration was internally managed by the LC system.

2.2.9.3 ESI-MS/MS analysis on the Q-ToF Ultima

Prior to analysis of each set of samples, the instrument was calibrated by direct injection of 10 fmol/µl GluFib. Correct calibration and functionality of the instrument was confirmed by a LC-MS/MS run of a BSA digest (injection of 50 fmol).

The instrument was operated in data dependent acquisition mode with a Top3 method. MS survey scans were recorded in the *m/z* range of 350 to 1600, acquisition time was 1 s. The three most intense precursors were chosen for fragmentation with CID (minimum intensity 40 counts; charge states 2, 3 and 4; isolation width +/- 1.5 *m/z*; *m/z* range 50-2000; acquisition time 3x 1 s) and subsequently excluded from re-fragmentation for 180 s (dynamic exclusion).

2.2.9.4 ESI-MS/MS analysis on the LTQ Orbitrap Velos

The instrument was operated in data dependent acquisition mode with a Top10 method. MS survey scans were recorded in the *m/z* range of 350 to 1600 at a resolution of 30 000. The ten most intense precursors were chosen for fragmentation with HCD (minimum intensity 5 000; charge states 2, 3 and 4; isolation width +/- 1 *m/z*; normalized collision energy 45) and subsequently excluded from re-fragmentation for 20 s (dynamic exclusion). MS/MS fragment spectra were recorded with a fixed first mass of 100 *m/z* and a resolution of 7 500.

2.2.9.5 ESI-MS/MS analysis on the Q Exactive

The instrument was operated in data dependent acquisition mode with a Top12 method. MS survey scans were recorded in the *m/z* range of 350 to 1600 at a resolution of 70 000. The twelve most intense precursors were chosen for fragmentation with HCD (minimum intensity 10 000; charge states 2, 3 and 4; isolation width +/- 1 *m/z*; normalized collision energy 30) and subsequently excluded from re-fragmentation for 20 s (dynamic exclusion). MS/MS fragment spectra were recorded with a fixed first mass of 100 *m/z* and a resolution of 17 500.

2.2.10 MS data analysis

2.2.10.1 Peptide identification with Mascot

Standard database search was typically done with Mascot (Matrix Science, [33]) as search engine. Raw data was processed and converted into text-based format. Q-ToF Ultima data was recalibrated and processed with MassLynx V4.0 (Waters) and converted into *.pkl*. Orbitrap data was converted into *.msm* by Raw2MSM version 1.10 [98].

Mascot search parameters are listed below. Typically, searches were performed against the respective taxonomy with phosphorylation as PTM to identify the latter, against the entire NCBI database without phosphorylation to identify additional contaminants, or against a reduced database containing U1 snRNP proteins and sequences for NusB, S10, She2p, She3p, and Cwc2 (respective sequences were added for the corresponding experiment). Searches against the entire NCBI database served as controls to exclude false positive cross-linking results if the corresponding spectrum resulted from a contamination like keratin.

Mascot search parameters

enzyme	trypsin
max. missed cleavages	2
precursor	monoisotopic
MSMS search	on
peptide charge	2+, 3+
peptide tolerance	50 ppm (Q-ToF) or 10 ppm (Orbitrap)
MSMS tolerance	0.2 Da (Q-ToF) or 0.02 Da (Orbitrap)
instrument	ESI-Quad-Tof (Q-ToF) or ESI-Trap (Orbitrap)
variable PTM	oxidation (M)
	carbamylation (N-term)
	carbamylation (K)
	(optional) phospho (ST)
	(optional) phospho (Y)

2.2.10.2 Identification of cross-links with 94 Da adducts using Mascot

After identification of the cross-linking product producing a 94 Da shift of peptide fragments, this particular observation was used to define an additional PTM in Mascot. The PTM was defined with a sum formula of $C_9H_{11}N_2O_8P$ (306.0253 Da, corresponding to [4SU $-H_2S$] or [U $-H_2O$]) and a neutral loss of $C_5H_9O_7P$ (212.0086 Da) which leaves $C_4H_2N_2O$ (94.0167 Da, corresponding to [(4SU)' $-H_2S$] or [U' $-H_2O$]) as a shift of the peptide fragments. The modification site was defined on the peptide C-terminus. This way, peptides cross-linked close to the C-terminus and showing this particular shift after fragmentation could be identified in searches against small databases or the entire proteome.

2.2.10.3 Online tools for calculation of monoisotopic masses

Monoisotopic masses of peptides and their fragments, RNA oligonucleotides and their fragments, as well as peptide–RNA adducts were calculated with several tools available online.

ProteinProspector	University of California, San Francisco; http://prospector.ucsf.edu
Peptide Mass Calculator	University of Leuven; http://rna.rega.kuleuven.ac.be/masspec/pepcalc.htm
Mongo Oligo Mass Calculator	University at Albany; http://rna-mdb.cas.albany.edu/RNAmods/masspec/mongo.htm
Molecular Mass Calculator	University at Albany; http://rna-mdb.cas.albany.edu/RNAmods/masspec/mole.htm

2.2.10.4 Identification of cross-links by manual spectra interpretation

MS/MS spectra of noncross-linked peptides were excluded by a standard database search. Spectra of residual RNA were excluded by characteristic groups of fragments 18 Da apart (neutral loss of water). Additionally, precursor and fragment masses of noncross-linked RNA have a smaller fractional mass than peptides. Remaining unassigned spectra of reasonable quality were annotated manually. Typically, a peptide sequence tag was derived from fragment series in the higher *m/z* range. The sequence tag was compared to the sequences of the proteins under investigation. Theoretical fragment masses of corresponding tryptic peptides were manually compared to the experimental spectrum. Once a match was confirmed, the cross-linked RNA was derived from the difference between experimental precursor and calculated peptide mass.

2.2.10.5 Identification of cross-linked peptides after precursor variant generation by a perl script

LC-ESI-MS/MS data of cross-linking experiments with NusB–S10 and ASH1 were analyzed after precursor variant generation with a perl (www.perl.org) script written by Dr. Petra Hummel (IT & Electronics Service). This script was developed, tested and optimized in the course of this thesis and details can be found in the results section (see 3.1.2.4). Briefly, the LC-MS/MS data was analyzed as described below.

Input for the perl script was data in *.pkl* format created by MassLynx V4.0 after processing of MS raw data.

The presence of RNA marker ions was not used as a filter. For the report of observed marker ions, mass deviation was set to 0.1 Da, the (relative) intensity threshold to 15% (NusB–S10) or 20% (ASH1) and 5% for the second A marker. The (absolute) intensity threshold for noise filtering was set to 3.

For NusB–S10, precursor mass variants were created for all combinations of A, C, G, U and 4SU. Separate searches were performed with combinations of A, C, G and U and combinations of all five nucleotides with at least one 4SU in the sequence. For ASH1, precursor mass variants were created for all oligonucleotides length 1-4 from the RNA sequence 5'-AUG GAU AAC UGA AUC UCU UUC AAC UAA UAA GAG ACA UUA UCA CGA AAC AAU-3'.

The output of the perl script contained several files: One *.csv* file summarized parameters chosen while running the script, including all RNA sequences used to create the precursor mass variants. The file also listed all precursors whose fragment spectra were filtered or did not contain any marker ions. A second *.csv* file contained all precursors with RNA marker ions above the threshold in the corresponding spectrum. For each spectrum not excluded by the low mass or fractional mass filter, one *.pkl* file was created that contained the experimental precursor and all its variants, each with the MS/MS fragment information reduced by the noise filter.

.pkl files were searched with Mascot (parameters see above). Initial searches were performed against a small database (see above). Spectra of cross-link candidates were researched against the respective proteome (*E. coli* for NusB–S10 or *S. cerevisiae* for ASH1).

2.2.10.6 Identification of cross-linked peptides with OpenMS and OMSSA

Cross-linking experiments of ASH1, Cwc2 and yeast protein–mRNA complexes after TAP tag isolation were analyzed with OpenMS [99, 100] and OMSSA [34] as search engine. Data analysis workflows were developed in the course of this work and are explained in more details in the results section. Workflows are based on OpenMS tools written especially for our purpose as well as existing tools. Code was written and TOPPAS pipelines were assembled by Timo Sachsenberg (Prof. Oliver Kohlbacher, Applied Bioinformatics Group, Eberhard Karls University, Tübingen).

MS data in Thermo proprietary *.raw* format was converted into the open *.mzML* format [101] with msconvert, part of the ProteoWizard [102] software bundle. Q Exactive data was processed with the OpenMS tool FileFilter with the option "sort" for correct assignment of MS1 and MS2 spectra. MS data recorded in profile mode, i.e. MS1 spectra of Velos measurements and both MS1 and MS2 of Q Exactive measurements, were centroided with the OpenMS tool PeakPickerHiRes. If automatic XIC filtering was desired later, an additional processing step was included: LC-MS data of control and UV irradiated sample were aligned to correct for small retention time shifts. The corresponding pipeline is shown in Figure 2.1. The pipeline requires the *.mzML* files of both control and UV irradiated sample as input. Output file is the control *.mzML* with transformed retention times.

After data processing and before creating precursor mass variants, the MS data was reduced by identification (ID) and extracted ion chromatogram (XIC) filters if desired. The ID filter pipeline (Figure 2.2) performed a standard database search with OMSSA to identify noncross-linked peptides, the corresponding MS/MS fragment spectra were removed from the MS data file. The database contained contaminant sequences (those distributed with MaxQuant [103]) as well as decoy sequences. The latter were used to determine a false discovery rate (FDR) and were created with the OpenMS DecoyDatabase tool by reversing the target sequences from the original database. A peptide hit was considered a confident match and subsequently used for filtering if the FDR was below 0.01. Parameters for the OMSSA search are listed below. Input file is an *.mzML*, output files are an *.idXML* file containing the peptide matches used for filtering, and a reduced *.mzML*. The output *.idXML* can be annotated to the input *.mzML* to retrace the peptide identifications.

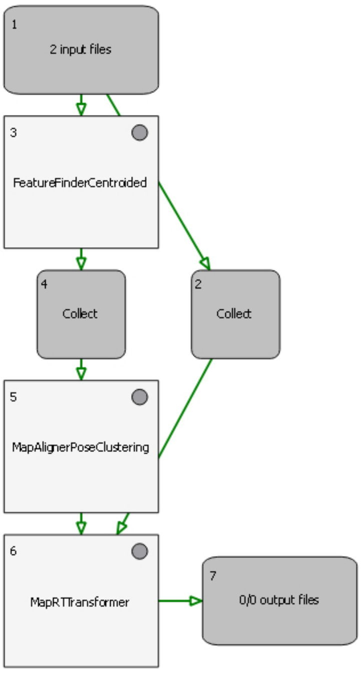


Figure 2.1: Pipeline for retention time alignment of LC-ESI-MS/MS data of control and UV irradiated sample (screenshot from TOPPAS). First, in both measurements peptides (features) are identified in the two-dimensional retention time versus m/z map by FeatureFinderCentroided. Based on the features, maps of both measurements are aligned by MapAlignerPoseClustering and the retention time transformations are applied by MapRTTransformer. Importantly, the control is transformed relative to the UV irradiated sample and not vice versa.

OMSSA search parameters

precursor mass tolerance	10 ppm
fragment mass tolerance	0.1 Da
min/max precursor charge	2/5
precursor charge determination	believe input file
variable modifications	oxidation (M)
	carbamylation (K), carbamylation (N-term)
	phospho (S), phospho (T), phospho (Y)
enzyme	trypsin
max number missed cleavages	2

The XIC filter was applied to remove MS/MS spectra of precursors that appeared in both control and UV irradiated sample at comparable intensity (default: fold change less than two). This filtering step was done with the OpenMS RNP^{xl}XICFilter specifically created for our purpose. Input are the *.mzML* files of both samples. The tool then calculates the intensity of a precursor in both control and UV irradiated sample in a small retention time window. If the intensity in the UV irradiated sample is less than twofold higher than in the control, the corresponding spectrum is filtered and not written into the output, the reduced *.mzML* file of the UV irradiated sample.

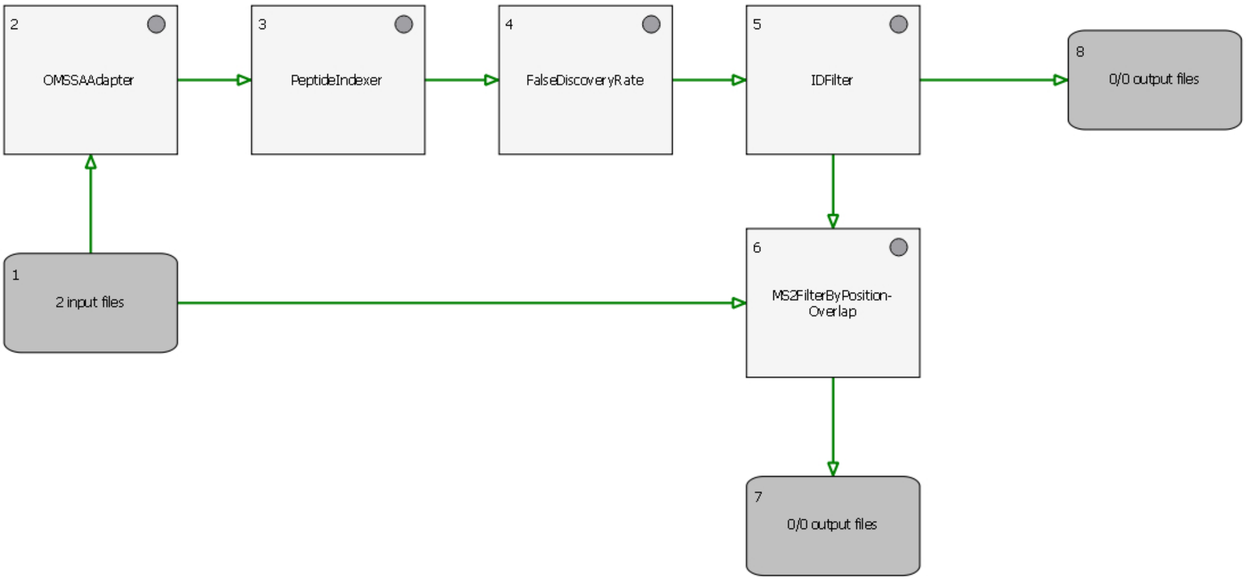
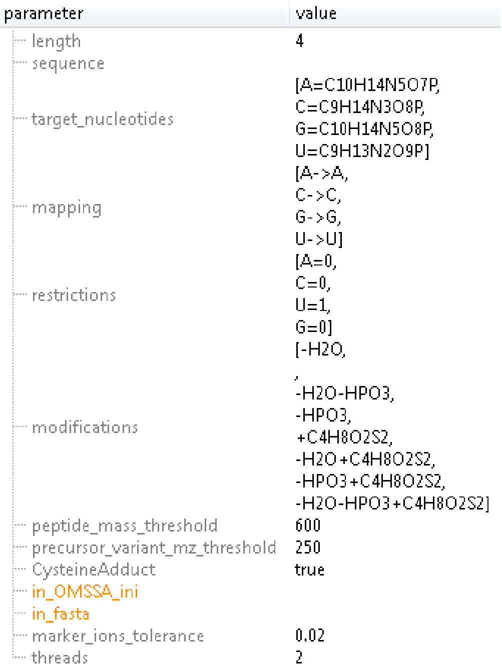


Figure 2.2: ID filter pipeline for removal of MS/MS spectra with confident peptide identification (screenshot from TOPPAS). OMSSAAdapter submits the OMSSA searches and retrieves the search result. PeptideIndexer determines whether identified peptides correspond to target or decoy sequences. FalseDiscoveryRate determines the false discovery rate for each identification. Finally, IDFilter keeps only identifications below a certain false discovery rate, typically 0.01. Confident identifications that pass this criterion are reported in an *.idXML* output file. Finally, MS2FilterByPositionOverlap removes the MS/MS spectra that gave rise to the confident identifications from the *.mzML* file, the reduced *.mzML* is output of the pipeline.

The crucial step of the data analysis, precursor mass variant generation and database searches, were performed with the RNP^{xl} tool, another OpenMS tool specifically created for our purpose. The tool takes an *.mzML* file as input. This file can be a reduced *.mzML* from any of the filtering steps described above or the original *.mzML* containing all raw data. Output files are an *.idXML* and a *.csv* file, both containing the database search results and RNA marker ion intensities for all MS/MS spectra contained in the input *.mzML*. The *.idXML* file can be used to annotate the search results to the MS data in *.mzML* in TOPPView, while the *.csv* file can be opened in programs like Microsoft Excel, e.g. to add notes about manual validation. Parameters for the RNP^{xl} tool are shown in Figure 2.3, the values correspond to the optimized parameters for yeast protein–RNA complexes after TAP tag purification. OMSSA search parameters are essentially as described for the ID filter with two important differences: The database is a limited database or the proteome of the respective organism, it does not contain contaminant or decoy sequences as those would increase analysis time and lead to false positive matches. For similar reasons, phosphorylation is not considered as a variable peptide modification.



parameter	value
length	4
sequence	

target_nucleotides	[A=C10H14N5O7P,
	C=C9H14N3O8P,
	G=C10H14N5O8P,
	U=C9H13N2O9P]

mapping	[A->A, C->C, G->G, U->U]

restrictions	[A=0,
	C=0,
	U=1,
	G=0]

modifications	[-H2O, , -H2O-HPO3, -HPO3, +C4H8O2S2, -H2O+C4H8O2S2, -HPO3+C4H8O2S2, -H2O-HPO3+C4H8O2S2]
peptide_mass_threshold	600
precursor_variant_mz_threshold	250
CysteineAdduct	true
in_OMSSA_ini	
in_fasta	
marker_ions_tolerance	0.02
threads	2

Figure 2.3: Parameters of the RNP$^{\mathrm{xl}}$ tool (screenshot from TOPPAS). *length* determines the maximum length of RNA combinations to be considered for precursor variant generation. *sequence* allows the input of a nucleotide sequence if only those combinations that appear in the sequence should be considered. When left empty, all combinations from the nucleotides defined below are calculated. *target_ nucleotides* allows the definition of any nucleotide (RNA, DNA, substituted or labeled with stable isotopes) by its sum formula. The *mapping* option is used to define an input sequence that is randomly labeled, then the labeled and the native nucleotide are mapped on the same letter in the input sequence. *restrictions* are used to require a certain nucleotide in all sequences considered for precursor mass variants. The parameters shown here would only allow sequences that contain at least one uracil. In the *modifications* field, all modifications are listed that should be considered for each of the nucleotide combinations. The parameters shown here resemble a standard experiment where the 152 adduct is also expected. All modifications have to be given as sum formulas. *precursor_ mass_ threshold* sets the (uncharged) threshold for the low mass filter, while *precursor_ variant_ m/z_ threshold* sets the m/z threshold for the precursor mass variants that are written in the output file. If *CysteinAdduct* is set to "true", 152 is considered as an adduct without any nucleotide. *in_ OMSSA_ ini* and *in_fasta* require the paths of the OMSSA parameter file and the database (in *.fasta* format), respectively. Finally, *marker_ ion_ tolerance* sets the mass tolerance for the determination of the presence and intensity of RNA marker ions.

2.2.10.7 Validation of cross-links

Cross-link candidates obtained from manual spectra interpretation or database search after precursor variants generation were validated in several steps. Validation criteria were refined and expanded in the course of this project and are described in detail in the results section. Important validation criteria are briefly listed below.

Correct assignment of monoisotopic peak and charge state were confirmed by evaluating the survey scan preceding the fragment spectrum under investigation. When data from a non-irradiated control was available, extracted ion chromatograms were compared to confirm that the precursor was not present in the control at significant intensity. Results of an independent Mascot search for peptide identification confirmed that the fragment spectrum did not yield any true positive hit for a noncross-linked peptide. Failure to meet any of the above mentioned criteria led to exclusion of the candidate as a false positive.

The experimental fragment spectrum was compared to predicted fragments of the candidate peptide. Peptide fragment masses were calculated from the amino acid sequence with ProteinProspector. In TOPPView, Orbitrap data was directly annotated with search results, experimental signals corresponding to calculated fragments were automatically highlighted. Remaining high intensity signals were manually compared to RNA fragments or peptide–RNA adducts.

Cross-link candidates were rejected when several high intensity signals could not be explained by calculated fragments of the candidate cross-link. Particular emphasis was on peptide fragment series in the higher m/z range, high intensity immonium ions, and RNA marker ions. Cross-linked RNA with two or more nucleotides should yield marker ions of significant intensity, marker ions for A, C and G base were expected to be dominating in the fragment spectrum if they appeared in the cross-linked RNA.

3 Results

UV induced protein–RNA cross-linking and its investigation by mass spectrometry is based on the following key steps:

- isolation or reconstitution of the protein–RNA complex(es)
- UV irradiation
- sample preparation for mass spectrometry (enrichment of cross-linked heteroconjugates)
- analysis by mass spectrometry
- data analysis

While several experimental strategies have been developed for UV cross-linking and mass spectrometry, there was further need for optimization and adaptation, especially for more complex biological systems. In addition, while advances in mass spectrometry instrumentation have led to great advances, they have also resulted in a call for adjustments and re-evaluations of existing experimental and data analysis strategies.

In the course of this work, all of the key steps were addressed. Experimental workflows were adjusted and optimized for ribonucleoproteins that had not been previously investigated by UV cross-linking and mass spectrometry. However, the major focus of this work was on data analysis. At the beginning of this project, MS data derived from cross-linking experiments was analyzed manually. MS/MS spectra were assigned by hand, a time-consuming process that requires considerable expertise in spectra interpretation. While feasible for small ribonucleoproteins and a limited number of spectra, increasing complexity and MS data amounts called for a new approach. Thus, in parallel with investigations of novel aspects in UV cross-linking and optimization of experimental workflows for several ribonucleoproteins, a data analysis strategy was developed and refined which eventually allowed the identification of cross-linked peptides in searches against entire proteomes.

3.1 Cross-linking products of 4-thio-uracil and a novel approach for automated data analysis

One of the major constraints in UV induced protein–RNA cross-linking is the low cross-linking yield. A strategy to increase the cross-linking yield is the use of photo-reactive nucleotides, e.g. 4-thio-uracil, 6-thio-guanine, or halopyrimidines such as 5-bromo-uracil.

In order to identify cross-linked peptide–RNA oligonucleotide heteroconjugates by mass spectrometry, the mass of the cross-linking product has to be known. For native RNA, cross-linking is mainly additive, i.e., the mass of the cross-linked heteroconjugate is the sum of the peptide and oligonucleotide masses (e.g. [65]). However, it was unknown whether the same is true for RNA substituted with carbonothioyl-containing bases.

We set out to address the two major constraints of cross-linking experiments: The use of a photo-reactive base-analogue, 4-thio-uracil (4SU), was investigated with a focus on cross-linking yield and mass of cross-linking products. In parallel, an approach for the automatization of data analysis was developed. For the intended experiments, a simple test system was needed. The NusB–S10 complex from *E. coli* was chosen since it had been investigated previously by protein–RNA cross-linking and mass spectrometry in our laboratory [71]. It plays an important role in transcription antitermination and has an enhanced affinity for BoxA-containing RNA. Co-expression of the protein complex had been established and could be reproduced. More importantly, the *rrn* BoxA-containing oligonucleotide used in the previous study is short and contains several uracils. Therefore, the variant of the same oligonucleotide synthesized with 4-thio-uracils at specific positions could be obtained.

More precisely, a 19mer RNA oligonucleotide containing the core *rrn* BoxA element (underlined) was cross-linked to the NusB–S10 complex. Cross-linking to the unsubstituted oligonucleotide (upper sequence) had been previously investigated [71]. We compared these results to cross-linking to the same oligonucleotide in which three uracils in the BoxA element were replaced by 4-thio-uracil (lower sequence).

5'-CAC UGC UCU UUA ACA AUU A-3'

5'-CAC UGC UC(4SU) (4SU)(4SU)A ACA AUU A-3'

3.1.1 Influence of 4-thio-uracil on the cross-linking yield of the NusB–S10-complex

The influence of 4-thio-uracil on the cross-linking yield of the NusB–S10 complex was investigated by cross-linking of ^{32}P-labeled oligonucleotides to the protein complex. Two 19mer oligonucleotides, with and without 4SU, were 5'-labeled with [γ-^{32}P]-ATP and cross-linked to the NusB–S10 complex. Cross-linking products were separated by SDS-PAGE and visualized by autoradiography (see Figure 3.1). UV irradiation of the proteins in complex with the unsubstituted oligonucleotide (lane 2) at 254 nm led to cross-linking products of both proteins, while no protein bands were observed in the non-irradiated control (lane 1). In contrast, the non-irradiated control of the complex with the 4SU-substituted oligonucleotide already contained cross-linking products (lane 3). This illustrates the high reactivity of 4SU: It cross-links under ambient light, even when protected from light as much as possible during the experiment. Increasing irradiation time at 365 nm (1, 2, 5, and 10 min; lanes

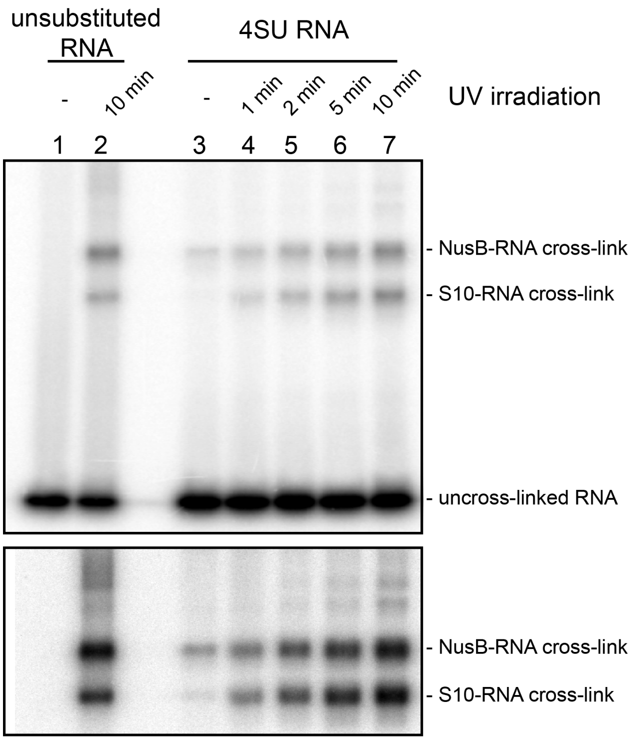


Figure 3.1: Autoradiography of NusB–S10 cross-linked to ^{32}P-labeled BoxA containing RNA oligonucleotides with and without 4-thio-uracil. The upper panel shows the autoradiography after 15 min exposure of a Phosphorimager screen, the lower panel shows details of the cross-linking products after 1 h exposure. Lanes 1 and 3 correspond to non-irradiated controls of complexes with unsubstituted and 4SU substituted RNA, respectively. Lane 2 shows cross-linking of NusB–S10 to unsubstituted RNA after 10 min irradiation at 254 nm. Lanes 4-7 show cross-linking products of the complex with 4SU-substituted RNA after UV irradiation at 365 nm for the time periods indicated above the gel lanes. Figure originally published in [104].

4–7) produced higher amounts of cross-linking products. However, a high excess of RNA remains uncross-linked, independent of substitution and irradiation time, and despite the high excess of protein used. This exemplifies the generally low yield of UV induced cross-linking.

The majority of cross-linking products observed after denaturing gel electrophoresis were binary protein–oligonucleotide complexes of either NusB or S10 and the oligonucleotide. Both unsubstituted and 4SU-containing RNA also showed higher-order cross-links. Their exact nature cannot be determined in our experiments.

Detailed investigation on the cross-linking products (lower panel in Figure 3.1) allowed for comparison of the cross-linking yields of the complexes with unsubstituted (lane 2) and 4SU-substituted (lane 7) RNA after the same irradiation period. Quantitative analysis of cross-linking product band intensities revealed that the cross-linking yield decreased by about 10% for NusB, while it increased by approximately 50% for S10. Thus, for the S10 protein, 4SU significantly enhances the cross-linking yield. At 254 nm, all nucleotides of the 19mer could undergo cross-linking. In contrast, only the three 4SU nucleotides were excited by irradiation at 365 nm. The slight decrease in the cross-linking yield of NusB could be due to it forming cross-links to nucleotides outside the triple U stretch. Upon substitution and irradiation at higher wavelengths, these cross-links might not form, consequently decreasing the cross-linking yield. However, our experiments clearly illustrate the potential of 4SU to increase the cross-linking yield for some proteins.

3.1.2 Development of a novel approach for automated data analysis

Identification of peptide–RNA oligonucleotide cross-links from mass spectrometry data has been done by manual spectra interpretation (see 2.2.10.4). Interestingly, the majority of fragments in the MS/MS spectra of cross-linked heteroconjugates correspond to the cross-linked peptide. In some cases, intense marker ions of the RNA bases or nucleotides are observed. Rarely, adducts of peptide and RNA or their fragments are detected. Based on these observations, we developed an idea for the identification of cross-linked peptides by database search.

3.1.2.1 Anticipated RNA combinations and modifications

We expected to identify cross-linked RNA with a maximum length of four nucleotides. No longer RNA sequences had been identified in cross-links after titanium dioxide enrichment and LC-ESI-MS/MS analysis. The number of RNA sequences to be considered can be calculated as k combinations of n elements with repetition according to the following equation [105]:

$$\bar{c}_n^k = \binom{n+k-1}{k} \tag{3.1}$$

In our case, n corresponds to the number of nucleotides and k to the oligonucleotide length. Consequently, the number of possible RNA sequences from the four standard nucleotides with a maximum oligonucleotide length of 1, 2, 3, or 4 is:

$$\binom{4}{1} + \binom{5}{2} + \binom{6}{3} + \binom{7}{4} = 69 \tag{3.2}$$

In addition, the RNA can have different modifications, i.e., the 5' and 3' end can be both hydroxyl, one hydroxyl and one phosphate, or both phosphate; all of these can additionally have a neutral loss of water, e.g. due to the formation of cyclic phosphates on the ribose. Therefore, each of the 69 sequences can have six modifications, which leads to a total of 414 potential RNA masses to be considered. The problem is further complicated when taking the 152 adduct (see 1.3.4) into account: Each of these combinations can be considered with and without additional 152. When also considering 152 alone as a cross-linking adduct, the result is 829 potential RNA masses. In case of a fifth nucleotide, Equation 3.2 changes to:

$$\binom{5}{1} + \binom{6}{2} + \binom{7}{3} + \binom{8}{4} = 125 \tag{3.3}$$

Multiplied with the six modifications (but disregarding the 152 adduct), this leads to 750 potential masses of the cross-linked RNA.

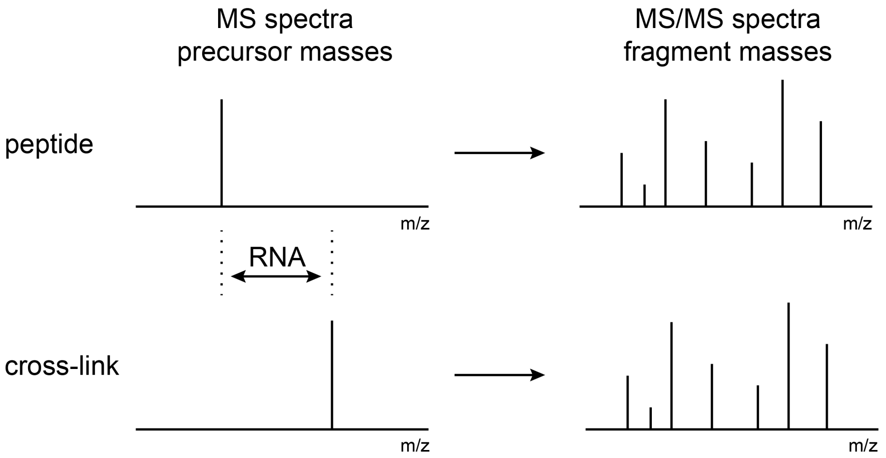


Figure 3.2: Schematic, simplified comparison of the MS and MS/MS spectra of the same peptide with and without cross-linked RNA. The peptide (precursor) mass is shifted by the mass of the RNA as a consequence of cross-linking. In contrast, the MS/MS spectra might not exhibit any obvious differences; in extreme cases the cross-link spectrum contains no trace of the RNA.

3.1.2.2 Cross-linked RNA and standard database search

Theoretically, MS/MS spectra of cross-linked heteroconjugates contain enough peptide fragments to enable the cross-linked peptide to be identified by algorithms employed in the analysis of standard proteomics experiments. In extreme cases, there might not be any obvious difference between the MS/MS spectrum of the same peptide, whether it is cross-linked or not. In such cases, the only evidence for the cross-linked RNA is the precursor mass shifted by the mass of the cross-linked RNA, schematically shown in Figure 3.2. This mass shift must be taken into account for database search.

In principle, the cross-linked RNA could be treated as any other post-translational modification (PTM). These are defined by their mass and the site(s) of modification. For example, phosphorylation is defined with an additional mass of 79.9663 Da (HPO_3) on typically either serine, threonine, or tyrosine. The mass of anticipated cross-linking products is known and we do not expect to identify novel products automatically. However, the number of potential RNA masses by far surpasses the number of PTMs considered in the analysis of a typical proteomics experiment.

The number of PTMs to be considered in a single database search is always limited, e.g. to nine PTMs in Mascot. Therefore, over 80 searches would have to be performed to consider all 750 potential RNA adduct masses in an experiment with 4SU-labeled RNA. But the combinatorial problem is even greater: The mass is only the first variable that needs to be defined. The second is the site of modification. While there are certain amino acids that seem to be more reactive in UV induced cross-linking, all amino acids could be modified. However, performing over 80 searches for each of the 20 amino acids is not feasible.

The problem could be simplified by defining the cross-linking site on the peptide N-terminus. Beam-type CID typically leads to a long y-series and only a few a- and b-ions containing the peptide N-terminus. Therefore, the N-terminal fragments could be disregarded for automated identification. In addition, the frequent loss of the cross-linked RNA from the peptide upon fragmentation could be accounted for by defining the RNA mass as a neutral loss. Similarly, loss of 97.9769 Da (H_3PO_4)

is defined as a neutral loss for phosphorylation as it is often observed. The database search engine then considers peptide fragments both with the PTM and with PTM and neutral loss.

This strategy could have been employed for the identification of cross-linked peptides with standard database search engines such as Mascot without the need for additional programs, scripts, etc. However, combining the results of many similar searches would not have been trivial. Therefore, we developed an alternative approach, described below. Cross-linked RNA as a PTM in database search was later applied successfully to one specific cross-linking product (see Sections 2.2.10.2 and 3.1.3).

3.1.2.3 The precursor variant approach

$$MW_{cross-link} = MW_{peptide} + MW_{RNA} \tag{3.4}$$

$$MW_{peptide} = M_{experimental} - MW_{RNA} \tag{3.5}$$

The strategy for PTM identification could not be transferred directly to cross-links. However, we developed an idea based essentially on the same observations: MS/MS spectra of cross-links contain mainly peptide fragments and the precursor mass shift by the cross-linked RNA has to be taken into account. For automatization, the problem was approached from the opposite direction: The masses of all potential RNA adducts could be subtracted from the experimental precursor mass (i.e., changing the perspective from Equation 3.4 to 3.5). The obtained precursor mass variants could then be submitted into database search. Only the precursor mass variant resulting from subtraction of the mass of the actual cross-linked RNA oligonucleotide should yield a true positive hit in the database search; all other variants should give no hits or false positive results.

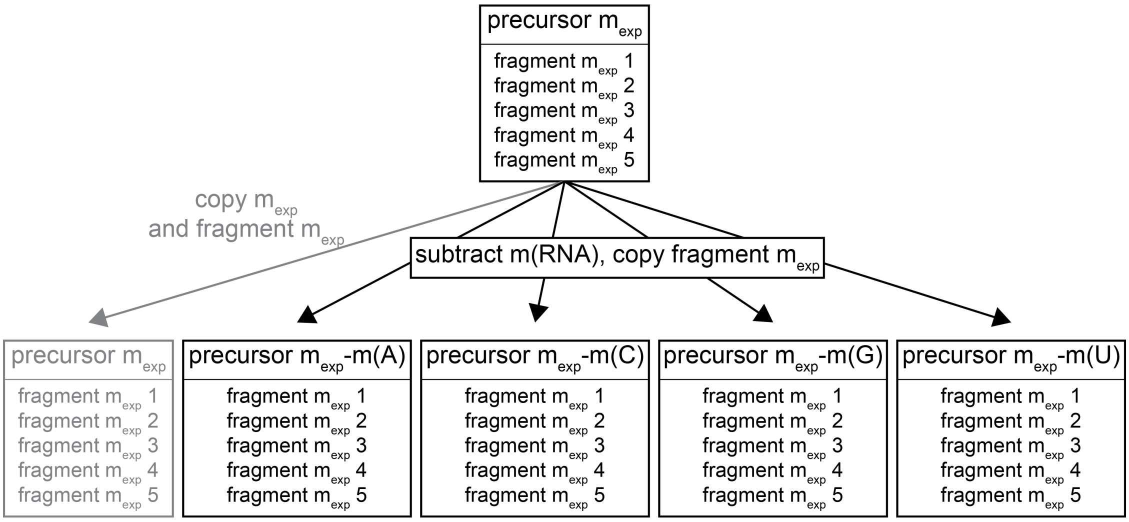


Figure 3.3: Schematic representation of precursor variant generation. The masses of potentially cross-linked RNA (here the nucleotides A, C, G, and U) are subtracted from the experimental precursor mass. The masses of the MS/MS fragments are copied without any modification. The original experimental precursor mass is kept as a control.

The idea is illustrated in Figure 3.3: Precursor mass variants are generated by subtraction of all potential masses of cross-linked RNA. The unaltered MS/MS information is copied to each of the precursor mass variants and to the original precursor mass. If the cross-linked RNA is a U nucleotide,

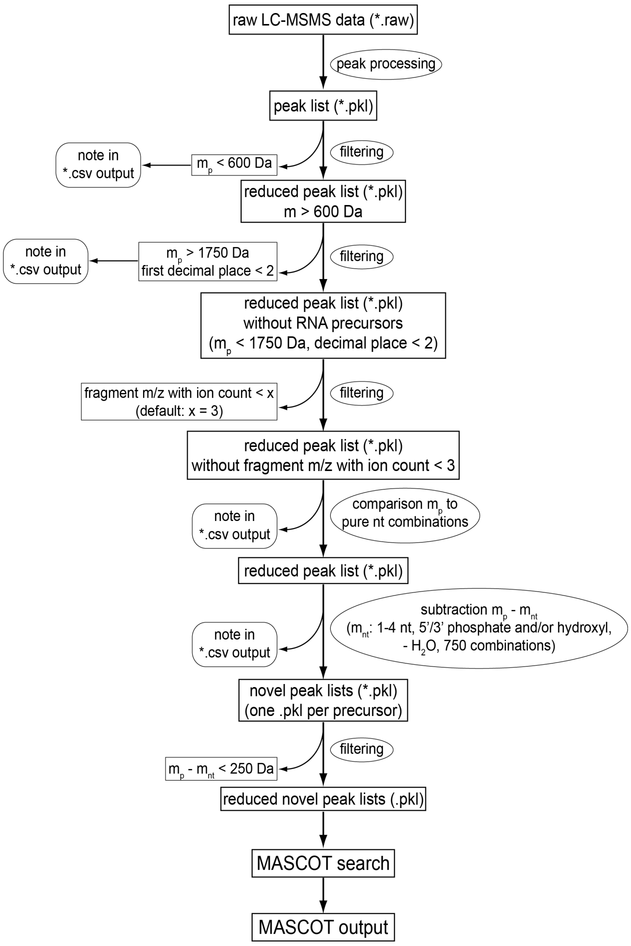

raw LC-MSMS data (*.raw)
peak processing
peak list (*.pkl)
filtering
m_p < 600 Da
note in *.csv output
reduced peak list (*.pkl) m > 600 Da
filtering
m_p > 1750 Da first decimal place < 2
note in *.csv output
reduced peak list (*.pkl) without RNA precursors (m_p < 1750 Da, decimal place < 2)
filtering
fragment m/z with ion count < x (default: x = 3)
reduced peak list (*.pkl) without fragment m/z with ion count < 3
comparison m_p to pure nt combinations
note in *.csv output
reduced peak list (*.pkl)
subtraction m_p - m_{nt} (m_{nt}: 1-4 nt, 5'/3' phosphate and/or hydroxyl, - H_2O, 750 combinations)
note in *.csv output
novel peak lists (*.pkl) (one .pkl per precursor)
filtering
m_p - m_{nt} < 250 Da
reduced novel peak lists (.pkl)
MASCOT search
MASCOT output

Figure 3.4:
Schematic description of data analysis with a perl script. Raw data is processed and converted into the text-based *.pkl* file format. Small precursors ($<600\,$Da) and precursors of short oligonucleotides (precursor $<1750\,$Da and fractional mass <0.2) are filtered. The remaining precursor masses are compared to masses of RNA oligonucleotides, and agreement is noted in a *.csv* file. Noise signals below a threshold of three counts are removed from the MS/MS spectra. Precursor variants are generated by subtracting the masses of all possible nucleotide combinations, i.e., 1-4 nucleotides, 5' and 3' end both phosphate, hydroxyl/phosphate, or both hydroxyl; and neutral loss of water. For partially 4SU-substituted RNA, this corresponds to 750 RNA masses. For each spectrum, the original precursor together with its variants are written in a separate *.pkl* output file. Precursor variants $<250\,$Da are filtered from the output *.pkl* files. All *.pkl* files are submitted into a Mascot search and Mascot search results are evaluated manually.
Figure originally published in [104].

we would only expect a database search result for the corresponding precursor variant (precursor m_{exp}-m[U]), all other variants should give no (or false positive) results. If the MS/MS spectrum results from fragmentation of a noncross-linked peptide, only the original precursor mass should yield a true positive database search result. Thus, the original precursor mass serves as a control.

3.1.2.4 Implementation of the precursor variant approach

For the implementation of the precursor variant approach, we received support from Dr. Petra Hummel (IT & Electronics Service). She realized the initial idea as well as all subsequent optimizations in the form of a perl script according to our specifications; its final workflow is outlined in Figure 3.4.

MS data from cross-linking experiments was converted into the text-based *.pkl* format by data processing in MassLynx V4.0, the vendor software of the Waters Q-ToF Ultima. For each MS/MS spectrum, the *.pkl* format contains a header line with precursor *m/z*, precursor intensity, and precursor charge state. Below, the fragment information is listed in the form of fragment *m/z* versus intensity. This format allowed for easy processing with the perl script.

Calculating the masses of all precursor variants and subtracting them from the precursor masses was easily implemented with the perl script. For each MS/MS spectrum, a separate *.pkl* output file was created with the original precursor mass and all its variants, each combined with the MS/MS fragment information. This was the only way to ensure that database search results could be traced back to the original precursor masses. Consequently, a separate search was performed for each spectrum, and only the best-scoring hit had to be evaluated further, while all other matches could be considered false positives.

When considering all combinations of 1–4 nucleotides, calculated RNA masses can be larger than the experimentally observed precursors. In addition, we do not expect to identify any peptide with fewer than three amino acids. The smallest tryptic peptide with three amino acids is GGK, with a monoisotopic mass of 260.1484 Da. The search of negative and extremely small precursor variants would be unnecessary. Thus, we included a filter that would prevent small precursor variants (below 250 Da) from being written into the output *.pkl* file.

In order to reduce the overall data amount, several filtering steps were added into the script. MS/MS data were reduced by filtering noise signals below a certain intensity threshold (default: 3 counts). During data acquisition, singly charged precursors were excluded from fragmentation and only precursors above *m/z* 350 were recorded. Both singly charged and small precursors are not expected to lead to confident peptide identifications and are typically excluded from fragmentation during data acquisition in any proteomics experiment. However, data reprocessing often revealed singly charged precursors that were sequenced due to incorrect charge state assignment during data acquisition. The corresponding spectra were of poor quality or corresponded to chemical contaminants. We included a filter into the perl script that would disregard all precursors with a mass below 600 Da. Finally, a filter removed precursor masses of small RNA oligonucleotides according to their fractional mass (precursor mass $<$ 1750 Da and fractional mass $<$ 0.2). These parameters were chosen conservatively according to published data (see [78, 80] and 1.3.4). For larger masses, the differentiation based on fractional mass is no longer unambiguous. Larger precursors were compared to the masses of oligonucleotides with a maximum length of 10; agreement was reported in a *.csv* file, but the precursors were not excluded from subsequent analysis.

Several approaches were taken to reduce the overall number of precursor mass variants to be created and thus later on searched. An option was included to accept an RNA sequence as input. Precursor mass variants are then generated for nucleotide combinations that actually appear in the input sequence. Especially for short RNAs, this can greatly reduce the number of precursor mass variants. Next, the precursor variants to be created were grouped:

1. all combinations of A, C, G, and U with the following modifications: none, –H_2O, –HPO_3, –H_3PO_4, +HPO_3, +HPO_3–H_2O
2. all combinations as in (1) with the 152 adduct
3. all combinations of A, C, G, U, and 4SU containing at least one 4SU in sequence, same modifications as in (1)

The perl script was set up to ask for an input sequence first and whether combinations should be calculated from the input sequence without and/or with 152 adducts, i.e., similar to group 1 and/or group 2, but limited to combinations appearing in the input sequence. Next, the user is asked whether combinations from group 1, 2, and/or 3 should be considered for the precursor mass variants. This way, the generated precursor variants can be chosen according to the experiment. If desired, precursor variants for each group can be created individually by running the script several times. While this leads to a decreased number of precursor variants in each search and thus might decrease the likelihood of false positives, the approach requires one search per group and spectrum.

In order to limit false positives and search time, we usually chose the different groups individually for precursor variant generation.

Another criterion to search for spectra containing RNA is the observation of RNA marker ions (see 1.3.3.3). Cross-links to RNA with at least two nucleotides typically produce intense marker ions, especially of the RNA bases adenine, guanine, and cytidine. The presence of both RNA marker ions and a peptide fragment series allows an experienced user to quickly assess a MS/MS fragment spectrum as a highly promising candidate for cross-link identification. However, this requires manual evaluation of all spectra contained in the measurement. Therefore, an automatic annotation of RNA marker ion intensities was included in the perl script.

To this end, signal intensities of the nucleic acid bases and nucleotides (see Table B.1) were checked within a certain mass tolerance. If one marker ion was observed above a certain threshold, e.g. 20% relative intensity, intensities for all marker ions present in the spectrum were reported in the output *.csv* file. These could be used to prioritize spectra for data analysis. Since the adenine marker ion and the tyrosine immonium ion have very similar masses (*m/z* 136.0623 and 136.0757, respectively), these can only be distinguished with high resolution instruments. The Q-ToF Ultima does not provide the necessary resolution. Therefore, the presence of adenine was only reported if both marker ions (base and nucleotide) were present: one above 20% and the second above 5% relative intensity.

Unfortunately, not all spectra of cross-links contain marker ions, especially those with single nucleotides. Therefore, the absence of marker ions cannot be used to exclude spectra of noncross-linked species. We still included it as an option in the perl script; if desired, precursor mass variants would only be created for spectra containing RNA marker ions. However, this option was not used in data analysis for this work.

3.1.3 Cross-linking products of 4-thio-uracil in the NusB–S10–BoxA RNA complex

In order to investigate the cross-linking products of 4-thio-uracil (4SU) by mass spectrometry and to test our novel data analysis approach, we cross-linked the NusB–S10 protein complex from *E. coli* to a synthetic RNA containing the *rrn* BoxA element. Three uracils of the oligonucleotide had been replaced by 4SU; the sequence was 5'-CAC $\underline{\text{UGC UC(4SU) (4SU)(4SU)A ACA}}$ AUU A-3'. The mass of the intact oligonucleotide and thus complete labeling was confirmed by ESI-MS.

Equal molar amounts of protein complex and RNA oligonucleotide were incubated for *in vitro* complex formation, cross-linked at 365 nm for five minutes, enriched with C18 and titanium dioxide chromatography according to our standard protocol, and investigated by LC-ESI-MS/MS on the Q-ToF Ultima. Data analysis was carried out with the perl script described above, assuming additive behavior of peptide and RNA masses in the cross-linking product, and manual spectra interpretation.

3.1.3.1 Additive cross-linking product of 4-thio-uracil

Initially, searches were performed after precursor mass variant generation with RNA combinations containing at least one 4SU nucleotide in the sequence. The only significant result obtained was a cross-link of the carbamylated NusB peptide SFGAEDSHKFVNGVLDK (S113–K129) to a 4SU

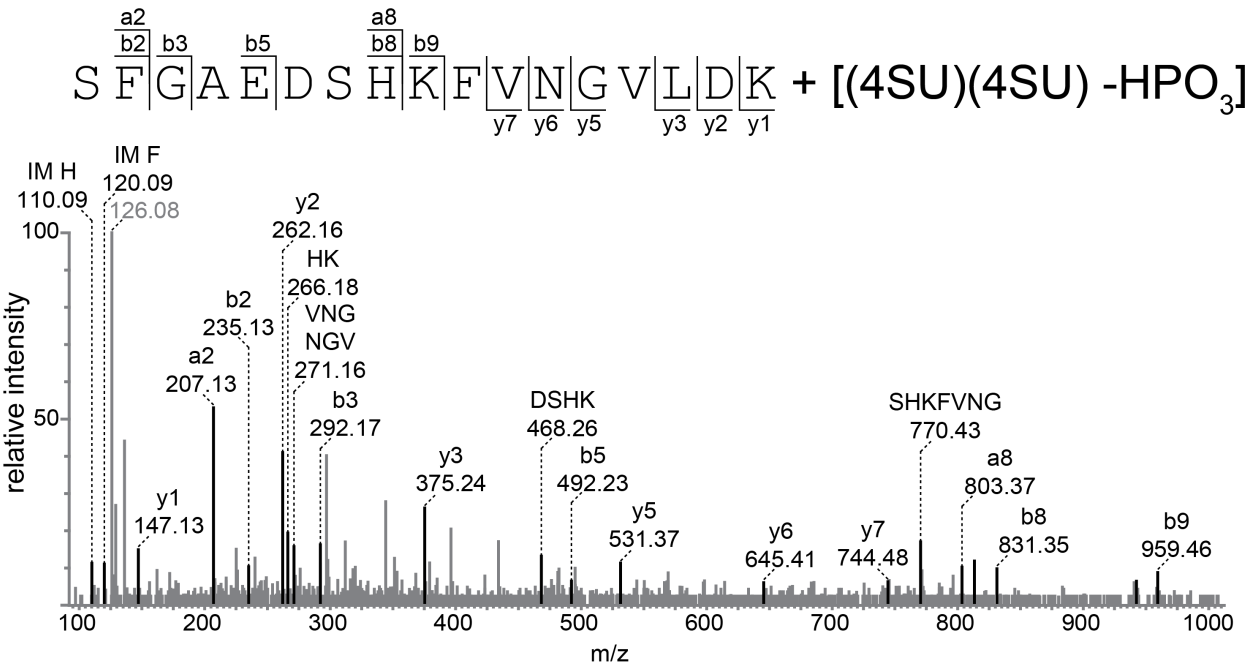


Figure 3.5: MS/MS fragment spectrum (smoothed and centroided) of carbamylated NusB peptide SFGAEDSHKFVNGVLDK (S113–K129) cross-linked to [(4SU)(4SU) –HPO_3]. Internal ions are indicated by their amino acid compositions, but not annotated to the peptide sequence above the spectrum. Since all peptide fragments and internal ions are observed without carbamylation, determination of the carbamylation site is inconclusive. The spectrum does not contain any obvious hint towards the cross-linked RNA or amino acid. The presence of carbamylation and the cross-linked 4SU dinucleotide without terminal phosphate is derived solely from the difference between the experimental precursor and calculated peptide mass.

dinucleotide without terminal phosphate. The cross-link was validated manually; the spectrum is shown in Figure 3.5 (general explanation of spectra annotation in B.1). However, some ambiguity remained: Mascot identified K121 as carbamylated; manually b9 and several internal ions containing K121 were annotated without carbamylation. Since several cross-links of the same peptide with and without carbamylation were identified later (see Table 3.5), it was concluded that the cross-link is a true positive result. Furthermore, the cross-link illustrates that 4SU can form cross-links that are additive with respect to mass, as it has been observed for unsubstituted RNA.

3.1.3.2 Identification of a novel, 4-thio-uracil specific cross-linking product

Several variations of the perl script and Mascot searches were carried out to test the feasibility of the approach. Among other things, searches were performed with precursor variants generated from unsubstituted nucleotides. Interestingly, we observed several matches for different peptides with a mass difference corresponding to a cross-linked [U –H_2O] nucleotide. Among these was the same peptide found cross-linked before, NusB peptide SFGAEDSHKFVNGVLDK. The corresponding spectrum is shown in Figure 3.6. A Mascot search of precursor mass variants generated for RNAs with at least one 4SU did not yield any search results for this spectrum, whereas a search with precursor variants for unsubstituted RNA resulted in a single match. The peptide was identified with a Mascot score of 16 and an E-value of 0.023. No other precursor variant led to a match.

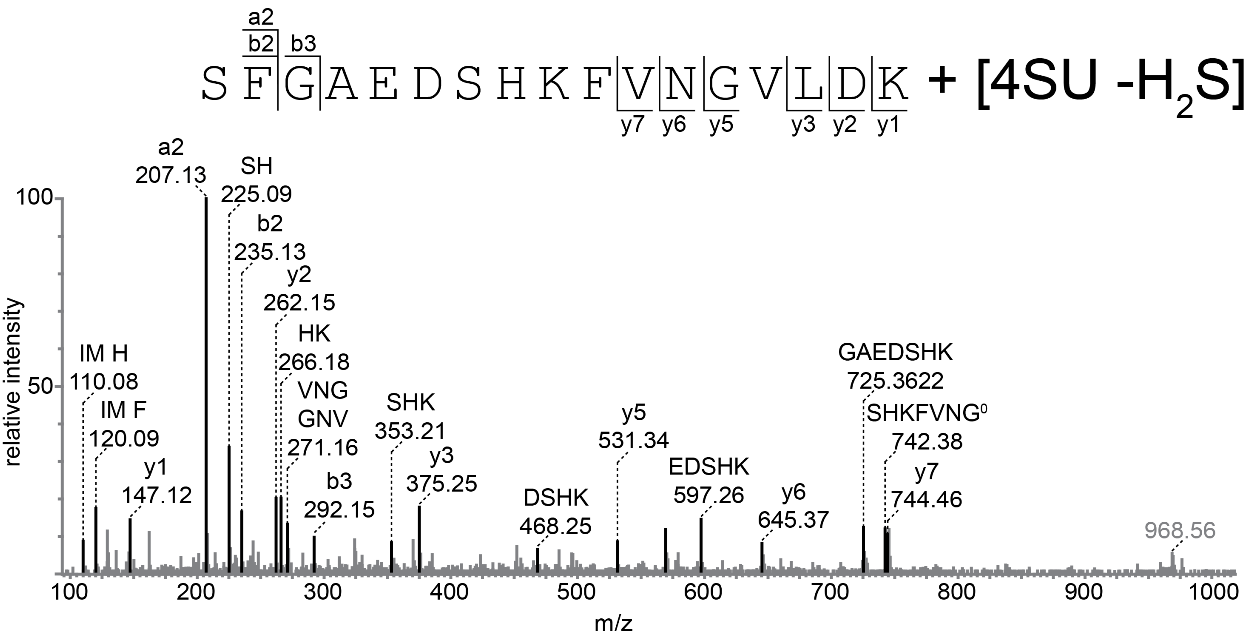


Figure 3.6: MS/MS fragment spectrum (smoothed and centroided) of NusB peptide SFGAEDSHKFVNGVLDK (S113–K129) cross-linked to [4SU $-H_2S$]. Internal ions are indicated by their amino acid compositions, but not annotated to the peptide sequence above the spectrum. The cross-linked RNA is derived from the difference between experimental precursor and calculated peptide mass. The spectrum does not allow unambiguous identification of the cross-linked amino acid. The distance between y7 and the signal at *m/z* 968.56 is 224.10 Da, this could correspond to phenylalanine (147 Da) plus 94 Da (see 3.8) minus NH_3. This could hint to F122 as the cross-linked amino acid.

The precursor mass variant that gave rise to the match was *m/z* 617.3228; the original experimental precursor mass was 719.3312 with a charge state of three. The difference between experimental precursor mass and precursor mass variant was 306.0252, corresponding to [U $-H_2O$]. The peptide was confirmed by manual validation; the cross-linked nucleotide could not be explained at that point. The Mascot search was repeated and the database was changed from a small database (see 2.2.10.1) to the *E. coli* proteome in the NCBI database. A search with 4SU-containing precursor variants led to nine matches, none of which were significant and thus all could be considered false positives. A search with precursor variants of unsubstituted RNA resulted in 45 matches. The validated peptide was the best-scoring hit, again with a score of 16, and an *E*-value of 11. The second best scoring hit had a score of 9 and an expect value of 16. Manual evaluation of both candidates revealed that the validated peptide had considerably more peptide fragment matches; thus, the second candidate was a random hit/false positive. In general, similar observations were made in the evaluation of Mascot search results for precursor mass variants: Candidates later validated manually typically were the only matches in the search against the small database. Searches against the *E. coli* proteome usually confirmed the hit.

Manual spectra interpretation, performed in parallel to the novel data analysis workflow, revealed another peptide with the same mass difference. The spectrum shown in Figure 3.7 contains a fragment series easily recognized by experts in spectra interpretation. A partial sequence tag, derived from fragments larger than *m/z* 500, quickly led to S10 peptide LIDQATAEIVETAKR (L17–R31). The difference between calculated peptide and experimental precursor mass was 306.0446 Da. This also hinted at [U $-H_2O$] (calculated mass 306.0253 Da) as the cross-linked nucleotide. More

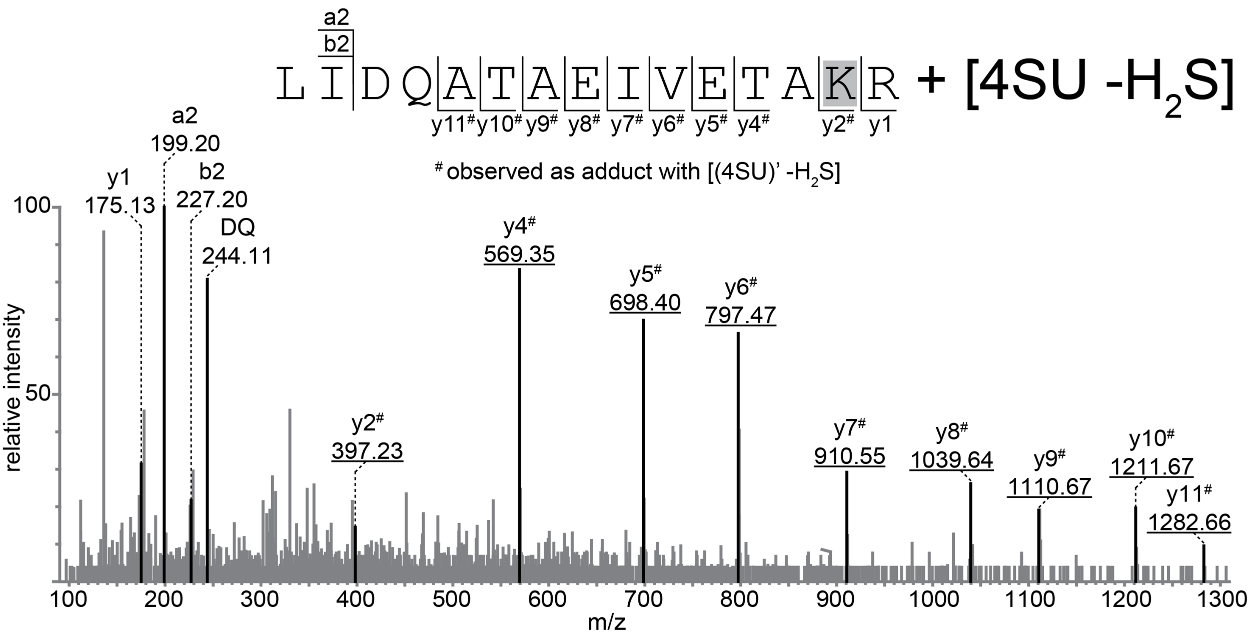


Figure 3.7: MS/MS fragment spectrum (smoothed and centroided) of S10 peptide LIDQATAEIVETAKR (L17–R31) observed as adduct with [4SU $-H_2S$]. All observed y-ions except for y1 are shifted by a mass of 94.0 Da, later interpreted as the 4SU base minus H_2S. Due to the shift, K30 was identified as the cross-linked amino acid.

astonishingly, the signals above *m/z* 500 were all shifted by 94 Da with respect to the calculated peptide fragment series. Closer inspection of the spectrum revealed that the same was true for all observed y-ions except for y1.

The possibility that the cross-linked nucleotide was actually [U $-H_2O$] had to be excluded. The absorption maxima of native nucleotides (250–270 nm) are far from that of 4-thio-uridine (330 nm) [35]. With UV irradiation carried out at the even higher wavelength of 365 nm, unsubstituted nucleotides are not excited and consequently not able to form cross-links. A modified nucleotide [4SU $-H_2S$] would have the same elemental composition as [U $-H_2O$] and, consequently, exactly the same mass. Moreover, cleavage of the N-glycosidic bond in [4SU $-H_2S$] would lead to a molecule with the composition $C_4H_2N_2O$ and a mass of 94.0167 Da. Cleavage of the N-glycosidic bond under our fragmentation conditions can be considered a likely fragmentation pathway, as the same cleavage leads to the nucleic acid base marker ions frequently observed. Potential structures for both the newly identified and the additive cross-linking products of 4-thio-uracil are shown in Figure 3.8.

Another peptide cross-linked to 4SU via the loss of H_2S is GPIPLPTR (G38–R44). Close manual inspection of the MS/MS fragment spectrum revealed b2, a3, and b3 shifted by 94 Da. The sequence ion y7 was observed without a shift, while the signal at *m/z* 944.57 corresponds to the intact peptide plus [(4SU)' $-H_2S$]. Therefore, the N-terminal G38 was identified as the cross-linked amino acid. This finding was particularly interesting since glycine had not been identified as the cross-linked amino acid before. Cross-linking of glycine might be a consequence of the considerably higher reactivity of 4SU.

The shift of the majority of peptide fragments as in the case of peptide LIDQATAEIVETAKR (see Figure 3.7) might hinder or completely prevent identification of the cross-link based on the regular peptide sequence ions. Therefore, a novel "post-translational" modification was defined in Mascot. The modification was defined with 306 Da on the peptide C-terminus. A neutral loss

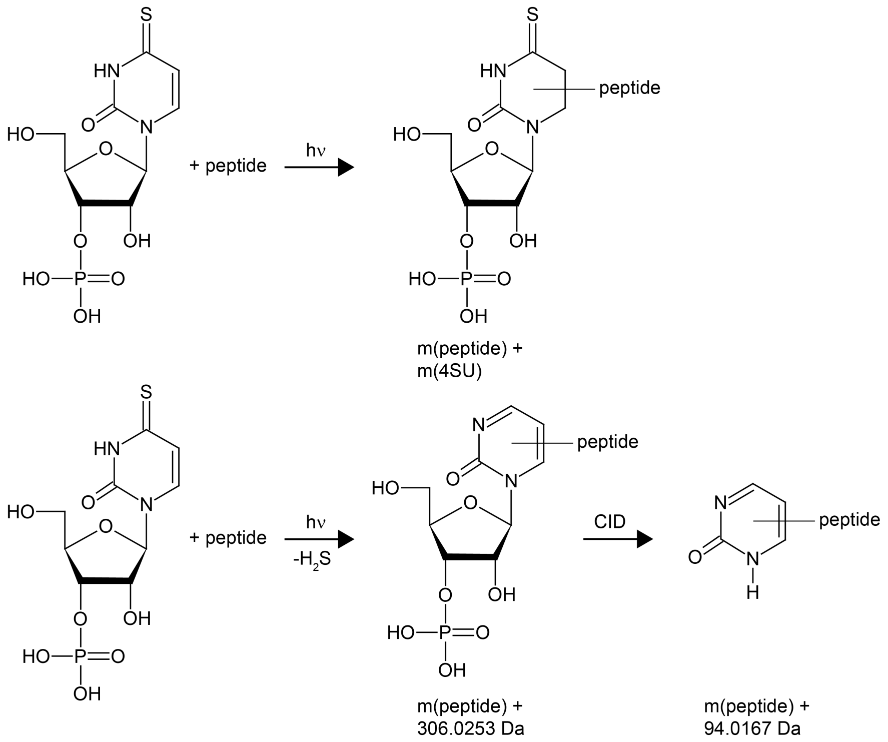


Figure 3.8: Possible structures of 4SU cross-linking products. Additive cross-linking (upper reaction) leads to a reaction product where the peptide (amino acid) is bound to 4SU, probably via C5 or C6 of the 4SU base. Cross-linking might lead to hydrolyzation of the C5–C6 double bond; the mass of the reaction product equals the sum of the masses of the single reactants. Alternatively, 4SU reacts with an amino acid residue under net loss of H_2S (lower reaction). The mass of the cross-linking product equals the mass of the peptide (amino acid) plus 306 Da. Neither the exact structure nor the exact cross-linking site on the base can be determined by mass spectrometry. Thus, the structure represents one of several possible product structures. Upon fragmentation with collision-induced dissociation (CID), the N-glycosidic bond is cleaved, leading to peptide fragments shifted by 94 Da.

(212 Da) was added that would leave 94 Da as the observable shift of peptide fragments (for details, see 2.2.10.2). When a regular database search was performed including this modification, the cross-linked peptide LIDQATAEIVETAKR was identified successfully. In contrast, the spectrum of cross-linked peptide GPIPLPTR (see Figure 3.9) contains only a few shifted peptide fragments and was identified after application of the precursor variant approach. In a regular database search with the newly defined modification, it would not be identified because it is cross-linked on its N-terminus. Therefore, the standard database search with the novel modification and the precursor variant approach present complementary strategies for the identification of peptides cross-linked to [4SU $-H_2S$] with a corresponding shift of peptide fragments by the mass of [(4SU)' –H2S].

3.1.3.3 Observation of peptides with a 258 Da adduct

In further tests of data analysis parameters, enzyme specificity was set to "unspecific". Interestingly, this yielded a result for DVPYKVAINEAIELAK (NusB); the difference between experimental precursor and calculated mass of the unspecific peptide suggested [G $-H_2O$] as the cross-linked nucleotide. As unsubstituted uracil, guanine should be unreactive under the irradiation conditions of this experiment. Manual assignment of the corresponding spectrum (Figure 3.10) led to the

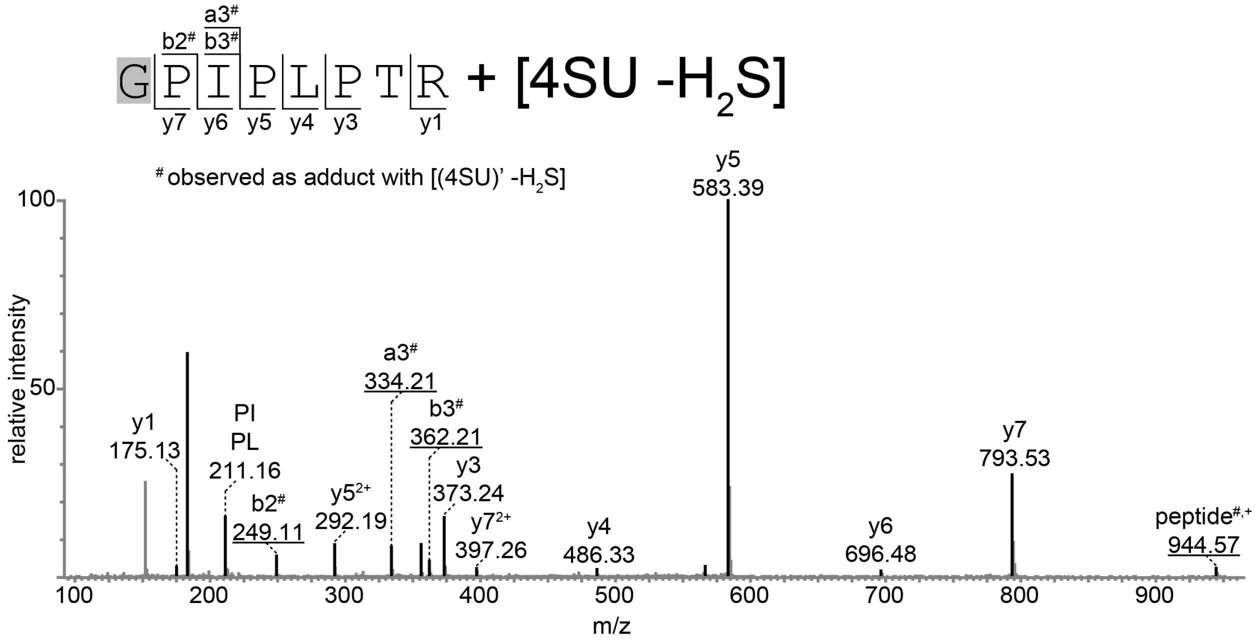


Figure 3.9: MS/MS fragment spectrum (smoothed and centroided) of S10 peptide GPIPLPTR (G38–R44) cross-linked to [4SU $-H_2S$]. Cleavage N-terminal to proline leads to high intensity signals as expected (proline effect). Interestingly, b2, a3, b3, and the intact peptide are observed as adduct with [(4SU)' $-H_2S$], i.e. shifted by a mass of 94 Da. This identifies G38 as the cross-linked amino acid.

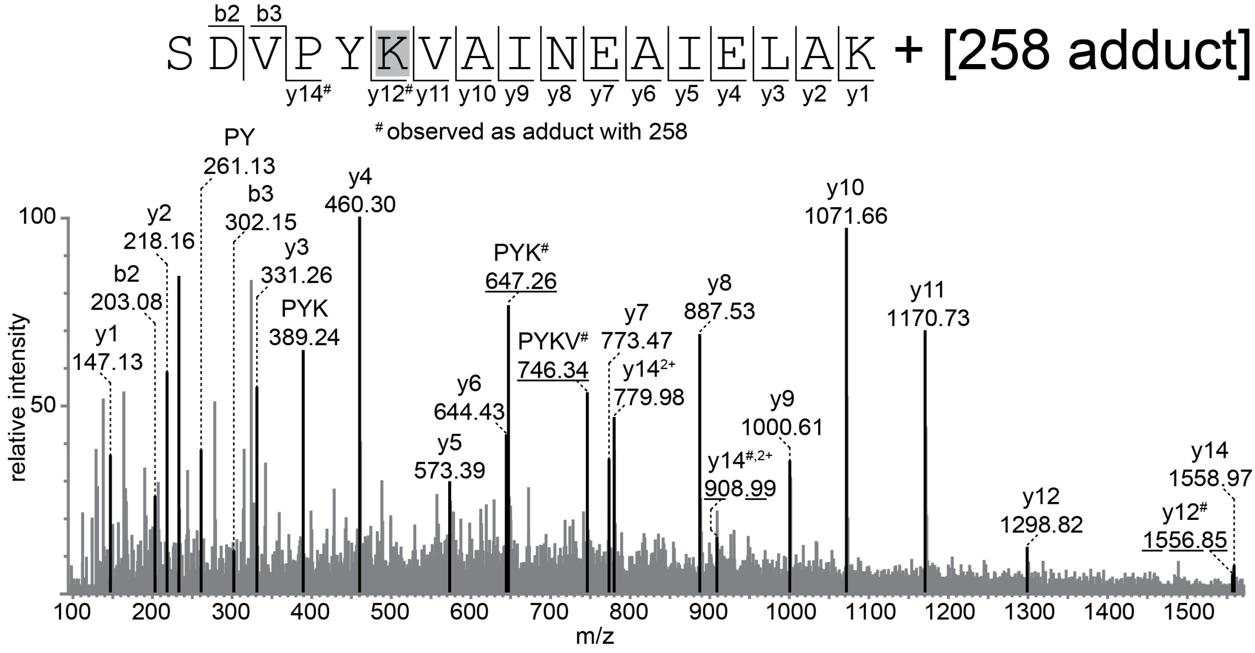


Figure 3.10: MS/MS fragment spectrum (smoothed and centroided) of NusB peptide SDVPYKVAINEAIELAK (S96–K112) observed as adduct with 258 Da. The peptide fragment signals y12 and y14 as well as internal ion PYK are observed as partially shifted by 258 Da. This suggests K101 as the cross-linked amino acid.

conclusion that the peptide was tryptic and not unspecific as suggested by the database search. In consequence, the mass difference between experimental precursor mass and calculated peptide mass changed to 258.0597 Da. Since y12, y14 and two internal ions containing K101 were observed shifted by the same mass, K101 was identified as the cross-linked amino acid. The same mass difference was found for several peptides after it was included as an additional option for precursor variant generation (see Table 3.1, Figures B.2 and B.1). Therefore, we initially interpreted this mass as a 4SU-specific cross-linking product [104]. However, it was later shown by Dr. Uzma Zaman (Bioanalytical Mass Spectrometry Group; unpublished) that this adduct is UV induced but observed independently from irradiation wavelength and RNA substitution.

3.1.3.4 Feasibility of the precursor variant approach

After the observed cross-linking products had been adapted into our data analysis workflow, either as modifications for standard Mascot search or as masses for precursor mass variation, all cross-linked peptides reported were identified unambiguously by the novel data analysis approach. If precursor variants were searched against a small database (14 proteins, see 2.2.10.1), the cross-linked peptide was the only hit. In searches against the *E. coli* proteome, the cross-linked peptide was typically the best-scoring match. Other hits could be excluded as false positives, either due to low scores, manual evaluation of peptide fragment matches, or simply because the corresponding protein was not contained in the sample. Table 3.1 summarizes all peptides of NusB and S10 that were found cross-linked and/or observed with the 258 Da adduct. Table A.1 additionally lists the corresponding mass values.

Table 3.1: Cross-links of the NusB–S10 complex

protein	position	peptide	aa	RNA	figure
NusB	I87–R95	IALYELSKR	K94	258 adduct	B.1
	S96–K112	SDVPYKVAINEAIELAK	K101	258 adduct	3.10
			K101	[4SU $-H_2S$]	-
	S113–K129	SFGAEDSHKFVNGVLDK	-	258 adduct	-
			F122	[4SU $-H_2S$]	3.6
			-	[(4SU)A $-H_2S$]	-
			-	[(4SU)C $-H_2S$]	-
		SFGAEDSHKFVNGVLDK	-	[4SU $-H_2S$]	-
		(carbamylated)	-	[(4SU)(4SU) $-HPO_3$]	3.5
			-	[(4SU)A $-H_2S$]	-
S10	L17–R31	LIDQATAEIVETAKR	K30	[4SU $-H_2S$]	3.7
	G38–R44	GPIPLPTR	G38	[4SU $-H_2S$]	3.9
	L73–R89	LVDIVEPTEKTVDALMR	-	258 adduct	B.2
		LVDIVEPTEKTVDALM(Ox)R	-	[4SU $-H_2S$]	-

protein: name of the cross-linked protein
position: position of the cross-linked peptide in the protein sequence
peptide: sequence of the cross-linked peptide
aa: one-letter code and position of the cross-linked amino acid
RNA: composition of the cross-linked RNA/observation of the 258 Da adduct
figure: reference to figure showing a representative spectrum of the cross-link

3.1.4 Comparison of obtained cross-linking results to cross-linking with unsubstituted RNA

In this section, the identified cross-links will be compared to the preceding study focused on the same biological system. In addition, the obtained insights on cross-linking of 4-thio-uracil will be summarized. A more general discussion on the cross-linking technique and method development will follow in Chapter 4.

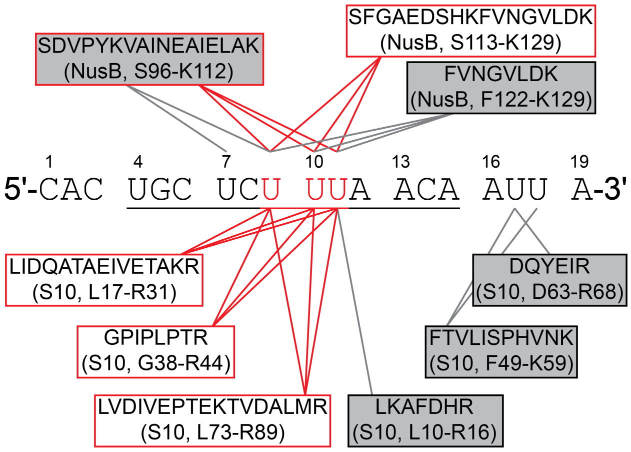


Figure 3.11: Comparison of NusB cross-links to unsubstituted [71] and 4-thio-uracil substituted *rrn* BoxA containing RNA oligonucleotides. The sequence of the oligonucleotide is given, the core BoxA element is underlined, and nucleotides substituted by 4SU are in red. Nucleotide numbers are indicated above the sequence. Cross-linked peptides with their positions in the respective protein are given, with NusB peptides above and S10 peptides below the sequence. Peptides found cross-linked to the 4SU-substituted oligonucleotide are shown in red boxes, peptides cross-linked to unsubstituted RNA with a gray background. Lines between the peptide and RNA sequence indicate potential cross-link positions on the RNA.

We identified several cross-links of the NusB–S10 complex to the *rrn* BoxA containing RNA oligonucleotide 5'-CAC UGC UCU UUA ACA AUU A-3' with three uracils (underlined) substituted by 4-thio-uracil. These results were compared to cross-linking results with the same, unsubstituted oligonucleotide previously obtained in our laboratory [71]. Since the 258 Da adduct is not a clear product of a cross-linking reaction to RNA, peptides found exclusively with this adduct are not considered for this comparison. As uracil is by far the most reactive nucleic acid base, we assume that all cross-links of the unsubstituted RNA occurred via U for the subsequent comparison. Identified cross-links of both experiments are summarized in Figure 3.11.

Our cross-linking results for the NusB protein show a great overlap with the cross-links to unsubstituted RNA. Peptide SDVPYKVAINEAIELAK (S96–K112) had been found cross-linked to a [CU] dinucleotide. Complementary experiments led to the conclusion that either the C6–U7 or the C8–U9 stretch of the *rrn* BoxA oligonucleotide was cross-linked [71]. We found the same peptide cross-linked to 4SU. Since the nucleotide at position 9 was one of the substituted nucleotides, it is possible that it formed cross-links to the peptide in both experiments.

The second NusB peptide identified as cross-linked to the 4SU-substituted RNA, SFGAEDSHK-FVNGVLDK (S113–K129), contains a missed cleavage site of trypsin. Previous experiments had identified the peptide FVNGVLDK (F122–K129) without a missed cleavage site as cross-linked to unsubstituted *rrn* BoxA containing RNA. The missed-cleaved peptide SFGAEDSHKFVNGVLDK had been found cross-linked to the γ NutR BoxA containing RNA oligonucleotide 5'-CAC CGC UCU UAC ACA AUU A-3'. There are two possible explanations for this observation: K121 might be not very accessible to trypsin, especially with a cross-link in close proximity, e.g. F122. Therefore, hydrolysis might be hindered or prevented. Alternatively, K121 itself might be cross-linked, preventing cleavage by trypsin completely, whereas unsubstituted *rrn* BoxA cross-links to the neighboring F122. Since neither data set allowed for unambiguous identification of the cross-linked residue, this hypothesis cannot be proven nor rejected.

Peptide FVNGVLDK (F122–K129) had been found cross-linked to a [UU] dinucleotide. Cross-linking of NusB alone had led to a cross-link of the same peptide to a [UUU] trinucleotide, which has to be the triple U stretch substituted in our experiment. Complementary experiments with the γ NutR BoxA oligonucleotide indicate U11 as the cross-linked nucleotide [71]. Peptide SFGAED-SHKFVNGVLDK (S113–K129) was found cross-linked to [(4SU)A $-H_2S$] as well as [(4SU)C $-H_2S$], indicating that it might cross-link to positions 9 and 11 of the oligonucleotide. The results for the unsubstituted RNA do not allow for the exclusion of the possibility that the protein region contacts both nucleotides.

Overall, the cross-linking results for NusB with both unsubstituted and 4SU-substituted *rrn* BoxA containing RNA oligonucleotides are in excellent agreement on both protein and RNA level. The NusB cross-linking yield did not differ significantly between unsubstituted and 4SU-substituted *rrn* BoxA containing RNA oligonucleotides (see Figure 3.1), which indicates that both protein regions exhibit high cross-linking reactivity, even with the less reactive unsubstituted uracils. The two NusB residues K101 and F122 identified in our study as cross-linked to 4SU were later shown to be significant for binding of NusB to BoxA by NMR chemical shift mapping. Moreover, a F122D mutant did not exhibit any detectable BoxA binding [106]. This illustrates the capability of UV induced cross-linking with mass spectrometry to identify direct interaction sites on an amino acid level.

In contrast to the results for NusB, cross-linking of S10 to the 4SU-substituted RNA differs substantially from cross-linking to unsubstituted RNA. S10 peptides LKAFDHR (L10–R16), FTVLISPH-VNK (F49–59), and DQYEIR (D63–R68) had been identified as cross-linked to the unsubstituted *rrn* BoxA containing RNA oligonucleotide [71]. In our study, peptides LIDQATAEIVETAKR (L17–R31), GPIPLPTR (G38–R44), and LVDIVEPTEKTVDALMR (L73–R89) were found cross-linked to the same oligonucleotide substituted with 4SU.

Peptide DQYEIR (D63–R68) had been identified as cross-linked to an [AAU] trinucleotide and it had been concluded that it was cross-linked to the AAU-stretch close to the 3' end of the oligonucleotide, probably via U17. Since the corresponding uracil was not one that was substituted in our experiment, we could not have identified the same cross-link. The same is true for peptide FTVLISPHVNK (F49–K59) that had been found cross-linked to [AAUU]: Oligonucleotides with this composition only appear at the 3' end of the *rrn* BoxA containing RNA oligonucleotide. In contrast, peptide LKAFDHR (L10–R16) had been found cross-linked to an [AU] dinucleotide. From results of complementary experiments [71], the cross-link had been mapped to the UA stretch containing the U11, substituted in our experiment. If the same cross-linking behavior was expected for both unsubstituted and 4SU-substituted RNA, this peptide should have been identified in our experiment. However, because we identified completely different peptides cross-linked to the 4SU-substituted RNA, the cross-linking behavior is clearly different.

There is currently no further satisfactory explanation for the discrepancy of S10 cross-linking results between unsubstituted and 4SU-substituted *rrn* BoxA containing RNA oligonucleotides. Interestingly, the cross-linked S10 residue G38 is in direct proximity to residues that interact with NusB (and 16S rRNA in the 30S subunit), namely residues R27 and P39 [71]. Therefore, it is located at the binding interface of both proteins. The role of this binding interface and the additional peptides found cross-linked to 4SU in RNA binding could only be answered by additional biochemical and structural studies.

An interesting detail of the cross-linking results with 4SU (see Table 3.1) is the high number of peptides with a missed cleavage site. More precisely, lysine residues with C-terminal peptide bonds that were not hydrolyzed by trypsin. Due to peptide sequence ions observed as adducts with RNA fragments, K101 of NusB and K30 of S10 were identified as cross-linked residues. This might indicate an increased reactivity between 4SU and lysine residues.

Overall, cross-linking experiments of the NusB–S10 complex to 4-thio-uracil substituted *rrn* BoxA containing RNA oligonucleotide provided valuable insights into UV-induced cross-linking in combination with mass spectrometry in general. It was shown that 4-thio-uracil has the potential to increase the cross-linking yield and that the cross-linking products can be identified by mass spectrometry. A non-additive cross-linking product of 4SU was identified; 4SU can undergo UV-induced reactions with amino acids via net loss of H_2S and the resulting cross-linking product displays interesting behavior under collision-induced dissociation conditions. The N-glycosidic bond between the modified base and ribose is cleaved, leaving a stable adduct of peptide (fragments) and a cross-linked base that is extremely useful in the identification of the cross-linked amino acid residue. Most importantly, the precursor variant approach proved feasible for automated identification of anticipated cross-links, greatly improving the data analysis procedure for future projects. Since the majority of cross-links could be identified in searches against the background of the entire proteome, the approach was a valuable step towards the identification of cross-links in more extended ribonucleoprotein complexes.

3.2 Instrumental and data analysis improvements and their implications for cross-link identification

The basic idea for automatization of data analysis based on precursor variant generation proved its feasibility in cross-link identification in the NusB–S10 complex. The practical realization with a perl script provided a useful tool, almost completely replacing *ab inicio* manual spectra interpretation. In collaboration with the group of Prof. Oliver Kohlbacher (Applied Bioinformatics Group, Universität Tübingen), the approach was next integrated into the mass spectrometry data processing environment OpenMS. This created the opportunity to directly combine precursor variant generation with preceding and subsequent data analysis steps, thus enabling further automatization of the entire data analysis workflow.

At the same time, the laboratory obtained an LTQ Orbitrap Velos mass spectrometer. Compared to the preceding LTQ Orbitrap generation, it had been reported to provide higher sensitivity and scan speed. In addition, the instrument contained an improved cell for higher-energy collisional dissociation (HCD) [107]. The Orbitrap Velos was evaluated for the analysis of cross-linking experiments.

Both developments in instrumentation and data analysis will be described in detail before illustrating the achieved improvements exemplified by cross-link identification in the ASH1 complex.

3.2.1 LTQ Orbitrap Velos mass spectrometer

Mass spectrometric analysis of UV induced protein–RNA cross-linking experiments had been carried out on a Q-ToF Ultima mass spectrometer. It was preferred to the LTQ Orbitrap XL mass spectrometer also available in our laboratory due to the unfavorable characteristics of CID spectra and insufficient sensitivity of HCD (see 1.3.3.4).

A newer generation of orbitrap instruments, the LTQ Orbitrap Velos [107], showed improvement in scan speed and sensitivity. Ion transmission from atmospheric pressure was improved by introducing the *stacked ring ion guide* or *S-lens.* More effective ion transmission in turn increased the overall sensitivity of the instrument. The HCD collision cell was re-engineered: Additional electrodes changed the electric field distribution to allow for more efficient extraction of ions from the HCD cell. Thus, sensitivity and scan speed of HCD fragmentation were significantly increased. In addition, CID sensitivity, fragmentation efficiency, and scan speed were increased by introduction of a dual-pressure linear ion trap.

Initial reports on the performance of the Orbitrap Velos suggested that the instrument was capable of analyzing complex samples with a *high-high* strategy, i.e. both MS survey and MS/MS (CID or HCD) fragment spectra are recorded in the orbitrap at high resolution [107]. Therefore, after an Orbitrap Velos became available in our laboratory, it was expected that the advancements would enable the analysis of cross-linking experiments on this instrument with HCD fragmentation.

3.2.2 Integration of the precursor variant approach into the OpenMS environment

A collaboration with the Applied Bioinformatics Group (Prof. Kohlbacher, Universität Tübingen) was initiated for further automatization of data analysis based on the established precursor variant approach. Together with other bioinformatics laboratories, they have developed and are constantly improving OpenMS [99,100], a library of algorithms and tools for the analysis of LC-ESI-MS/MS data. OpenMS offers various building blocks for different steps of LC-ESI-MS/MS data analysis that can be combined into analysis pipelines. Due to this feature, it was possible to create a specific building block for precursor variant generation that could then be combined with existing algorithms to create a data analysis routine for our cross-linking data. Programming of the necessary code was primarily done by Timo Sachsenberg (Kohlbacher laboratory).

OpenMS itself is a freely available, open source software package. Consequently, it is based on open mass spectrometry data formats like *.mzXML* [108] and *.mzML* [101]. Mass spectrometry data must be converted from vendor formats to one of these open formats before using OpenMS.

The development of an OpenMS data analysis workflow was first based on Q-ToF data. However, both the initiation of the collaboration and first measurements on the Orbitrap Velos were done almost in parallel, so the focus was quickly set entirely on Orbitrap data. The age of the Q-ToF Ultima implies an outdated data structure that presented various challenges in different stages of analysis with OpenMS. While initial results proved that Q-ToF data could be analyzed with OpenMS, different data structure issues would have had to be resolved for complete analysis. Since Velos data was more straightforward to process and soon proved to yield better cross-linking results (see below), complete reproduction of cross-link identifications obtained with the perl script was abandoned.

Similarly, the perl script was set up to generate precursor variants from the text-based *.msm* file obtained from Velos data after conversion with Raw2MSM [98]. The script was easily modified for *.msm* files by Petra Hummel. However, the significantly higher number of MS/MS fragment spectra made manual submission of Mascot searches unfeasible, e.g., a single measurement of the ASH1-short complex (see below) contained 2677 MS/MS spectra. Therefore, efforts were completely set on analysis of Velos data with OpenMS.

All data processing steps described for the perl script were essentially transferred into a novel OpenMS tool named RNP$^{\mathrm{xl}}$. Prior to precursor variant generation, precursors below 600 Da and precursors corresponding to small oligonucleotides according to their fractional mass were filtered. Precursor variants were then generated as described, i.e. for all RNA sequences of one to four nucleotides and with the modifications –H_2O, –HPO_3, –H_3PO_4, +HPO_3, and +HPO_3-H_2O. Additionally, all RNA combinations were considered with and without the 152 Da adduct. Precursor variants were omitted from further analysis if they were below *m/z* 250. As in the perl script, the original precursor mass was included with all its variants to identify noncross-linked peptides even after precursor variant generation.

Two steps contained in the perl script were omitted from RNP$^{\mathrm{xl}}$: The noise filter was not integrated because Orbitrap data contains considerably fewer noise signals. The comparison of precursor masses to larger RNA oligonucleotides was also omitted from the data analysis workflow because the information obtained had not provided any practical use.

Most importantly, and in contrast to the perl script, the RNP$^{\mathrm{xl}}$ tool was designed to automatically submit the database searches. The actual database search was carried out with OMSSA (Open Mass Spectrometry Search Algorithm [34]) instead of Mascot. The main reason for this change was the higher speed of the OMSSA search engine. In addition, OMSSA is a freely available open source

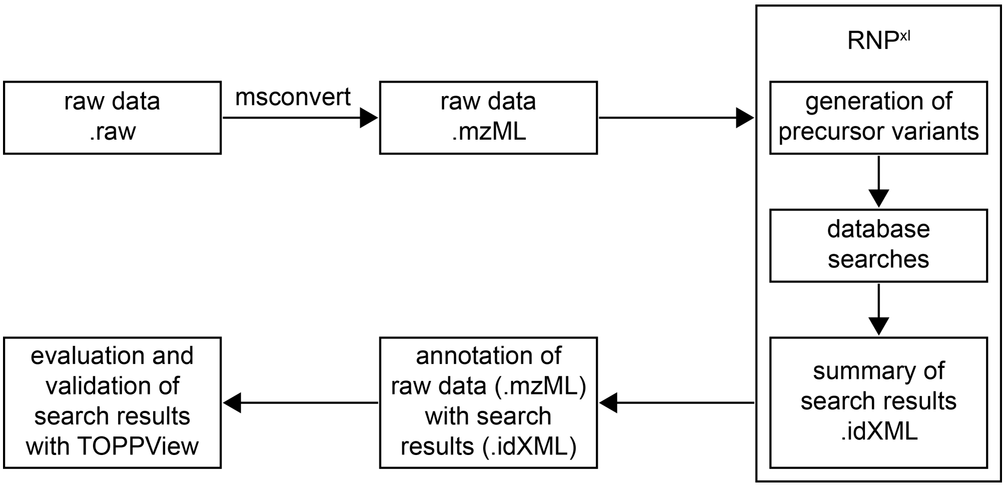

RNP^{xl}
raw data .raw
msconvert
raw data .mzML
generation of precursor variants
database searches
summary of search results .idXML
annotation of raw data (.mzML) with search results (.idXML)
evaluation and validation of search results with TOPPView

Figure 3.12: Data analysis workflow with the RNP^{xl} tool. Raw data in the *.raw* vendor format is converted into the open *.mzML* format by msconvert. The OpenMS tool RNP^{xl} generates the precursor variants, submits the database searches into OMSSA, and summarizes the search results into a single *.idXML* file. These can be annotated to the raw data in *.mzML* format in TOPPView.

program and can be installed alongside OpenMS, thus running both programs on the same machine. In contrast, Mascot is a commercial search engine and is typically run on separate dedicated servers. While OpenMS could also be set up to submit the searches to Mascot, the implementation is less straightforward due to the separate locations in which the programs are run.

In contrast to the perl script, the RNP^{xl} tool took the raw mass spectrometry data in *.mzML* format as input files. Internally, one MS/MS data file per spectrum containing the precursor variants was generated for submission of database searches. The search results were automatically summarized into an *.idXML* file, retaining only the best scoring match for each precursor variant *.mzML* file. In other words, only the best hit for each spectrum was reported in the *.idXML* file. This file was used to annotate the search results to the raw MS data in the OpenMS graphical user interface TOPPView. The entire data analysis workflow is outlined in Figure 3.12.

Data analysis of the cross-linking experiments described in this and the following section were done with the basic functionalities described above. In the developmental phase, precursor variant generation and database searches were performed by our collaborators. The final version of the RNP^{xl} tool was designed to enable analysis of a wide range of UV induced cross-linking experiments and to provide the user with many opportunities to influence the precursor variant generation.

Several options were included to allow the user to determine the RNA combinations that would be used for precursor variant generation. First of all, nucleotides can be defined by a one-letter code and their sum formula. Thus, experiments with photo-reactive nucleotides like 4-thio-uracil, DNA, or nucleotides labeled with heavy stable isotopes can be analyzed. As for the perl script, RNA combinations can be limited to those appearing in an input RNA sequence. RNA combinations can also be required to contain a certain nucleotide. For example, cross-linking experiments with 4-thio-uracil and UV irradiation at 365 nm should only yield cross-links to RNA containing at least one 4SU. For such experiments, the tool can be set up to only create precursor mass variants for corresponding RNA oligonucleotides, disregarding all combinations without 4SU. The maximum length expected for the cross-linked oligonucleotide can also be chosen freely. Additionally, the modifications to be considered for each RNA combination can be defined. Therefore, any effect observed in cross-linking experiments can be integrated, e.g. the loss of H_2S from 4SU upon cross-link formation.

The thresholds for the mass filters, i.e., the filters of small precursor masses prior to precursor variant generation and the *m/z* threshold for precursor variants to be included in the output file, were also included as parameters that can be adjusted by the user. The database search parameters, e.g. protein database and mass deviation, have to be defined by the user as well. The final version of RNP^{xl} is described in more technical details in 2.2.10.6, with Figure 2.3 showing the parameters.

3.2.3 Cross-linking of ASH1

Cross-linking of a model complex for *ASH1* mRNA transport in budding yeast was carried out with both experimental and data analysis setups, i.e. the combination of Q-ToF Ultima and perl script as well as Orbitrap Velos and OpenMS.

The ASH1 model complex had been studied by our collaborators, the group of Prof. Dierk Niessing (Institute of Structural Biology, Helmholtz Zentrum München). This tertiary complex contained the two proteins She2p and She3p, and a 51 nucleotide section of zip code element E3 of the *ASH1* mRNA. She2p is an RNA-binding protein that interacts with all four zip code elements of *ASH1* mRNA. *ASH1*-E3 zip code alone can mediate mRNA transport *in vivo.* Our collaborators identified She3p as a novel RNA-binding protein and showed synergistic binding of She2p and She3p to zip code containing RNAs with significantly higher affinity and specificity than the individual proteins. UV induced cross-linking with labeled RNA and denaturing gel electrophoresis proved that both proteins directly interact with RNA (see [96] and references therein).

Analysis of cross-linking products with mass spectrometry was intended to provide more detailed information on the interaction sites of both proteins. Two model complexes were prepared by Roland Heym (group of Dierk Niessing): The first complex, termed ASH1-FL, contained full-length She3p, She2p, and the RNA. The second complex, termed ASH1-short, contained the C-terminus of She3p (92 amino acids, positions 334–425, termed She3p-short, sufficient for synergistic RNA binding [96]), She2p, and the RNA. We obtained complexes that had been isolated by gel filtration. The complexes were UV irradiated at 254 nm for 10 minutes and enriched according to our standard protocol (see 2.2.8.4). Starting amounts for experiments with Q-ToF analysis were 120 µg ASH1-FL and 100 µg ASH1-short (protein amounts). Experiments with analysis on the Orbitrap Velos were carried out with significantly lower amounts due to the higher sensitivity of the instrument, i.e. 20 µg ASH1-FL and 25 µg ASH1-short.

Data from the Q-ToF Ultima was analyzed after precursor variant generation by the perl script. Mass variants were only created for RNA sequences that actually appear in the RNA used in the experiment, namely 5'-AUG GAU AAC UGA AUC UCU UUC AAC UAA UAA GAG ACA UUA UCA CGA AAC AAU-3'. This reduced the number of RNA combinations from 69 to 45. Database searches were performed with Mascot. Orbitrap data was analyzed with OpenMS and OMSSA searches without limiting the RNA combinations.

3.2.4 ASH1 cross-links identified after LC-ESI-MS/MS measurement on the Q-TOF and data analysis with the perl script

All cross-links identified in the ASH1-FL and the ASH1-short complexes in the initial experiments with the Q-ToF/perl script setup are listed in Table 3.2, with corresponding mass values given in Table A.2. In both complexes, She2p peptide IGSNLLDLEVVQFAIK (I164–K179) was found cross-linked to uracil. No cross-link of She3p was identified in ASH1-FL. The cross-linked peptide GPLGSMGNSSNNK identified after cross-linking of ASH1-short contains the N-terminus of the shortened She3p (positions G334–K340) and a stretch of six amino acids (underlined) that are not part of She3p, but remained after cleavage of a GST tag introduced for protein isolation.

Table 3.2: Cross-links of the ASH1 complexes identified after analysis on the Q-ToF Ultima.

protein	position	peptide	aa	RNA	figure	FL	short
She2p	I164–	IGSNLLDLEVVQFAIK	F181	[U]	3.13	+	+
	K179	I*GSNLLDLEVVQFAIK *carbamylated	F181	[U]	-	+	+
She3p	G334–	GPLGSMGNSSNNK	S337–N339	[AAU –HPO$_3$]	3.14b	-	+
-short	K340	G*PLGSMGNSSNNK *carbamylated	-	[U]	3.14a	-	+
		GPLGSM*GNSSNNK *oxidized	-	[U]	-	-	+

protein: cross-linked protein
position: position of the cross-linked peptide in the protein sequence
peptide: sequence of the cross-linked peptide
aa: one-letter code and position of the cross-linked amino acid
RNA: composition of the cross-linked RNA
figure: reference to figure showing representative MS/MS fragment spectrum
FL/short: identification of the cross-link in ASH1-FL and ASH1-short, respectively

Cross-links of She2p peptide IGSNLLDLEVVQFAIK were identified after database searches with high confidence. In all cases, Mascot searches after precursor variant generation unambiguously identified the peptide as the best-scoring match in searches against a reduced database (see 2.2.10.1) or the *S. cerevisiae* proteome as contained in the NCBI database. The cross-linked peptide was identified with high Mascot scores, the unmodified peptide scored between 62 and 72, and scores for the carbamylated peptide were 49 to 59.

A spectrum of this cross-link is shown in Figure 3.13. Close manual inspection revealed that the y-series was partially shifted by the mass of [U –H$_3$PO$_4$] and [U]. Since the smallest y-ion observed with an RNA adduct is y4, it was concluded that F181 was the cross-linked amino acid.

The spectrum also exemplifies the influence of RNA-adducts on the fractional mass of peptide fragments. Fractional masses of fragments in the same m/z range can indicate absence or presence of RNA (fragments). For example, the signals at m/z 1029.44, 1030.51, and 1046.64 show increasing fractional masses. The signal at m/z 1046.64 was annotated as y9 in the Mascot search results. Manual spectra interpretation revealed the signal at m/z 1030.51 corresponding to an adduct of y7 with [U –H$_3$PO$_4$] and m/z 1029.44 as y6 with [U]. Thus, the effect on the fractional mass typically increases with the size of the RNA adduct. This observation can be extremely helpful in manual spectra interpretation. It must be noted that a- and b-ions also exhibit a fractional mass smaller

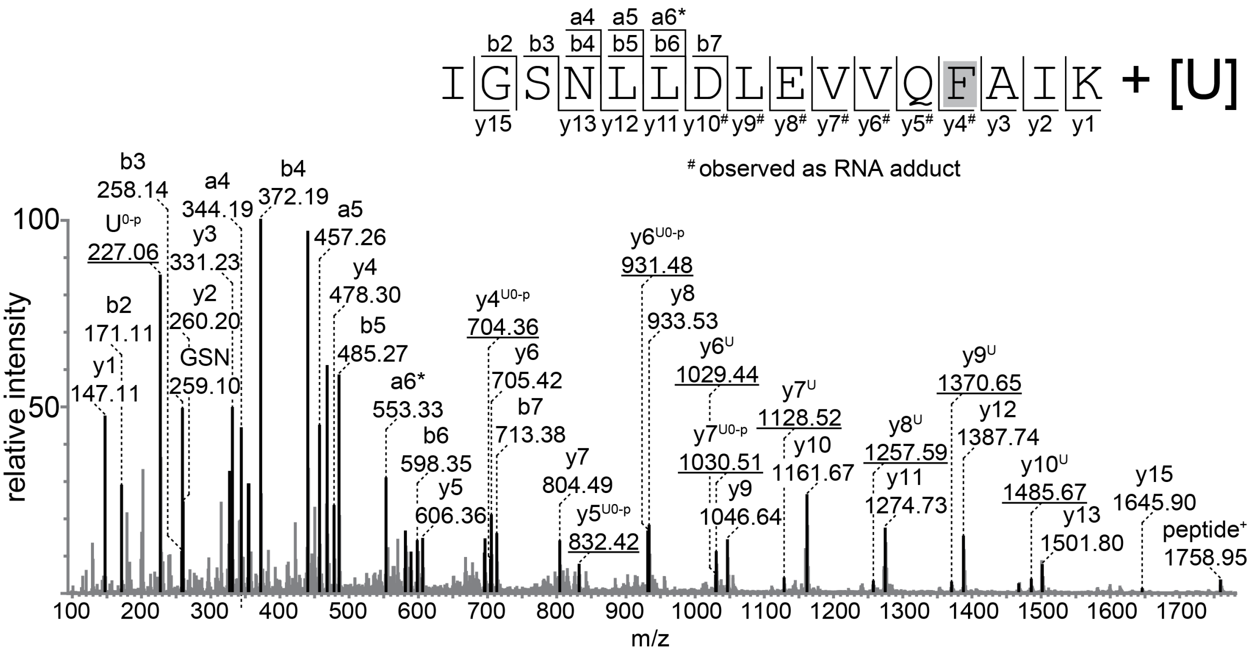
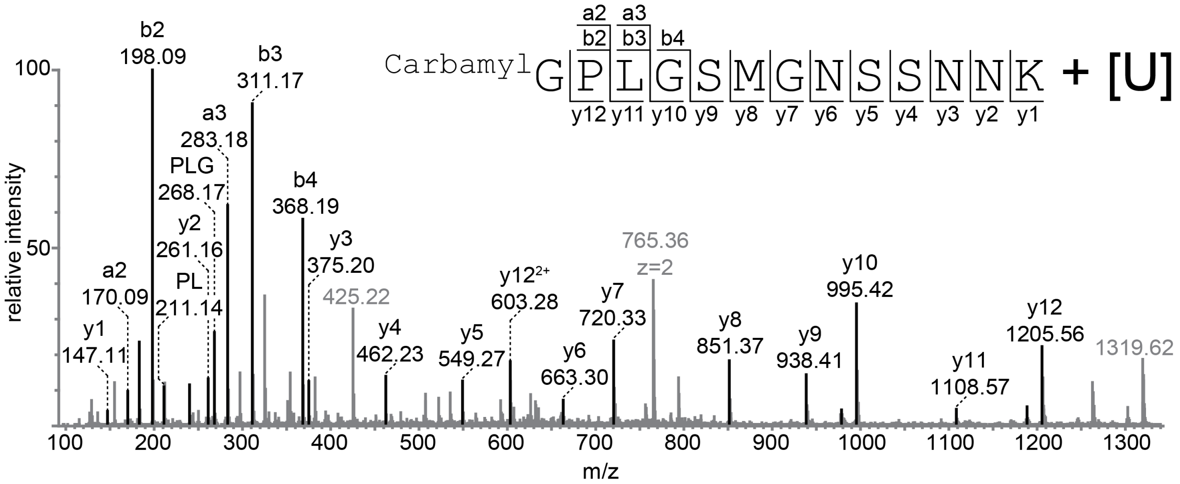


Figure 3.13: MS/MS fragment spectrum (smoothed and centered) of She2p peptide IGSNLLDLEVVQFAIK (I164–K179) cross-linked to [U]. The y-series is partially shifted by [U $-H_3PO_4$] or [U] starting with y4. This identifies F181 as the cross-linked amino acid. An intense signal is observed at *m/z* 227.06, which corresponds to the RNA fragment [U $-H_3PO_4$] .
(Details of spectra annotation are described in B.1.)

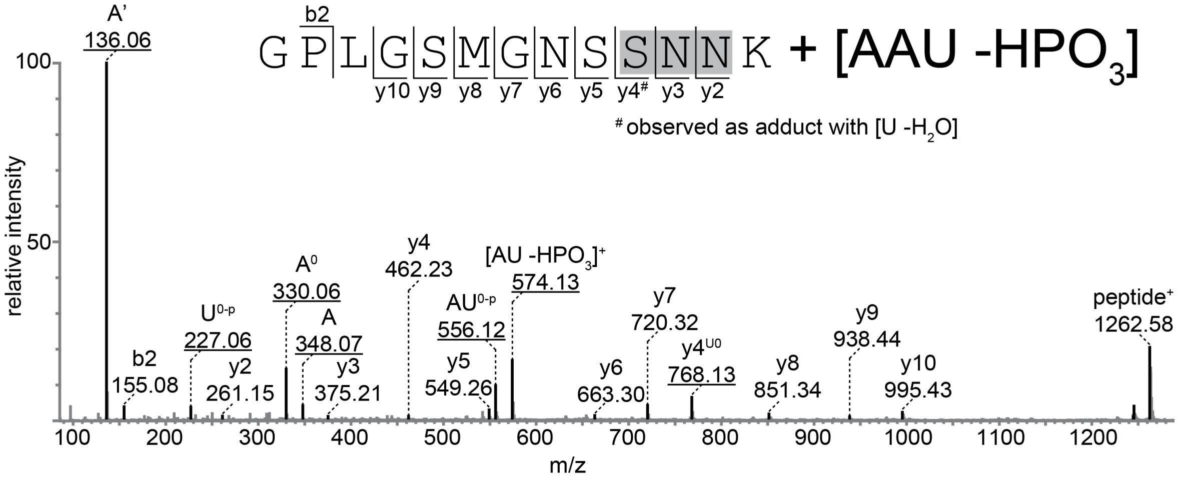
than that of y-ions (e.g., compare a5/b5 pair at *m/z* 457.26 and 485.27 to y4 at *m/z* 478.30). a- and b-ions are rarely observed in the higher *m/z* range after beam-type CID (see 1.3.3.4). Therefore, large signals with fractional masses significantly below those of calculated/observed y-ions of the cross-linked peptide are very likely to correspond to RNA adducts.

Two spectra of She3p-short peptide GPLGSMGNSSNNK (GPLGSM from cleavage of GST and She3p positions G334–K340) are shown in Figure 3.14. In 3.14a, the peptide was found carbamylated on its N-terminus. The spectrum does not contain any trace of the cross-linked RNA; its presence and composition was deducted solely from the difference between experimental precursor and calculated peptide mass. In contrast, the MS/MS spectrum of the same, unmodified peptide cross-linked to [AAU $-HPO_3$] (3.14b) is dominated by RNA signals. The marker ion for adenine at *m/z* 136.06 is the most intense signal. RNA signals for adenosine, adenosine after loss of water, uridine after loss of phosphoric acid, as well as for the [AU] dinucleotide with loss of HPO_3 and H_3PO_4 were observed. They all exhibit a higher intensity than any of the peptide fragments; only the intact peptide produces an intense signal. The signal at *m/z* 768.13 corresponds to an adduct of y4 with [U $-H_2O$], this places the cross-link on either S337, N338, or N339. Due to the low intensity of peptide fragment signals, the cross-linking site cannot be narrowed down further, as it cannot be excluded that other peptide–RNA adducts were present but not detected.

In summary, it can be concluded that both She2p and She3p-short directly interact with *ASH1* mRNA zip-code element E3 in the ASH1-short complex. The results from analysis on the Q-ToF do not allow for final conclusions about whether the same is true for the ASH1-FL complex.



(a) Cross-link to [U].



(b) Cross-link to [AAU $-HPO_3$].

Figure 3.14: MS/MS fragment spectra (smoothed and centered) of She3p-short peptide GPLGSM-GNSSNNK (G334–K340) cross-linked to (a) [U] and (b) [AAU $-HPO_3$]. While the spectrum of the cross-link to [U] does not contain any RNA signals, the cross-link to [AAU $-HPO_3$] produces dominating RNA (marker) ions.

3.2.5 ASH1 cross-links identified after LC-ESI-MS/MS measurement on the Orbitrap Velos and data analysis with OpenMS

Cross-linking experiments of the model complexes for *ASH1* mRNA transport were the first to be analyzed on the Orbitrap Velos and consequently the first Orbitrap data that were evaluated after precursor mass variant generation with OpenMS and OMSSA database searches.

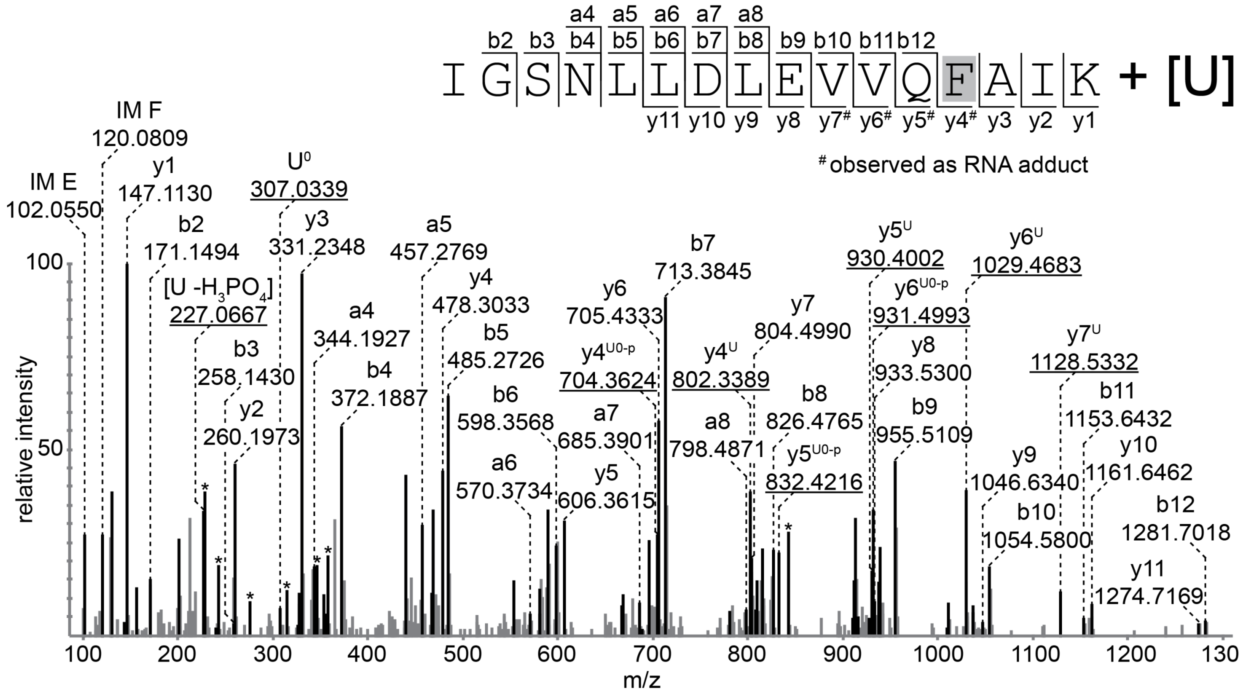


Figure 3.15: MS/MS fragment spectrum of She2p peptide IGSNLLDLEVVQFAIK (I164–K179) cross-linked to [U]. The y-series is partially shifted by [U] and/or [U –H_3PO_4] starting with y4. Thus, F181 is identified as the cross-linked amino acid.

The same cross-link of She2p peptide IGSNLLDLEVVQFAIK to [U] that had been identified in the previous experiments analyzed with the Q-ToF Ultima was found cross-linked. The corresponding fragment spectra are highly similar (compare Figures 3.13 and 3.15). In both cases, the y-series was observed as partially shifted by [U] and/or [U –H_3PO_4]. Because the shift was observed for y4 but not for y3, it was concluded that F181 was the cross-linked amino acid. The exact number of observed a-, b-, and y-ions, as well as relative signal intensities, do show differences between both spectra. However, it has to be noted that the precursor charge state observed on the Q-ToF was 2 while it was 3 in the measurement on the Orbitrap Velos. MS/MS fragment spectra of the same species but obtained after fragmentation of different charge states often show similar differences in number and relative intensities of peptide fragments and RNA–peptide adducts, even if they are recorded within the same measurement on the same instrument (data not shown). Therefore, the similarities between the MS/MS fragment spectra after either beam-type CID in the Q-ToF or HCD in the Orbitrap Velos, as exemplified here with the cross-link of IGSNLLDLEVVQFAIK to [U], led to the conclusion that fragmentation of cross-links under both conditions are comparable.

The cross-link of the N-terminus of She3p-short identified in the Q-ToF measurement was also confirmed in the analysis on the Orbitrap Velos (see Figure B.3 and Table 3.3), although with different RNA sequences. This could be due to less reproducible RNA hydrolysis or cleavage during enrichment, chromatography, or ionization. On the peptide level, previous cross-linking results

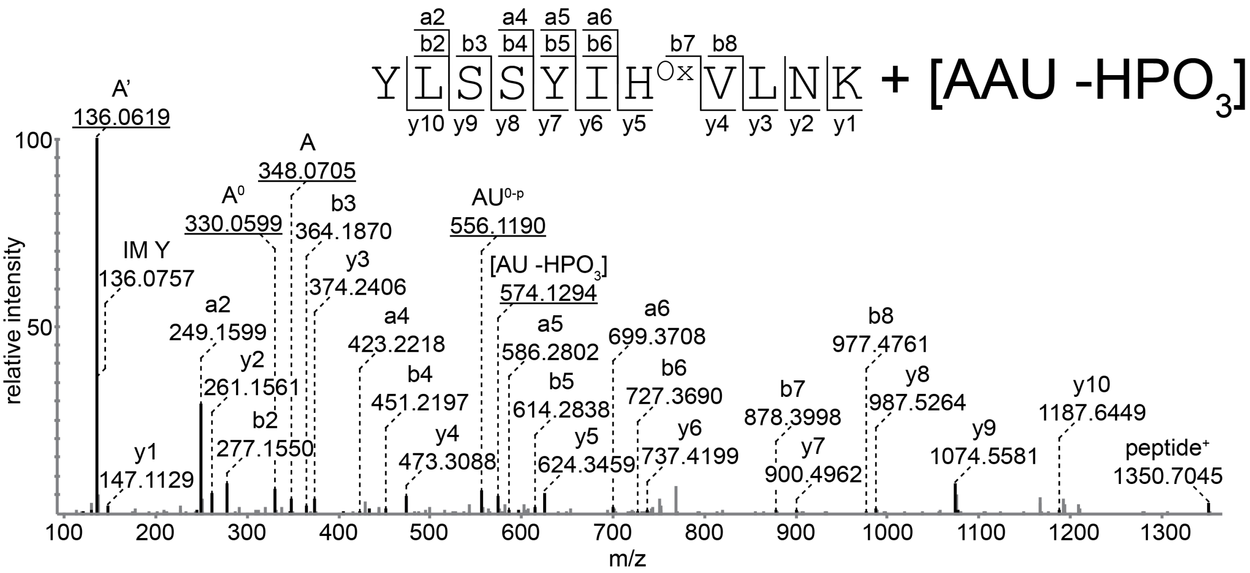


Figure 3.16: MS/MS fragment spectrum of She2p peptide YLSSYIHVLNK (Y27–K37), oxidized at H33, cross-linked to [AAU –HPO$_3$].

could be completely reproduced with the Orbitrap Velos, even though the amounts of starting material were considerably lower (only 17% for ASH1-FL and 25% for ASH1-short). In addition, a surprisingly high number of additional cross-links were found after analysis on the Orbitrap Velos that were not identified before. Most cross-links were confidently identified and manually validated. Example spectra for each cross-linked peptide are shown in the Appendix, observed cross-links are summarized in Table 3.3, and corresponding mass values are given in Table A.3. However, three interesting cross-links will be described in more detail below.

Figure 3.16 shows the MS/MS fragment spectrum of She2p peptide YLSSYIHVLNK (Y27–K37). It was identified as cross-linked to [GGU –H$_3$PO$_4$] after a database search of the precursor variants. However, the spectrum clearly shows an intense adenine marker ion and several signals corresponding to an [AU] dinucleotide; in contrast, no G marker ions are observable. The experimental precursor m/z was 751.6308 (z=3); subtraction of the calculated peptide mass of 1335.7186 Da leads to a mass difference of 916.1504 Da. Thus, the cross-linked RNA should contain three nucleotides. From the marker ions it was concluded that the RNA should contain only A and U. Consequently, the RNA should be either [AAU] or [AUU]. [AAU –HPO$_3$] has a calculated mass of 902.1746 Da and thus lies in the same range as the observed mass difference. The synthetic RNA oligonucleotide used in the experiments has AAU at its 3' end and does not contain a 3' phosphate, therefore [AAU –HPO$_3$] could originate from there. Subtracting the calculated mass of [AAU –HPO$_3$] from the mass difference between experimental precursor and calculated peptide mass leaves 13.9758 Da. Interestingly, close inspection of the fragment spectrum revealed all peptide fragments containing H33 were shifted by the same mass, i.e., y5–y10 and b7–b8, as well as the signal at m/z 1350.7045 corresponding to the intact peptide plus 13.9781 Da. Initially, it was assumed that the shift could result from methylation with a monoisotopic mass of 14.0156 Da. However, the mass deviation between the calculated mass of the methylated peptide cross-linked to [AAU –HPO$_3$] and the experimental mass would be 17.7 ppm. The mass accuracy of the Orbitrap Velos is typically well below 5 ppm; therefore, methylation had to be excluded. Finally, it was concluded that the histidine was likely oxidized to 2-oxo histidine. The corresponding mass shift would be 13.9793 Da, which is in excellent agreement with the observed differences in both peptide fragments and precursor mass. The mass difference between the calculated mass of the peptide containing 2-oxo histidine cross-linked to [AAU $-\mathrm{HPO_3}$] and the experimental mass is 1.6 ppm, which lies within the expected mass accuracy of the Orbitrap Velos. This example illustrates the importance of cross-link validation and how an iterative approach in testing and confirming or excluding different assumptions leads to the final conclusion about cross-linked peptide, RNA, and their modifications.

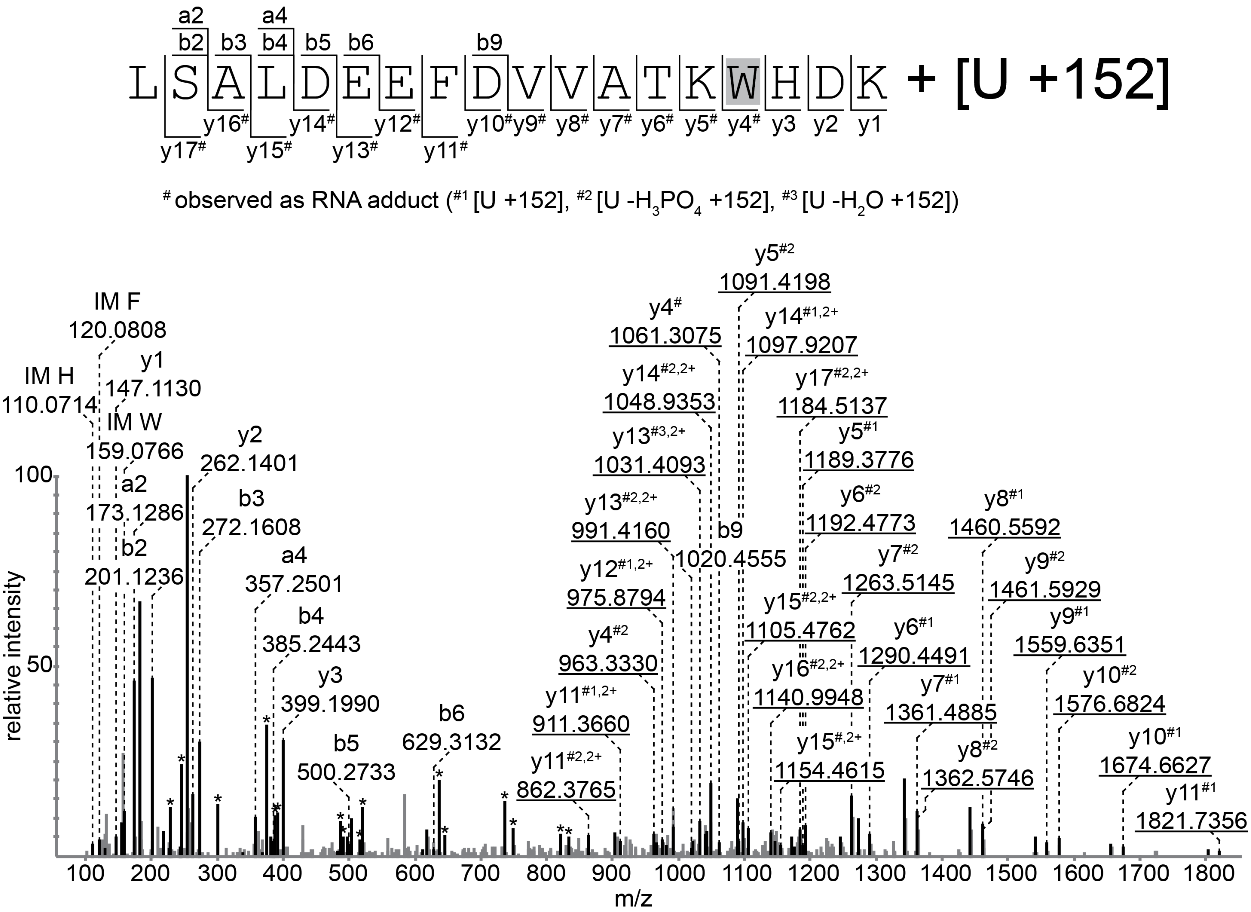
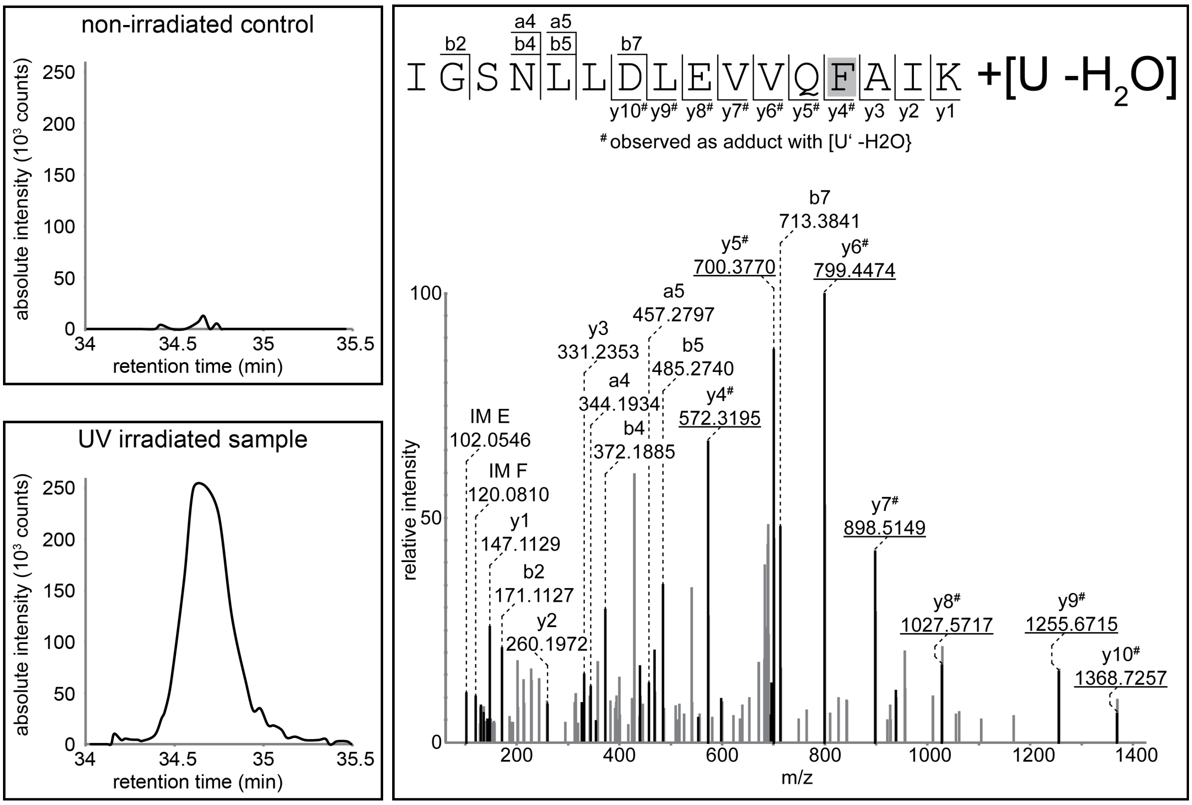


Figure 3.17: MS/MS fragment spectrum of She2p peptide LSALDEEFDVVATKWHDK (L223–K240) cross-linked to [U +152]. Signals corresponding to internal ions are marked with an asterisk and not annotated further. The y-series is completely shifted by different RNA adducts starting with y4, which leads to identification of W237 as the cross-linked amino acid.

Another intriguing observation was that She2p peptide LSALDEEFDVVATKWHDK (L223–K240) cross-linked to [U +152]. In contrast to all other peptides observed with the 152 adduct, this peptide does not contain a cysteine residue. Close manual inspection of the spectrum revealed a high number of y-ions shifted by different RNA adducts, all containing 152 Da as an additional mass. Since regular y-ions were only observed for y1 to y3, it was concluded that W237 was the cross-linked amino acid. This cross-link presents the first (and so far only) observation of a cross-link with the additional mass of 152 Da that cannot be connected to a cysteine residue.

3.2.5.1 Introduction of non-irradiated controls for validation

Negative controls are common for many analytical methods to exclude false positive results due to unexpected effects. For example, negative, i.e. non-irradiated, controls were included in the gel-



(a) XIC of *m/z* 689.0138. (b) Annotated fragment spectrum.

Figure 3.18: Extracted ion chromatograms (XICs) and MS/MS spectrum of She2p peptide IGSNLLDLEVVQFAIK (I164–K179) cross-linked to [U $-H_2O$]. The XIC of the UV irradiated sample shows a clear signal, while the control XIC shows only background noise. The peptide sequence is confirmed by almost complete coverage with N-terminal a- and b-ions and C-terminal y-type fragments. The y-series is shifted by 94 Da starting with y4, corresponding to the uracil base minus water. Through the shift, F181 is identified as the cross-linked amino acid.

based comparison of cross-linking yields of native versus 4SU-substituted RNA to the NusB–S10 complex (see 3.1.1). However, such controls were not previously included in mass spectrometric analysis of cross-linking experiments. The low sensitivity of the Q-ToF Ultima together with the low cross-linking yield and the usually very limited sample amounts did not permit any further division of the starting material.

The significantly higher sensitivity of the LTQ Orbitrap Velos led us to prepare non-irradiated controls along with the irradiated sample in every experiment. This strategy was first tested for feasibility with the ASH1 complex.

Measurement of a non-irradiated control followed by the cross-linked sample allows for a straightforward and unambiguous validation of cross-link candidates: Extracted ion chromatograms (XICs) are calculated for the precursor mass of the cross-link candidate in both the control and UV irradiated sample. In other words, intensities of all ions in a certain mass range are summed and plotted against retention time. The mass range is adjusted to the mass error of the instrument around the calculated precursor mass of the cross-link.

An example for XIC comparison is shown in Figure 3.18a. XICs for the precursor mass of a cross-link candidate are calculated in both the control and UV irradiated sample. While the chromatogram of the irradiated sample shows a clear signal, the intensity in the control is an order of magnitude smaller. Therefore, the precursor must be of a species formed as a consequence of UV irradiation. The peptide sequence obtained from the database search was confirmed by comparison of predicted and experimental fragments (Figure 3.18b). Interestingly, the 94 Da shift observed for 4SU-substituted RNA was also found here (see below).

3.2.5.2 4-thio-uracil and native uracil form similar cross-linking products

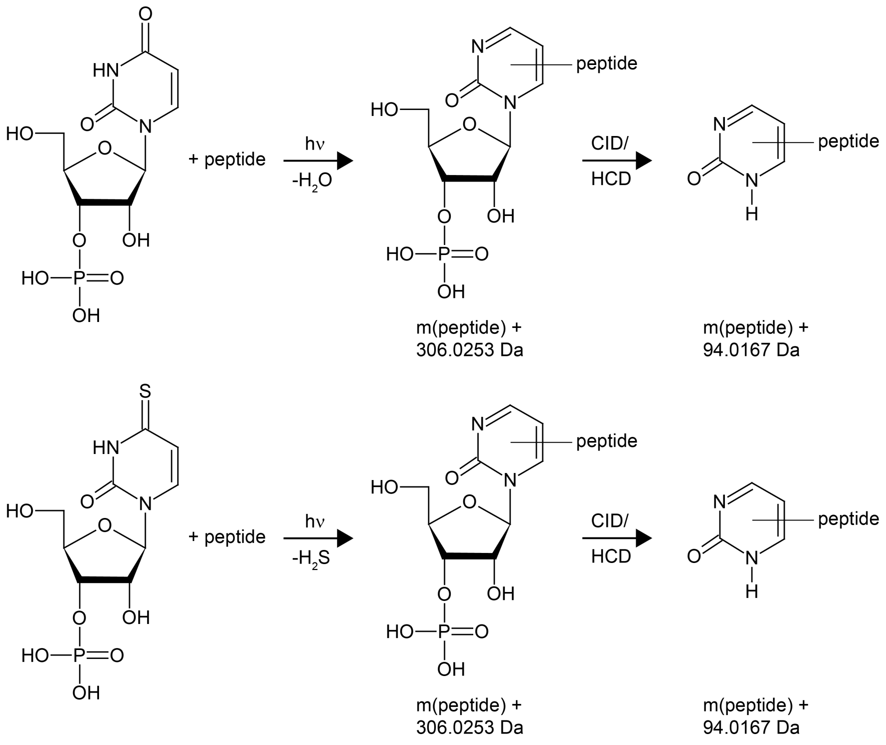


Figure 3.19: Possible structures of 4SU and uracil cross-linking products. Upon UV irradiation, both uracil and 4SU cross-link with a net loss of H_2O or H_2S, respectively. The resulting cross-linking product has a mass of 306.0253 Da. Upon fragmentation with CID or HCD, an adduct of 94.0167 Da remains on the peptide, corresponding to cleavage of the N-glycosidic bond.

In previous experiments on cross-linking 4SU-substituted RNA to NusB–S10, several cross-links were observed that showed an overall mass of the cross-linked RNA of 306 Da and a partial shift of the peptide fragments by 94 Da (see 3.1.3.2). Our conclusion was that H_2S was lost upon cross-linking and fragmentation of the N-glycosidic bond but not the cross-linking bond led to the observed adduct in the peptide fragments.

In our cross-linking experiments of the ASH1 complex, we observed the same mass adducts, i.e., 306 Da for the precursor and 94 Da for the peptide fragments. An example is shown in Figure 3.18b. Therefore, we believe that both 4SU and native uracil undergo similar or even the same cross-linking reactions with a net loss of H_2S or H_2O, respectively (see Figure 3.19). Since it is clear that the sulfur atom is lost in 4SU, we propose that the carbonyl oxygen at C4 is lost in native uracil. However, this could only be proven in additional experiments, e.g. with ^{18}O labeled uracil. Mass spectrometry allows only for conclusions about the elemental composition of the products, the exact structure and connectivity, especially to the cross-linked peptide, cannot be determined.

Therefore, the actual cross-linking product could be an isomer of the structure shown in Figure 3.19.

3.2.5.3 Significant improvement of cross-link identification after measurement on the Orbitrap Velos and data analysis with OpenMS

Table 3.3 summarizes all cross-links identified after measurement on the Orbitrap Velos and data analysis with OpenMS. Overall, six regions of She2p, four regions of She3p and the N-terminus of She3p-short were identified as cross-linked, combining the results of both ASH1-FL and ASH1-short. Compared to the single peptide regions of both She2p and She3p-short identified after analysis with the Q-ToF Ultima, there was a remarkable improvement. Since starting amounts were decreased significantly in experiments analyzed with the Orbitrap Velos, the results illustrate the increased sensitivity and sequencing speed of the Orbitrap Velos even more drastically. Assuming that the same cross-links were formed in both experiments, identification with the Q-ToF Ultima could have been hampered for several reasons: (1) Due to the lower sensitivity, cross-link precursors might have been below the detection limit of the instruments. (2) If detected, intensities might not have been sufficient to trigger MS/MS fragmentation or to produce a fragment spectrum of reasonable quality. (3) The significantly lower sequencing speed might have prevented low intensity precursors from being chosen for fragmentation at all. For example, 382 MS/MS spectra were recorded in the ASH1-short measurement on the Q-ToF Ultima, while the analysis of cross-links from the same complex on the Orbitrap Velos produced 2677 MS/MS spectra (identical LC gradients). Therefore, the Orbitrap Velos is considerably more likely to fragment low intensity precursors.

This significant increase in spectra numbers also illustrates that both improvements on the instrumental and data analysis side were largely dependent on each other: While analysis of Orbitrap Velos data would in principle be possible with precursor mass generation by the perl script and Mascot searches, the high number of spectra would make this approach even more time-consuming, rendering it unfeasible. Equally, data analysis of Q-ToF data with OpenMS was theoretically possible, but hampered by several characteristics of the antiquated format of the Q-ToF data and the low sequencing speed. These disadvantages were not encountered with the Orbitrap Velos data.

The most obvious advantages of the integration of the precursor variant approach into the OpenMS environment are the automated submission of database searches and the retrieval and summary of search results. The duration of these steps in the data analysis workflow was substantially shortened. However, OpenMS included another advantage crucial for straightforward evaluation and validation of cross-link candidates: TOPPView, the graphical user interface of OpenMS for viewing MS data, can be used to annotate the database search results in *.idXML* format onto the raw data in *.mzML* format. This procedure has many advantages for validation. Previously, Mascot search results had to be compared manually to the raw MS data, i.e., two separate programs had to be used in order to perform one task. TOPPView automatically highlights signals of the raw data that overlap with theoretical fragments of the peptide match resulting from the database search. Additional signals, e.g. RNA marker ions, can be easily annotated with the corresponding m/z and a label. In addition, TOPPView allows for calculation of distances between two signals. This simple functionality is extremely useful in manual spectra interpretation and not offered by the programs provided by the mass spectrometer vendors. Suggestions for minor improvements of annotation collected while evaluating cross-linking search results were integrated into TOPPView by our collaborators. Finally, in order to determine the RNA composition of the cross-link candidate from Mascot search results, the difference between experimental precursor mass and the precursor variant that gave rise to a match had to be calculated manually. In contrast, TOPPView was set up to be able to display the RNA composition directly, allowing, for example, quick comparison of the expected and experimentally observed marker ions. Therefore, TOPPView overall is extremely useful for validation, greatly improving and speeding up the process.

Table 3.3: Cross-links of the ASH1 complexes identified after analysis on the Orbitrap Velos.

Cross-links of She2p

position	peptide	aa	RNA	figure	FL	short
M1–K3	GPLGSMSK	-	[U –H_2O]	B.3	-	+
		-	[AU –H_2O]	-	-	+
	GPLGSM*SK	-	[U –H_2O]	-	+	-
	*oxidized	-	[AU –H_2O]	-	+	+
Y27–K37	YLSSYIH*VLNK	-	[AAU –HPO_3]	3.16	-	+
	*oxidized					
F64–K82	FYNDCVLSYNASE-	-	[U]	-	-	+
	FINEGK	C68	[U +152]	B.4	-	+
		C68	[AU +152]	-	+	+
		C68	[AAU +152 –HPO_3]	-	-	+
		C68	[AAU +152]	-	-	+
F64–K94	FYNDCVLSYNASEFINE-	C68	[U +152]	-	-	+
	GKNELDPEADSFDK	C68	[AU +152]	-	-	+
C106–K123	CVETFDLLNYYLTQSLQK	C106	[U +152]	B.5	-	+
		C106	[AU +152]	-	+	+
I164–K179	IGSNLLDLEVVQFAIK	F176	[U –H_3PO_4]	-	+	-
		F176	[U –H_2O]	3.18b	+	+
		F176	[U]	3.15	+	+
		-	[GU]	-	+	+
L223–K240	LSALDEEFDVVATKWHDK	W237	[U +152]	3.17	-	+
		-	[AU +152]	-	-	+

Cross-links of She3p

position	peptide	aa	RNA	figure	FL	short
M130–K138	M*DQLSKLAK	K135	[U –H_2O]	B.6	+	-
	*oxidized					
N139–K150	NSSAIEQSCSEK	C147	[U +152]	B.9	+	-
		C147	[AU +152]	-	+	-
		C147	[AAU +152 –HPO_3]	-	+	-
G283–K291	GAVVQTLKK	K290	[U –H_2O]	B.7	+	-
T383–R405	TNVTHNNDPSTSPTISV-	-	[GU]	B.8	-	+
	PPGVTR	-	[AAU –HPO_3]	-	-	+
G334–K340	GPLGSMGNSSNNK	-	[U]	B.3	-	+
		-	[GU]	-	-	+

For results obtained with the Q-ToF Ultima and column labels see 3.2.

3.2.6 Summary and functional implications of obtained cross-linking results

This section will discuss the results obtained in cross-linking of model complexes for *ASH1* mRNA transport in a biological context. A broader view on technique and method development will be taken in Chapter 4.

Neither She2p nor She3p show homology to known RNA binding domains. Therefore, identification of peptides or even amino acids contributes considerably to the knowledge about RNA-interacting regions for both proteins.

When comparing cross-linking results of ASH1-FL and ASH1-short, there are several differences. It must be noted that only positive identifications are indicated in Table 3.3. The absence of an identification does not necessarily imply that a particular cross-link was not formed during UV irradiation. Due to the low abundance of cross-links in general, the corresponding precursors are not always chosen for fragmentation during data acquisition or the MS/MS fragment spectra are not of sufficient quality for confident identification. Therefore, data analysis distinguished between identified and detected cross-links. The term detected indicates here that the measurement (but not the corresponding control) contained a species with the same mass-to-charge ratio and retention time as a cross-link identified in another MS data set of the same or similar protein–RNA complex.

In the case of She2p, all cross-linked peptides were detected in both ASH1-FL and ASH1-short. This hints that She2p–RNA interactions in both complexes are highly similar. In contrast, all She3p cross-links were exclusively found in one of the complexes. This could indicate that She3p–RNA interactions are slightly different in ASH1-FL and ASH1-short. Due to the high specificity of UV induced cross-linking, even minor changes in the three-dimensional structure could prevent or enable the formation of cross-links. Therefore, regions or residues identified as cross-linked in one complex could still participate in protein–RNA interactions in the other. It could be speculated that none of the She3p peptides found cross-linked are responsible for RNA-binding specificity.

The first publication including our findings about direct interactions as derived from UV induced protein–RNA cross-linking with mass spectrometric analysis was based on early experiments with the ASH1-short complex and analysis on the Q-ToF Ultima [96]. The much more detailed results from the improved Orbitrap Velos/OpenMS setup were not available at that point, but are the basis for ongoing studies by our collaborators. Therefore, the comparison to functional data is focused on this and other published studies.

The cross-linked She2p peptide IGSNLLDLEVVQFAIK is part of a helix that is essential for interactions of She2p with She3p and *ASH1*-E3, as deletion mutants failed to bind She3p or to form the ternary complex with RNA. In contrast, deletion of the She2p C-terminus weakens but not prevents synergistic RNA binding, confirming the observation of the very C-terminus cross-linking to RNA. Both mutants do not show localization *in vivo* [96]. Mutation of She2p residue C68, identified as cross-linked in our study, to tyrosine prevents dimerization and efficient RNA binding [109]. This confirms the direct interactions of C68 with RNA implied by the cross-linking results. Comparison of the cross-linking results to the structure of the She2p dimer [109] indicates that the cross-linked residues C68 and W237 as well as F176 of the respective other monomeric subunit lie adjacent to the proposed RNA binding surface (data not shown).

The 20 N-terminal amino acids of the She3p-short construct (positions 334–353 of the construct comprising positions 334–425 of wild type She3p) were found to be essential for formation of the ternary She3p-short, She2p and *ASH1*-E3 complex [96]. While other residues in this stretch, especially S348, were found to be important for *ASH1* mRNA localization *in vivo* [110], the identified direct interactions of residues S337, N338, or N339 with *ASH1* mRNA might add to synergistic binding.

Overall, the significant increase in the number of identified cross-links after measurement on the Orbitrap Velos and data analysis with OpenMS illustrates that both improvements are crucial in the advancement of the investigation of UV induced cross-linking by mass spectrometry.

3.3 Application of the automated data analysis workflow to the spliceosomal protein Cwc2 in complex with RNA

Cwc2 is a yeast splicing factor and related to the nineteen complex (NTC) that joins the spliceosome during its activation. In isolated, catalytically active spliceosomes, it cross-links to U6 snRNA [111]. Our collaborators (Macromolecular Crystallography Group, MPI for Biophysical Chemistry) had solved the crystal structure of the functional core of Cwc2. Comparison to conserved sequence motifs and physical properties suggested cooperative RNA-binding by several domains (see below and [97]).

We performed UV cross-linking and mass spectrometry to assess which domains of Cwc2 are involved in direct interactions with RNA. Analysis on the Orbitrap Velos together with cross-link identification after precursor variant generation with OpenMS allowed for comparatively fast evaluation of several different experiments performed (see below). Therefore, cross-linking of Cwc2 presents an excellent example for application of the improved mass spectrometry/data analysis workflow.

3.3.1 Cross-linking of Cwc2 to U6 snRNA

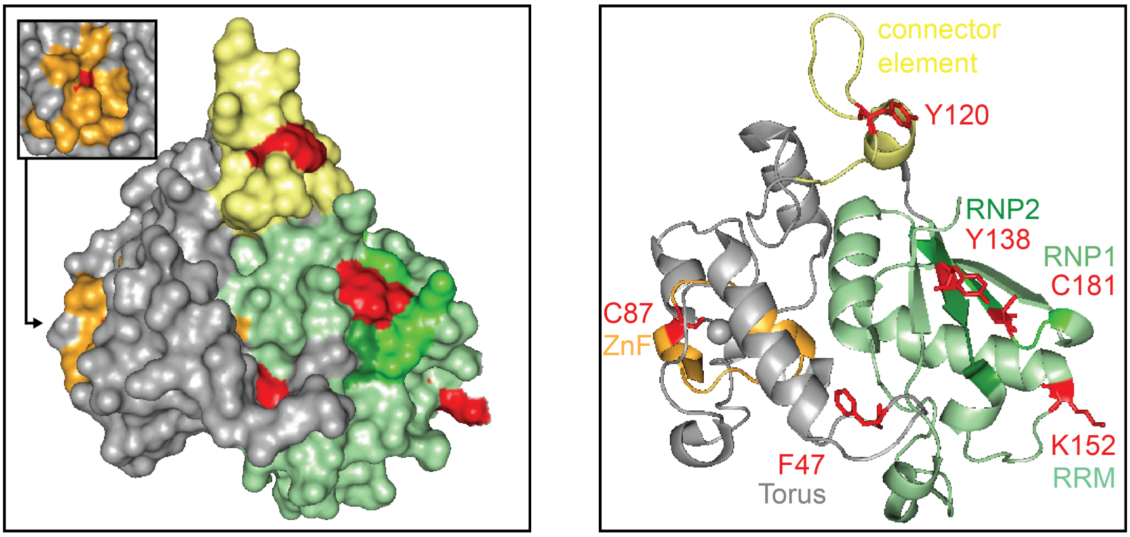
UV induced cross-linking of the *in vitro* reconstituted, binary Cwc2–U6 snRNA complex and analysis of cross-linking products by mass spectrometry was performed to identify regions of Cwc2 that directly interact with RNA. To this end, reconstituted complexes were UV irradiated at 254 nm for 10 minutes and enriched according to the standard protocol (see 2.2.8.5). LC-ESI-MS/MS analysis was carried out with the Orbitrap Velos and data analysis was performed with our newly established workflow based on OpenMS.

Overall, six regions of the functional core of Cwc2 were found cross-linked (see Table 3.4). Close inspection of the corresponding spectra enabled identification of the cross-linked amino acid residue in all cases. Representative spectra for each of the cross-linked region are shown in the Appendix (Figures B.10 to B.15).

Cross-linked residues in both RNP1 and RNP2 are in excellent agreement with conserved aromatic residues at these positions that are involved in stacking interactions with RNA [1]. Cross-linked residue Y138 represents the conserved aromatic residue in RNP2. The first aromatic residue of RNP1 is replaced by C181 in Cwc2; however, C181 was found cross-linked, suggesting that the cysteine still participates in important protein–RNA interactions. The second aromatic residue in the RNP1 consensus sequence corresponds to F183 in Cwc2. Although F183 was not identified as a cross-linked residue, it is contained in the cross-linked peptide NCGFVK (N180–K185). It cannot be completely excluded that some of the identified cross-links might stem from F183 instead of C181 as the latter was not unambiguously identified as the cross-linked amino acid in all cross-links of NCGFVK listed in Table 3.4.

The third cross-linked residue within the RNA recognition motif (RRM) of Cwc2 is K152. Residue C87, also identified as cross-linked, is one of the three cysteines coordinating to zinc in the CCCH zinc finger domain, confirming direct RNA interaction in Cwc2 by this canonical RNA-binding domain.

The connector element (residues 116–133) links the C-terminal part of the Torus domain to the N-terminus of the RRM. It contains several conserved residues, three aromatic and three positively



(a) Surface representation.

(b) Cartoon representation.

Figure 3.20: Structure of Cwc2. (a) shows the protein surface while (b) shows secondary structure elements. Domains are colored as indicated in (b): RRM light green, RNP2 dark green, RNP1 green, zinc finger (ZnF) orange, connector element yellow, and Torus light gray. Cross-linked residues are highlighted in red and shown as sticks in (b).
(a) All cross-linked residues lie on the surface of Cwc2. The insert shows a representation of the zinc finger, rotated around 90°, with C87 slightly buried but nonetheless solvent-exposed.
(b) Representation of cross-linked residues within secondary structure elements. Side chains of Y120, Y138, K152 and C181 are exposed and might easily bind RNA. In contrast, the side chain of F47 is buried. C87 coordinates to zinc (gray sphere).

charged, suggesting a potential RNA-binding property. One of the conserved aromatic residues in Cwc2, F120, was found cross-linked; therefore, the connector element presents an additional RNA-binding element of Cwc2. The sixth cross-linked residue, F47, lies within the Torus domain of Cwc2.

The cross-linked residues all lie on the same side of Cwc2 (see Figure 3.20) and are distributed over a large, positively charged [97] surface. Residue Y120 in the connector element as well as residues K152, Y138 and C181 (latter two RNP2 and RNP1, respectively) of the RRM are solvent-exposed and can therefore easily form interactions to RNA. The side chain of F47, which is part of the Torus domain, faces away from the protein surface. Since it lies within a loop region, it might flip out upon RNA binding. Zinc-coordinating residue C87 might bind a nucleotide together with Y89 (not shown, see [97]).

Our collaborators performed mutation studies on residues identified as directly interacting with RNA in our cross-linking experiments and tested RNA-binding by electrophoretic mobility shift assays (EMSAs). Single point mutations did not significantly change RNA-binding affinity. Double mutants Y138A/Y120A and Y138A/C181A exhibited a significant decrease in RNA affinity, while RNA-binding was not considerably affected in Y138A/F47A, Y138A/K152A, and Y120A/F47A. These findings provide evidence for the importance of RNP2 and RNP1 as well as the connector element for RNA-binding of Cwc2 to U6 snRNA *in vitro*, while K152 (RRM) and F47 (Torus) did not appear to be crucial for RNA-binding [97]. Mutation of C87 is not tolerated due to its important structural role [112].

Table 3.4: Cross-links of Cwc2 to U6 snRNA

domain	position	peptide	aa	RNA	spectrum
Torus	W37–K61	WSQGFAGNTRFVSPFALQPQLHSGK	F47	[UU]	-
				[AUU]	-
	F47–K61	FVSPFALQPQLHSGK	F47	[U –H_2O]	B.10
				[U]	-
				[UU]	-
zinc finger	G79–K101	GM(Ox)CCLGPKCEYLHHIPDEEDIGK	C87	[U +152 –H_2O]	-
	C87–K101	CEYLHHIPDEEDIGK	C87	[U]	-
				[U +152]	-
				[AU +152]	B.11
				[AAU +152]	-
				[GU +152]	-
ce	F117–R131	FADYREDMGGIGSFR	Y120	[U]	B.12
				[AU]	-
RNP2	T136–K149	TLYVGGIDGALNSK	Y138	[U]	B.13
				[AU]	-
				[AAU]	-
				[AUU]	-
				[GU]	-
				[GGU]	-
				[AGU]	-
				[UU]	-
				[CU]	-
				[ACU]	-
				[CGU]	-
RRM	H150–R159	HLKPAQIESR	K152	[U –H_2O]	B.14
				[AU –H_2O]	-
RNP1	N180–K185	NCGFVK	C181	[U]	-
				[U +152]	B.15
				[AU]	-
				[AU +152]	-
				[AAU +152]	-
				[AUU +152]	-
				[GU]	-
				[GU +152]	-
				[AGU +152]	-
				[UU]	-
				[UU +152]	-

domain: domain/motif according to [97] (ce: connector element)
position: amino acid position of the cross-linked peptide in the protein sequence
peptide: amino acid sequence of the cross-linked peptide
aa: one-letter code and position of the cross-linked amino acid
RNA: composition of the cross-linked RNA (oligo)-nucleotide with modifications
spectrum: reference to figure showing representative spectrum

3.3.2 Cross-linking of Cwc2 to U4 snRNA and U6 internal stem loop

In vitro, Cwc2 binds several spliceosomal snRNAs [112]. Therefore, additional experiments were performed, comparing cross-linking of Cwc2 to U6 snRNA with cross-linking to U4 snRNA. Experiments were performed as described above for U6 snRNA. Irradiation time was usually 10 minutes, and each binary complex was additionally analyzed after UV irradiation for two minutes at 254 nm.

Cross-links identified in the binary Cwc2–U6 snRNA complex (see 3.4) were evaluated in all subsequent experiments on two levels: First, it was examined whether the same cross-link was identified in the new experiment after data analysis with OpenMS. If a previously observed cross-link was not identified, it was next tested whether a species with the same precursor mass and retention time was detected. To this end, extracted ion chromatograms were calculated in both the UV irradiated sample and non-irradiated control. The following assumption was made: Detection of a species with the same precursor mass, charge state, and retention time indicated presence of the cross-link. Observation in the UV irradiated sample but not in the corresponding control supported the assumption. Generation of a different species by UV irradiation that had the same mass and retention time was considered extremely unlikely, especially in the simple binary system. The intensity of the cross-link was believed, however, insufficient to trigger MS/MS fragmentation in the new experiment. Without fragment information, the cross-link could not be validated completely. Previously identified cross-links were therefore denoted as either identified, detected, or not detected in each new experiment. This approach allows for a rough comparison of cross-linking behavior of one protein to different RNAs.

After cross-linking of Cwc2 to either U6 or U4 snRNA, the number of identified cross-links was comparable for Y138 (RNP2) and C181 (RNP1) (see Table 3.5 and Supplementary Table 2 in [97]). In contrast, cross-links of K152 (RRM) and Y120 (connector element) were observed less frequently with U4 snRNA. Differences in cross-link detection of C87 (zinc finger) and F47 (Torus domain) were more pronounced. This observation might indicate a preference of the zinc finger and the Torus domain to bind U6 snRNA, while the RRM could bind RNA with less selectivity.

Comparing the cross-linking results for each complex irradiated for 2 or 10 minutes did not reveal any major differences. As expected, the overall number of identified and detected cross-links was lower after the shortened time due to decreased cross-linking yield. It was anticipated that cross-links observed with low abundance after 10 minutes irradiation could fall below the detection limit of the instrument after decreasing the irradiation period. However, for U6 at least one cross-link for each region was identified, except for the cross-link in the RRM (K152), where it was only detected but not fragmented. After cross-linking of Cwc2–U4 snRNA for 2 minutes, only cross-links in RNP2 (Y138) and RNP1 (C181) were identified, but cross-links of the four other regions were detected. Overall, the number of identified or detected cross-links decreased to a comparable extent in both Cwc2–U6 and Cwc2–U4 when irradiation time was shortened.

Table 3.5: Cross-links of Cwc2 to U6 snRNA, U4 snRNA and U6 internal stem loop

			cross-linked RNA	U6		U4		ISL
			irradiation time (min)	10	2	10	2	10
domain	peptide	aa	RNA					
Torus	WSQGFAGNTR-FVSPFALQPQLHSGK	F47	[UU]	+	-	-	-	-
			[AUU]	+	-	-	-	-
	FVSPFALQPQLHSGK	F47	[U –H_2O]	+	+	(+)	(+)	-
			[U]	+	(+)	(+)	-	-
			[UU]	+	(+)	(+)	-	-
zinc finger	GM(Ox)CCLGPK-CEYLHHIPDEEDIGK	C87	[U +152 –H_2O]	+	-	-	-	(+)
	CEYLHHIPDEEDIGK	C87	[U]	+	(+)	(+)	(+)	-
			[U +152]	+	+	+	(+)	-
			[AU +152]	+	+	(+)	-	(+)
			[AAU +152]	+	-	-	-	-
			[GU +152]	+	(+)	(+)	(+)	-
ce	FADYREDMGGIGSFR	Y120	[U]	+	+	+	(+)	(+)
			[AU]	+	(+)	(+)	(+)	(+)
RNP2	TLYVGGIDGALNSK	Y138	[U]	+	+	+	+	+
			[AU]	+	+	+	+	+
			[AAU]	+	+	+	+	(+)
			[AUU]	+	(+)	-	(+)	-
			[GU]	+	+	+	+	+
			[GGU]	+	(+)	-	-	-
			[AGU]	+	(+)	(+)	(+)	+
			[UU]	+	+	+	+	+
			[CU]	+	+	-	+	+
			[ACU]	+	(+)	(+)	(+)	-
			[CGU]	+	(+)	(+)	(+)	-
RRM	HLKPAQIESR	K152	[U –H_2O]	+	(+)	(+)	(+)	+
			[AU –H_2O]	+	-	-	-	-
RNP1	NCGFVK	C181	[U]	+	+	+	+	+
			[U +152]	+	+	+	+	(+)
			[AU]	+	+	+	+	+
			[AU +152]	+	+	+	+	+
			[AAU +152]	+	+	+	(+)	-
			[AUU +152]	+	+	(+)	(+)	-
			[GU]	+	(+)	+	(+)	-
			[GU +152]	+	+	+	+	(+)
			[AGU +152]	+	+	(+)	(+)	(+)
			[UU]	+	(+)	(+)	+	-
			[UU +152]	+	+	+	+	(+)

cross-linked RNA: RNA utilized in the experiment; U6 snRNA, U4 snRNA, or U6 internal stem loop (ISL)
irradiation time (min): period for which the Cwc2–RNA complex was UV irradiated
+: identified; (+): detected; -: not detected
additional column legends, peptide positions, reference to representative spectra in 3.4

Cross-links identified after mass spectrometric analysis typically reveal only single nucleotides or short sequences of the RNA that do not allow for identification of the cross-linking site on the RNA. Additional cross-linking experiments were performed with a short RNA oligonucleotide resembling the internal stem loop (ISL) of U6. The aim was to investigate which regions of Cwc2 interact with U6 snRNA stretches within or outside of the internal stem loop. Interestingly, the results were similar to those obtained for U4 snRNA: While no substantial differences were observed for RNP2, RNP1, and the RRM, cross-links to the connector element were only detected and not identified, indicating lower abundance. This effect was even more pronounced for the zinc finger, while no cross-links at all were observed for the Torus domain.

Overall, comparative cross-linking of Cwc2 to U6 and U4 snRNA as well as U6 ISL indicated that Cwc2 might bind U6 snRNA at the internal stem loop with low sequence specificity while the other RNA binding domains (zinc finger, connector element, and Torus domain) could contribute to the specificity of binding. As the approach can only be the basis for a rough comparison and other factors might influence specificity in the complex environment of the spliceosome, further biochemical experiments should be performed to confirm these observations.

Cross-linking of Cwc2 presented an excellent example of the applicability and advantages of the novel workflow of mass spectrometric analysis on the Orbitrap Velos and data evaluation based on OpenMS. Cross-link identification was significantly more straightforward and the novel workflow provided the tool necessary to quickly assess whether previously observed cross-links were also identified in a larger number of similar experiments. Data evaluation without any external tools or based on the perl script would have required considerably more time. Since many cross-links were of low abundance, it is very likely that they would not have been detectable on the Q-ToF Ultima instrument.

3.4 Application of the automated data analysis workflow to complex systems

After our data analysis approach proved excellent feasibility for simple, *in vitro* reconstituted binary and ternary complexes, we wanted to prove its applicability to more extended systems. Since cross-links from these restricted systems could also be identified in searches against the respective proteome (*E. coli* or *S. cerevisiae*), we expected the same for more complex systems. In order to verify this assumption, an appropriate protocol for isolation and cross-linking of more extended protein–RNA complexes and suitable preparation for mass spectrometric analysis had to be developed.

3.4.1 Isolation of protein–RNA complexes by TAP tag purification

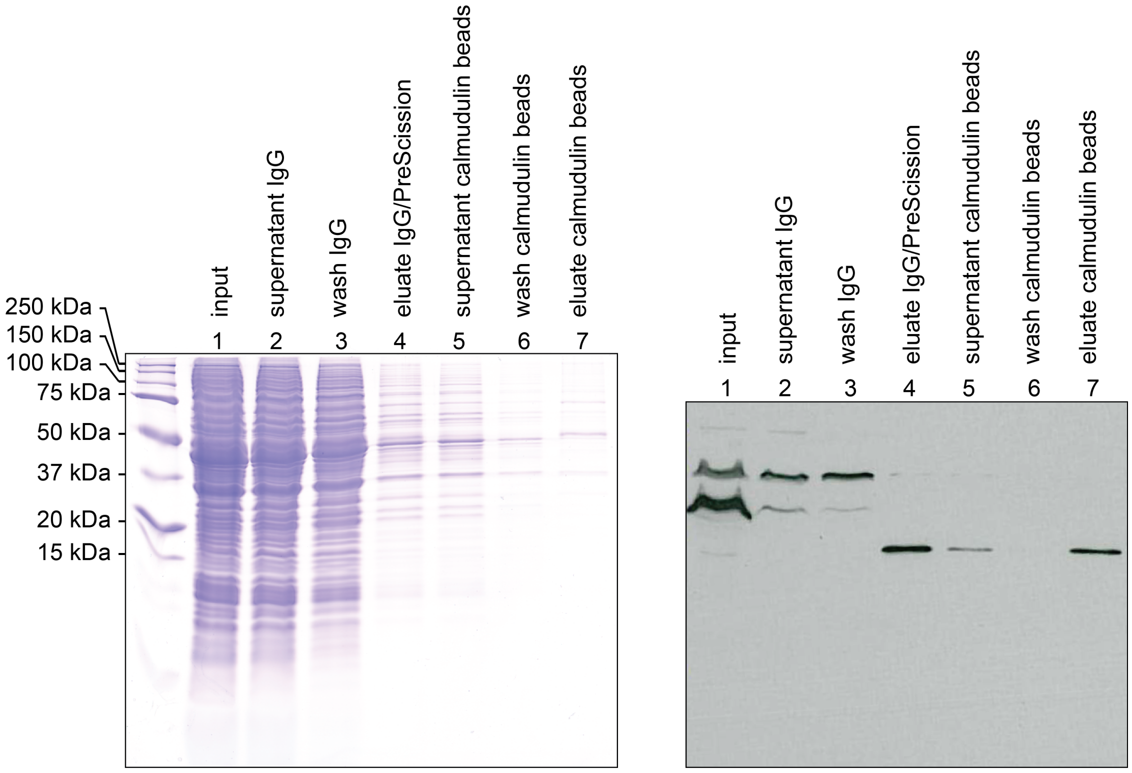
In searching for an appropriate test system for our data analysis workflow, we decided on the organisms *S. cerevisiae.* It has a proteome of medium complexity (6 607 predicted proteins in the Saccharomyces Genome Database, www.yeastgenome.org, according to [113]). In addition, the yeast proteome does not contain nearly as many phosphorylation sites as the human proteome (3 620 phosphorylation sites identified in yeast compared to 24 262 in human according to PHOSIDA [114]). Enrichment strategies based on the properties of the RNA phosphate backbone, such as titanium dioxide enrichment, always co-enrich phosphopeptides. These increase sample complexity, which might be disadvantageous for MS analysis.

Next, a protocol for isolation of protein–RNA complexes from yeast extracts had to be chosen. The strategy finally applied was based on the fact that most eukaryotic mRNAs carry a cap structure that is bound by the cap-binding complex comprising proteins Cbp20 and Cbp80.

One method frequently applied for the purification of a certain protein with its interaction partners from yeast is inclusion of a tandem affinity purification (TAP) tag on the C- or N-terminus of the investigated protein. The TAP tag contains two IgG binding domains of protein A (ProtA) from *Staphylococcus aureus* and a calmoduline binding peptide (CBP) separated by a cleavage site for TEV protease. Due to the strong interactions between protein A and IgG, under native conditions the tagged protein can only be released by TEV protease. In the second affinity purification step, the complexes are trapped on calmoduline-coated beads in the presence of calcium and released by incubation with the chelating agent EGTA [93]. We set out to isolate protein–RNA complexes by TAP tag purification from a yeast strain containing a TAP tag on the cap binding protein Cbp20.

The commercially available strain with a regular TAP tag was obtained by EUROSCARF. However, in our hands, small-scale experiments failed to elute the complexes from IgG beads by TEV protease (data not shown). Therefore, an alternative approach had to be found. Dr. Kum-Loong Boon (Department of Cellular Biochemistry) suggested the introduction of a TAP tag with a PreScission instead of a TEV cleavage site into a wild type yeast strain. He also provided protocols and helpful advice for transformation and optimized TAP tag purification. After integration of the TAP tag cassette with a PreScission cleavage site by homologous recombination, the modified TAP tag purification protocol was successfully applied.

Figure 3.21 shows a Coomassie stained protein gel of the different purification steps. The decrease of protein amount in the sample during isolation of protein–RNA complexes is illustrated in Figure



(a) Protein gel. (b) Western blot.

Figure 3.21: Protein gel and Western blot of different steps of TAP tag purification. Samples were split in half for two SDS-PAGE gels, one was stained with Coomassie (left image) and one was used for Western blotting with an anti-Calmoduline binding peptide antibody (right image). Volumes of samples for each gel relative to overall volume: Input (yeast cell extract, lane 1) and supernatant after incubation with IgG beads (lane 2) 0.04%; wash IgG beads (lane 3) 1% of first ml; eluate IgG after incubation with PreScission (lane 4) and supernatant after incubation with Calmoduline beads (lane 5) 0.67%; wash Calmoduline beads (lane 6) 2% of first ml; eluate Calmoduline (lane 7) 3%.

3.21a. The Western blot in Figure 3.21b shows that this decrease is not solely due to isolation of the mRNA-binding proteins from all proteins in the cell extract. The isolation comes at the cost of significant sample lost, illustrated by the decreased band intensities in bands 1, 4, and 7. The fraction of the overall sample volume that was used for the different lanes was increased from 0.04% (lane 1) to 0.67% (lane 4) and 3% (lane 7); thus, the effect is greater than it appears on the gel. One of the main reasons for this is that binding to both IgG and Calmoduline beads was not complete, the supernatant of both purification steps still had significant amounts of the tagged protein (see lanes 2 and 5). The same is true for the washing step after binding to IgG (lane 3). The band shift from lanes 1-3 to lanes 4, 5, and 7 is due to the cleavage of the ProtA part of the TAP tag by PreScission protease.

3.4.2 Optimization of extract preparation, complex isolation, cross-linking, sample preparation, and LC-ESI-MS/MS

In order to maximize the number of cross-links that could be identified in our experiments, several variations and protocols were evaluated and will be described below. In contrast to *in vitro* reconstituted complexes, irradiation time was shortened to two minutes in order to prevent RNA damage. Since the isolation is based on the purification of capped mRNA, extensive RNA damage would decrease the amount of isolated complexes.

For initial experiments, protein–RNA complexes were isolated by TAP tag purification; ethanol precipitated; and hydrolyzed with RNases A and T1, benzonase, and finally trypsin. Samples were desalted and enriched according to the standard protocol, except that one sample was split on two C18 columns to prevent overloading. Samples were measured under standard conditions and analyzed with the RNP$^{\text{xl}}$ tool. For each experiment, a corresponding control was prepared in parallel, treated exactly as the sample except for UV irradiation.

For the decision on the optimal conditions, equal numbers of high confidence cross-links were evaluated. Therefore, the numbers mentioned below do not represent the final number of cross-links contained in the measurement.

3.4.2.1 UV cross-linking

First, the optimal point for UV irradiation within the experimental workflow was evaluated. To this end, cross-linking was carried out (1) on cell extract, (2) after the first step of TAP tag purification, i.e. on the IgG eluate, and (3) after the second purification step, i.e. on the Calmoduline eluate. The yeast extract was prepared in AGK buffer, which contains 10% glycerine. Glycerine is a radical scavenger and diminishes or prevents cross-linking. Therefore, cross-linking on cell extract was performed after dialysis against AGK buffer without glycerin. All three samples were purified in parallel by the complete TAP tag purification protocol and irradiated at the steps of the purification protocol indicated above. Cross-links were identified in all three samples. However, the highest number of cross-linked peptides was identified for the sample irradiated after IgG elution (10 cross-links versus 3 and 2 in extract and Calmoduline eluate, respectively). Therefore, in all following experiments UV irradiation was performed at this point.

3.4.2.2 Preparation of yeast extract

Initial results showed a high number of cross-links from ribosomal proteins. Therefore, an additional ultracentrifugation step was carried out in preparation of cell extract to separate polysomes by sedimentation. However, cross-links of ribosomal proteins were still predominant and the additional centrifugation was omitted from further experiments to prevent the considerable sample loss of this step.

3.4.2.3 Sample preparation for LC-ESI-MS/MS analysis

One of the most crucial steps in the investigation of UV cross-linking experiments with mass spectrometry is the enrichment of cross-linked peptide–RNA oligonucleotide heteroconjugates prior to MS analysis. In comparison to small, *in vitro* reconstituted complexes, the excess of noncross-linked proteins and, consequently, peptides in this protein–RNA complex purification protocol is larger. The irradiation time was decreased from ten to two minutes to avoid irradiation damage. This also decreases the cross-linking yield. In addition, the isolation workflow does not discriminate between primary and secondary interactions, i.e., proteins not directly interacting with RNA but with another RNA-binding protein are also isolated. Therefore, we set out to compare strategies for the enrichment of cross-linked peptide–oligonucleotide heteroconjugates.

In initial experiments carried out to evaluate extract preparation and cross-linking conditions, our standard protocol for cross-link enrichment with C18 and TiO_2 chromatography proved successful. However, a substantial number of noncross-linked peptides was identified even after enrichment. Therefore, we wanted to evaluate another protocol frequently employed in large-scale purification of cross-linked heteroconjugates: the isolation by size exclusion (SE) chromatography. After purification of protein–RNA complexes, proteins are hydrolyzed under denaturing conditions. Intact RNA, a fraction of which has been cross-linked to peptides, is isolated from noncross-linked peptides by size exclusion chromatography. Fractions absorbing at 254 nm (RNA) and 280 nm (peptides) are collected, the RNA is hydrolyzed, and cross-linked heteroconjugates are further purified (see 1.3.2).

For comparison of the standard protocol with C18 and TiO_2 enrichment and a SE protocol, two samples were processed in parallel. After the first step of TAP tag isolation with IgG beads, the samples were UV irradiated. One sample was further processed by binding to Calmoduline beads, hyrolization, and the standard protocol with C18 and TiO_2 chromatography. For the second sample, further isolation with Calmoduline beads was omitted from the protocol to prevent the associated sample loss. Complexes were ethanol precipitated and redissolved in the presence of 1% SDS. For proteolysis with trypsin, the sample was diluted to a final concentration of 0.1% SDS. Afterwards, the sample was directly injected onto the size exclusion column. A typical elution profile can be found in Figure 3.22. Importantly, no significant differences were observed between the control and UV irradiated samples. None were expected, as only a small part of RNA should be cross-linked. High absorbance at 280 nm for early fractions of the control could indicate incomplete complex disassembly and/or hydrolysis.

Fractions of the first chromatographic peak, typically fractions 3 to 6 of several runs, were pooled and ethanol precipitated. Further sample preparation was analogous to the standard protocol, i.e., hydrolysis of RNA and proteins/peptides in the presence of urea and desalting with C18 columns (two columns per sample). As the majority of peptides was removed during size exclusion chromatography, titanium dioxide enrichment was omitted from the sample preparation workflow. Comparison between the standard workflow (C18, TiO_2) and the size exclusion plus C18 preparation revealed that a higher number of cross-links could be identified after the SE protocol (26 versus 12 cross-links).

We next investigated whether the combination of both sample preparation strategies, i.e. SE, C18, and TiO_2, would be even more beneficial. The sample was prepared as described above, with an additional titanium dioxide enrichment following the standard protocol after C18 desalting and prior to LC-MS/MS analysis with a 120 min gradient (see below). As expected, the number of noncross-linked peptides identified decreased dramatically for the combination SE/C18/TiO_2. However, this did not result in an increased number of identified cross-links, but in a greater number of sequencing events (MS/MS spectra) per cross-link. In both experiments, 93 cross-links were identified. After SE and C18, the overall number of MS/MS sequencing events combined for these 93 cross-links was 241. Two cross-links of the 40S ribosomal protein S5 peptide TIAETLAEELINAAK were identified in 23 and 18 spectra, respectively. All other cross-links had 9 or fewer MS/MS sequencing events. In the SE/C18/TiO_2 workflow, the 93 cross-links were identified in a total of 600 fragment

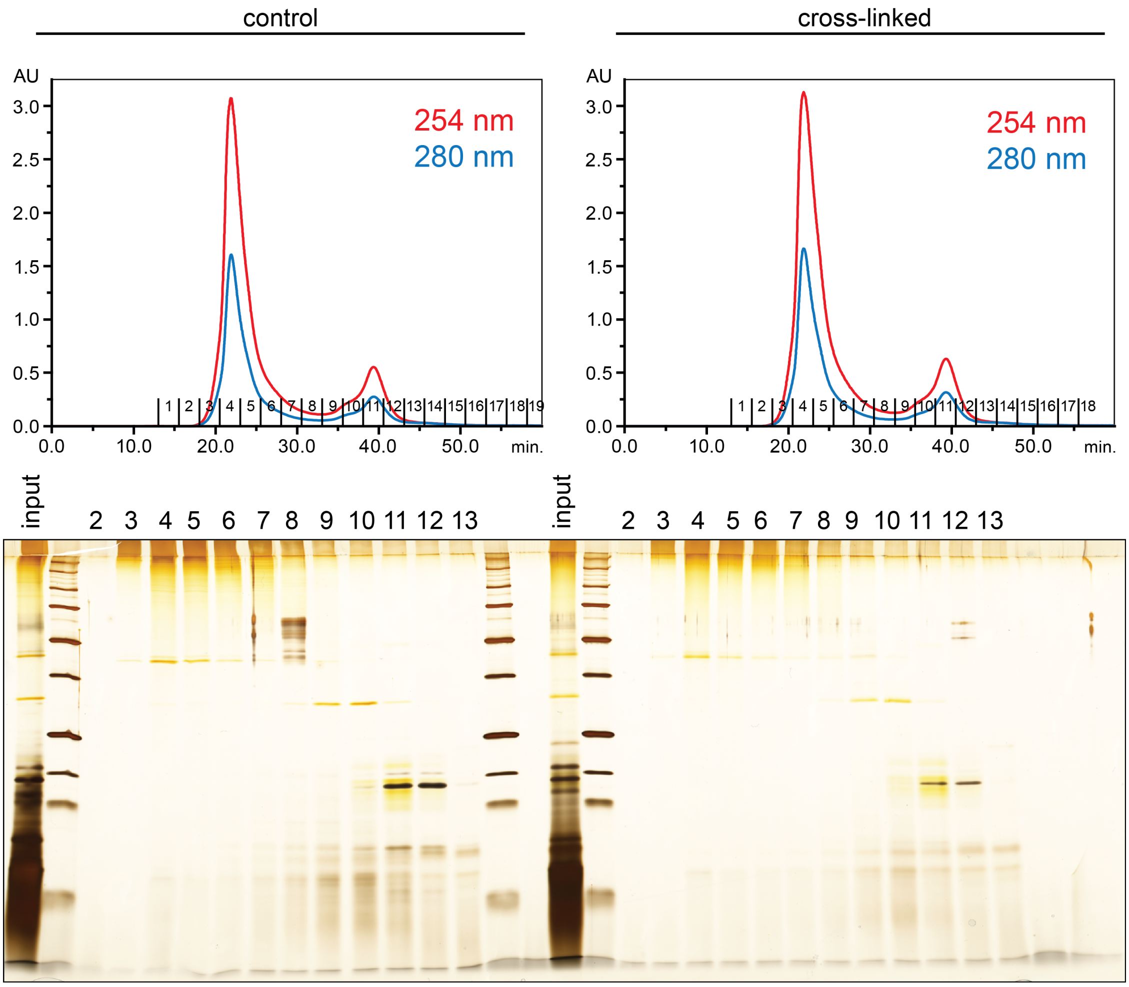


Figure 3.22: Size exclusion chromatogram and gel of corresponding fractions. The upper part shows two size exclusion chromatograms with the absorption at both 254 nm (red, RNA) and 280 nm (blue, peptides). Left is the absorption profile of a non-irradiated control, right the chromatogram of a UV irradiated sample. In the lower part, a protein gel of the most important fractions is shown. Fraction numbers are annotated at the bottom of the chromatograms and above the gel lanes. Neither chromatograms nor gel lanes show significant differences between the control and UV samples.

spectra. A single cross-link of 60S ribosomal protein L16-B peptide AEALNISGEFFR was identified in 107 spectra, and four cross-links had between 20 and 40 sequencing events each. Since a high number of MS/MS spectra for the same cross-link does not increase confidence and the overall number of identified cross-links remained the same, titanium dioxide enrichment was not performed in subsequent experiments.

One interesting observation in comparing both workflows was the frequent identification of cross-links that had lost water after the SE/C18 sample preparation. This might be due to formation of cyclic phosphate on the RNA 3' end and is apparently reversible, since the SE/C18/TiO_2 sample preparation did not display the same effect. The basic elution conditions of titanium dioxide enrichment might lead to hydrolization of the 3' phosphates.

3.4.2.4 Gradient for LC-ESI-MS/MS analysis

In the last experiments for optimization of sample preparation, we also compared the length of the LC gradient in LC-MS/MS analysis. A longer gradient increases peak separation and enables the mass spectrometer to collect fragment information for more precursors. On the other hand, elution profiles are broadened and, consequently, peak intensities drop. Lower intensities decrease the signal intensities in the mass spectra, and precursors as well as their fragments might fall below the detection limit. In order to evaluate which effect would outweigh the other, the same sample was measured with 60- and 120-minute gradients. The number of identified cross-links was significantly higher for the 120-minute gradient, rising from 32 to 93. Therefore, longer LC gradients should be performed for complex cross-linking samples.

3.4.2.5 MS instrumentation

All experiments described above were analyzed on a LTQ Orbitrap Velos. Another orbitrap instrument, an Orbitrap Exactive, became available for LC-MS/MS analysis towards the end of the project and was also evaluated. In contrast to the Velos, the Exactive does not contain a linear ion trap and only allows fragmentation with HCD. The scan speed is significantly faster compared to the Velos, potentially leading to sequencing of more low abundant precursors. For a comparison of both instruments, a sample was prepared with the optimized protocol and split for measurements on both instruments. The Exactive did indeed acquire a significantly higher number of fragment spectra compared to the Velos (18 914 vs. 11 883). We observed a substantially higher number of low-quality spectra that did not allow any identification in the Exactive measurement. In a rough evaluation, the Velos measurement yielded the identification of only 25 cross-links, while the Exactive measurement resulted in 35 cross-link IDs. Therefore, the Q Exactive mass spectrometer should be more closely evaluated and might replace the LTQ Velos as the instrument of choice for cross-linking experiments in the future.

3.4.3 Data analysis and integration of additional filters

All experiments described above were analyzed with the RNP^{xl} tool. Overall, 18 experiments (see Table 3.6, each with corresponding controls) were carried out for optimization and with the optimized protocol. During evaluation of the different variations of the protocol, we began to collect a library of identified cross-links that was continuously expanded. The majority of cross-links were identified in detailed analysis of measurements with the optimized protocol.

3.4.3.1 Validation of cross-link candidates: Extracted ion chromatogram and independent database search

The output of RNP^{xl} contained a long list of peptides and potential cross-links. Particularly low quality spectra often lead to false positive search results, i.e. random matches, after precursor variant generation. Therefore, careful validation of cross-link candidates is essential. While several validation criteria have been mentioned in previous sections, the validation process was again optimized for complex samples as described here. The significantly higher number of identified

Table 3.6: Overview on experiments in yeast

experiment	UZ	XL	IgG	CB	SE	C18	TiO_2	MS	gradient
1	-	extract	+	+	-	+	+	Velos	60 min
2	-	extract	+	+	-	+	+	Velos	60 min
3	-	CB eluate	+	+	-	+	+	Velos	60 min
4	-	CB eluate	+	+	-	+	+	Velos	60 min
5	-	IgG eluate	+	+	-	+	+	Velos	60 min
6	-	IgG eluate	+	+	-	+	+	Velos	60 min
7	-	IgG eluate	+	+	-	+	+	Velos	60 min
8	-	IgG eluate	+	+	-	+	+	Velos	120 min
9	+	IgG eluate	+	+	-	+	+	Velos	60 min
10	+	IgG eluate	+	+	-	+	+	Velos	120 min
11	+	IgG eluate	+	-	+	+	-	Velos	60 min
12	+	IgG eluate	+	-	+	+	-	Velos	120 min
13	-	IgG eluate	+	-	+	+	-	Velos	60 min
14	-	IgG eluate	+	-	+	+	-	Velos	120 min
15	-	IgG eluate	+	-	+	+	-	Velos	120 min
16	-	IgG eluate	+	-	+	+	+	Velos	120 min
17	-	IgG eluate	+	-	+	+	-	Exactive	100 min
18	-	IgG eluate	+	-	+	+	+	Exactive	100 min

UZ: ultracentrifugation; CB: Calmoduline beads

cross-links compared to previous experiments, together with a considerable increase of experience in evaluation of cross-link spectra, allowed to derive more general rules for validation as well as an extended summary of observed RNA signals and RNA adducts (see below). Finally, validation criteria were integrated in automated filtering tools (see below).

High-scoring cross-links were initially validated by comparison of the extracted ion chromatograms (XICs) in irradiated sample and the corresponding control (see 3.2.5.1). Typically, the control did not contain any precursor with the same mass and a comparable retention time, while the sample XIC showed a clear signal (as exemplified in Figure 3.18a). If the cross-link candidate was only observed in the irradiated sample, it could be concluded that it had to be a species formed as a consequence of UV irradiation. Otherwise, the cross-link candidate was discarded as a false positive.

In addition, a search against the entire NCBI database was performed with Mascot as search engine. Cross-link candidates should not yield any significant, true positive hit in this search, otherwise they would be false positives. This way, contaminant peptides with sequences included in the NCBI database used for the Mascot search but not in the UniProt yeast database used for cross-link identification could be identified. If a sample was contaminated in contrast to the control, this validation step would exclude them, although they were exclusively observed in the irradiated sample.

Additionally, post-translational modifications (PTMs), especially phosphorylation, could be considered in the Mascot search. In principle, PTMs could also be included in the precursor variant searches for cross-link identification. However, additional PTMs greatly increase analysis time and number of false positive results. In the case of phosphorylation, ambiguity is introduced, as loss of HPO_3 is usually considered as a modification of RNA in the generation of precursor mass variants. When including phosphorylation as a PTM for cross-linked peptides, we frequently observed that phosphorylated peptides were reported as cross-linked to an RNA oligonucleotide without terminal phosphate. Manual evaluation revealed that the assignment of the phosphorylation on the peptide was wrong; the cross-link was manually assigned as the same peptide without phosphorylation cross-linked to an RNA with terminal phosphate. In principle, the database search considers both options, i.e. cross-link of unmodified peptide and oligonucleotide, and phosphorylated peptide cross-linked to RNA without terminal phosphate. Since the unmodified peptide can be clearly identified manually by the peptide sequence ions and the absence of any sequence ion shifted by the mass of HPO_3, the database search engine should also yield a higher score for the unmodified peptide. It remains elusive why this was not the case. In order to limit data analysis time, number of false positives, and ambiguity, the number of PTMs considered was minimized. Only oxidation of methionines and carbamylation of primary amines, i.e. lysine residues and peptide N-termini, were considered.

3.4.3.2 Validation of cross-link candidates: Mass spectra

After the cross-link candidate was confirmed by manual validation of XICs and comparison to an independent database search, the corresponding mass spectra were evaluated in detail. Correct assignment of monoisotopic mass and charge state of the precursor were confirmed on the basis of the MS spectrum preceding the MS/MS fragment spectrum under investigation. In addition, isotopic pattern and co-eluting precursors in the selection window were assessed. In case of ambiguities, the candidate was discarded or an alternative spectrum of the same candidate was chosen for further analysis.

The last important step in validating a cross-link candidate was the evaluation of the MS/MS fragment spectrum. The search results from the RNP^{xl} tool were annotated onto the raw MS data in TOPPView. Each cross-link candidate was evaluated individually. MS/MS signals corresponding to peptide fragments were automatically annotated. Next, the presence of RNA signals was investigated (observed signals with corresponding m/z values are given in Table B.1). Cross-links to a single U nucleotide usually do not show RNA marker ions. In contrast, the vast majority of cross-link spectra with oligonucleotides is dominated by RNA marker ions corresponding to the nucleic acid bases A, G, and/or C. The only exception are cross-links to poly(U) oligonucleotides due to the low proton affinity of uracil. Peaks remaining unassigned were further investigated if they were of high intensity and/or showed the pattern of an amino acid sequence. Peaks in the lower m/z range are often internal peptide fragments or adducts of immonium ions and RNA (see Table B.3). Peaks in the higher m/z range often correspond to RNA adducts of peptide fragments; observed mass shifts are listed in Table B.2.

After manual annotation of unassigned fragments, the final judgment on the peptide candidate was made according to the following criteria: (1) All high intensity ions, especially in the mass range $m/z > 400$, should be explained by the cross-link. (2) The a2/b2 ion pair and frequently observed immonium ions should be present unless the corresponding amino acids were cross-linked. (3) The C-terminus of the peptide should be covered by corresponding y-ions. Cross-link candidates fulfilling all criteria described above were considered as true positive, manually validated hits. Candidates violating any of the rules were disregarded as false positives.

3.4.3.3 Integration of exclusion criteria into automated filtering tools

The RNP^{xl} tool proved highly useful in data analysis of cross-linking experiments. However, manual validation by comparison of extracted ion chromatograms and independent database searches was still necessary and presented a time-consuming task. Therefore, we set out to implement these two important criteria into the data analysis workflow that are used to exclude spectra of noncross-linked species. More precisely, we wanted to remove these fragment spectra early in the data analysis workflow, rather than discarding them as noncross-linked species retrospectively when following the established procedure. The main benefit would be decreased effort for manual filtering of these false positives. Additionally, processing time would be decreased, as noncross-linked species do not need to be considered in the precursor variant searches.

Fragment spectra of noncross-linked peptides are identified as such by the RNP^{xl} tool if their sequences and modifications are considered in the corresponding search. However, their identification in a standard database search is more straightforward, and generation of precursor mass variants for noncross-linked peptides represents an unnecessary processing step. In addition, the number of protein sequences and post-translational modifications should be limited in the precursor variant searches in order to minimize false positive results. To identify noncross-linked peptides, including those with less frequent PTMs (e.g. phosphorylation, see above) or peptides from contaminating proteins (e.g. keratins), we included an additional filtering step. A standard database search was performed on the MS data obtained after the cross-linking experiment with OMSSA in the OpenMS environment. The database contained the protein sequences anticipated for cross-linked peptides, i.e. here the *S. cerevisiae* proteome, and contaminant sequences, e.g. keratins and enzymes. We used the contaminant sequences distributed with MaxQuant [103]. In order to determine false positive rates with the target-decoy strategy, a target-decoy version of the database was created. Phosphorylation of serine, threonine, and tyrosine; oxidation of methionine; and carbamylation of lysines and peptide N-termini were considered. All peptide identifications below a false discovery rate of 1% were considered as valid IDs and the corresponding spectra were filtered from the MS data file. The dedicated OpenMS pipeline for this purpose is described in more technical detail in 2.2.10.6.

The second important validation criterion is the comparison of cross-link candidate XICs in the UV irradiated sample to XICs of the same precursors in the control sample (see 3.2.5.1). All species showing comparable intensities in both measurements are not connected to UV irradiation and can be excluded from further analysis. As no existing OpenMS tool allows such assessment, our collaborators scripted a new tool called RNP^{xl}XICfilter. This tool calculates extracted ion chromatograms in both measurements (control and UV) for all precursors fragmented in the UV irradiated sample. Precursor intensities are summed up in a narrow retention time window. If the same precursor appears in both measurements with comparable intensity, i.e. difference smaller than a factor of 2, the corresponding fragment spectrum is not written in the output file. A more technical and detailed description of the XIC filter pipeline can be found in 2.2.10.6.

Figure 3.23 illustrates the benefit of both ID and XIC filtering. Typically, the ID filter identifies between 20 and 40% of all spectra as peptides with high confidence (FDR $< 1\%$). Obvious outliers are experiments 16 and 18. In these experiments, isolation of cross-linked heteroconjugates was achieved with size exclusion, C18, and titanium dioxide chromatography. Therefore, the low number of confident identifications by the ID filter is in excellent agreement with our previous observation

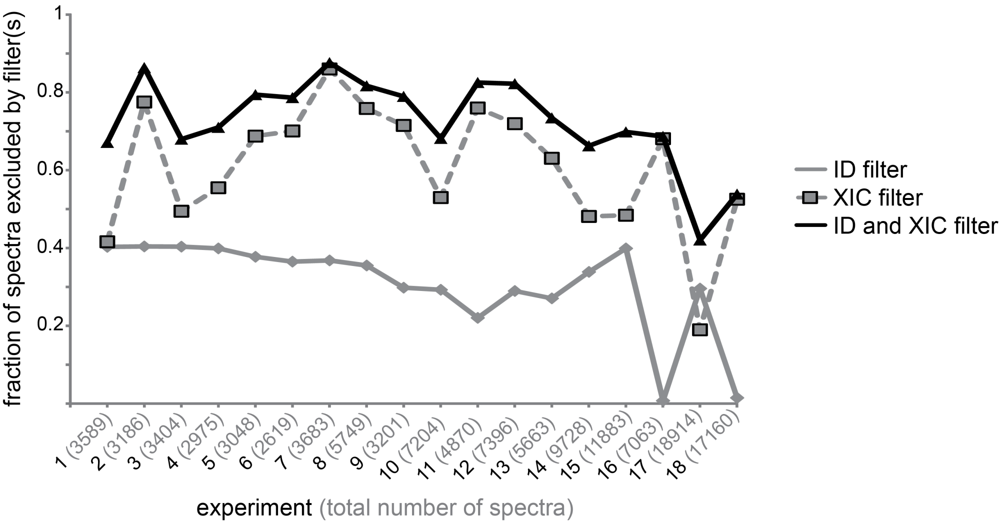


Figure 3.23: Effect on spectra numbers after filtering with ID and XIC filter across experiments. The fraction of spectra excluded from further analysis after application of the respective filter(s) are shown for each of the 18 experiments. Experiments are numbered according to Table 3.6. The total number of spectra in each measurement is given after the experiment number. The ID filter typically removed between 20 and 40% of spectra, the XIC filter between 50 and 80%. Combination of both filters typically excludes over 60% of all fragment spectra from further analysis.

that the combination of all three isolation methods dramatically decreases the number of noncross-linked peptides after enrichment (see 3.4.2.3).

The XIC filter removes a higher fraction of spectra than the ID filter, between 50 and 80% in most measurements. The XIC filter excludes all spectra of species appearing in both the UV irradiated sample and negative control. This can be spectra of noncross-linked peptides, RNA oligonucleotides, or other contaminants. The latter two are not considered in the ID filter. As it does not rely on identifications, the XIC filter can remove peptides indiscriminant of sequence, PTMs, protease specificity, and quality of fragment spectra. Thus, the XIC filter excludes more spectra than the ID filter.

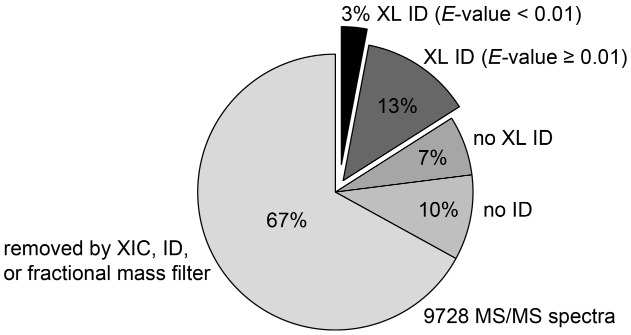


Figure 3.24: Effect on spectra numbers after filtering with ID and XIC filter in a single experiment. A total number of 9728 MS/MS spectra were acquired for experiment 14. Two thirds were filtered by XIC, ID, and fractional mass filter. 17% did not yield a potential cross-link identification in database search after precursor variant generation. 13% were identified as potentially cross-linked, but with a very low score (E-value above 0.01). The remaining 318 spectra (3%) had a good score (E-value below 0.01) and were further evaluated manually.

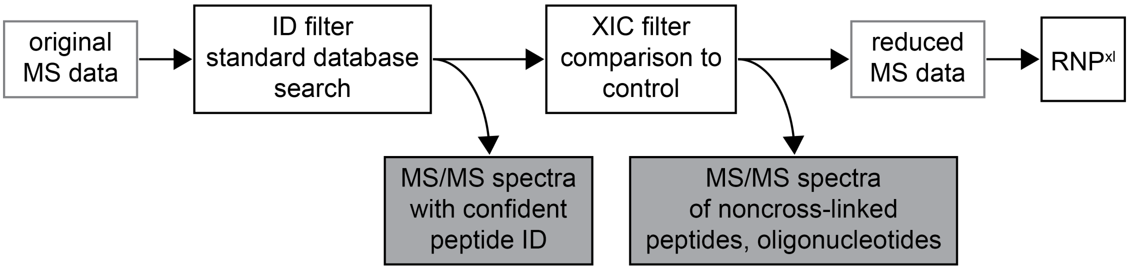

original MS data
ID filter standard database search
XIC filter comparison to control
reduced MS data
RNP^xl
MS/MS spectra with confident peptide ID
MS/MS spectra of noncross-linked peptides, oligonucleotides

Figure 3.25: Schematic workflow for automated filtering of cross-linking data. The original MS data is subjected to a standard database search (ID filter). MS/MS spectra that give rise to confident peptide identifications with a low false discovery rate are removed from the data set. Next, extracted ion chromatograms of precursors with remaining MS/MS spectra are calculated in both the control and UV irradiated samples. MS/MS spectra of species appearing in both samples at comparable intensities are filtered from the MS data set. The reduced data can then be submitted for subsequent analysis with $\mathrm{RNP}^{\mathrm{xl}}$.

The benefit of the applied filters is further illustrated in Figure 3.24. Out of the total number of 9728 MS/MS spectra acquired during MS analysis of experiment 14, only 318 were considered for manual validation. The vast majority of spectra were excluded as peptides, oligonucleotides, or contaminants by the ID and XIC filter; or did not yield a potential cross-link hit with a reasonable score in the database search after precursor variant generation.

The most efficient strategy is the combination of both filters as outlined in Figure 3.25. We first apply the ID filter because we can retrieve the peptide sequences corresponding to the filtered spectra. The XIC filter is applied second; this order can be chosen freely. As visible in Figure 3.23, the combination of both filtering strategies excluded over 60% of all spectra in experiments 1 to 15. On average, the combination of both removed 12% more spectra than the XIC filter alone. Outliers are experiments 16 and 18, which used SE, C18, and $\mathrm{TiO_2}$ enrichment, and experiments 17 and 18, which were analyzed with the Q Exactive mass spectrometer. Due to the low number of peptide identifications in experiments 16 and 18, the combination of both filters naturally does not significantly exceed the results of the XIC filter alone. We currently have no explanation why the XIC filter excluded considerably fewer spectra in the Exactive measurements (experiments 17 and 18), this observation needs to be explored further.

3.4.4 Cross-links identified after TAP tag purification and isolation of cross-linked heteroconjugates

Overall, we have identified 184 cross-links after TAP tag purification of protein–RNA complexes from yeast. They are listed in detail with corresponding calculated and experimental masses in Tables A.5, A.6, A.7, and A.8. In this context, cross-links refers to unique combinations of peptide and RNA oligonucleotide, thus counting cross-links of several oligonucleotides to the same peptide, as well as peptides with and without missed cleavage sites covering the same protein region.

The 184 cross-links identified correspond to 64 unique RNA-binding protein regions, i.e. disregarding cross-links of several oligonucleotides to the same peptide and misscleaved peptides covering the same region as a peptide without missed cleavage site. These 64 protein regions were mapped to 49 proteins. In 37 of the 64 unique protein regions, the cross-linked amino acid residue could be identified. Cysteines were by far the most cross-linked residues with 22 identifications, followed by phenylalanine with four, histidine and tyrosine with three, tryptophane with two, and finally threonine, lysine, and isoleucine with one cross-link each.

The applied experimental workflow does not allow for the identification of the cross-linked RNA. For a confident identification of the cross-linked peptide, the RNA has to be hydrolyzed to single nucleotides or short oligonucleotides (see 1.3.2). These short sequences are not sufficient to identify the cross-linked RNA. In addition, protein–RNA complex isolation by TAP tag purification of capped mRNA bound by Cbp20 was performed under native conditions. Therefore, not only proteins (or RNA) cross-linked or directly interacting with capped mRNA were purified. Ribonucleoprotein complexes interacting with mRNA were also isolated. Cross-links within those complexes are also enriched and can be identified in the mass spectrometric analysis. For example, the high number of cross-linked ribosomal proteins suggests that large amounts of ribosomes were isolated with this experimental workflow. Cross-links within the ribosome, e.g. to ribosomal RNA, were thus also identified (see below).

Finally, the cross-linked proteins were sorted according to their annotated functions. If not noted otherwise, protein information has been derived from the UniProt database [115] and the Saccharomyces Genome Database (Stanford University, www.yeastgenome.org). The majority of cross-links were found for ribosomal proteins, specifically 11 proteins from the small ribosomal subunit and 23 proteins from the large ribosomal subunit. In addition, the ribosome-related ribosome biogenesis protein RLP7 was identified as cross-linked to RNA. Six proteins were found that have annotated polynucleotide-binding function to either RNA or DNA. Eight proteins were identified as cross-linked to RNA that did not have any polynucleotide-binding function annotated in the UniProt database. Cross-links of the three functional groups will be examined in more detail below while a general discussion of the technique can be found in chapter 4.

3.4.4.1 Cross-links of ribosomal proteins

Protein–RNA complex isolation by TAP tag purification with a tagged version of the cap-binding protein Cbp20 was carried out under native conditions. Consequently, all macromolecules with primary or secondary interactions with capped RNA were isolated with the applied protocol. Isolation of ribosomes, e.g. in the process of translating a capped mRNA, was therefore expected. The high number of identified cross-links of ribosomal proteins was, at least in part, a consequence of their high abundance within the cell. Since the applied protocol might have isolated intact ribosomes, observed cross-links could have resulted from interactions with messenger RNA (mRNA), ribosomal RNA (rRNA), or even transfer RNA (tRNA, see above).

A high resolution structure of the *S. cerevisiae* ribosome is available [116], and cross-linking results could be compared with structural data. Sebastian Klinge (laboratory of Prof. Nenad Ban, ETH Zurich) provided helpful advice about visualization of the ribosomal structure within PyMOL.

Due to the large number of cross-links in the ribosome (see Table 3.7), only four representative examples are described below and shown in the structure of the ribosome. Cross-links of ribosomal proteins were roughly sorted into three categories after comparison with the three-dimensional structure: Most cross-linked amino acid residues are in close proximity to nucleotides of the ribosomal RNA in the available structure. Several cross-linked residues are in flexible regions on the

Table 3.7: Cross-linked ribosomal proteins

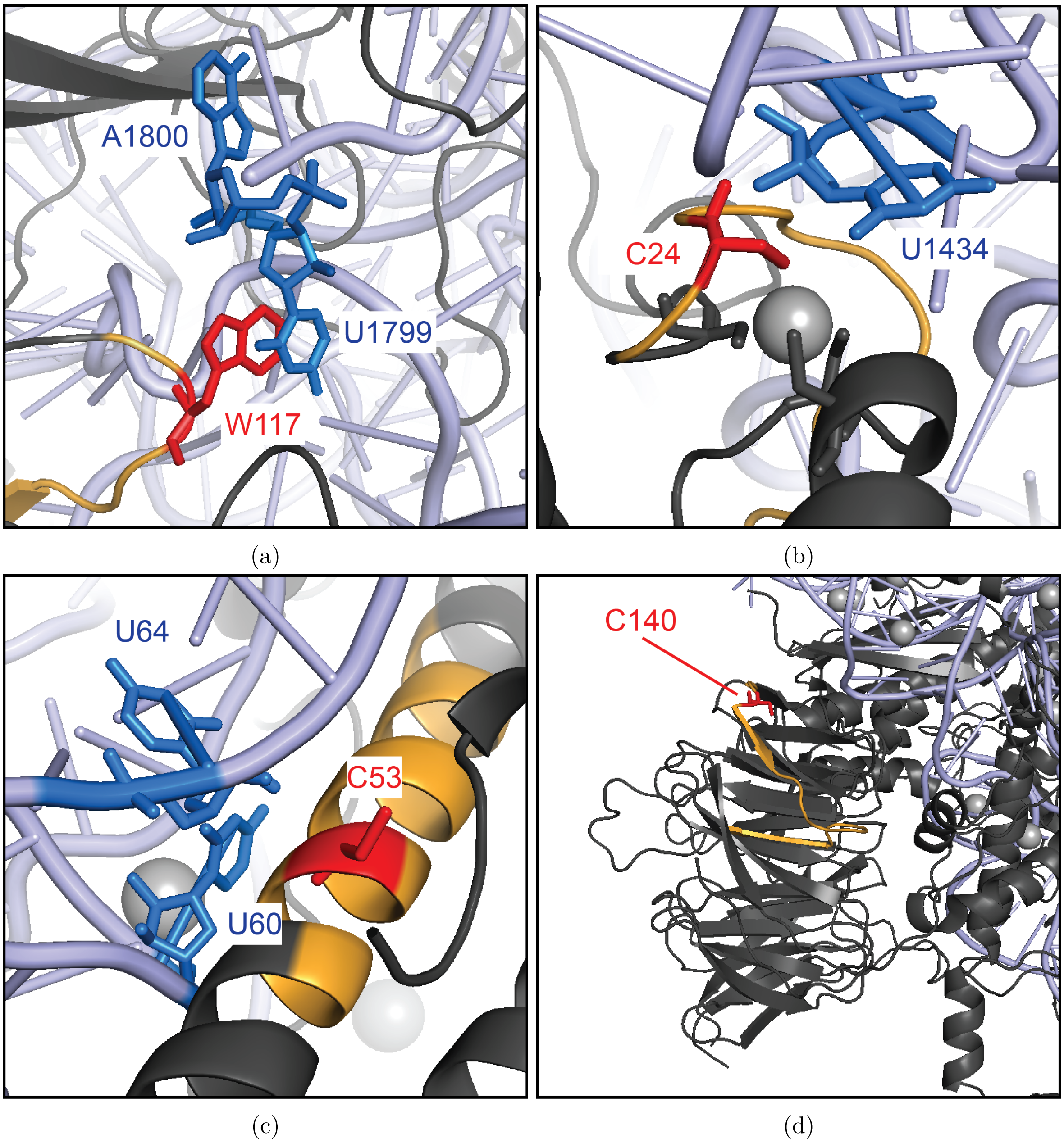
proteins of the 40S subunit		proteins of the 60S subunit	
protein	cross-linked residue	protein	cross-linked residue
S1-A/-B	W117	L1-A/-B	C80
S3	C134	L2-A/-B	Y133
S5	T189	L3	C251
S11-A/-B	C128	L4-A/-B	C94, H243, I290
S14-A/-B	-	L5	-
S16-A/-B	H74	L6-A	W9
S17-A/-B	C35, H56	L6-B	W9
S24-A/-B	K117	L8-A	-
S29-A	C24	L8-B	-
S29-B	C24	L16-A/-B	Y149
Rack1*	C140	L16-A	-
		L16-B	F38
		L18-A/-B	C121
		L23-A/-B	C122
		L26-B	-
		L28	Y48
		L31-A/-B	F25
		L33-A/-B	-
		L35-A/-B	C53
		L37-A	-
		L37-B	-
		L40**	C115
		L42-A/-B	C88/K89

recommended name (UniProt):
*guanine nucleotide-binding protein subunit beta-like protein
**ubiquitin 60S ribosomal protein L40

surface of the ribosome. In the crystal structure, no nucleotides are in close proximity to these amino acids. Therefore, it is not clear whether the cross-link was to flexible rRNA regions not resolved in the structure, or to mRNA or tRNA. The smallest group of cross-links lies in proximity to RNA, but distances between amino acid residues and nucleotides appear to be too large for cross-link formation.

Examples for cross-links within the ribosome are shown in Figure 3.26. Peptide KWQTLIEANVTVK (K116–K128) of ribosomal protein S1 was found to cross-link via W117 (see Figure B.16). In the three-dimensional structure, W117 was found in a loop of S1 and close to nucleotide U1799 of 18S rRNA (see Figure 3.26a). The spatial arrangement of W117 and U1799 suggests that both participate in stacking interactions. In addition, one cross-link of the peptide was to RNA with the composition [AU $-HPO_3$]. This is in excellent agreement with A1800 being the 3' end of 18S rRNA and not containing a 3' phosphate.

Another protein from the small ribosomal subunit, S29, was found to cross-link via C24 (see spectra in Figures B.28 and B.29). Comparison to the crystal structure indicates that C24 coordinates to zinc and is in close spatial proximity to nucleotide U1434 of 18S rRNA (see Figure 3.26b). Formation of a covalent bond between C24 and U1434 as a consequence of UV irradiation would be possible, especially as C24 lies within a loop of S29.



(a) (b)

(c) (d)

Figure 3.26: Representative examples for cross-links in the ribosome. Ribosomal proteins are shown in gray, cross-linked peptides in orange, and cross-linked amino acid residues in red. Ribosomal RNA is shown in light blue, highlighted nucleotides in dark blue. Zinc ions are depicted as gray spheres.
(a) W117 of ribosomal protein S1 stacks with U1799 of 18S rRNA.
(b) Zinc-coordinating residue C24 of ribosomal protein S29 is in close spatial proximity to U1434 of 18S rRNA.
(c) Ribosomal protein L35 residue C53 is in proximity to 5.8S rRNA, e.g. nucleotides U60 and U64.
(d) Residue C140 of the protein Rack1 lies within a flexible region of the ribosome, the crystal structure does not show any RNA in spatial proximity.

In contrast to the previous examples, cross-linked residue C53 of ribosomal protein L35 (see spectrum shown in Figure B.58) lies within an α-helix, and the three dimensional structure does not show any uridines in close spatial proximity (see Figure 3.26c). The 5.8S rRNA lies adjacent to the cross-linked peptide and contains two uridines (U60 and U64) in this stretch, one of which might be the cross-linked nucleotide. However, formation of a covalent bond upon UV irradiation probably requires smaller distances between the nucleotide and the amino acid residue. One possibility would be that the structure of the active ribosome in solution might rearrange in this area to bring C53 closer to the RNA.

Rack1 (guanine nucleotide-binding protein subunit beta-like protein) peptide GQCLATLLGHNDWVSQVR (G138–R155) was found cross-linked via C140 (spectrum shown in Figure B.30). The protein lies at the head of the small ribosomal subunit and close to the mRNA exit channel. In the structure, no RNA is found in proximity to residue C140 (see Figure 3.26d). No final conclusion can be drawn whether the protein cross-links to flexible regions of the rRNA that are not resolved in the crystal structure, or to mRNA in actively translating ribosomes.

3.4.4.2 Cross-links of RNA-/DNA-binding proteins

Overall, six proteins were identified as cross-linked that had annotated functions as RNA- (or DNA-) binding proteins. These proteins are listed in Table 3.8. Unfortunately, no structural data is available which would allow comparison of the cross-linking data with protein–RNA contacts on a molecular level.

Table 3.8: Cross-linked RNA-/DNA-binding proteins

	recommended name	gene	synonym	motif
1	cruciform DNA-recognizing protein 1	CRP1	Crp1p	
2	elongation factor 1-alpha	TEF1/TEF2	eEF1A	
3	nucleolar protein 3	NPL3	Nop3	RRM
4	nucleolar protein 13	NOP13	Nop13	RRM
5	polyadenylate-binding protein	PAB1	PABP	RRM
6	single-stranded nucleic acid-binding protein	SBP1	Sbp1p	RRM

Numbers are indicated at the corresponding paragraphs in the text.

(1) Peptide IPEAGGLLCGKPPR (I105–R118) was found to be cross-linked to RNA via residue C113 (spectrum see B.64). This peptide is found in a protein termed cruciform DNA-recognizing protein 1 because it was isolated with DNA templates resembling cruciform DNA [117]. There are no reports about additional functional roles of this protein. Crp1p is cleaved into an approximately 160 amino acid long N-terminal, DNA-binding subpeptide; residues 120 to 141 are necessary for DNA binding. Neither the DNA-binding region nor the protein show significant homology to other cruciform DNA-binding domains or proteins [117]. The identification of Crp1p as an RNA-binding protein indicated an additional function of the protein, more specifically, to the region directly upstream of the DNA-binding region.

(2) A protein related to ribosomal translation that was found cross-linked to RNA was elongation factor 1-alpha. eEF1A is a component of the eukaryotic elongation factor 1 complex and delivers aminoacyl-tRNA to the A-site of the ribosome. Peptide FVPSKPMCVEAFSEYPPLGR (F402–R421; spectrum see B.65) was found to directly interact via C409. In the primary as well as the

tertiary structure (data not shown), the cross-linked region is on the opposite side to the GTP binding residues (14–21, 91–95, 153–156). The cross-linked RNA cannot be identified from the mass spectrometry data (see above), but one possibility is that eEF1A cross-linked to tRNA and was purified while bound to ribosomes translating capped mRNA.

The other four proteins with annotated RNA-binding function that were found cross-linked all contain RNA recognition motifs (RRMs). Interestingly, in all cases, the cross-linked peptides were within these RRMs:

(3) Two regions of nucleolar protein 3 were found cross-linked. The first, peptide ILNGFAFVEFEEAESAAK (I156–K173, spectrum shown in B.66), lies within the first RRM of the protein (positions 125–195). The second cross-linked peptide, ENSLETTFSSVNTR (E222–R235, spectrum see B.67), is part of the second RRM (positions 200–275). Nop3 has various functions, including nuclear export of poly(A) mRNA and in splicing.

(4) Nucleolar protein 13 cross-links via peptide ILFVGNLSFDVTDDLLR (I240–R256), more precisely via F242 (see spectrum in B.68). The peptide is at the N-terminus of the second RRM domain of the protein (positions 239–317). The protein is found in preribosomal complexes.

(5) Polyadenylate-binding protein peptide YQGVNLFVK (Y319–K327) was found to cross-link via F325 (spectrum shown in B.69). The cross-linked residue lies within the fourth RRM of the protein (positions 322–399). Interactions between PABP and polyadenylated RNA are important for mRNA export into the cytoplasm, mRNA stability, and translation.

(6) Peptide SKDTLYINNVPFK (S184–K196) of single-stranded nucleic acid-binding protein was found cross-linked (see spectrum in B.70). This region lies at the N-terminus of the second RRM of Sbp1p (positions 186–274). The protein plays a role in translational repression and decapping.

3.4.4.3 Cross-links of proteins without annotated polynucleotide binding function

The third group of proteins found cross-linked to RNA contains proteins without annotated RNA (or DNA) binding properties. Interestingly, this group contains exclusively metabolic enzymes (summarized in Table 3.9). Three of those enzymes, adenosylhomocysteinase, alcohol dehydrogenase, and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) contain a Rossmann fold domain. Since this domain has been proposed to have RNA-binding properties (discussed in 4.1.5), the structure and function of these three proteins will be described in more detail.

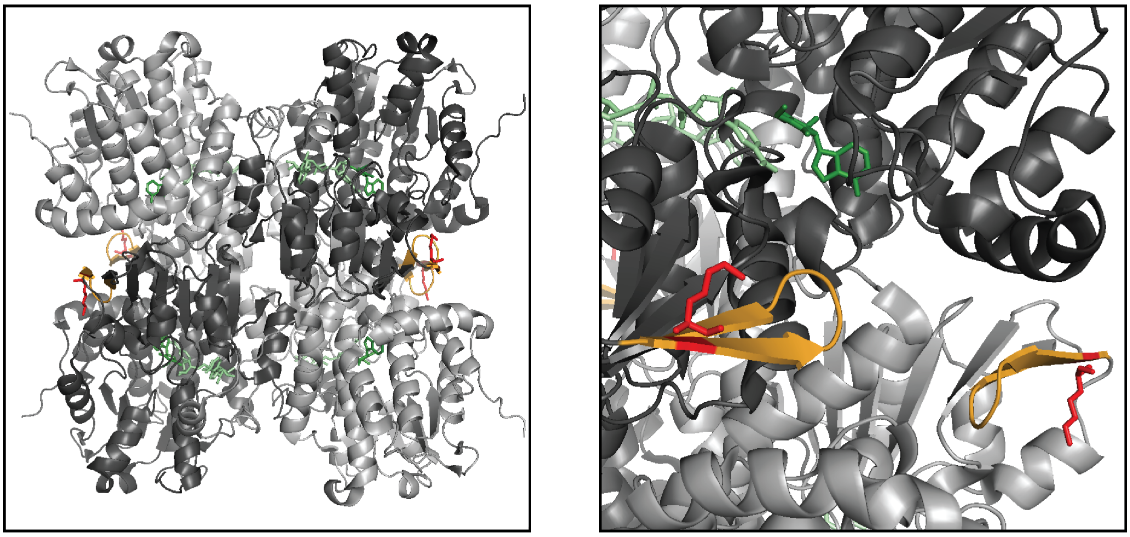
(1) Adenosylhomocysteinase (Sah1p) peptide ECINIKPQVDR (position E320–R330) was found to cross-link via C321 (spectrum shown in Figure B.71). Sah1p hydrolyzes S-adenosyl-homocystein (AdoHcy) into homocystein and adenosine. AdoHcy is a competitive inhibitor of S-adenosyl-L-methionine-dependent methyl transferase reactions. As the latter are important for formation of the cap structure on viral mRNAs, Sah1p inhibition leads to antiviral activity. A structure of the human homolog S-adenosylhomocystein hydrolase is available and shown in Figure 3.27. Yeast and human protein share 70% sequence identity, and the cross-linked peptide is almost identical to the human sequence EKVNIKPQVDR (E320–R330). The only exception is the cross-linked C321, the human protein has a lysine residue at this position. The peptide lies within the NAD binding Rossmann fold domain but does not contact the cofactor directly (see Figure 3.27b) [118].

(2) A cross-linked peptide of the sequence YSGVCHTDLHAWHGDWPLPVK was identified (see Figure B.72). This sequence can be found in two highly homologous alcohol dehydrogenases (around

Table 3.9: Cross-linked proteins without annotated RNA-binding function

	recommended name	gene	synonym	protein feature
1	adenosylhomocysteinase	SAH1	Sah1p	Rossmann fold
2	alcohol dehydrogenase 1/3	ADH1/ADH3	ADH1/ADH3	Rossmann fold
3	glyceraldehyde-3-phosphate dehydrogenase 2/3	TDH2/TDH3	GAPDH 2/ GAPDH 3	Rossmann fold
4	enolase 1/2	ENO1/ENO2	Eno1p/Eno2p	
5	inorganic pyrophosphatase	IPP1	IPP1	
6	peroxiredoxin TSA1	TSA1	Tsa1p	
7	phosphoglycerate kinase	PGK1	PGK	
8	pyruvate kinase 1	CDC19	PK	

Numbers are indicated at the corresponding paragraphs in the text.



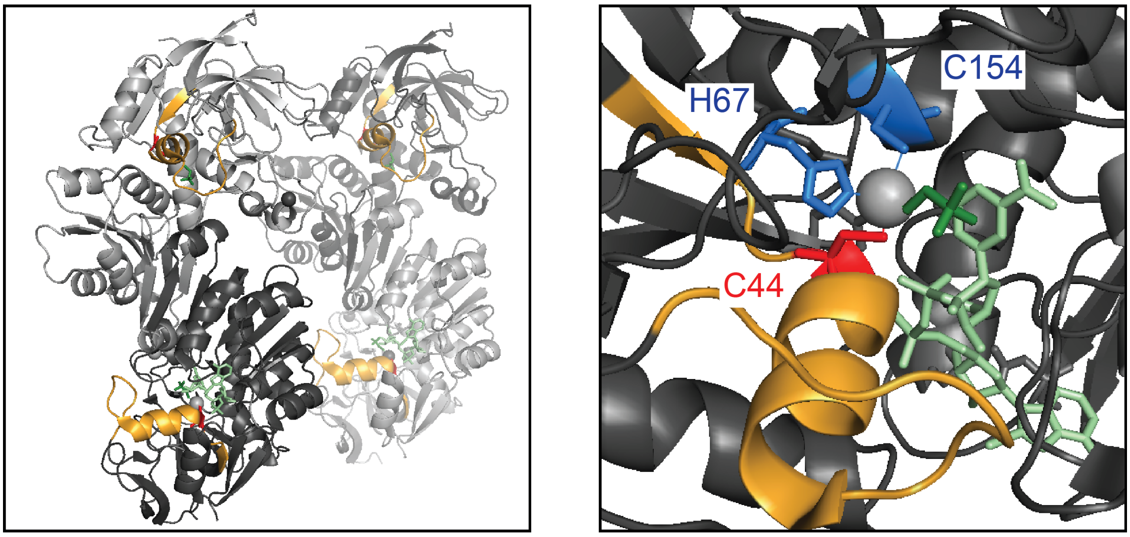
(a) Overview of the tetrameric structure.

(b) Zoom on the cross-linked peptide.

Figure 3.27: Structure of human S-adenosylhomocysteine hydrolyase in complex with nicotinamide-adenine-dinucleotide (NAD, light green) and the inhibitor fluoroneplanocin A (dark green). In the tetrameric structure, each subunit binds one molecule NAD and inhibitor.
pdb 3NJ4 [118]

80% amino acid identity), namely, alcohol dehydrogenases 1 and 3 (Adh1p: Y40–K60, Adh3p: Y67–K86). Alcohol dehydrogenases catalyze the oxidation of an alcohol to its corresponding aldehyde or ketone by NADH. In yeast fermentation, they additionally catalyze the reverse reaction, e.g. the reduction of acetaldehyde to ethanol. Adh1p is the main ADH expressed and active during anaerobic fermentation; Adh3p is a mitochondrial protein. A crystal structure for Adh1p is available (B.V. Plapp, B.R. Savarimuthu, S. Ramswamy; PDB ID 2HCY; no related publication). In the tetramer (see Figure 3.28a), each subunit binds two zinc ions although only one is catalytically active. The cross-linked cysteine residue C44 (Adh3p: C71) lies in a catalytic pocket and coordinates zinc together with H67 and C154 (see Figure 3.28b). All these residues are conserved between Adh1p and Adh3p [119]. The cross-linked peptide does not lie within the C-terminal Rossmann fold.

(3) Another pair of highly homologous metabolic enzymes, glyceraldehyde-3-phosphate dehydrogenases (GAPDH) 2 or 3, are identified by the common peptide ETTYDEIKK (E250–K258) cross-



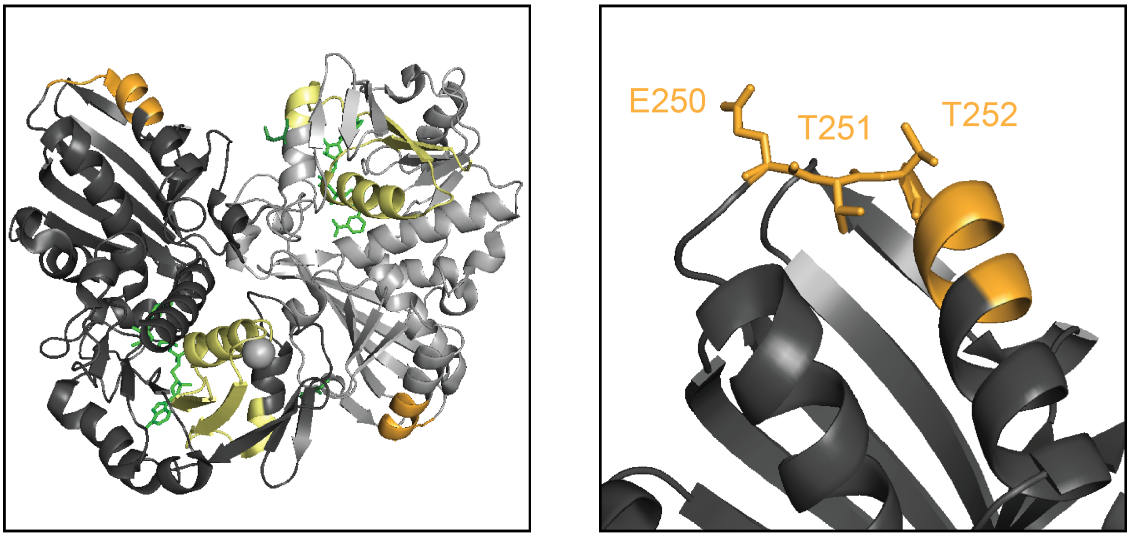
(a) Overview of the homotetrameric structure. (b) Zoom on one catalytic pocket.

Figure 3.28: Structure of yeast alcohol dehydrogenase 1 in complex with nicotinamide-8-iodo-adenine-dinucleotide (light green) and trifluoroethanol (dark green). The cross-linked peptide is shown in orange, and the cross-linked cysteine residue in red. Additional Zn-coordinating residues are in dark blue, with residues involved in NAD-binding in light blue.
(pdb 2HCY, B.V. Plapp, B.R. Savarimuthu, S. Ramaswamy, no related publication.)

linked to [AU $-HPO_3$] (spectrum in Figure B.74). GAPDH is involved in the sixth reaction of glycolysis, namely, the oxidation and phosphorylation of glyceraldehyde-3-phosphate to 1,3-bisphosphoglycerate by NAD^+ and orthophosphate. It was shown that the 43 N-terminal amino acids of the human protein are sufficient to preserve RNA-binding activity of the respective GST fusion protein, but do not preserve the protein's preference to AU-rich elements [120]. A structure is available for GAPDH 3 (I. Garcia-Saez, F. Kozielski, D. Job, C. Boscheron; PDB ID 3PMY; no related publication). The structure shows a homodimer, with each subunit containing a NAD binding Rossmann fold (see Figure 3.29a); however only tetramers are catalytically active. The cross-linked peptide can be found on the surface of the protein, with the peptide N-terminus in a loop region containing amino acids E250, T251, and T252 (see Figure 3.29b). It is proximal to the Rossmann fold of the respective other subunit. The N-terminal protein region homologous to the human protein fragment sufficient for RNA-binding is highlighted in yellow in Figure 3.29a. The cross-link identified could point to the protein region responsible for RNA binding selectivity.

In addition to GAPDH, three other glycolytic enzymes were found to cross-link to RNA, namely enolase, phosphoglycerate kinase, and pyruvate kinase. While these proteins are best known for their role in glycolysis, several other functions and interactions have been described (reviewed in [121]).

(4) Both yeast enolases Eno1p and Eno2p are highly homologous (96% sequence identity). Therefore, the peptide IGLDCASSEFFK (I244–K255) found cross-linked to [U +152 $-H_2O$] via C248 (see spectrum in Figure B.73) could be of either or both enolases. Enolase partakes in the ninth step of the glycolytic pathway, catalyzing the dehydration of 2-phosphoglycerate (2PG) to phosphoenolpyruvate (PEP).



(a) Overview of the dimeric structure.

(b) Zoom on the cross-linked peptide.

Figure 3.29: Structure of yeast glyceraldehyde-3-phosphate dehydrogenases (GAPDH) 3 in complex with meso-erythritol (dark green), sodium and NAD (light green). The 41 N-terminal amino acids connected with RNA-binding are shown in yellow. The cross-linked peptide is shown in orange.
(pdb 3PYM, I. Garcia-Saez, F. Kozielski, D. Job, C. Boscheron, no related publication.)

(5) Yeast inorganic pyrophosphatase is a cytoplasmic protein and hydrolyzes phosphoanhydride. Its peptide NCFPHHGYIHNYGAFPQTWEDPNVSHPETK (N83–K112) was found cross-linked via residue C84 (see spectrum in B.75).

(6) Peroxiredoxin TSA1 is a cytoplasmic protein and belongs to a family of thiol-specific peroxidases. Interestingly, C171, the cross-linked residue in the peptide NGTVLPCNWTPGAATIKPTVEDSK (N165–K188; spectrum shown in B.76), is the resolving cysteine in the catalytic reaction. After the peroxidatic cysteine (C47 in TSA1) is oxidized, it reacts with the resolving cysteine to form a disulfide bond which, in turn, is reduced by thioredoxin to complete the catalytic cycle [122].

(7) Phosphoglycerate kinase (PGK) catalyzes the seventh reaction of glycolysis, transferring a phosphate group from 1,3-bisphosphoglycerate (1,3-BPG) to ADP, yielding 3-phosphoglycerate (3PG) and ATP. PGK peptide YVLEHHPR (Y49–R56) was found cross-linked to RNA; the cross-linked amino acid residue could not be determined (spectrum shown in Figure B.77).

(8) Pyruvate kinase (PK) catalyzes the tenth and final step of glycolysis, conversion of phosphoenolpyruvate (PEP) to pyruvate, thereby transferring phosphate from PEP to ADP to yield ATP. Two regions of pyruvate kinase were identified to interact with RNA: Peptide NCTPKPTSTTETVAASAVAAVFEQK (N370–K394) was found cross-linked via C371 and peptide YRPNCPIILVTR (Y414–R425) via C418 (see spectra in Figures B.78 and B.79).

3.4.5 Summary

This study is the first report of UV induced protein–RNA cross-linking with identification of interaction sites on a peptide or amino acid level in a complex system. In contrast to the *in vitro* reconstituted particles described in the previous sections, sample complexity presented the major challenges for both experimental and data analysis procedures. We have successfully developed an experimental workflow for isolation and enrichment of cross-linked peptide–RNA oligonucleotide heteroconjugates after TAP tag purification of protein–RNA complexes. Key steps of the procedure, i.e. extract preparation, cross-linking, protein–RNA complex isolation, enrichment of cross-linked heteroconjugates and mass spectrometric analysis, were optimized. Most importantly, the data analysis approach based on precursor variant generation proved to be feasible in the identification of cross-links in searches against the entire yeast proteome.

The extent of the collected data on cross-links is unprecedented. Never before has such a high number of proteins been shown to directly interact with RNA on a peptide or even amino acid level from a single set of very similar experiments. In addition, the data has considerably increased the knowledge about the induced dissociation of cross-linked heteroconjugates in the gas phase. Validation criteria for cross-link candidates have been refined, and an extended list of RNA adducts observed after fragmentation has been collected (see Tables B.2 and B.3).

Overall, this feasibility study paves the way for future experiments of UV induced protein–RNA cross-linking with identification of direct interaction sites on a peptide or amino acid level based on mass spectrometric analysis as it will be described in the following chapter.

4 Discussion

UV induced protein–RNA cross-linking with mass spectrometric analysis: advantages, limitations, and future perspectives

The presented work focused on the investigation of protein–RNA interactions by UV cross-linking and mass spectrometry. The obtained results have been discussed in terms of biological context and methodological advances in the previous chapter. In this chapter, the different steps of the experimental methodology and technical aspects will be recapitulated, highlighting prerequisites, challenges and future directions.

In the past, UV induced cross-linking and the identification of protein regions, peptides or amino acids directly interacting with RNA by mass spectrometry had been proven to yield valuable insights into contact sites on a molecular level. In contrast to methods based on DNA sequencing that focus on binding sites on the RNA, the MS based approach has been limited to relatively simple macromolecules; human U1 and U2 snRNPs were the most extended complexes investigated [78]. The long-term goal, the identification of contact sites after *in vivo* cross-linking, does require significant advancements of the method on many levels. In the presented work, considerable progress towards this application was achieved. In addition, valuable experience and knowledge was collected in investigations of different protein–RNA complexes.

The successful identification of protein–RNA interaction sites by mass spectrometry depends on a large number of factors, concerning the biological system under investigation, the experimental strategy and technical aspects of the bioanalytical approach. If a cross-link is observed, it clearly points at a direct protein–RNA interaction due to the high specificity of the UV induced reaction. The reverse conclusion, i.e. that the absence of cross-links corresponds to the lack of binding, is incorrect. Not all proteins that bind RNA form cross-links upon UV irradiation. One example is the protein 15.5K of the human spliceosome. The protein was shown to bind the 5' stem-loop of U4 snRNA [123], and the structure of 15.5K co-crystallyzed with an oligonucleotide resembling the U4 5' stem-loop was solved [124]. However, no cross-link between 15.5K and U4 snRNA was observed, since none of the amino acids interacting with U4 is reactive in UV cross-linking [49]. Similarly, if none of the nucleotides involved in interactions exhibits sufficient reactivity, no (detectable) cross-links will be formed.

4.1 Experimental insights and potentials of UV induced protein–RNA cross-linking

The unprecedented number of protein–RNA cross-links reported in this work allows a more detailed evaluation of the UV induced reaction. The acquired insights on cross-linking on the molecular level will be described, as well as questions that will remain to be answered.

4.1.1 Reactivity of nucleotides

There are considerable differences in the reactivities of nucleic acid bases in UV induced cross-linking. Among all cross-links reported in this work, there is only one example of a cross-link to an oligonucleotide that does not contain a single uridine, namely peptide SKDTLYINNVPFK of the single stranded nucleic acid binding protein (positions S184–K196) cross-linked to a [AC] or [CA] dinucleotide. Cytosine is in general more reactive than adenosine (see 1.3). However, the MS/MS fragment spectrum (shown in Figure B.70) does not contain any signals that would allow conclusions about the cross-linked nucleotide.

Since all except one of the over 250 cross-links reported in this work were observed to uridine or uridine-containing RNA, it is concluded that the difference in reactivity between uridine and the other nucleotides is substantial. Consequently, it would be unlikely to identify cross-linked heteroconjugates by mass spectrometry if the protein–RNA interactions are based on contacts of adenosine, guanosine or cytidine with amino acid residues. If protein–RNA interactions are assumed to involve regions with little or no uridines, incorporation of photoreactive nucleotides might be considered, for example 6-thio-guanine or 5-bromo-cytosine.

For *in vivo* cross-linking approaches, the incorporation of 4-thio-uracil (and 6-thio-guanine; 4SU and 6SG) also provides considerable advantages: In many cases, the cross-linking yield is increased. Since 4SU and 6SG have absorption maxima at a longer wavelength (330 and 340 nm compared to 250–270 nm for unsubstituted nucleotides), irradiation is typically carried out at 365 nm. At this wavelength, UV damage is significantly reduced. Therefore, UV irradiation periods can be increased, potentially raising the yield even further.

Very recently, several MS based studies identified RNA binding proteins after *in vivo* cross-linking. Cells were grown in the presence of 4SU [14] or 4SU and 6SG [15]. After cross-linking and cell lysis, proteins covalently linked to polyadenylated RNA were purified under stringent conditions with oligo(dT). The RBPs were finally identified by mass spectrometry. The study published by the Hentze laboratory (EMBL Heidelberg) compared cross-linking of unlabeled and 4SU-labeled cells. The overlap of identified RBPs between both approaches was approximately two thirds [14]. This implies that the different strategies yield complementary results. A later study identified yeast RBPs following the same complex isolation workflow but without incorporation of photoreactive nucleotides [16]. These surveys demonstrate that mass spectrometry can be applied to identify RNA binding proteins after *in vivo* cross-linking. The technical differences between identification of entire RBPs and cross-linking sites on a peptide or amino acid level will be described in more detail below.

After incorporation of 4SU and cDNA sequencing, thymidine to cytidine transitions were frequently observed [54]. The observations that the sulfur is lost upon cross-linking (see 3.1.3 and Figure 3.8) provides a possible explanation: A hydrogen bond acceptor site is missing in the reaction product, altering base pairing properties. If the cross-linked amino acid residue is covalently attached to the 4 position on the pyrimidine through an hydrogen bond donor, e.g. the amine of lysine, the resulting product resembles cytosine.

4.1.2 Reactivity of amino acids

The reactivity of amino acids in UV induced cross-linking also differs. Interestingly, the chemical functionalities do not seem to be a reliable indicator for cross-linking reactivity. Functional groups might be assumed to increase reactivity, in consequence aliphatic residues would be expected to be rather unreactive. In contrast, one of the most prominent cross-links in the human U1 small ribonucleoprotein particle is L175 of the U1 70K protein (first reported in [65]).

Combining the cross-linking results of the model complex for *ASH1*-mRNA transport, the spliceosomal protein Cwc2 and the yeast RNA binding proteins, 81 unique protein regions were found cross-linked. In 50 of these regions, the cross-linked amino acid could be identified, i.e. in more than 60%. About half of the cross-links (27) were formed via cysteine residues. The other residues identified as cross-linked were phenylalanine (6x), tyrosine (5x), lysine (4x), histidine (3x), tryptophane (3x), threonine (1x), and isoleucine (1x).

The high number of observed cross-links involving cysteine is very interesting. The identity of the other residues indicates a high reactivity in UV induced cross-linking for all aromatic residues.

Importantly, the majority of cross-links to cysteine residues involved the 152 Da mass adduct whose origin and composition is still unclear (see also 1.3.4). For each cross-link, an experimental value for this mass adduct can be determined by subtracting the calculated peptide and (oligo)nucleotide masses from the experimental precursor mass. Based on 38 cross-links identified in yeast, the average mass of this adduct was calculated to be 151.9938 Da. The fractional mass proves that the species must have a high content of atoms with a mass deficiency, e.g. oxygen or phosphorus (see 1.3.4). After fragmentation, the adduct was found to shift the mass of peptide fragment ions together with uracil (e.g. spectrum shown in Figure B.24). This indicates that the species links the amino acid residue and the nucleic acid base. Despite tremendous efforts, the identity of the species producing this mass shift could not be determined (F. Richter, C. Endler, K.K., U. Zaman, H. Urlaub, Bioanalytical Mass Spectrometry group, unpublished).

Without this knowledge, no final conclusions about the reactivity of cysteines and the structural interpretation of the resulting cross-links can be drawn. The species could correspond to a nucleic acid base derivative and result from formation of two covalent bonds, i.e. between two bases as well as a base and a residue. In this case, the cross-link would still be highly specific. Alternatively, the adduct could originate from incorporation of a small molecule. This would mean that the molecule could bridge a distance between cysteine and base much greater than one covalent bond. In addition, the reaction could take place on a slower timescale. Increased distance and lower reaction speed would decrease specificity.

In some cases, both cross-links with and without the 152 Da adduct were observed for the same cysteine. For example, cross-links of cysteine residues in Cwc2 (C87 and C181) were found both with and without the additional mass of 152 Da (see Table A.4). This suggests that at least some if not all cross-links with this adduct exhibit the same specificity as regular cross-links. Several efforts to clarify the origin and nature of the adduct are presently taken in our laboratory and will hopefully enable full interpretation of the obtained results in the near future.

It would be desirable to extend the knowledge about relative amino acid reactivity based on a higher number of cross-links in a variety of biological systems. In addition, the type of interactions between the nucleic acid base and the amino acid residue might affect cross-link formation and reactivity (see below).

4.1.3 Influence of the three-dimensional structure

Studies on the reactivity of nucleotides and amino acids as well as the reaction mechanism are based on simplified systems, e.g. polynucleotides and single amino acids or even bases with small molecules resembling amino acid residues (e.g. [35] and references therein). In protein–RNA complexes, many interactions might have an additional influence on both reactivity and mechanism. Systematic experimental investigation in these complex environments is very challenging. This is mainly due to the low cross-linking yield and the fact that often samples are only available in limited amounts.

It is generally accepted that UV induced cross-linking is highly specific and occurs between nucleotides and amino acid residues that are involved in direct interactions, i.e. which are in close spatial proximity. However, nothing is known whether the type of interaction has an influence. For example, aromatic residues can stack with or in between nucleic acid bases. Alternatively, hydrogen bonds are formed, e.g. involving amine, carbonyl and alcohol functionalities. It is possible that the type and strength of interaction influences the mechanism and consequently or independently the yield of the cross-linking reaction.

In addition, the involvement of a nucleotide or amino acid residue in intramolecular interactions might influence cross-link formation. For example, nucleotides within double-stranded stem loops are involved in both hydrogen bonds and stacking interactions with opposite and neighboring bases, respectively. Consequently, proteins bind to double-stranded DNA or RNA mostly without any sequence specificity as there is no space for corresponding interactions. Protein secondary structure might also prevent or hinder cross-link formation, and most cross-linked residues are found within flexible loop regions [49].

The relation between UV cross-linking and the type of interactions could be investigated in cross-linking experiments of protein–RNA complexes with available three-dimensional structures, ideally in various complexes from different origins.

4.1.4 Identification of the cross-linking site on the RNA level

In MS/MS fragment spectra of cross-linked heteroconjugates, longer RNA moieties lead to increasing suppression of peptide sequence ions while mostly dominating RNA fragments are observed (see for example Figure 3.14). Therefore, thorough RNA hydrolysis is part of the sample preparation for MS analysis. Consequently, cross-linked RNA typically contains one to three nucleotides. In rare cases, cross-links to four nucleotides can be identified. However, the corresponding spectra typically show such a low number of peptide fragments that the cross-link can only be validated in simple systems or by comparison to cross-links of the same peptide to shorter RNA (data not shown).

In most cases, only the RNA composition but not the order of nucleotides can be derived from the MS data. Only if the RNA employed in the experiment is short and a single stretch within corresponds to the oligonucleotide composition determined by MS, the cross-linking site can be determined unambiguously. The cross-linked nucleotide can be derived if an RNA is labeled with photoreactive nucleotides site-specifically and UV irradiation is performed at a wavelength that excludes cross-linking of the native nucleotides. An example for both cases is cross-linking of the NusB–S10 complex to short, BoxA containing nucleotides. In the first study, the cross-linked RNA region could be determined in several cases where longer oligonucleotides were identified [71]. In the experiments described in 3.1, only the stretch of three 4SU nucleotides was excited by irradiation at 365 nm and could form cross-links.

The cross-linking site on the RNA can usually only be deduced by complementary experiments such as primer extension. Longer RNA stretches have been sequenced by MALDI-MS after preparative isolation of peptide–oligonucleotide heteroconjugates and limited alkaline hydrolysis. However, the available sample amounts are often not sufficient for purification on a preparative scale. The recent advances in sensitivity and sequencing speed of MS instruments might enable a similar approach without extensive isolation workflows. For example, full-length RNA with cross-linked peptides could be separated by size exclusion chromatography as described. The sample could then be split for complete and limited RNA hydrolysis, e.g. treating one half with both RNases A and T1 and the other only with RNase T1. Cross-linked peptides could be identified after MS analysis of the first sample. The second sample could then be checked for cross-links of the same peptide to longer oligonucleotides, i.e. by adding the masses of nucleotides to the mass of the cross-link identified in the first sample.

This strategy should first be tested and might be limited to small complexes. The sample complexity increases with the size and number of proteins and RNAs. It might reach a limit above which validation of cross-links is impossible. An increasing oligonucleotide length considerably decreases the number and intensity of observed peptide fragments. If only a small number of low intensity peptide fragments is observed, the derived information might not be sufficient for unambiguous identification of the peptide in a complex system. Therefore, combination of a similar approach with *in vivo* cross-linking does not seem feasible at this point. In this context, the combination of DNA sequencing methods and mass spectrometry for the identification of the cross-linking site on the RNA and protein, respectively, might be pursued.

4.1.5 RNA-binding metabolic enzymes: Rossmann fold domains as RNA-binding motifs

Several metabolic enzymes have been found cross-linked to RNA by UV induced cross-linking with mass spectrometry (see 3.4.4.3). This illustrates the advantage of the method as an unbiased approach capable of identifying novel RNA-binding domains. Here, the example of RNA-binding metabolic enzymes in general and the Rossmann fold in particular will be shortly described.

RNA-binding properties have been reported for several metabolic enzymes. However, the relationship between metabolic and RNA-related function is mostly unclear. The multifunctionality of metabolic enzymes might result from changes in the oligomeric state. For example, GAPDH functions as a tetramer in glycolysis and as a monomer in transcriptional regulation. Changes in post-translational modifications and molecular interactions might also induce different protein functions. In addition, an unexpected localization of several glycolytic enzymes in the nucleus has been reported and further suggests a role in transcription or DNA replication (see [125] and references therein).

RNA-binding capabilities have been reported for several NAD$^+$ dependent dehydrogenases with (di)nucleotide-binding Rossmann fold domains. We found two such dehyrogenases, alcohol dehydrogenase (ADH) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH), cross-linked to RNA. RNA-binding in GAPDH is competitive with NAD-binding, and increasing RNA concentration decrease the enzyme's activity (see [120] and references therein). The same study also provides evidence for RNA-binding abilities of ADH. Recently, Barbas *et al.* investigated RNA binding of the UDP-glucose dehydrogenase (UgdG) from *Sphingomonas elodea* [126]. They demonstrate that RNA and NAD$^+$ binding regions do not overlap and that UgdG exhibits ribonuclease activity. However, this enzyme contains a second, catalytically inactive Rossmann fold, which could be responsible for RNA-binding. Another Rossmann fold containg enzyme, adenosylhomocysteinase (Sah1p), was found cross-linked to RNA in our study. Therefore, our results support the suggestion to add Rossmann fold containing metabolic enzymes to the list of RNA-binding proteins and the Rossmann fold to RNA-binding domains [120].

In general, *in vivo* cross-linking and identification of RNA-binding sites on a peptide or amino acid level would enable the unbiased identification of RNA-binding domains. This would add valuable knowledge about protein–RNA interactions, as many RNA-binding proteins do not contain classical RNA-binding motifs. This adds a valuable application to the approach of UV induced cross-linking with mass spectrometric analysis.

4.2 Mass spectrometry and data analysis

4.2.1 Instrumentation

The performance of UV induced cross-linking with mass spectrometric analysis has been tightly connected to improvements on the instrumental side. The low cross-linking yield presents the major challenge in collecting MS information of a cross-linked sample. Sensitivity, sequencing speed and resolution/mass accuracy have a great influence on MS analysis of cross-linking experiments:

(1) Instrument sensitivity, especially in MS/MS acquisition, is very important. Failure to detect cross-links or to record MS/MS spectra of sufficient quality prevents cross-link identification.

(2) Higher sequencing speed of instruments enables the collection of fragment information on more precursors. Therefore, sequencing of a low abundant cross-link in data-dependent acquisition is more likely.

(3) The resolution and consequently mass accuracy of the instrument influences cross-link identification significantly. For example, the mass accuracy of the precursor (intact cross-link) determines the search space for peptide identification in the precursor variant approach. Higher mass deviation leads to evaluation of more candidates which can increase the number of false positive results.

In this work, cross-link identification after analysis on a Q-ToF Ultima or a LTQ Orbitrap Velos was compared. Due to the rapid technological advances, the Q-ToF Ultima has to be considered an outdated instrument although it is only ten years old. Therefore, it was not surprising that it was outperformed by the Orbitrap Velos. Several vendors offer Q-ToF instruments that can compete with orbitraps in terms of sensitivity, resolution and sequencing speed. It will be interesting to compare modern Q-ToF and orbitrap instruments for analysis of cross-linking experiments. The influence of future instrumental advances on bioanalytics in general and more dedicated applications as UV induced cross-linking will be interesting to follow.

Further insight into the cross-linked amino acid residues in cases where no RNA-adduct was observed would be desirable. After a fast data analysis workflow has been established, this might be addressed in more detail with mass spectrometry. Investigation of fragmentation conditions on observability of adducts could be approached, e.g. by varying the collision energy. In addition, fragmentation patterns of the unmodified and the cross-linked peptide could be compared in detail. In some cases, expected peptide fragments were not observed or of very low intensity if they contained a potentially cross-linked amino acid. For example, the a2-/b2-ion pair is typically observed at medium to high intensity after beam-type CID of peptides or cross-linked heteroconjugates. Absence or low intensity of this ion pair sometimes correlated with an amino acid reactive in cross-linking at the corresponding positions in the peptide. However, these observations have to be investigated in more detail and subsequently confirmed biochemically before they could be included as criteria for identification of the cross-linked amino acid.

Alternative fragmentation techniques such as electron capture dissociation (ECD [127]) and electron transfer dissociation (ETD [128]) could be evaluated for the identification of cross-linked heteroconjugates. Both methods are based on activation of the precursor ion by generation of an odd-electron species. The predominant fragments are c- and z-ions that result from cleavage of the amino alkyl (N-Cα) bond. In comparison to CID, ECD and ETD are more likely to leave labile peptide modifications intact. For example, the loss of phosphoric acid is frequently observed after CID fragmentation of phosphopeptides and presents a challenge for the identification of the exact phosphorylation site. In contrast, the phosphate group is mostly retained during ECD and ETD fragmentation. Therefore, ECD and ETD have been demonstrated to yield valuable information complementary to that of CID in phosphopeptide analysis (see [129] and references therein).

Both ECD and ETD were demonstrated to increase the peptide fragment information in MS/MS spectra of a model peptide–oligonucleotide heteroconjugate compared to ion trap CID [73]. However, the model heteroconjugate used in this study contained five arginine residues on the 14 amino acid long peptide moiety. In addition, the synthetic $(CH_2)_6$ link between peptide and oligonucleotide connected the carboxyl group of an aspartic acid residue to the 5' phosphate of the RNA. Neither properties of the peptide and the linking bond are well comparable to heteroconjugates originating from UV induced protein–RNA cross-linking experiments. Peptides obtained after hydrolysis with the endoproteinase trypsin are unlikely to contain five basic residues as trypsin cleaves C-terminal to lysine and arginine. The cross-linking bond formed as a consequence of UV irradiation connects the nucleic acid base and not the phosphate backbone to the amino acid residue. Therefore, the comparison of fragmentation techniques could be repeated with more appropriate model molecules or heteroconjugates derived from irradiated complexes to obtain more meaningful results.

4.2.2 Development and feasibility of the precursor variant approach

The novel workflow for automated identification of cross-linked heteroconjugates from mass spectrometry data is based on the subtraction of calculated RNA masses from the experimental precursor mass. Therefore, it was termed *precursor variant approach*. The basic idea arose during manual assignment of fragment spectra in the beginning of this project. Practical realization of the approach was initially accomplished in close collaboration with Petra Hummel (IT & Eletronics Service, MPI for Biophysical Chemistry). The approach was proven feasible in cross-linking studies of the NusB–S10 complex to 4-thio-uracil substituted RNA (Section 3.1) as well as a model complex for *ASH1* mRNA transport in budding yeast (Section 3.2).

Further automatization was achieved by integration of the approach into a novel tool in the OpenMS environment. The necessary bioinformatic knowledge for programming was provided by our collaborators, the Applied Bioinformatics Group (Prof. Oliver Kohlbacher, Universität Tübingen). Reinvestigation of the ASH1 complex (Section 3.2), comparative cross-linking of the spliceosomal protein Cwc2 to U6 and U4 snRNA (Section 3.3) as well as the identification of a large number of RNA binding proteins after isolation of protein–RNA complexes by TAP tag purification of the cap-binding protein Cbp20 (Section 3.4) proved feasibility and improvements of the approach.

This recapitulation of development and application of the precursor variant approach illustrates that testing and optimization of data analysis workflows was a constant process during most of the presented project. This included extensive testing of different developmental versions of the algorithms in application to actual cross-linking data. The collaborators were responsible for programming the necessary algorithms in the respective environments (perl and C++) and provided helpful suggestions. However, the major concepts from the basic idea to the framework for functionalities and parameters were solely developed as part of this project.

The precursor variant approach was developed based on the assumption that fragment spectra of cross-links exhibit great similarity to spectra of noncross-linked peptides. For the majority of cross-links, the corresponding unmodified peptide is not observed within the same measurement since it was separated during titanium dioxide enrichment or size exclusion chromatography. Therefore, the assumption could only be tested systematically by either measuring the sample prior to enrichment or by fragmenting synthetic peptides. However, comparison between fragmentation of the cross-linked and the unmodified peptide was possible in a few cases, one example is shown in Figure 4.1. Here, the fragment spectra of the unmodified and cross-linked peptide exhibit only minor differences.

Shifts of peptide sequence ions by covalently linked RNA (fragments) increase the differences between spectra. More importantly, the corresponding signals are not recognized by the database search engine. Therefore, the scores given to these spectra underestimate the agreement between the cross-linked peptide and the spectrum. Interestingly, the cross-linked peptides are nonetheless identified with reasonable scores, provided the quality of the spectrum is high.

Frequently observed shifts could be defined as post-translational modifications (PTMs) for a standard database search. For example, shifts of uridine (fragments) with the 152 Da adduct could be defined as PTMs of cysteines. In Mascot searches, cross-links to [U $-H_2O$] or [4SU $-H_2S$] producing an extensive shift of 94 Da due to the corresponding base remaining on the peptide fragment have been identified successfully after definition of a corresponding modification. Initial tests have shown that shifts can be integrated as PTMs in OMSSA searches and annotated in TOPPView (data not shown). This approach requires further testing and optimization. Especially the integration into the data analysis workflow has to be optimized. Parallel searches with various parameters produce different results, the scores are not necessarily comparable. Therefore, this step has to be integrated carefully to avoid bias as well as false positive or false negative results.

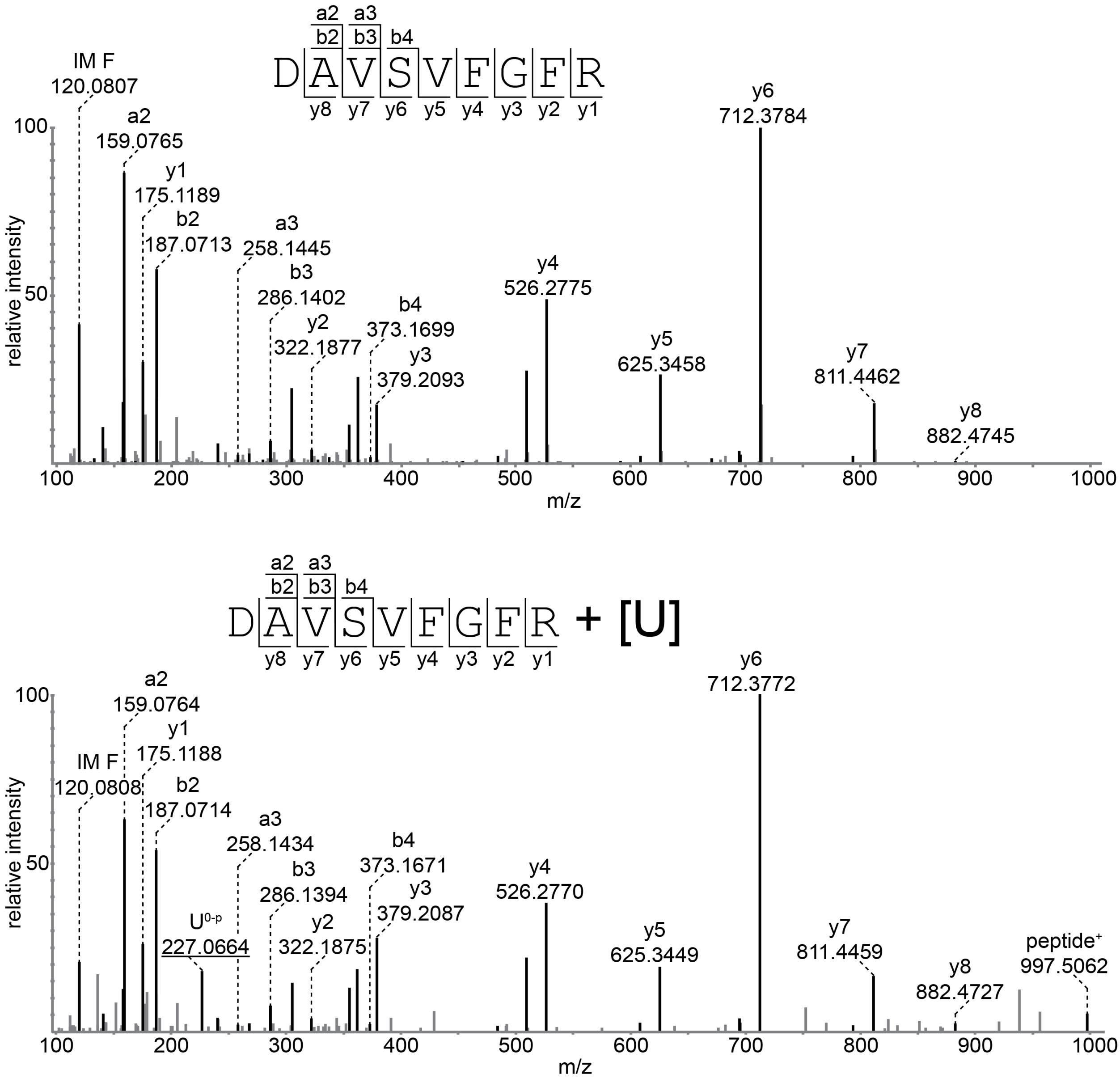


Figure 4.1: Comparison of MS/MS fragment spectra of the unmodified S24-A/-B peptide DAVSVFGFR (D53–R61, upper panel) and the same peptide cross-linked to uridine (lower panel). Both spectra are strikingly similar. The number of observed peptide fragments is exactly the same and relative fragment intensities do not show significant differences except for the phenylalanine immonium ion (drop of 40% to 20% relative intensity in the fragment spectrum of the cross-link). In the spectrum of the cross-linked peptide, additional signals corresponding to the RNA fragment [U $-H_3PO_4$] as well as the intact peptide are observed.

The precursor variant approach is based on subtracting masses of anticipated nucleotide combinations. Consequently, cross-links with mass adducts that were not expected cannot be identified. For example, cross-links of the cap binding protein Cbp20 to the cap structure could not have been identified because methylated guanine was not defined as a nucleotide for precursor variant generation. The observation of this cross-link seemed highly unlikely. First, guanine is rather unreactive in UV induced cross-linking. Secondly, the three phosphate groups connecting the 5' position of the 7-methyl-guanine to the 5' position of the next nucleotide would considerably hinder ionization in positive ion mode. However, the structure and consequently mass of 7-methyl-guanine is known and could have been integrated if desired.

Unexpected variations from the masses of the common cross-links cannot be identified automatically. These can result from unanticipated modifications of the protein and RNA as well as loss (or gain) of atoms or molecules during the cross-linking reaction. Novel cross-linking products reported in this thesis were identified manually, e.g. the cross-linking product of uracil and 4-thio-uracil which originates from loss of oxygen or sulfur, respectively, from the 4 position of the base (see 3.1.3.2 and 3.2.5.2). However, the additional filters developed to rule out spectra of pure peptides and spectra of species appearing in the non-irradiated control (see 3.4.3.3) would be extremely useful for further manual identification of novel modifications or cross-linking products. High quality spectra remaining after filtering and identification of anticipated cross-links are good candidates for a manual search. While this strategy has not been applied, it will be considered for future experiments.

The developed data analysis strategy represents one key development that finally enables unbiased identification of cross-linked peptides after UV irradiation *in vivo*.

4.3 Mass spectrometry and *in vivo* cross-linking

One of the major limitations of UV induced protein–RNA cross-linking in combination with mass spectrometric analysis is the required sample amounts. For example, immunodetection of cross-linked proteins exhibits a considerably higher sensitivity. The same applies to cross-linking experiments with radiolabeled RNA. Both methods therefore require lower sample amounts. Methods based on reverse transcription and sequencing of the resulting DNA have the advantage that the cross-linking signal is enhanced. Consequently, several protocols have been established for *in vivo* cross-linking and the analysis of contact sites on the RNA level by DNA sequencing. These can be combined with incorporation of photoreactive nucleotides, as it is done in the approach termed Photoactivatable-Ribonucleoside-Enhanced Cross-Linking (PAR-CLIP) [54].

Nonetheless, several studies published in the past 12 months demonstrate the ability of mass spectrometry to identify several hundred RNA-binding proteins after UV cross-linking (see also above, [14–16]). The common experimental approach is based on isolation of proteins cross-linked to polyadenylated RNA by hybridization with oligo(dT). This procedure permits stringent purification conditions, i.e. interruption of all non-covalent protein–RNA interactions and selective isolation of proteins cross-linked to RNA. After hydrolyzation of both RNA and proteins, the latter were identified by mass spectrometry.

There are several crucial differences between the identification of the cross-linking site on a peptide or amino acid level and this approach. The cross-linked protein is identified by unmodified peptides.

Therefore, standard proteomics data analysis algorithms can be utilized, avoiding the challenges involved in identification of peptides covalently linked to RNA. In addition, disadvantages due to less efficient ionization of cross-linked heteroconjugates are circumvented. On the contrary, fragmentation of one unique peptide might yield sufficient information for unambiguous identification of the cross-linked protein. Therefore, a low abundant protein might be detectable by mass spectrometry if it comprises a peptide with properties advantageous for ionization and fragmentation, e.g. containing a few basic residues that are easily protonated and having a length between 10 and 20 amino acids. Consequently, the required sample amounts are expected to be lower compared to the approach described in this thesis. However, the obtained information is limited to the protein level and these approaches should be considered as complementary to the method described in this work.

As a follow-up of the presented work, a collaboration was started between the Hentze laboratory (EMBL Heidelberg) and our group. Dr. Benedikt Beckmann performed *in vivo* cross-linking of 4SU labeled yeast cells and isolated polyadenylated RNA with cross-linked proteins by oligo(dT). Enrichment with C18 and titanium dioxide chromatography as well as mass spectrometric and data analysis were performed in our laboratory. Since the results are very preliminary, they were not included in this work, but they indicate that it is indeed possible to identify peptides and amino acids interacting with RNA after cross-linking *in vivo*. Optimization of this and other *in vivo* approaches will be important future projects of the laboratory.

Other members of the Bioanalytical Mass Spectrometry Group have been conducting experiments with proteome-wide searches for cross-linked peptides enabled by the data analysis approach reported here. Importantly, Saadia Qamar demonstrated that cross-links can be identified in searches against the human proteome. The UniProt database for the human proteome is roughly ten times larger than that of *S. cerevisiae* which was used in the presented work. Dr. Uzma Zaman carried out experiments with yeast cells grown in the presence of 4SU and proved that the data analysis approach is also feasible for RNA labeled with photoreactive nucleotides. The data analysis approach, the cross-links obtained for yeast RNA binding proteins without 4SU labeling which were presented in this work, and the two surveys mentioned above are currently summarized for publication.

Bibliography

[1] Maris, C., C. Dominguez and F.H. Allain: *The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression.* FEBS J, 272(9):2118–2131, May 2005.

[2] Cléry, A., M. Blatter and F.H.-T. Allain: *RNA recognition motifs: boring? Not quite.* Curr Opin Struct Biol, 18(3):290–298, Jun 2008.

[3] Valverde, R., L. Edwards and L. Regan: *Structure and function of KH domains.* FEBS J, 275(11):2712–2726, Jun 2008.

[4] Quintal, S.M., Q. Antonia dePaula and N.P. Farrell: *Zinc finger proteins as templates for metal ion exchange and ligand reactivity. Chemical and biological consequences.* Metallomics, 3(2):121–139, Feb 2011.

[5] Ding, J., M.K. Hayashi, Y. Zhang, L. Manche, A.R. Krainer and R.M. Xu: *Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA.* Genes Dev, 13(9):1102–1115, May 1999.

[6] Braddock, D.T., J.L. Baber, D. Levens and G.M. Clore: *Molecular basis of sequence-specific single-stranded DNA recognition by KH domains: solution structure of a complex between hnRNP K KH3 and single-stranded DNA.* EMBO J, 21(13):3476–3485, Jul 2002.

[7] Lu, D., M.A. Searles and A. Klug: *Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition.* Nature, 426(6962):96–100, Nov 2003.

[8] Calero, G., K.F. Wilson, T. Ly, J.L. Rios-Steiner, J.C. Clardy and R.A. Cerione: *Structural basis of m7GpppG binding to the nuclear cap-binding protein complex.* Nat Struct Biol, 9(12):912–917, Dec 2002.

[9] Mazza, C., A. Segref, I.W. Mattaj and S. Cusack: *Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex.* EMBO J, 21(20):5548–5557, Oct 2002.

[10] Price, S.R., P.R. Evans and K. Nagai: *Crystal structure of the spliceosomal U2B"-U2A' protein complex bound to a fragment of U2 small nuclear RNA.* Nature, 394(6694):645–650, Aug 1998.

[11] Matunis, M.J., W.M. Michael and G. Dreyfuss: *Characterization and primary structure of the poly(C)-binding heterogeneous nuclear ribonucleoprotein complex K protein.* Mol Cell Biol, 12(1):164–171, Jan 1992.

[12] Siomi, H., M.J. Matunis, W. . Michael and G. Dreyfuss: *The pre-mRNA binding K protein contains a novel evolutionarily conserved motif.* Nucleic Acids Res, 21(5):1193–1198, Mar 1993.

[13] Ankö, M. and K.M. Neugebauer: *RNA-protein interactions in vivo: global gets specific.* Trends Biochem Sci, 37(7):255–262, Jul 2012.

[14] Castello, A., B. Fischer, K. Eichelbaum, R. Horos, B.M. Beckmann, C. Strein, N.E. Davey, D.T. Humphreys, T. Preiss, L.M. Steinmetz, J. Krijgsveld and M.W. Hentze: *Insights into RNA biology from an atlas of mammalian mRNA-binding proteins.* Cell, 149(6):1393–1406, Jun 2012.

[15] Baltz, A.G., M. Munschauer, B. Schwanhäusser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich and M. Landthaler: *The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts.* Mol Cell, 46(5):674–690, Jun 2012.

[16] Mitchell, S.F., S. Jain, M. She and R. Parker: *Global analysis of yeast mRNPs.* Nat Struct Mol Biol, 20(1):127–133, Jan 2013.

[17] Fenn, J.B., M. Mann, C.K. Meng, S.F. Wong and C.M. Whitehouse: *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 246(4926):64–71, Oct 1989.

[18] Fenn, J.B., M. Mann, C.K.Meng, S.F.Wong and C.M.Whitehouse: *Electrospray ionization: principles and practice.* Mass Spectrom Rev, 9(1):37–70, Jan 1990.

[19] Tanaka, K., H. Waki, Y. Ido, S. Akita, Y. Yoshida and T. Yoshida: *Protein and polymer analysis up to m/z 100,000 by laser ionization time-of-flight mass spectrometry.* Rapid Commun Mass Spectrom, 2(8):151–153, Aug 1988.

[20] *The Nobel Prize in Chemistry 2002.* (http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2002/; accessed 8.4.2013).

[21] Karas, M. and F. Hillenkamp: *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Anal Chem, 60(20):2299–2301, Oct 1988.

[22] Dass, C.: *Fundamentals of contemporary mass spectrometry.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2007.

[23] Mallick, P. and B. Küster: *Proteomics: a pragmatic perspective.* Nat Biotechnol, 28(7):695–709, Jul 2010.

[24] Altelaar, A.F.M., J. Munoz and A.J.R. Heck: *Next-generation proteomics: towards an integrative view of proteome dynamics.* Nat Rev Genet, 14(1):35–48, Jan 2013.

[25] Thermo Fisher Scientific: *LTQ Orbitrap Velos Biotech Operations (Training Course Manual).*

[26] Roepstorff, P. and J. Fohlman: *Proposal for a common nomenclature for sequence ions in mass spectra of peptides.* Biomed Mass Spectrom, 11(11):601, Nov 1984.

[27] Biemann, K.: *Contributions of mass spectrometry to peptide and protein structure.* Biomed Environ Mass Spectrom, 16(1-12):99–111, Oct 1988.

[28] Wysocki, V.H., G. Tsaprailis, L.L. Smith and L.A. Breci: *Mobile and localized protons: a framework for understanding peptide dissociation.* J Mass Spectrom, 35(12):1399–1406, Dec 2000.

[29] Paizs, B. and S. Suhai: *Fragmentation pathways of protonated peptides.* Mass Spectrom Rev, 24(4):508–548, 2005.

[30] McLuckey, S.A., G.J. Van Berkel and G.L. Glish: *Tandem mass spectrometry of small, multiply charged oligonucleotides.* J Am Soc Mass Spectrom, 3:60–70, 1992.

[31] Wu, J. and S.A. McLuckey: *Gas-phase fragmentation of oligonucleotide ions.* Int J Mass Spectrom, 237:197–241, 2004.

[32] Nesvizhskii, A.I.: *A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics.* J Proteomics, 73(11):2092–2123, Oct 2010.

[33] Perkins, D.N., D.J. Pappin, D.M. Creasy and J.S. Cottrell: *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 20(18):3551–3567, Dec 1999.

[34] Geer, L.Y., S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, W. Shi and S.H. Bryant: *Open mass spectrometry search algorithm.* J Proteome Res, 3(5):958–964, 2004.

[35] Meisenheimer, K.M. and T.H. Koch: *Photocross-linking of nucleic acids to associated proteins.* Crit Rev Biochem Mol Biol, 32(2):101–140, 1997.

[36] Urlaub, H., V. Kruft, O. Bischof, E.C. Müller and B. Wittmann-Liebold: *Protein-rRNA binding features and their structural and functional implications in ribosomes as determined by cross-linking studies.* EMBO J, 14(18):4578–4588, Sep 1995.

[37] Shetlar, M.D., J. Christensen and K. Hom: *Photochemical addition of amino acids and peptides to DNA.* Photochem Photobiol, 39(2):125–133, Feb 1984.

[38] Shetlar, M.D., J. Carbone, E. Steady and K. Hom: *Photochemical addition of amino acids and peptides to polyuridylic acid.* Photochem Photobiol, 39(2):141–144, Feb 1984.

[39] Shetlar, M.D., K. Home, J. Carbone, D. Moy, E. Steady and M. Watanabe: *Photochemical addition of amino acids and peptides to homopolyribonucleotides of the major DNA bases.* Photochem Photobiol, 39(2):135–140, Feb 1984.

[40] Cadet, J. and P. Vigny: *Bioorganic photochemistry.*, chapter The photochemistry of nucleic acids., pages 1–272. Wiley Interscience Publication, New York, 1990.

[41] Urlaub, H., V.A. Raker, S. Kostka and R. Lührmann: *Sm protein-Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure.* EMBO J, 20(1-2):187–196, Jan 2001.

[42] Expert-Bezançon, A., A. Sureau, P. Durosay, R. Salesse, H. Groeneveld, J.P. Lecaer and J. Marie: *hnRNP A1 and the SR proteins ASF/SF2 and SC35 have antagonistic functions in splicing of beta-tropomyosin exon 6B.* J Biol Chem, 279(37):38249–38259, Sep 2004.

[43] Möller, K. and R. Brimacombe: *Specific cross-linking of proteins S7 and L4 to ribosomal RNA, by UV irradiation of Escherichia coli ribosomal subunits.* Mol Gen Genet, 141(4):343–355, Dec 1975.

[44] Bonnal, S., F. Pileur, C. Orsini, F. Parker, F. Pujol, A. Prats and St. Vagner: *Heterogeneous nuclear ribonucleoprotein A1 is a novel internal ribosome entry site trans-acting factor that modulates alternative initiation of translation of the fibroblast growth factor 2 mRNA.* J Biol Chem, 280(6):4144–4153, Feb 2005.

[45] Urlaub, H., K. Hartmuth and R. Lührmann: *A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles.* Methods, 26(2):170–181, Feb 2002.

[46] Urlaub, H., B. Thiede, E. C. Müller, R. Brimacombe and B. Wittmann-Liebold: *Identification and sequence analysis of contact sites between ribosomal proteins and rRNA in Escherichia coli 30 S subunits by a new approach using matrix-assisted laser desorption/ionization-mass spectrometry combined with N-terminal microsequencing.* J Biol Chem, 272(23):14547–14555, Jun 1997.

[47] Nottrott, S., H. Urlaub and R. Lührmann: *Hierarchical, clustered protein interactions with U4/U6 snRNA: a biochemical role for U4/U6 proteins.* EMBO J, 21(20):5527–5538, Oct 2002.

[48] Schmidt, C., K. Kramer and H. Urlaub: *Investigation of protein-RNA interactions by mass spectrometry–Techniques and applications.* J Proteomics, 75(12):3478–3494, Jun 2012.

[49] Urlaub, H., E. Kühn-Hölsken and R. Lührmann: *Analyzing RNA-protein crosslinking sites in unlabeled ribonucleoprotein complexes by mass spectrometry.* Methods Mol Biol, 488:221–245, 2008.

[50] Botelho, D., M.J. Wall, D.B. Vieira, S. Fitzsimmons, F. Liu and A. Doucette: *Top-down and bottom-up proteomics of SDS-containing solutions following mass-based separation.* J Proteome Res, 9(6):2863–2870, Jun 2010.

[51] Fecko, C.J., K.M. Munson, A. Saunders, G. Sun, T.P. Begley, J.T. Lis and W.W. Webb: *Comparison of femtosecond laser and continuous wave UV sources for protein-nucleic acid crosslinking.* Photochem Photobiol, 83(6):1394–1404, 2007.

[52] Steen, H. and O.N. Jensen: *Analysis of protein-nucleic acid interactions by photochemical cross-linking and mass spectrometry.* Mass Spectrom Rev, 21(3):163–182, 2002.

[53] Meisenheimer, K.M., P.L. Meisenheimer and T.H. Koch: *Nucleoprotein photo-cross-linking using halopyrimidine-substituted RNAs.* Methods Enzymol, 318:88–104, 2000.

[54] Hafner, M., M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A. Jungkamp, M. Munschauer, A. Ulrich, G.S. Wardle, S. Dewell, M. Zavolan and T. Tuschl: *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.* Cell, 141(1):129–141, Apr 2010.

[55] Dix, I., C.S. Russell, R.T. O'Keefe, A.J. Newman and J.D. Beggs: *Protein-RNA interactions in the U5 snRNP of Saccharomyces cerevisiae.* RNA, 4(12):1675–1686, Dec 1998.

[56] Golden, M.C., K.A. Resing, B.D. Collins, M.C. Willis and T.H. Koch: *Mass spectral characterization of a protein-nucleic acid photocrosslink.* Protein Sci, 8(12):2806–2812, Dec 1999.

[57] Chepanoske, C.L., O.A. Lukianova, M. Lombard, M. Golinelli-Cohen and S.S. David: *A residue in MutY important for catalysis identified by photocross-linking and mass spectrometry.* Biochemistry, 43(3):651–662, Jan 2004.

[58] Ule, J., K.B. Jensen, M. Ruggiu, A. Mele, A. Ule and R.B. Darnell: *CLIP identifies Nova-regulated RNA networks in the brain.* Science, 302(5648):1212–1215, Nov 2003.

[59] Jensen, K.B. and R.B. Darnell: *CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins.* Methods Mol Biol, 488:85–98, 2008.

[60] Granneman, S., G. Kudla, E. Petfalski and D. Tollervey: *Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs.* Proc Natl Acad Sci U S A, 106(24):9613–9618, Jun 2009.

[61] Granneman, S., E. Petfalski, A. Swiatkowska and D. Tollervey: *Cracking pre-40S ribosomal subunit structure by systematic analyses of RNA-protein cross-linking.* EMBO J, 29(12):2026–2036, Jun 2010.

[62] Kühn-Hölsken, E., O. Dybkov, B. Sander, R. Lührmann and H. Urlaub: *Improved identification of enriched peptide RNA cross-links from ribonucleoprotein particles (RNPs) by mass spectrometry.* Nucleic Acids Res, 35(15):e95, 2007.

[63] Lenz, C., E. Kühn-Hölsken and H. Urlaub: *Detection of protein-RNA crosslinks by NanoLC-ESI-MS/MS using precursor ion scanning and multiple reaction monitoring (MRM) experiments.* J Am Soc Mass Spectrom, 18(5):869–881, May 2007.

[64] Urlaub, H., K. Hartmuth, S. Kostka, G. Grelle and R. Lührmann: *A general approach for identification of RNA-protein cross-linking sites within native human spliceosomal small nuclear ribonucleoproteins (snRNPs). Analysis of RNA-protein contacts in native U1 and U4/U6.U5 snRNPs.* J Biol Chem, 275(52):41458–41468, Dec 2000.

[65] Kühn-Hölsken, E., C. Lenz, B. Sander, R. Lührmann and H. Urlaub: *Complete MALDI-ToF MS analysis of cross-linked peptide-RNA oligonucleotides derived from nonlabeled UV-irradiated ribonucleoprotein particles.* RNA, 11(12):1915–1930, Dec 2005.

[66] Kühn-Hölsken, E., C. Lenz, A. Dickmanns, H. Hsiao, F.M. Richter, B. Kastner, R. Ficner and H. Urlaub: *Mapping the binding site of snurportin 1 on native U1 snRNP by cross-linking and mass spectrometry.* Nucleic Acids Res, 38(16):5581–5593, Sep 2010.

[67] Geyer, H., R. Geyer and V. Pingoud: *A novel strategy for the identification of protein-DNA contacts by photocrosslinking and mass spectrometry.* Nucleic Acids Res, 32(16):e132, 2004.

[68] Urlaub, H., B. Thiede, E.C. Müller and B. Wittmann-Liebold: *Contact sites of peptide-oligoribonucleotide cross-links identified by a combination of peptide and nucleotide sequencing with MALDI MS.* J Protein Chem, 16(5):375–383, Jul 1997.

[69] Steen, H., J. Petersen, M. Mann and O.N. Jensen: *Mass spectrometric analysis of a UV-cross-linked protein-DNA complex: tryptophans 54 and 88 of E. coli SSB cross-link to DNA.* Protein Sci, 10(10):1989–2001, Oct 2001.

[70] Larsen, M.R., T.E. Thingholm, O.N. Jensen, P. Roepstorff and T.J.D. Jørgensen: *Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns.* Mol Cell Proteomics, 4(7):873–886, Jul 2005.

[71] Luo, X., H. Hsiao, M. Bubunenko, G. Weber, D.L. Court, M.E. Gottesman, H. Urlaub and M.C. Wahl: *Structural and functional analysis of the E. coli NusB-S10 transcription antitermination complex.* Mol Cell, 32(6):791–802, Dec 2008.

[72] Richter, F.M., H. Hsiao, U. Plessmann and H. Urlaub: *Enrichment of protein-RNA crosslinks from crude UV-irradiated mixtures for MS analysis by on-line chromatography using titanium dioxide columns.* Biopolymers, 91(4):297–309, Apr 2009.

[73] Krivos, K.L. and P.A. Limbach: *Sequence analysis of peptide:oligonucleotide heteroconjugates by electron capture dissociation and electron transfer dissociation.* J Am Soc Mass Spectrom, 21(8):1387–1397, Aug 2010.

[74] Ghalei, H., H. Hsiao, H. Urlaub, M.C. Wahl and N.J. Watkins: *A novel Nop5-sRNA interaction that is required for efficient archaeal box C/D sRNP formation.* RNA, 16(12):2341–2348, Dec 2010.

[75] Petit, V. W, S. Zirah, S. Rebuffat and J. Tabet: *Collision induced dissociation-based characterization of nucleotide peptides: fragmentation patterns of microcin C7-C51, an antimicrobial peptide produced by Escherichia coli.* J Am Soc Mass Spectrom, 19(8):1187–1198, Aug 2008.

[76] Andersen, T.E., . Kirpekar and K.F. Haselmann: *RNA fragmentation in MALDI mass spectrometry studied by H/D-exchange: mechanisms of general applicability to nucleic acids.* J Am Soc Mass Spectrom, 17(10):1353–1368, Oct 2006.

[77] Jensen, O.N., S.Kulkarni, J.V. Aldrich and D.F. Barofsky: *Characterization of peptide-oligonucleotide heteroconjugates by mass spectrometry.* Nucleic Acids Res, 24(19):3866–3872, Oct 1996.

[78] Richter, F.M.: *Mass spectrometry based analysis of protein-RNA and protein-protein interactions in spliceosomal complexes.* PhD thesis, Humbold-Universität Berlin, 2009.

[79] Bley, C.J., X. Qi, D.P. Rand, C.R. Borges, R.W. Nelson and J.J.-L. Chen: *RNA-protein binding interface in the telomerase ribonucleoprotein.* Proc Natl Acad Sci U S A, 108(51):20333–20338, Dec 2011.

[80] Pourshahian, S. and P.A. Limbach: *Application of fractional mass for the identification of peptide-oligonucleotide cross-links by mass spectrometry.* J Mass Spectrom, 43(8):1081–1088, Aug 2008.

[81] Mozaffari-Jovin, S., K.F. Santos, H. Hsiao, C.L. Will, H. Urlaub, M.C. Wahl and R. Lührmann: *The Prp8 RNase H-like domain inhibits Brr2-mediated U4/U6 snRNA unwinding by blocking Brr2 loading onto the U4 snRNA.* Genes Dev, 26(21):2422–2434, Nov 2012.

[82] Shaw, A.A., A.M. Falick and M.D. Shetlar: *Photoreactions of thymine and thymidine with N-acetyltyrosine.* Biochemistry, 31(45):10976–10983, Nov 1992.

[83] Shivanna, B.D., M.R. Mejillano, T.D. Williams and R.H. Himes: *Exchangeable GTP binding site of beta-tubulin. Identification of cysteine 12 as the major site of cross-linking by direct photoaffinity labeling.* J Biol Chem, 268(1):127–132, Jan 1993.

[84] Jensen, O.N., D.F. Barofsky, M.C. Young, P.H. von Hippel, S.Swenson and S.E. Seifried: *Direct observation of UV-crosslinked protein-nucleic acid complexes by matrix-assisted laser desorption ionization mass spectrometry.* Rapid Commun Mass Spectrom, 7(6):496–501, Jun 1993.

[85] Bennett, S.E., O.N. Jensen, D.F. Barofsky and D.W. Mosbaugh: *UV-catalyzed cross-linking of Escherichia coli uracil-DNA glycosylase to DNA. Identification of amino acid residues in the single-stranded DNA binding site.* J Biol Chem, 269(34):21870–21879, Aug 1994.

[86] Farrow, M.A., F. Aboul-ela, D. Owen, A. Karpeisky, L. Beigelman and M.J. Gait: *Site-specific cross-linking of amino acids in the basic region of human immunodeficiency virus type 1 Tat peptide to chemically modified TAR RNA duplexes.* Biochemistry, 37(9):3096–3108, Mar 1998.

[87] Dybkov, O., C.L. Will, J. Deckert, N. Behzadnia, K. Hartmuth and R. Lührmann: *U2 snRNA-protein contacts in purified human 17S U2 snRNPs and in spliceosomal A and B complexes.* Mol Cell Biol, 26(7):2803–2816, Apr 2006.

[88] Sambrook, J. and D.W. Russell: *Molecular cloning. A laboratory manual.* Cold Spring Harbor Laboratory Press, 3rd edition, 2001.

[89] Studier, F.W.: *Protein production by auto-induction in high density shaking cultures.* Protein Expr Purif, 41(1):207–234, May 2005.

[90] Bradford, M.M.: *A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.* Anal Biochem, 72:248–254, May 1976.

[91] Neuhoff, V., N. Arold, D. Taube and W. Ehrhardt: *Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250.* Electrophoresis, 9(6):255–262, Jun 1988.

[92] Amberg, D.C., D.J. Burke and J.N. Strathern: *Methods in yeast genetics.* Cold Spring Harbor Laboratory Press, 2005.

[93] Puig, O., F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm and B. Séraphin: *The tandem affinity purification (TAP) method: a general procedure of protein complex purification.* Methods, 24(3):218–229, Jul 2001.

[94] Ito, H., Y. Fukuda, K. Murata and A. Kimura: *Transformation of intact yeast cells treated with alkali cations.* J Bacteriol, 153(1):163–168, Jan 1983.

[95] Gietz, R.D. and R.H. Schiestl: *High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method.* Nat Protoc, 2(1):31–34, 2007.

[96] Müller, M., R.G. Heym, A. Mayer, K. Kramer, M. Schmid, P. Cramer, H. Urlaub, R. Jansen and D. Niessing: *A cytoplasmic complex mediates specific mRNA recognition and localization in yeast.* PLoS Biol, 9(4):e1000611, Apr 2011.

[97] Schmitzová, J., N. Rasche, O. Dybkov, K. Kramer, P. Fabrizio, H. Urlaub, R. Lührmann and V.Pena: *Crystal structure of Cwc2 reveals a novel architecture of a multipartite RNA-binding protein.* EMBO J, 31(9):2222–2234, May 2012.

[98] Olsen, J.V., L.M.F. de Godoy, G. Li, B. Macek, P. Mortensen, R. Pesch, A. Makarov, O. Lange, S. Horning and M. Mann: *Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap.* Mol Cell Proteomics, 4(12):2010–2021, Dec 2005.

[99] Kohlbacher, O., K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff and M. Sturm: *TOPP–the OpenMS proteomics pipeline.* Bioinformatics, 23(2):e191–e197, Jan 2007.

[100] Sturm, M., A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert and O. Kohlbacher: *OpenMS - an open-source software framework for mass spectrometry.* BMC Bioinformatics, 9:163, 2008.

[101] Martens, L., M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W.H. Tang, A. Römpp, S. Neumann, A.D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P. Binz and E.W. Deutsch: *mzML–a community standard for mass spectrometry data.* Mol Cell Proteomics, 10(1):R110.000133, Jan 2011.

[102] Chambers, M.C., B. Maclean, R. Burke, D. Amodei, D.L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T.A. Baker, M. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S.L. Seymour, L.M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E.W. Deutsch, R.L. Moritz, J.E. Katz, D.B. Agus, M. MacCoss, D.L .Tabb and P. Mallick: *A cross-platform toolkit for mass spectrometry and proteomics.* Nat Biotechnol, 30(10):918–920, Oct 2012.

[103] Cox, J. and M. Mann: *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 26(12):1367–1372, Dec 2008.

[104] Kramer, K., P. Hummel, H. Hsiao, X. Luo, M. Wahl and H. Urlaub: *Mass-spectrometric analysis of proteins cross-linked to 4-thio-uracil- and 5-bromo-uracil-substituted RNA.* Int J Mass Spectrom, 304(2-3):184–194, Jul 2011.

[105] Papula, L.: *Mathematische Formelsammlung für Ingenieure und Naturwissenschaftler, 8. Auflage.* Vierwegs Fachbücher der Technik, 2003.

[106] Stagno, J.R., A.S. Altieri, M. Bubunenko, S.G. Tarasov, J. Li, D.L. Court, R.A. Byrd and X. Ji: *Structural basis for RNA recognition by NusB and NusE in the initiation of transcription antitermination.* Nucleic Acids Res, 39(17):7803–7815, Sep 2011.

[107] Olsen, J.V., J.C. Schwartz, J. Griep-Raming, M.L. Nielsen, E. Damoc, E. Denisov, O. Lange, P. Remes, D. Taylor, M. Splendore, E.R. Wouters, M. Senko, A. Makarov, M. Mann and S. Horning: *A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed.* Mol Cell Proteomics, 8(12):2759–2769, Dec 2009.

[108] Pedrioli, P.G.A., J.K. Eng, R. Hubley, M. Vogelzang, E.W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R.H. Angeletti, R. Apweiler, K. Cheung, C.E. Costello, H. Hermjakob, S. Huang, R.K. Julian, E. Kapp, M.E. McComb, S.G. Oliver, G. Omenn, N.W. Paton, R. Simpson, R. Smith, C.F. Taylor, W. Zhu and R. Aebersold: *A common open representation of mass spectrometry data and its application to proteomics research.* Nat Biotechnol, 22(11):1459–1466, Nov 2004.

[109] Niessing, D., S. Hüttelmaier, D. Zenklusen, R.H. Singer and S.K. Burley: *She2p is a novel RNA binding protein with a basic helical hairpin motif.* Cell, 119(4):491–502, Nov 2004.

[110] Landers, S.M., M.R. Gallas, J. Little and R.M. Long: *She3p possesses a novel activity required for ASH1 mRNA localization in Saccharomyces cerevisiae.* Eukaryot Cell, 8(7):1072–1083, Jul 2009.

[111] Rasche, N., O. Dybkov, J. Schmitzová, B. Akyildiz, P. Fabrizio and R. Lührmann: *Cwc2 and its human homologue RBM22 promote an active conformation of the spliceosome catalytic centre.* EMBO J, 31(6):1591–1604, Mar 2012.

[112] McGrail, J.C., A. Krause and R.T. O'Keefe: *The RNA binding protein Cwc2 interacts directly with the U6 snRNA to link the nineteen complex to the spliceosome during pre-mRNA splicing.* Nucleic Acids Res, 37(13):4205–4217, Jul 2009.

[113] Picotti, P., M. Clément-Ziza, H. Lam, D.S. Campbell, A. Schmidt, E.W. Deutsch, H. Röst, Z. Sun, O. Rinner, L. Reiter, Q. Shen, J.J. Michaelson, A. Frei, S. Alberti, U. Kusebauch, B. Wollscheid, R.L. Moritz, A. Beyer and R. Aebersold: *A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis.* Nature, 494(7436):266–270, Feb 2013.

[114] Gnad, F., J. Gunawardena and M. Mann: *PHOSIDA 2011: the posttranslational modification database.* Nucleic Acids Res, 39(Database issue):D253–D260, Jan 2011.

[115] Consortium, UniProt: *Reorganizing the protein space at the Universal Protein Resource (UniProt).* Nucleic Acids Res, 40(Database issue):D71–D75, Jan 2012.

[116] Ben-Shem, A., N. Garreau de Loubresse, S. Melnikov, L. Jenner, G. Yusupova and M. Yusupov: *The structure of the eukaryotic ribosome at 3.0 Å resolution.* Science, 334(6062):1524–1529, Dec 2011.

[117] Rass, U. and B. Kemper: *Crp1p, a new cruciform DNA-binding protein in the yeast Saccharomyces cerevisiae.* J Mol Biol, 323(4):685–700, Nov 2002.

[118] Lee, K.M., W.J. Choi, Y. Lee, H.J. Lee, L.X. Zhao, H.W. Lee, J.G. Park, H.O. Kim, K.Y. Hwang, Y. Heo, S. Choi and L.S. Jeong: *X-ray crystal structure and binding mode analysis of human S-adenosylhomocysteine hydrolase complexed with novel mechanism-based inhibitors, haloneplanocin A analogues.* J Med Chem, 54(4):930–938, Feb 2011.

[119] Leskovac, V., S. Trivić and D. Pericin: *The three zinc-containing alcohol dehydrogenases from baker's yeast, Saccharomyces cerevisiae.* FEMS Yeast Res, 2(4):481–494, Dec 2002.

[120] Nagy, E., T. Henics, M. Eckert, A. Miseta, R.N. Lightowlers and M. Kellermayer: *Identification of the NAD(+)-binding fold of glyceraldehyde-3-phosphate dehydrogenase as a novel RNA-binding domain.* Biochem Biophys Res Commun, 275(2):253–260, Aug 2000.

[121] Kim, J. and C.V. Dang: *Multifaceted roles of glycolytic enzymes.* Trends Biochem Sci, 30(3):142–150, Mar 2005.

[122] Iraqui, I., G. Kienda, J. Soeur, G. Faye, G. Baldacci, R.D. Kolodner and M. Huang: *Peroxiredoxin Tsa1 is the key peroxidase suppressing genome instability and protecting against cell death in Saccharomyces cerevisiae.* PLoS Genet, 5(6):e1000524, Jun 2009.

[123] Nottrott, S., K. Hartmuth, P. Fabrizio, H. Urlaub, I. Vidovic, R. Ficner and R. Lührmann: *Functional interaction of a novel 15.5kD [U4/U6.U5] tri-snRNP protein with the 5' stem-loop of U4 snRNA.* EMBO J, 18(21):6119–6133, Nov 1999.

[124] Vidovic, I., S. Nottrott, K. Hartmuth, R. Lührmann and R. Ficner: *Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment.* Mol Cell, 6(6):1331–1342, Dec 2000.

[125] Hernández-Pérez, L., F. Depardón, F. Fernández-Ramírez, A. Sánchez-Trujillo, R.M. Bermúdez-Crúz, L. Dangott and C. Montañez: *a-Enolase binds to RNA.* Biochimie, 93(9):1520–1528, Sep 2011.

[126] Barbas, A., A. Popescu, C. Frazão, C.M. Arraiano and A.M. Fialho: *Rossmann-fold motifs can confer multiple functions to metabolic enzymes: RNA binding and ribonuclease activity of a UDP-glucose dehydrogenase.* Biochem Biophys Res Commun, 430(1):218–224, Jan 2013.

[127] Zubarev, R.A., N.K. Kelleher and F.W. McLafferty: *Electron capture dissociation of multiply charged protein cations: a nonergodic process.* J Am Chem Soc, 120(13):3265–3266, Mar 1998.

[128] Syka, J.E.P., J.J. Coon, M.J. Schroeder, J. Shabanowitz and D.F. Hunt: *Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.* Proc Natl Acad Sci U S A, 101(26):9528–9533, Jun 2004.

[129] Palumbo, A.M., S.A. Smith, C.L. Kalcic, M. Dantus, P.M. Stemmer and G.E. Reid: *Tandem mass spectrometry strategies for phosphoproteome analysis.* Mass Spectrom Rev, 30(4):600–625, 2011.

Acknowledgements

First and foremost I would like to thank my supervisor Henning Urlaub for the possibility to work on this interesting and challenging project, the continuous support including advice and infrastructure, and the chance to attend several interesting international meetings.

I am grateful to Jörg Stülke and Claudia Höbartner for their support as members of my thesis committe, and to Reinhard Lührmann, Holger Stark, and Kai Tittmann for participation in my extended committee.

My appreciation also goes to the Göttingen Graduate School for Neurosciences, Biophysics, and Molecular Biosciences for providing unique chances to broaden my knowledge, all administrative support and gracious travel grants.

I would like to thank all (past and present) members of the Bioanalytical Mass Spectrometry group for the great working atmosphere and joyful un-sciency times. Especially, I would like to thank Carla, Miro and He-Hsuan for the help and advice, particularly in the beginning; Romina for countless discussions and the not less important moral support; Caro for being an amazing master student both scientifically and personally; all cross-linkers (Florian, Uzma, Saadia, Kundan); Dominika, Samir, Christof, Ingo, Ilian, Sunit, Jasmin, Chung-Tien; and last but certainly not least Moni, Uwe, Johanna, and Lisa for the excellent technical assistance.

I wish to thank Reinhard Lührmann and all members of the Department of Cellular Biochemistry for the cooperative working atmosphere. Many people have answered questions and introduced me to new methods. Also I very much appreciated the continuous supply of clean glassware ensured by Uschi, Gertrud, and Claudia, and the support in administrative issues by Juliane. Special thanks to Boon for sharing knowledge and skill and (upon special request) for the lab dance.

I appreciate the work of and discussion with all collaborators in the laboratories of Markus Wahl, Dierk Niessing, Vlad Pena, and Oliver Kohlbacher. I am especially grateful for the cooperative teamwork with Petra Hummel and Timo Sachsenberg in the development of the data analysis approaches. Thank you also: Xiao, Roland, Jana, Olex, Andreas.

Thanks go to all professors and supervisors that guided me through my studies of Chemistry at the Philipps-Universität Marburg. Especially, thank you Uwe for teaching me so much about mass spectrometry, a knowledge that was invaluable during my doctorate.

Many people have supported me and my work in the background in administrative and organizational issues which still did not go unnoticed. Therefore, I would like to thank the IT & Electronics service (Günni, Sebastian, Heiko), the media service (Eddie, Herr Sebesse, especially for their skill and support in preparing figures), and the administration of the MPI for Biophysical Chemistry.

Thank you for proof-reading and suggestions: Romina, Miro, Christof, Boon; and Sebastian for correcting my English, and especially Rachael for her amazingly thorough search for all spelling and grammar mistakes in my results and all not-science-related conversations in the past weeks.

I would like to thank my parents for their continuous financial support during my studies, raising me like they did and listening to all joys and problems, as small as they sometimes were. To my sister and all my old friends, especially Felix, Marianne and Britta, thank you for being in my life.

Appendices

A Masses of identified cross-links

On the following pages, all cross-links identified in the course of this thesis are listed with the corresponding calculated and experimental mass values. Mass values were calculated with online tools listed in Section 2.2.10.3. Experimental masses were determined by averaging several MS survey scans, if possible all spectra exceeding 50% intensity of the extracted ion chromatogram signal, otherwise around the time point when the MS/MS fragment spectrum was recorded. *m/z* values and mass deviation were calculated according to the following equations:

$$m/z = \frac{m + z * m(H)}{z} \tag{A.1}$$

m(H) corresponds to the mass of a hydrogen atom (proton).

$$\text{mass deviation } [ppm] = \frac{m(exp) - m(calc)}{m(calc)} * 10^6 \tag{A.2}$$

The columns of the following tables are described in more detail below. Theoretical masses of cross-links with the 152 Da adduct were calculated assuming the exact mass to be 151.9938 Da.

protein	protein name in case of yeast proteins identified after TAP tag purification: recommended name according to UniProt [115]
UniProt	unique UniProt ID (for yeast proteins identified after TAP tag purification)
position	position of cross-linked peptide in the protein sequence
peptide	amino acid sequence of the cross-linked peptide
aa	position of cross-linked amino acid residue
RNA	composition of cross-linked RNA
fig	reference to figure of annotated MS/MS fragment spectrum
m(peptide)	calculated mass of cross-linked peptide
m(RNA)	calculated mass of cross-linked RNA
m(XL)	calculated mass of cross-link
z	charge state in which cross-link was observed
m/z	calculated *m/z* with observed charge state
m/z exp	experimentally observed *m/z*
Δm	mass deviation between calculated and experimental cross-link mass in ppm for all regular cross-links absolute deviation in Da for cross-links with the 152 Da adduct

Table A.1: Overview of cross-links of NusB and S10 and the corresponding mass values.

protein	position	peptide	aa	RNA	fig	m(peptide)	m(RNA)	m(XL)	z	m/z	m/z exp	Δm
NusB	I87–R95	IALYELSKR	K94	258 adduct	B.1	1091.6338	-	-	2	-	675.8458	*258.0422 Da*
	S96–K112	SDVPYKVAINEAIELAK	K101	258 adduct	3.10	1859.0039	-	-	2	-	1059.5396	*258.0597 Da*
			K101	[4SU –H_2S]	-	1859.0039	306.0253	2165.0292	2	1083.5224	1083.5555	30.5 ppm
	S113–K129	SFGAEDSHKFVNGVLDK	-	258 adduct	-	1848.9005	-	-	3	-	703.3201	*258.0364 Da*
			F122?	[4SU –H_2S]	3.6	1848.9005	306.0253	2154.9258	3	719.3164	719.3298	18.6 ppm
			-	[(4SU)A –H_2S]	-	1848.9005	635.0778	2483.9783	3	829.0006	829.0264	31.1 ppm
			-	[(4SU)C –H_2S]	-	1848.9005	611.0666	2459.9671	3	820.9968	821.0108	17.1 ppm
		SFGAEDSHKFVNGVLDK (carbamylated)	-	[4SU –H_2S]	-	1891.9063	306.0253	2197.9316	3	733.6517	733.6676	21.7 ppm
			-	[(4SU)(4SU) –HPO_3]	3.5	1891.9063	582.0491	2473.9554	3	825.6596	825.6838	29.3 ppm
			-	[(4SU)A –H_2S]	-	1891.9063	635.0778	2526.9841	3	843.3358	843.3553	23.1 ppm
S10	L17–R31	LIDQATAEIVETAKR	K30	[4SU –H_2S]	3.7	1656.9046	306.0253	1962.9299	3	655.3178	655.3242	9.8 ppm
	G38–R44	GPIPLPTR	G38	[4SU –H_2S]	3.9	849.5072	306.0253	1155.5325	2	578.7741	578.7913	29.7 ppm
	L73–R89	LVDIVEPTEKTVDALMR	-	258 adduct	B.2	1928.0288	-	-	3	-	729.6940	*258.0298 Da*
		LVDIVEPTEKTVDALM(Ox)R	-	[4SU –H_2S]	-	1944.0237	306.0253	2250.0490	3	751.0241	750.9923	42.3 ppm

Table A.2: Overview of cross-links of She2p and She3p identified after measurement on the Q-ToF Ultima and the corresponding mass values.

protein	position	peptide	aa	RNA	fig	m(peptide)	m(RNA)	m(XL)	z	*m/z*	*m/z* exp	Δm
She2p	I164–K179	IGSNLLDLEVVQFAIK	F181	[U]	3.13	1757.9926	324.0359	2082.0285	2	1042.0221	1041.9695	50.5 ppm
		IGSNLLDLEVVQFAIK (carbamylated)	F181	[U]	-	1800.9984	324.0359	2125.0343	2	1063.5250	1063.4763	45.8 ppm
She3p-short	G334–K340	GPLGSMGNSSNNK	S337–N339	[AAU –HPO_3]	3.14b	1261.5720	902.1746	2163.7466	2	1082.8811	1082.8723	8.1 ppm
		GPLGSMGNSSNNK (carbamylated)	-	[U]	3.14a	1304.5778	324.0359	1628.6137	2	815.3147	815.3373	27.7 ppm
		GPLGSM(Ox)GNSSNNK	-	[U]	-	1277.5669	324.0359	1601.6028	2	801.8092	801.8643	69.7 ppm

Table A.3: Overview of cross-links of She2p and She3p identified after measurement on the Orbitrap Velos and the corresponding mass values.

protein	position	peptide	aa	RNA	fig	m(peptide)	m(RNA)	m(XL)	z	*m/z*	*m/z* exp	Δm
She2p	M1–K3	GPLGSMSK	-	[U –H_2O]	B.3	775.3898	306.0253	1081.4151	2	541.7154	541.7147	1.3 ppm
			-	[AU –H_2O]	-	775.3898	635.0778	1410.4676	2	706.2416	706.2408	1.1 ppm
		GPLGSM(Ox)SK	-	[U –H_2O]	-	791.3847	306.0253	1097.4100	2	549.7128	549.7119	1.6 ppm
			-	[AU –H_2O]	-	791.3847	635.0778	1426.4625	2	714.2391	714.2380	1.5 ppm
	Y27–K37	YLSSYIH(Ox)VLNK	-	[AAU –HPO_3]	3.16	1349.6979	902.1746	2251.8725	3	751.6320	751.6308	1.6 ppm
	F64–K82	FYNDCVLSYNASEFINEGK	-	[U]	-	2211.9782	324.0359	2536.0141	2	1269.0149	1269.0138	0.87 ppm
			C68	[U +152]	B.4	2211.9782	*476.0297*	*2688.0079*	3	*897.0104*	897.0116	*151.9973 Da*
			C68	[AU +152]	-	2211.9782	*805.0822*	*3017.0604*	3	*1006.6946*	1006.6960	*151.9980 Da*
			C68	[AAU +152 –HPO_3]	-	2211.9782	*1054.1684*	*3266.1466*	3	*1089.7233*	1089.7249	*151.9985 Da*
			C68	[AAU +152]	-	2211.9782	*1134.1347*	*3346.1129*	3	*1116.3788*	1116.3802	*151.9981 Da*
	F64–K94	FYNDCVLSYNASEFINEGKNELDPEADSFDK	C68	[U +152]	-	3572.5565	*476.0297*	*4048.5862*	3	*1350.5365*	1350.5366	*151.9940 Da*
			C68	[AU +152]	-	3572.5565	*805.0822*	*4377.6387*	3	*1460.2207*	1460.2207	*151.9938 Da*
	C106–K123	CVETFDLLNYYLTQSLQK	C106	[U +152]	B.5	2177.0714	*476.0297*	*2653.1011*	3	*885.3748*	885.3761	*151.9976 Da*
			C106	[AU +152]	-	2177.0714	*805.0822*	*2982.1536*	3	*995.0590*	995.0603	*151.9977 Da*
	I164–K179	IGSNLLDLEVVQFAIK	F176	[U –H_3PO_4]	-	1757.9926	226.0590	1984.0516	3	662.3583	662.3576	1.8 ppm
			F176	[U –H_2O]	3.18b	1757.9926	306.0253	2064.0179	3	689.0138	689.0135	0.44 ppm
			F176	[U]	-	1757.9926	324.0359	2082.0285	3	695.0173	695.0178	0.72 ppm
			-	[GU]	-	1757.9926	669.0833	2427.0759	3	810.0331	810.0328	0.37 ppm
	L223–K240	LSALDEEFDVVATKWHDK	W237	[U +152]	3.17	2102.0319	*476.0297*	*2578.0616*	3	*860.3617*	860.3622	*151.9954 Da*
			-	[AU +152]	-	2102.0319	*805.0822*	*2907.1141*	4	*727.7863*	727.7868	*151.9957 Da*
She3p	M130–K138	M(Ox)DQLSKLAK	K135	[U –H_2O]	B.6	1048.5586	306.0253	1354.5839	2	678.2998	678.2979	2.8 ppm
	N139–K150	NSSAIEQSCSEK	C147	[U +152]	B.9	1281.5506	*476.0297*	*1757.5803*	2	*879.7980*	879.7984	*151.9947 Da*
			C147	[AU +152]	-	1281.5506	*805.0822*	*2086.6328*	2	*1044.3242*	1044.3260	*151.9974 Da*
			C147	[AAU +152 –HPO_3]	-	1281.5506	*1054.1684*	*2335.7190*	3	*779.5808*	779.5807	*151.9933 Da*
	G283–K291	GAVVQTLKK	K290	[U –H_2O]	B.7	942.5861	306.0253	1248.6114	2	625.3135	625.3105	4.8 ppm
	T383–R405	TNVTHNNDPSTSPTISVPPGVTR	-	[GU]	B.8	2390.1825	669.0833	3059.2658	3	1020.7631	1020.7627	0.39 ppm
			-	[AAU –HPO_3]	-	2390.1825	902.1748	3292.3573	3	1098.4602	1098.4598	0.36 ppm
She3p-short	G334–K340	GPLGSMGNSSNNK	-	[U]	-	1261.5720	324.0359	1585.6079	2	793.8118	793.8105	1.6 ppm
			-	[GU]	-	1261.5720	669.0833	1930.6553	2	966.3355	966.3368	1.3 ppm

Table A.4: Overview of cross-links of Cwc2 and the corresponding mass values.

domain	position	peptide	aa	RNA	fig	m(peptide)	m(RNA)	m(XL)	z	m/z	m/z exp	Δm
Torus	W37–K61	WSQGFAGNTRFVSPFALQPQLHSGK	F47	[UU]	-	2759.3931	630.0611	3389.4542	3	1130.8259	1130.8278	1.7 ppm
				[AUU]	-	2759.3931	959.1137	3718.5068	3	1240.5101	1240.5109	0.64 ppm
	F47–K61	FVSPFALQPQLHSGK	F47	[U –H_2O]	B.10	1654.8830	306.0253	1960.9083	2	981.4620	981.4644	2.4 ppm
				[U]	-	1654.8830	324.0359	1978.9189	3	660.6474	660.6451	3.5 ppm
				[UU]	-	1654.8830	630.0611	2284.9441	3	762.6558	762.6529	3.8 ppm
zinc finger	G79–K101	GM(Ox)CCLGPKCEYLHHIPDEEDIGK	C87	[U +152 –H_2O]	-	2602.1323	*458.0191*	*3060.1514*	4	*766.0457*	766.0461	*151.9956 Da*
	C87–K101	CEYLHHIPDEEDIGK	C87	[U]	-	1796.8039	324.0359	2120.8398	3	707.9544	707.9526	2.5 ppm
				[U +152]	-	1796.8039	*476.0297*	*2272.8336*	3	*758.6190*	758.6209	*151.9995 Da*
				[AU +152]	B.11	1796.8039	*805.0822*	*2601.8861*	3	*868.3032*	868.3057	*152.0014 Da*
				[AAU +152]	-	1796.8039	*1134.1347*	*2930.9386*	4	*733.7425*	733.7439	*151.9996 Da*
				[GU +152]	-	1796.8039	*821.0771*	*2617.8810*	3	*873.6348*	873.6370	*152.0004 Da*
connector element	F117–R131	FADYREDMGGIGSFR	Y120	[U]	B.12	1719.7674	324.0359	2043.8033	2	1022.9095	1022.9121	2.5 ppm
				[AU]	-	1719.7674	653.0884	2372.8558	3	791.9597	791.9588	1.1 ppm
RNP2	T136–K149	TLYVGGIDGALNSK	Y138	[U]	B.13	1406.7405	324.0359	1730.7764	2	866.3960	866.3984	2.8 ppm
				[AU]	-	1406.7405	653.0884	2059.8289	3	687.6174	687.6177	0.44 ppm
				[AAU]	-	1406.7405	982.1409	2388.8814	3	797.3016	797.3022	0.50 ppm
				[AUU]	-	1406.7405	959.1137	2365.8542	3	789.6259	789.6241	2.3 ppm
				[GU]	-	1406.7405	669.0833	2075.8238	2	1038.9197	1038.9208	1.1 ppm
				[GGU]	-	1406.7405	1014.1307	2420.8712	3	807.9649	807.9637	1.5 ppm
				[AGU]	-	1406.7405	998.1358	2404.8763	3	802.6332	802.6319	1.6 ppm
				[UU]	-	1406.7405	630.0611	2036.8017	2	1019.4087	1019.4069	1.8 ppm
				[CU]	-	1406.7405	629.0772	2035.8177	2	1018.9167	1018.9154	1.3 ppm
				[ACU]	-	1406.7405	958.1297	2364.8702	3	789.2979	789.2960	2.4 ppm
				[CGU]	-	1406.7405	974.1246	2380.8651	3	794.6295	794.6267	3.5 ppm
RRM	H150–R159	HLKPAQIESR	K152	[U –H_2O]	B.14	1177.6567	306.0253	1483.6820	3	495.5685	495.5680	1.0 ppm
				[AU –H_2O]	-	1177.6567	635.0778	1812.7345	3	605.2526	605.2525	0.17 ppm
RNP1	N180–K185	NCGFVK	C181	[U]	-	666.3159	324.0359	990.3518	2	496.1837	496.1826	2.2 ppm
				[U +152]	B.15	666.3159	*476.0297*	*1142.3456*	2	*572.1806*	572.1823	*151.9972 Da*
				[AU]	-	666.3159	653.0884	1319.4043	2	660.7100	660.7098	0.30 ppm
				[AU +152]	-	666.3159	*805.0822*	*1471.3981*	2	*736.7069*	736.7100	*152.0001 Da*
				[AAU +152]	-	666.3159	*1134.1347*	*1800.4506*	2	*901.2331*	901.2363	*152.0002 Da*
				[AUU +152]	-	666.3159	*1111.1075*	*1777.4234*	2	*889.7195*	889.7199	*151.9946 Da*
				[GU]	-	666.3159	669.0833	1335.3992	2	668.7074	668.7055	2.8 ppm
				[GU +152]	-	666.3159	*821.0771*	*1487.3930*	2	*744.7043*	744.7059	*151.9970 Da*
				[AGU +152]	-	666.3159	*1150.1296*	*1816.4455*	2	*909.2306*	909.2303	*151.9933 Da*
				[UU]	-	666.3159	630.0611	1296.3770	2	649.1963	649.1958	0.77 ppm
				[UU +152]	-	666.3159	*782.0549*	*1448.3708*	2	*725.1932*	725.1958	*151.9990 Da*

Table A.5: Overview of cross-links from the 40S small ribosomal subunit and the corresponding mass values.

protein	UniProt	position	peptide	aa	RNA	fig	m(peptide)	m(RNA)	m(XL)	z	m/z	m/z exp	Δm
40S ribosomal protein S1-A/-B	P33442/ P23248	K116–K128	KWQTLIEANVTVK	W117	[U $-H_2O$]	B.16	1528.8613	306.0253	1834.8866	2	918.4511	918.4503	0.87
				-	[AU $-HPO_3$]	-	1528.8613	573.1221	2101.9834	3	701.6689	701.6686	0.48
40S ribosomal protein S3	P05750	G95–K108	GLSAVAQAESMKFK	-	[GU]	B.17	1465.7598	669.0833	2134.8431	3	712.6222	712.6211	1.50
				-	[UU]	-	1465.7598	630.0612	2095.8210	3	699.6148	699.6144	0.57
		G133–K141	GCEVVVSGK	C134	[U +152 $-H_2O$]	B.18	876.4374	*458.0191*	*1334.4565*	2	*668.2361*	668.2366	*151.9949*
40S ribosomal protein S5	P26783	N186–K203	NIKTIAETLAEELINAAK	-	[U $-H_2O$]	-	1941.0782	306.0253	2247.1035	3	750.0423	750.0426	0.40
				-	[GU $-H_2O$]	-	1941.0782	651.0727	2592.1509	3	865.0581	865.0582	0.12
		T189–K203	TIAETLAEELINAAK	T189	[U $-H_3PO_4$]	-	1585.8562	226.0590	1811.9152	3	604.9795	604.9788	1.21
				T189	[U $-H_2O$]	-	1585.8562	306.0253	1891.8815	3	631.6350	631.6342	1.21
				T189	[U]	-	1585.8562	324.0359	1909.8921	2	955.9539	955.9514	2.56
				T189	[GU $-HPO_3$]	-	1585.8562	589.1170	2174.9732	3	725.9989	725.9988	0.09
				T189	[GU $-H_2O$]	B.19	1585.8562	651.0727	2236.9289	3	746.6508	746.6512	0.58
				T189	[GU]	-	1585.8562	669.0833	2254.9395	3	752.6543	752.6537	0.80
				-	[ACGU]	-	1585.8562	1303.1771	2889.0333	3	964.0189	964.0182	0.73
40S ribosomal protein S11-A/-B	P0CX47/ P0CX48	V117–K133	VQVGDIVTVGQCRPISK	C128	[U $-H_2O$]	-	1797.9770	306.0253	2104.0023	3	702.3419	702.3400	2.71
				C128	[AU $-H_2O$]	B.20	1797.9770	635.0778	2433.0548	3	812.0261	812.0257	0.45
40S ribosomal protein S14-A/-B	P06367/ P39516	I19–K36/ I20–K37	IYASFNDTFVHVTDLSGK	-	[CU]	-	2012.9843	629.0772	2642.0615	3	881.6950	881.6939	1.21
				-	[GU]	-	2012.9843	669.0833	2682.0676	3	895.0303	895.0294	1.04
				-	[UU]	B.21	2012.9843	630.0612	2643.0455	3	882.0230	882.0220	1.10
		A50–K70/ A51–K71	ADRDESSPYAAMLAAQDVAAK	-	[GU]	B.22	2179.0215	669.0833	2848.1048	3	950.3761	950.3746	1.54
40S ribosomal protein S16-A/-B	P0CX51/ P0CX52	V69–R82	VTGGGHVSQVYAIR	H74	[U $-H_2O$]	B.23	1442.7629	306.0253	1748.7882	3	583.9372	583.9366	1.03
				H74	[U]	-	1442.7629	324.0359	1766.7988	3	589.9407	589.9399	1.41
40S ribosomal protein S17-A/-B	P02407/ P14127	L34–K44	LCDEIATIQSK	C35	[U +152]	B.24	1219.6118	*476.0297*	*1695.6415*	2	*848.8286*	848.8292	*151.9951*
				C35	[AAU +152]	-	1219.6118	*1134.1347*	*2353.7465*	3	*785.5900*	785.5899	*151.9936*
		I50–K59	IAGYTTHLMK	H56	[U $-H_2O$]	B.25	1133.5902	306.0253	1439.6155	2	720.8156	720.8147	1.18
				-	[U]	-	1133.5902	324.0359	1457.6261	3	486.8832	486.8824	1.57
			IAGYTTHLM(Ox)K	H56	[U]	-	1149.5851	324.0359	1473.6210	3	492.2148	492.2138	2.03
40S ribosomal protein S24-A/-B	P0CX31/ P0CX32	D53–R61	DAVSVFGFR	-	[U]	B.26	996.5028	324.0359	1320.5387	2	661.2772	661.2767	0.68
		D115–K123	DKKIFGTGK	K117	[U $-H_2O$]	-	992.5654	306.0253	1298.5907	2	650.3032	650.3015	2.54
				K117	[U]	-	992.5654	324.0359	1316.6013	3	439.8749	439.8743	1.36
				K117	[CU]	B.27	992.5654	629.0772	1621.6426	2	811.8291	811.8289	0.25
40S ribosomal protein S29-A	P41057	V23–R32	VCSSHTGLIR	C24	[U +152 $-H_2O$]	B.28	1071.5495	*458.0191*	*1529.5686*	2	*765.7921*	765.7914	*151.9924*
40S ribosomal protein S29-B	P41058	V23–R32	VCSSHTGLVR	C24	[U +152 $-H_2O$]	B.29	1057.5338	*458.0191*	*1515.5529*	2	*758.7843*	758.7819	*151.9891*
				C24	[AU +152 $-H_2O$]	-	1057.5338	*787.0716*	*1844.6054*	3	*615.8763*	615.8770	*151.9960*
Guanine nucleotide-binding protein subunit beta-like protein	P38011	G138–R155	GQCLATLLGHNDWVSQVR	C140	[U +152 $-H_2O$]	B.30	1995.9948	*458.0191*	*2454.0139*	3	*819.0124*	819.0124	*151.9937*

Table A.6: Overview of cross-links from the 60S large ribosomal subunit (proteins L1 to L8) and the corresponding mass values.

protein	UniProt	position	peptide	aa	RNA	fig	m(peptide)	m(RNA)	m(XL)	z	*m/z*	*m/z* exp	Δm
60S ribosomal protein L1-A/-B	P0CX43/	S79–K91	SCGVDAMSVDDLK	C80	[U +152 –H_2O]	-	1338.5795	*458.0191*	*1796.5986*	2	*899.3071*	899.3074	*151.9944*
	P0CX44	S79–K92	SCGVDAMSVDDLKK	C80	[U +152 –H_2O]	B.31	1466.6744	*458.0191*	*1924.6935*	3	*642.5723*	642.5729	*151.9956*
				C80	[U +152]	-	1466.6744	*476.0297*	*1942.7041*	3	*648.5758*	648.5759	*151.9940*
60S ribosomal protein L2-A /-B	P0CX45/	A129–K145	ASGNYVIIIGHNPDENK	Y133	[U]	B.32	1839.9114	324.0359	2163.9473	3	722.3236	722.3225	1.48
	P0CX46	G201–K221	GVAMNPVDHPHGGGNHQHIGK	-	[AAGU –H_2O]	B.33	2158.0238	1309.1778	3467.2016	4	867.8082	867.8088	0.70
60S ribosomal protein L3	P14126	V249–R266	VACIGAWHPAHVMWSVAR	C251	[U –H_2O]	B.34	1989.9817	306.0253	2296.0070	3	766.3435	766.3433	0.22
60S ribosomal protein L4-A/-B	P10664/	S85–R95	SGQGAFGNMCR	C94	[U –H_2O]	B.35	1126.4647	306.0253	1432.4900	2	717.2528	717.2521	0.98
	P49626			C94	[U]	-	1126.4647	324.0359	1450.5006	2	726.2581	726.2570	1.51
				C94	[U +152 –H_2O]	-	1126.4647	*458.0191*	*1584.4838*	2	*793.2497*	793.2503	*151.9950*
				C94	[AU –H_2O]	-	1126.4647	635.0778	1761.5425	2	881.7791	881.7783	0.85
		N221–R246	NVPGVETANVASLNLLQLAPGAHLGR	H243	[AU –H_2O]	B.36	2610.4241	635.0778	3245.5019	3	1082.8418	1082.8425	0.68
		I289–K308	IINSSEIQSAIRPAGQATQK	I290	[GU]	B.37	2111.1334	669.0833	2780.2167	3	927.7467	927.7455	1.29
60S ribosomal protein L4-A	P10664	T347–K360	TGTKPAAVFTETLK	-	[AU –H_2O]	B.38	1462.8031	635.0778	2097.8809	3	700.3014	700.3012	0.33
				-	[AU]	-	1462.8031	653.0884	2115.8915	3	706.3050	706.3044	0.80
				-	[AAU]	-	1462.8031	982.1409	2444.9440	3	815.9891	815.9880	1.39
				-	[AAAU –HPO_3]	-	1462.8031	1231.2271	2694.0302	3	899.0179	899.0172	0.74
60S ribosomal protein L4-B	P49626	T347–K360	TGTKPAAVFAETLK	-	[AU –H_2O]	B.39	1432.7925	635.0778	2067.8703	3	690.2979	690.2975	0.58
				-	[AU]	-	1432.7925	653.0884	2085.8809	3	696.3014	696.3005	1.34
				-	[AAAU –HPO_3]	-	1432.7925	1231.2271	2664.0196	4	667.0127	667.0128	0.15
60S ribosomal protein L5	P26321	S197–R218	SYIFGGHVSQYMEELADDDEER	-	[U –H_2O]	-	2589.0965	306.0253	2895.1218	3	966.0484	966.0481	0.31
				-	[U]	B.40	2589.0965	324.0359	2913.1324	3	972.0519	972.0510	0.96
			SYIFGGHVSQYM(Ox)EELADDDEER	-	[U]	-	2605.0914	324.0359	2929.1273	3	977.3836	977.3825	1.09
		S197–K224	SYIFGGHVSQYMEELADDDEERFSELFK	F	[U –H_2O]	-	3340.4869	306.0253	3646.5122	3	1216.5119	1216.5121	0.19
60S ribosomal protein L6-A/-B	Q02326/	L30–R48	LRASLVPGTVLILLAGRFR	-	[GU –H_2O]	B.41	2051.2730	651.0727	2702.3457	3	901.7897	901.7901	0.44
	P05739	A32–R48	ASLVPGTVLILLAGRFR	-	[GU –H_2O]	-	1782.0879	651.0727	2433.1606	3	812.0613	812.0611	0.29
		A32–K50	ASLVPGTVLILLAGRFRGK	-	[GU –H_2O]	-	1967.2043	651.0727	2618.2770	3	873.7668	873.7681	1.49
60S ribosomal protein L6-A	Q02326	A6–K19	APKWYPSEDVAALK	-	[AU]	-	1573.8139	653.0884	2226.9023	3	743.3086	743.3072	1.84
		W9–K19	WYPSEDVAALK	-	[AU]	-	1277.6291	653.0884	1930.7175	3	644.5803	644.5792	1.71
		W9–K20	WYPSEDVAALKK	W9	[AU –H_2O]	B.42	1405.7241	635.0778	2040.8019	3	681.2751	681.2739	1.76
				W9	[AU]	-	1405.7241	653.0884	2058.8125	3	687.2786	687.2774	1.79
		H57–K70	HLEDNTLLISGPFK	-	[U –H_2O]	B.44	1582.8354	306.0253	1888.8607	2	945.4382	945.4375	0.69
				-	[U]	-	1582.8354	324.0359	1906.8713	2	954.4435	954.4419	1.62
				-	[GU]	-	1582.8354	669.0833	2251.9187	3	751.6474	751.6470	0.49
60S ribosomal protein L6-B	P05739	T2–K19	TAQQAPKWYPSEDVAAPK	-	[AU]	-	1985.9846	653.0884	2639.0730	3	880.6988	880.6977	1.25
		W9–K19	WYPSEDVAAPK	W9	[AU –H_2O]	B.43	1261.5978	635.0778	1896.6756	3	633.2330	633.2324	0.95
				-	[AU]	-	1261.5978	653.0884	1914.6862	3	639.2365	639.2369	0.57
				-	[AGU]	-	1261.5978	998.1358	2259.7336	3	754.2523	754.2523	0.04
		H57–K70	HLEDNTLLVTGPFK	-	[U –H_2O]	B.45	1582.8354	306.0253	1888.8607	2	945.4382	945.4378	0.37
				-	[U]	-	1582.8354	324.0359	1906.8713	2	954.4435	954.4418	1.73
				-	[AAU –H_2O]	-	1582.8354	964.1303	2546.9657	3	849.9964	849.9956	0.90
				-	[AAU]	-	1582.8354	982.1409	2564.9763	3	855.9999	855.9990	1.05
60S ribosomal protein L8-A	P17076	Y134–K146	YGLNHVVALIENK	-	[GU –H_2O]	-	1468.8037	651.0727	2119.8764	3	707.6333	707.6331	0.44
		Y134–K147	YGLNHVVALIENKK	-	[U]	-	1596.8987	324.0359	1920.9346	3	641.3193	641.3188	1.49
				-	[GU –H_2O]	B.46	1596.8987	651.0727	2247.9714	3	750.3316	750.3312	1.76
60S ribosomal protein L8-B	P29453	Y134–K146	YGLNHVVSLIENK	-	[U –H_2O]	-	1484.7986	306.0253	1790.8239	3	597.9491	597.9486	0.69
				-	[U]	-	1484.7986	324.0359	1808.8345	2	905.4251	905.4269	0.95
				-	[GU –H_2O]	B.47	1484.7986	651.0727	2135.8713	3	712.9649	712.9644	0.37
				-	[GU]	-	1484.7986	669.0833	2153.8819	3	718.9684	718.9671	1.73
				-	[UU –H_2O]	-	1484.7986	612.0506	2096.8492	3	699.9575	699.9589	0.90
				-	[UU]	-	1484.7986	630.0612	2114.8598	3	705.9611	705.9597	1.05
		Y134–K147	YGLNHVVSLIENKK	-	[U –H_2O]	-	1612.8936	306.0253	1918.9189	3	640.6474	640.6472	0.24

Table A.7: Overview of cross-links from the 60S large ribosomal subunit (proteins L16 to L42) and RPL7 and the corresponding mass values..

protein	UniProt	position	peptide	aa	RNA	fig	m(peptide)	m(RNA)	m(XL)	z	*m/z*	*m/z* exp	Δm
60S ribosomal protein L16-A/-B	P26784/ P26785	L141–K155/ L140–K154	LSTSVGWKYEDVVAK	Y149	[U –H$_2$O]	B.48	1680.8722	306.0253	1986.8975	3	663.3070	663.3068	0.25
				Y149	[U]	-	1680.8722	324.0359	2004.9081	3	669.3105	669.3098	1.05
				-	[UU]	-	1680.8722	630.0612	2310.9334	3	771.3189	771.3179	1.34
60S ribosomal protein L16-A	P26784	A38–R49	AEELNISGEFFR	-	[AU]	-	1410.6779	653.0884	2063.7663	3	688.9299	688.9280	2.76
				-	[ACU]	-	1410.6779	958.1297	2368.8076	3	790.6103	790.6097	0.80
		A38–K51	AEELNISGEFFRNK	-	[AU]	-	1652.8158	653.0884	2305.9042	3	769.6425	769.6420	0.69
				-	[CU]	-	1652.8158	629.0772	2281.8930	3	761.6388	761.6380	1.03
				-	[ACU –HPO$_3$]	-	1652.8158	878.1633	2530.9791	3	844.6675	844.6663	1.44
				-	[ACU]	B.49	1652.8158	958.1297	2610.9455	3	871.3230	871.3229	0.08
60S ribosomal protein L16-B	P26785	A37–R48	AEALNISGEFFR	-	[U]	-	1352.6724	324.0359	1676.7083	2	839.3620	839.3600	2.32
				-	[AU]	-	1352.6724	653.0884	2005.7608	3	669.5947	669.5938	1.39
				F38	[CU]	B.50	1352.6724	629.0772	1981.7496	3	661.5910	661.5906	0.58
				-	[ACU]	-	1352.6724	958.1297	2310.8021	3	771.2752	771.2733	2.42
		A37–K50	AEALNISGEFFRNK	-	[AU –H$_2$O]	-	1594.8103	635.0778	2229.8881	3	744.3038	744.3031	0.99
				-	[AU]	-	1594.8103	653.0884	2247.8987	3	750.3074	750.3066	1.02
				-	[CU]	-	1594.8103	629.0772	2223.8875	3	742.3036	742.3031	0.70
				-	[ACU –HPO$_3$]	-	1594.8103	878.1633	2472.9736	3	825.3323	825.3317	0.78
				-	[ACU]	-	1594.8103	958.1297	2552.9400	3	851.9878	851.9864	1.64
60S ribosomal protein L18-A/-B	P0CX49/ P0CX50	A51–K56	ALFLSK	-	[AU –H$_2$O]	-	677.4111	635.0778	1312.4889	2	657.2523	657.2517	0.84
				-	[AU]	B.51	677.4111	653.0884	1330.4995	2	666.2576	666.2566	1.43
				-	[AGU –H$_2$O]	-	677.4111	980.1253	1657.5364	2	829.7760	829.7755	0.60
				-	[AGU]	-	677.4111	998.1358	1675.5469	2	838.7813	838.7802	1.26
		A117–R130	AGGECITLDQLAVR	C121	[U +152 –H$_2$O]	B.52	1444.7343	*458.0191*	*1902.7534*	2	*952.3845*	952.3873	*151.9994*
60S ribosomal protein L23-A/-B	P0CX41/ P0CX42	E121–R128	ECADLWPR	C122	[U +152 –H$_2$O]	B.53	988.4436	*458.0191*	*1446.4627*	2	*724.2392*	724.2391	*151.9937*
60S ribosomal protein L26-B	P53221	K17–R27	KAYFTAPSSER	-	[GU]	B.54	1255.6196	669.0833	1924.7029	3	642.5754	642.5745	1.45
60S ribosomal protein L28	P02406	I43–K55	INMDKYHPGYFGK	Y48	[U –H$_2$O]	B.55	1568.7445	306.0253	1874.7698	3	625.9311	625.9310	0.11
60S ribosomal protein L31-A/-B	P0C2H8/ P0C2H9	L20–K26	LHGVSFK	-	[U –H$_2$O]	-	786.4388	306.0253	1092.4641	2	547.2399	547.2391	1.37
				-	[U]	-	786.4388	324.0359	1110.4747	2	556.2452	556.2440	2.07
				-	[GU]	-	786.4388	669.0833	1455.5221	2	728.7689	728.7675	1.85
				F25	[UU –H$_2$O]	B.56	786.4388	612.0506	1398.4894	2	700.2525	700.2517	1.14
				-	[UU]	-	786.4388	630.0612	1416.5000	2	709.2578	709.2567	1.55
				-	[UUU]	-	786.4388	936.0865	1722.5253	2	862.2705	862.2705	0.06
		L20–K27	LHGVSFKK	F25	[U –H$_2$O]	-	914.5337	306.0253	1220.5590	3	407.8608	407.8603	1.23
				-	[U]	-	914.5337	324.0359	1238.5696	3	413.8643	413.8635	2.01
				-	[UU –H$_2$O]	-	914.5337	612.0506	1526.5843	3	509.8692	509.8685	1.44
				-	[UU]	-	914.5337	630.0612	1544.5949	3	515.8728	515.8721	1.29
				-	[UUU]	-	914.5337	936.0865	1850.6202	2	926.3179	926.3172	0.76
60S ribosomal protein L33-A/-B	P05744/ P41056	I49–R60	IAYVVYRASKEVR	-	[AU –H$_2$O]	B.57	1453.8041	635.0778	2088.8819	3	697.3018	697.3024	0.91
				-	[AU]	-	1453.8041	653.0884	2106.8925	3	703.3053	703.3042	1.56
60S ribosomal protein L35-A/-B	P0CX84/ P0CX85	S50–R63	SIACVLTVINEQQR	C53	[U –H$_2$O]	B.58	1572.8293	306.0253	1878.8546	2	940.4351	940.4341	1.06
				-	[U]	-	1572.8293	324.0359	1896.8652	2	949.4404	949.4378	2.74
60S ribosomal protein L37-A	P49166	R73–K85	RFKNGFQTGSASK	-	[ACU]	-	1426.7316	958.1297	2384.8613	3	795.9616	795.9624	1.05
			RFKN(deamidated)GFQTGSASK	-	[ACU]	-	1427.7157	958.1297	2385.8454	3	796.2896	796.2881	1.88
		F74–K85	FKNGFQTGSASK	-	[ACU]	B.59	1270.6305	958.1297	2228.7602	3	743.9279	743.9263	2.11
			FKN(deamidated)GFQTGSASK	-	[ACU]	-	1271.6145	958.1297	2229.7442	3	744.2559	744.2537	2.91
60S ribosomal protein L37-B	P51402	F74–K84	FKNGFQTGSAK	-	[ACU]	B.60	1183.5985	958.1297	2141.7282	3	714.9172	714.9162	1.40
Ubiquitin-60S ribosomal protein L40	P0CH08/ P0CH09	C115–K124	CGHTNQLRPK	-	[U –H$_2$O]	-	1152.5822	306.0253	1458.6075	2	730.3116	730.3107	1.16
				C115	[U +152 –H$_2$O]	B.61	1152.5822	*458.0191*	*1610.6013*	3	*537.8749*	537.8739	*151.9908*
60S ribosomal protein L42-A/-B	P0CX27/ P0CX28	C88–K97	CKHFELGGEK	C88/K89	[U –H$_2$O]	B.62	1146.5491	306.0253	1452.5744	2	727.2950	727.2944	0.82
				C88/K89	[AU]	-	1146.5491	653.0884	1799.6375	3	600.8870	600.8858	1.94
				C88/K89	[UU]	-	1146.5491	630.0612	1776.6103	2	889.3130	889.3123	0.73
Ribosome biogenesis protein RLP7	P40693	G149–K161	GPLAVNIPNKAFK	-	[CUU]	B.63	1367.7924	935.1025	2302.8949	3	768.6394	768.6394	0.04

Table A.8: Overview of cross-links of enzymatic and RNA-binding proteins (excluding ribosomal) and the corresponding mass values.

protein	UniProt	position	peptide	aa	RNA	fig	m(peptide)	m(RNA)	m(XL)	z	*m/z*	*m/z* exp	Δm
Cruciform DNA-recognizing protein 1	P38845	I105–R118	IPEAGGLLCGKPPR	C113	[U –H_2O]	B.64	1406.7703	306.0253	1712.7956	2	857.4056	857.4047	1.05
Elongation factor 1-alpha	P02994	F402–R421	FVPSKPM(Oxidation)CVEAFSEYPPLGR	C409	[U +152]	B.65	2269.0910	*476.0297*	*2745.1207*	3	*916.0480*	916.0476	*151.9925*
Nucleolar protein 3	Q01560	I156–K173	ILNGFAFVEFEEAESAAK	F	[U –H_2O]	B.66	1970.9624	306.0253	2276.9877	3	760.0037	760.0039	0.26
		E222–R235	ENSLETTFSSVNTR	-	[U]	B.67	1583.7427	324.0359	1907.7786	2	954.8971	954.8956	1.57
Nucleolar protein 13	P53883	I240–R256	ILFVGNLSFDVTDDLLR	F242	[GU –H_2O]	B.68	1936.0305	651.0727	2587.1032	3	863.3755	863.3756	0.08
Polyadenylate-binding protein	P04147	Y319–K327	YQGVNLFVK	F325	[U –H_2O]	-	1066.5811	306.0253	1372.6064	2	687.3110	687.3096	2.04
				F325	[U]	B.69	1066.5811	324.0359	1390.6170	2	696.3163	696.3158	0.72
				F325	[AU –H_2O]	-	1066.5811	635.0778	1701.6589	2	851.8373	851.8365	0.88
				F325	[AU]	-	1066.5811	653.0884	1719.6695	2	860.8426	860.8416	1.10
				F325	[UU]	-	1066.5811	630.0612	1696.6423	2	849.3290	849.3282	0.88
				F325	[AAU]	-	1066.5811	982.1409	2048.7220	3	683.9151	683.9141	1.51
Single-stranded nucleic-acid binding protein	P10080	S184–K196	SKDTLYINNVPFK	-	[U]	-	1537.8140	324.0359	1861.8499	3	621.6244	621.6236	1.34
				-	[AC]	B.70	1537.8140	652.1044	2189.9184	3	730.9806	730.9801	0.68
				-	[AU –H_2O]	-	1537.8140	635.0778	2172.8918	3	725.3051	725.3041	1.33
				-	[AU]	-	1537.8140	653.0884	2190.9024	3	731.3086	731.3078	1.09
				-	[GU]	-	1537.8140	669.0833	2206.8973	3	736.6402	736.6388	1.95
				-	[UU]	-	1537.8140	630.0612	2167.8752	3	723.6329	723.6320	1.20
				-	[AUU]	-	1537.8140	959.1137	2496.9277	3	833.3170	833.3163	0.88
				-	[UUU]	-	1537.8140	936.0865	2473.9005	3	825.6413	825.6410	0.36
Adenosylhomocysteinase	P39954	E320–R330	ECINIKPQVDR	C321	[U +152]	B.71	1313.6761	*476.0297*	*1789.7058*	3	*597.5764*	597.5761	*151.9929*
				C321	[AU +152]	-	1313.6761	*805.0822*	*2118.7583*	3	*707.2606*	707.2610	*151.9951*
Alcohol dehydrogenase 1/3	P00330/ P07246	Y40–K60/ Y67–K86	YSGVCHTDLHAWHGDWPLPVK	C44/71	[U +152 –H_2O]	B.72	2417.1374	*458.0191*	*2875.1565*	4	*719.7969*	719.7972	*151.9949*
				C44/71	[U +152]	-	2417.1374	*476.0297*	*2893.1671*	4	*724.2996*	724.2968	*151.9827*
Enolase 1/2	P00924/ P00925	I244–K255	IGLDCASSEFFK	C248	[U]	-	1315.6118	324.0359	1639.6477	2	820.8317	820.8316	0.06
				C248	[U +152 –H_2O]	B.73	1315.6118	*458.0191*	*1773.6309*	2	*887.8233*	887.8223	*151.9919*
				C248	[U +152]	-	1315.6118	*476.0297*	*1791.6415*	2	*896.8286*	896.8318	*152.0003*
				C248	[AU +152]	-	1315.6118	*805.0822*	*2120.6940*	2	*1061.3548*	1061.3540	*151.9922*
				C248	[GU +152]	-	1315.6118	*821.0771*	*2136.6889*	3	*713.2374*	713.2369	*151.9922*
				C248	[AAU +152]	-	1315.6118	*1134.1347*	*2449.7465*	3	*817.5900*	817.5893	*151.9918*
Glyceraldehyde-3-phosphate dehydrogenase 2/3	P00358/ P00359	E250–K258	ETTYDEIKK	-	[AU –HPO_3]	B.74	1125.5553	573.1221	1698.6774	2	850.3465	850.3468	0.38
Inorganic pyrophosphatase	P00817	N83–K112	NCFPHHGYIHNYGAFPQTWEDPNVSHPETK	C84	[U +152 –H_2O]	B.75	3521.5635	*458.0191*	*3979.5826*	5	*796.9243*	796.9234	*151.9892*
Peroxiredoxin TSA1	P34760	N165–K188	NGTVLPCNWTPGAATIKPTVEDSK	C171	[U +152]	B.76	2498.2474	*476.0297*	*2974.2771*	3	*992.4335*	992.4326	*151.9911*
Phosphoglycerate kinase	P00560	Y49–R56	YVLEHHPR	-	[AU –HPO_3]	B.77	1049.5406	573.1221	1622.6627	2	812.3392	812.3396	0.55
Pyruvate kinase 1	P00549	N370–K394	NCTPKPTSTTETVAASAVAAVFEQK	C371	[U +152 –H_2O]	B.78	2550.2635	*458.0191*	*3008.2826*	3	*1003.7687*	1003.7679	*151.9915*
				C371	[U +152]	-	2550.2635	*476.0297*	*3026.2932*	3	*1009.7722*	1009.7711	*151.9905*
				C371	[AU +152 –H_2O]	-	2550.2635	*787.0716*	*3337.3351*	3	*1113.4528*	1113.4537	*151.9964*
				C371	[AU +152]	-	2550.2635	*805.0822*	*3355.3457*	3	*1119.4564*	1119.4543	*151.9876*
				C371	[GU +152 –H_2O]	-	2550.2635	*803.0666*	*3353.3301*	3	*1118.7845*	1118.7837	*151.9915*
				C371	[GU +152]	-	2550.2635	*821.0771*	*3371.3406*	3	*1124.7880*	1124.7865	*151.9893*
				C371	[UU +152 –H_2O]	-	2550.2635	*764.0444*	*3314.3079*	3	*1105.7771*	1105.7791	*151.9998*
				C371	[UU +152]	-	2550.2635	*782.0550*	*3332.3185*	3	*1111.7806*	1111.7790	*151.9889*
		Y414–R425	YRPNCPIILVTR	C418	[U +152 –H_2O]	B.79	1443.8020	*458.0191*	*1901.8211*	3	*634.9482*	634.9485	*151.9948*
				C418	[U +152]	-	1443.8020	*476.0297*	*1919.8317*	3	*640.9517*	640.9521	*151.9950*
				C418	[AU +152 –H_2O]	-	1443.8020	*787.0716*	*2230.8736*	2	*1116.4446*	1116.4462	*151.9970*
				C418	[AU +152]	-	1443.8020	*805.0822*	*2248.8842*	3	*750.6359*	750.6364	*151.9954*
				C418	[UU +152]	-	1443.8020	*782.0550*	*2225.8570*	3	*742.9601*	742.9595	*151.9919*

B MS/MS fragment spectra of identified cross-links

B.1 Annotation of MS/MS spectra of cross-linked peptides

B.1.1 Peptide sequence ions

Peptide sequence ions are annotated according to the established nomenclature (see 1.2.1.3).

Neutral losses of sequence ions are annotated with an asterisk for ammonia and a superscripted 0 for water. Annotations are only given if the signal without neutral loss is not detected.

Internal ions are annotated with capital letters and usually not indicated within the peptide sequence unless they provide evidence for the peptide sequence not covered by the a-, b-, and y-ions. Immonium ions are annotated as IM X where X is the one letter code of the corresponding amino acid.

Signals corresponding to the intact peptide are annotated as peptide^{+}, peptide^{2+} etc. This simplyfies the commonly used nomenclature that would also give the additional protons responsible for the charge, e.g. [peptide $+2$H]$^{2+}$. Similarly, the precursor ion is annotated without the additional protons, e.g. [M]$^{4+}$ instead of [M $+4$H]$^{4+}$.

B.1.2 RNA marker ions and fragments

RNA marker ions (see Table B.1) are annotated as X’ for the nucleic acid bases and as X^0 for the nucleotide minus water. The corresponding *m/z* value is underlined. As for intact peptides, the additional protons that produce the charge are ommitted from the annotation.

Table B.1: Calculated monoisotopic masses of RNA (marker) ions

RNA marker ions of the nucleic acid bases

marker	symbol	calculated *m/z*
cytosine	C’	112.0511
uracil	U’	113.0351
adenine	A’	136.0623
guanine	G’	152.0572

RNA marker ions of the nucleotides

marker	symbol	calculated *m/z*
cytidine*	C^0	306.0491
uridine*	U^0	307.0331
adenosine*	A^0	330.0603
guanosine*	G^0	346.0553

Additional RNA ions

formula	symbol	calculated *m/z*
[U –H_2O –U’]	[U^0 –U’]	195.0058
[C –H_3PO_4]	$C^{0\text{-p}}$	226.0828
[U –H_3PO_4]	$U^{0\text{-p}}$	227.0667
[U’ +152]		265.0289**
[A]		348.0709
[U +152 –H_3PO_4]	[$U^{0\text{-p}}$ +152]	379.0606**
[G –H_2O +HPO_3]	$G^{0+\text{p}}$	426.0216
[U +152 –H_2O]	[U^0 +152]	459.0269**
[AX –H_2O –X’]	[AX^0 –X’]	524.0584
[GX –H_2O –X’]	[GX^0 –X’]	540.0533
[AU –H_3PO_4]	[$AU^{0\text{-p}}$]	556.1193
[AU –HPO_3]		574.1299

X: any nucleotide
* with loss of water
** assuming 151.9938 for 152 Da adduct

B.1.3 RNA adducts of peptides and their sequence ions

RNA adducts of intact peptides or their fragment ions are given in an abbreviated code described in Table B.2. Loss of metaphosphoric acid HPO_3 is indicated with "-p". The *m/z* values of peptide–RNA adducts are underlined.

Table B.2: Annotation of peptide–RNA adducts

adduct	calculated adduct mass	abbreviated annotation	amino acid	example
[C_3O]	51.9949	$y5^{\#}$	K	B.27, p. 163
			Y	B.55, p. 175
[U' –H_2O]	94.0167	$y5^{U'0}$	F	B.56, p. 175
[U']	112.0273	$y5^{U'}$	I	B.37, p. 168
			K	B.27, p. 163
			Y	B.32, p. 165
[U –H_3PO_4]	226.0590	$y5^{U0\text{-}p}$	T	B.19, p. 160
			F	B.50, p. 173
[U' +152]	264.0211*	$y5^{\#}$	C	B.24, p. 162
[U –H_2O]	306.0253	$y5^{U0}$	C	B.20, p. 160
			H	B.23, p. 161
			F	B.50, p. 173
			Y	B.48, p. 172
			W	B.16, p. 159
[4SU –H_2S]	306.0253	$y5^{\#}$	K	3.7, p. 66
			G	3.9, p. 68
[U]	324.0359	$y5^{U}$	F	3.13, p. 78
[U +152 –H_3PO_4]	378.0528	$y5^{\#}$	C	B.5, p. 154
			W	3.17, p. 82
[U +152 –HPO_3]	396.0633	$y5^{\#}$	C	B.9, p. 155
[U +152 –H_2O]	458.0191*	$y5^{\#}$	C	B.72, p. 182
			W	3.17, p. 82
[U +152]	476.0297*	$y5^{\#}$	C	B.4, p. 153
			W	3.17, p. 82
[AU –H_2O –A']	500.0233	$y5^{\#}$	C	B.20, p. 160
			W	B.42, p. 170
[XU –H_2O –X']	500.0233	$y5^{\#}$		

* assuming 151.9938 for 152 Da adduct

In some cases, immonium or other internal ions of single amino acids are observed as shifted by the cross-linked RNA or fragments thereof. Table B.3 summarizes the observed ions and gives the calculated m/z values of the respective ions.

Table B.3: RNA-adducts of single amino acids

adduct	symbol	calculated m/z	example
IM F + [U' $-H_2O$]	IM $F^{U'0}$	214.0980	B.56, p. 175
IM Y + [U']	IM $Y^{U'}$	248.1035	B.32, p. 165
IM C + [U $-H_2O$]	IM C^{U0}	382.0474	B.58, p. 176
IM H + [U $-H_2O$]	IM H^{U0}	416.0971	B.23, p. 161
IM F + [U $-H_2O$]	IM F^{U0}	426.1066	B.66, p. 180
(IM W $-CHNH_2$) + [U $-H_2O$]	(IM W $-CHNH_2)^{U0}$	436.0910	B.16, p. 159
IM W + [U $-H_2O$]	IM W^{U0}	465.1175	B.16, p. 159
IM C + [U +152 $-H_2O$]	IM $C^{\#}$	534.0412*	B.31, p. 165
IM C + [U +152]	IM $C^{\#}$	552.0518*	B.5, p. 154
IM W + [AU $-H_2O$ –A']	IM $W^{\#}$	659.1155	B.42, p. 170

* assuming 151.9938 for 152 Da adduct

B.2 MS/MS fragment spectra of NusB–S10 peptides observed as adducts with 258 Da

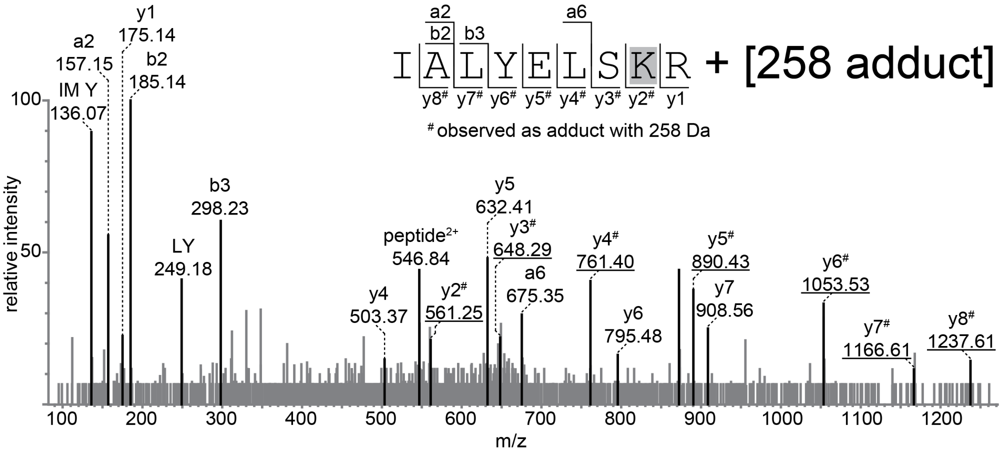


Figure B.1: MS/MS fragment spectrum (smoothed and centroided) of NusB peptide IALYELSKR (I87–R95) observed as adduct with 258 Da. All y-ions except for y1 are observed as partially shifted by the adduct mass, 258 Da. This leads to the conclusion that K94 was the amino acid that reacted to yield the 258 Da adduct.

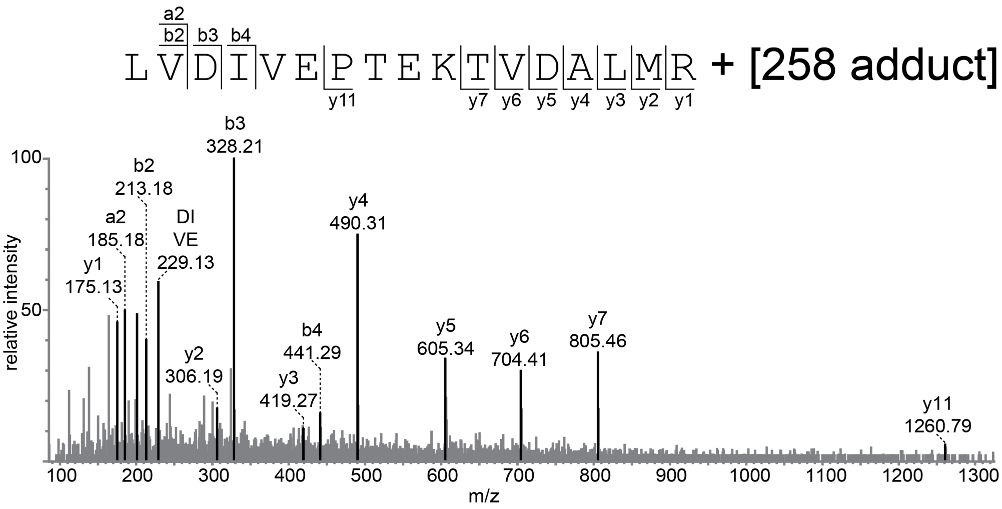


Figure B.2: MS/MS fragment spectrum (smoothed and centroided) of S10 peptide LVDIVEPTEKTVDALMR (L73–R89) observed as adduct with 258 Da. The spectrum does not contain any hint of the adduct mass, no shifted peptide fragments are observed.

B.3 MS/MS fragment spectra of cross-links from the ASH1 complexes

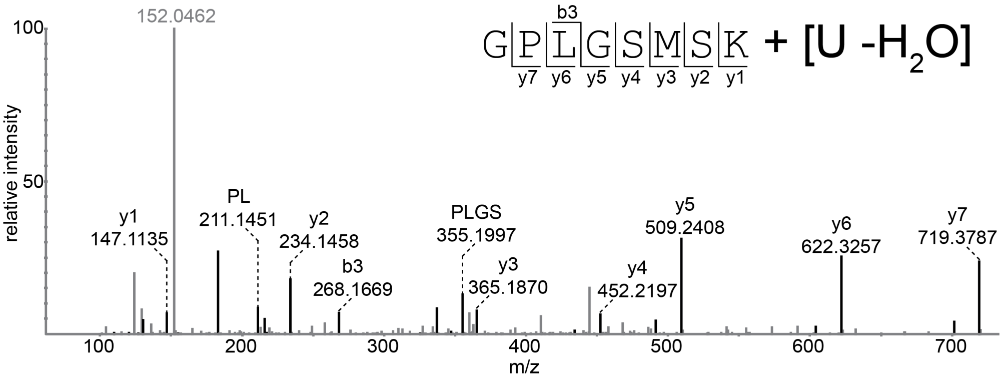


Figure B.3: MS/MS fragment spectrum of She2p peptide GPLGSMSK (M1–K3) cross-linked to [U $-H_2O$]. The underlined part of the peptide sequence is not part of She2p but remains after cleavage of a GST tag. The spectrum does not contain any hint towards the cross-linked amino acid.

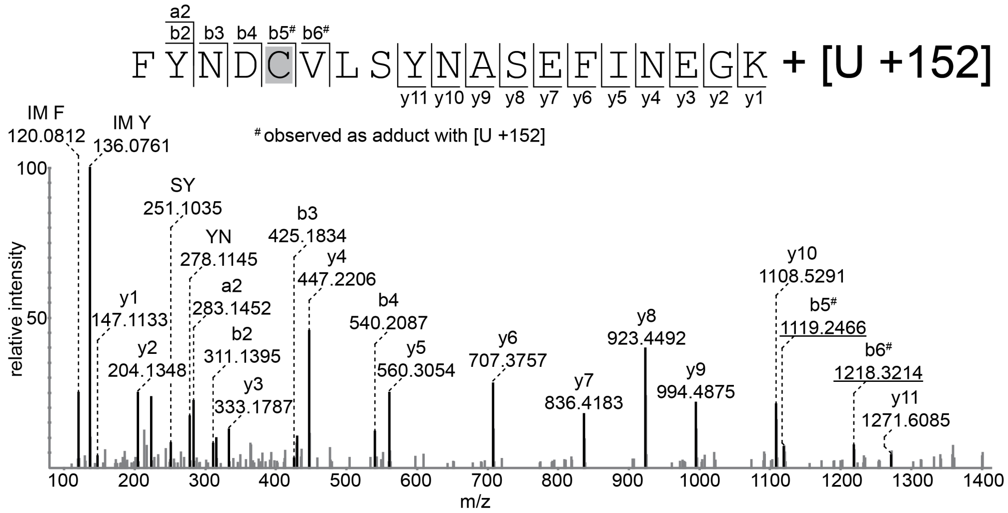


Figure B.4: MS/MS fragment spectrum of She2p peptide FYNDCVLSYNASEFINEGK (F64–K82) cross-linked to [U +152]. The shift of b5 and b6 by [U +152] as well as the presence of a 152 Da adduct in general identify C68 as the cross-linked amino acid.

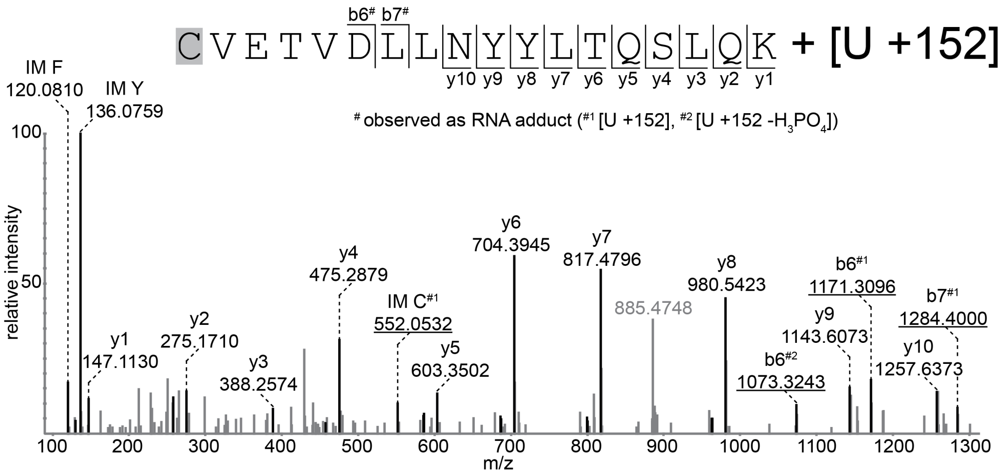


Figure B.5: MS/MS fragment spectrum of She2p peptide CVETFDLLNYYLTQSLQK (C106–K123) cross-linked to [U +152]. The observation of the immonium ion as adduct with [U +152] as well as the presence of a 152 Da adduct in general identify C103 as the cross-linked amino acid.

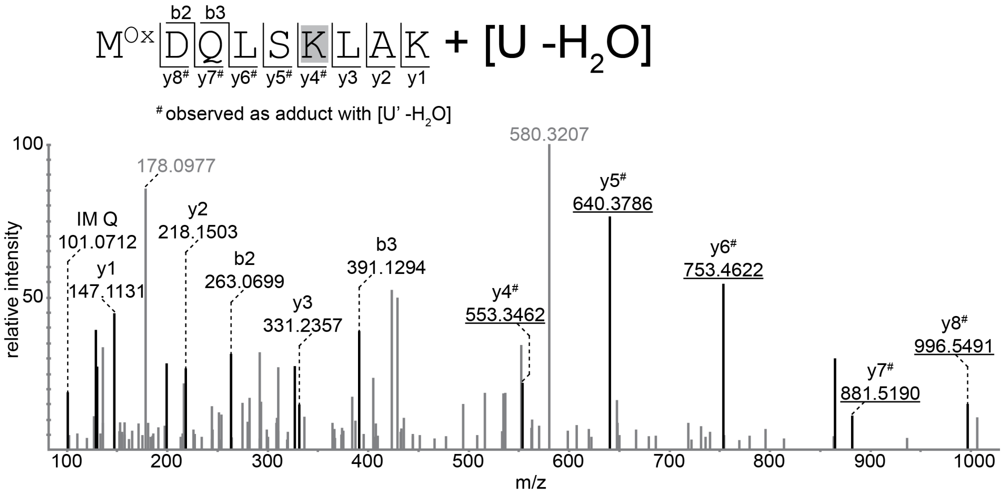


Figure B.6: MS/MS fragment spectrum of She3p peptide MDQLSKLAK (M130–K138), oxidized at M130, cross-linked to [U $-H_2O$]. The complete shift of the y-series by [U' $-H_2O$] starting with y4 identifies K135 as the cross-linked amino acid.

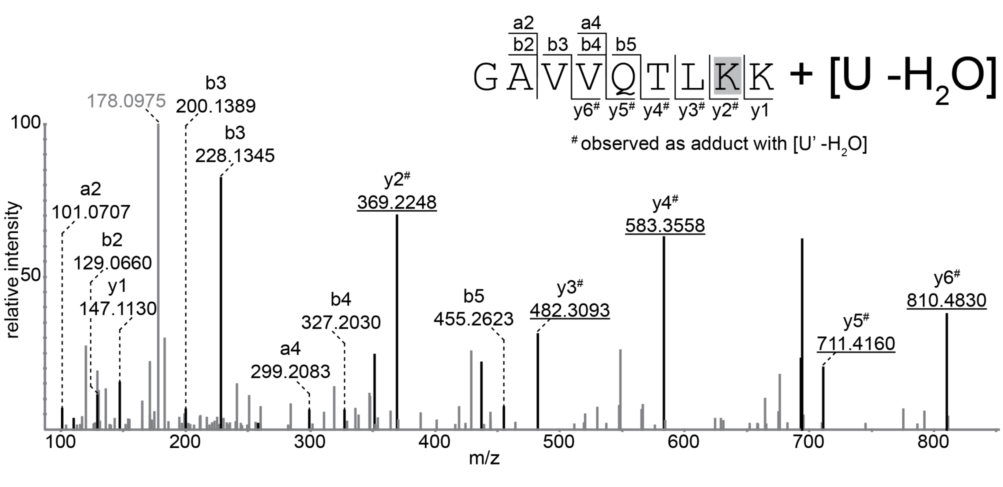


Figure B.7: MS/MS fragment spectrum of She3p peptide GAVVQTLKK (G283–K291) cross-linked to [U $-H_2O$]. Peptide fragments y2 to y6 are completely shifted by [U' $-H_2O$], pointing at K290 as the cross-linked amino acid.

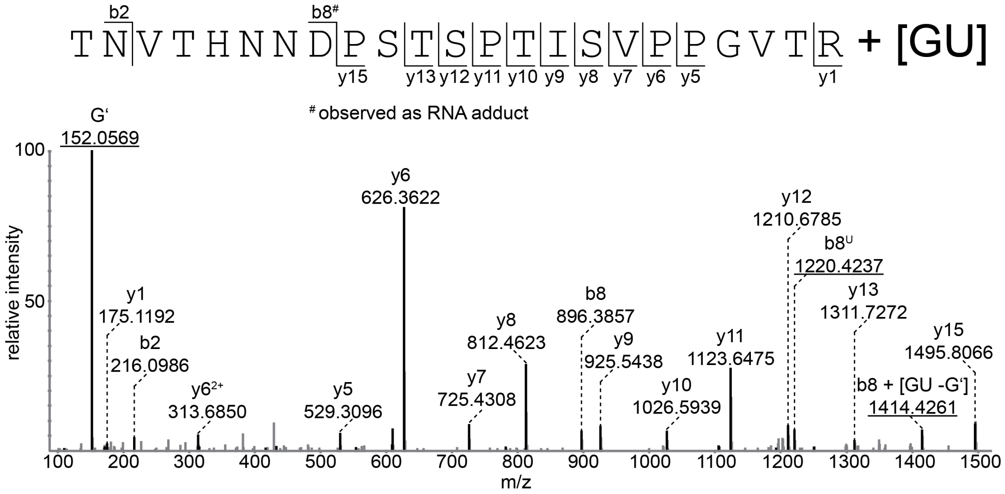


Figure B.8: MS/MS fragment spectrum of She3p peptide TNVTHNNDPSTSPTISVPPGVTR (T383–R405) cross-linked to [GU]. An intense RNA marker ion for guanine is observed. RNA adducts of b8 place the cross-link on the N-terminal part of the peptide, the exact cross-linking site cannot be determined.

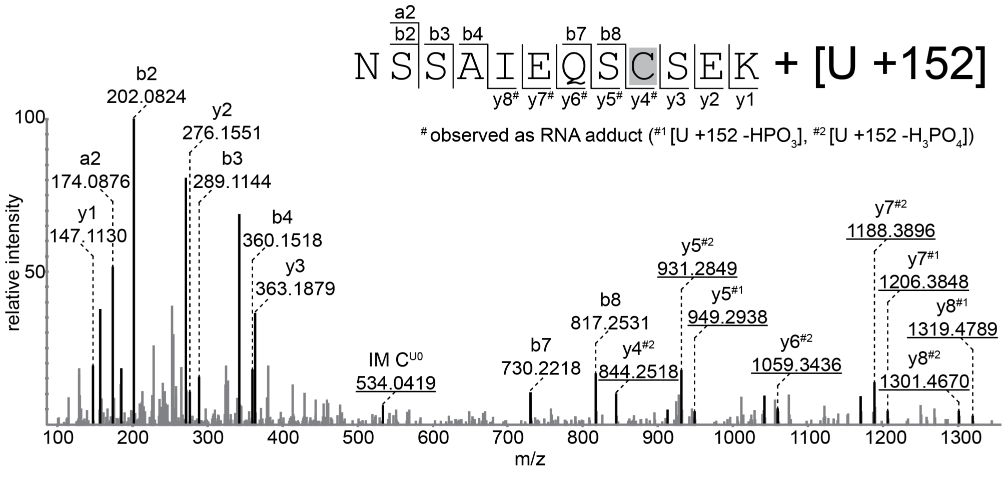


Figure B.9: MS/MS fragment spectrum of She3p peptide NSSAIEQSCSEK (N139–K150) cross-linked to [U +152]. Several observations identify C147 as the cross-linked amino acid: The immonium ion of cystein is observed as adduct with [U $-H_2O$] and all y-ions containing C147 are shifted by the cross-linked RNA. In addition, the 152 Da adduct in general points at cysteine as the cross-linked amino acid.

B.4 MS/MS fragment spectra of cross-links from Cwc2

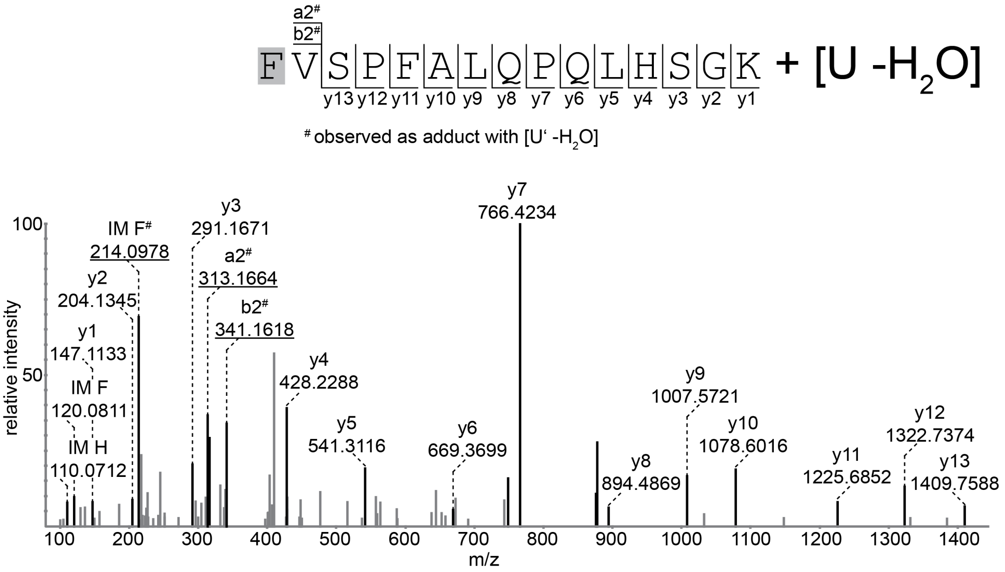


Figure B.10: MS/MS fragment spectrum of Cwc2 peptide FVSPFALQPQLHSGK (F47–K61) cross-linked to [U $-H_2O$]. F47 was identified as the cross-linked amino acid due to observation of its immonium ion as an adduct with [U' $-H_2O$].

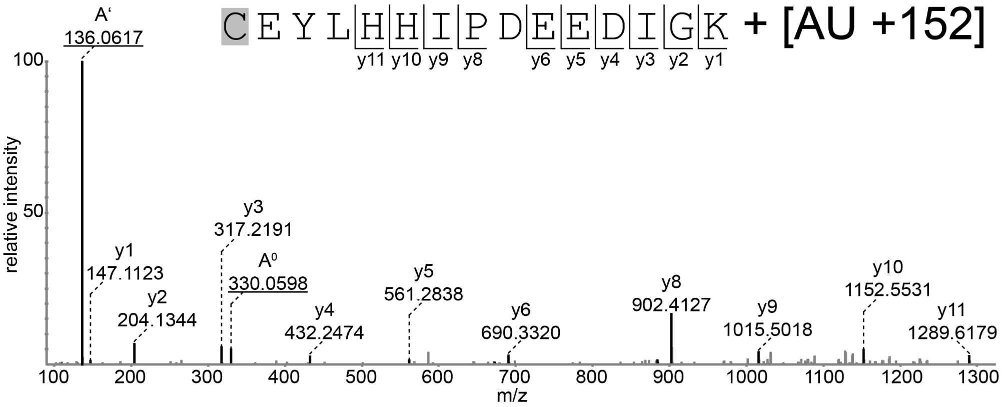


Figure B.11: MS/MS fragment spectrum of Cwc2 peptide CEYLHHIPDEEDIGK (C87–K101) cross-linked to [AU +152]. The spectrum is dominated by the adenine marker ions, peptide fragments are suppressed. The observation of the 152 Da adduct suggests C87 as the cross-linked amino acid. The absence of any a- or b-ions confirms this assumption.

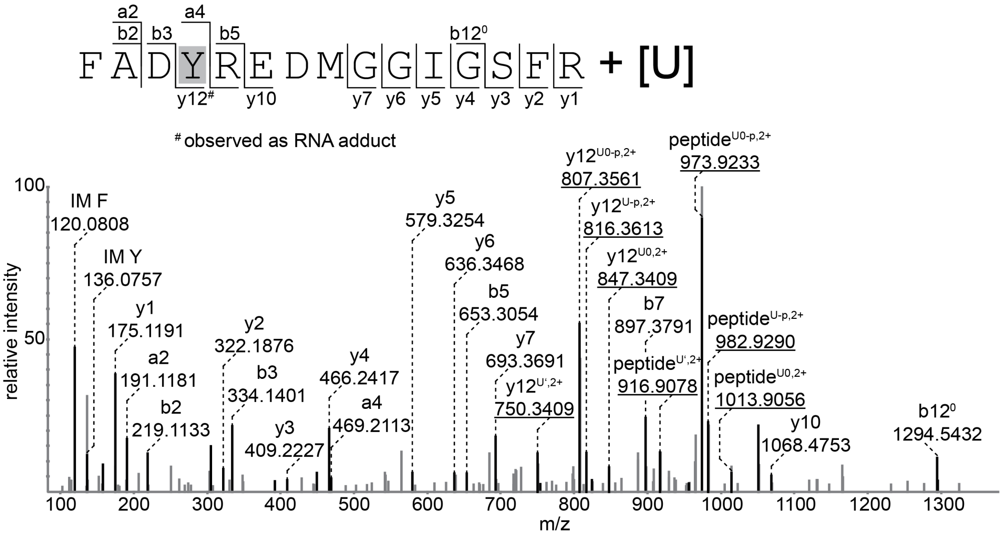


Figure B.12: MS/MS fragment spectrum of Cwc2 peptide FADYREDMGGIGSFR (F117–R131) cross-linked to [U]. y12 is observed as an adduct with several fragments of [U], therefore Y120 is identified as the cross-linked amino acid.

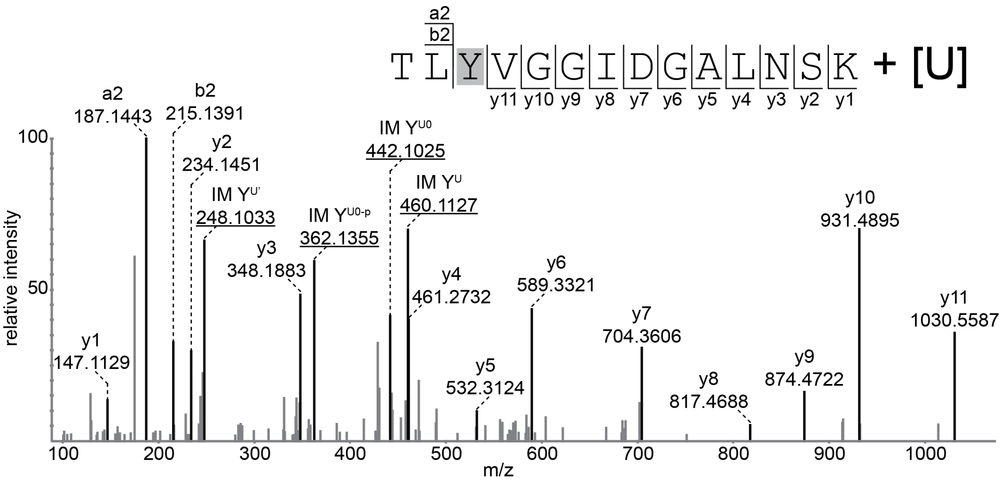


Figure B.13: MS/MS fragment spectrum of Cwc2 peptide TLYVGGIDGALNSK (T136–K149) cross-linked to [U]. Adducts of the immonium ion of tyrosine with [U] and several of its fragments unambiguously identify Y138 as the cross-linked amino acid.

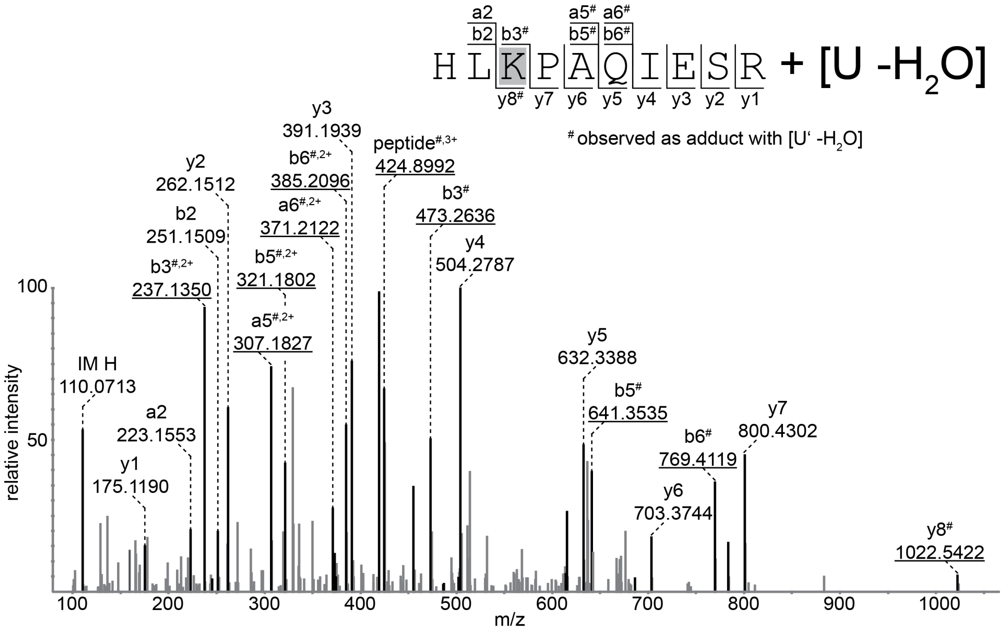


Figure B.14: MS/MS fragment spectrum of Cwc2 peptide HLKPAQIESR (H150–R159) cross-linked to [U $-H_2O$]. All observed peptide fragments containing K152, i.e. b-ions starting with b3 and y8, are shifted by [U' $-H_2O$]. Therefore, K152 is the cross-linked amino acid.

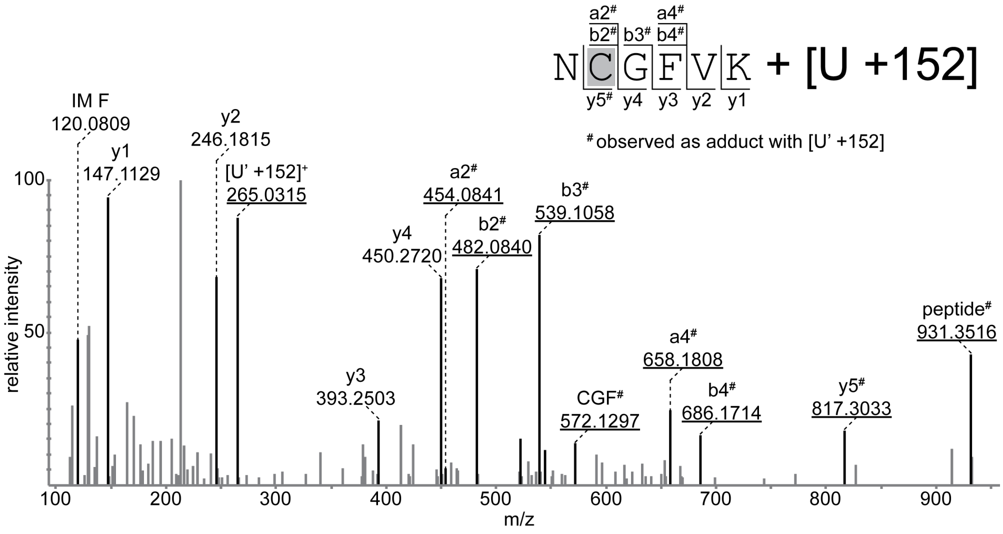


Figure B.15: MS/MS fragment spectrum of Cwc2 peptide NCGFVK (N180–K185) cross-linked to [U +152]. The observation of the overall mass adduct of 152 Da as well as all peptide fragments containing C181 shifted by [U' +152] point at C181 as the cross-linked amino acid.

B.5 MS/MS fragment spectra of cross-links from yeast after TAP tag isolation

B.5.1 Cross-links of the 40S small ribosomal subunit

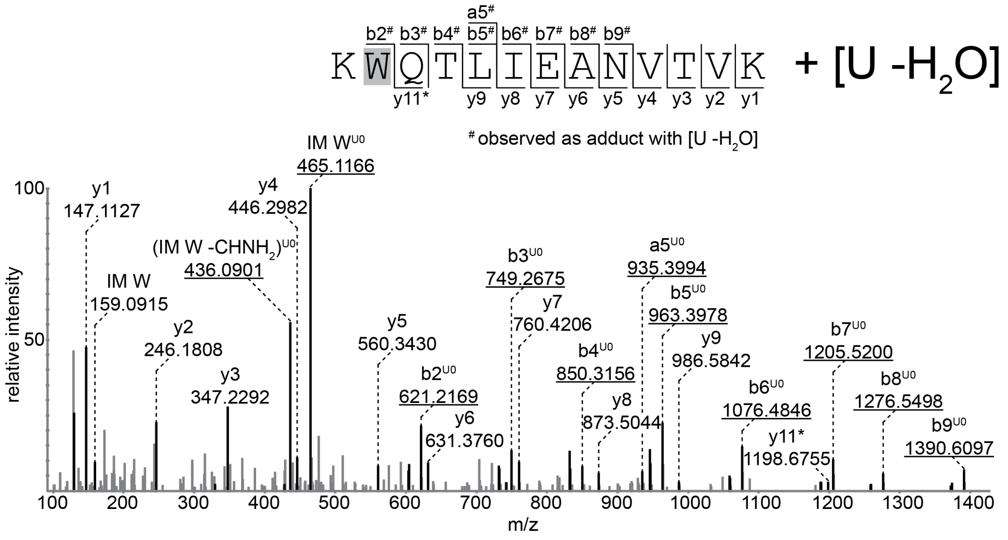


Figure B.16: MS/MS fragment spectrum of 40S ribosomal protein S1-A/-B peptide KWQTLIEANVTVK (K116–K128) cross-linked to [U $-H_2O$]. Intense signals of the tryptophan immonium ion and one of its internal fragments shifted by the cross-linked RNA [U $-H_2O$] as well as the shifted b-ion-series pinpoint to W117 as the cross-linked amino acid.

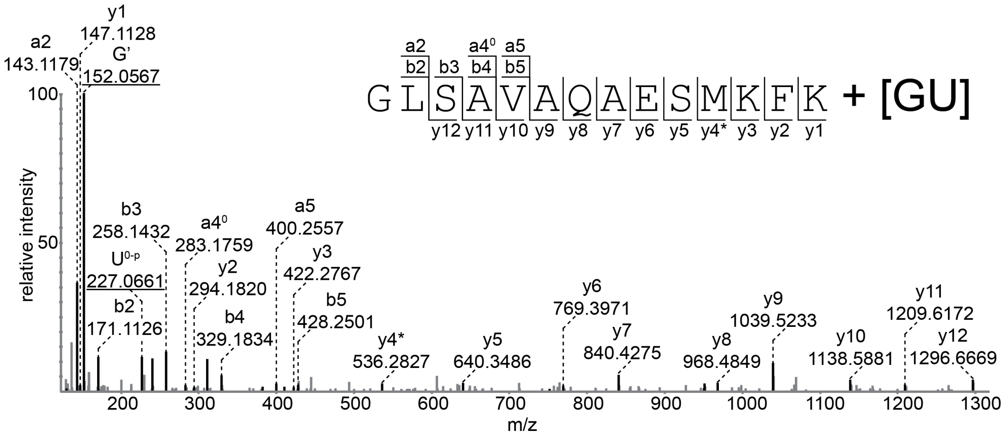


Figure B.17: MS/MS fragment spectrum of 40S ribosomal protein S3 peptide GLSAVAQAESMKFK (G95–K108) cross-linked to [GU]. A dominant G' marker ion is observed, as well as a weaker ion for [U $-H_3PO_4$]. However, no RNA adducts of peptide fragments are observed that would allow the identification of the cross-linked amino acid.

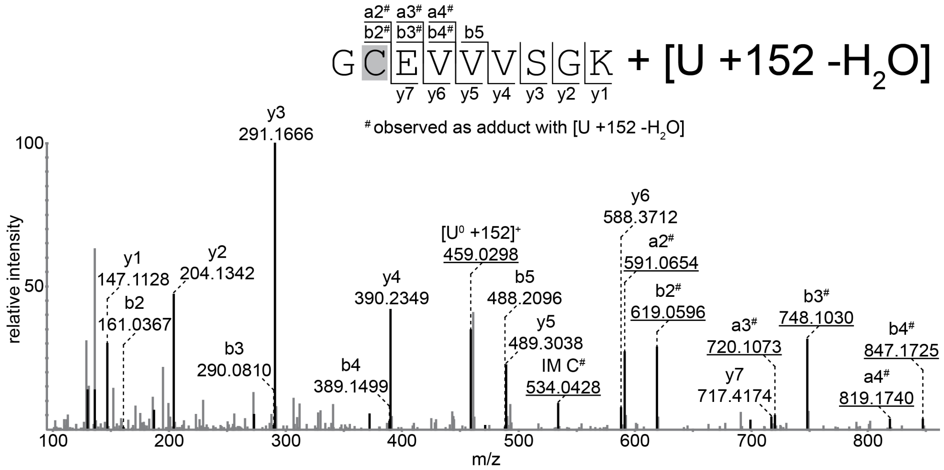


Figure B.18: MS/MS fragment spectrum of 40S ribosomal protein S3 peptide GCEVVVSGK (G133–K141) cross-linked to [U +152 $-H_2O$]. The immonium ion of cysteine as well as the majority of a- and b-ions are observed as shifted by [U +152 $-H_2O$], thus it can be concluded that C134 is the cross-linked amino acid.

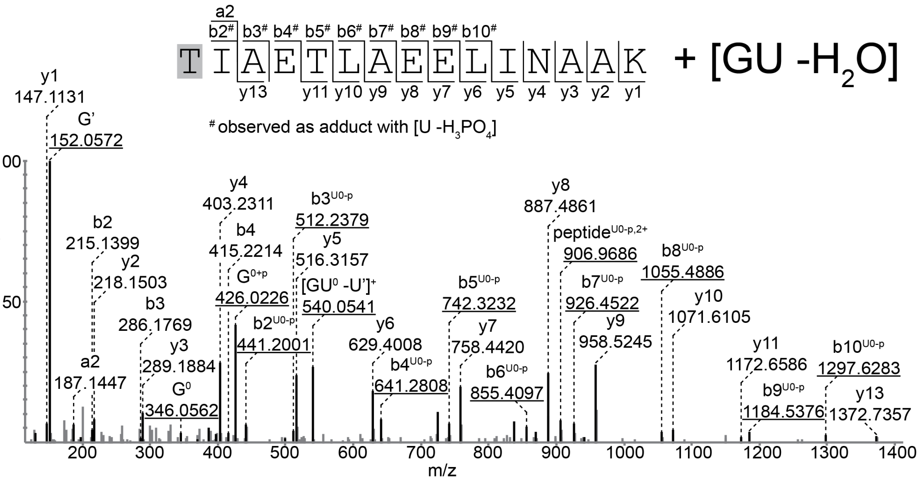


Figure B.19: MS/MS fragment spectrum of 40S ribosomal protein S5 peptide TIAETLAEELINAAK (T189–K203) cross-linked to [GU $-H_2O$]. The spectrom is dominated by the G' marker ion, all peptide fragment ions are supressed below 25%. b-ions 2–10 are shifted by [U $-H_3PO_4$]. Therefore, either T189 or I190 could be the cross-linked amino acid. As isoleucine is expected to be much less reactive in UV cross-linking, T189 is likely to be cross-linked.

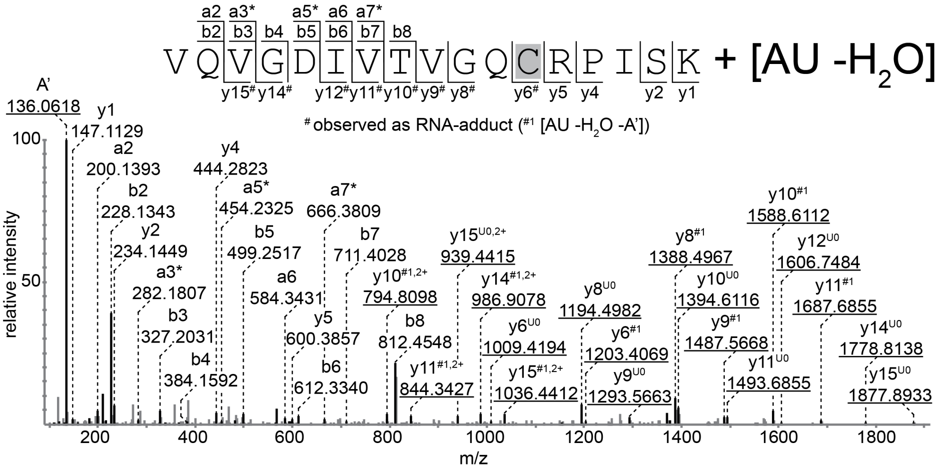


Figure B.20: MS/MS fragment spectrum of 40S ribosomal protein S11-A/-B peptide VQVGDIVTVGQCRPISK (V117–K133) cross-linked to [AU $-H_2O$]. An intense A' marker ion is observed with all peptide fragment ions below 40% relativ intensity. All y-ions, starting with y6, are shifted by RNA fragments, identifying C128 as the cross-linked amino acid.

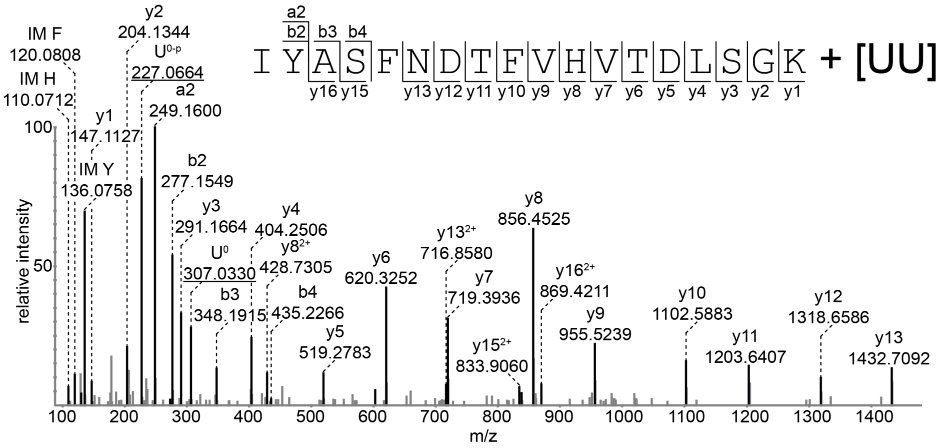


Figure B.21: MS/MS fragment spectrum of 40S ribosomal protein S14-A/-B peptide IYASFNDTFVHVTDLSGK (A: I19–K36, B: I20–K37) cross-linked to [UU]. RNA-marker ions for [U $-H_3PO_4$] and [U $-H_2O$] are observed. However, the cross-linked amino acid cannot be identified.

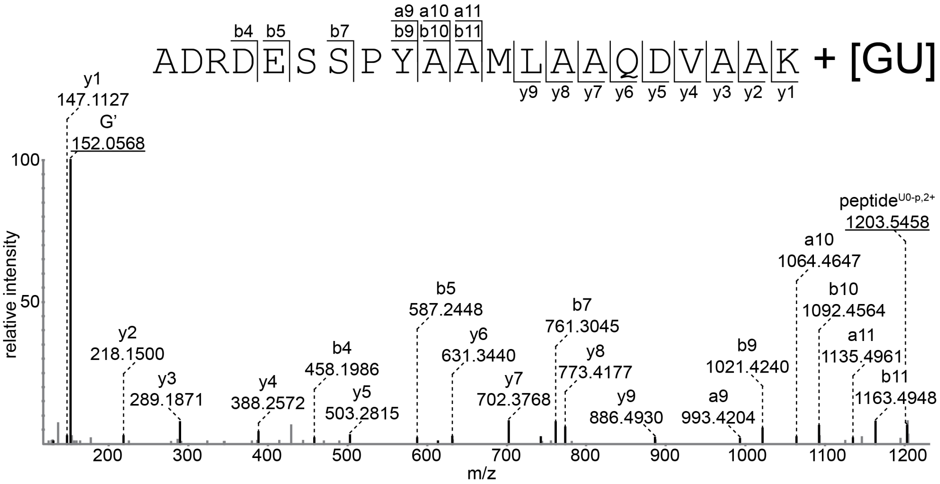


Figure B.22: MS/MS fragment spectrum of 40S ribosomal protein S14-A/-B peptide ADRDESSPYAAMLAAQDVAAK (A: A50–K70, B: A51–K71) cross-linked to [GU]. The spectrum is dominated by the G' marker, the peptide fragments are supressed below 10% relative intensity. The actual cross-linked amino acid residue cannot be identified.

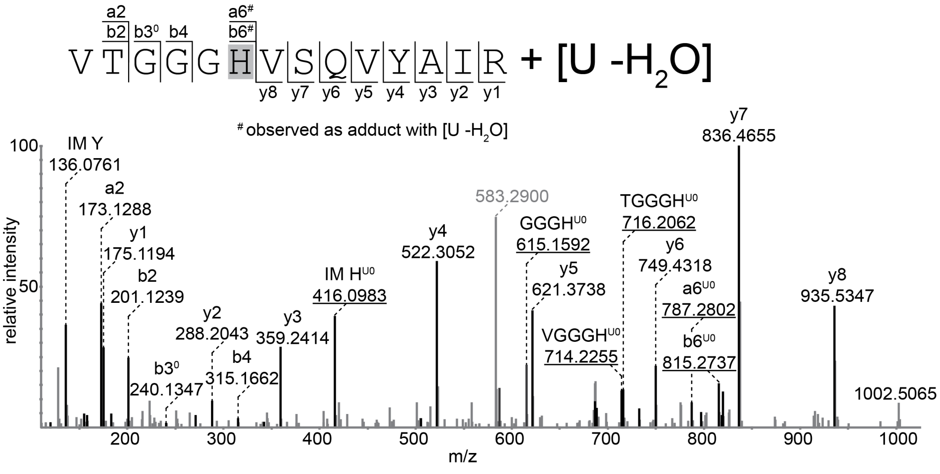


Figure B.23: MS/MS fragment spectrum of 40S ribosomal protein S16-A/-B peptide VTGGGHVSQVYAIR (V69–R82) cross-linked to [U $-H_2O$]. The immonium ion of histidine and the a6/b6 ion pair is shifted by the mass of the cross-linked RNA, as are a number of internal ions. This points at H74 as the cross-linked amino acid.

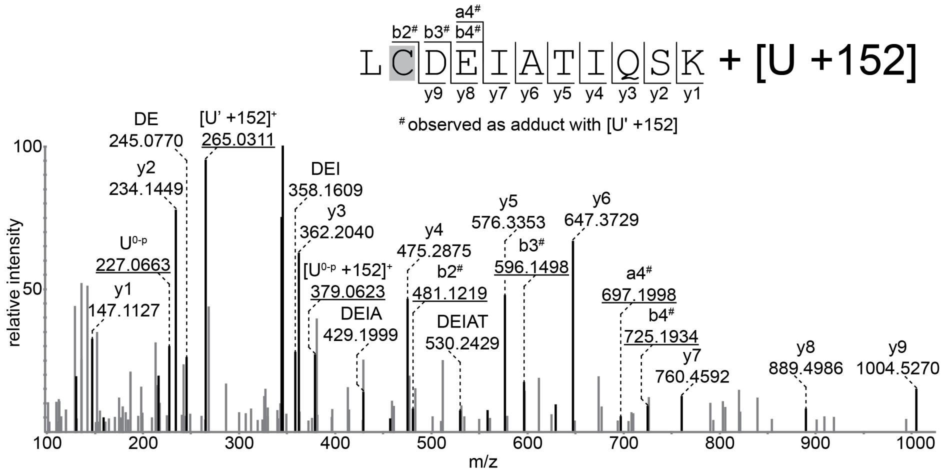


Figure B.24: MS/MS fragment spectrum of 40S ribosomal protein S17-A/-B peptide LCDEIATIQSK (L34–K44) cross-linked to [U +152]. All peptide fragments containing the cross-linked C35 residue are shifted by [U' +152]. In addition, several internal ions are observed, which originate from cleavage C-terminal to the cross-linked cysteine. Finally, several RNA-signals are observed, namely signals corresponding to [U $-H_3PO_4$], [U' +152], and [U +152 $-H_3PO_4$].

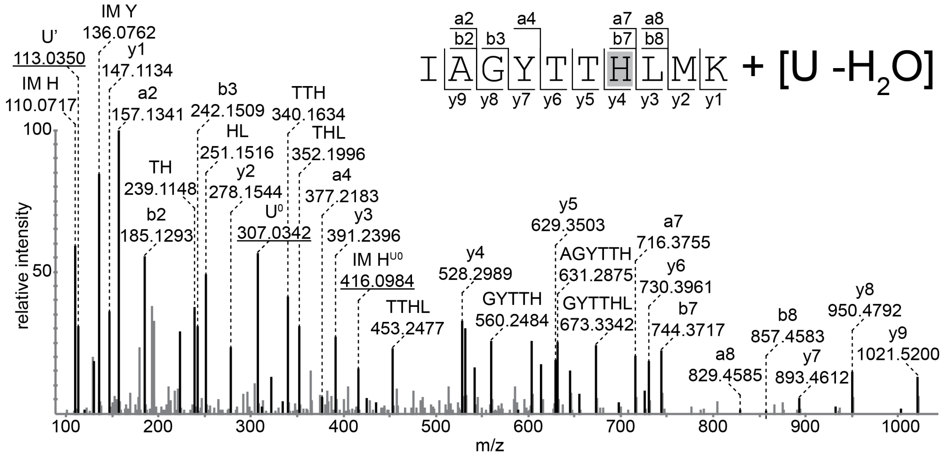


Figure B.25: MS/MS fragment spectrum of 40S ribosomal protein S17-A/-B peptide IAGYTTHLMK (I50–K59) cross-linked to [U $-H_2O$]. The only peptide fragment shifted by RNA is the immonium ion of histidine, thus identifying H56 as the cross-linked amino acid. The observation of U' and U^0 marker are unusual for a cross-link to a single U nucleotide, as well as the large number of internal ions.

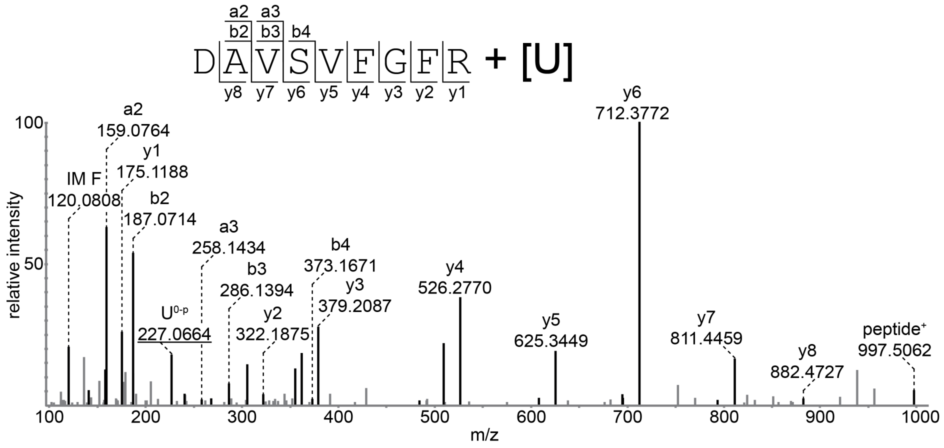


Figure B.26: MS/MS fragment spectrum of 40S ribosomal protein S24-A/-B peptide DAVSVFGFR (D53–R61) cross-linked to [U]. An RNA marker for [U $-H_3PO_4$] was observed but no adduct that would allow to derive the cross-linked amino acid.

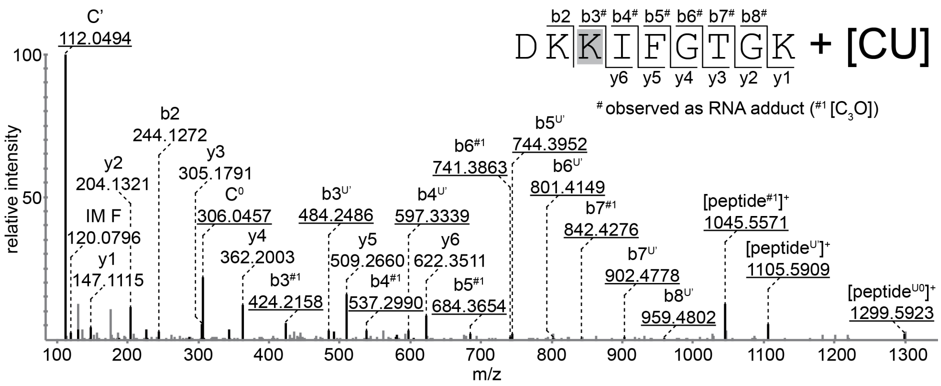


Figure B.27: MS/MS fragment spectrum of 40S ribosomal protein S24-A/-B peptide DKKIFGTGK (D115–K123) cross-linked to [CU]. The peptide is identified by a full b-ion series shifted by [C_3O] and [U']; the y-series is complete until the cross-linked amino acid, K117. RNA markers of C' and C^0 are clearly observable, the peptide fragments are below 15% relative intensity.

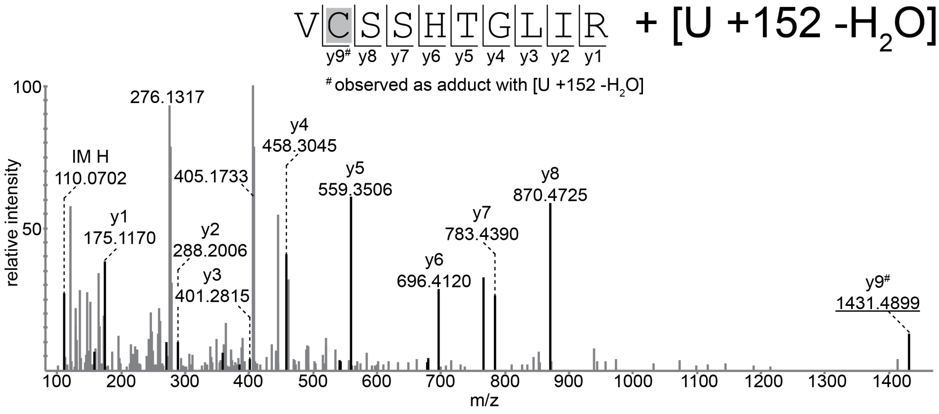


Figure B.28: MS/MS fragment spectrum of 40S ribosomal protein S29-A peptide VCSSHTGLIR (V23–R32) cross-linked to [U +152 –H_2O]. Additional description see Figure B.29.

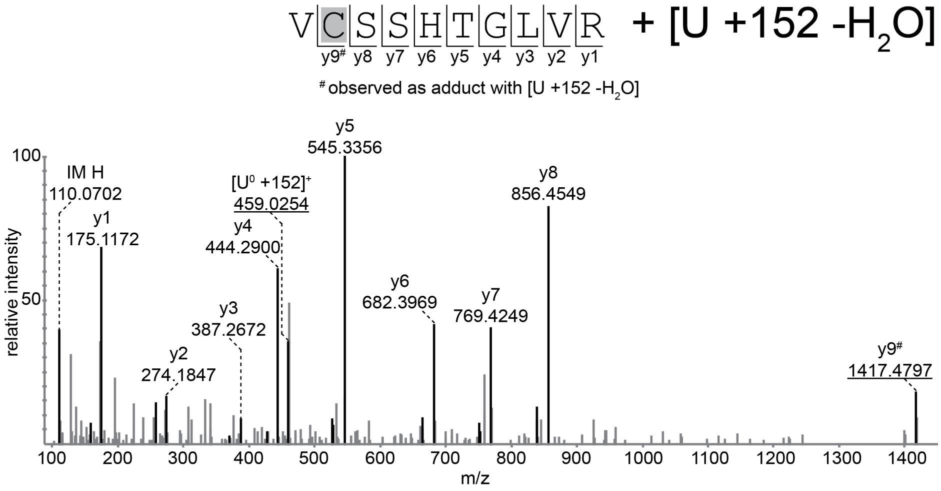


Figure B.29: MS/MS fragment spectrum of 40S ribosomal protein S29-B peptide VCSSHTGLVR (V23–R32) cross-linked to [U +152 –H_2O]. S29-A peptide VCSSHTGLIR (spectrum shown in B.28) and S29-B peptide VCSSHTGLVR (position V23–R32 in both proteins) differ in position 31, S29-A containing isoleucine and S29-B valine. Therefore, both peptides have a different mass and also different y-ion-series which allows a confident discrimination of both protein forms. Both spectra contain a y9 ion shifted by [U +152 –H_2O], confirming C24 as the cross-linked amino acid. In both cases, no a- or b-ions are observed, possibly due to the cross-link to the peptides' second position.

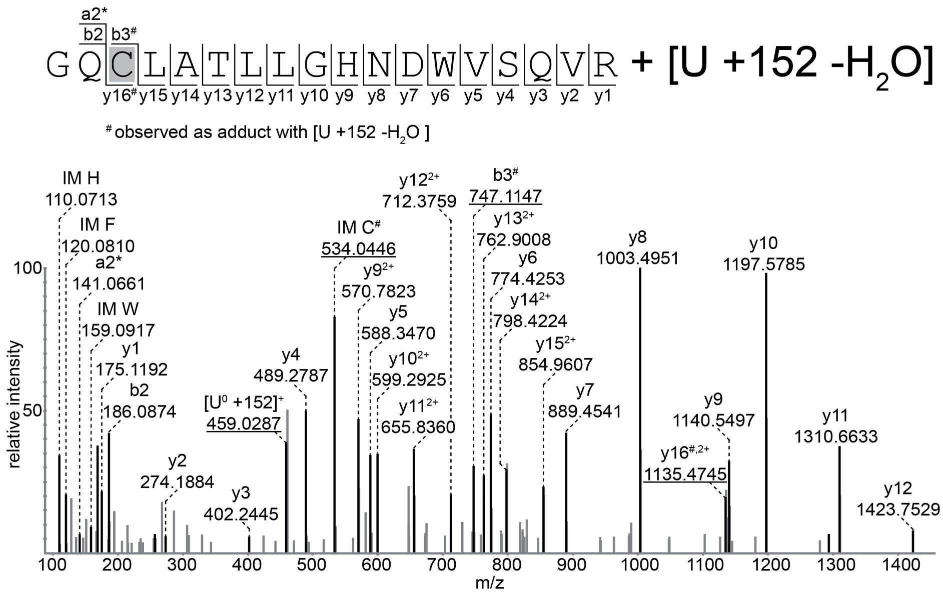


Figure B.30: MS/MS fragment spectrum of guanine nucleotide-binding protein subunit beta-like protein (Rack1) peptide GQCLATLLGHNDWVSQVR (G138–R155) cross-linked to [U +152 –H_2O]. C140 is identified as the cross-linked amino acid by the immonium ion, b3 and y16 shifted by [U +152 –H_2O]. The peptide spans between the WD repeats 3 (aa 105–145) and 4 (aa 147–191) of the protein.

B.5.2 Cross-links of the 60S large ribosomal subunit

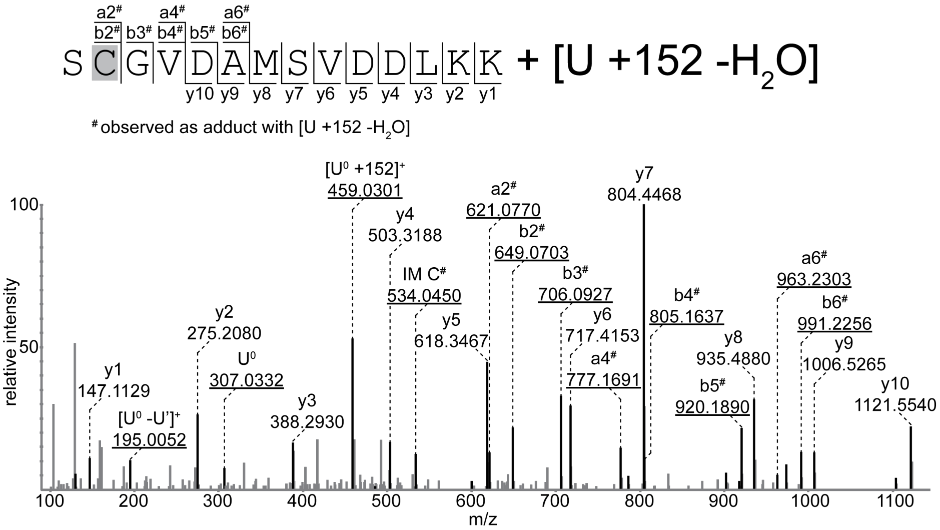


Figure B.31: MS/MS fragment spectrum of 60S ribosomal protein L1-A/-B peptide SCGVDAMSVDDLKK (S79–K92) cross-linked to [U +152 –H_2O]. Observed a- and b-ions are completely shifted by the mass of the cross-linked RNA, as is the immonium ion of cysteine. Thus it can be concluded that C80 is the cross-linked amino acid. Interestingly, beside an ion corresponding to the cross-linked RNA, signals of uridine are observed that result from cleavage from the 152 adduct.

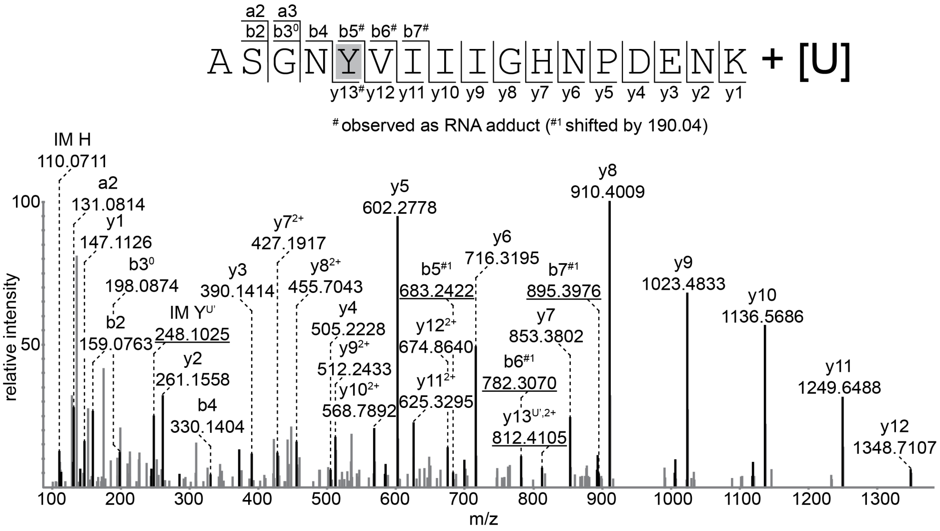


Figure B.32: MS/MS fragment spectrum of 60S ribosomal protein L2-A/-B peptide ASGNYVIIIGHNPDENK (A129–K145) cross-linked to [U]. Y133 is the actual cross-linked amino acid as is shown by its immonium ion observed as adduct with [U'] and a shift of the corresponding b- and y-ions. The b-series is shifted by 190.04 Da, a mass that cannot be explained.

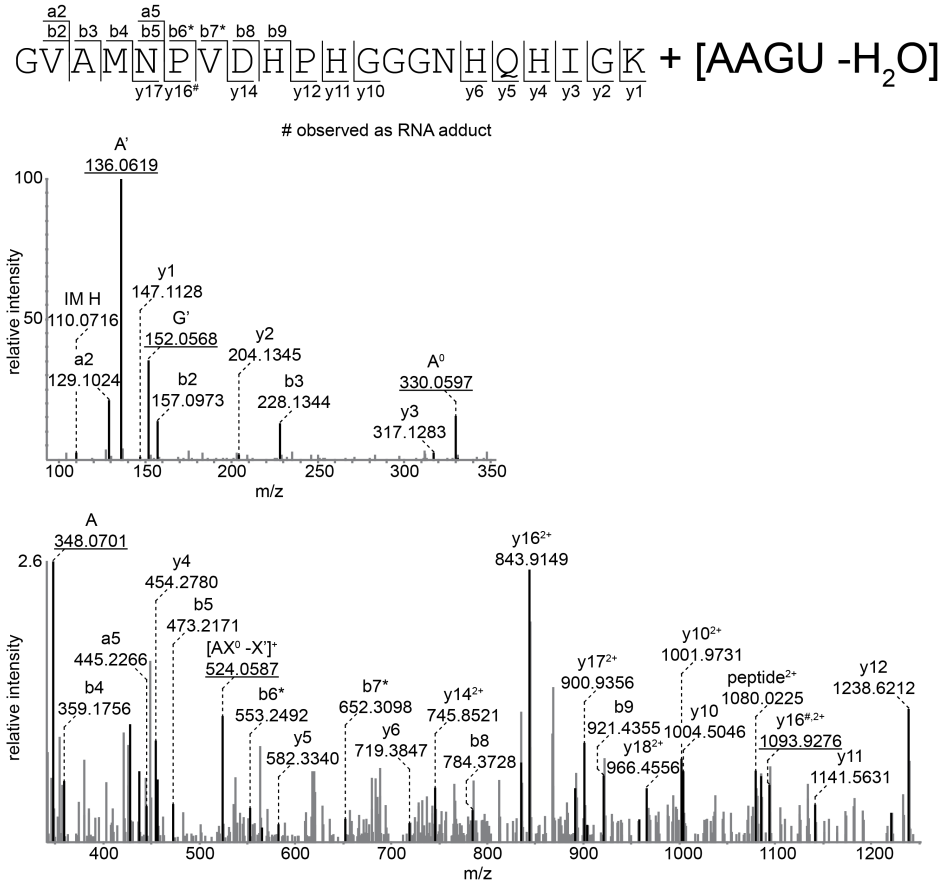


Figure B.33: MS/MS fragment spectrum of 60S ribosomal protein L2-A/-B peptide GVAMNPVDHPHGGGNHQHIGK (G201–K221) cross-linked to [AAGU $-H_2O$]. The spectrum is shown in two views due to the high complexity. The upper pane shows the mass range up to *m/z* 350, the lower pane shows the higher mass range where the maximum intensity is 2.6% relative to the signal of the A’ marker ion. While the fragments in the lower mass range (upper pane) show high intensity RNA marker ions and a few peptide fragments with reasonable intensity, all other fragments in the higher mass range are of low intensity (lower pane). Importantly, fragments resulting from cleavage N-terminal to proline are observed (y12, y16) and both peptide termini are covered by the corresponding sequence ions. The signal at 1093.9276 corresponds to a doubly charged RNA adduct of y16 with the composition [UX^0 –X’]; since the base of the second nucleotide is cleaved off, its nature cannot be determined. Since proline itself is thought to be rather unreactive towards UV cross-linking and no y15 is observed, no clear conclusion about the cross-linked amino acid can be drawn.

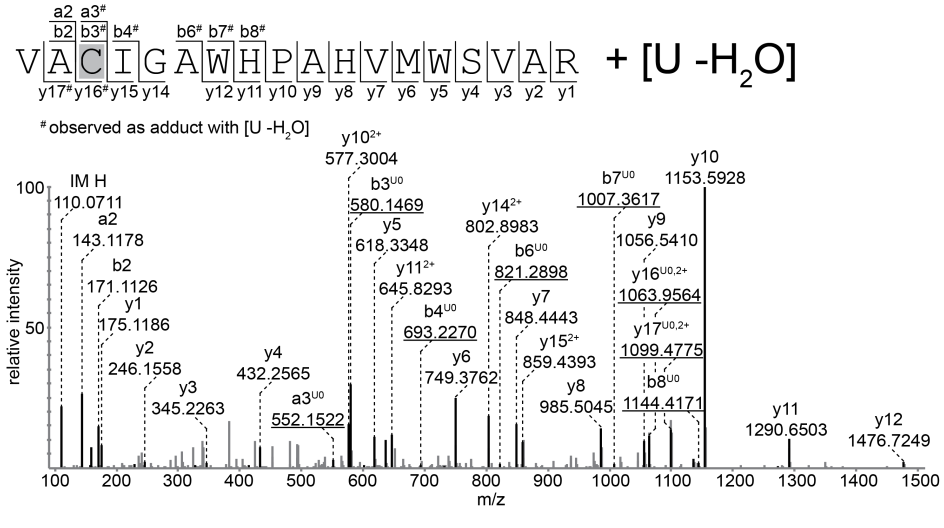


Figure B.34: MS/MS fragment spectrum of 60S ribosomal protein L3 peptide VACIGAWHPAHVMWSVAR (V249–R266) cross-linked to [U $-H_2O$]. All peptide fragments containing C251 are shifted by [U $-H_2O$], pinpointing the cross-link to this residue.

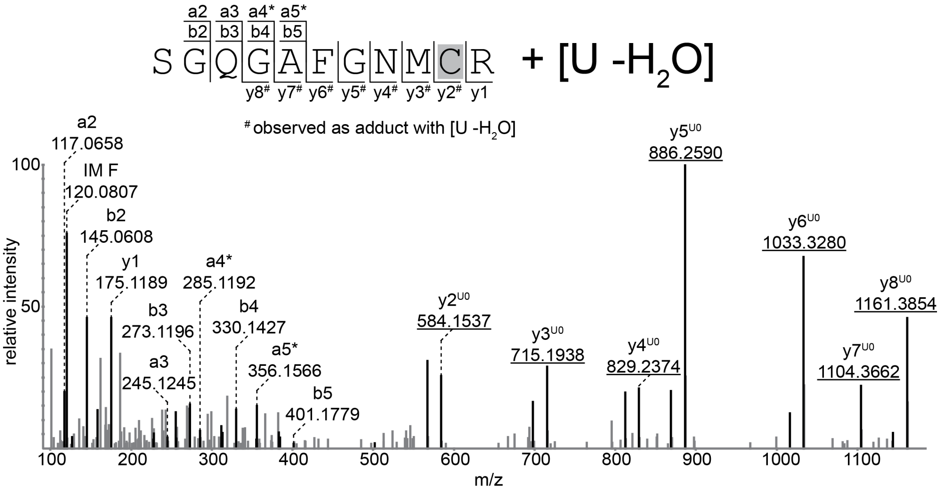


Figure B.35: MS/MS fragment spectrum of 60S ribosomal protein L4-A/-B peptide SGQGAFGNMCR (S85–R95) cross-linked to [U $-H_2O$]. The y-series is completely shifted by [U $-H_2O$] starting with y2, clearly identifying C94 as the cross-linked amino acid.

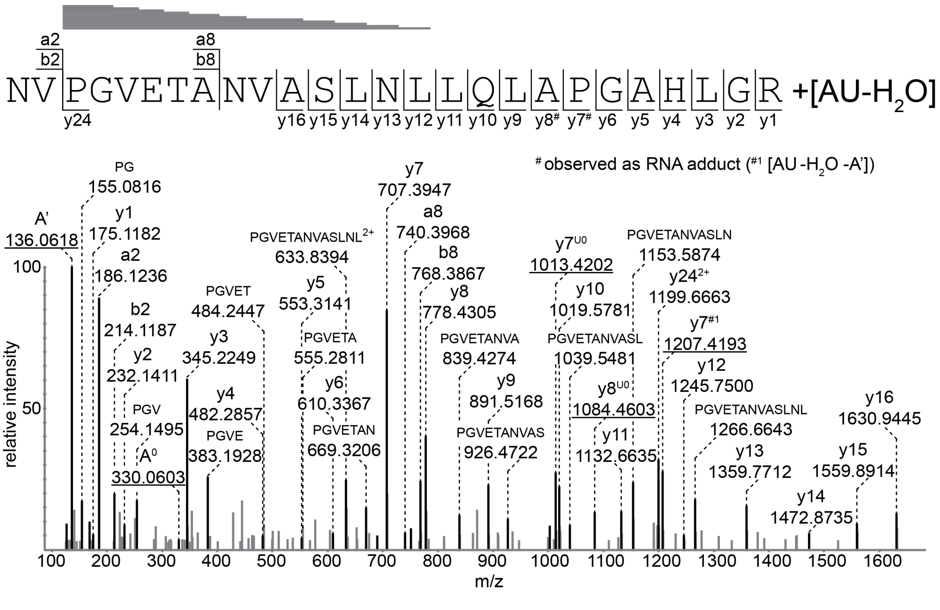


Figure B.36: MS/MS fragment spectrum of 60S ribosomal protein L4-A/-B peptide NVPGVETANVASLNLLQLAPGAHLGR (N221–R246) cross-linked to [AU $-H_2O$]. From cleavage N-terminal of P223 originate a great number of internal ions. The amino acid stretches contained in each of these internal fragments are indicated as individual lines above the peptide sequence and add confidence to the identification as only two a-/b-ion pairs are observed. y7 and y8 are partially shifted by RNA. However, this does not allow a clear conclusion about the cross-linked amino acid. The RNA-adducts are of much lower intensity than the original sequence ions, therefore RNA-adducts of smaller, less intense y-ions might be below the detection limit. The normally well observable immonium ion of histidine was not detected, which could be a consequence of a cross-linked H243. Overall, the observed shifts only allow the conclusion that the cross-link must be on the C-terminal part of the peptide.

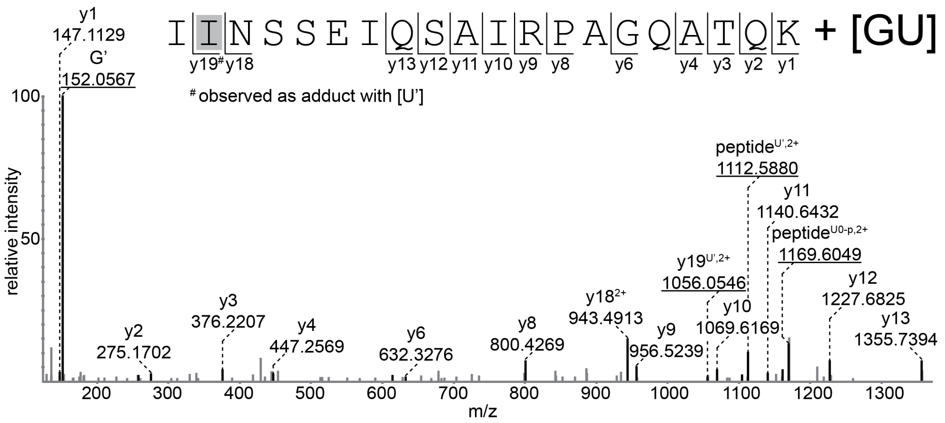


Figure B.37: MS/MS fragment spectrum of 60S ribosomal protein L4-A/-B peptide IINSSEIQSAIRPAGQATQK (I289–K308) cross-linked to [GU]. Interestingly, y19 is observed as an adduct with [U']. A cross-link at this position could also explain the absence of the usually well observable a2/b2 ion pair. Therefore, I290 is likely to be the cross-linked amino acid.

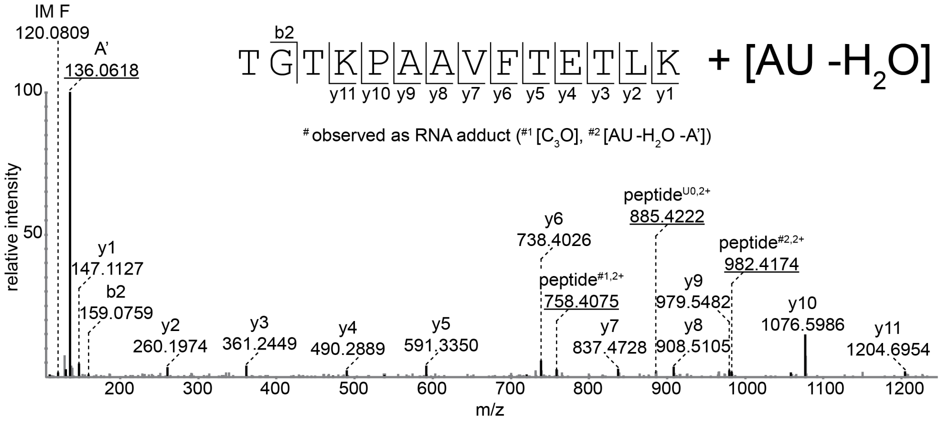


Figure B.38: MS/MS fragment spectrum of 60S ribosomal protein L4-A peptide TGTKPAAVFTETLK (T347–K360) cross-linked to [AU $-H_2O$]. Further description below Figure B.39.

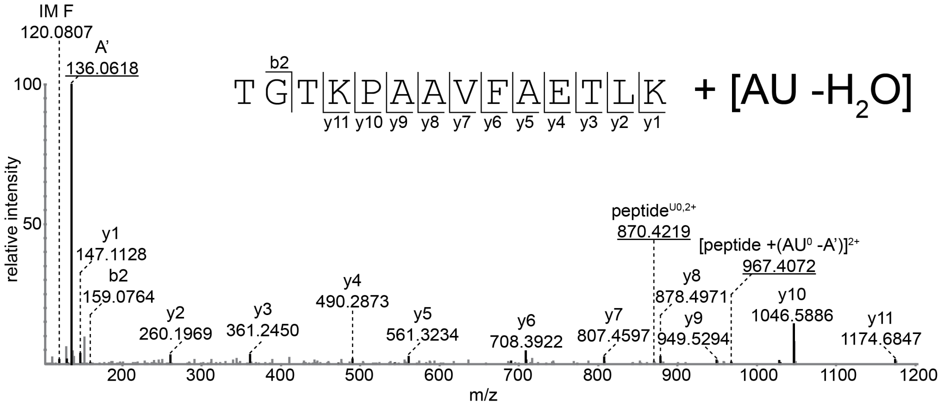


Figure B.39: MS/MS fragment spectrum of 60S ribosomal protein L4-B peptide TGTKPAAVFAETLK (T347–K360) cross-linked to [AU $-H_2O$]. The two cross-linked peptides of L4-A and -B (spectrum of L4-A shown in B.38) differ by only one amino acid, namely T356 (A-form) or A356 (B-form). Both peptides differ in their overall mass; also the different masses of the corresponding peptide sequence ions, i.e. y5 to y11, are clearly observable. Apart from this, both cross-link spectra share many similarities: Both are cross-links to the same RNA, [AU $-H_2O$]. The A' marker ion is dominant with the peptide fragments below 20% relative intensity. Finally, only RNA-adducts of the intact peptide are observed, therefore the cross-linked amino acid cannot be determined.

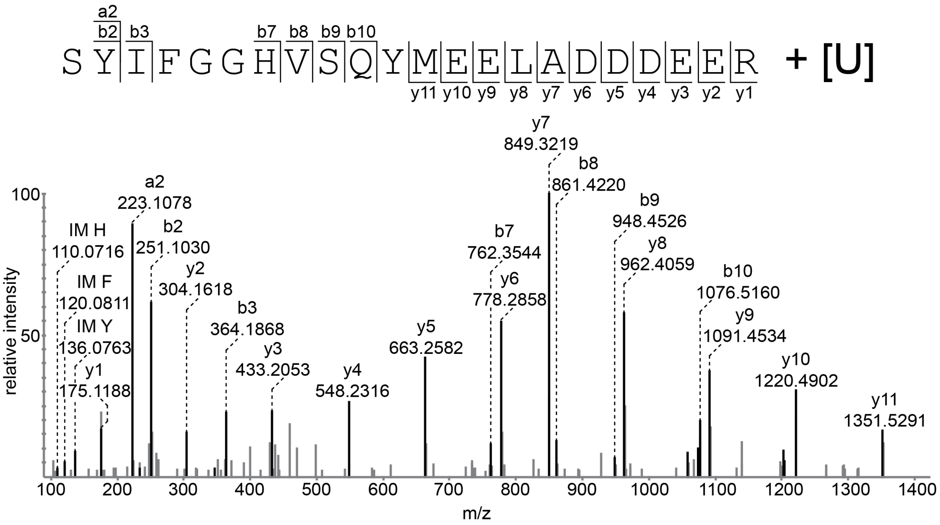


Figure B.40: MS/MS fragment spectrum of 60S ribosomal protein L5 peptide SYIFGGHVSQYMEELADDDEER (S197–R218) cross-linked to [U]. The spectrum shows no trace of the cross-linked RNA. Consequently, the cross-linked amino acid cannot be identified.

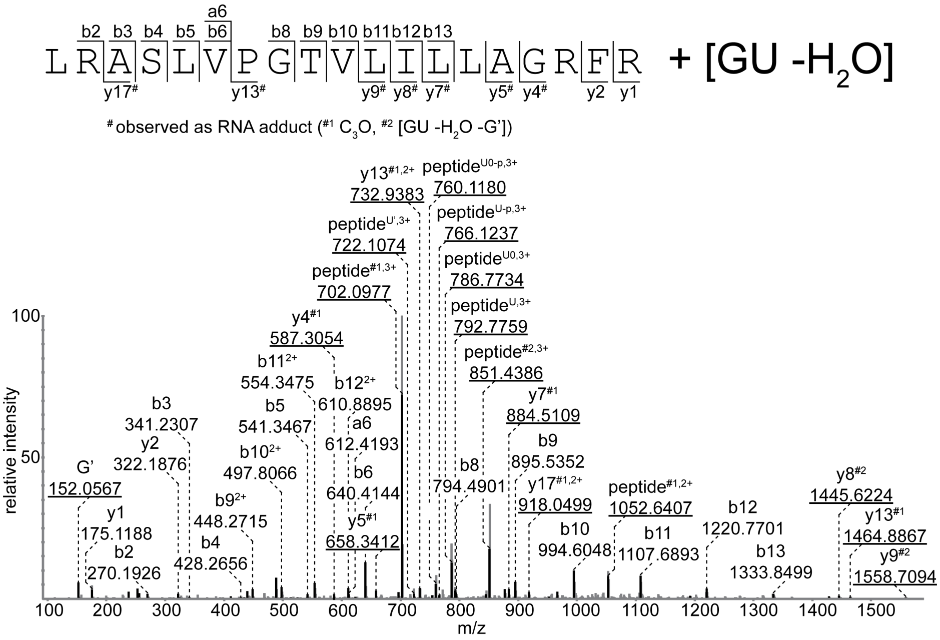


Figure B.41: MS/MS fragment spectrum of 60S ribosomal protein L6-A/-B peptide LRASLVPGTVLILLAGRFR (L30–R48) cross-linked to [GU $-H_2O$]. All y-ions starting with y4 are observed as RNA-adducts, which places the cross-link either on G45 or R46. In addition, the spectrum shows a number of RNA adducts of the intact peptide. The G' marker ion has an relative intensity well below 10% which is very unusual, typically adenine, cytosine and guanine produce high intensity marker ions.

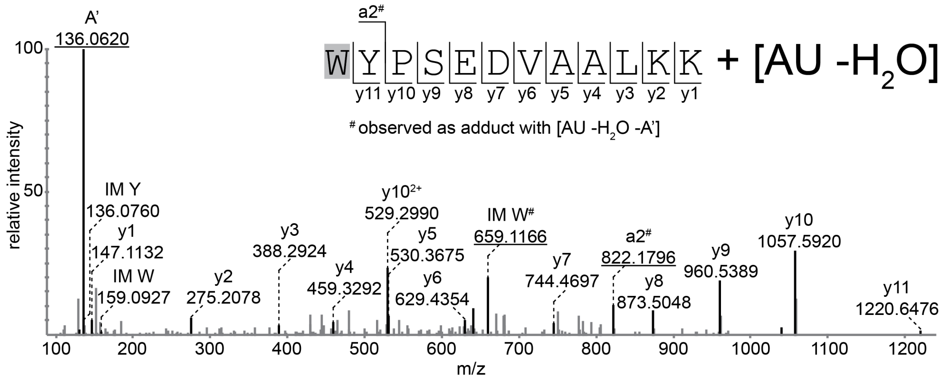


Figure B.42: MS/MS fragment spectrum of 60S ribosomal protein L6-A peptide WYPSEDVAALKK (W9–K20) cross-linked to [AU –H_2O]. Further description below Figure B.43.

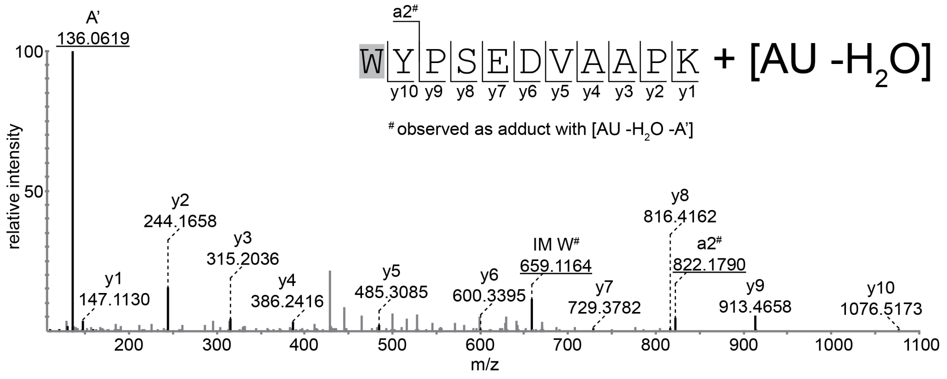


Figure B.43: MS/MS fragment spectrum of 60S ribosomal protein L6-B peptide WYPSEDVAAPK (W9–K19) cross-linked to [AU –H_2O]. Peptide WYPSEDVAALKK (W9–K20) of L6-A (spectrum shown in B.42) contains a leucine at position 18 which is missing in L16-B peptide WYPSEDVAAPK (W9–K19). In addition, the L6-A peptide contains a missed cleavage site. In both cases, the spectrum is dominated by the A' marker ion. Both peptides are identified by a full y-series and a2 shifted by the cross-linked RNA. In addition, the immonium ion of tryptophan is observed with the same shift, unambiguously identifying W9 as the cross-linked amino acid.

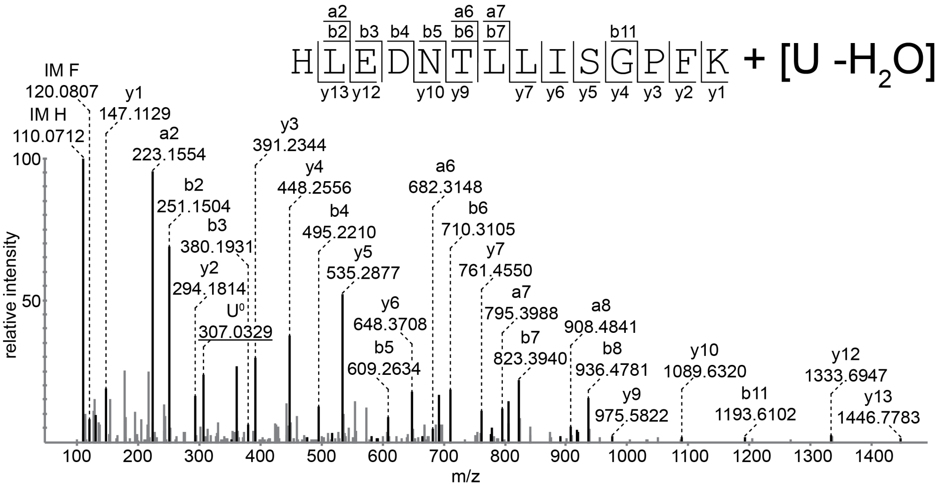


Figure B.44: MS/MS fragment spectrum of 60S ribosomal protein L6-A peptide HLEDNTLLISGPFK (H57–K70) cross-linked to [U $-H_2O$]. More detailed description below Figure B.45.

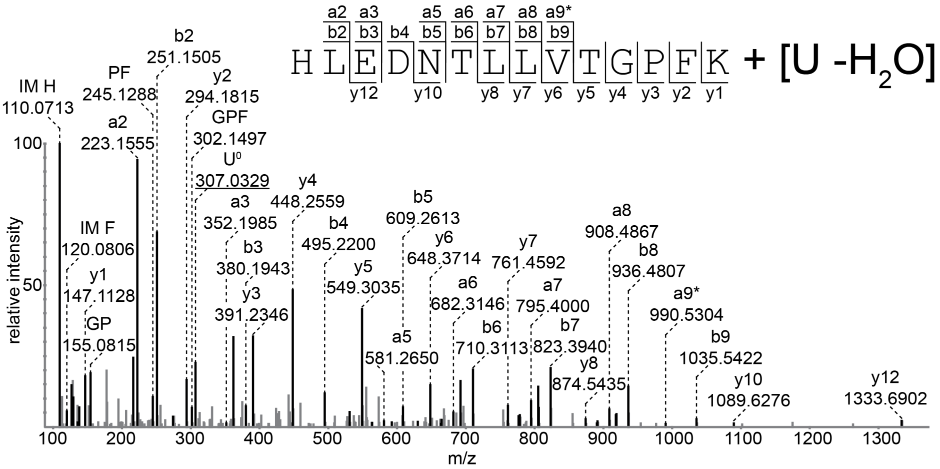


Figure B.45: MS/MS fragment spectrum of 60S ribosomal protein L6-B peptide HLEDNTLLVTGPFK (H57–K70) cross-linked to [U $-H_2O$]. The cross-linked peptides of L6-A (spectrum shown in B.44) and L6-B differ at positions 65 and 66, the A-form containing IS and the B-form VT at these positions. Interestingly, both peptides are isobaric, i.e. they have exactly the same elemental composition and consequently the same mass. However, both peptides can be clearly distinguished in the corresponding spectra of their cross-links to [U $-H_2O$] (compare Figures B.44 and B.45). The peptide sequence ion between the residues 65 and 66, i.e. y5 (and additionally a9/b9 in Figure B.45) allows differentiation of both homologs. Apart from that, both spectra are very similar, containing a relatively complete y-series and a number of a- and b-ions, a very intense histidine immonium ion and a uridine marker ion. However, no RNA adducts of peptide fragments are observed that would allow the identification of the cross-linked amino acid.

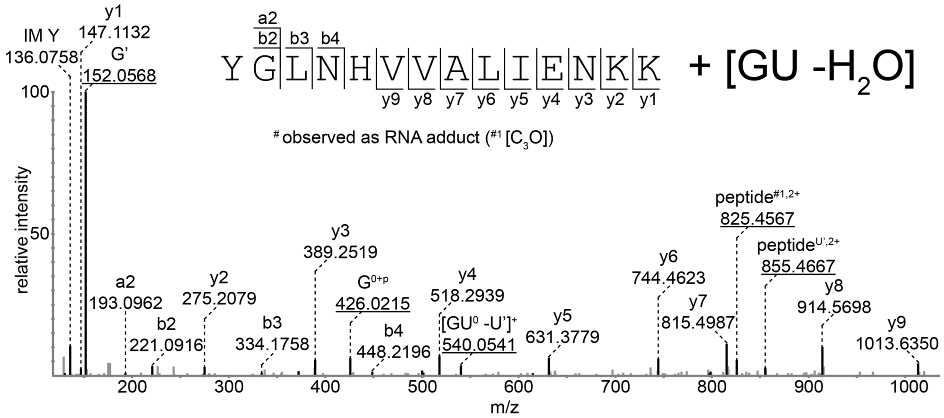


Figure B.46: MS/MS fragment spectrum of 60S ribosomal protein L8-A peptide YGLNHVVALIENKK (Y134–K147) cross-linked to [GU –H_2O] More detailed description below Figure B.47.

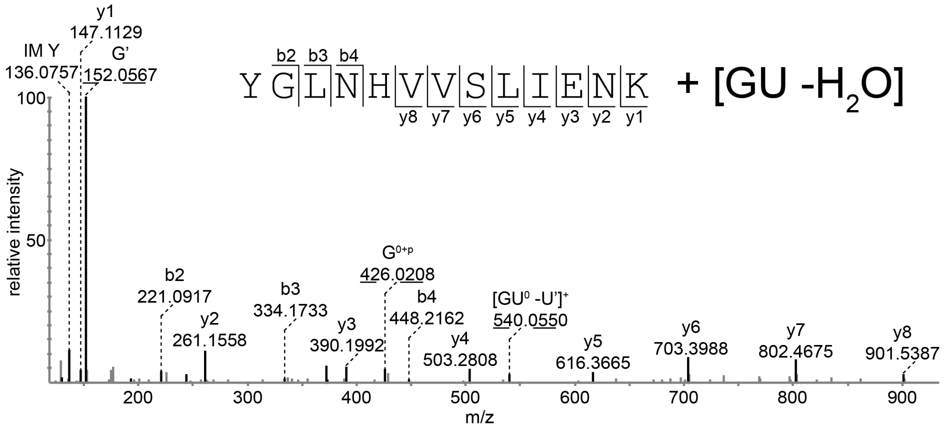


Figure B.47: MS/MS fragment spectrum of 60S ribosomal protein L8-B peptide YGLNHVVSLIENK (Y134–K146) cross-linked to [GU –H_2O]. The peptides of 60S ribosomal proteins L8-A and -B differ at position 141. L8-A peptide YGLNHVVALIENKK (Y134–K147, spectrum in B.46) contains a leucine while L8–B peptide YGLNHVVSLIENK (Y134–K146) contains a serine at this position. The cross-link of the A-form peptide contains a missed cleavage site. However, both spectra of the cross-links to [GU –H_2O] are very similar: Both are dominated by the G' marker ion and contain two additional RNA signals. The peptides' N- and C-termini are confidently identified by a series of the corresponding sequence ions. No RNA-adducts of peptide signals are observed, therefore the cross-linked amino acid cannot be identified.

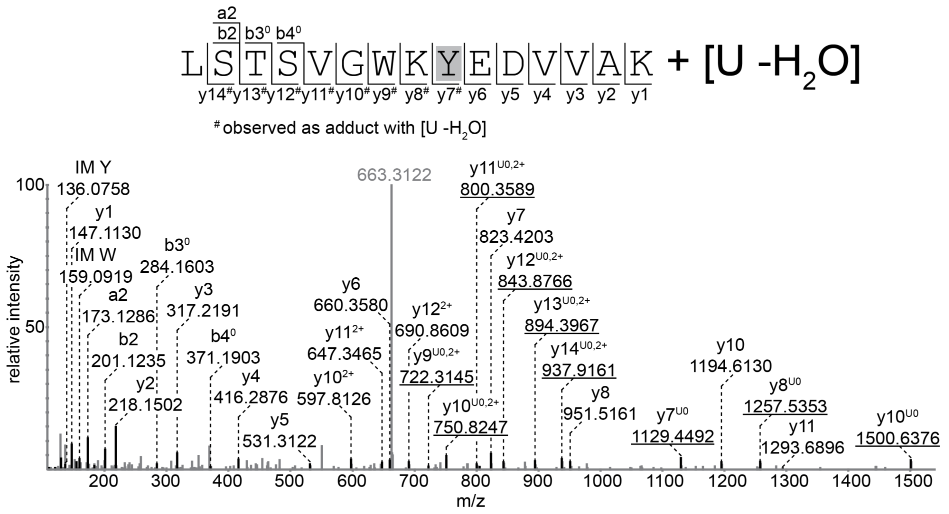


Figure B.48: MS/MS fragment spectrum of 60S ribosomal protein L16-A/-B peptide LSTSVGWKYEDVVAK (A: L141–K155, B: L140–K154) cross-linked to [U –H_2O]. The y-series is shifted by the mass of the cross-linked RNA starting with y7, identifying Y149 as the cross-linked amino acid.

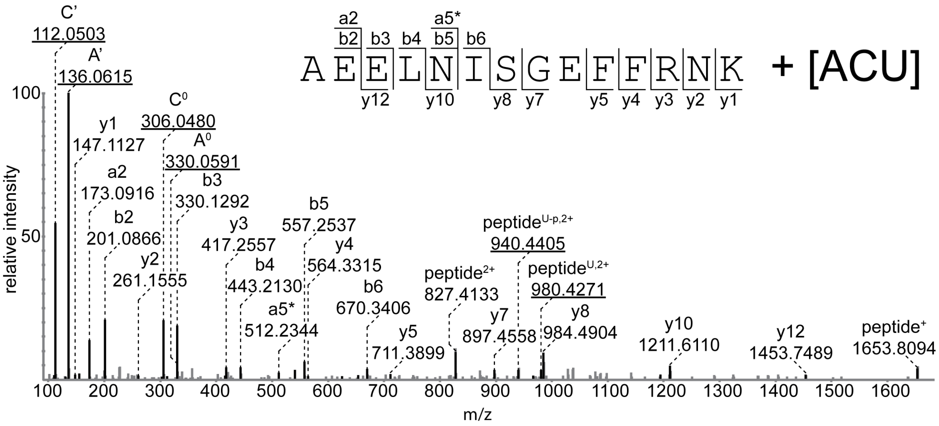


Figure B.49: MS/MS fragment spectrum of 60S ribosomal protein L16-A peptide AEELNISGEFFRNK (A38–K51) cross-linked to [ACU]. As expected, the fragment spectrum is dominated by RNA marker ions, namely A', C', and the corresponding nucleotides minus water, confirming the cross-linked RNA. However, the cross-linked amino acid cannot be determined. See also Figure B.50.

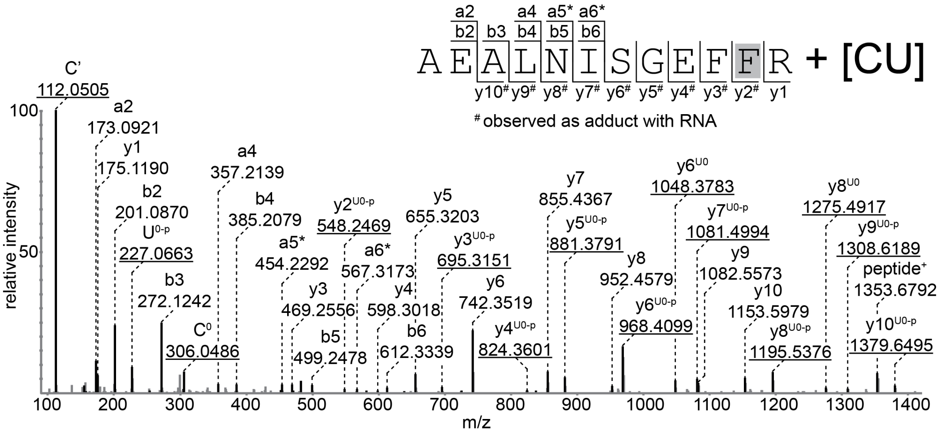


Figure B.50: MS/MS fragment spectrum of 60S ribosomal protein L16-B peptide AEALNISGEFFR (A37–R48) cross-linked to [CU]. This cross-link of L16-B covers the same region as the cross-link of L16-A shown in B.49. In comparison, the spectrum exhibits some substantial differences, it shows an intense C' marker ion and additionally the [C $-H_2O$] and [U $-H_3PO_4$] marker ions. The y-series is observed as partially shifted by RNA starting with y2, identifying F38 as the cross-linked amino acid residue. A spectrum very similar to the cross-link of the homologue L16-A peptide was also identified (data not shown).

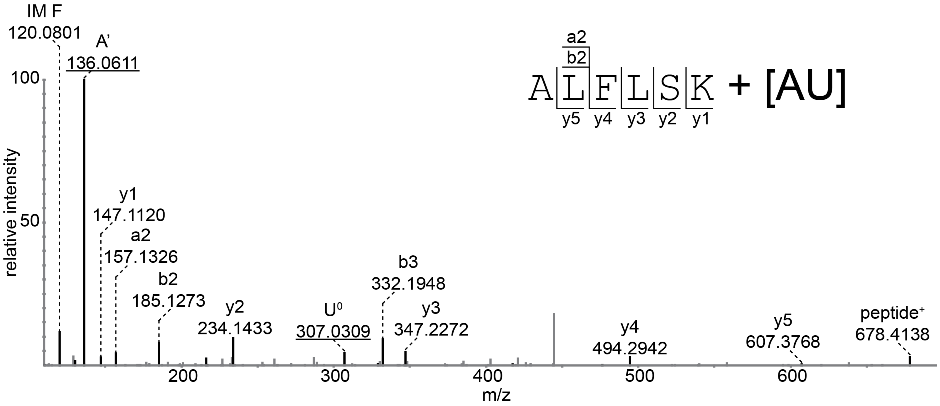


Figure B.51: MS/MS fragment spectrum of 60S ribosomal protein L18-A/-B peptide ALFLSK (A51–K56) cross-linked to [AU]. The cross-linked peptide is not unique. It appears in both 60S ribosomal proteins L18-A and -B and in the glycolipid 2-alpha-mannosyltransferase. Since another cross-link of the L18 proteins was identified and the majority of cross-links originate from ribosomal proteins, we assume that this cross-link is indeed from the L18 proteins. As expected from the cross-linked RNA [AU], an intense A' marker is identified. No adducts are observed which would allow the identification of the cross-linked amino acid.

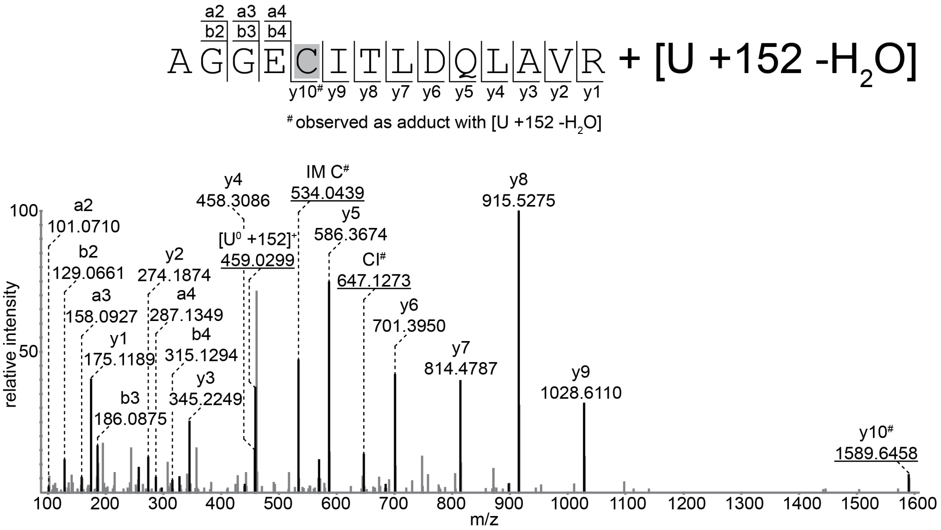


Figure B.52: MS/MS fragment spectrum of 60S ribosomal protein L18-A/-B peptide AGGECITLDQLAVR (A117–R130) cross-linked to [U +152 –H_2O]. The immonium ion of cysteine, y10, and an internal ion CI are all found shifted by the mass of the cross-linked RNA, thus identifying C121 as the cross-linked amino acid residue.

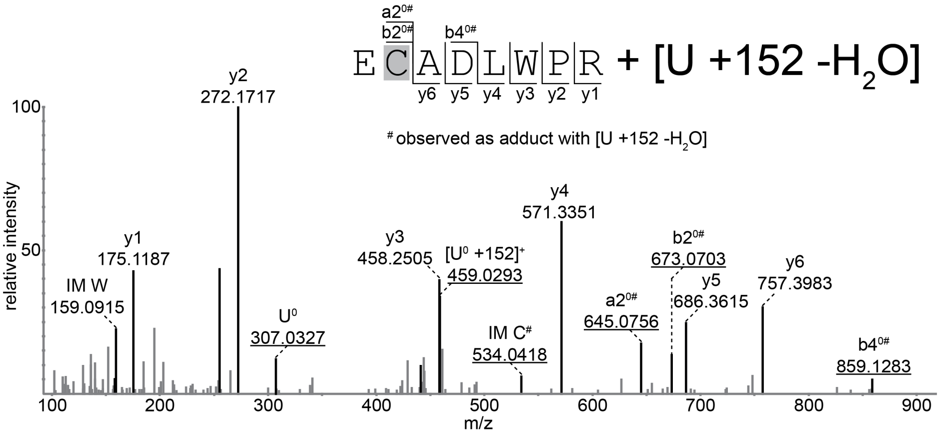


Figure B.53: MS/MS fragment spectrum of 60S ribosomal protein L23-A/-B peptide ECADLWPR (E121–R128) cross-linked to [U +152 –H_2O]. All ions containing the N-terminus are shifted by [U +152 –H_2O], the cross-linked RNA. Since the immonium ion of cysteine is shifted by the same mass, it can be concluded that C122 is the cross-linked amino acid.

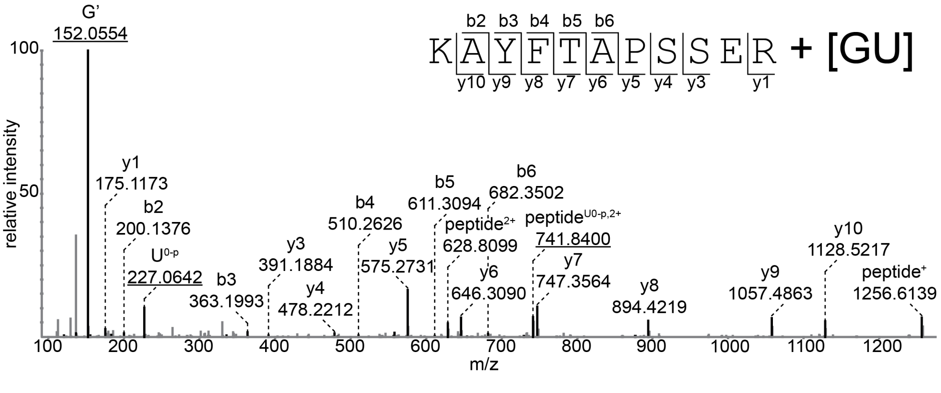


Figure B.54: MS/MS fragment spectrum of 60S ribosomal protein L26-B peptide KAYFTAPSSER (K17–R27) cross-linked to [GU]. The cross-linked RNA is confirmed by the intense G' marker and a signal for [U –H_3PO_4]. However, no peptide–RNA adduct is observed that would allow the identification of the cross-linked amino acid.

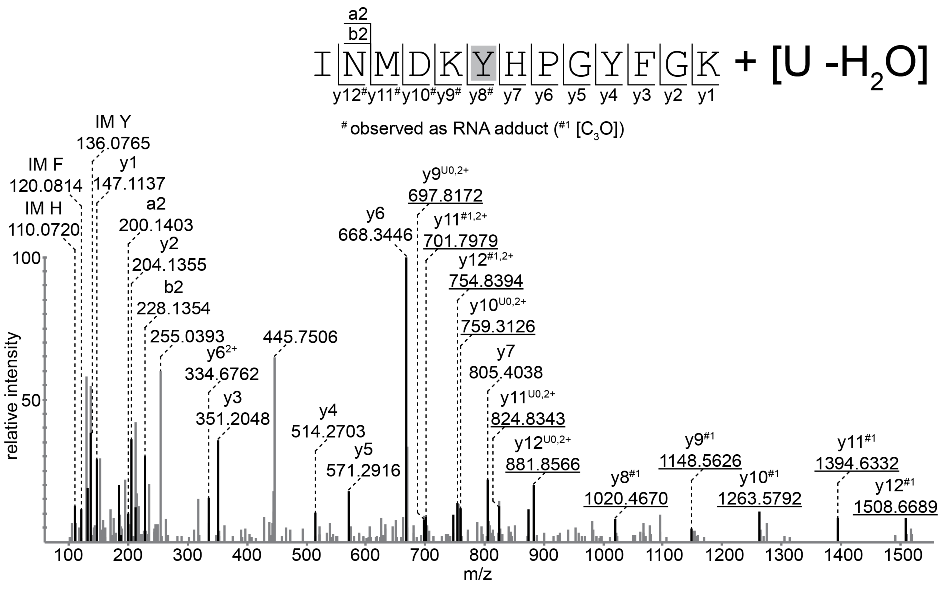


Figure B.55: MS/MS fragment spectrum of 60S ribosomal protein L28 peptide INMDKYHPGYFGK (I43–K55) cross-linked to [U $-H_2O$]. All y-ions from y8 on are observed as completely shifted by RNA adducts, either uracil fragment [C_3O] or the cross-linked RNA [U $-H_2O$]. This puts the cross-link on Y48.

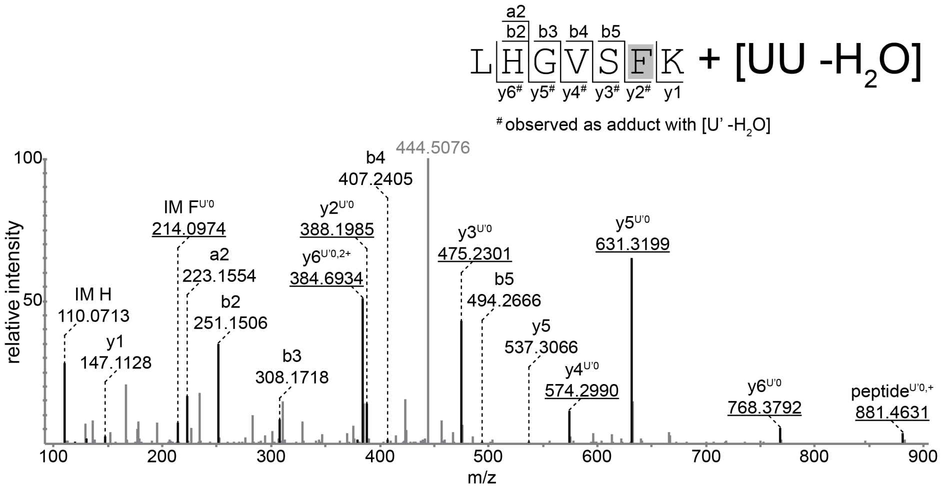


Figure B.56: MS/MS fragment spectrum of 60S ribosomal protein L31-A/-B peptide LHGVSFK (L20–K26) cross-linked to [UU $-H_2O$]. F25 is the cross-linked amino acid residue because all y-ions containing this residue are observed as an adduct with [U' $-H_2O$].

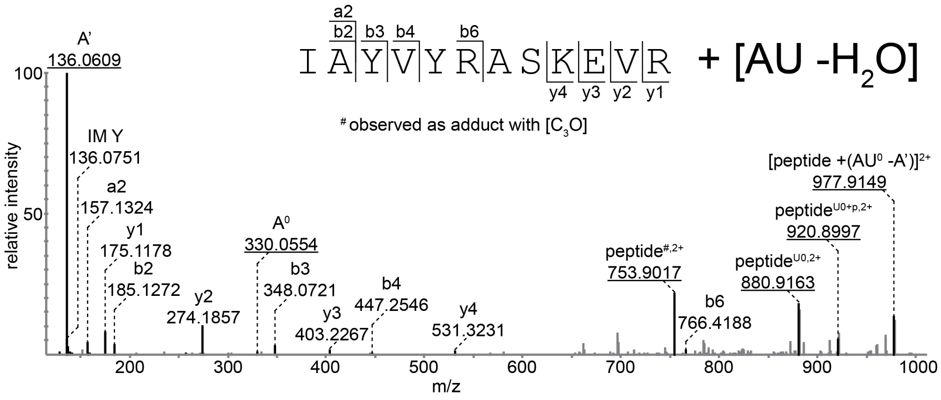


Figure B.57: MS/MS fragment spectrum of 60S ribosomal protein L33-A/-B peptide IAYVYRASKEVR (I49–R60) cross-linked to [AU $-H_2O$]. The spectrum is dominated by the A' marker ion and only a limited number of peptide fragments (y1–y4, a2/b2, b3–b4, and b6) are observed. In addition, fragmentation of the cross-linked RNA on the intact peptide is observed, leading to a number of peptide adduct with RNA fragments (see higher m/z range).

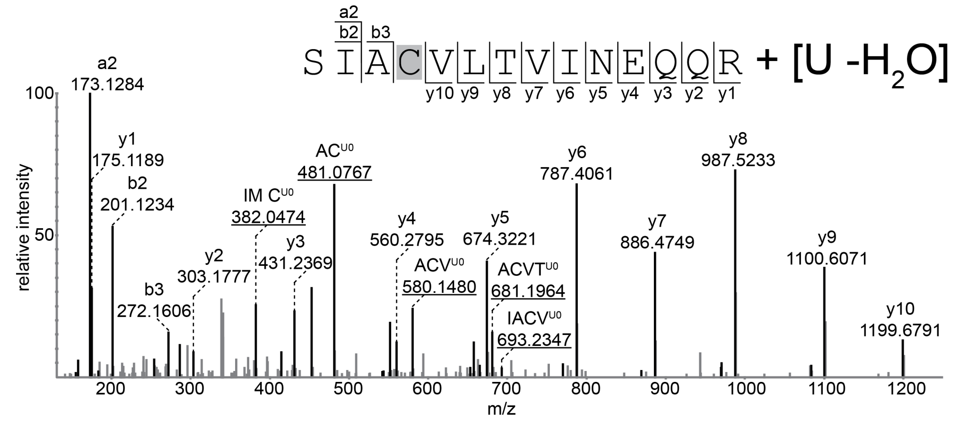


Figure B.58: MS/MS fragment spectrum of 60S ribosomal protein L35-A/-B peptide SIACVLTVINEQQR (S50–R63) cross-linked to [U $-H_2O$]. The immonium ion of cysteine shifted by [U $-H_2O$] clearly identifies C53 as the cross-linked residue. Interestingly, no regular peptide sequence ions containing the cross-linked C53 are observed. Instead, a number of internal ions with said cysteine are observed, all shifted by the cross-linked RNA.

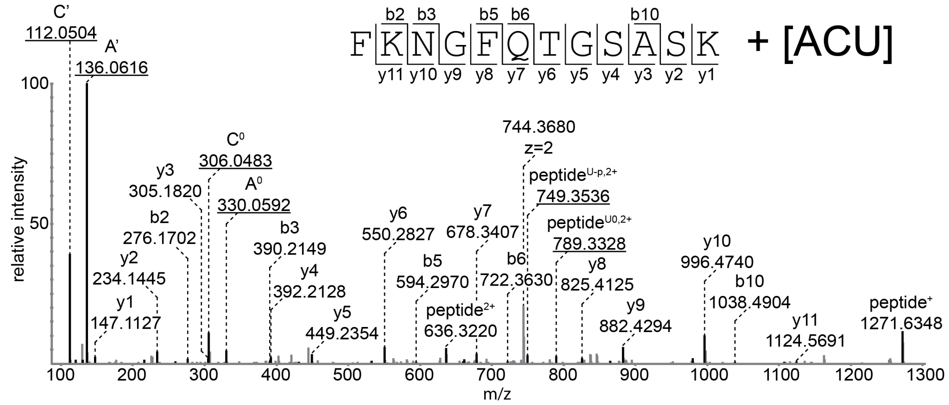


Figure B.59: MS/MS fragment spectrum of 60S ribosomal protein L37-A peptide FKNGFQTGSASK (F74–K85) cross-linked to [ACU]. More detailed description below Figure B.60.

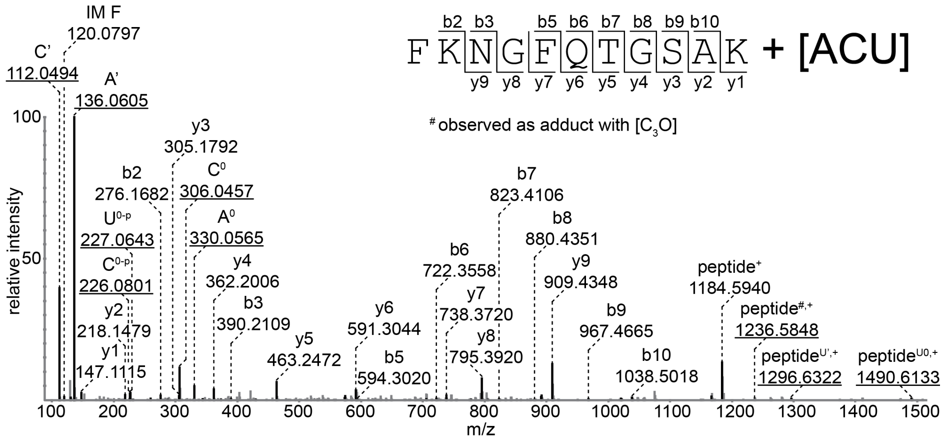


Figure B.60: MS/MS fragment spectrum of 60S ribosomal protein L37-B peptide FKNGFQTGSAK (F74–K84) cross-linked to [ACU]. The L37-A peptide FKNGFQTGSASK (F74–K85, spectrum see B.59) contains a serine at position 84 which is missing from L37-B peptide FKNGFQTGSAK (F74–K84). Both peptides are identified by almost complete y-series and a number of b-ions. Apart from the different overall masses, both peptides can be easily distinguished by their y-series differing from y2. As expected from the cross-linked RNA [ACU], an intense A' marker is observed, while the C' marker is clearly visible but only at 40% relative intensity. The C^0 and A^0 marker ions are also observed. Both spectra do not contain any shifted peptide fragments but only RNA adducts of the intact peptide, therefore the cross-linked amino acid cannot be identified.

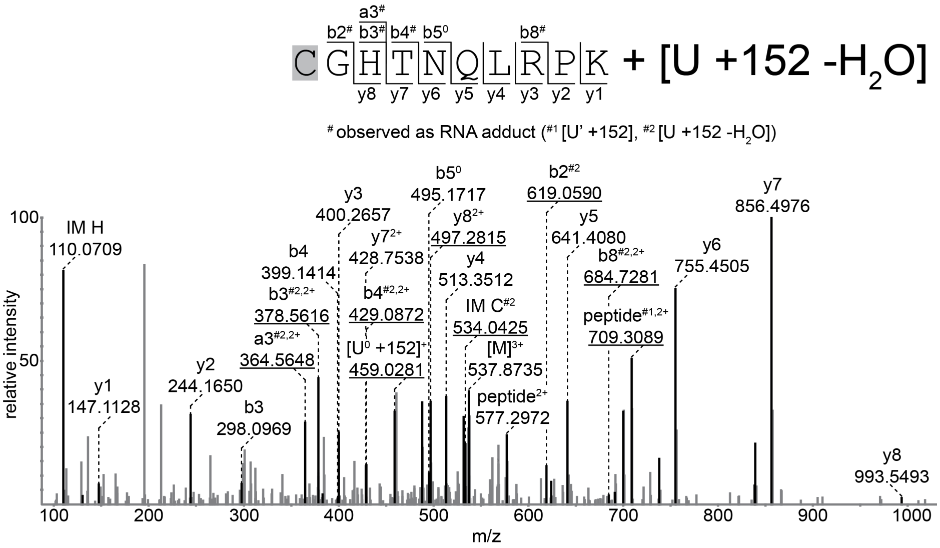


Figure B.61: MS/MS fragment spectrum of Ubiquitin-60S ribosomal protein L40 peptide CGHTNQLRPK (C115–K124) cross-linked to [U +152 –H_2O]. The a- and b-ions are partially shifted by either the cross-linked RNA or its cleavage product [U' +152]. A signal for [U +152 –H_2O] is also observed. The 152 adduct as well as its immonium ion shifted by the cross-linked RNA point to C115 as the cross-linked amino acid.

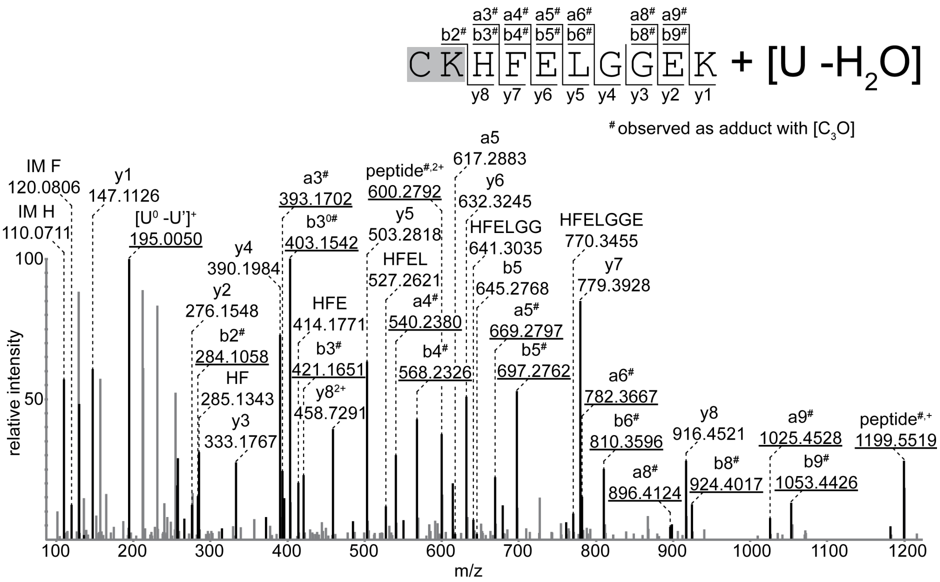


Figure B.62: MS/MS fragment spectrum of 60S ribosomal protein L42-A/-B peptide CKHFELGGEK (C88–K97) cross-linked to [U –H_2O]. All observed a- and b-ions are partially shifted by uracil fragment [C_3O]. This places the cross-link on either C88 or K89. The absence of y9 and the observation of several internal fragments originating from cleavage C-terminal to K89 might hint to it being cross-linked, however there is no unambiguous proof.

B.5.3 Cross-links of ribosome-related proteins

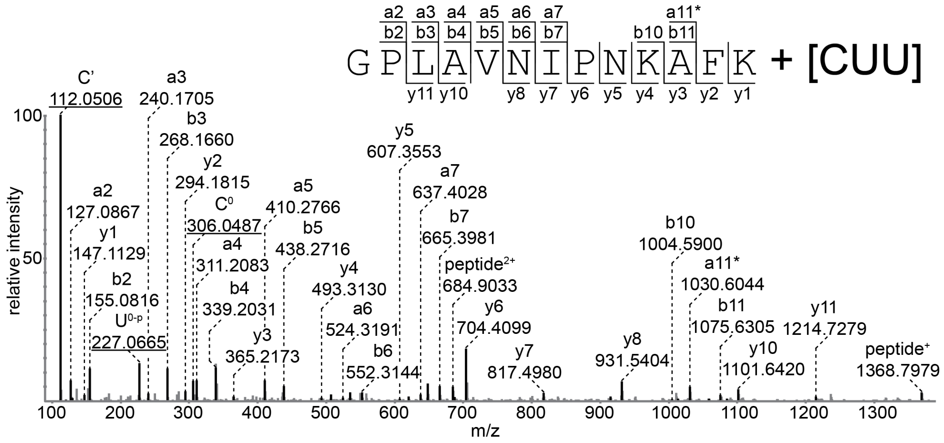


Figure B.63: MS/MS fragment spectrum of ribosome biogenesis protein RLP7 peptide GPLAVNIPNKAFK (G149–K161) cross-linked to [CUU]. The spectrum is dominated by the C' marker ion, the peptide fragments are suppressed below 15% relative intensity. Additional RNA marker ions for [U $-H_3PO_4$] and C^0 are observed. However, there is no observable shift of peptide fragments that would allow to pinpoint to the cross-linked amino acid.

B.5.4 Cross-links of polynucleotide-binding proteins

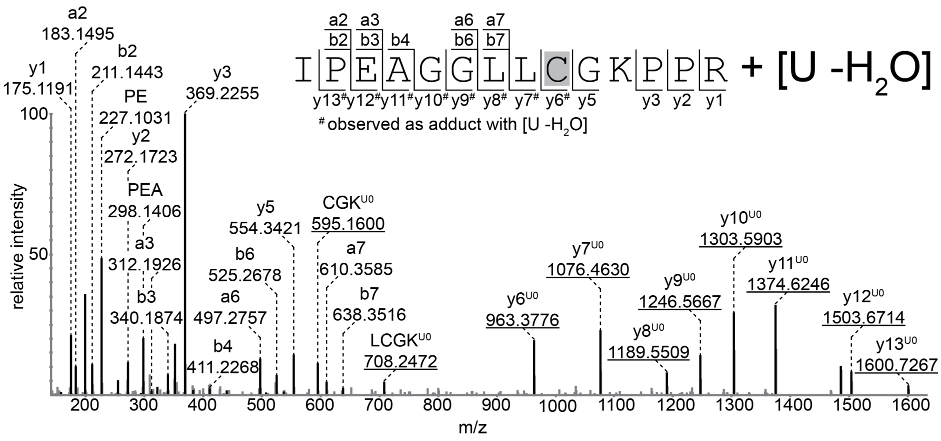


Figure B.64: MS/MS fragment spectrum of cruciform DNA-recognizing protein 1 peptide IPEAGGLLCGKPPR (I105–R118) cross-linked to [U $-H_2O$]. All y-ions from y6 on are shifted by the mass of [U $-H_2O$], which points at C113 as the cross-linked amino acid.

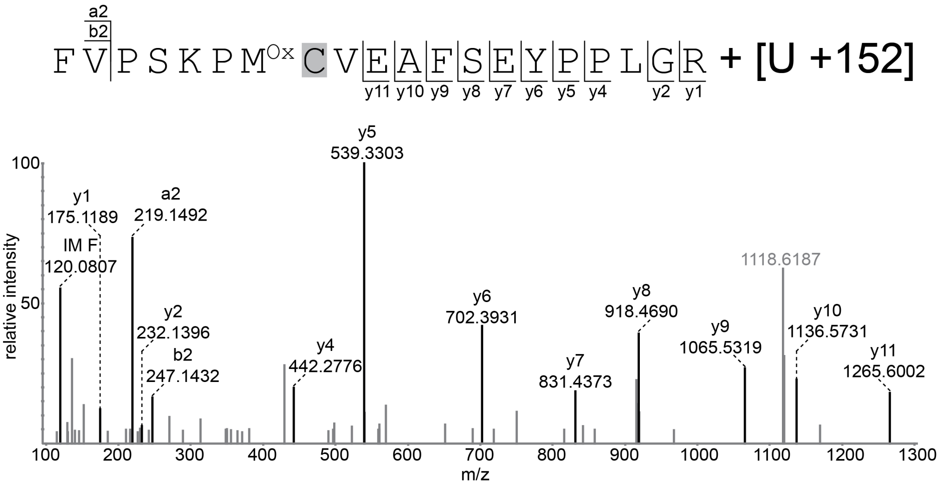


Figure B.65: MS/MS fragment spectrum of elongation factor 1-alpha peptide FVPSKPMCVEAFSEYPPLGR (F402–R421) cross-linked to [U +152]. The oxidation on M408 as well as the cross-linked RNA sequence [U +152] are derived from the difference between experimental precursor and calculated peptide mass. The mass shift of 152 points at C409 as the cross-linked amino acid.

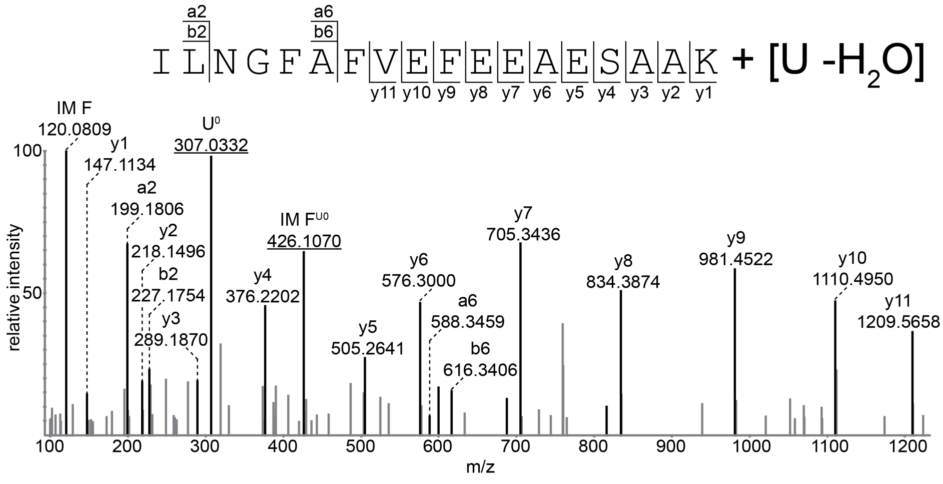


Figure B.66: MS/MS fragment spectrum of nucleolar protein 3 peptide ILNGFAFVEFEEAESAAK (I156–K173) cross-linked to [U –H_2O]. An intense U^0 marker is observed. The immonium ion of phenylalanine is observed partially shifted by the cross-linked RNA. However, the spectrum does not allow to distinguish whether the cross-link is via F160, F162, or F165. The peptide lies within RRM 1 of Nop3p (positions 125–195).

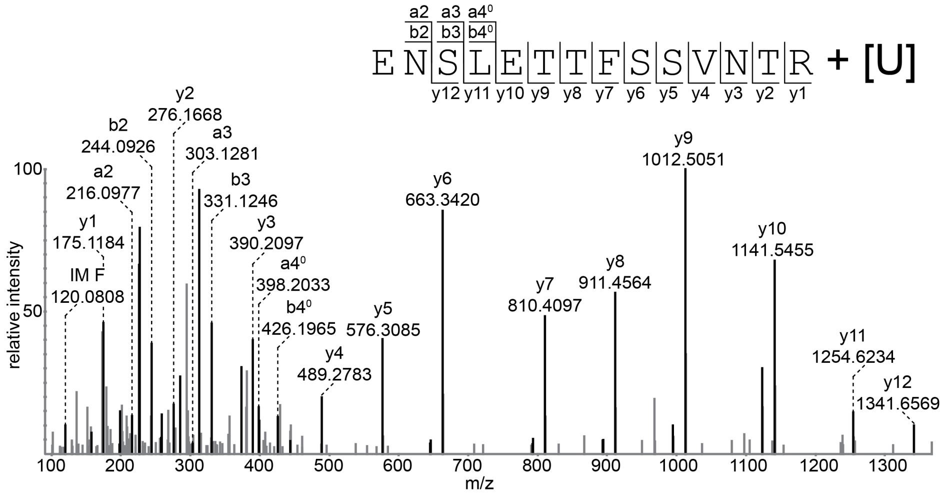


Figure B.67: MS/MS fragment spectrum of nucleolar protein 3 peptide ENSLETTFSSVNTR (E222–R235) cross-linked to [U]. The spectrum does not contain any RNA signals, the cross-linked nucleotide is solely deduced from the difference between the experimental precursor and the calculated peptide mass. The cross-linked amino acid cannot be determined. The peptide lies within RRM 2 (positions 200–275) of Nop3p.

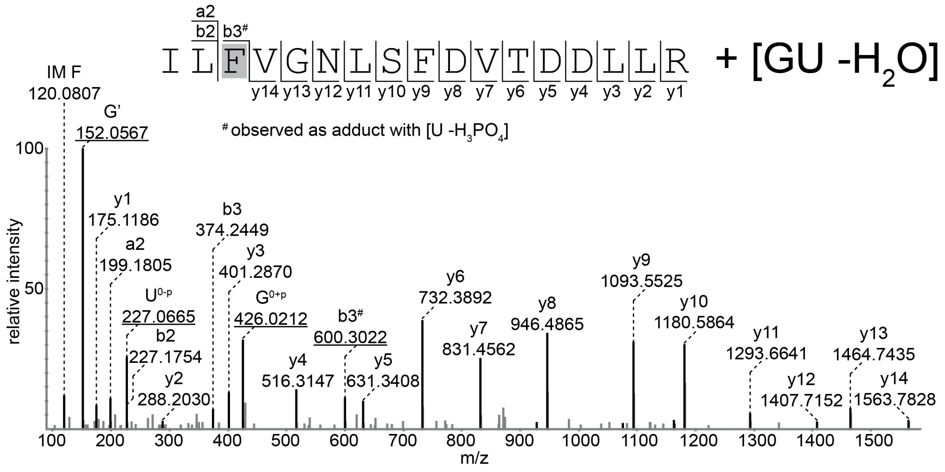


Figure B.68: MS/MS fragment spectrum of nucleolar protein 13 peptide ILFVGNLSFDVTDDLLR (I240–R256) cross-linked to [GU –H_2O]. b3 is observed as an adduct with [U –H_3PO_4], thus F242 is the cross-linked amino acid residue. The G' marker is the most intense signal, the peptide fragments are below 40% relative intensity. In addition, marker ions for [U –H_3PO_4] and [G –H_2O +HPO_3] are observed at 35% and 40% relative intensity, respectively. The peptide lies at the beginning of the protein's RRM 2 (position 239–317).

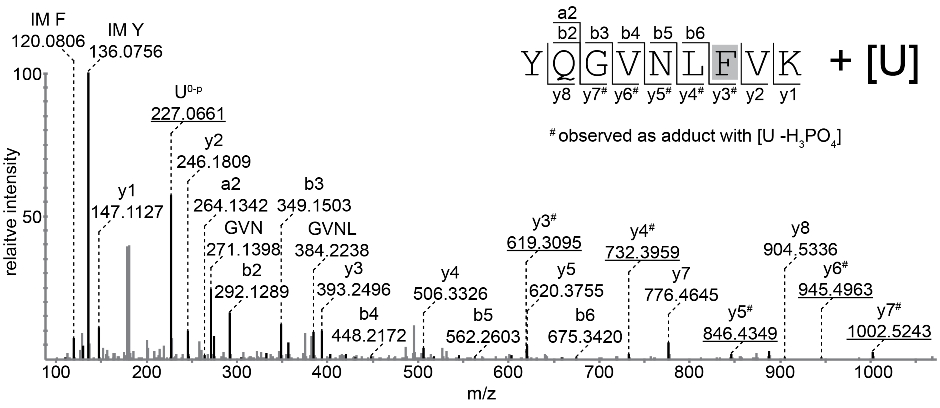


Figure B.69: MS/MS fragment spectrum of polyadenylate-binding protein peptide YQGVNLFVK (Y319–K327) cross-linked to [U]. An RNA marker ion for [U $-H_3PO_4$] is observed. y3 to y7 are partially shifted by the same RNA. Thus, F325 is the cross-linked amino acid.

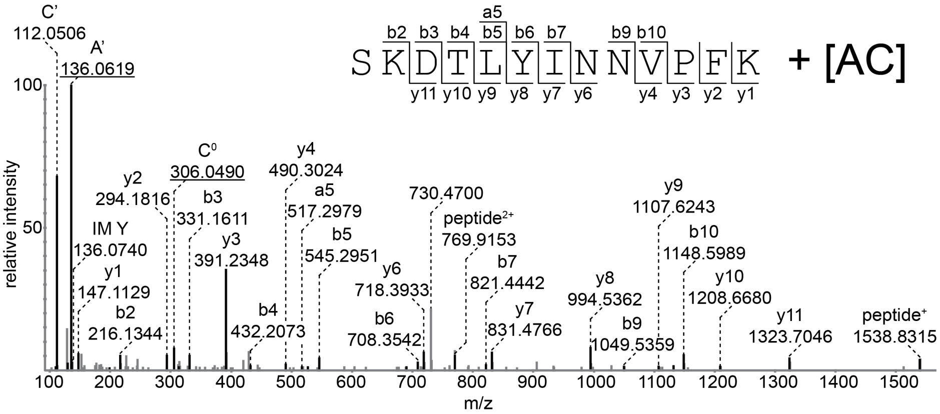


Figure B.70: MS/MS fragment spectrum of single-stranded nucleic-acid binding protein peptide SKDTLYINNVPFK (S184–K196) cross-linked to [AC]. This cross-link represents the only example for a cross-link without uridine in this experiment. The A' marker is the most intense signal, C' is observed at 70% relative intensity. In addition, the C^0 marker is observed. No peptide–RNA adduct is visible that would allow identification of the cross-linking site on the peptide or the cross-linked nucleotide.

B.5.5 Cross-links of proteins without any annotated polynucleotide-binding function

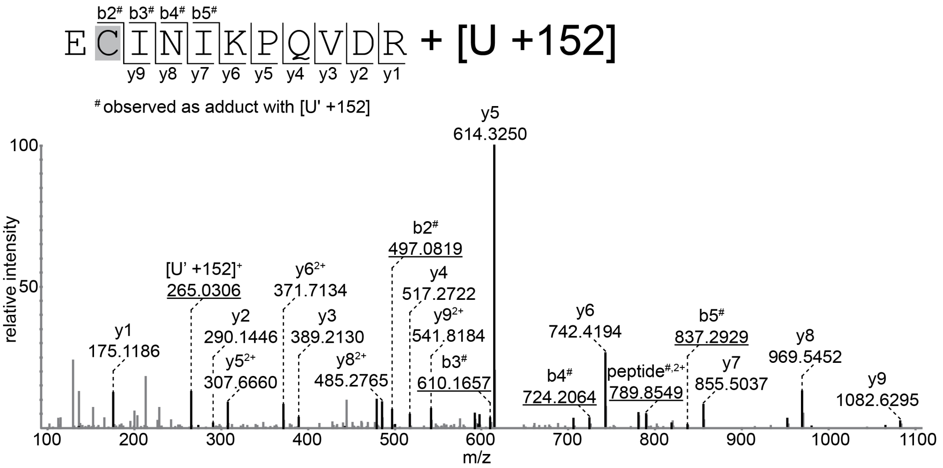


Figure B.71: MS/MS fragment spectrum of adenosylhomocysteinase peptide ECINIKPQVDR (E320–R330) cross-linked to [U +152]. The observation of the 152 Da adduct hints at C321 as the cross-linked amino acid and is confirmed by the b-ion series shifted by [U' +152].

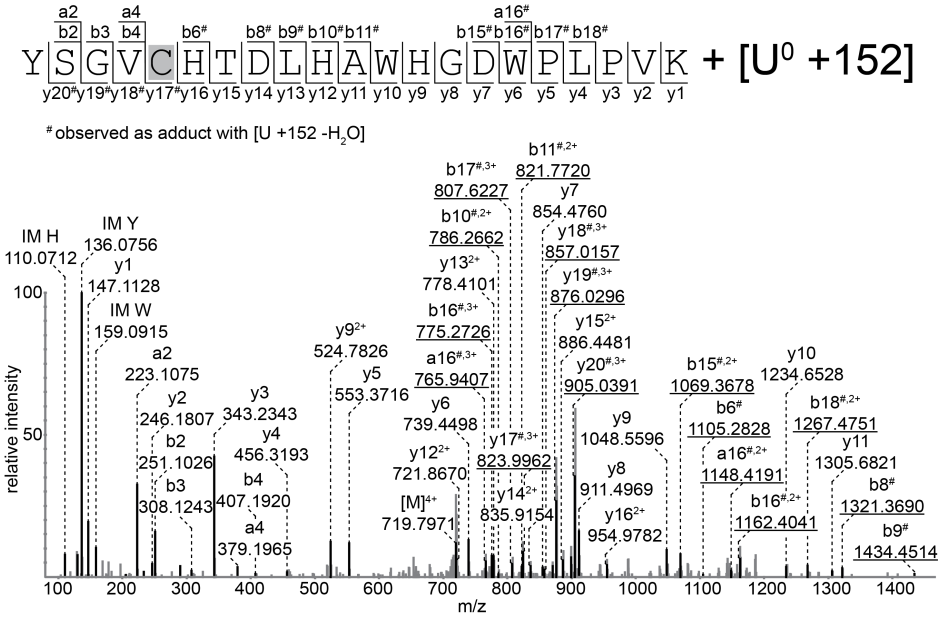


Figure B.72: MS/MS fragment spectrum of alcohol dehydrogenase 1/3 peptide YSGVCHTDLHAWHGDWPLPVK (Adh1p: Y40–K60, Adh3p: Y67–K86) cross-linked to [U +152 –H_2O]. Fragment ions containing the cross-linked cysteine (Adh1p: C44, Adh3p: C71), i.e. y-ions y17 to y20 and a-/b-ions starting with b6, are shifted by the cross-linked RNA fragment, [U +152 –H_2O].

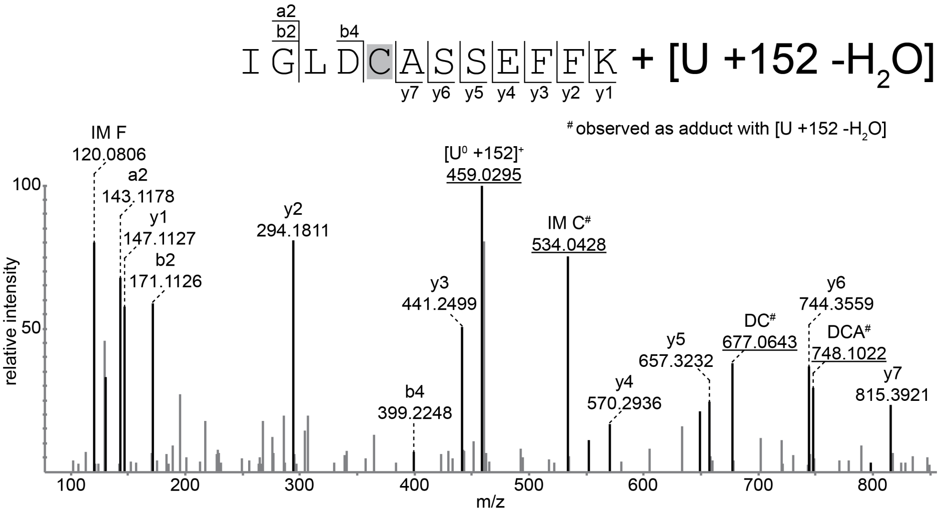


Figure B.73: MS/MS fragment spectrum of enolase 1/2 peptide IGLDCASSEFFK (I244–K255) cross-linked to [U +152 –H_2O]. C248 is identified as the cross-linked amino acid by its immonium ion observed as adduct with the cross-linked RNA, [U +152 –H_2O]. Interestingly, no peptide sequence ions containing the cross-linked cysteine are observed, but two internal ion which exhibit the same shift as the immonium ion. The most intense cross-link fragment is an RNA signal corresponding to [U +152 –H_2O].

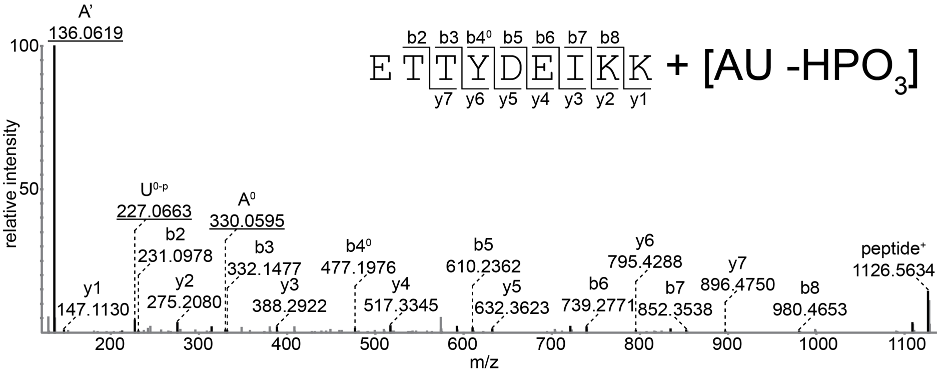


Figure B.74: MS/MS fragment spectrum of glyceraldehyde-3-phosphate dehydrogenase 2/3 peptide ETTYDEIKK (E250–K258) cross-linked to [AU –HPO_3]. The adenine marker ion dominates the spectrum, peptide fragments are supressed below 5% relative intensity. The cross-linked amino acid cannot be determined.

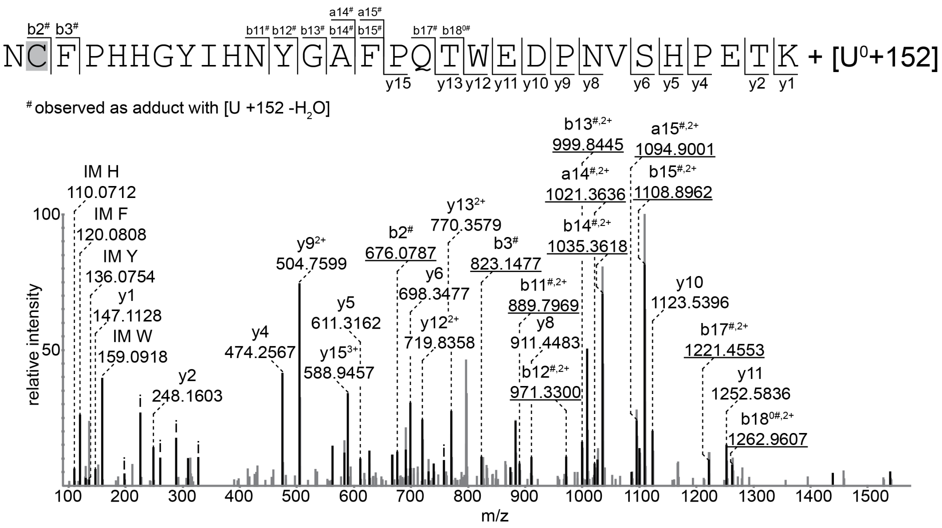


Figure B.75: MS/MS fragment spectrum of inorganic pyrophosphatase peptide NCFPHHGYIHNYGAFPQTWEDPNVSHPETK (N83–K112) cross-linked to [U +152 –H_2O]. All observed a- and b-ions are completely shifted by the cross-linked RNA. The 152 adduct identifies C84 as the cross-linked amino acid which is confirmed by the shift of a- and b-ions. Due to high number of signals, several signals are just highlighted and not annotated: if a fragment was observed in several charge states, only the most abundant signal is annotated. In addition, internal ions except for immonium ions are marked with an "i" only and are not annotated further.

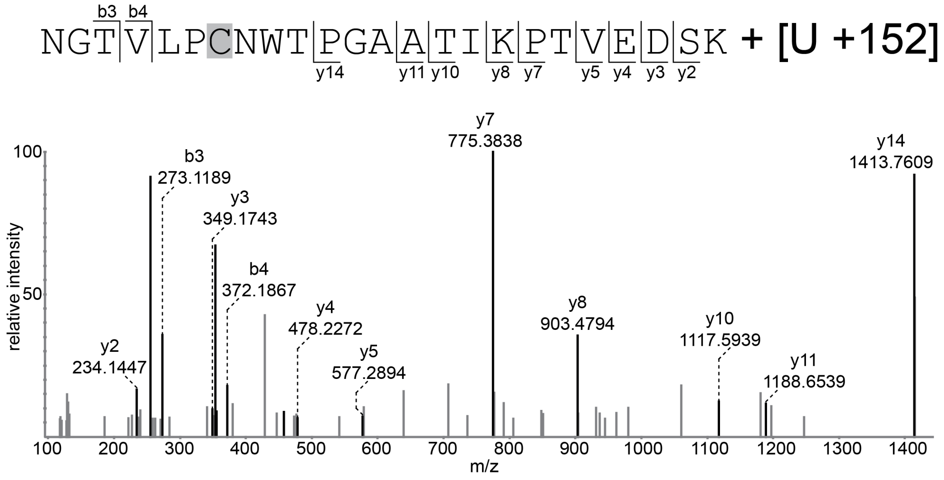


Figure B.76: MS/MS fragment spectrum of peroxiredoxin TSA1 peptide NGTVLPCNWTPGAATIKPTVEDSK (N165–K188) cross-linked to [U +152]. The overall intensity of the observed fragments is low, the cross-link is apparently of very low abundance. As a consequence, only very few peptide fragments are observed. Importantly, fragments from cleavage N-terminal to proline residues, here y7 and y14, are highly abundant as expected. The mass adduct of 152 points at C171 as the cross-linked amino acid.

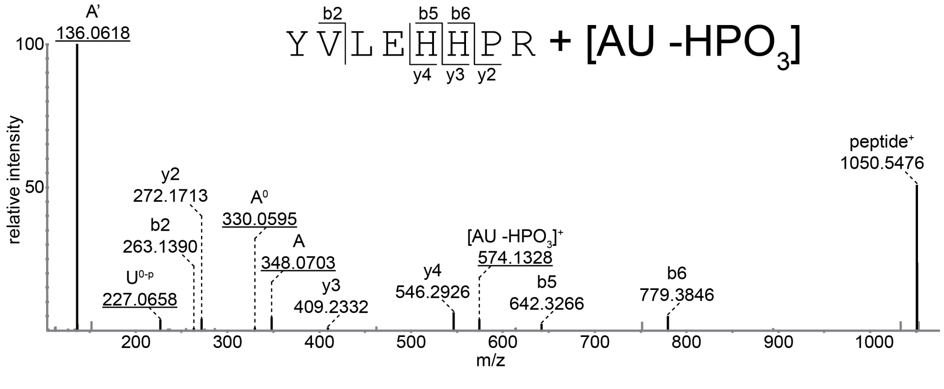


Figure B.77: MS/MS fragment spectrum of phosphoglycerate kinase peptide YVLEHHPR (Y49–R56) cross-linked to [AU –HPO$_3$]. The spectrum is dominated by the A' marker ion, the peptide fragments are suppressed well below 10% relative intensity. In addition, a signal for the intact peptide with over 50% relative intensity is observed. This can explain why only a small number of peptide fragments is observed (y2–y4, b2, and b5–b6). No shifted peptide sequence ions are observed which would allow the identification of the cross-linked amino acid residue.

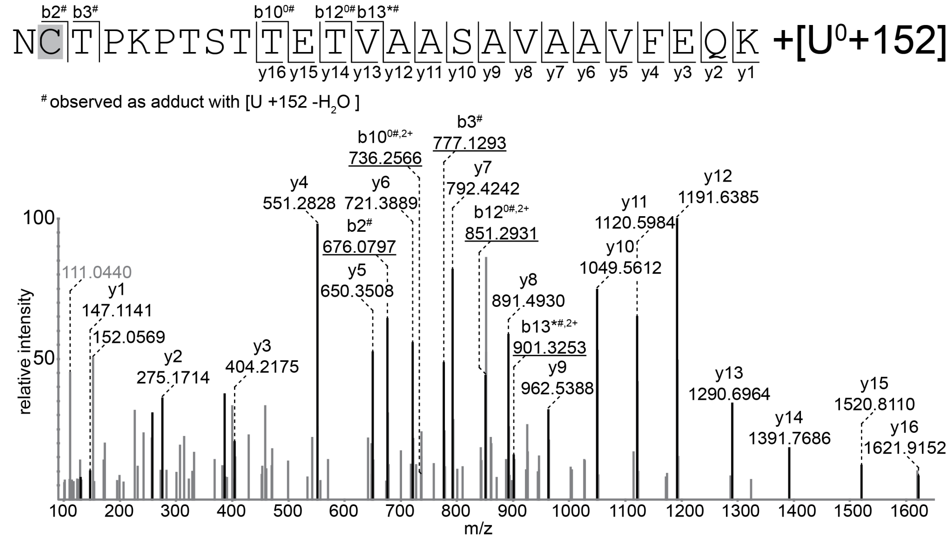


Figure B.78: MS/MS fragment spectrum of pyruvate kinase 1 peptide NCTPKPTSTTETVAASAVAAVFEQK (N370–K394) cross-linked to [U +152 –H_2O]. All observed b-ions are shifted by the cross-linked RNA, confirming C371 as the cross-linked amino acid.

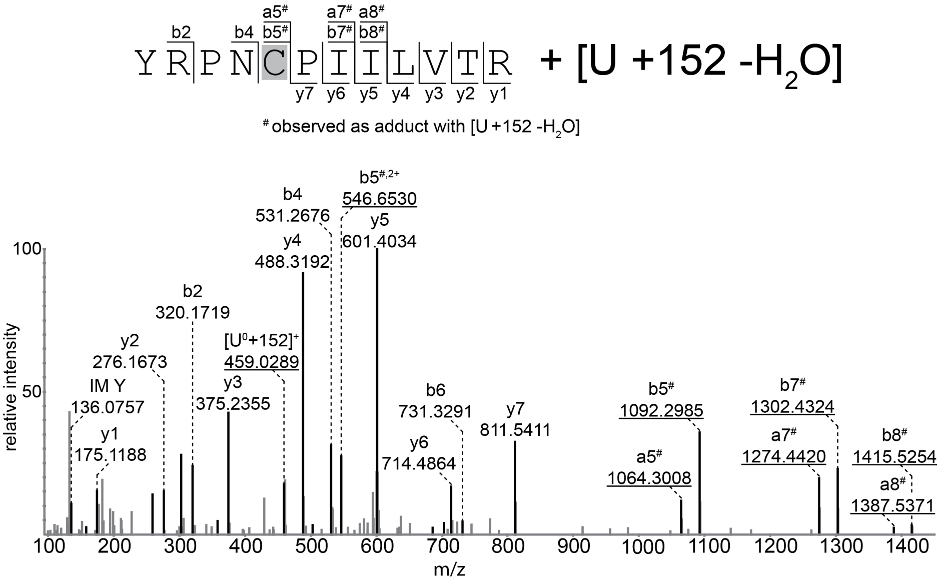


Figure B.79: MS/MS fragment spectrum of pyruvate kinase 1 peptide YRPNCPIILVTR (Y414–R425) cross-linked to [U +152 –H_2O]. The 152 mass adduct points towards C418 as the cross-linked amino acid, which is confirmed by the shifted a- and b-ions.

C Abbreviations

2PG	2-phosphoglycerate
4SU	4-thio-uracil
AC	alternating current
ACN	acetonitrile
ADH	alcohol dehydrogenase
AdoHcy	adenosylhomocysteine
APS	ammonium peroxodisulfate
ATP	adenosine-5'-triphosphate
CID	collision-induced dissociation
CIP	calf intestinal alkaline phosphatase
DC	direct current
DDA	data dependent acquisition
DEAE	diethylaminoethanol
DHB	2,5-dihydroxybenzoic acid
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide-5´-triphosphate
DTT	dithiothreitol
ECD	electron capture dissociation
ECL	enhanced chemiluminescence
E. coli	*Escherichia coli*
EDTA	ethylene diamine tetraacetic acid
EGTA	ethylene glycol tetraacetic acid
EMSA	electrophoretic mobility shift assay
ESI	electrospray ionization
ETD	electron transfer dissociation
FA	formic acid
FDR	false discovery rate
FT-ICR	fourier transform ion cyclotron resonance
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GST	glutatione *S*-transferase
HCD	higher energy collisional dissociation
Hepes	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
hnRNP	heterogeneous nuclear RNP
HPLC	high performance liquid chromatography
IM	immonium ion
IMAC	immobilized metal-ion affinity chromatography
KH domain	K homology domain
LB	lysogeny broth

LC	liquid chromatography
LTQ	linear trap quadrupole
M	molar, mol/l
m/z	mass-to-charge (ratio)
MALDI	matrix assisted laser desorption/ionization
miRNA	microRNA
MRM	multiple reaction monitoring
mRNA	messenger RNA
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NAD	nicotinamide adenine dinucleotide
NADH	nicotinamide adenine dinucleotide (reduced form)
NMR	nuclear magnetic resonance
NTA	nitrilotriacetic acid
OD	optical density
OMSSA	Open Mass Spectrometry Search Algorithm
PAGE	polyacrylamide gel electrophoresis
PCI	phenol-chloroform-isoamyl alcohol
PCR	polymerase chain reaction
PEG	polyethylene glycol
PEP	phosphoenolpyruvate
PNK	polynucleotide kinase
ppm	parts per million
ProtA	protein A (*Staphylococcus aureus*)
PSM	peptide-to-spectrum match
PTM	post-translational modification
Q-ToF	quadrupole time-of-flight
Q-trap	triple quadrupole/linear ion trap
RBD	RNA binding domain
RBP	RNA binding protein
RF	radio frequency
RNA	ribonucleic acid
RNP	ribonucleoprotein (particle or domain)
RP	reversed phase (-HPLC)
rpm	rounds per minute
RRM	RNA recognition motif
rRNA	ribosomal RNA
S. cerevisiae	*Saccharomyces cerevisiae*
SDS	sodium dodecyl sulfate
siRNA	small interfering RNA
snRNA	small nuclear RNA
snRNP	small nuclear RNP
TAP	tandem affinity purification
TCA	trichloroacetic acid
TEMED	tetramethylethylenediamine

TFA	trifluoroacetic acid
THAP	2,4,5-trihydroxyacetophenone
Tris	tris(hydroxymethyl) aminomethane
tRNA	transfer RNA
UV	ultraviolet (light)
WB	Western blot
XIC	extracted ion chromatogram
YPD	yeast extract, peptone, dextrose

D Curriculum Vitae

Personal data

Name	Katharina Kramer
Date of birth	28.11.1983
Place of birth	Fulda
Nationality	German

Education

5/2009 – present	doctoral research study Bioanalytical Mass Spectrometry Group Max Planck Institute for Biophysical Chemistry, Göttingen, Germany
2/2009	Diploma in Chemistry Diploma thesis entitled: "H/D-Austausch-Massenspektrometrie zur Strukturanalytik von einem Peptidyl-Carrier-Protein und BabA"
10/2003 – 2/2009	Studies of Chemistry at the Philipps-Universität Marburg, Germany compulsory optional subject: Analytical Chemistry
6/2003	Abitur, Winfriedschule Fulda
8/2000 – 6/2001	exchange student, Center Grove High School, Greenwood, IN, USA
8/1994 – 6/2003	Gymnasium, Winfriedschule Fulda
8/1990 – 6/1994	Grundschule Langenbieber

Travel stipends

ASMS Sanibel Conference: "From Fragmentation Mechanisms to Sequencing: Tandem Mass Spectrometry Based Peptide and Protein Identification" (1/2011; awarded by GGNB)

60th ASMS Conference on Mass Spectrometry and Allied Topics (5/2012; awarded by GGNB)

17th Annual Meeting of the RNA Society (5-6/2012, awarded by the RNA Society)

EMBL symposium: "The Complex Life of mRNA" (10/2012; awarded by GGNB)

Oral conference presentation

"Investigation of protein–RNA cross-linking by mass spectrometry", 23.1.2011, ASMS Sanibel Conference, St. Pete Beach, FL, USA

Publications

Müller, M., R.G. Heym, A. Mayer, K. Kramer, M. Schmid, P. Cramer, H. Urlaub, R. Jansen and D. Niessing: *A cytoplasmic complex mediates specific mRNA recognition and localization in yeast.* PLoS Biol, 9(4):e1000611, Apr 2011.

Kramer, K., P. Hummel, H. Hsiao, X. Luo, M. Wahl and H. Urlaub: *Mass spectrometric analysis of proteins cross-linked to 4-thio-uracil- and 5-bromo-uracil-substituted RNA.* Int J Mass Spectrom, 304(2-3):184-194, Jul 2011.

Schmitzová, J., N. Rasche, O. Dybkov, K. Kramer, P. Fabrizio, H. Urlaub, R. Lührmann and V.Pena: *Crystal structure of Cwc2 reveals a novel architecture of a multipartite RNA-binding protein.* EMBO J, 31(9):2222-2234, May 2012.

Schmidt, C., K. Kramer and H. Urlaub: *Investigation of protein-RNA interactions by mass spectrometry — Techniques and applications.* J Proteomics, 75(12):3478-3494, Jun 2012.

Kramer, K.*, T. Sachsenberg*, S. Qamar, U. Zaman, K. Boon, R. Lührmann, O. Kohlbacher, H. Urlaub: *RNP^{xl}: A novel tool for the identification of peptides cross-linked to RNA by UV irradiation.* Manuscript in preparation (*equal contribution).