THE UNIVERSITY *of* EDINBURGH

3D proteomics: Analysis of proteins and protein complexes by chemical cross-linking and mass spectrometry

Zhuo A. Chen

Thesis for the Degree of Doctor of Philosophy

The University of Edinburgh

August 2011

DECLARATION

I hereby declare that the work presented in this thesis was carried out by me under the supervision of Prof. Juri Rappsilber at the University of Edinburgh between April 2007 and May 2011. No part of this thesis has been previously submitted at this or any other university for any other degree or professional qualification

Zhuo Chen

August 2011

ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisor Prof. Juri Rappsilber for his kind guidance, advice and continuous support during my Ph.D. It has been a great experience to be his student.

I also would like to thank everyone in the Rappsilber lab who has immensely contributed to my professional and personal time at the University of Edinburgh. Thanks to Lutz, Andy, Adam, Heather, Jimi, Karen, Lauri, Salman and Sally for correcting my writings. And thanks to everybody who helped me with my Ph.D.

I would like to thank my second supervisor, Professor Paul N Barlow, for his generous help on the C3 and C3b project. Thanks to Professor Patrick Cramer and his group for the collaboration on the Pol II-TFIIF project. I thank Dr.Kevin Hardwick, Sjaak van der Sar and Dr. Paul McLaughlin for their support on my work with the affinity purified protein complexes.

Big love to my family, especially my mum, without their support, I would not have managed my Ph.D.

CONTENTS

DECLARATION I

ACKNOWLEDGEMENTS II

LIST OF FIGURES X

LIST OF TABLES

ABBREVIATIONS XIII

ABSTRACT XV

Chapter 1 INTRODUCTION 1

1.1 Integrated structural biology and 3D proteomics 1

1.1.1 Integrated structural analysis of large protein complexes and assemblies 1

1.1.2 Applications of mass spectrometry in protein structural analysis 3

1.1.3 3D proteomics 4

1.2. Chemical cross-linking 8

1.2.1 Cross-linking reagents 8

1.2.1.1 Cross-linking chemistry 8

1.2.1.2 Cross-linking reagent design 15

1.2.1.3 Functionalized cross-linking reagents 16

1.2.2 Cross-linking reaction 18

1.2.3 In vivo cross-linking 20

1.3 Enrichment of cross-linked peptides 20

1.3.1 Separation and digestion of cross-linked protein samples 20

1.3.2 Enrichment of cross-linked peptides 23

1.4 Analysis of cross-linked peptides by mass spectrometry 24

1.4.1 Mass spectrometric analysis of cross-linked samples 24

1.4.2 Fragmentation of cross-linked peptides 27

1.5 Identification of cross-linked peptides 30

1.6 Current application of 3D proteomics 33

1.7 Project aim 36

Chapter 2 METHODS AND MATERIALS 37

2.1 Cross-linking analysis of synthetic peptides 37

2.1.1 Cross-linking of synthetic peptides 37

2.1.2 Strong cation exchange (SCX) fractionation 38

2.1.2.1 SCX-HPLC fractionation 38

2.1.2.2 SCX-StageTip fractionation 39

2.1.3 Analysis via Mass spectrometry 40

2.1.3.1 Sample preparation 40

2.1.3.2 LC-MS/MS analysis 40

2.1.4 Database searching 42

2.2 Cross-linking analysis of Pol II and Pol II-TFIIF complexes 44

2.2.1 The Pol II complex and the Pol II-TFIIF complex 44

2.2.2 Cross-linking titration of Pol II and Pol II-TFIIF complexes 45

2.2.3 Cross-linking of Pol II and Pol II-TFIIF complexes 48

2.2.4 Sample preparation for mass spectrometric analysis 48

2.2.5 Mass spectrometry 49

2.2.6 Database searching 50

2.3 Quantitative 3D proteomic analysis of C3 and C3b samples 51

2.3.1 Protein cross-linking for quantitative analysis 51

2.3.2 Sample preparation for mass spectrometric analysis 52

2.3.3 Mass spectrometric analysis 52

2.3.4 Identification of cross-linked peptides 53

2.3.5 Quantitation of cross-linkages 53

2.3.6 Comparison between cross-linking data and crystal structures 54

2.4. Structural analysis of affinity purified protein complexes by 3D proteomics 54

2.4.1 Affinity purified tagged endogenous protein complexes 54

2.4.2 'On-beads' cross-linking procedure 55

2.4.3 Sample preparation for mass spectrometric analysis 55

2.4.4 Mass spectrometric analysis 56

2.4.5 Database searching 56

2.4.6 Surveillance of inter-complex cross-links 57

2.5 Supplementary Information and experimental procedures 58

2.5.1 Supplementary Information 58

2.5.1.1 Supplier information 58

2.5.1.2 StageTips 58

2.5.2 Preparation of trypsin digested E.coli extract 58

2.5.2.1 Preparation of E.coli extract 58

2.5.2.2 In gel digestion of E.coli extract 59

2.5.3 Preparation of trypsin digested yeast extract 59

2.5.4 Protocol for silver staining 59

2.5.4.1 Solutions for silver staining 59

2.5.4.2 Silver staining procedure 60

Chapter 3 DEVELOPMENT OF A 3D PROTEOMICS ANALYTICAL WORKFLOW 61

3.1 Summary 61

3.2 Introduction 63

3.3 Analysis of cross-linked peptide library 65

3.3.1 Design of a cross-linked peptide library 65

3.3.2 LC-MS/MS analysis scheme for cross-linked peptides 67

3.3.3 Data base searching for cross-linked peptides 69

3.4 CID fragmentation of cross-linked peptides 70

3.4.1 Manual annotation of cross-linked peptide fragmentation spectra 70

3.4.2 High resolution fragmentation spectra of cross-linked peptides 70

3.4.3 The influence of different cross-linkers on the fragmentation of cross-linked peptides 74

3.4.4 The impact of resolution for MS^2 spectra on interpretation and identification of fragmentation spectra of cross-linked peptides 74

3.4.5 Automated interpretation of MS^2 spectra of cross-linked peptides 79

3.5 Validation of cross-linked peptide identification 79

3.5.1 Confidence criteria of cross-linked peptide identification 79

3.5.2 A large dataset of cross-linked peptides 80

3.6 Charge based enrichment strategy for cross-linked peptides 82

3.6.1 Strong cation exchange chromatography and cross-linked peptides enrichment 84

3.6.2 Selective fragmentation of highly charged precursor ions in mass spectrometric analysis increases detection of cross-linked peptides 85

3.7 Cross-linked peptide library and advanced 3D proteomics analytical workflow 89

3.8 Other applications of the cross-linked peptide library 89

Chapter 4 ARCHITECTURE OF THE RNA POLYMERASE II-TFIIF COMPLEX REVEALED BY 3D PROTEOMICS 91

4.1 Summary 91

4.2 Introduction 92

4.3 3D proteomics analysis of the Pol II complex 96

4.3.1 Cross-linking/MS analysis of the Pol II complex 96

4.3.2 Cross-linking and protein-protein interactions 98

4.4 Cross-linking/MS analysis of the Pol II-TFIIF complex 99

4.4.1 Cross-linking/MS data of the Pol II-TFIIF complex 99

4.4.2 Yeast TFIIF domain structures 102

4.4.3 Location of TFIIF on Pol II 104

4.4.4 Possible conformation changes of Pol II in the Pol II –TFIIF complex 109

4.5 Discussion 112

4.5.1 Architecture of the Pol II-TFIIF complex and TFIIF functions 112

4.5.2 Study architectures of large multi-protein complexes using 3D proteomics 115

Chapter 5 QUANTITATIVE 3D PROTEOMICS DETECTED CONFORMATIONAL DIFFERENCES BETWEEN C3 AND C3B IN SOLUTION AND GAVE INSIGHT INTO THE CONFORMATION OF SPONTANEOUSLY HYDROLYZED C3 117

5.1 Summary 117

5.2 Introduction 118

5.3 Quantitative 3D proteomics analysis of C3 and C3b samples 122

5.3.1 Cross-linking of C3 and C3b 122

5.3.2 Identification and quantitation of Cross-linked peptides 124

5.3.3 Quantified cross-linkages suggested differences between C3 and C3b samples 128

5.4 Quantitative cross-link data is in agreement with the crystal structures of C3 and C3b 129

5.4.1 Cross-linking data and the crystal structures agreed on residue proximity 129

5.4.2 Cross-linking data confirmed in solution the structural similarities and differences between C3 and C3b characterized by crystal structures 131

5.5 Quantitative cross-link data uncovered hydrolyzed C3 in the presence of C3 and C3b 136

5.6 Domain architecture of C3(H_2O) 141

5.7 Flexibility of the TED domain in C3b and C3(H_2O) 143

5.8 Cross-link data contradicts a false C3b crystal structure 144

5.9 Discussion 146

5.9.1 C3b-like functional domain arrangement and the function of C3(H_2O) 146

5.9.2 Outlook for quantitative 3D proteomics 147

Chapter 6 STRUCTURAL ANALYSIS OF TAGGED PROTEIN COMPLEXES BY 3D PROTEOMICS 148

6.1 Summary 148

6.2 Introduction 149

6.3 Cross-linking analysis of TAP-tagged endogenous protein complexes 150

6.3.1 'On-beads' cross-linking and digestion procedure 150

6.3.2 SILAC control experiments 153

6.4 Cross-links observed from low microgram amounts of endogenous protein complexes 155

6.4.1 Composition of purified tagged protein complex samples 155

6.4.2 Identification of cross-linked peptides from affinity purified complex samples 159

6.5 Organization of the Mad1-Mad2 complex 163

6.6 Cross-link data revealed a conserved loop region in Ndc80. 167

6.7 From AP-MS to AP-3DMS 172

Chapter 7 SUMMARY AND PERSPECTIVE 174

7.1 Summary 174

7.2 Perspective 176

APPENDIX 178

A.1 Observation of C3 contamination in the C3b sample 178

A.1.1 Detection of C3 contamination 178

A.1.1.1 Experimental procedure 178

A.1.1.1.1 Denaturing gel electrophoresis 178

A.1.1.1.2 Mass spectrometric analysis 178

A.1.1.2 Results 179

A.1.2 Quantitation of C3 contamination 180

A1.2.1 1 Experimental procedure 180

A1.2.2 Results 180

A.1.3 Discussion 180

A.2 Supplementary figures 184

A.3 Supplementary Tables 188

A.4 Publications 211

CITED LITERATURE 212

LIST OF FIGURES

Figure 1.1	Analytical strategies for 3D proteomics	5
Figure 1.2	Amine-reactive cross-linkers	10
Figure 1.3	Reaction scheme of sulfhydryl-reactive cross-linking with maleimides	11
Figure 1.4	Reaction schemes of a 'zero-length' cross-linker EDC including the reaction in combination with sulfo-NHS	12
Figure 1.5	Reaction schemes of most commonly used photoreactive cross-linking reagents	13
Figure 1.6	Chemical structures of four photoreactive amino acid analogues	14
Figure 1.7	Chemical structures of deuterated amine-reactive cross-linker BS^3-d4 in comparison with its unlabelled analogue BS^3-d0	17
Figure 1.8	Nomenclature of common products of chemical cross-linking reactions.	22
Figure 1.9	Fragment ions observed in MS^2 spectrum	28
Figure 2.1	Titration of BS3 cross-linking reactions for Pol II complex and Pol II-TFIIF complex	47
Figure 3.1	Design of the cross-linked peptide library	66
Figure 3.2	LTQ-Orbitrap hybrid mass spectrometer	68
Figure 3.3	Annotation of fragmentation spectra of cross-linked peptides	71
Figure 3.4	Peptide fragmentation patterns are similar in cross-linked and linear status	73
Figure 3.5	Impact of cross-linker on fragmentation	75
Figure 3.6	High and low resolution MS^2 spectra of cross-linked peptides	77
Figure 3.7	Validation of cross-linked peptide fragmentation spectra matches	81

Figure 3.8	Cross-linked peptide enrichment by SCX chromatographic fractionation	87
Figure 3.9	Precursor charge selection and cross-linked peptide enrichment	88
Figure 4.1	Important domains of Pol II	95
Figure 4.2	3D proteomics analysis of the Pol II complex	97
Figure 4.3	3D proteomics analysis reveals predominantly direct pairwise interaction between Pol II subunits.	100
Figure 4.4	Cross-linking reaction of Pol II –TFIIF complex	101
Figure 4.5	Cross-links observed within TFIIF and structures of TFIIF domains	103
Figure 4.6	Cross-links between Pol II and TFIIF	105
Figure 4.7	Cross-linking footprints of TFIIF subunits on the surface of Pol II structure	106
Figure 4.8	Alternative position of Tfg2 C-terminal region (linker, WH domain and C-terminal) on the Pol II surface	108
Figure 4.9	Architecture of Pol II-TFIIF in preinitiation complex	110
Figure 4.10	Cross-links within Pol II observed in Pol II-TFIIF complex	111
Figure 5.1	The experimental scheme of quantitative 3D proteomics analysis of C3 and C3b conformational changes in solution	123
Figure 5.2	Cross-linking of the C3 and C3b samples	125
Figure 5.3	Quantitation of cross-links	127
Figure 5.4	Cross-links observed in C3 and C3b samples	130
Figure 5.5	Quantitative cross-link data reflects similarities and differences between C3 and C3b	133
Figure 5.6	Domain architectures of C3 and C3b as derived from cross-link data	135
Figure 5.7	Quantitative cross-link data suggested that an alternative conformation existed in the C3 sample	137
Figure 5.8	Domain architecture of C3(H_2O)	142

Figure 5.9	Cross-link data contradicts a fraudulent C3b crystal structure	145
Figure 6.1	Workflow of the 'on-beads' process for 3D proteomics analysis	151
Figure 6.2	Scheme of SILAC control experiment for monitoring the occurrence of inter-complex cross-links	154
Figure 6.3	Validation of cross-linked peptide identification in MS^1 spectra	160
Figure 6.4	Spectra of cross-links between Mad1 molecules in the Mad1-Mad2 complex	165
Figure 6.5	Organization of the *S. cerevisiae* Mad1-Mad2 complex	166
Figure 6.6	Internal architecture of the S. cerevisiae Ndc80 complex	170
Figure 7.1	Draft of expected versatile applications of 3D proteomics in the future	177
Figure A1.1	SDS-PAGE gel image of the C3 and C3b	183
Figure A1.2	An example MS^1 spectrum of C3a peptide	183
Figure S1	Mass accuracy of Orbitrap mass analyzer at different resolutions	185
Figure S2	Inconsistency between crystallographic and cross-linking data on the Pol II complex	186

LIST OF TABLES

Table 1.1	Commonly used techniques for characterizing structures of protein complexes and protein assemblies	2
Table 2.1	SCX-StageTip fractionation	39
Table 2.2	Mass spectrometric acquisition methods for cross-linked synthetic peptide samples	42
Table 2.3	Search parameters for linear peptides samples in Mascot search	43
Table 2.4	Search parameters for cross-linked peptides samples in Xmass search	44
Table 2.5	Experimental plan for Pol II complex cross-linking titration	45
Table 2.6	Experimental plan for Pol II-TFIIF complex cross-linking titration	46
Table 2.7	Acquisition parameters for mass spectrometric analysis of the cross-linked Pol II and Pol II-TFIIF samples using the LTQ-Orbitrap mass spectrometer	50
Table 2.8	Search parameters used for database search for cross-linked peptides in Xi	51
Table 3.1	Summary of manually annotated cross-linked peptide identifications	83
Table 5.1	Interpretation of clustered cross-links	140
Table 6.1	Composition of affinity-purified protein complex samples	157
Table 6.2	Influence of sample amount on cross-linking detection	162
Table A.1.1	Identified C3a peptides from the C3b sample	181
Table A.1.2	Proteins identified from the C3b sample using Mascot	181
Table A1.3	Quantitation of cross-linker modified C3a peptides	182
Table S1	List of 49 synthetic peptides	189
Table S2	List of high confidence cross-links observed from the Pol II complex sample	191
Table S3	List of high confidence cross-links observed from the Pol	194

II-TFIIF complex sample

Table S4	Quantified cross-linkages in conformational comparison of C3 and C3b by quantitative 3D proteomics	204
Table S5	Ten most intense proteins identified from the affinity purified *S. cerevisiae* Mad1-Mad2 complex	206
Table S6	Ten most intense protein identified from the affinity purified *S. cerevisiae* Ndc80 complex	206
Table S7	List of cross-links observed from the affinity purified *S. cerevisiae* endogenous Mad1-Mad2 complex	207
Table S8	List of cross-links observed from the affinity purified *S. cerevisiae* endogenous Ndc80 complex	209

ABBREVIATIONS

1D	1 dimension
3D	3 dimension
ABC	ammonium bicarbonate
ACN	acetonitrile
AP-MS	affinity purification-mass spectrometry
BS^2G	Bis[sulfosuccinimidyl] glutarate
BS^3	Bis[sulfosuccinimidyl] suberate
CID	collision-induced dissociation
DEB	1,3-diformyl-5-ethynylbenzene
DMF	*N*,*N*-dimethylformamide
DMSO	dimethyl sulfoxide
DPI	dual polarization interferometry
DSG	disuccinimidyl glutarate
DSS	disuccinimidyl suberate
DTT	dithiothreitol
EDC	1-ethyl-3-[3-dimethylaminopropyl]carbodiimide hydrochloride
EM	electron microscope
ESI	electrospray ionization
ET	electron transfer
ETD	electron-transfer dissociation
FDR	false discovery rate
FP	fluorescence polarization
FRET	fluorescence resonance energy transfer
FT	Fourier transform
FTICR	Fourier transform ion cyclotron resonance mass spectrometry
HPLC	high-performance liquid chromatography
IAA	iodoacetamide
LC-MS/MS	liquid chromatography–tandem mass spectrometry
LIT	linear ion trap
LRET	luminescence resonance energy transfer

LTQ	linear trap quadrupole
MALDI	matrix-assisted laser desorption/ionization
MES	2-(*N*-morpholino)ethanesulfonic acid
MOPS	3-(N-morpholino)propanesulfonic acid
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MS^1	full scan (spectrum)
MS^2	fragmentation scan (spectrum)
NHS-ester	N-hydroxysuccinimide ester
NMR	nuclear magnetic resonance
PIC	preinitiation complex
PIR	protein interaction reporter
Pol II	RNA polymerase II
PTM	post translational modification
-Q-	quadrupole
RNA	ribonucleic acid
SBC	*N*-succinimidyl *p*-benzoyldihydrocinnamate
SCX	strong cation exchange
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SILAC	stable isotope labelling with amino acids in cell culture
Stage-Tip	stop-and-go-extraction tips
Sulfo-SMCC	sulfosuccinimidyl-4-(*N*-maleimidomethyl)cyclohexane-1-carboxylate
TEA	thriethanolamine
TFA	trifluoroacetic acid
TFIIB	transcription factor IIB
TFIID	transcription factor IID
TFIIF	transcription factor IIF
-TOF	time-of-flight mass spectrometry
Tris	2-Amino-2-hydroxymethyl-propane-1,3-diol
UV	ultraviolet
XDB	cross-link database

ABSTRACT

The concept of 3D proteomics is a technique that couples chemical cross-linking with mass spectrometry and has emerged as a tool to study protein conformations and protein-protein interactions. In this thesis I present my work on improving the analytical workflow and developing applications for 3D proteomics in the structural analysis of proteins and protein complexes through four major tasks.

I. As part of the technical development of an analytical workflow for 3D proteomics, a cross-linked peptide library was created by cross-linking a mixture of synthetic peptides. Analysis of this library generated a large dataset of cross-linked peptides. Characterizing the general features of cross-linked peptides using this dataset allowed me to optimize the settings for mass spectrometric analysis and to establish a charge based enrichment strategy for cross-linked peptides. In addition to this, 1185 manually validated high resolution fragmentation spectra gave an insight into general fragmentation behaviours of cross-linked peptides and facilitated the development of a cross-linked peptide search algorithm.

II. The advanced 3D proteomics workflow was applied to study the architecture of the 670 kDa 15-subunit Pol II-TFIIF complex. This work established 3D proteomics as a structure analysis tool for large multi-protein complexes. The methodology was validated by comparing 3D proteomics analysis results and the X-ray crystallographic data on the 12-subunit Pol II core complex. Cross-links observed from the Pol II–TFIIF complex revealed interactions between the Pol II and TFIIF at the peptide level, which also reflected the dynamic nature of Pol II -TFIIF structure and implied possible Pol II conformational changes induced by TFIIF binding.

III. Conformational changes of flexible protein molecules are often associated with specific functions of proteins or protein complexes. To quantitatively measure the differences between protein conformations, I developed a quantitative 3D proteomics strategy which combines isotope labelling and cross-linking with mass spectrometry and

database searching. I applied this approach to detect in solution the conformational differences between complement component C3 and its active form C3b in solution. The quantitative cross-link data confirmed the previous observation made by X-ray crystallography. Moreover, this analysis detected the spontaneous hydrolysis of C3 in both C3 and C3b samples. The architecture of hydrolyzed C3 -C3(H_2O) was proposed based on the quantified cross-links and crystal structure of C3 and C3b, which revealed that C3(H_2O) adopted the functional domain arrangement of C3b. This work demonstrated that quantitative 3D proteomics is a valuable tool for conformational analysis of proteins and protein complexes.

IV. Encouraged by the achievements in the above applications with relatively large amounts of highly purified material, I explored the application of 3D proteomics on affinity purified tagged endogenous protein complexes. Using an on-beads process which connected cross-linking and an affinity purification step directly, provided increased sensitivity through minimized sample handling. A charge-based enrichment step was carried out to improve the detection of cross-linked peptides. The occurrence of cross-links between complexes was monitored by a SILAC based control. Cross-links observed from low micro-gram amounts of single-step purified endogenous protein complexes provided insights into the structural organization of the *S. cerevisiae* Mad1-Mad2 complex and revealed a conserved coiled-coil interruption in the *S. cerevisiae* Ndc80 complex.

With this endeavour I have demonstrated that 3D proteomics has become a valuable tool for studying structure of proteins and protein complexes.

Chapter 1

INTRODUCTION

1.1 Integrated structural biology and 3D proteomics

1.1.1 Integrated structural analysis of large protein complexes and assemblies

Protein complexes and their network of interactions play essential roles in cellular function and regulation. Structural characterization of protein complexes and large protein assemblies underline the mechanistic understanding of cellular processes. To properly characterize the structure of a protein complex or assembly, the following information is required:

1) Characters of all subunits
2) Stoichiometry of subunits in the protein complex (protein assembly)
3) Assembling of subunits
4) Structural dynamics of the protein complex (protein assembly).

Rarely, single structural biology techniques alone can achieve such comprehensive characterization, especially for large protein complexes and assemblies. However, these structural information can be gathered using different techniques. These include high and low resolution structural biology techniques such as X-ray crystallography, nuclear magnetic resonance (NMR), electron microscopy, electron tomography, small angle scattering, mass spectroscopy and advanced light microscopy. In addition a wide range of physical, chemical, biochemical, molecular biological characterization and computational techniques can be used (Sali *et al.*, 2003) (Table 1.1). Moreover, computational tools that can integrate all this

information for modelling structures of protein complexes and assemblies have become available in recent years (Sali *et al.*, 2003; Alber *et al.*, 2007).

Table 1.1 - Commonly used techniques for characterizing structures of protein complexes and protein assemblies.

Structural features		**Commonly used techniques**
Characters of subunits	Subunit primary sequence	Edman sequencing, Mass spectrometry
	PTMs	Mass spectrometry
	Subunit shape	X-ray crystallography, NMR, Electron microscopy, Electron tomography, Protein structure prediction, Small angle scattering, Ion mobility-mass spectrometry.
	Subunit structure	X-ray crystallography, NMR, Protein structure prediction
Stoichiometry of subunits		X-ray crystallography, Quantitative proteomics analysis, Quantitative immuno-blotting.
Assembling of subunits	Subunit-subunit contact	X-ray crystallography, NMR, Electron microscopy, Electron tomography, Mass spectrometry, Chemical cross-linking/MS, Affinity purification-mass spectrometry, FRET, Site-directed mutagenesis, Yeast two-hybrid system, Computational docking
	Subunit proximity	X-ray crystallography, Electron microscopy, Electron tomography, Immuno-eletron microscopy, Chemical cross-linking/MS, Affinity purification-mass spectrometry, FRET, Yeast two-hybrid system
	Assembly structure	X-ray crystallography
	Assembly shape	X-ray crystallography, NMR, Electron microscopy, Electron tomography, Small angle scattering
	Assembly symmetry	X-ray crystallography, NMR, Electron microscopy, Electron tomography, Immuno-eletron microscopy, Small angle scattering
Dynamics of assemblies	Compositional dynamics	Affinity purification-mass spectrometry, Quantitative proteomics
	Conformational dynamics	X-ray crystallography, NMR, Electron microscopy, Electron tomography, Small angle scattering, Chemical-cross-linking/MS, Light microscopy techniques

1.1.2 Applications of mass spectrometry in protein structural analysis.

Today mass spectrometry plays important roles in structural biology studies. Mass spectrometry based proteomics has been very successful in identifying proteins in complexes and organelle, and hundreds of proteins can now be analyzed in a single experiment (Aebersold and Mann, 2003).Additionally, mass spectrometry has also been able to reveal protein post-translational modifications (PTMs) (Mann and Jensen, 2003) which often play important roles in dynamics of protein structures. Consequentially mass spectrometry has become a key tool for studying primary protein structures. Its combination with affinity purification (AP-MS) has significantly advanced our understanding of protein complex composition (Gingras *et al.*, 2007).

However, applications of mass spectrometry have not been restricted to analyzing protein primary sequences. Mass spectrometric analysis of intact and partially disassociated protein complexes can provide information on subunit packing and interaction networks (Zhou and Robinson, 2010). Applications of ion mobility mass spectrometry on intact protein complexes and subunits may give rise to additional topology constraints for structural modelling of protein complexes (Ruotolo *et al.*, 2008; Jurneczko and Barran, 2011).

In the past decade, chemical cross-linking has been introduced to mass spectrometry based proteomics workflows, which have provided constraints on residue proximity in native structures of proteins and protein complexes. Distinguished from standard proteomics, which focuses on detecting primary sequences of proteins, this new cross-linking/MS approach provides additional information on spatial folding of proteins and protein-protein interactions. As a consequence, in this thesis, it has been designated with the term 3D proteomics. In recent applications, 3D proteomics data has played an essential role in integrated structural analysis of the Pol II-TFIIF complex (Chen *et al.*, 2010) and the 26S proteasome (Bohn *et al.*, 2010).

1.1.3 3D proteomics

As a technique for studying the structure of proteins and protein complexes, 3D proteomics consists of two major elements: chemical cross-linking and identification of cross-linked residues using mass spectrometry. Chemical cross-linking is aimed to convert proximity between amino acid residues in native protein structures and non-covalent protein-protein interactions into stable covalent bonds with distance constraints. Tracing back to 1970s, cross-linking treatment has been used in combination with electrophoretic analysis to study protein-protein interaction in ribosome (Clegg and Hayes, 1974; Sun *et al.*, 1974). Currently it is also used to stabilize protein complexes for electron microscopies analysis and affinity purifications (Gingras *et al.*, 2007). However, the identification of cross-links was not reported until the end of the1990s (Rappsilber *et al.*, 2000; Young *et al.*, 2000). Over the past 20 years, a series of technical breakthroughs made mass spectrometry an indispensable tool in proteomics and in all fields of the life sciences. Mass spectrometry provides amazing power to study protein sequences and determine protein modifications which also make it possible to reveal the location of cross-links in protein sequences. Cross-linked residue pairs with distance constraint carry much structural information of proteins and protein complexes, such as low resolution protein folding, topology of protein complexes and transient protein-protein interactions.

In order to identify cross-links, the technique of shotgun proteomics has been adopted for mass spectrometric analysis. In this strategy, cross-linked proteins are enzymatically digested into peptides and then analyzed by mass spectrometry. The cross-linked peptides are subsequently identified through database searching and linkage sites are assigned based on fragmentation data of the cross-linked peptides. This strategy is also known as the 'bottom-up' approach (Figure 1.1).

There is another strategy for mass spectrometric analysis of cross-linked proteins, which is the 'top-down' approach. In this technique intact cross-linked proteins are analyzed.

The accurate measurement of the mass of proteins reveals the number of cross-links occurred. The cross-linked residues are assigned based on fragmentation information. So far applications of this approach are only restricted to single purified proteins. This approach is not employed and will not be discussed further in this thesis (Figure 1.1).

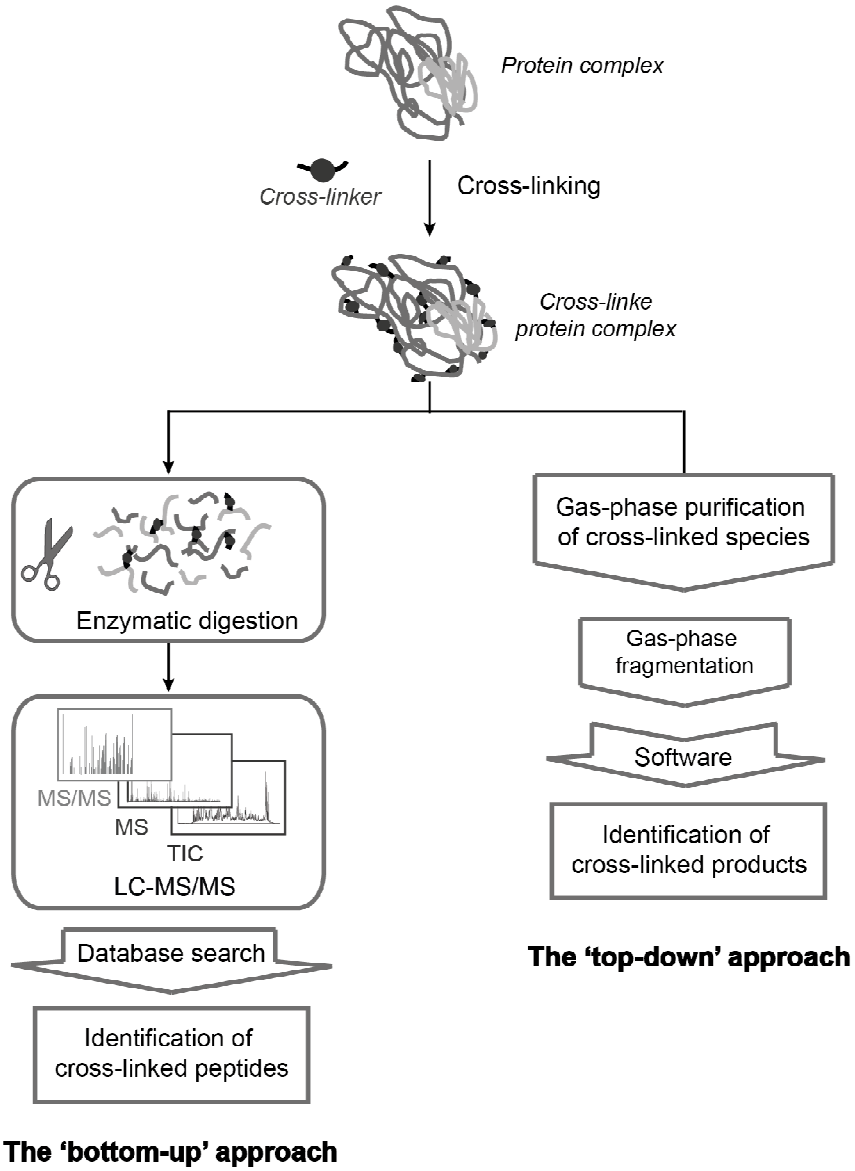


Figure 1.1 - Analytical strategies for 3D proteomics.

The 'bottom-up (left)' and the 'top-down' strategies for 3D proteomics analysis are demonstrated with a protein complex sample.

As with any technique, 3D proteomics has its strengths and limitations. The principle of 3D proteomics conveys several inherent advantages:

1) Proteins and protein complexes are studied in solution under favourable circumstances that are close to physiological condition (in terms of pH, ion strength *etc.*).

2) 3D proteomics is applicable to wide range of structural motifs, including the otherwise hard to study coiled-coil structures (Maiolica *et al.*, 2007) and flexible loop regions. However some folding is required to obtain specific cross-link data (Chen *et al.*, 2010).

3) The cross-linked proteins and protein complexes are analyzed as proteolytic peptides. Theoretically the mass and size of analyzed protein and protein complexes are not limited. Protein post translational modifications are maintained and can be identified by mass spectrometry.

4) Sample heterogeneity caused by the existence of multiple conformations or other proteins will increase the complexity of a sample and challenge the detection and data processing. However they will not principally impair the analysis (Rappsilber, 2011).

5) Analysis is generally fast, and requires only femtomole to picomole amounts of material.

6) There is a wide range of cross-linking reagents with different reaction specificities and spacers which offer the possibility to perform a wide range of experiments (Huermanson, 1996).

Inevitably, these advantages are accompanied by several inherent disadvantages:

1) 3D proteomics analysis gives rise to paired residues with distance constraints which only provide only low resolution structural information.

2) Non-homogeneous distribution as well as variable availabilities and accessibility of reactive sites in protein structures can lead to patchy incomplete nature of cross-linking data. However, applications of different cross-linking chemistries can to some extent increase the coverage of cross-linking data for a protein structure.

3) The structure of proteins and protein complexes are captured *via* chemical cross-linking reactions. The speed of these reactions place limits on the time scale of protein conformations and protein-protein interactions that can be characterized by 3D proteomics.

4) Multiple conformations of a protein will not be distinguished by standard 3D proteomics analysis, since mass spectrometry detects populations other than individuals. Instead, they will be detected as an overlapped image.

Despite these disadvantages, 3D proteomics still can be a powerful tool for studying the structure of proteins and protein complexes, especially due to its great potential on studying large protein complexes and high throughput analysis. However two major technical challenges have impeded the application of this technique to complex protein samples. The first is the difficulty in detecting the relatively low stoichiometric cross-linked peptides in mixtures with a large excess of non-cross-linked linear peptides. Secondly, the quadratically expanded search space that accompanies increased sample complexity poses a computational challenge for a search algorithm to correctly identify cross-linked peptides (Rinner *et al.*, 2008; Rappsilber, 2011). In the past ten years, progress has been made by our group and others to overcome these technical limitations and technical developments are still ongoing. The evolution of the field in the last decade was reviewed by (Young *et al.*, 2000; Back *et al.*, 2003; Sinz, 2006; Jin Lee, 2008; Leitner *et al.*, 2010; Singh *et al.*, 2010; Sinz,

2010). In the following stages, I will introduce the developments which took place in each step of the analytical workflow which typically included cross-linking reactions, protein digestion, mass spectrometric analysis and identification of cross-linked peptides.

1.2 Chemical cross-linking

The main purpose of chemical cross-linking is to generate covalent bonds between two spatially proximate residues within or between protein molecules. This process involves amino acids (normally through their side chains) and a cross-linker. A typical cross-linker contains two reactive groups that are connected by a spacer. Cross-linkers typically react with functional groups in amino acids (*e.g*. primary amine, sulfhydryls, and carboxylic acid) which result in bridges between residues. The maximum distance between cross-linked residues is defined by the length of the spacers. Recently a number of reviews have been published focusing on chemical cross-linking reagents and application protocols (Brunner, 1993; Kluger and Alagic, 2004; Melcher, 2004; Kodadek *et al.*, 2005; Sinz, 2006).

1.2.1 Cross-linking reagents

1.2.1.1 Cross-linking chemistry

There are hundreds of cross-linkers described in the literature (Wong, 1991; Huermanson, 1996) and offered commercially, however they are only based on several different organic chemical reactions.

I. Amine-reactive cross-linkers

In protein molecules, the most common target for cross-linking reactions are primary amine groups, such as free *N*-terminus and ε-amino groups in lysine side chains. Amine group targeted cross-linking takes advantage of high frequency (>6%) of lysine residue in proteins which consequently increases the yield of cross-links.

i) *N-hydroxysuccinimide (NHS) esters*. N-hydroxysuccinimide (NHS) esters are almost exclusively used as reactive groups for amine reactive cross-linkers. They react with nucleophiles to release the NHS group to create stable amide and imide bonds with primary or secondary amines (Sinz, 2006) (Figure 1.2 A). Many NHS esters are insoluble in aqueous buffers and need to be dissolved in a small volume of an organic solvent such as DMSO or DMF before being added to the sample in an aqueous buffer. Alternatively, the sulfo analogues of NHS esters (sulfo-NHS) are used since they are more water-soluble (Figure 1.2 C). NHS esters have high reaction rates with amine groups, but at the same time they are susceptible to rapid hydrolysis with a half-life in the order of hours under physiological pH conditions (pH 7.0–7.5). Both hydrolysis and amine reactivity increase when the pH and temperature are raised (Huermanson, 1996). The hydrolysis of NHS esters limits the cross-linking reaction time and reduces the yield of desired cross-linking products. Side reactions of NHS ester with serine, threonine and tyrosine residues have been reported however under alkaline conditions (pH 8.4) they were found to react preferentially with the *N*-terminus and lysine amine groups. Under carefully controlled reaction condition (pH, protein to reagent ratio, and reaction time) the side reactions may not occur at relevant level (Chen *et al.*, 2010).

ii) *Imidoesters*. Imidoesters are also used to construct cross-linkers for protein conjugation (Figure 1.2B). The imidate functional group has high specificity towards primary amines. However at physiological pH, imidoesters have a lower cross-linking efficiency than NHS esters (Dihazi and Sinz, 2003) (Sinz, 2006).

iii) *Other amine-reactive cross-linkers.* Recently new amine specific cross-linkers using N-hydroxyphthalimide, hydroxybenzotriazole, and 1-hydroxy-7-azabenzotriazole as function groups were reported to react 10 time faster and with higher efficiency than NHS esters in comparison to disuccinimidyl suberate (DSS) (Bich *et al.*, 2010).

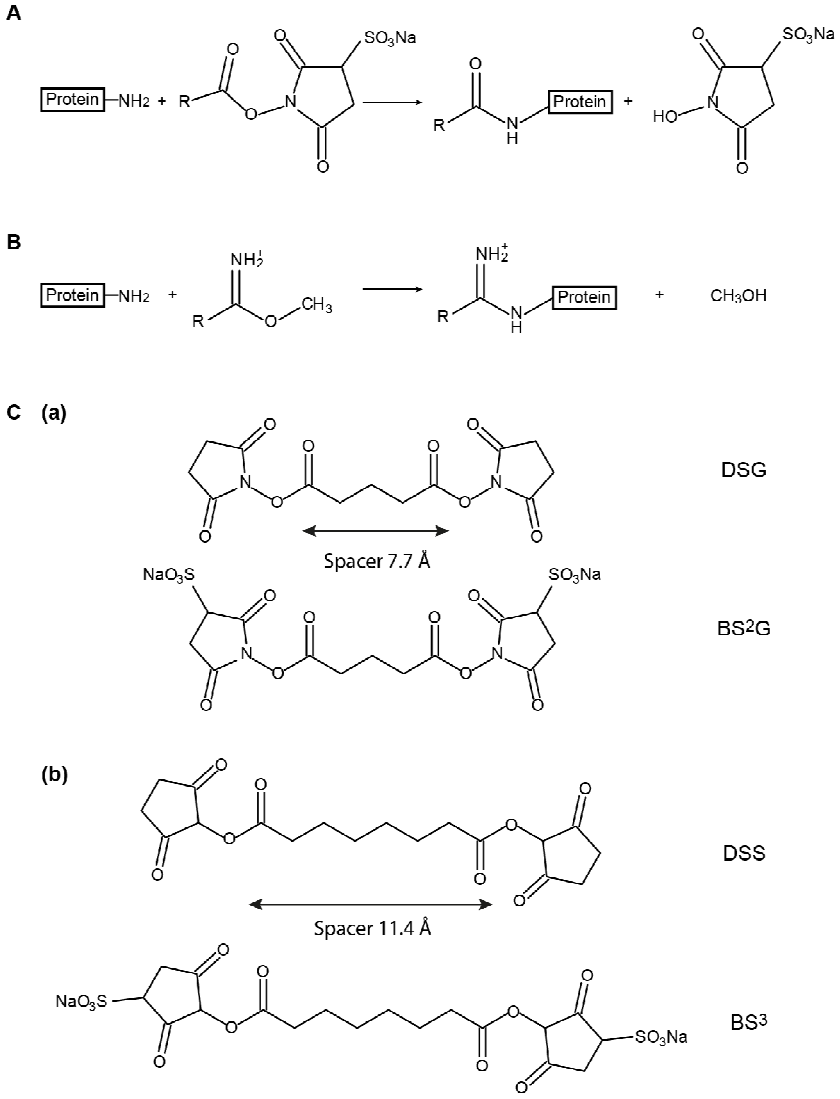


Figure 1.2 - Amine-reactive cross-linkers.

Reaction schemes of two commonly applied amine-reactive cross-linking reagents are shown in A (NHS ester) and B (imidates). Chemical structures of two most commonly used amine-reactive cross-linkers, DSB (a) and DSS (b), and their sulfo analogues BS^2G and BS^3 are shown in C.

II. Sulfhydryl-reactive cross-linkers

Alternatively, the cross-linking reaction can target on sulfhydryl group (cysteine side chain). The commonly used maleimides have rather high specificity towards sulfhydryls (Figure 1.3) at pH range of 6.5 to 7.5, but especially at pH7. However the low abundance of cysteine (<2%) and frequent involvement of sulfhydryl group in disulfide-bond formation in native protein structure make this option less attractive.

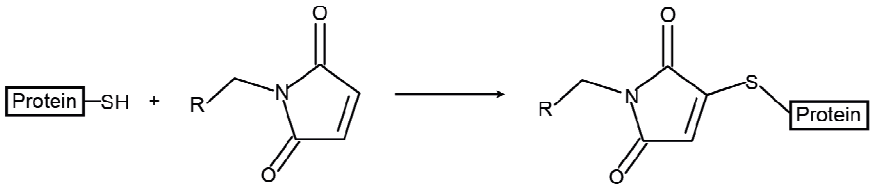


Figure 1.3 - Reaction scheme of sulfhydryl-reactive cross-linking with maleimides.

III Zero-length cross-linkers

Carbodiimides such as 1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) can mediate amide bond formation between carboxylic acids (aspartate, glutamate, protein C-terminus) and amines (lysine, protein N-terminus) without introducing a spacer chain into the protein. Therefore they were called 'zero-length' cross-linkers. Zero-length cross-linking requires very close proximity between linked function groups (<3 Å). A second reagent such as sulfo-NHS ester could be added to improve the cross-linking efficiency (Pierce 2003/2004; Sinz 2006) (Figure 1.4).

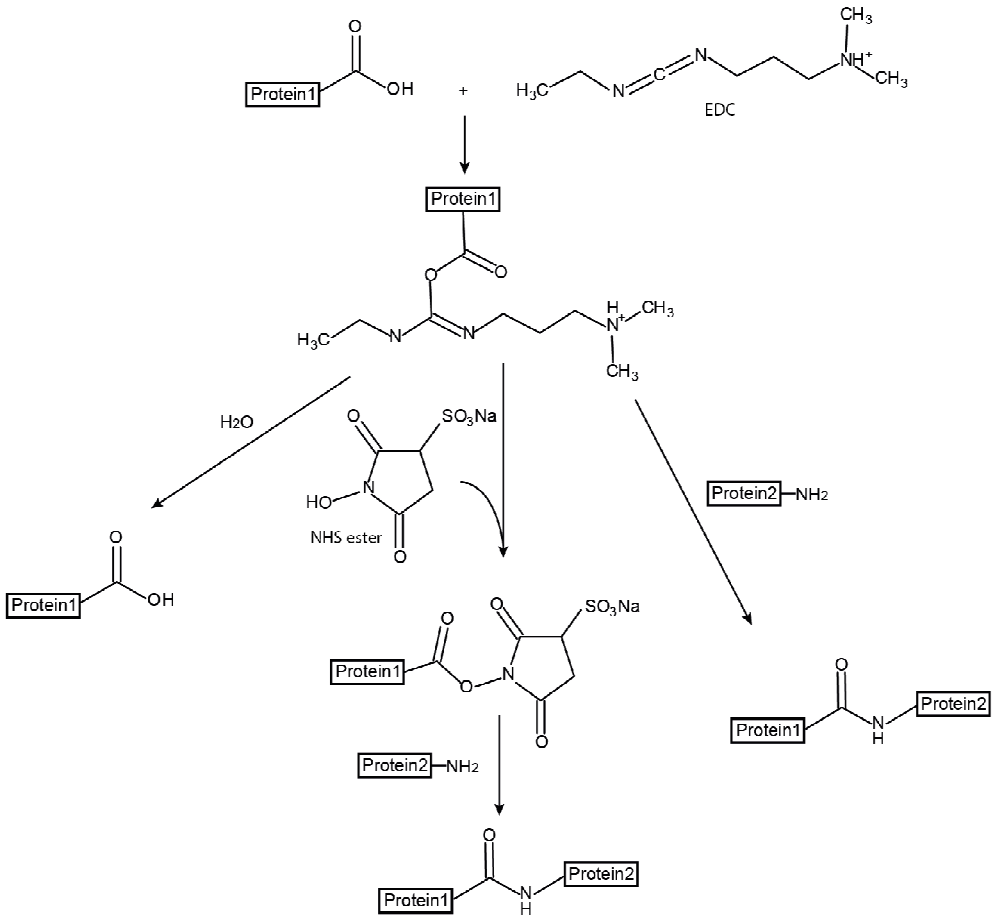


Figure 1.4 - Reaction schemes of a 'zero-length' cross-linker EDC including the reaction in combination with sulfo-NHS.

IV Formaldehyde

Formaldehyde is often used to rapidly cross-link protein complexes. It contains a single aldehyde group, connecting two amino acid side chains via a two-step reaction (Leitner *et al.*, 2010). Formaldehyde has low specificity towards individual amino acid residues. There is no report about its use in cross-linking sites analysis (Jin Lee, 2008). Recent investigation discovered that lysine, tryptophan and protein termini were primarily targeted when limited to formaldehyde exposure for 10 min (Sutherland *et al.*, 2008).

V. Photoreactive

Photoreactive cross-linkers can react with target molecules when induced by exposure to UV light. Aryl azides (also called phenylazides) (Figure 1.5A) were the most popular photo-reactive chemical group used in cross-linking; diazirines (Figure 1.5 B) are a new class of photo-reactive chemical groups with better photostability than phenyl azide groups and more easily and efficiently activated with long-wave UV light. Both of them have no specificity towards certain functional groups (Pierce, 2003/2004). Benzophenones (Figure 1.5 C) have a completely different photochemistry compared to former two reactive groups, and show a certain specificity towards methionine (Sinz, 2006) (Wittelsberger *et al.*, 2006). Photoreactive cross-linkers are mostly heterobifunctional reagents, with the other end targeting the amine or sulfhydryl group, and react in a stepwise manner (Pierce, 2003/2004). For example the NHS ester first reacts with primary amine in the protein molecule followed by a reaction of the photoreactive benzophenone moiety to a nearby residue that is induced by UV irradiation (Krauth *et al.*, 2009).

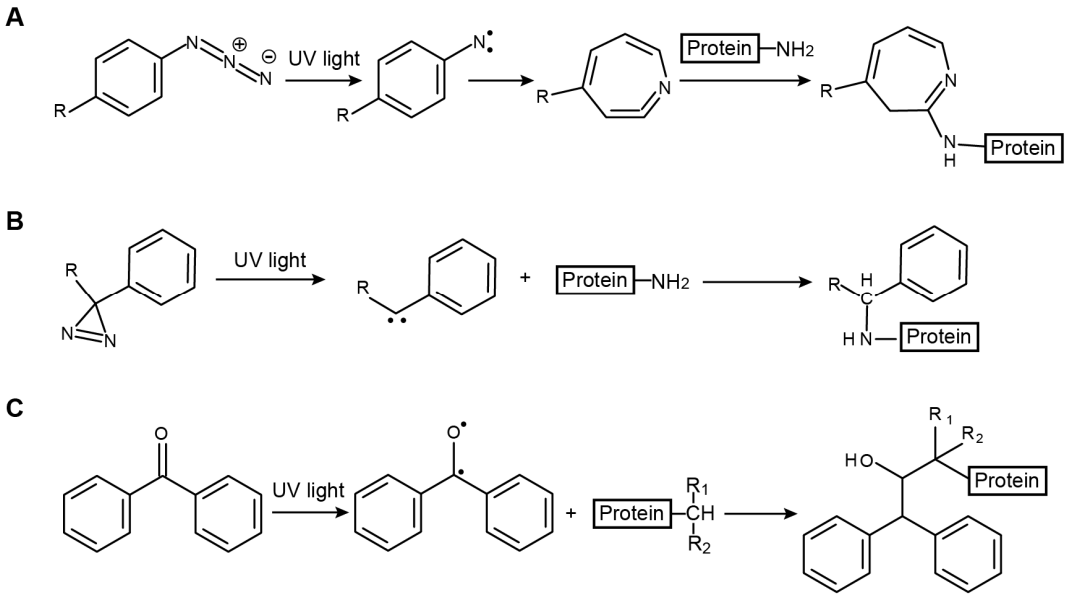


Figure 1.5 - Reaction schemes of most commonly used photoreactive cross-linking reagents.

A. Aryla azides; B. Diazirines and C. Benzophenones

VI Photoreactive amino acid analogues

Recently, another interesting approach has been introduced, the incorporation of photoreactive amino acid analogues into the protein sequence. Photo-methionine, photo-leucine and photo-isoleucine (Figure 1.6) were incorporated into proteins by the cell's normal translation machinery due to their structural similarity to the natural amino acids. Activation by UV light induced covalent cross-linking of interacting proteins (Suchanek *et al.*, 2005). Vila-Perello and co-workers introduced photo-Met and phospho-Ser into multiple sites of Smad2 HM2 domain using semi-synthesis. By activating the photo-Met, the transient phosphorylation dependent protein–protein interactions were covalently captured by photo-cross-linking (Vila-Perello *et al.*, 2007)…Incorporation of another non-natural photoreactive amino acid *p*-benzoyl-L-phenylalanine (Bpa) (Figure 1.6) was applied to reveal the interaction between transcription factor IIF on RNA polymerase II surface (Chen *et al.*, 2007)

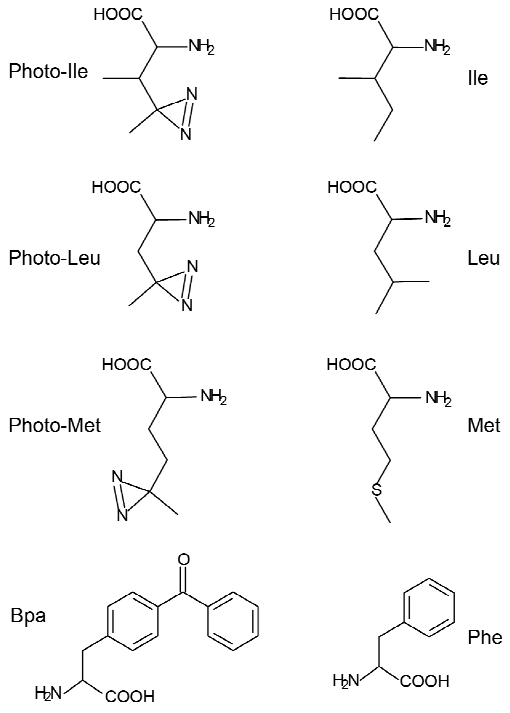


Figure 1.6 - Chemical structures of four photoreactive amino acid analogues

Chemical structures of 'Photo-Ile', 'Photo-Leu', 'Photo-Met' and Bpa (left) are shown in comparison to the natural amino acid Ile, Leu, Met and Phe (right).

1.2.1.2 Cross-linking reagents design

Conventional cross-linkers typically contain a spacer and two reactive groups at each end. Homobifunctional cross-linkers have identical reactive groups at either end of a spacer while heterobifunctional cross-linkers possess different reactive groups at either end. Homobifunctional cross-linkers have the advantage of single step conjugation. Heterobifunctional cross-linkers require for sequential (two-step) reaction. However this can minimize undesirable polymerization or self-conjugation. The most widely used heterobifunctional cross-linkers are those with an amine-reactive at one end and a sulfhydryl-reactive group on the other end, for example in Sulfo-SMCC, the unstable NHS ester is reacted first, subsequently the maleimide group is reacted after the removal of excessive cross-linkers (Lee *et al.*, 2007). Cross-linkers may also contain three reactive groups. However they have not been used in structural analysis so far, mainly because the identification of cross-linked peptides involving three cross-linked residues presents a huge challenge (Rappsilber, 2011). Therefore, most of the trifunctional cross-linkers used in 3D proteomics analysis have affinity or antibody handles as the third functionality, for the purpose of enrichment (further discussed in *1.2.1.3*).

The spacer of a cross-linker is typically an alkyl chain. Its length can affect solubility of a cross-linker and determines the distance constraint between cross-linked residues. The scale of this distance constraint is essential for structural analysis. As described before, the short cross-linkers such as the zero-length EDC require close proximity between cross-linked functional groups, which may result in low reaction efficiency. Generally, longer spacers allow for more residue pairs in protein structures to be cross-linked. However, an increase in spacer length will reduce the accuracy in determining the spatial distance between cross-linked residues. A linker with a ~8-15Å distance is the preferred length, as it is considered to provide the most useful distance geometry information for the threading calculation (Collins *et al.*, 2003). Currently the most widely used cross-

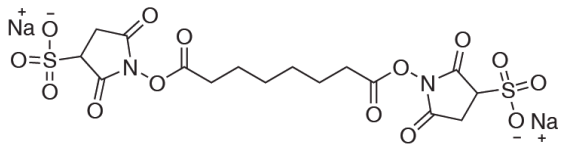
linkers are disuccinimidyl suberate (DSS,, spacer length 11.4Å) and disuccinimidyl glutarate (DSG, spacer length 7.7Å) as well as their sulfo analogues bis(sulfosuccinimidyl) suberate (BS^3) and bis(sulfosuccinimidyl) glutarate (BS^2G). Considering the length of lysine side chain of ~6 Å and it flexibility, the theoretical distance between two alpha-carbon (C-α) atoms of cross-linked residues can reach 24 Å for DSS (BS^3) and 19 Å for DSG (BS^2G). In the literature, the maximum cross-linkable distances of cross-linkers are often defined as the distances between the two reactive groups in a fully extended conformation (Pierce Chemical Company). However, stochastic molecular dynamics simulations showed that cross-linkers can achieve a broader range of end-to-end distances (Green *et al.*, 2001). When mapped onto the crystal structure, the measured distances of 108 experimentally observed BS^3 cross-links from the Pol II complex displayed a natural distribution between 6 and 29 Å, central at ~16 Å (Chen *et al.*, 2010). In the literature, it is frequently proposed that using cross-linkers with same chemistry but different spacer length may refine the distance constraints. However, when Leitner and co-workers cross-linked 7 proteins with known 3D structures with DSS and DSG, the distances of cross-linked residues determined from PDB data did not show differences between these two cross-linkers. Only fewer cross-links were observed with the DSG experiment.

1.2.1.3 Functionalized cross-linking reagents

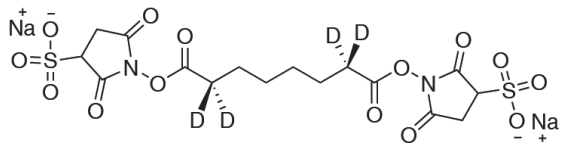
Besides the conventional cross-linking reactivity, additional functions have been introduced into cross-linking reagents to facilitate the analysis of cross-linking products by mass spectrometry. These include stable isotope-labelled cross-linkers, cross-linkers with affinity tags and cleavable cross-linkers.

Cross-linking using a 1:1 mixture of stable isotope labelled (heavy) cross-linkers and their mono-isotopic (light) form were introduced first by Muller *et al.* (Mueller *et al.*, 2001). The cross-linking products will display a distinctive isotopic signature in the mass spectra.

Different types of stable isotope labelled cross-linkers can be obtained commercially from several suppliers, such as Creative Molecules and Thermo Scientific. The most common products are deuterated BS^3 and BS^2G (BS^3-d4 and BS^2G -d4) (Figure 1.7). Cross-linking with an equal amount mixture of BS^3-d0 and BS^3-d4 followed by enzymatic digestion results in doublet signals in the MS^1 spectra with a 4 Da difference for the cross-linker containing species. This allowed them to be detected easily, even if they occurred with low abundance (Schmidt *et al.*, 2005). However, for the large (> 2 kDa) cross-linked peptides, it is harder to distinguish the isotope clusters of heavy and light species with 4 Da distance, as the isotope clusters might become overlaid. Moreover, the dilution of cross-linked peptide abundance may to some extent reduce the sensitivity of detection (Lee *et al.*, 2007).



BS3-d0
M.W. 572.43
Spacer arm 11.4Å



BS3-d4
M.W. 576.45
Spacer arm 11.4Å

Figure 1.7 - Chemical structures of deuterated amine-reactive cross-linker BS^3-d4 in comparison with its unlabelled analogue BS^3-d0.

Affinity tags were introduced to the cross-linkers in addition to two reactive groups for enrichment of the cross-linker containing species. The biotin group is frequently used and can be purified through avidin affinity chromatography (Trester-Zedlitz *et al.*, 2003; Kang *et al.*, 2009). With a different chemistry, another reported enrichment method was based on the covalent capture of azide-containing-cross-linker reacted peptides by azide-reactive cyclooctyne resin (Nessen *et al.*, 2009). In another strategy, peptides modified by an amine specific cross-linker that carried a thiol group were enriched using beads that were modified by a cross-linker with a reactive iodoacetyl group and an additional photocleavage site (Yan *et al.*, 2009). Currently, the application of affinity cross-linkers is still restricted to model studies, mainly due to the complicated synthesis and the risk of steric hindrance cause by the large affinity groups (Leitner *et al.*, 2010).

Cross-linkers that contain labile bonds can be easily cleaved during MS/MS experiments, for example the protein interaction reporter (PIR) introduced by Bruce and co-workers (Anderson *et al.*, 2007). The cleavage of cross-linker bonds is normally involved in the release of cross-linked peptides and the generation of diagnostic ions. The sequence information for individual peptides may be obtained using MS3 experiments. It is common that several functional designs can be combined in the one cross-linker. The PIR mentioned above also contained a biotin affinity tag. Petrotchenko *et al.* also reported several multifunctional cross-linkers (Petrotchenko *et al.*, 2005; Petrotchenko *et al.*, 2009; Petrotchenko *et al.*, 2011). However, wide applications of these newly developed cross-linkers have not yet been reported.

1.2.2 Cross-linking reaction

There is no standard protocol for cross-linking as reaction conditions may vary depending on different reagents and applications. However for a successful experiment, the cross-linking condition must be carefully controlled in order to yield appropriate cross-linking products for

structural analyses. There are several key parameters that need to be considered to refine and optimize reactions:

1) Buffer pH and composition. During cross-linking reactions, native states of proteins and protein complexes have to be preserved. In most cases, this prerequisite restricts the pH of cross-linking buffer in the range of 6.5-8.5 (Leitner *et al.*, 2010). Buffers may contain salt or low concentrations of DTT or EDTA that increase the stability of protein samples. However none of these components should interfere with the cross-linking reaction. For cross-linkers that require dissolution, the final concentration of organic solvent should not exceed 8% in volume.

2) Protein concentration. Low protein concentration can minimize unwanted oligomerization. However, this also decreased the cross-linking efficiency, especially for the most frequently used NHS ester cross-linkers that have high hydrolysis rates. Protein concentrations in the mg/ml range have proved to yield efficient cross-linking without promoting oligomerization (Bohn *et al.*, 2010; Chen *et al.*, 2010).

3) Reaction temperature. Cross-linking reactions can be carried out at different temperatures but are often in the range of 4-37℃. The actual temperature very much relies on the sample stability. Generally, at higher temperature, the cross-linkers will show higher reactivity towards proteins, but may result in undesired side reactions.

4) Substrate to cross-linker ratio and reaction time. Substrate to cross-linker ratios may vary significantly for different protein samples. Titration experiments are very useful to determine the optimal substrate to cross-linker ratio and reaction time for the desired product. For example, as shown by Dihazi *et al.*, both high cross-linker to protein ratios and long reaction time promoted oligomerization of cytochrome C (Dihazi and Sinz, 2003).It is common that active cross-linkers are still present after

designated reaction time. However they are normally quenched before further processing.

In practice, after cross-linking, cross-linked protein samples can be analyzed by one-dimension gel electrophoresis (SDS-PAGE) or by mass spectrometry to monitor cross-linking efficiency and cross-linking products.

1.2.3 In vivo cross-linking

Cross-linking analysis is normally performed on isolated protein or protein complex samples. Recently attempts have been reported to cross-link interacting proteins in living cells. In one strategy, *in vivo* cross-linking reactions were achieved by using photoreactive amino acid analogues that were incorporated into the protein (Suchanek *et al.*, 2005). In another two reports, the cross-linkers formaldehyde (Vasilescu *et al.*, 2004) and PIR (Zhang *et al.*, 2009) both efficiently permeated the cell membrane and generated cross-links between proteins. Among these attempts, only the cross-linking with PIR allowed for the identifications of the cross-links. More than 20 PIR cross-linked peptides were identified, mainly involving membrane proteins (Zhang *et al.*, 2009).

1.3 Enrichment of cross-linked peptides

1.3.1 Separation and digestion of cross-linked protein samples

After cross-linking, cross-linked protein samples are subjected to proteolytic digestion prior to mass spectrometric analysis. Before digestion, the protein level separations are normally performed to isolate the desired cross-linking products from uncross-linked material and unwanted oligomers which may provide false information for structural analysis. One dimensional gel electrophoresis (SDS-PAGE) is most commonly used. This process is applicable in cases where the cross-linked products can form clearly defined bands on gels,

for example the Pol II complex and C3 protein discussed in Chapters 4 and 5. Gel bands corresponding to cross-linked protein or protein complexes are cut from the gels and digested. For samples that are not suitable for gel based enrichment, digestion can be carried out in solution. Trypsin is the most commonly used enzyme for digestion. Both in-gel and in-solution digestion of cross-linked protein samples follow the procedures well established in standard proteomics.

After digestion, there are typically several types of cross-linking products in the peptide mixture (Schilling *et al.*, 2003). The different nomenclatures for these cross-linking products are summarized in Figure 1.8.

1) The cross-linked peptide (the type 2 cross-links): two peptide chains that are cross-linked by a cross-linker.

2) The modified peptide (the type 0 cross-link): a single peptide chain that is modified by a cross-linker with its reactive group on the other end inactivated by hydrolysis or reaction with scavenging reagents.

3) The loop linked peptide (the type 1 cross-link): a single peptide chain with two residues cross-linked by a cross-linker.

4) The higher order cross-linked peptide: peptides that are cross-linked with more than one of above three cross-linking formats. Possible combinations of cross-linking for this type of cross-linking products are theoretically unlimited.

5) The linear peptide: non-cross-linked peptide. Linear peptides form the majority of peptides in mixtures of cross-linked protein samples.

All three basic cross-linked products (1, 2, and 3) can be identified by mass spectrometric analysis and they all carry structural information. Modified peptides can reflect surface solvent accessibility, loop linked peptides may reveal local structure such as α-helix, while cross-linked peptides will indicate proximity between residues that are far

separated on sequence level or from different polypeptide chains. Cross-linked peptides are the most informative for 3D structural analysis, as they imply protein folding and protein-protein interactions. Hence in this thesis, I mainly focus on the analysis of cross-linked peptides.

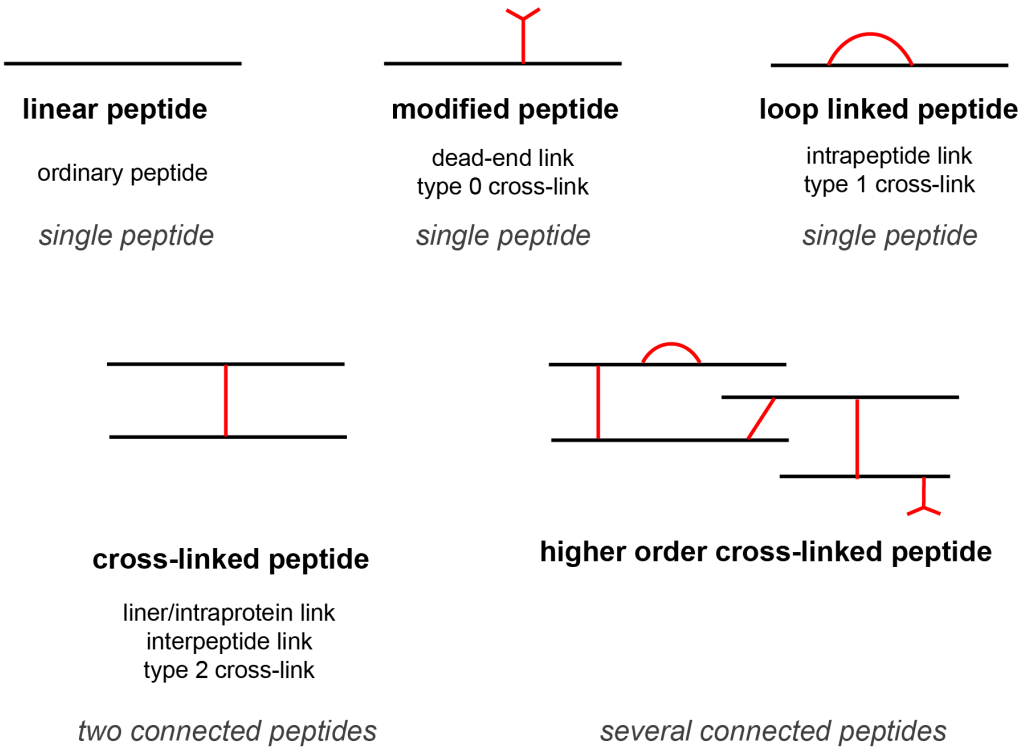


Figure 1.8 - Nomenclature of common products of chemical cross-linking reactions.

The terminology of the linear peptide, the cross-linked peptide, the modified peptide, the loop linked peptide and the higher order cross-linked peptide (black, peptide chains; red, cross-linkers). Alternative terms for these cross-linking products are also listed. This figure is modified from Rappsilber and Leitner (Leitner *et al.*, 2010; Rappsilber, 2011).

An isotope labelling procedure reported by (Back *et al.*, 2002), can be introduced during the trypsin digestion to facilitate the identification of cross-linked peptides. When cross-linked protein samples are digested in ^{18}O enriched water, two ^{18}O atoms will be incorporated to each C-terminus of lysine or arginine containing peptide. All cross-linked peptides will show an 8 Da shift in their mass spectra due to the possession of two C-termini while all non-cross-linked linear peptides, cross-linker modified peptides and loop linked

peptides will only shift by 4 Da since they all have only one C-terminus. In this way, cross-linked peptides can be readily distinguished. However this method does rely on the complete incorporation of ^{18}O.

1.3.2 Enrichment of cross-linked peptides

In the peptide mixture of a digested protein sample, cross-linked peptides are normally in relative low stoichiometry compared to linear peptides. There are three main reasons for this. Firstly, limited by chemical relativities of cross-linkers, cross-linking reactions are usually formed with relatively low efficiency and often incomplete. Secondly, as described before, cross-linking reactions may result in multiple types of cross-linking products. Finally, a residue may form different cross-links with several proximal residues. The low stoichiometry of cross-linked peptides determines their inferior position in competition with abundant linear peptides in mass spectrometric analysis. Particularly in complex protein samples, the difficulty in detecting cross-linked peptides has been described as finding the needle in the haystack (Sinz, 2006).

Selective isolation of cross-linked peptides would enhance the detection of cross-linked peptides. As discussed in *1.2.1.3,* cross-linkers containing an affinity handle have been developed, which allow for affinity purification of the cross-linker containing peptides. However, routine applications of this type for cross-linker have not been reported.

Therefore two other strategies were developed to enrich for cross-linked peptides. Firstly, a charge based strategy was developed based on the fact that cross-linked peptides tend to have more basic groups, particularly when digested by trypsin. This property enables enrichment of cross-linked peptides using strong cation exchange chromatography, and selective fragmentation during the mass spectrometric acquisitions. This strategy has been developed and applied by our group and others (Maiolica *et al.*, 2007; Rinner *et al.*, 2008; Chen *et al.*, 2010) and will be further discussed in Chapter 3. It is worth mentioning that

Trnka *et al.* introduced a new amine specific cross-linker 1,3-diformyl-5-ethynylbenzene (DEB) (Trnka and Burlingame, 2010) that contains two additional protonation sites, which was applied to improved the charge based separation of cross-linked peptides from linear peptides.

Secondly, peptide-level size exclusion chromatography has also been used for enrichment. Cross-linked peptides with higher molecular weight and larger size tend to be eluted with lower retention volume (Bohn *et al.*, 2010; Leitner *et al.*, 2010).

Finally, reversed-phase chromatography (on-line or off-line to MS) and iso-electric focusing can also reduce sample complexity. They will increase the chance for cross-linked peptides to be detected and fragmented (if applicable) in a mass spectrometer. However these separation steps will not specifically enrich for cross-linked peptides.

1.4 Analysis of cross-linked peptides by mass spectrometry

1.4.1 Mass spectrometric analysis of cross-linked samples

Peptide mixtures of digested cross-linked protein samples are subjected to MS and MS/MS mass spectrometric analysis to identify cross-linked peptides and subsequently assign the linkage sites. There are two major strategies for analyzing cross-linked peptides. The first is the peptide mass mapping approach. It commonly involves comparison of a control and a cross-linked sample. The novel peaks in the cross-linked sample are considered to be candidates for cross-linked peptides. The cross-linked peptides are identified by matching the mass of the observed candidates to the masses of all possible peptides combinations. In addition MS/MS analysis can be conducted with the cross-linked peptides to obtain fragmentation information that may verify the identity and assign the cross-linked residues. As discussed, cross-linked peptides have low abundance in the peptide mixtures. The use of isotopically labelled cross-linkers generates distinctive isotopic doublets pattern for cross-

linker containing species which makes them easily recognizable in complex spectra even at very low intensity. Furthermore this isotopic signature can increase the specificity of the identification of cross-linked peptides. Analysis by MALDI-TOF (matrix-assisted laser desorption/ionization-time-of-flight) mass spectrometry allows for the straightforward comparison between control and cross-linked samples, It is widely used in the 3D proteomics studies on simple protein samples (Lee *et al.*, 2007). Some isotope labelling strategies are also applied to distinguish cross-linked peptides from background. Besides the ^{18}O incorporation during trypsin digestion, a method that was designated as "mixed isotope cross-linking" was devoloped. This strategy involved the use of a 1:1 mixture of ^{15}N-labeled and unlabeled (^{14}N) protein. The cross-linked peptides display 1:2:1 triplet signals in MS^1 spectra and all the other peptides show as 1:1 doublets (Taverner *et al.*, 2002).

In the second strategy, peptide mixtures of cross-linked samples are subjected to LC-MS/MS (liquid chromatography–tandem mass spectrometry) analysis. Normally peptide mixtures are separated by reverse-phase chromatography and the eluted peptides are directly injected into a mass spectrometer via electrospray ionization. This method is chosen to obtain more information on the amino acid sequence of cross-linked peptides and also on cross-linking sites. More importantly, this approach allows for high throughput analysis of complex samples. With LC-MS/MS data, cross-linked peptides are identified based on both precursor mass and fragmentation species using database searches. Even with an additional LC separation step, the competition with large excess of abundant linear peptides significantly reduces the yield of fragmentation spectra of cross-linked peptides, particularly in complex protein samples. Therefore several labelling and enrichment schemes are applied in LC-MS/MS analysis to improve the detection of cross-linked peptides. Doublet signals of peptides cross-linked with stable isotope labelled cross-linkers were used to direct peptide fragmentation in a repeated acquisition by using inclusion lists. When electrospray

ionization is applied, charge selective fragmentation in a data dependent MS/MS acquisition can promote sequencing of cross-linked peptides that are normally higher charged.

Theoretically, cross-linking can be expected to occur between any two peptides from proteins in the sample. To identify cross-linked peptides, all possible peptide combinations need to be considered. With increased protein sequence database size, the number of possible combinations increases quadratically. An accurate mass measurement (<10 ppm) at MS^1 level is essential especially when searching against a large database and particularly when only MS^1 data (peptide mass) is used for the identification (Leitner *et al.*, 2010). However even with high mass accuracy, in a search space that contains combinations of 49 *E.coli* ribosomal peptide sequences, MS^1 data alone is not enough for unambiguous matches (further discussed in Chapter 3). For unambiguous identification of cross-linked peptides from large database fragmentation information (MS^2) is necessary.

In earlier studies, analyses of single protein and simple protein complex samples were mainly conducted on MALDI-TOF, MALDI-TOF/TOF, ESI-TOF (electrospray ionization-TOF) and ESI-Q-TOF (ESI-linear trap quadrupole-TOF) instruments (Sinz, 2006; Jin Lee, 2008; Leitner *et al.*, 2010). Most high throughput analyses were carried out using LC-MS/MS. Analysis using FTMS (Fourier transform mass spectrometry) provides high mass accuracy data, however its low scan speed in CID (collision-induced dissociation) MS/MS is not compatible with on-line LC applications. The use of hybrid instruments like ESI-LIT-FTICR (ESI- linear ion trap-Fourier transform ion cyclotron resonance mass spectrometry) and ESI-LIT-Orbitrap overcame this issue by acquiring CID MS/MS in a low resolution linear ion trap. Q-TOF and Orbitrap also allow the high resolution CID MS/MS measurement in the LC time scale (Jin Lee, 2008). Moreover the new generation instruments like LTQ-Orbitrap Velos (Thermo Scientific) allow for high resolution MS/MS sequencing at high speed (Olsen *et al.*, 2009; McAlister *et al.*, 2010).

1.4.2 Fragmentation of cross-linked peptides

In a typical LC-MS/MS analysis, peptides are first separated by on-line reverse-phase liquid chromatography (LC). The eluted peptides are continuously injected to a mass spectrometer and are analyzed in cycles. Each MS/MS acquisition cycle involves multiple steps. Firstly, mass to charge ratio of eluted peptides are detected and recorded in an MS^1 spectrum. Then a selection of peptides in the mixture are isolated and fragmented according to criteria such as their intensity. Cleavages of peptide bonds are induced and the peptide chain falls into fragments. The mass spectrometer then detects for each precursor ions collection of fragment ions that are cleaved at different amide bonds and records them in an MS^2 spectrum.

Peptide fragmentation can be induced in different ways, however for each fragmentation method, peptides break following certain rules and generate certain type of fragment ions. For example, the collision induced dissociation (CID) typically generates b type ions and y type ions while the electron-transfer dissociation (ETD) generates c and z type ions (Roepstorff and Fohlman, 1984; Biemann, 1988) (Figure 1.9). Fragmentation can to some extent be controlled so that for the majority of peptide molecules only one cleavage event occurs. Theoretically each peptide fragment differs from its neighbour by one amino acid. Hence a series of fragments encode the peptide sequence and may also reveal modifications. Searching a MS^2 spectrum containing a series of fragment ions against protein sequence database allows for identification of peptides according to both precursor mass detected in the MS^1 scan and fragments in the MS^2 spectrum (Steen and Mann, 2004).

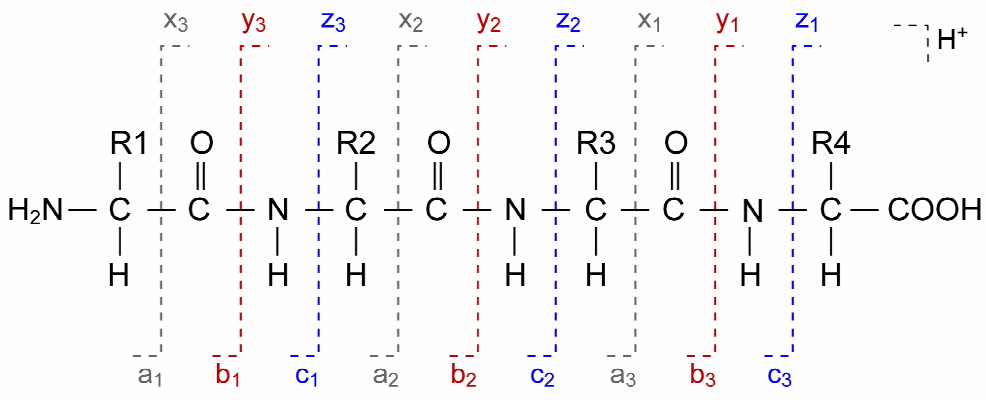


Figure 1.9 - Fragment ions observed in MS^2 spectrum.

Nomenclature for fragment ions were proposed by Roepstorff and Fohlman (Roepstorff and Fohlman, 1984), and subsequently modified by Johnson *et al.* (Richard S. Johnson, 1987).

This strategy has been adopted in the analysis of cross-linked peptides in order to identify cross-linked peptides and assign cross-linking sites. Fragmentation spectra of cross-linked peptides have been reported in several studies (Pearson *et al.*, 2002; Huang *et al.*, 2004; Bhat *et al.*, 2005; Petrotchenko *et al.*, 2005; Huang and Kim, 2006; Jacobsen *et al.*, 2006; Kitatsuji *et al.*, 2007; Lee *et al.*, 2007; Maiolica *et al.*, 2007). However due to the lack of large data sets, current knowledge of fragmentation behaviours of cross-linked peptides is based on limited observations from analyses of simple protein samples and a few synthetic peptides (Schilling *et al.*, 2003; Gaucher *et al.*, 2006; Gardner and Brodbelt, 2008; Iglesias *et al.*, 2009). So far, cross-linked peptides have mainly been fragmented using CID. From observed MS^2 spectra, fragments are mainly generated by a single cleavage on peptide backbone which follows the general fragmentation rules observed for linear peptides. In cross-linked peptides, fragments are observed for both cross-linked peptides, however they are not always evenly distributed between the two peptides. Another fragmentation technique, electron-transfer dissociation (ETD) has been reported which gives more efficient fragmentation for large peptides with 3+ or higher charge states, of which cross-linked peptides could be described. However this technique has not been widely applied in analysis

of cross-linked peptides, which could be because that for a long time ETD was restricted to the FTICR platform and low resolution ion traps (Leitner *et al.*, 2010). Recently, an application of ETD in conjunction with CID yielded complementary results and improved confidence level of identification of cross-linked peptides (Chowdhury *et al.*, 2009). Truka also reported an application of ETD for analyzing peptides cross-linked by DEB (Trnka and Burlingame, 2010). A new fragmentation method, higher-energy C-trap dissociation (HCD) has now become available on the LTQ-Orbitrap instrument (Olsen *et al.*, 2007). It has been shown that HCD gave better sequence coverage for a naturally cross-linked SUMO2-SOMO2 peptide when compared to CID (Waanders *et al.*, 2008). However, the application of HCD for the analysis of cross-linked peptides analysis has not yet been reported.

Even though cross-linked peptides follow certain fragmentation rules, in reality it is not easy to interpret MS^2 spectra of cross-linked peptides. Comparing to linear peptides, MS^2 spectra of cross-linked peptides are much more complicated by the existence of fragment ions which have originated from two peptides. Fragments derived from each partner peptide include both linear fragments and fragments that contain linkage sites. Moreover, due to the application of electrospray ionization, these fragment ions can be present in different charge states. The complexity of MS^2 spectra are often also accompanied by low intensities. These features made it extremely hard to interpret cross-linked peptides manually. Therefore, in recent years, bioinformatics tools have been developed to automate interpretation of cross-linked peptide MS^2 spectra (Jin Lee, 2008). In fact`, some reporter ions that are specific for the cross-linker containing peptides and only cross-linked peptide have been observed for certain cross-linkers. Detection of these reporter ions in the MS^2 spectra will improve the specificity of identification of cross-linked peptides (Seebacher *et al.*, 2006; Iglesias *et al.*; Iglesias *et al.*, 2010).

1.5 Identification of cross-linked peptides

To identify cross-linked peptides, peptide pairs need to be calculated in the database searches. This can not be achieved by using the search algorithms that are established in standard proteomics analysis tools such as Mascot. In the past ten years, quite a few programs and algorithms to identify and match spectra and candidate cross-linked peptides have been reported and have recently been reviewed (Sinz, 2006; Jin Lee, 2008; Leitner *et al.*, 2010).

The MS^1 based algorithms identify cross-liked peptide candidates by comparing the MS data from a cross-linked sample and a non-cross-linked control, or by recognizing the signature of cross-linked peptides generated using isotope labelling strategies (*e.g.* ^{18}O digestion (Back *et al.*, 2002) and "mixed isotope cross-linking" (Taverner *et al.*, 2002)). Typical search programs calculate all possible peptide combinations according to the user defined protein sequences, protease, cross-linker, modifications, permitted missed cleavages *etc*. The candidates then are matched to these peptide combinations by their mass with a certain error tolerance (Gao *et al.*, 2006; Seebacher *et al.*, 2006). As discussed in *1.4.1*, application of MS^1 based analysis is limited by sample complexity. It is worth noting that the recently developed analysis strategy using the PIR cross-linker also identifies cross-linked peptides based on accurate mass. However the mass of the two cross-linked peptides they described are matched separately rather than as a combination. This process reduced search space and allowed for the application in more complex samples (Anderson *et al.*, 2007). Although high mass accuracy data can increase the specificity of matches, the peptide mixtures of cross-linked samples are much more complex than equivalent *in silico* digestion products. Unknown modifications, non-specific cleavages *etc*, can give rise to unexpected peaks in the mass spectra. The probability that these peaks match to cross-linked peptides in database randomly by mass can not be excluded. Therefore verification of identified cross-linked peptides by MS2 is necessary. Some of search programs are able to

match the MS^2 spectra and cross-linked peptide candidates. However the validation of these matches mainly relies on skilled manual work (Gao *et al.*, 2006; Seebacher *et al.*, 2006).

Algorithms designed for searching LC-MS/MS data follow the basic logic used to identify linear peptides in standard proteomics analysis. Firstly, candidates of an observed peptide that match on precursor mass within certain error tolerance are filtered out. Then the *in silico* fragmentation of these candidates are compared to the observed fragmentation spectrum to understand which peptide sequence best explains the observed spectra. The returned matches are normally accompanied by a score that reflects how well the theoretical MS^2 spectra match the observed MS^2 spectra.

In order to search for cross-linked peptides, all possible peptide combinations need to be considered in database. For a database containing n peptides, the possible combination is $(n^2+n)/2$. This means when the number of peptides increase linearly, the number of combination increase quadratically, and this number defines the required search space. When searching against large databases, the search algorithm needs to handle an overwhelming numbers of candidate spectra. This presents a computational challenge for the search engines. This problem is simplified when working with single proteins or simple protein complexes, since very few proteins need to be considered. Several algorithms and programs have been reported to automate and match MS^2 spectra of cross-linked peptides in such smaller databases.

In an algorithm reported by Maiolica *et al.*, a purpose-built cross-link database (XDB) was built in which each possible cross-linked peptide pair was converted into two merged linear sequences with the cross-linker mass as a modification on the cross-linking sites. These linear sequences have the same mass as the cross-linked peptide pair and cover all possible single peptide bond fragments of the cross-linked peptides. Searching against this database, cross-linked peptides were identified using Mascot, a search algorithm used in standard proteomics workflows, whereby a score is assigned to each peptide/protein match.

This is the first automatic search algorithm that identified cross-linked peptides from a multi-protein complex (the Ndc80 complex) and this approach has recently been integrated into a commercial search engine, Phenyx. However this approach did not solve the search space issue for large databases.

Recently several search strategies were developed to use restricted databases to decrease the search space. One strategy was used by the Protein Prospector package (Singh *et al.*, 2008) and CXMS pipeline (Chu *et al.*, 2010). In principle, they regard a cross-linked peptide as two linear peptides with the cross-linker and their cross-linking partner as modifications. The two peptides in a cross-linked pair can be searched separately. Therefore the quadratic search space is converted into a succession of two linear searches. This process has been achieved by conducting open modification search to retrieve candidate combinations, the MS^2 spectra are matched among retrieved candidate combinations using a complete set of fragment ions.

The newly developed xQuest is designed for identification of cross-linked peptides from large databases. It has two optional modes: firstly in the enumeration mode it follows the logic of a standard linear peptide identification workflow, and considers all peptide combinations. It allows for searches against 100 protein sequences, although it requires extensive computational time. Secondly in the ion-tag mode, it requires the use of isotope labelled cross-linkers, and for each cross-linked peptide, both heavy and light ions need to be fragmented. In this mode heavy and light versions of spectra are compared for a cross-linked peptide. The linear fragments are overlaid in both while the cross-linkage site containing fragment will show a mass shift caused by isotope labelled cross-linkers. Hence signals in MS^2 spectra can be sorted into two groups. The linear fragments were used to get candidates from database, and the complete set of fragment ions are use to identify cross-linked peptides among the retrieved candidates. Using this approach, cross-linked peptides were identified from whole cell *E.coli* lysate (Rinner *et al.*, 2008).

Beside the search space, the probability of false matches also increased quadratically. In addition to the increased search space, there is also a difficulty in separating the true positive and random matches from the overwhelming number of possible candidates of cross-linked peptides. Currently, manual interrogation still plays a major role in verification and validation of cross-linked peptides identification even in large datasets (Chen *et al.*, 2010). Although it is labour intensive it is still manageable. However, when the analysis moves towards the whole proteome level, manual interrogation will not be possible. Although different approaches have been attempted following the work with linear peptides to estimate the false positive rate (Rappsilber, 2011), further development is required. .

1.6 Current application of 3D proteomics

Since the combination of chemical cross-linking with mass spectrometry was introduced at the end of the 1990s, 3D proteomics has grown dramatically after one decade's developments. The applications of this technique have changed from the proof of principle studies in model proteins (Lee *et al.*, 2007) to analyses of large multi-protein complexes (Bohn *et al.*, 2010; Chen *et al.*, 2010). These applications cover a wide range of biology, including bacteria (Nechifor and Wilson, 2007), yeast (Chen *et al.*, 2010), plant (Nyarko *et al.*, 2007) and human (Kitatsuji *et al.*, 2007).

3D proteomics has provided information on protein low resolution folding through intra protein cross-links (Young *et al.*, 2000; Dihazi and Sinz, 2003; Huang *et al.*, 2004; Sinz, 2006; Jin Lee, 2008). It also revealed the conformational change of a serine/threnine kinase, Akt during its activation (Bhat *et al.*, 2005). This technique was also used for mapping binding proteins. This has been shown by Rappsilber *et al.* in combining chemical cross-linking, electrophoresis and mass spectrometry to provide a topological map of Nup84 complex. Kitatsuji *et al.* idenfied interaction between oxidized human neuroglobin (Ngb)

and G protein α-subunit through cross-links. More sophisticated cross-linking site analyses revealed proximity between gamma, epsilo and III subunits of spinach chloroplasts ATP synthases (Gertz *et al.*, 2007). Yield changes of certain cross-links reflected the impact of nucleotides and Mg(2+) on special arrangement of these subunits. Using the MIX (mixed isotope cross-linking) strategy, 3D proteomics revealed spatially proximal residues between cytokine interleukin-6 (IL-6D) chains in a homodimer (Taverner *et al.*, 2002). Analysis of the NDEL1complex and the Ndc80 complex demonstrated the internal organization of these two complexes and revealed directionality of the homodimeric coiled-coil of the NDEL1 complex, as well as the register of the heterodimeric coiled-coils in the NDC80 complex. Similar applications were carried out on a group of protein complexes with two to five proteins and revealed spatial organization and the interaction suface between subunits in complexes (Chang *et al.*, 2004; Chu *et al.*, 2004; Chu *et al.*, 2006; Schulz *et al.*, 2007; Pimenova *et al.*, 2008; Dimova *et al.*, 2009). The application on the 12-subunit Pol II complex and 15-subunit Pol II-TFIIF made a breakthrough on the size of protein complexes that can be handled by 3D proteomics (Chen *et al.*, 2010). This application was followed by the analyses of the phi29 connector/scaffolding complex (Fu *et al.*, 2010), and the 21-protein GroEL–GroES chaperonin complex (Trnka and Burlingame, 2010). Very recent study on the 26S proteasome in creased the size record to 2.5 MDa (Bohn *et al.*, 2010). These applications indicated that 3D proteomics has become a valuable tool for structural analysis of large multi-protein complexes.

When combinded with other methods, 3D proteomics has become a more powerful tool. The full length stucture of Acyl-CoA thioesterase 7 was assembled using C-terminal and N-terminal crystal structures, cross-linking data and computational simulation (Forwood *et al.*, 2007). A structural model of apoA-II in reconstituted HDL was built using 3D proteomics and internal reflection infrared spectroscopy (Silva *et al.*, 2007). In large protein complex studies, cross-links located TFIIF on the Pol II surface, a model of Pol II in

complex with TFIIF dimerization domain was built using the Pol II crystal structure, distance constraints carried by cross-linking data and the homology model of the TFIIF dimerization domain which was also validated using cross-linking data (Chen *et al.*, 2010). A data-driven docking model, using input from 3D proteomics and mutagenesis data, suggested an interaction between the scaffolding protein and the connector dodecamer (Fu *et al.*, 2010). In the study on 26S proteasome, 3D proteomics data clarified the topology of the AAA-ATPase module in the 19S regulatory particle and its spatial relationship to the α-ring of the 20S core particle (Bohn *et al.*, 2010). These applications suggested a valuable role of 3D proteomics in the integrated structural biology. They also indicated the "beginning of a beautiful friendship" between 3D proteomics and modelling of proteins and multi-protein complexes (Rappsilber, 2011). The experience and perspective of protein modelling using 3D proteomics data have been recently reviewed (Leitner *et al.*, 2010; Rappsilber, 2011).

There have also been two other exciting applications that represented important technical progress in the field in the last three years. Rinner and coworkers published a new search algorithm, xQuest, which allowed for the identification of inter-protein and intra-protein cross-links from a total *E.coli* lysate, searching against the total *E.coli* protein database (Rinner *et al.*, 2008). Bruce and coworkers carried out *in vivo* cross-linking in *Shewanella oneidensis* bacterial cells using a PIR cross-linker which resulted in the identification of a set of protein-protein interactions and their contact regions (Zhang *et al.*, 2009). Although these two applications were still proof of principle studies, they demonstrated the possibility of *in vivo* 3D proteomic studies and proteomic scale applications.

1.7 Project aim

3D proteomics combines chemical cross-linking, mass spectrometry and database searching. It detects spatial proximity between protein residues and provides information about low resolution protein structures and protein-protein interactions. The analytical principle of 3D proteomics determines that it can be a powerful tool for studying structures of proteins and protein complexes. However technical limitations have impeded its application on complex protein samples. The major challenges are on the detection and correct identification of cross-linked peptides.

My work aims:

1) to improve the analytic workflow of 3D proteomics
2) to develop applications of 3D proteomics on large multi-protein complexes
3) to explore the integration of 3D proteomic data in the structural biological study of proteins and protein complexes
4) to investigate possible combinations of 3D proteomics and other structural biology and proteomic techniques.
5) to investigate the possibility of obtaining structural information from protein complexes in a background of a complex protein mixture.

In this thesis, I present my work designed to achieve these aims through four major tasks. The work has provided new insights into architecture of Pol II-TFIIF complex, conformation of C3(H_2O), and architecture of *S. cerevisiae* endogenous Mad1-Mad2 complex and Ndc80 complex obtained from the application of 3D proteomics.

Chapter 2

METHODS AND MATERIALS

2.1 Cross-linking analysis of synthetic peptides

2.1.1 Cross-linking the synthetic peptides

49 tryptic peptides with sequences derived from *E.coli* ribosome 30S subunits were purchased from (Sigma). All synthetic peptides contained one non-C-terminal lysine residue within their sequence; peptide N-termini were Fmoc protected. The C-terminal amino acid residue of these peptides was arginine, except for four peptides derived from ribosomal protein C-termini. Each synthetic peptide was dissolved in 400 µl DMF, the concentrations of these peptides ranged from 2.6 nmol/µl to 7.6 nmol/µl and were on average 4 nmol/µl. These 49 peptides were mixed in equal volume to give a 49-peptide mixture. 200 µl of 49-peptide mixture was cross-linked with bis[sulfosuccinimidyl] glutarate (BS^2G, Thermo Scientific) in a 1:2 peptide to cross-linker ratio. The cross-linker solution was freshly prepared in DMF, and 2 µl 25% TEA in water was added to the reaction. The reaction was carried out at 60°C for 24 hours and quenched with 5 µl of 2.7 M ammonium bicarbonate (ABC) at 60°C for 2 hours. Peptide N-termini were de-protected overnight in three volumes (600 µl) of 20% piperidine in DMF, after which the sample was diluted 100-fold with 0.1% TFA and cleaned on C18-StageTips (Rappsilber *et al.*, 2003) following a published protocol (Rappsilber *et al.*, 2007). The cleaned peptide mixture was eluted from C18-StageTips with 80% acetonitrile (ACN) for follow-on SCX-HPLC fractionation.

In parallel, 150 µl of 49 peptides mixture was cross-linked with Bis[sulfosuccinimidyl] suberate (BS^3 , Thermo Scientific) following the same experimental procedure as before.

2.1.2 Strong cation exchange (SCX) fractionation

2.1.2.1 SCX-HPLC fractionation

Half of the BS^2G cross-linked synthetic peptides (corresponding to ~100 µl of the 49-peptide mixture) and one third of BS^3 cross-linked synthetic peptides (corresponding to ~50 µl of the 49-peptide mixture) were fractionated using a 200×2.1 mm PolySULFOETHYL A column (5 µm particles; 200 Å pore size; Poly LC, Columbia, MD USA). Chromatography was performed on an Ultimate 3000 HPLC system (Dionex, Sunnyvale, CA) using a salt gradient of buffer A (5 mM KH_2PO_4, 10% ACN, pH 3.0) and buffer B (buffer A with 1 M KCl). Peptides eluted from C18-StageTips with 80% ACN were diluted to 20% ACN and acidified to pH3.0 by buffer A and loaded on the SCX column with solvent A at a flow rate of 200 µl/min. Subsequently, peptides from two cross-linked samples were eluted at a flow rate of 200 µl/min by 2 slightly varied gradients. Both used gradients were curve gradients with the curve 8 equation (CHROMELEON 6.80; Dionex) in which the buffer B increased from 0% to 60% in 20 min. For the BS^2G cross-linked sample, the whole gradient contained 5min at 100% buffer A before the curve gradient and 1min at 60% buffer B afterwards. For the BS^3 cross-linked sample, the gradient stayed at 100% buffer A for only 1min before the curve gradient, and the buffer B increased to 70% linearly in 1 min after the curve gradient and stayed at 70% for an additional 1 min. The elution flow rate was 200 µl/min and the chromatograms were recorded as UV absorbance measured at 214 nm. Fractions were collected for one minute intervals.

The behaviour of non-cross-linked linear peptides during the SCX separation was mimicked by using trypsin digested *E.coli* and budding yeast extract. 100 µg of *E.coli* extract digest was separated on SCX column with the same gradient used for the BS^2G cross-linked synthetic peptides sample whereas 200 µg of yeast extract digest was fractionated using a gradient consisting of 1min at 100% buffer A followed by the curve 8 gradient that

increased from 0% buffer B to 70% buffer B in 35 min. The fractions were collected at one minute internals.

All collected fractions were diluted with 0.1% TFA to 5% ACN and desalted using C18-StageTips (Rappsilber *et al.*, 2003; Rappsilber *et al.*, 2007) prior to mass spectrometric analysis.

2.1.2.2 SCX-StageTip fractionation

An aliquot of BS^2G cross-linked sample corresponding to about 10 µl of 49-peptide mixture was fractionated using SCX-StageTip (Rappsilber *et al.*, 2003) following the protocol described for linear peptides (Rappsilber *et al.*, 2007). Briefly, the cross-linked peptides that were eluted in 80% ACN were diluted and acidified to pH3.0 and 20% ACN with 0.5% acetic acid before loading to SCX-StageTips. The flow-through was collected as fraction 0, peptides were eluted with a 4-step gradient as listed in Table 2.1. All five fractions were diluted 4-fold with 0.1% TFA and desalted using C18-StageTips. To characterize the behaviour of linear peptides, 10 µg of *E.coli* trypsin digest was fractionated with the same procedure.

Table 2.1 - SCX-StageTip fractionation

Elution buffer	Buffer composition	Eluate volume	Fraction
SCX buffer 1	0.5% Acetic acid, 20% ACN, 20 mM NH_4OAc	150 µl (50 µl X 3 times)	fraction 1
SCX buffer 2	0.5% Acetic acid, 20% ACN, 50 mM NH_4OAc	150 µl (50 µl X 3 times)	fraction 2
SCX buffer 3	0.5% Acetic acid, 20% ACN, 100 mM NH_4OAc	150 µl (50 µl X 3 times)	fraction 3
SCX buffer 4	0.5% Acetic acid, 20% ACN, 500 mM NH_4OAc	150 µl (50 µl X 3 times)	fraction 4

2.1.3 Analysis via Mass spectrometry

2.1.3.1 Sample preparation

The following samples were prepared for LC-MS/MS analysis:

1. All SCX-HPLC fractions of BS^2G cross-linked synthetic peptides,
2. All SCX-HPLC fractions of trypsin digested yeast extract,
3. Fractions 13 to 25 of BS^3 cross-linked synthetic peptides,
4. SCX-StageTip fractions of BS^2G cross-linked synthetic peptides,
5. SCX-StageTip fractions of trypsin digested *E.coli* extract,
6. 1 µg aliquots of non-fractionated trypsin digested *E.coli* extract.

All fractions were desalted on the C18-StageTips as described before, peptides were eluted with 20 µl 80% ACN in 0.1% TFA. The ACN was removed in vacuum (Concentrator 5301, Eppendorf AG, Hamburg, Germany), and the sample volume was adjusted to 5 µl with 0.1% TFA for the LC-MS/MS analysis. For trypsin digested *E.coli* extract, peptides corresponding to 10 µg of digest were eluted with 80% ACN. ACN was then removed in vacuum and the peptides were diluted with 0.1% TFA to final concentration of 0.2 µg/µl. A 5 µl aliquot was used for each injection.

2.1.3.2 LC-MS/MS analysis

LC-MS/MS analysis was performed using an HPLC system (1100 binary nano pump, Agilent, Palo Alto, CA) coupled to a hybrid LTQ-Orbitrap mass spectrometer (Thermo Scientific). An analytical column was packed in/house with C18 material (ReproSil-Pur C18-AQ 3 µm; Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) in a spray emitter (75 µm ID, 8 µm opening, 120 mm length; New Objectives, USA) using an air-pressure pump (Proxeon Biosystems, Odense, Denmark). Mobile phase A consisted of water, 5% acetonitrile and 0.5 % acetic acid and mobile phase B consisted of acetonitrile and 0.5%

acetic acid. Samples were loaded onto the column at a flow rate of 700 nL/min, peptides were separated using a linear gradient at 300 nL/min flow rate. For all SCX-HPLC fractions and 1 µg *E.coli* extract samples, the gradient consisted of a 5 min linear variation from 0% to 5% solvent B, followed by a separating gradient to 23% solvent B over 80 min. Separating gradient was followed by a rapid rise to 80% solvent B in 10 minutes, 80% solvent B was kept for 4min before re-equilibration to starting conditions. For the SCX-StageTip fractions, the linear increase of B from 5% to 23% in the gradient was extended to 135 min.

Acquisition methods in the LTQ Orbitrap mass spectrometer were customized according to samples (Table 2.2). In summary, all MS^1 spectra were acquired in the Orbitrap (FTMS) with a resolution of 100,000 at m/z 400 and recorded in profile mode; the selected ions from the full scans were fragmented in the ion trap by collision induced dissociation (CID) with a normalized collision energy of 35% and a isolation window of 2 Th. For all cross-linked samples, up to 3 of the most intense ions in the MS scans were fragmented in each acquisition cycle, the fragment spectra (MS^2 spectra) were recorded in the Orbitrap in centroid mode (resolution 7500 at m/z 400, AGC target 750,000, Max fill time 1 s). For BS^2G cross-linked synthetic peptide SCX-HPLC fractions, the MS^2 spectra were additionally recorded in the ion trap. For linear peptide containing samples, the MS^2 spectra of up to 6 of the most intense ions in each cycle were acquired in the ion trap. Two of the 1 µg *E.coli* extract samples were analyzed with and without excluding singly and doubly charged precursor ions. For BS^2G cross-linked synthetic peptide SCX-StageTip fractions and BS^3 cross-linked synthetic peptide samples, singly and doubly-charged precursors were excluded for MS^2 acquisitions. For all acquisitions, the lock mass of polymethylsiloxane (m/z 445.12005) was used. The dynamic exclusion function was enabled for all analysis using following parameters: repeat count, 1; repeat duration, 30 second; exclusion duration, 90 second for using the 80min gradient and 120 second for the 135 min gradient.

Table 2.2 - Mass spectrometric acquisition methods for cross-linked synthetic peptide samples.

Sample	Full scan		MS^2 scan				
	Detector	**Resolution at m/z 400**	**Top n intense peaks**	**precursor ion charge selection**	**Fragmentation**	**Detector**	**Resolution**
1[1]	Orbitrap	100,000	3	1+ excluded	CID in ion trap	Orbitrap & ion trap	7500 & N/A
3	Orbitrap	100,000	3	1+ & 2+ excluded	CID in ion trap	Orbitrap	7500
4	Orbitrap	100,000	3	1+ & 2+ excluded	CID in ion trap	Orbitrap	7500
2	Orbitrap	100,000	6	1+ excluded	CID in ion trap	Ion trap	N/A
5	Orbitrap	100,000	6	Not enabled	CID in ion trap	Ion trap	N/A
6 (1)	Orbitrap	100,000	6	1+ excluded	CID in ion trap	Ion trap	N/A
6 (2)	Orbitrap	100,000	6	1+ & 2+ excluded	CID in ion trap	Ion trap	N/A

[1] Sample codes are defined in *2.1.3.1*

2.1.4 Database searching

MS^2 spectra peak lists were generated from the raw data files using the Quant module of MaxQuant version 1.0.11.2 (Cox and Mann, 2008). For low resolution spectra (acquired in the ion trap) all parameters were set to default, while for the high resolution spectra (acquired in the Orbitrap) for cross-linked samples, the "Top MS/MS peaks per 100Da" was set to 200.

Linear peptides in the *E.coli* extract and yeast extract samples were identified by searching the spectra against the SwissProt database using Mascot v2.2 (www.matrixscience.com). The search parameters are listed in Table 2.3. The spectra of cross-linked synthetic peptides samples were also searched against SwissProt database for the identification of non-cross-linked peptides with modified searching parameters: the MS/MS tolerance was set to 0.06Da; the "Max. Missed cleavages" was set to 0; hydrolyzed and ammonia reacted cross-linker (BS^2G or BS^3 corresponding to the sample) on lysine, and Fmoc modified N-termini were set as an additional variable modification while the fixed

modification was set to none. The returned peptide candidates with Mascot score above 25 were accepted as positive identifications.

Table 2.3 - Search parameters for linear peptides samples in Mascot search.

Search parameters	**Settings**
Enzyme	Trypsin
Max. Missed cleavages	3
Fixed modifications	Carbamidomethyl (Cys)
Variable modifications	Oxidation (Met)
Mass value	monoisotopic mass
Peptide tolerance	± 6 ppm
MS/MS tolerance	± 0.6Da
Peptide charge	1+,2+ and 3+

For the identification of cross-linked peptide identification, high resolution spectra were searched with in-house software XMass (Salman Tahir, unpublished) against 49 synthetic peptide sequences using the search parameters listed in Table 2.4. Matched spectra and cross-linked peptide candidates were returned by XMass in pairs. 30 spectra were annotated and validated by hand. .All other candidates were validated manually using automated annotation software Xaminatrix (Morten Rasmussen). The database search was repeated with a 50 ppm MS tolerance and 100 ppm MS^2 tolerance, while the other parameters were the same.

Table 2.4 - Search parameters for cross-linked peptides samples in XMass search.

Searching parameters	Settings
Cross-linker	BS2G/BS3 [1]
Enzyme	Trypsin
Max. Missed cleavages	0
Fixed modifications	Carbamidomethyl (Cys)
Variable modifications	Oxidation (Met), BS^2G/BS^3-OH (Lys)[2], BS^2G/BS^3-NH2 (Lys) [3]
Mass value	monoisotopic mass
Peptide tolerance	± 6 ppm
MS/MS tolerance	± 20 ppm

[1] The cross-linker and corresponding cross-linker modifications were set according to different samples.

[2] "-OH" indicates modification by hydrolyzed cross-linkers..

[3] "-NH2" indicates modification by ammonia reacted cross-linkers.

2.2 Cross-linking analysis of Pol II and Pol II-TFIIF complexes

2.2.1 The Pol II complex and the Pol II-TFIIF complex

The Pol II and Pol II-TFIIF complexes were provided by Patrick Cramer and were prepared as described previously (Chen *et al.*, 2010). The Pol II complex was affinity purified through a Hexahistidine -tagged Rpb3 subunit followed by further chromatographic separations. From a yeast strain that over-expresses TFIIF subunits and using tandem affinity purification (TAP) tag on Tfg2, the Pol II-TFIIF complex was purified as a pure, homogeneous, stoichiometric and catalytically active complex with the Pol II Rpb1 C-terminal repeat domain (CTD) non-phosphorylated (Sydow *et al.*, 2009; Chen *et al.*, 2010). 0.7 µg/µl Pol II and 0.8 µg/µl Pol II-TFIIF samples were in a buffer consisting of 10 mM HEPES pH 8.0, 200 mM potassium acetate, 1 mM EDTA, 10 % glycerol and 1 mM DTT were prepared for cross-linking.

2.2.2 Cross-linking titration of Pol II and Pol II-TFIIF complexes

To find out cross-linker-to-protein complex ratios that allowed for efficient cross-linking of Pol II and Pol II-TFIIF complexes, the cross-linking reactions of targeted complexes were titrated on small aliquots of samples. 2.5 µg aliquots of Pol II sample were cross-linked by Bis(Sulfosuccinimidyl) suberate (BS^3, Thermo Scientific) with a series of three-fold increases in cross-linker to protein molar ratio (from 1: 200 up to 1:16200) while the other experimental conditions were kept identical (Table 2.5). The protein complex was mixed with freshly prepared BS^3 solution in cross-linking buffer (10 mM HEPES pH 8.0, 200 mM Potassium acetate) and the cross-linking reaction was carried out for 2 h on ice. For the control sample, only buffer was added without cross-linker. Cross-linking reactions were quenched with 1 µl 2.7 M ABC for 45 min on ice. The cross-linking products from titration reactions were separated on a NuPAGE 4-12% Bis-Tris gel (Invitrogen) using MES running buffer and were silver stained using the protocol described in *2.5.4*. Comparing the migration of cross-linked complexes to the non cross-linked control on a denaturing gel, it was apparent that most of the bands in the control sample corresponding to individual subunits of Pol II complex got efficiently converted to a major band at high MW range. Moreover, there was no significant change on the cross-linking products pattern between 1:5400 and 1:16200 ratios (Figure 2.1 A). This suggested that a protein to cross-linker ratio of around 1:5400 would already efficiently cross-link most of the subunits in the Pol II complexes.

Table 2.5 - Experimental plan for Pol II complex cross-linking titration

Protein/cross-linker (molar/molar)	**Protein**		**Cross-linker**		**Total volume**
	Amount	**Volume**	**Amount**	**Volume**	
1:0 (control)	5 pmol (~2.5 µg)	3.5 µl	0	8 µl	11.5 µl
1:200	5 pmol	3.5 µl	1 nmol	8 µl	11.5 µl
1:600	5 pmol	3.5 µl	3 nmol	8 µl	11.5 µl
1:1800	5 pmol	3.5 µl	9 nmol	8 µl	11.5 µl
1:5400	5 pmol	3.5 µl	27 nmol	8 µl	11.5 µl
1:16200	5 pmol	3.5 µl	81 nmol	8 µl	11.5 µl

Based on the Pol II sample titration, the complex to cross-linker ratios for Pol II-TFIIF sample were titrated in a narrower range using only 1:1800, 1:3600 and 1:7200 ratios (Table 2.6). The cross-linking procedure was kept the same as for the Pol II complex titration. The denaturing gel electrophoresis analysis showed that both 1:3600 and 1:7200 complex to BS^3 ratios could efficiently cross-link subunits and no obvious changes on the pattern of cross-linking products between these two ratios were observed (Figure 2.1B).

Table 2.6 - Experimental plan for Pol II-TFIIF complex cross-linking titration

Protein/cross-linker (molar/molar)	**Protein**		**Cross-linker**		**Total volume**
	Amount	**Volume**	**Amount**	**Volume**	
1:0 (control)	5 pmol (~3.3 µg)	4.2 µl	0	8 µl	12.2 µl
1:1800	5 pmol	4.2 µl	9 nmol	8 µl	12.2 µl
1:3600	5 pmol	4.2 µl	18 nmol	8 µl	12.2 µl
1:7200	5 pmol	4.2 µl	36 nmol	8 µl	12.2 µl

The optimal reaction conditions for larger scale samples were established based on both titrated complex to cross-linker ratios and the cross-linker concentration in reactions. Overly high cross-linker to protein ratios were avoided, as they might cause undesirable over-linkage of the sample and may also induce formation of oligomers.

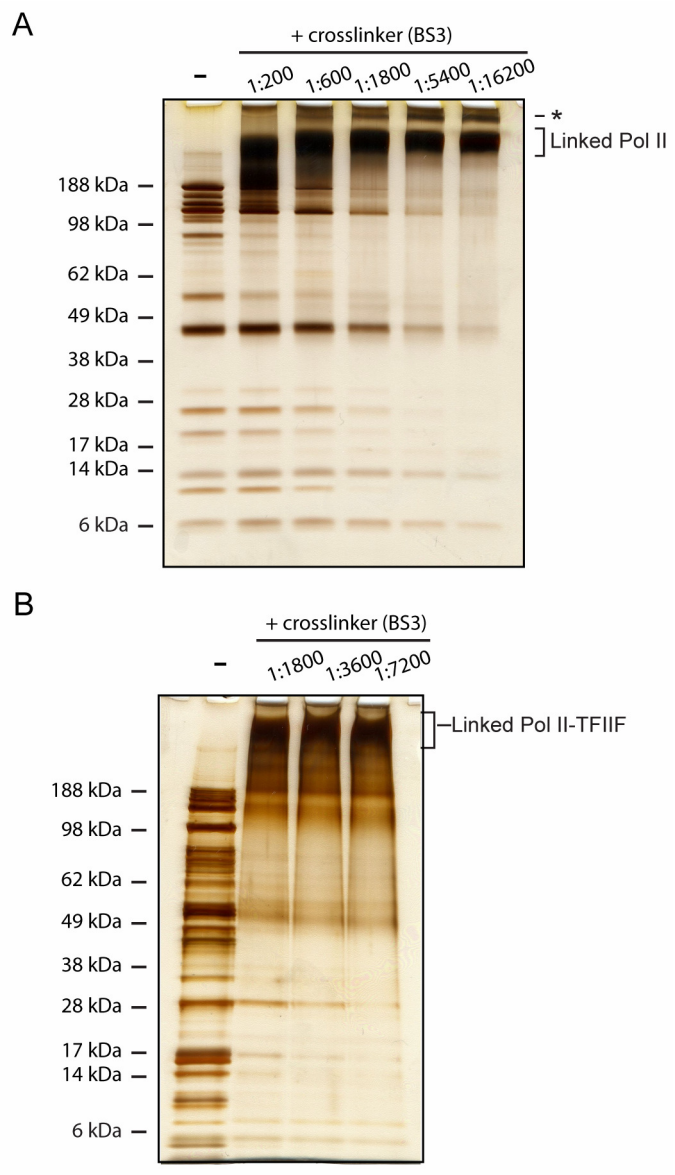


Figure 2.1 - Titration of BS3 cross-linking reactions for Pol II complex and Pol II-TFIIF complex

A. Separation of titrated BS3 cross-linker titrated Pol II complex sample on SDS-PAGE gel. The individual subunits of Pol II resolved into bands as seen in lane 1 disappear with increasing concentration of cross-linker. A higher order linkage product (asterisk) could be a Pol II dimmer or cross-linked Pol II with CTD in different phosphorylation state.

B. Separation of titrated BS3 cross-linker titrated Pol II-TFIIF complex sample on SDS-PAGE gel.

2.2.3 Cross-linking of Pol II and Pol II-TFIIF complexes

35 µg of purified Pol II complex was cross-linked with 150 µg BS^3 (Thermo Scientific) (dissolved in 70 µl cross-linking buffer consisting of 10 mM HEPES pH 8.0, 200 mM Potassium acetate) for 2 hours on ice. The reaction was stopped by the addition of 5 µl 2.7 M ABC and incubation for 45 min on ice. In order to monitor the cross-linking reaction, two 2.5 µg aliquots of cross-linked Pol II sample were separated on a denaturing NuPAGE 4-12% Bis-Tris gel (Invitrogen) using MES running buffer (Invitrogen) and on a native PAGE 3-12% Bis-Tris gel (Invitrogen), proteins were visualized by silver staining (*2.5.4*). The rest of cross-linked sample was separated on a NuPAGE 4-12% Bis-Tris gel (Invitrogen) using MES running buffer (Invitrogen). The gel was fixed in 50% methanol, 5% acetic acid and stained using colloidal blue kit following manufacturer's instructions (Invitrogen).

The purified TFIIF-Pol II complex (250 µl containing 200 µg) was mixed with 1 mg BS^3 (Thermo Scientific) that was dissolved in 470 µl cross-linking buffer (10 mM HEPES pH 8.0, 200 mM potassium acetate) and incubated on ice for 2 hrs. The reaction was stopped by adding 50 µl of 2.5 M ammonium bicarbonate and incubating for 45 min on ice. The efficiency of cross-linking reaction was checked using both native gel and SDS-PAGE gel with 5 µg aliquots of cross-linked Pol II-TFIIF the same way as for the Pol II sample. The remaining of cross-linking products were separated on a NuPAGE 4-12% Bis-Tris gel (Invitrogen) using MOPS running buffer (Invitrogen) and subsequently fixed in 50% methanol, 5% acetic acid and stained using colloidal blue kit (Invitrogen).

2.2.4 Sample preparation for mass spectrometric analysis

Bands from the SDS-PAGE gel corresponding to cross-linked complexes were excised for trypsin digestion. The cross-linked Pol II-TFIIF migrated into less defined bands especially on the lower molecular weight edge. The gel area above the largest subunit Rpb1 to the clear upper edge of the cross-linked complex was excised and further divided into the

"Upper" fraction (higher molecular weight) and the "Lower" fraction (lower molecular weight) for the analysis. In-gel digestion was performed based on the protocol described in Maiolica *et al.*, 2007: Excised gel bands were reduced using 50 mM DTT in 50 M ABC for 30 min at 37°C, followed by alkylation with 55 mM iodoacetamide for 30 min at room temperature in the dark. Proteins were digested using trypsin (proteomics grade, Sigma) with a 1:20 enzyme to protein mass ratio at 37°C for 15 hours. The cross-linked Pol II digest was extracted from the gel with ACN, diluted and acidified with 2.5% acetic acid to 20% ACN and pH 3.0. The peptide mixture was fractionated using SCX-StageTips to give 5 fractions as described in *2.1.2.2* and desalted with C18 StageTips prior to LC-MS/MS analysis. The digested 'Upper' and 'Lower' fractions of cross-linked Pol II-TFIIF were also extracted from gel pieces and adjusted to 20% ACN and pH3.0 for SCX-HPLC fractionation. The peptide mixtures were fractionated using strong cation exchange chromatography as described in *2.1.2.1*, with the gradient consisting of 5 min at 100% solvent A followed by 20 min transition to 60% solvent B with a curve gradient (curve 8 equation, CHROMELEON software v.6.80; Dionex), and 1 min at 60% solvent B. Fractions were collected at one minute intervals. All Pol II SCX StageTip fractions and Pol II-TFIIF SCX fractions 14 to 26 were desalted using C18 StageTips for subsequent LC-MS/MS analysis.

2.2.5 Mass spectrometry

LC-MS/MS analyses were performed as described for cross-linked synthetic peptide samples *2.1.3.2*. For all samples, peptides were first separated by reverse-phase chromatography which was coupled to MS with a solvent flow rate at 300 nl/min. The same elution gradient used for cross-linked synthetic peptide samples (*2.1.3.2*) was applied with little modification: the separating part of the gradient (5% to 23% solvent B) was extended to 90 min. Mass spectrometric acquisition parameters applied are given in Table 2.7.

Table 2.7 - Acquisition parameters for mass spectrometric analysis of the cross-linked Pol II and Pol II-TFIIF samples using the LTQ-Orbitrap mass spectrometer.

Parameters		Settings
MS	Detector	Orbitrap
	Resolution at m/z 400	100,000
	Acquisition mode	profile
	AGC target	750,000
	Max. fill time	500 ms
Fragmentation	Selection of top n intense peaks	3
	Isolation window	2 Th
	Precursor ion charge selection	1+ & 2+ excluded
	Dynamic exclusion	repeat count, 1; repeat duration, 30 second; exclusion duration, 90 s
	Fragmentation method	CID
	Normalized collision energy	35%
MS/MS	Detector	Orbitrap
	Resolution at m/z 400	7500
	Acquisition mode	centroid
	AGC target	100,000
	Max. fill time	1s

2.2.6 Database searching

The mass spectrometric raw files were processed into peak lists using the Quant module of MaxQuant version 1.0.11.2 software package (Cox and Mann, 2008) at default parameters apart from "Top MS/MS Peaks per 100Da" being set to 200. The in-house developed program Xi was used to search spectra against a database containing the sequences of the 12 Pol II subunits and the three TFIIF subunits from *S. cerevisiae*. Search parameters were listed in Table 2.8. The reversed sequences of proteins that were in the database were used for a target-decoy approach for false discovery rate (FDR) estimations. No linkage sites were specified for BS^3 in the search, cross-linking sites were determined by the best match of fragmentation spectra to cross-linked peptide sequence in the search algorithm. The returned candidate cross-linked peptide pairs with assigned linkage sites were manually validated using the in-house developed Xaminatrix program and sorted into high and low confidence

according to the criteria described in *3.5.1*. All matches had to be top ranking and unambiguous in the target-decoy search.

Table 2.8 - Search parameters used for database searching for cross-linked peptides in Xi.

Searching parameters	Settings
Cross-linker	BS^3
Enzyme	Trypsin
Max. Missed cleavages	4
Fixed modifications	Carbamidomethyl (Cys)
Variable modifications	Oxidation (Met), BS^3-OH (Lys & N-terminus); BS^3-NH2 (Lys & N-terminus)
Mass value	monoisotopic mass
Peptide tolerance	± 6 ppm
MS/MS tolerance	± 20 ppm

2.3 Quantitative 3D proteomics analysis of C3 and C3b samples

2.3.1 Protein cross-linking for quantitative analysis

C3 and C3b (Complement Technology, Inc. USA), 2 pmol/µl in cross-linking buffer (20 mM HEPES-KOH pH 7.8, 20 mM NaCl, 5 mM $MgCl_2$) were cross-linked with either Bis[Sulfosuccinimidyl] suberate-d0 (BS^3-d0) (Thermo Scientific) or its deuterated analogue Bis[Sulfosuccinimidyl] 2,2,7,7-suberate-d4 (BS^3-d4), at a protein to cross-linker molar ratio of 1:1000. After 2 h incubation on ice, reactions were quenched with 5 µl 2.5 M ABC for 45 min on ice. Cross-linking reactions were monitored using SDS PAGE gel electrophoresis. 5 pmol of cross-linked material from all 4 reactions were separated on a NuPAGE 4-12% Bis-Tris gel with MOPS running buffer. Protein bands were visualized using colloidal blue kit (Invitrogen).

Equimolar amounts of BS3-d0 cross-linked C3 and BS3-d4 cross-linked C3b were mixed as 'forward labelled' samples for quantitative analysis (300 pmol, sample 1; 150 pmol, sample2); while the 'reverse labelled' samples was equimolar mixture of BS3-d0 cross-linked C3b and BS3-d4 cross-linked C3 (300 pmol, sample 1R; 150 pmol, sample2R). These four samples were separated on NuPAGE 4-12% Bis-Tris gels using MOPS running buffer. The gels were fixed in 50% methanol, 5% acetic acid and stained using colloidal blue kit (Invitrogen). Bands corresponding to monomers of cross-linked C3 and C3b were isolated for subsequent conformational analysis.

2.3.2 Sample preparation for mass spectrometric analysis

Proteins were in-gel reduced/alkylated and digested using trypsin as described in *2.2.4*. Peptides from Sample1 and Sample 1R were fractionated using strong cation exchange chromatography as described in *2.1.2.1*, with minor alterations. Peptides were separated at a flow rate of 200 µl/min using a gradient consisting of 1 min at 100% solvent A followed by 12 min transition to 60% solvent B with a curve gradient (curve 8 equation, CHROMELEON software v.6.80; Dionex), then a 1 min linear gradient to 70% solvent B which was keep at 70% solvent B for 1 min. Fractions were collected at one minute intervals. High salt fractions 11 to 17 were desalted using C18-StageTips (Rappsilber *et al.*, 2003) for subsequent LC-MS/MS analysis. Sample 2 and Sample 2R were fractionated using SCX-StageTips (Ishihama *et al.*, 2006) as described in *2.1.2.2* and desalted using C18-StageTips before mass spectrometric analysis.

2.3.3 Mass spectrometric analysis

Fractionated samples were analyzed with the same LC-MS/MS setup as described in *2.1.3*.2. Peptides were separated on the C18 reverse-phase analytical column. The same elution

gradient applied for cross-linked synthetic peptide samples was used. For SCX HPLC fractionated samples, the separating part of the gradient (5% to 23% solvent B) was set to 80 min. For all SCX-Stage tip fractions, this part of the gradient was extended to 135 min. Peptides were eluted at a 300 nl/min flow rate directly into the LTQ-Orbitrap mass spectrometer and analyzed with the same parameter settings listed in Table 2.7 except for the dynamic exclusion duration, which was set to 90 sec for the samples with 80 min separating gradient, and 120 sec for samples with 135 min separating gradient.

2.3.4 Identification of cross-linked peptides

Mass spectrometric raw files were processed into peak lists using MaxQuant version 1.0.11.2 (Cox and Mann, 2008) at default parameters except "Top MS/MS Peaks per 100Da" set to 200. Peak lists were searched against the sequences of C3 and C3b using the in-house developed program Xi. The search parameters were set according to those listed in Table 2.8: All samples were searched with both cross-linkers (BS^3-d0 and BS^3-d4). The cross-linked peptide candidates returned from the Xi search were manually validated using the in-house developed program Xaminatrix and were sorted into high and low confidence as described previously (*2.2.6*).

2.3.5 Quantitation of cross-linkages

The quantitation was conducted at the linkage site level. The C3 to C3b signal ratio for a certain cross-link was calculated as the intensity weighted average of all detected cross-linked peptides corresponding to it. For each cross-linked peptide, the summed intensity of the three most intense isotope peaks in the isotope clusters of both heavy and light signals were used to calculate the C3 to C3b signal ratio. Peak intensities that were represented by the elution peak area were read out manually from raw data using "peak detection" function

in Xcalibur (version 2.1.0, Thermo Scientific). For each cross-linked peptide, if more than one charge state was observed, the signals from different charge states were summed prior to calculating the ratio.

2.3.6 Comparison between cross-linking data and crystal structures

Crystal structures of C3 (PDB|2a73) (Janssen *et al.*, 2005) and C3b (PDB|2i07; PDB|2hr0) (Abdul Ajees *et al.*, 2006; Janssen *et al.*, 2006) were retrieved from the protein data bank. Cross-links were displayed in each crystal structure using Pymol (DeLano, 2002). The proximity between cross-linked residues in crystal structures was determined by the distances between C-α atoms.

2.4 Structural analysis of affinity purified protein complexes by 3D proteomics

2.4.1 Affinity purified tagged endogenous protein complexes

The endogenous *S.cerevisiae* Mad1-Mad2 complex and NDC80 complex were purified by Sjaak van der Sar from Kevin Hardwick's Lab. In short, a duplex protein A affinity tag ('ZZ tag') was infused to the C-terminal of Mad1 in the Mad1-Mad2 complex and NDC80 in the NDC80 complex by standard genetic procedures. The bait proteins were purified from the prepared extract of mitotically arrested yeast cells by Rabbit IgG antibodies (Sigma) that were covalently coupled to M270 Epoxy Dynabeads (Invitrogen). Dynabeads and associated complexes were washed four times with cold prep buffer (50 mM Bis-Tris Propane-HCl pH 7.6, 100 mM KCl, 10% glycerol, 5 mM EGTA with additions: complete mini EDTA-free protease inhibitors (Roche), 1 mM Pefabloc SC (Roche), 10 µg/mL leupeptin, pepstatin and chymostatin, 5 mM NaN_3, 0.4 mM Na_3VO_4, 2 µM microcystin-LR, 20 mM β-glycerophosphate and 1% Triton-X) and four times with cross-linking buffer (100 mM HEPES-KOH pH 7.0 and 50 mM KCl).

2.4.2 'On-beads' cross-linking procedure

About 12 µg of purified Mad1-Mad2 complex was directly cross-linked onto Dynabeads in a buffer containing 50 mM HEPES-KOH (pH 7.6) and 100 mM KCl. The freshly prepared cross-linker solution with a 1:1 mixture of BS^2G-d0 and its deuterated analogue BS2G-d4 was used for cross-linking, with a complex to cross-linker mass ratio of 1:50. After cross-linking on ice for 1.5 hour, the reaction was quenched with 10 µl 2.7 M ABC for 45 min on ice.

About 9 µg of purified NDC80 complex on Dynabeads was cross-linked with 50-fold excess of BS2G-d0 (mass ratio). The reaction lasted for 2 hour on ice and was quenched by 10 µl 2.7 M ABC for 45 min on ice.

As a parallel control, protein complexes purified from a 1:1 mixture of SILAC labelled cells (lysine $^{13}C_6$) and unlabelled cells were used to monitor the occurrence of inter-complex cross-links. 3.5 µg of Mad1-Mad2 complex (incorporation rate of SILAC labelling < 95%) was cross-linked under the same conditions as described above. This experiment was repeated with 5 µg material, with an improved lysine C13 incorporation rate (~98%). For the NDC80 complex, 5 µg of material (lysine $^{13}C_6$, incorporation rate ~98%) was applied.

2.4.3 Sample preparation for mass spectrometric analysis

After cross-linking, the protein complex samples were digested 'on-beads' in 25 mM ABC buffer. Trypsin was added with an enzyme-to-protein mass ratio of 1:30. Samples were incubated at 37°C for 15 hours.

After digestion, peptides mixtures were separated from the Dynabeads and acidified to pH 3 with acidic acid. The 12 µg of non-SILAC labelled Mad1-Mad2 material was divided to Sample I with 4 µg complex and Sample II with 8 µg complex. Digested samples

were further fractionated using SCX StageTips as described in *2.1.2.2*. Fractions were desalted using C18-StageTips (Rappsilber *et al.*, 2003) prior to mass spectrometric analysis.

2.4.4 Mass spectrometric analysis

As described previously in *2.1.3.2*, peptides were separated on the on-line reverse-phase analytical column prior to injection into the mass spectrometer. The same elution gradient as used for cross-linked synthetic peptide samples (*2.1.3.2*) was applied. The separating part of the gradient (5%-23% solvent B) was set to 135 min. Mass spectrometric analysis was conducted using the settings listed in Table 2.7 except that the dynamic exclusion duration was set to 120 sec.

2.4.5 Database searching

The peak lists of MS^2 spectra were extracted from mass spectrometric raw files using MaxQuant (Version 1.0.11.2) (Cox and Mann, 2008).Top 200 MS/MS peaks per 100 Da were included into the peak list, while all other parameters were kept at default settings. Peak lists were searched against the *Saccharomyces* Genome Database using Mascot for identification of proteins in the sample preparations. The search parameters were set according to Table 2.3 with the following exceptions:

1) the cross-linker was BS^2G-d0 (BS^2G-d4 in the cases that labelled cross-linker used) modification on Lysine and protein N-termini;
2) for SILAC labelled samples, lysine $^{13}C_6$ label was also included in variable modifications.

Based on the Mascot search results, the final protein list for each protein complex preparation was generated using MaxQuant with default settings and allowed for 1% FDR.

For Mascot searches, the data from cross-linked Mad1-Mad2 complex Sample 1 (4 μg) and Sample 2 (8 μg) were combined.

The in-house program Xi was used for database searching for cross-linked peptides. For Mad1-Mad2 complex, data was searched against Mad1 and Mad2 sequence; For the Ndc80 complex, data was searched against a database containing Ndc80, Nuf2, Spc24 and Spc25. The tag sequence was included as part of the protein sequences. Search parameters were set as listed in Table 2.8, however the cross-linker and corresponding cross-linker modifications were set according to samples. Additionally, for SILAC labelled sample, lysine $^{13}C_6$ labelling was set as variable modification. As a control, these data were also searched with the same parameters but against the ten most intense proteins identified in corresponding sample preparations for the Mad1-Mad2 complex and NDC 80 complex. Spectra of all cross-linked peptides returned from Xi search were manually validated using in-house program Xaminatrix. Moreover, for the samples cross-linked with 1:1 mixture of BS^2G-d0 and BS^2G-d4, the cross-linked peptides were confirmed with 4 Da different doublets signals at MS^1 spectra level by hand.

2.4.6 Surveillance of inter-complex cross-links

In the SILAC labelled control samples that were designed to detect the occurrence of inter-complex cross-links, pattern of the heavy (lysine C13 labelled) and light (unlabelled) signals in MS^1 spectra for all identified cross-linked peptides were checked manually using Xcalibur (version 2.1.0, Thermo Scientific). The mass differences between the heavy and light signals were also verified by the number of labelled residues in the identified peptide sequences. For those cross-linked peptides that were only identified in non-labelled sample,, their MS signals in the corresponding SILAC control samples were retrieved based on the SCX fraction, m/z value and retention time.

2.5 Supplementary information and experimental procedures

2.5.1 Supplementary Information

2.5.1.1 Supplier information

All chemical reagents used in this study (if not stated otherwise) were supplied by Sigma-Aldrich and Fisher Scientific. All solvent used for HPLC and Nano-LC-MS/MS analysis were HPLC or LC-MS grade. Milli-Q water is used in all experiments. Trypsin (proteomics grade) was supplied by Sigma-Aldrich.

2.5.1.2 StageTips

All StageTips (Rappsilber *et al.*, 2003) used in this study were prepared in-house as described in (Rappsilber *et al.*, 2007); the Empore high performance extraction disks for preparing the StageTips were supplied by 3M.

2.5.2 Preparation of trypsin digested E.coli extract

2.5.2.1 Preparation of E.coli extract

Two 5 ml *E.coli* MRE 600 cultures were grown overnight in the LB medium at 37°C and were subsequently diluted to 2L of LB medium and grown to around 0.5 OD at 37°C. The *E.coli* culture was centrifuged at 5000 rpm with JLA 10.5 Rotor for 20 min at 4°C. After washing with 150 ml ddH_2O twice, the *E.coli* pellet was frozen at -80°C for over night. After thawing slowly on ice, the pellet was suspended in the lysis buffer containing 25 mM MOPS, pH7.5, 300 mM NaCl and 1 mM DTT with a ratio of 5ml buffer for 100ml 0.5OD culture. Lysozyme (Sigma-Aldrich) dissolved in ddH_2O was added to *E.coli* suspension to a final concentration of 1 mg/ml, incubating on ice for 30 min. The *E.coli* suspension was sonicated with 8 short burst of 15 sec followed by intervals of 30 sec for cooling on ice. The cell debris was removed by centrifugation at 10500 rpm with JA 25.5 rotor, 4°C for 50min.

The supernatant was collected and the protein concentration in the crude extract was determined using a Bradford assay.

2.5.2.2 In-gel digestion of E.coli extract

500 µg of *E.coli* extract was loaded on a NuPAGE 4-12% Bis-Tris gel (Invitrogen) with 50 µg material *per* lane. The short electrophoresis was stopped till the protein mixture have entered the gel. The gel was fixed in 50% methanol, 5% acetic acid and stained using a colloidal blue kit (Invitrogen). The bands of concentrated protein mixture were excised and combined in one tube. The in-gel reduction/alkylation and trypsin digestion was conducted as described in *2.2.4*. Trypsin was added with a 1:20 enzyme to protein mass ratio. After digestion, the peptide mixture was acidified with 0.1% TFA. Aliquots corresponding to 50 µg of starting material were loaded on a C18 StageTip for desalting (as described in *2.1.2.2*). Peptides were washed with 100 µl 0.1%TFA and StageTips were stored at -20°C.

2.5.3 Preparation of trypsin digested yeast extract

The yeast extract was provided by Sjaak van der Sar from Kevin Hardwick Lab. The prepared yeast lysate was digested by trypsin following the same procedure as for the *E.coli* extract. Desalted peptides on C18-StageTips were stored at -20°C in 50 µg aliquots.

2.5.4 Protocol for silver staining

2.5.4.1 Solutions for silver staining

a) Fixing solution: 40% ethanol, 10% acetic acid, 50% ddH_2O
b) Washing solution: 30% ethanol, 70% ddH_2O
c) Sensitizing solution: 0.02% $Na_2S_2O_3$
d) Silver nitrate solution: 0.2% $AgNO_3$, 0.02% formaldehyde

e) Developing solution: 3% Na_2CO_3, 0.05% formaldehyde

f) Termination solution: 5% acetic acid: 5% acetic acid

2.5.4.2 Silver staining procedure

1) Fix the gel with the fixing solution for 1 hour.
2) Wash the gel in the washing solution for 20 min, twice.
3) Wash the gel with ddH_2O for 20 min.
4) Incubate the gel in the sensitize gel: 0.02% $Na_2S_2O_3$ for 1 min.
5) Wash the gel with ddH_2O for 20 sec, 3 times.
6) Incubate the gel in the pre-cold silver nitrate solution for 20 min at 4ºC.
7) Wash the gel with ddH_2O for 20 sec, 3 times.
8) Change the gel to a new container.
9) Wash the gel with ddH_2O for 1 min.
10) Develop the stain with the developing solution until the protein bands appear.
11) Wash the gel with ddH_2O for 20 sec
12) Terminate the staining with the termination solution.
13) Wash the gel with ddH_2O for 10 min 3 times
14) Store the gel can in 1% acetic acid

Chapter 3

DEVELOPMENT OF A 3D PROTEOMICS ANALYTICAL WORKFLOW

3.1 Summary

For long time, detecting and correctly identifying cross-linked peptides have been challenges for 3D proteomics analysis and, as much, restricted its applications to single proteins and simple protein complexes. Enrichment of cross-linked peptides, advanced mass spectrometers, and improved data interpretation may all contribute to address this challenge. To facilitate the development of an advanced analytical workflow for 3D proteomics, a cross-linked peptide library was designed in order to generate large datasets of cross-linked peptides. A mixture of 49 synthetic peptides was cross-linked *via* the lysine residues. The theoretical size of the resulting library of cross-linked peptides was over 1600, of which 508 were identified with high confidence by mass spectrometry. For this cross-linked peptide library to represent the features of cross-linked peptides generated from trypsin digested proteins, the sequences of the synthetic peptides were derived from observed tryptic peptides from *E.coli* ribosome. This library was then used in three ways:

1) The charge distribution of cross-linked peptides revealed their preference to 3+ and higher charger states. This resulted in the establishment of a charge based enrichment strategy for cross-linked peptides;

2) Manual validation of high resolution MS^2 spectra of cross-linked peptides gave insight into the general CID fragmentation behaviour of cross-linked peptides and allowed for the development of a automated interpretation tool for fragmentation spectra of cross-linked peptides;

3) 1185 manually validated high resolution fragmentation spectra provided a representative and statistically meaningful dataset for further systematic studies on fragmentation rules of cross-linked peptides, which is fundamental for the development of a cross-linked peptide search algorithm.

This work with synthetic cross-linked peptides laid the foundation for the strategies in terms of enrichment, mass spectrometric acquisition and spectra annotation used in subsequent chapters for the analysis of multi-protein complexes.

3.2 Introduction

As introduced in Chapter 1, technical limitations on detecting and correctly identifying cross-linked peptides have impeded applications of 3D proteomics on complex protein samples. Progress has been made by us and others to overcome these limitations, and technical developments are still ongoing. To advance the application of 3D proteomics on more complex systems, for example large multi-protein complexes, our group further developed our experimental and computational workflows based on the previously reported workflow that was applied in the study of the 180 kDa Ndc80 complex, the largest protein complex analyzed by 3D proteomics at the time (Maiolica *et al.*, 2007). The new experimental procedure was designed by Professor Juri Rappsilber and I. In this new procedure, the lysine specific cross-linking chemistry and trypsin protein digestion were maintained (Maiolica *et al.*, 2007). Three key elements of this new procedure are the application of the LTQ-Orbitrap hybrid mass spectrometer, a charge based enrichment strategy for cross-linked peptides and a high-high acquisition scheme (high resolution MS and high resolution MS^2 measurements).

Using a library of cross-linked synthetic peptides, I have

1) Followed the general behaviour of cross-linked peptides through this experimental pipe-line and optimize the experimental settings;
2) Investigated the fragmentation data of cross-linked peptides generated with an established experimental setup;
3) Created datasets that allowed us to establish fragmentation rules of cross-linked peptides in a statistical fashion.

The construction of a library of cross-linked synthetic peptides constitutes a major advancement as only a very limited amount of cross-linked peptides can be obtained from single proteins or small model proteins complexes (for example 69 high confident cross-

linked peptide-matches were obtained from the 180 kDa Ndc80 complex (Maiolica *et al.*, 2007)). Cross-linking a mixture of 49 synthetic peptides allowed for more than 1600 theoretical peptide combinations (including modifications). On the basis of such a large library, the properties of cross-linked peptides could be studied in a representative way.

Analyzing the cross-linked peptide library by our experimental pipe-line led to a number of key insights.

- Firstly, cross-linked tryptic peptides tend to have a larger positive charge than non-cross-linked tryptic peptides. Consequently, cross-linked peptides enrich in high-salt SCX fractions and among species observed with charge states ($z \geq 3$) in the mass spectrometer. This forms the base of a charge-based enrichment strategy.
- Secondly, manual interpretation of fragmentation spectra suggested that cross-linked peptides follows the dissociation rules established for non-cross-linked linear peptides fragmented by CID.
- Thirdly, in contrast to low resolution data, high resolution fragmentation spectra with high mass accuracy and charge recognition can reduce the random matches in database searching and allowed for simplification of MS^2 spectra. This supported the application of the high-high acquisition scheme (high resolution MS and high resolution MS^2 measurements).
- Finally, manual validation using an automated annotation tool that was developed based on the manual interpretation gave rise to over a thousand high resolution fragmentation spectra of cross-linked peptides, which added statistical meaning to the dataset and proved essential for computational developments.

3.3 Analysis of cross-linked peptides library

3.3.1 Design of a cross-linked peptide library

To evaluate and optimize the advanced analytical workflow of 3D proteomics, a cross-linked peptide library was developed. This cross-linked peptide library was constructed by cross-linking a mixture of 49 synthetic peptides (Figure 3.1B, Table S1). The peptide sequences were retrieved (using scripts developed by Morten Rasmussen) from surface peptides of the *E.coli* ribosome 30S subunit crystal structure (PDB|3FIH, Figure 3.1A) and fulfilled the criteria of fully tryptic peptides. These peptides ranged in length from 5 amino acids up to 15 amino acids. Each of them contained one cross-linkable lysine residue and ended with a C-terminal arginine residue (except for four peptides that were each the C-terminal peptides of a protein sequence). The primary amine specific cross-linkers can couple two peptides *via* the side chain amine group of the lysine residue. The reactions to peptide N-termini were prevented by N-terminal Fluorenylmethyloxycarbonyl (Fmoc) protection. The N-terminal protection was removed after the cross-linking reaction, so that the cross-linked peptides may precisely mimic the trypsin digested cross-linked proteins. Cross-linking of this mixture of 49 synthetic peptides theoretically allows for combination of any two peptides. If the peptides with oxidized methionine were considered as distinct sequences to the non-modified counterparts, in total 1653 possible cross-linked peptides would be expected. These peptide combinations determine the database search space required for identification of all possible cross-linked peptides. The database size is comparable to the number of peptide combinations that need to be considered for identifying amine cross-linked peptides generated by trypsin digestion of cytochrome C (1595 peptide combination with maximum 3 missed cleavages allowed). In this study, the 49 synthetic peptide mixtures were cross-linked using BS^2G (Thermo Scientific, spacer length 7.7 Å) and BS^3 (Thermo Scientific, spacer length 11.4 Å) respectively.

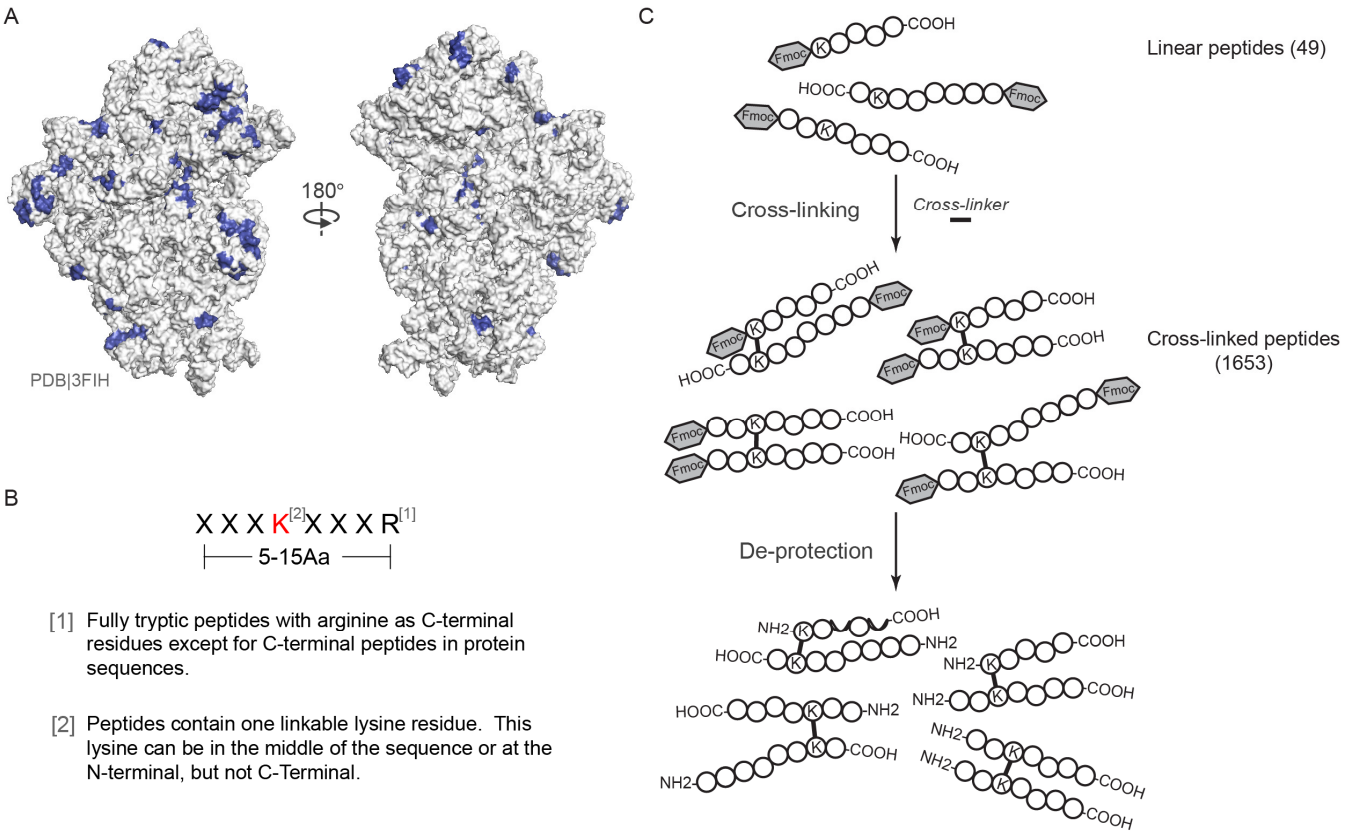


Figure 3.1 - Design of the cross-linked peptide library

A. 49 peptides (highlighted in blue) that were selected for construction of a cross-linked peptide library were taken from the structure of the *E.coli* ribosome 30S subunit (PDB|3FIH). The structure is shown in standard front and back view.

B. Sequence requirements of selected peptides

C. Cross-linking scheme for the cross-linked peptide library

3.3.2 LC-MS/MS analysis scheme for cross-linked peptides

Using a LTQ-Orbitrap hybrid mass spectrometer (Thermo Scientific) (Figure 3.2) is one of key elements of the advanced 3D proteomics analytical workflow. In this hybrid mass spectrometer, a linear ion trap (LTQ) is coupled with a novel mass analyzer Orbitrap which converts the time-domain signals of orbiting ions into a mass-to-charge spectrum using a fast Fourier transform (FT) algorithm (Makarov *et al.*, 2006). The LTQ-Orbitrap instrument is capable of LC-MS/MS. Automated high throughput MS/MS analysis is supported by the data dependent acquisition. As described in *1.4.2*, peptides eluted from the on-line LC were analyzed continuously in acquisition cycles. The ions fragmented in each cycle can be selected based on the intensity, the charge state and the m/z value. Fragmentation is carried out in the ion trap. Both MS^1 and MS^2 spectra can be recorded in either of two mass analyzers. The ion trap has advantages on speed and sensitivity, while the Orbitrap features high resolution and high mass accuracy.

In the analysis of cross-linked peptide samples, a high-high acquisition scheme was applied taking advantage of the high resolution and high mass accuracy of the Orbitrap. Since a <10 ppm mass accuracy is crucial for identification of cross-link peptides from a large database (Leitner *et al.*, 2010) the MS^1 spectra were acquired in the Orbitrap with resolution at 100,000. Three most intense peaks in each MS^1 spectra were fragmented by CID. From previous experience with linear peptides, the CID fragmentation conditions were set to generate single peptide bond cleavages. The fragmentation spectra were also acquired in the Orbitrap, however with a resolution set to 7500 to reduce the scanning time. Despite this, the high resolution MS/MS detection still compromises on the scanning speed and sensitivity in comparison with low resolution (2000) ion trap detection. In order to compare the impact of high and low resolution MS^2 spectra on the downstream data process, two types of MS^2 spectra were recorded in parallel for each precursor for some samples.

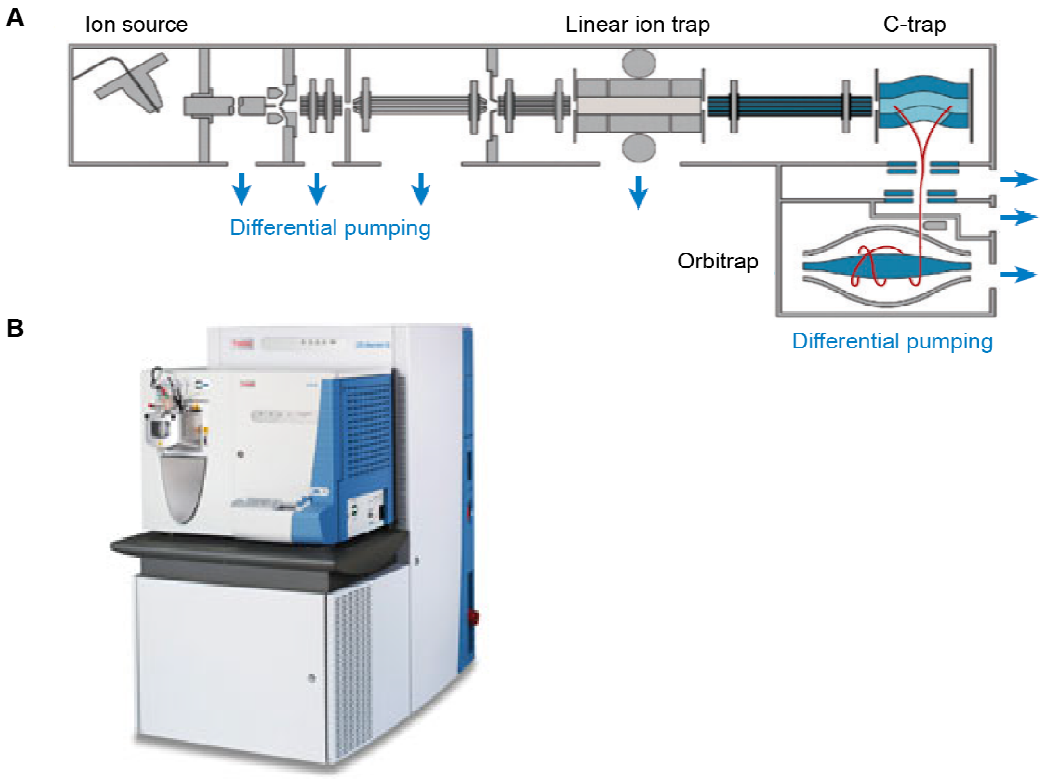


Figure 3.2 - LTQ-Orbitrap hybrid mass spectrometer

A. Overall diagram of the LTQ-Orbitrap (Hardman and Makarov, 2003).

B. A photograph of a LTQ-Orbitrap hybrid mass spectrometer (Thermo Scientific)

3.3.3 Database searching for cross-linked peptides

To identify cross-linked peptides, the high resolution fragmentation spectra of cross-linked synthetic peptide samples were searched against 1653 possible peptide combinations using in-house software XMass (Salman Tahir, unpublished). As the first step, cross-linked peptide candidates for each spectrum were extracted from the database by mass. In this study, the Orbitrap MS detection gave rise to <6 ppm mass accuracy (Makarov *et al.*, 2006)(Figure S1). Despite this high mass accuracy, there were still 219 among the 1653 peptides combinations that had the same mass as at least one other cross-linked peptide within a 6 ppm error tolerance. Hence the subsequent fragment matches were crucial for unambiguous identification of cross-linked peptides. The possible fragments for each candidate were calculated and compared to the observed signals in MS^2 spectra. The XMass algorithm calculated all possible b- and y- ions that are normally generated by CID fragmentation for both peptides in each candidate combination. The lysine residues were considered as the default cross-linkage sites and for each peptide the cross-linker together with the cross-linking partner were regarded as a modification on the cross-linked lysine residue. Since the electrospray ionization was applied, the precursors of collected MS^2 spectra carried charges from 1+ to 7+, so the charge states of fragments were calculated up to the precursor charge state. The measurement of fragment spectra of cross-linked peptides in the Orbitrap at 7500 resolution allowed for matching fragment ions with a 20 ppm error tolerance (Figure S1). XMass returned the matched MS^2 spectra and candidate peptide combinations in pairs. For spectra that had more than one matche, the matched candidates were ranked based on the number of matched fragment ions, and the top 3 matches were reported. Since a scoring scheme that can distinguish the true and false matches had not yet been established for this search algorithm, further manual validation was required to determine the confidence of identifications.

3.4 CID fragmentation of cross-linked peptides

3.4.1 Manual annotation of cross-linked peptide fragmentation spectra

30 identified high resolution fragmentation spectra of cross-linked peptides were manually annotated based on the sequences of cross-linked peptides returned by XMass. The selected spectra had precursors with charge states from 2+ to 6+, including spectra of three peptide pairs that were detected at different charge states, as well as spectra of three cross-linked peptides that were cross-linked with different cross-linkers (BS^2G and BS^3). Annotation of this representative collection of spectra did not only provide a general impression of fragmentation spectra acquired with our instrument settings, but also shed light on the CID fragmentation behaviour of cross-linked peptides, as well as the impact of different precursor charge states and cross-linkers on such behaviour.

3.4.2 High resolution fragmentation spectra of cross-linked peptides

MS^2 spectra acquired in the Orbitrap mass analyzer led to two major benefits for the interpretation of fragmentation spectra. First of all, the *m/z* of fragment ions were measured with <20 ppm error (Figure 3.6 A). Secondly, in high resolution spectra, the isotope clusters of fragment ions could be resolved and their charged state could be recognized (Steen and Mann, 2004). Therefore the signals of fragment ions in MS^2 spectra can be assigned with both accurate mass and correct charge state, which significantly increase the accuracy of annotation.

In these spectra, the fragment ions were predominantly derived from single cleavages of peptide bonds. Series of b and y ions that encoded the peptide sequences were observed for both peptides and allowed for confident and unambiguous identification of them (an example is shown in Figure 3.3). Besides the b- and y-ions, neutral losses ions of b- or y-ions and precursor ions were observed, which included the loss of water and

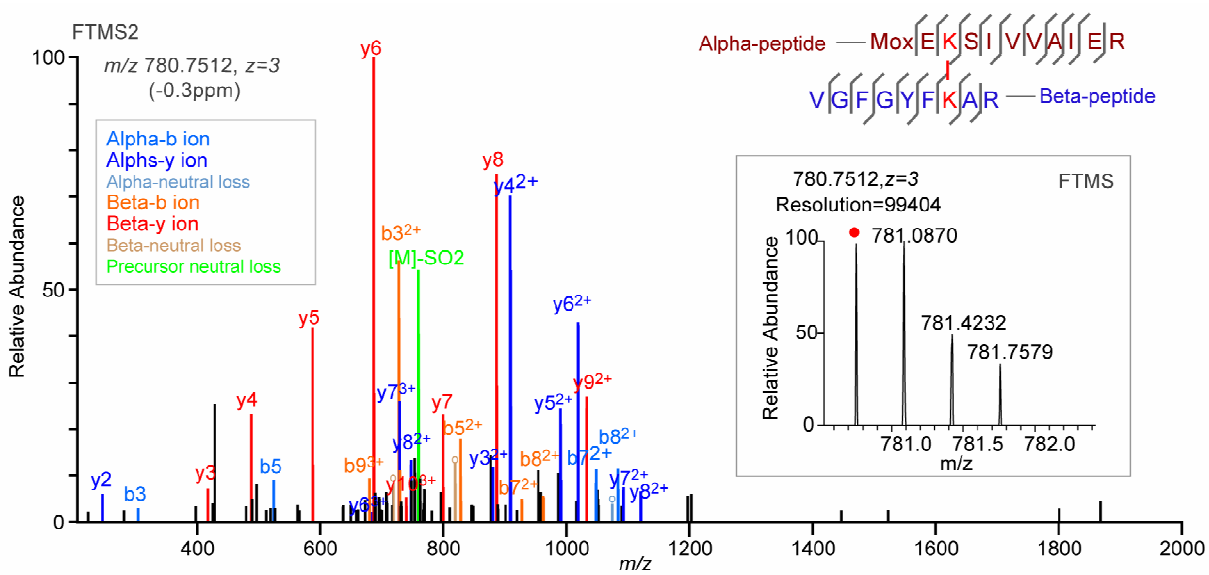


Figure 3.3 - Annotation of fragmentation spectra of cross-linked peptides

The annotated high-resolution fragmentation spectrum of cross-linked peptide MoxEK(xl)SIVVAIER-VGFGYFK(xl)AR. The annotated peaks are labelled and coloured. The b- and y-ions are labelled with the fragment name. For ions with charge state higher than 1+, the charge states are indicated as superscript. The neutral loss from precursor is labelled as "[M]-" and the loss group; neutral loss from b- and y- ions are labelled with a circle. The ions from different peptides are distinguished by colour. The observed fragments are indicated in the corresponding peptides sequence. The magnified view of the precursor signal is shown in the inlet.

ammonia, but also of the –SO2 group in the peptides that contained oxidized methionine. The spectra of a cross-linked peptide with different precursor charge states revealed that the charge states of fragmentation ions tended to span from 1+ up to the charge of the precursors. The majority of fragment ions that contain cross-linked lysine were with charge state 2+ or higher.

To distinguish two cross-linked peptides in annotation, they were named as the 'α-peptide', which had relatively more fragment ions observed in the spectrum, and the 'β-peptide' (only b- and y-ions were counted). As the example in Figure 3.4 shows, the cross-linked peptide spectra were aligned on *m/z* axis with the spectra of the individual α and β peptides, acquired with the same instrument settings. The fragmentation pattern of either peptide in a cross-linked pair was similar to that of its linear counterpart in not only the observed b- and y-ions, but also the intensity distribution of these ions, and even the detected neutral loss ions. This proves that cross-linked peptides follow the same fragmentation rules established for linear peptides.

For both cross-linked peptides, the fragments upstream to the cross-linked residue represented only the mass of the included residues, while the mass of fragments downstream to the linked residues included the additional mass of cross-linker and cross-linking partner. Therefore, detection of a series of fragments that fall upstream and downstream of the linked residue will narrow down the linkage site in peptide sequences. Observation of a pair of subsequent fragments that locate right before and after the cross-linked residue can point out the exact cross-linkage site. An example for this is the y8/y9 fragments of the α-peptide and y2/y3 fragments of the β-peptide in the example spectrum in Figure 3.2. The only type of fragments involving cross-linker fragmentation was derived from a cleavage at the N-O bond between the Nε of lysine side chain and the cross-linker spacer chain. This cleavage was also observed for another four amine specific cross-linkers (Gaucher *et al.*, 2006).

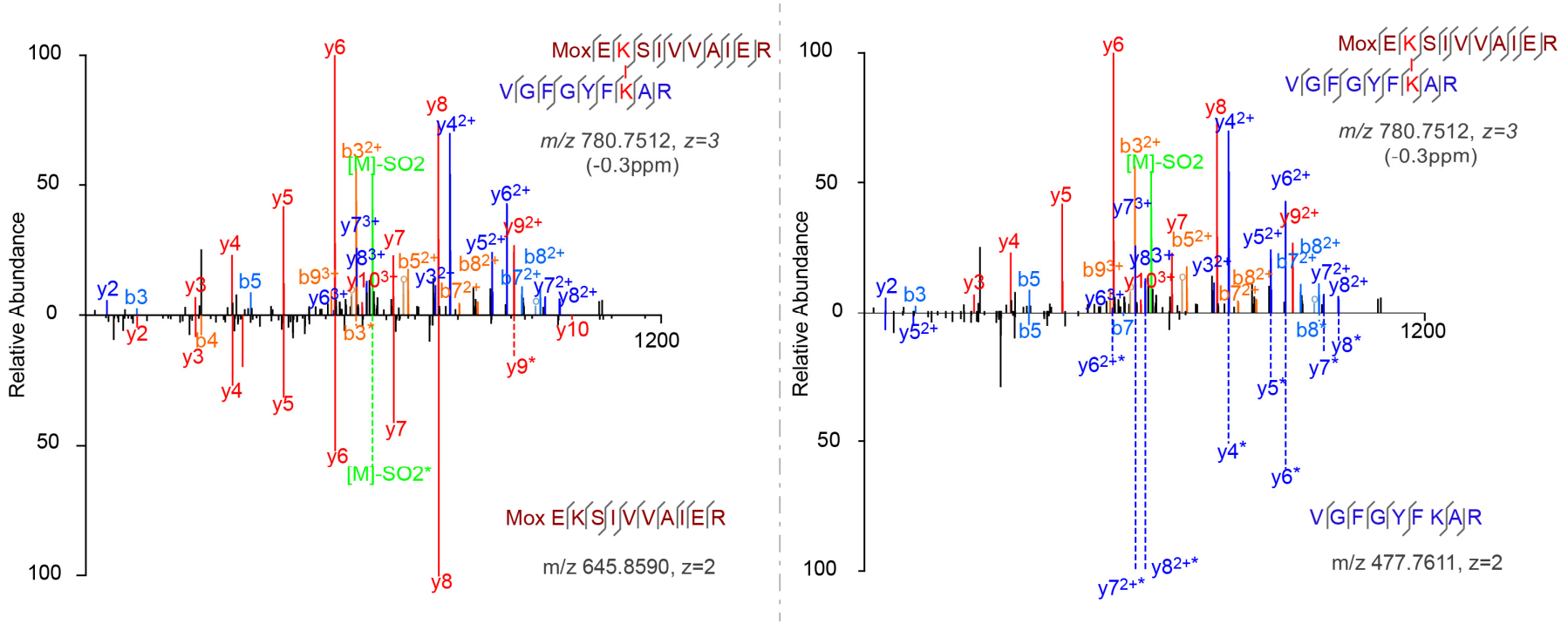


Figure 3.4 - Peptide fragmentation pattern are similar in cross-linked and linear status

The fragment ions observed from cross-linked peptides MoxEK(xl)SIVVAIER-VGFGYFK(xl)AR (the topmost spectrum) were compared to the fragments observed from each linear peptide (bottom spectrum). The fragments that do not contain the cross-link site coincide with linear peptide fragments in *m/z* and charge state; the signals in the linear peptide spectra corresponding to fragments that contain the cross-link site have been shifted and aligned with the corresponding peaks in the cross-linked peptide spectra (indicated with dashed lines).

However, these fragments were neither observed in all cross-linked peptides nor dominant in signal intensity.

3.4.3 The influence of different cross-linkers on the fragmentation of cross-linked peptides

The influence of different cross-linkers on the fragmentation of cross-linked peptides was investigated by comparing the fragmentation of the identical peptide combinations cross-linked with BS^2G and BS^3. As shown in Figure 1.2 C, these two amine-reactive cross-linkers have the same function groups and only differ on spacer lengths. The chosen spectrum pairs were observed with the same charge states and same total number of fragmentation ions. As demonstrated in Figure 3.5, the presence of different cross-linkers showed no obvious effect on the fragmentation pattern. This observation is consistent with a previous study on other amine-reactive cross-linkers (Gaucher *et al.*, 2006). These two example spectra matched on not only the observed fragment ions, but also the intensity and charge distribution of these ions, including the neutral loss peaks. The fragment ions that did not contain the cross-linker were nearly perfect aligned, while the cross-linker containing ions exhibited a 42 mass shift, which originated from the mass difference of BS^2G and BS^3. As mentioned previously, a cleavage between the cross-linker and the side chain of the linked lysine has been observed for both BS^2G and BS^3, but not necessarily in every spectrum.

3.4.4 The impact of resolution for MS^2 spectra on interpretation and identification of fragmentation spectra of cross-linked peptides

High mass accuracy and high resolution significantly increases the accuracy of interpretation of fragmentation spectra of cross-linked peptides. However, this requires compromise on the

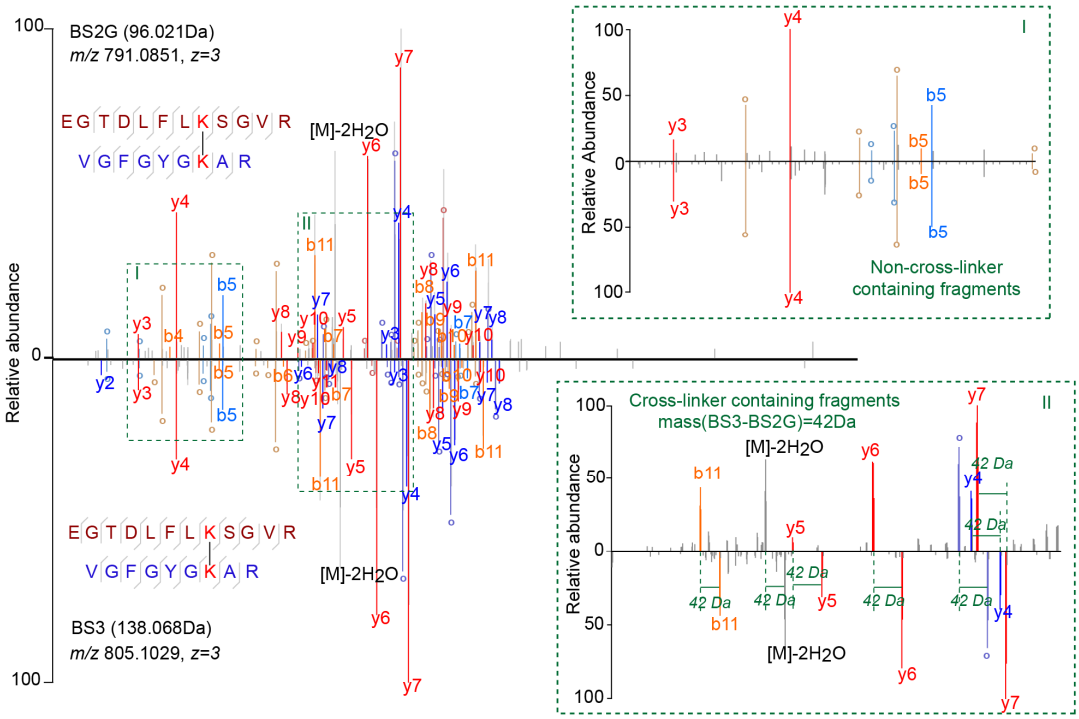


Figure 3.5 - Impact of cross-linker on fragmentation

A high resolution fragmentation spectrum of BS^2G cross-linked peptides were aligned on *m/z* axis with the high resolution fragmentation spectrum of same peptides combination that were cross-linked by BS^3. The two spectra were detected with the same precursor charge states and same total ion counts. Two zoomed in views on the aligned spectra are shown inset.

scanning speed and detection sensitivity. The advances of high resolution fragmentation spectra in the downstream data process for identifying cross-linked peptides in database searches were evaluated by the comparison with low resolution MS^2 of cross-linked peptides acquired in the ion trap. This comparison was carried out using 27 manually annotated high resolution MS^2 ($FTMS^2$) spectra of cross-linked peptides and their corresponding ion trap MS^2 ($ITMS^2$) spectra. The paired high and low MS^2 spectra were acquired parallel for the same precursors in the same acquisition cycle.

As expected there were more peaks detected in their corresponding $ITMS^2$ spectra (Figure 3.6 B). However the intense peaks (with more than 10% relative intensity in the spectra) in $ITMS^2$ spectra overlapped with those of corresponding $FTMS^2$ spectra to a large extent. Manual annotation showed that the peaks in the $ITMS^2$ spectra can be assigned to the cross-linked peptide fragments on *m/z* value although with larger mass error; however the matches on the charge states were seldom achieved, especially for higher charged fragment ions (Figure 3.6 B).

The complexity of MS^2 spectra of cross-linked peptide caused by the presence of fragmentation ions from two peptide sequences and multiple charge states of these fragment ions would result in an increased chance of false matches (random matches) in database searches, whereas for the $FTMS^2$ data this probability can be diminished by both high mass accuracy and assigned charge states of fragment ions (Figure3.6C). Although identifying cross-linked peptides from large sequence databases using $ITMS^2$ spectra can be achieved (Rinner *et al.*, 2008), the computational strategy relied on the use of isotopically coded cross-linkers. This method limits the use of isotope labelling for quantitative analysis and thus reduces the value of 3D proteomics (Chapter 5).

Moreover, the signals of each recognized isotope cluster in the $FTMS^2$ spectra can be simplified to the neutral mass of detected fragments. This de-charging and de-isotoping process can reduce the complexity of the fragmentation spectra and results in a smaller

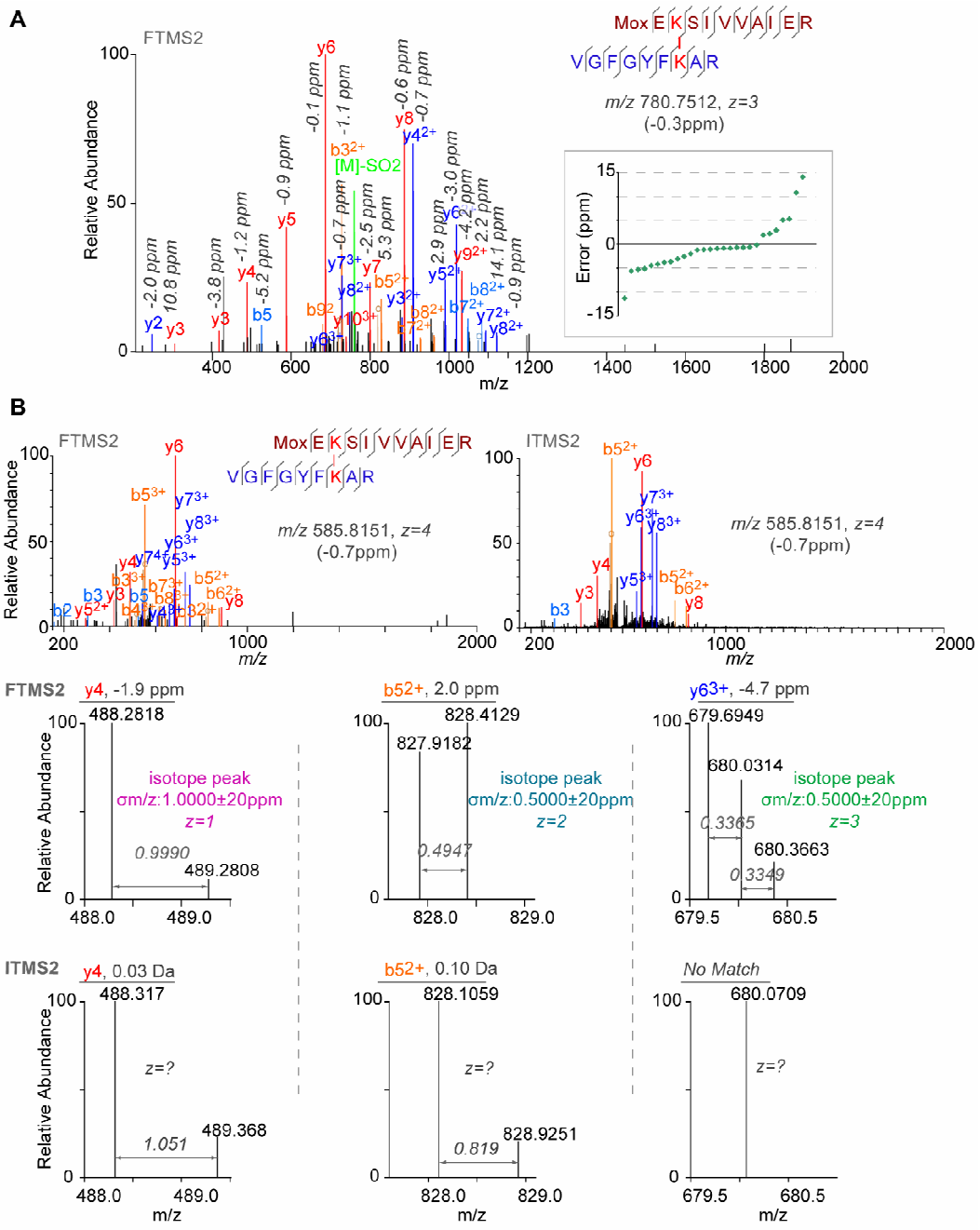


Figure 3.6 -High and low resolution MS^2 spectra of cross-linked peptides.

A. The mass errors of fragment ions are labelled for high intense peaks in the annotated cross-linked peptides spectrum presented in Figure 3.2A. The distribution of mass errors for all observed fragments ions are plotted inset.

B. High resolution and low resolution fragmentation spectra of cross-linked peptide MoxEK(xl)SIVVAIER-VGFGYFK(xl)AR acquired from same precursor. Peaks with >10% relative intensity are annotated in both spectra. Isotope clusters of y4, b52+ and y63+ ions are displayed in the magnified view with charge state annotation. The corresponding signals in the low resolution spectra are displayed berneath.

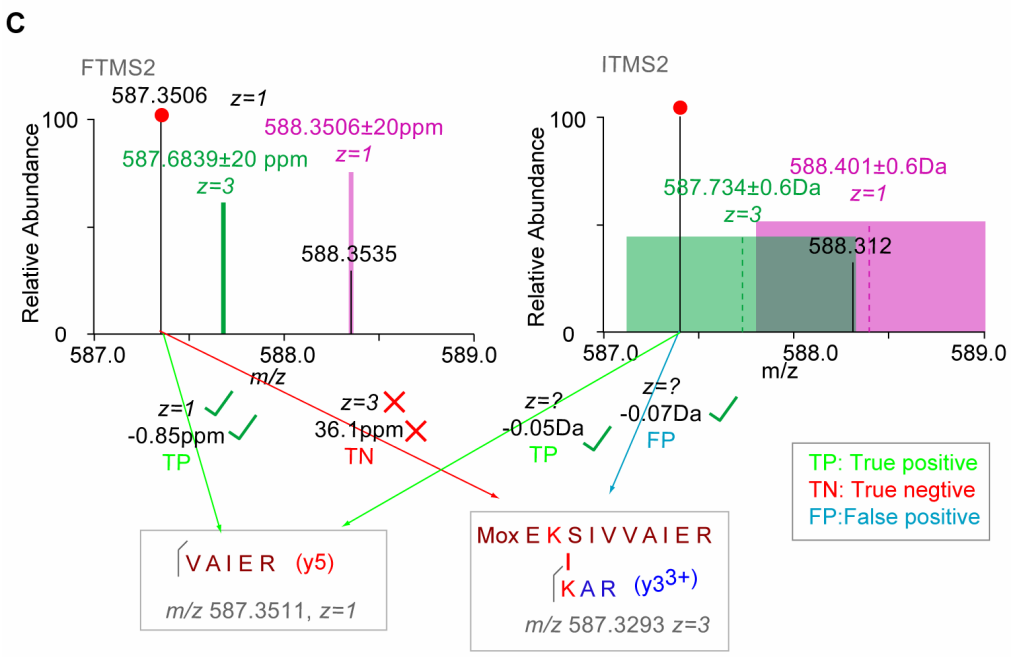


Figure 3.6 continued - High and low resolution of cross-linked peptides.

C. High resolution fragmentation spectra can avoid false fragment matches in database searches. The corresponding high ($FTMS^2$) and low ($ITMS^2$) resolution fragmentation spectra in figure 3.3 B are shown magnified to in a 578.0-589.0 *m/z* window. In the $ITMS^2$ spectrum, the signal at *m/z* 587.401 can match both y5 and $y3^{3+}$ ions of the candidate cross-linked peptide with accepted mass error (<0.6Da). Its corresponding signal in $FTMS^2$ was detected at *m/z* 587.3506 with a resolved isotope peak at 588.3535 which assigned the 1+ charge state of this fragment ion. Both the charge state and 20 ppm error tolerance excluded the match to the $y3^{3+}$ ion and unambiguously assigned this signal to the y5 fragment. (The *m/z* window for predicted 1+ (pink) and 3+ (green) isotope peaks within mass error tolerance for the "587" peak were highlighted in both spectra)

database search space and decreased computation time. Furthermore, it will also reduce the probability of random hits in database searches.

3.4.5 Automated interpretation of MS^2 spectra of cross-linked peptides

Comprehensively interpreted spectra are a prerequisite for manual validation of cross-linked peptides identification. Manual annotation is a labour intensive and time consuming task, and not practical for large datasets. Based on the observation in the 30 manually annotated spectra and in collaboration with Morten Rasmussen as programmer, the in-house software Xaminatrix was developed to carry out an automated interpretation of cross-linked peptide spectra. The m/z values of b- and y-ions and observed neutral loss ions were calculated up to the charge state of the precursor. The cleavage between cross-linker and lysine side chain was also included. Calculation for the double fragmentation ions caused by two cleavages in the peptide backbone, were also implemented, but included as an optional function, since it requires more computation time. The automated computation shortened the annotation time for each spectrum from about 45 min to about 2-10 seconds, this improvement made it possible to manually validate cross-linked peptide identification for large datasets.

3.5 Validation of cross-linked peptide identification

3.5.1 Confidence criteria of cross-linked peptide identification

Since the returned matches between spectra and cross-linked peptide candidates were not marked as true or false identifications by the search algorithm, the confidence of these matches needed to be determined by manual interrogation. The quality of a match was judged by firstly the portion of observed fragments among all predicted backbone cleavages from both peptides sequences (b- and y- type ions in this case), and secondly the percentages of both absolute number and relative intensity of explained peaks in the spectra. Then the

matches were categorized into 3 confidence levels according to the following criteria. As demonstrated in Figure 3.7, a match will be marked with "Confidence A" when at least 80% expected peptide backbone cleavages for each of the peptides are observed in a spectrum and 80% of intense peaks (with over 10% relative intensity) in the spectrum are annotated by predicted b- or y-ions and neutral loss ions derived from them. If a match does not fulfil the criteria for A level confidence, but more than 60% of intense peaks in the spectrum is explained and at least one sequence encoding fragment series that contains no less than 4 serial fragments (3 for peptides with only 5 or less amino acids) is detected for both peptides, it will be scored as "Confidence B". Both "Confidence A" and "Confidence B" matches were considered as high confidence matches. However two cross-linked peptides were not always equally fragmented, particularly in the cases where one peptide is much shorter than the other. Frequently in these cases, one peptide was fragmented predominantly while the other had only few non-sequential fragments or even none. When the better characterized peptide among the two fulfilled the criteria for "Confidence A", and the other was identified unambiguously by mass, the candidate was still considered as a true hit because the mass of peptides was measured with high accuracy. These matches will be marked as "Confidence C" and regarded as low confidence matches

3.5.2 A large dataset of cross-linked peptides

Using the automated annotation tool Xaminatrix, manual validation was performed for identified cross-linked peptides from three sample sets:

1) SCX-HPLC fractions of the BS^2G cross-linked 100 µl (about 8 nmol per peptides) of 49-synthetic peptide mixture.
2) SCX-HPLC fractions of the BS^3 cross-linked 50 µl of 49-synthetic peptide mixture.

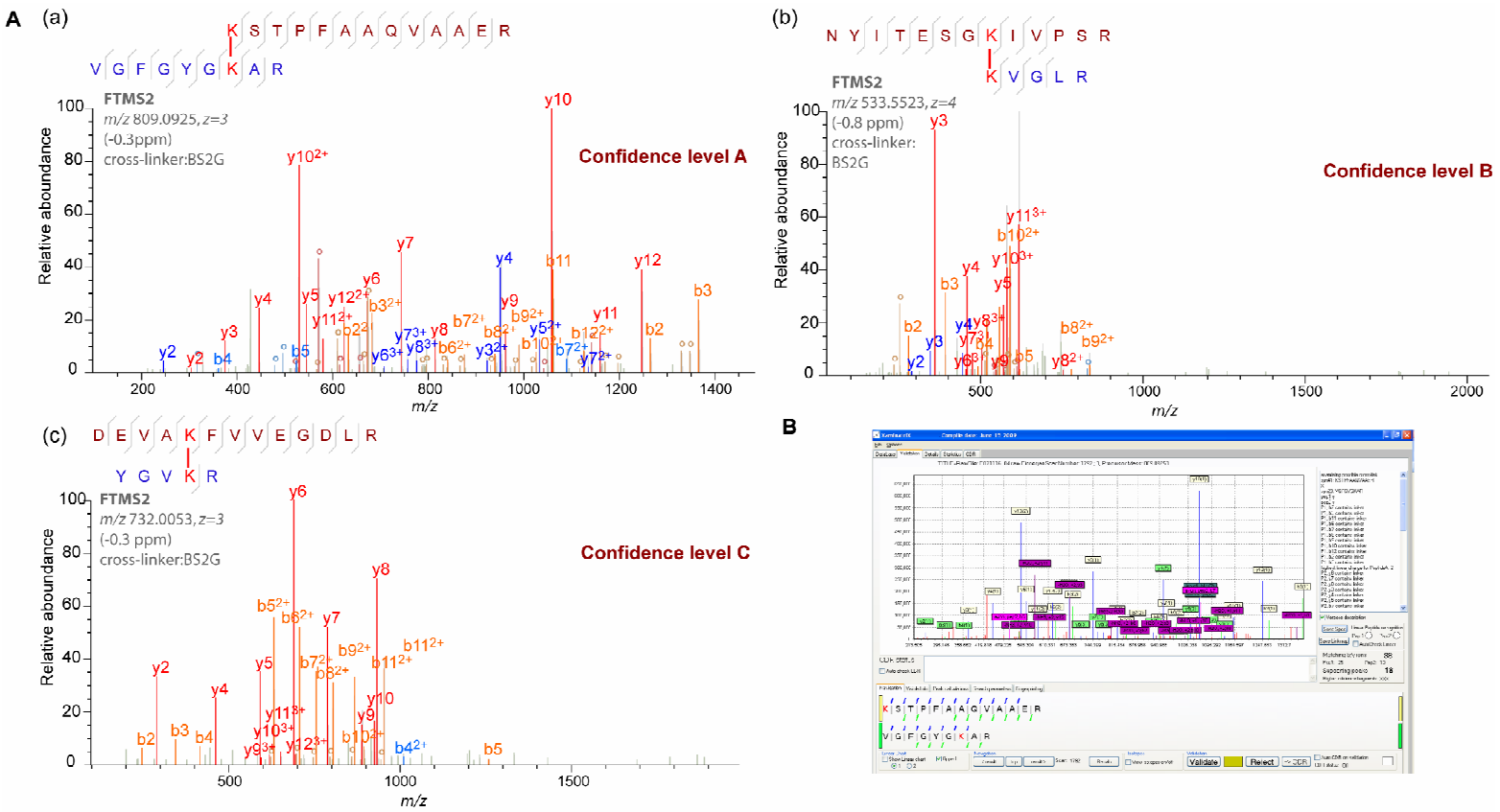


Figure 3.7 - Validation of cross-linked peptide fragmentation spectra matches.

A. Examples of annotated high resolution fragmentation spectra of cross-linked peptides from 3 validation confidence levels.

B. Screenshot of an annotated high resolution fragmentation spectrum of cross-linked peptides in automated annotation software Xaminatrix

3) SCX-Stage-Tip fractions of BS^2G cross-linked 10 µl of 49-synthetic peptide mixture.

In total, 1185 identified cross-linked peptide spectra were validated as true matches; 75% (891) of them were marked as high confident matches. These spectra identified 508 unique cross-linked peptides, and 223 of them were identified from both BS^2G and BS^3 cross-linked material. The details of identified spectra and cross-linked peptides from these three samples are summarized in Table 3.1. The number of both identified spectra and unique cross-linked peptides showed direct correlation to the amount of material analyzed.

The 1185 manually annotated high resolution CID fragmentation spectra of cross-linked peptides with annotation information were stored in a database for further statistical studies. The identification of 508 cross-linked peptides qualified the cross-linked peptide library as a model for characterizing the general behaviour of cross-linked peptides in various experimental procedures.

3.6 Charge-based enrichment strategy for cross-linked peptides

As discussed in *1.3.2*, enrichment of cross-linked peptides plays important role in the analytical workflow of 3D proteomics to improve detection of cross-linked peptides. Several features of cross-linked peptides can be utilized for enrichment. In this advanced experimental procedure, we made use of the feature that trypsin digested cross-linked peptides tend to carry more positive charges under acidic conditions. A typical tryptic peptide has two positive charges that locate on the N-terminal and the C-terminal residues. When two tryptic peptides are connected by cross-linking, the resulting molecule carries twice the number of basic sites. Theoretically, cross-linked peptides are more likely to be observed with relatively high charge states (>2+) in mass spectrometers and present higher affinity to cation ion exchange matrix in chromatography. Therefore we designed a charge based enrichment strategy including the selective fragmentation of > 2+ ions in mass

Table 3.1 - Summary of manually annotated cross-linked peptide identifications

Sample information			Unique cross-linked peptides	Spectra			
49 synthetic peptides	**Cross-linker**	**Fractionation**		**Spectra**	**Confidence A**	**Confidence B**	**Confidence C**
50 µl	BS^3	SCX-HPLC	295	347	162	101	84
100 µl	BS^2G	SCX-HPLC	397	838	291	337	210
10 µl	BS^2G	SCX-StageTip	103	135	50	20	65
Total			**508**	**1185**	**453**	**438**	**294**

spectrometric analysis and fractionation using SCX chromatography. Here, the performance of this enrichment strategy is evaluated by using a cross-linked peptide library.

3.6.1 Strong cation exchange chromatography and cross-linked peptide enrichment

Fractionation by strong cation exchange chromatography was conducted for 100 µl BS^2G cross-linked synthetic peptides and 50 µl BS^3 cross-linked synthetic peptides. The peptide mixtures were eluted with a salt gradient. Fractions were analyzed by LC-MS/MS. The distribution of identified cross-linked peptides suggested that cross-linked peptides were enriched in the high salt fractions (Figure 3.8 A (a) and (b)). In the BS^2G cross-linked sample, 99% of cross linked peptides were identified in the fractions eluted with >7% buffer B (equal to 70 mM KCl), and the cross-linked peptides identified in the BS^3 cross-linked sample were all eluted in the same part of the gradient. Apparently after the cross-linking reaction, cross-linked peptides were not the only components in the peptide mixture. They were accompanied by a large excess of non-cross-linked or cross-linker modified linear peptides. These peptides were identified using Mascot (Matrix Science); and it is worthwhile to mention that there was no peptide with FMOC on the N-terminal identified, which proved are fully tryptic status of all peptides in the mixture. As reflected on the chromatograms, more than half of the material was concentrated in the low salt fractions where no cross-linked peptide was identified. This peptide mixture to some extent reflected the composition of digested cross-linked protein samples, however not on the complexity of linear peptides. In order to better imitate the distribution of linear peptides in the SCX fractionation, 100 µg of trypsin digested *E.coli* extract was separated with the identical gradient used for the BS^2G cross-linked synthetic peptide sample. Based on the chromatogram, 48% of material fell in the cross-links enriched eluent range (Figure 3.8 A (c)). Therefore with the experimental setting presented here, two-fold enrichment for cross-linked peptides in a complex peptide mixture by SCX chromatography can be expected.

According to our experience from linear peptide samples, at least 50 µg of material is required for complex peptide mixtures to efficiently recover from SCX HPLC fractionations for the subsequent mass spectrometric analysis. In order to accomplish the SCX based enrichment for small amounts of cross-linked material (less than 30ug), a substitute procedure, using SCX-StageTips (Rappsilber *et al.*, 2003), was developed and tested with the cross-linked peptide library. 10 µl BS^2G cross-linked synthetic peptide mixture was eluted with a step salt gradient from a SCX-StageTip into 5 fractions (flow through as fraction 0; fraction 1, eluted with 20 mM NH4AcO; fraction 2, eluted with 50 mM NH4AcO; fraction 3, eluted with 100 mM NH4AcO; fraction 4, elute with 500 mM NH4AcO). 94% of 103 cross-linked peptides were identified in the later two high salt fractions. The separation of linear peptide mixture was mimicked using 10 µg of trypsin digested *E.coli* extract. Based on the number of identified peptides in each fraction, less than half of the material remained in the two high salt fractions, which gave a similar enrichment scale as using the HPLC system (Figure 3.8 B). Consequently, the SCX-StageTip proved capable to perform enrichment for cross-linked peptide in small amounts of material.

3.6.2 Selective fragmentation of highly charged precursor ions in mass spectrometric analysis increases detection of cross-linked peptides.

Although SCX chromatography fractionation can enrich for cross-linked peptides, it does not isolate them from linear peptides. Competition with linear peptides will cause reduced yield of fragmentation spectra of cross-linked peptides, since they generally have a low abundance relative to linear peptides, and in a data dependent acquisitions, the selection of ion for fragmentation mainly depends on the ion intensity. When there was no restriction on the precursor ion charge states during the acquisition, 93% of all identified cross-linked peptide spectra had triply or higher charged precursors corresponding to 98% of identified cross-linked peptides from the SCX-HPLC fractionated BS^2G cross-linked synthetic peptide

sample (Figure 3.9A). The charge distribution of identified linear peptides in each SCX-HPLC fraction of 200 μg trypsin digested yeast extract showed that higher charged peptides tend to be enriched in high salt fractions. However in the fractions where 99% of cross-linked peptides were enriched, still 50% of yeast peptides were identified with 1+ or 2+ charge states (Figure3.9B). An even higher proportion was detected in the high salt SCX-StageTip fractions of 1 μg *E.coli* extract sample (Figure 3.8 B). When acquiring with no precursor charge selection, only 24.8% of identified spectra was with precursor charge 3+ or higher, while in a repeat acquisition with 1+ and 2+ changed precursor excluded, the identification of triply and higher charged ions was increased by 103% (Figure 3.8C). Hence implementation of precursor charge selection in the data dependent acquisition for mass spectrometric analysis can efficiently increase the fragmentation of triply and higher charged ions that include most cross-linked peptides. This mass spectrometric acquisition level enrichment has been previously used in the analysis of the Ndc80 complex (Maiolica *et al.*, 2007). A combination with the SCX chromatography fractionation further improved the efficiency of enrichment.

This charge based enrichment strategy has no limit on selection of cross-linkers and can be applied in wide range of cross-linked protein samples. This charge based enrichment strategy was also reported later by Rinner and co-workers (Rinner *et al.*, 2008). Its application facilitated identification of cross-linked peptides from a total *E.coli* lysate. However, the performance of this enrichment strategy can be significantly affected by the complexity of protein samples and the efficiency of trypsin digestion. Because highly charged non-cross-linked peptides, caused by existence of another chargeable residue histidine and missed cleavages in trypsin digestion, cannot be separated from cross-linked peptides.

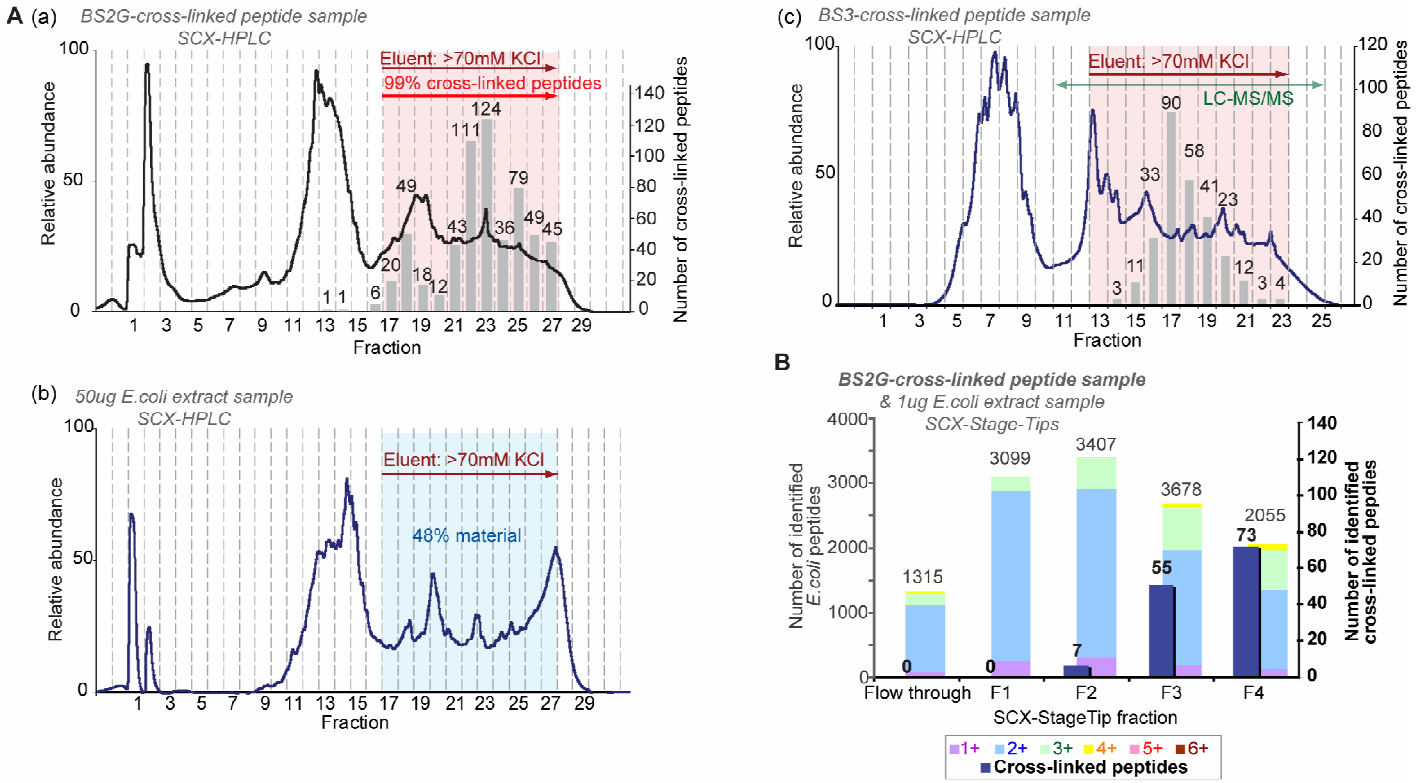


Figure 3.8 - Cross-linked peptide enrichment by SCX chromatographic fractionation.

A. The distribution of identified cross-linked peptides in SCX HPLC fractionations of BS^2G (a) and BS^3 (c) cross-linked synthetic peptide mixture samples are displayed on top of the chromatogram (Baseline extracted). Cross-linked peptides showed enrichment in fractions eluted with >70mM KCl (highlighted in pink). The distribution of linear peptides in fractions was mimicked with 50 µg *E.coli* extract. The fractions where cross-linked peptides were enriched in are highlighted (blue).

B. Distribution of cross-linked peptides (foreground) and linear peptides (background) in SCX-Stage-Tip fractions. The cross-linked peptides were identified from BS^2G cross-linked 10 µl synthetic peptides sample. The material distribution of 1 µg *E.coli* extract sample is represented by the number of identified linear peptide spectra in the fractions. The charge distribution of these identified linear peptide spectra in each fraction are shown in stacked columns.

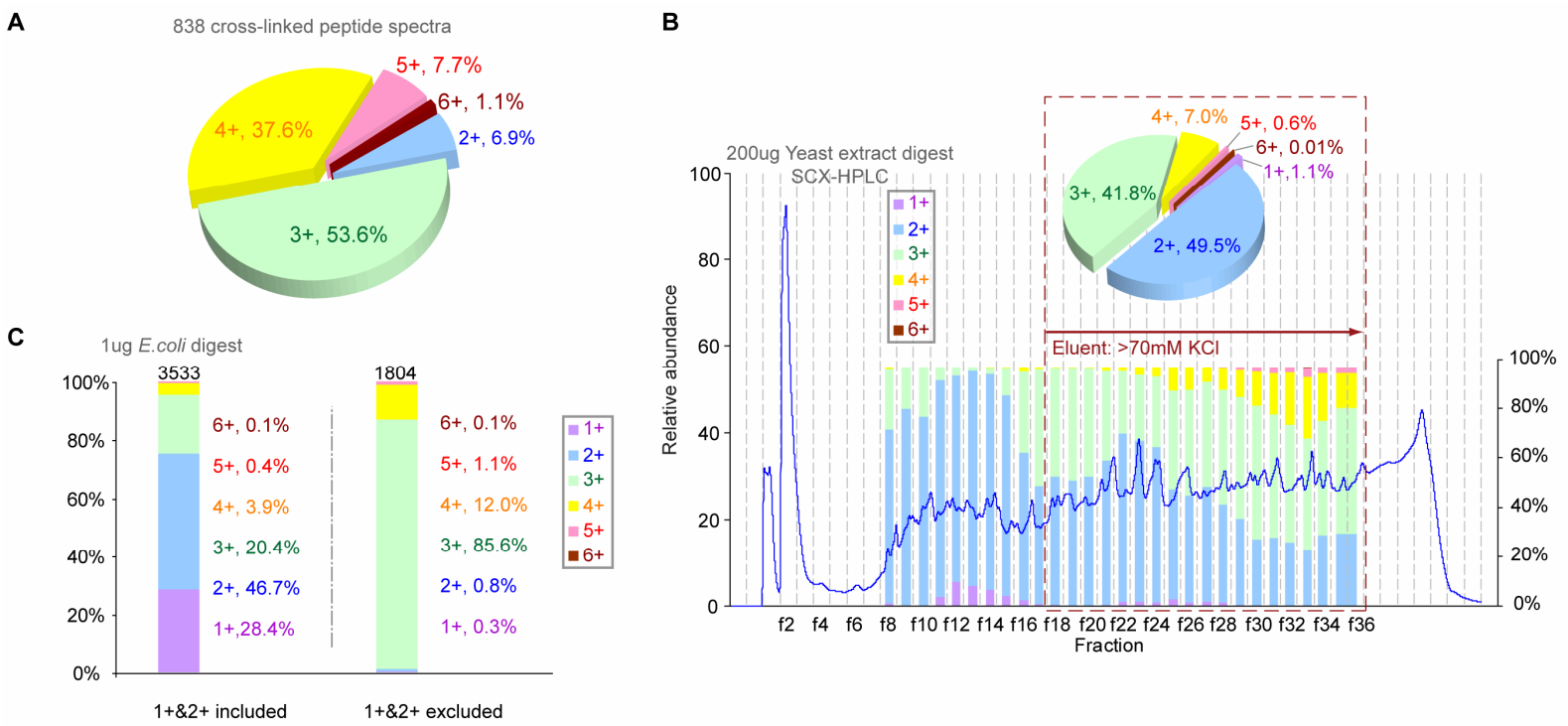


Figure 3.9 - Precursor charge selection and cross-linked peptide enrichment

A. The charge distribution of identified cross-linked peptide spectra across SCX-HPLC fractions of 100 µl BS^2G cross-linked peptide library sample.

B. The charge distribution of identified linear peptides from 200 µg yeast extract sample in SCX-HPLC fraction are shown in percentage stacked columns displayed on top of chromatogram. The summed charge distribution for spectra identified in fractions eluted with >70mM KCl are shown in inset.

C. Influence of precursor charge selection on charge distribution of identified peptide spectra. 1 µg *E.coli* extract sample was analyzed by LC-MS/MS with and without excluding the 1+ and 2+ ions for fragmentation. The charge distribution of identified peptide spectra from both acquisitions are displayed in percentage stacked columns.

3.7 Cross-linked peptide library and advanced 3D proteomics analytical workflow

In this work, the advanced analytical workflow of 3D proteomics was evaluated and optimized using a library of cross-linked synthetic peptides. The charge based enrichment strategy improved detection of cross-linked peptides by mass spectrometry. The LC-MS/MS setup using the LTQ-Orbitrap hybrid mass spectrometer allowed for acquisition of both MS^1 and MS^2 spectra with high resolution. High mass accuracy and charge recognition benefited from the high resolution fragmentation spectra will significantly reduce the probability of false match in database searches and also allowed for simplification of MS^2 spectra in data process. Additionally, the analysis of the cross-linked peptide library provided a large dataset with more than 1500 high quality high resolution MS^2 spectra of cross-linked peptides. A statistical study on the fragmentation rules of cross-linked peptides using this dataset (Morten Rasmussen and Lutz Fischer, unpublished) contributed fundamentally to the development of Xi, a new search algorithm for cross-linked peptides that allows for searching again large protein databases in the scale of whole human proteome (Salman Tahir *et al.*, unpublished). Moreover, based on the experience of manual interpretations, automated annotation software for high resolution MS^2 spectra of cross-linked peptides has been developed (in collaboration with Morten Rasmussen) which enables the manual validation of cross-linked peptide matches for large datasets.

3.8 Applications of the cross-linked peptide library

In this study, the cross-linked peptide library was developed to provide a large dataset of cross-linked peptides for the evaluation and optimization of the advanced analytical workflow of 3D proteomics. Although this library was constructed using synthetic peptides, the native peptide sequences and their variety in peptide length, amino acid composition, location of cross-link site in the sequence *et cetera*, made this peptide mixture a likely

representative dataset for trypsin digested cross-linked peptides from biological protein samples. Moreover, identification of 508 unique cross-linked peptide pairs of 46 synthetic peptides (in total 49) by mass spectrometry further confirmed the statistical value of this dataset. Therefore, beside the application in this work, this cross-linked peptide library has also been used as a standard tool by others in the Rappsilber group, for example, to investigate the performance of the newly released LTQ Orbitrap Velos instrument on cross-linked peptide detection; to characterize the dissociation behaviours of cross-linked peptides with HCD fragmentation. Moreover, it has contributed to the discovery of a series of reporter ions of cross-linking products for the amine-reactive cross-linkers (Helena Barysz, Lutz Fishcher, unpublished).

Chapter 4

ARCHITECTURE OF THE RNA POLYMERASE II-TFIIF COMPLEX REVEALED BY 3D PROTEOMICS

4.1 Summary

In this study, 3D proteomics was applied to analyze the RNA polymerase II (Pol II) in complex with general transcription factor II F (TFIIF). The methodology was validated by examining the consistency between 3D proteomics analysis results and X-ray crystallographic data on the 513 kDa 12-subunit Pol II complex. Following this, the analysis was applied to the 670 kDa 15-subunit Pol II-TFIIF complex. The results revealed interactions between Pol II and TFIIF with peptide resolution. The location of TFIIF on Pol II allows for further discussion on TFIIF functions during transcription initiation. Moreover, cross-link data reflected the dynamic nature of Pol II-TFIIF structure and implied possible Pol II conformational changes induced by TFIIF binding. Consequently, this work has established 3D proteomics as a structural analysis tool for large multi-protein complexes.

Note: part of the work presented in this chapter has been published (Chen *et al.*, 2010)

4.2 Introduction

As presented in last chapter, the analytical workflow of 3D proteomics has been improved to handle more complex systems. In this chapter, I present the application of this analytical workflow on a 670 kDa 15-subunit multi-protein complex, the RNA polymerase II-transcription factor IIF (Pol II-TFIIF) complex.

RNA polymerase II (Pol II) is one of three eukaryotic RNA polymerases (Pol I-III); it transcribes pre-mRNA. Pol II initiated transcription begins with assembly of the preinitiation complex (PIC) on the promoter DNA which requires general transcription factors TFIIB, TFIID (containing the TATA box–binding protein, TBP), TFIIE, TFIIF and TFIIH (Reinberg *et al.*, 1998; Kornberg, 1999; Lee and Young, 2000; Orphanides and Reinberg, 2002). These general transcription factors facilitate correct positioning of Pol II on the transcription start site; after the promoter melting and transcription, they also help Pol II on promoter clearance and procession into the elongation phase (Dvir *et al.*, 2001; Woychik and Hampsey, 2002; Hahn, 2004). Studying the assembly and structure of PIC is essential for understanding the mechanisms of the transcription machinery and its regulation during the initiation process. X-ray crystallography analysis has revealed the high-resolution structure of the initiation-competent 12-subunit Pol II complex (Armache *et al.*, 2003; Bushnell and Kornberg, 2003; Armache *et al.*, 2005). However the architecture of the preinitiation complex remains under debate (Asturias, 2004; Hahn, 2004; Cramer, 2007).

Here I explored the structure of Pol II in complex with general transcription factor IIF (TFIIF). TFIIF is one of the general transcription factors that directly interact with Pol II. Among all these factors, TFIIF has the strongest affinity to Pol II (Burton *et al.*, 1988; Flores *et al.*, 1989; Bushnell *et al.*, 1996). In yeast *Saccharomyces cerevisiae*, about 50% of Pol II is associated with TFIIF (Rani *et al.*, 2004). Yeast TFIIF contains two essential subunits Tfg1 and Tfg2, which are homologues to Rap74 and Rap30 in human TFIIF; while the third yeast TFIIF subunit Tfg3 is functionally non-essential and does not have a

counterpart in mammals (Henry *et al.*, 1994). Rap74 is organized into an N-terminal region that binds Rap30 (Wang and Burton, 1995), a highly charged central domain and a C-terminal domain that is essential for full activity and binding of the CTD phosphatase Fcp1 (Chambers *et al.*, 1995; Kobor *et al.*, 2000). Rap30 contains an N-terminal region that interacts with Rap74 (Yonaha *et al.*, 1993), a control Pol II-binding region (Sopta *et al.*, 1989; McCracken and Greenblatt, 1991) and a DNA-binding C-terminal domain (Garrett *et al.*, 1992; Tan *et al.*, 1994). Structural analysis revealed that the interacting N-terminal regions of Rap74 and Rap30 form a "triple barrel" fold dimerization domain (Gaiser *et al.*, 2000) and the C-terminal domains of both Rap74 and Rap30 form Winged-helix (WH) domains (Groft *et al.*, 1998; Kamada *et al.*, 2001).

TFIIF is required for accurate transcription from promoters, with or without TATA boxes (Burton *et al.*, 1988). It is one of a minimal set of basal factors for accurate initiation of transcription by Pol II (Parvin and Sharp, 1993; Tyree *et al.*, 1993) and is involved in subsequent recruitment of TFIIE and TFIIH (Flores *et al.*, 1991; Maxon *et al.*, 1994). TFIIF function in the recognition of the transcriptional start site (Sun and Hampsey, 1995; Ghazy *et al.*, 2004; Freire-Picos *et al.*, 2005). TFIIF decreases the affinity of Pol II to DNA non-promoter sites and therefore prevents non-specific binding to DNA (Garrett *et al.*, 1992; Killeen and Greenblatt, 1992). During the initial transcription, it stimulates an early phosphodiester bond formation and the stabilization of a short RNA-DNA hybrid in the Pol II active centre (Funk *et al.*, 2002; Khaperskyy *et al.*, 2008). TFIIF also facilitates promoter escape (Yan *et al.*, 1999). During elongation, TFIIF has been shown to increase the transcript elongation rate *in vitro*. It was also reported that TFIIF can work along with TFIIS, suppressing Pol II pausing (Izban and Luse, 1992; Tan *et al.*, 1994; Zhang and Burton, 2004). Finally, TFIIF is involved in Pol II recycling through stimulation of the Pol II CTD phosphatase Fcp1 (Chambers *et al.*, 1995).

Detailed structural knowledge of how TFIIF binds to Pol II is key in understanding TFIIF functions, PIC architecture and mechanism of action. Previously, Electron Microscopy (EM) data on yeast Pol II TFIIF complex at ~18 Å resolution placed Tfg1 around the Pol II subunit complex Rpb4/Rpb7 and on the clamp (Figure 4.1), whereas Tfg2 was positioned along the Pol II active centre cleft (Chung *et al.*, 2003). However, a biochemical study using a photoreactive cross-linker, Bpa, located the TFIIF dimerization domain on the Rpb2 lobe and protrusion directly above the cleft and opposite the clamp(Chen *et al.*, 2007). Additionally, as indirect evidence for TFIIF location, the TFIIF subunit RAP 30 was detected to interact with template DNA on both sides of the TATA box and RAP74 only interacts with DNA downstream of TATA box (Kim *et al.*, 1997; Hahn, 2004). Moreover, TFIIF was reported to interact with other PIC components such as TFIIB, TFIIE, and Pol II CTD phosphatase Fcp1 (Maxon *et al.*, 1994; Tan *et al.*, 1995; Kamada *et al.*, 2003; Nguyen *et al.*, 2003; Chen *et al.*, 2007).

In this study, 3D proteomics was first applied on the 513 kDa, 12-subunit Pol II complex as an initial benchmark experiment. Agreement between cross-linking data with the crystal structure indicated that 3D proteomics was capable of analyzing such large multi-protein complexes. This approach was then applied to the structural analysis of the Pol II-TFIIF complex (670 kDa, 15 subunits) purified from yeast cells. Cross-links revealed proximity between TFIIF subunits and Pol II core, and allowed for sketching the location of TFIIF on the surface of the Pol II crystal structure which provides insights into TFIIF functions during transcription.

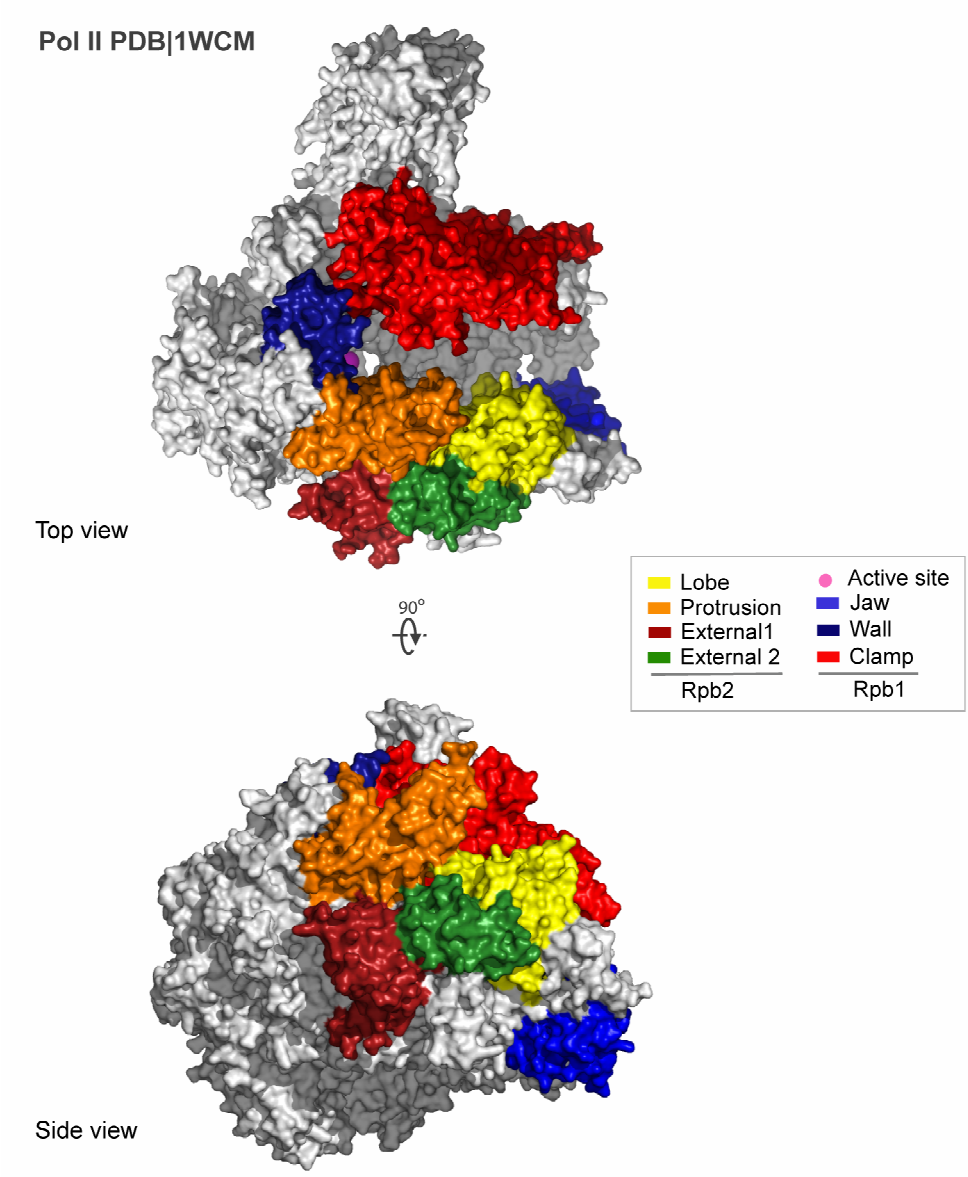


Figure 4.1 - Important domains of Pol II

Important domains of Pol II are highlighted in the Pol II structure (PDB|1WCM). The top view and the side view are shown (Armache *et al.*, 2005).

4.3 3D proteomics analysis of the Pol II complex

4.3.1 Cross-linking/MS analysis of the Pol II complex

30 μg of purified Pol II complex was cross-linked with amine-reactive cross-linker BS^3 in solution. The cross-linking products were analyzed by electrophoresis. On SDS-PAGE, bands of individual subunits in control sample were converted into a larger band at high molecular weight after cross-linking indicating that the Pol II complex was efficiently cross-linked (Figure 4.2A). On native PAGE, the cross-linked Pol II showed similar electrophoretic mobility as non-cross-linked complexes, confirming that no aggregation was introduced by chemical cross-linking (Figure 4.2B). The charge based enrichment strategy for cross-linked, tryptic peptides was performed using SCX-StageTip fractionation and precursor charge selection during mass spectrometric acquisition. In total, 429 spectra of cross-linked peptides were identified and validated manually which gave rise to 146 unique cross-linked residue pairs. Among these residue pairs, 108 links were supported by at least one high confidence match (*3.5.1*).

To evaluate the cross-linking data obtained from such a large multi-protein complex, cross-linking results were compared to the crystal structure of the yeast Pol II complex (PDB 1WCM)(Armache *et al.*, 2005). Distances between cross-linked residues in the crystal structure were compared to the cross-linking limit which is determined by the spacer length of the cross-linker, in this case BS^3. Theoretically, the distances between cross-linked residue's alpha-carbons (C-α distance) should not be greater than 24.4 Å (spacer length 11.4 Å plus two times the average length of lysine side chains 6.5Å). However considering an estimated coordinate error for mobile surface residues (1.5 Å), BS^3 should be able to bridge lysine residues up to 27.4 Å apart in the crystal structure. 80 out of 108 high confidence cross-links had both cross-linked residues present in the crystal structure. The distribution of C-α distances distinguished the observed cross-linked lysine pairs from a random selection

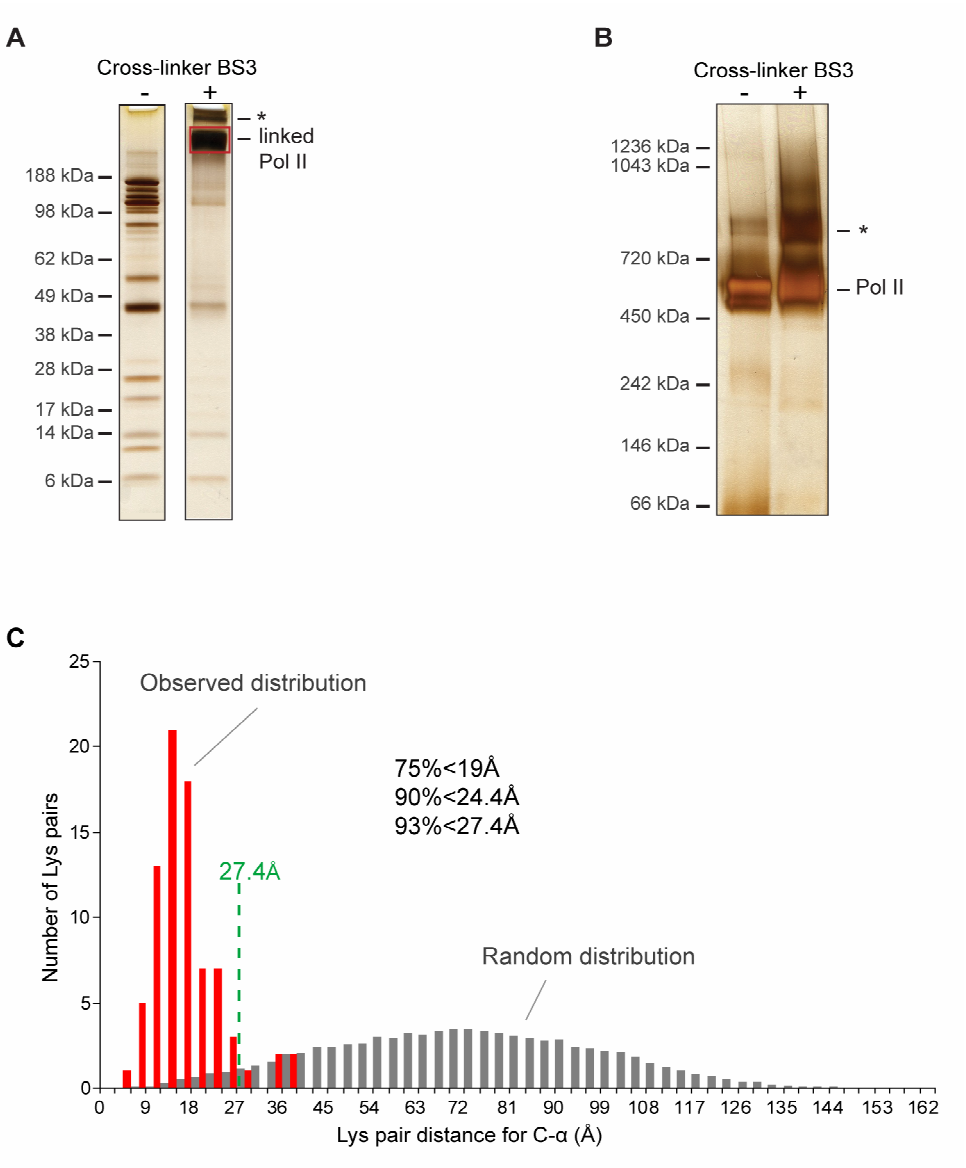


Figure 4.2 - 3D proteomics analysis of the Pol II complex

A. SDS-PAGE analysis of Pol II and BS^3 cross-linked Pol II. Cross-linked Pol II was excised from the SDS-PAGE gel and analyzed (red box). The minor cross-linking product band (asterisk) was excluded, which is likely corresponding to a Pol II dimmer or Pol II with different phosphorylation status on Rpb1 CTD. It was also observed on the native gel with and without cross-linking (asterisk)

B. Native gel electrophoresis of Pol II and BS^3 cross-linked Pol II.

C. C-α distance distribution of 80 high confidence cross-linked Lys-Lys pairs (red bars) versus distribution of random Lys-Lys pairs (grey bars) within Pol II crystal structure (PDB 1WCM). The theoretical maximum cross-link limit for BS^3 of 27.4 Å is indicated by a dashed line. The observed links beyond this limit are potentially disagreeing with the crystal structure.

Part of the data presented in this figure has been published (Chen *et al.*, 2010).

of all possible pairs in the structure (p-value of 3x10-87) and the cross-link data accurately reflected the structural features of Pol II. The C-α distances of 75 pairs fell below 27.4 Å, 72 (90%) fell below 24.4 Å, and 79 (75%) fell below 19 Å. 19 Å is predicted to be the feasible link limit of BS^3 based on computational simulation (Ye *et al.*, 2004)(Figure 4.2 B). Five cross-links bridged residues that have a C-α distance beyond the theoretical cross-linking limit (between 29.3 and 38.2Å in the crystal structure). These cross-links involved nine residues, six located in the flexible loop structure and five with a B-factor over 100, which suggested high mobility of these residues in the crystal structure. Moreover, given that the cross-linking reaction was conducted in solution, where protein molecules are more flexible than in a crystal, is may not matter that the five observed over-length cross-links conflicted with the crystal structure (Figure S2). Accepting 26 low confidence cross-links that can be displayed in the crystal structure brought in two additional over length cross-links. One of them appeared to be false since the required path of cross-linker through the Pol II molecule is not possible. Inclusion of low-confidence cross-links gave rise to a 1% error rate in the 106 cross-links that can be displayed in the crystal structure of Pol II.

The high consistency of cross-link data with crystal structure allowed me to conclude that 3D proteomics analysis is able to provide accurate residue proximities in the context of a large, multi-protein complex. All the subsequent analysis in this study was based only on high confidence cross-link data.

4.3.2 Cross-linking and protein-protein interactions

I obtained 108 high confidence cross-links from the analysis of the Pol II complex. These cross-links span the entire Pol II structure, and were detected from eleven of twelve subunits. 44 linkages between subunits connected these 11 subunits and sketched a network between them (Figure 4.3 A). The interactions between Pol II subunits in the crystal structure PDB|1WCM were defined using the Protein Interfaces, Surfaces and Assemblies service

(PISA) at the European Bioinformatics Institute (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html) (Krissinel and Henrick, 2007). Among 29 subunit interactions in the crystal structure, 16 were reflected by 42 (95.5%) of 44 inter-subunit cross-links, while two other cross-links occurred only due to the proximity between subunits. Therefore, in this case, 3D proteomics detected the interaction between twelve Pol II subunits with 89% specificity and 55% sensitivity (Figure 4.3 B). The cross-linked residues were predominantly distributed surrounding interfaces as shown in Figure 4.3 C. Multiple cross-links between subunits provide clues on the relative position of subunits in the whole complex structure. This suggests that it is feasible to study protein-protein interactions using cross-linking/MS data.

4.4 Cross-linking/MS analysis of the Pol II-TFIIF complex

4.4.1 Cross-linking/MS data of the Pol II-TFIIF complex

To investigate the interactions between Pol II and the three subunits of TFIIF, the 670 kDa Pol II-TFIIF complex was cross-linked and analyzed by mass spectrometry. Denaturing gel electrophoresis analysis of the cross-linked complex showed efficient cross-linking of all subunits from both Pol II and TFIIF (Figure 4.4 A). Native PAGE indicated a predominantly homogenous product of approximately 700 kDa with no oligomer formation (Figure 4.4 B). Using 200 micrograms of purified complex allowed for more comprehensive analysis with elaborate fractionation than was the case for the Pol II analysis. From this material we identified and validated 1891 cross-linked peptide spectra, and obtained 413 unique high-confidence cross-links from all 15 Pol II-TFIIF subunits: 189 within the Pol II core complex, 133 within TFIIF and 91 between Pol II and TFIIF (Table S3).

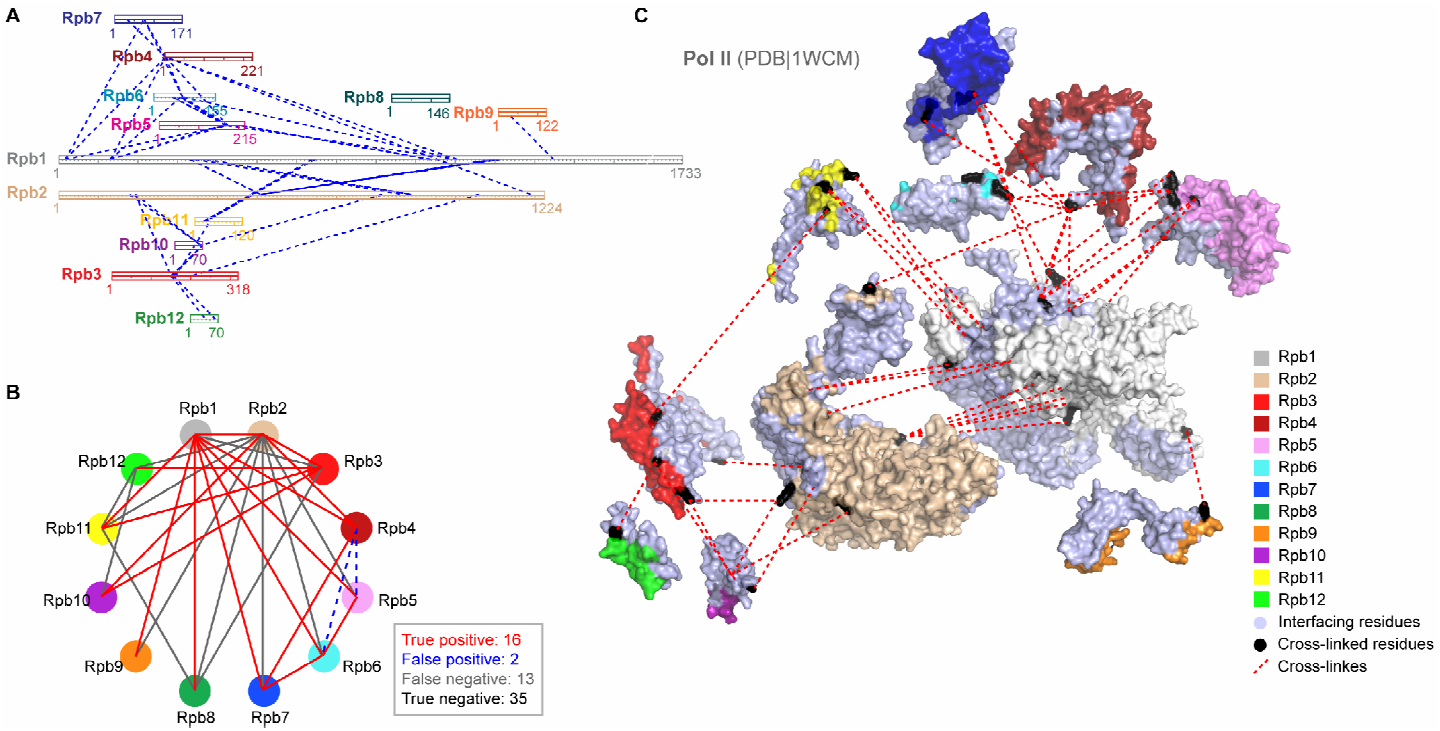


Figure 4.3 - 3D proteomics analysis reveals predominantly direct pairwise interaction between Pol II subunits.

(Poll II subunits in these figures are highlighted in the canonical colours)

A. The cross-linkages (blue dashed lines) between Pol II subunits (coloured bars) exhibited a network between subunits.

B. Cross-linkages detected 16 of 29 interactions between Pol II subunits (coloured spheres) established in the crystal structure (PDB|1WCM) (the interactions were indicated as stroke, with cross-linking analysis depicted in red and the rest in grey). While two pairs of subunits that were cross-linked, they were actually not interacting in the crystal structure.

C. Cross-linkages (red dashed lines) observed between Pol II subunits are displayed in a disassociated Pol II structure (PDB|1WCM) shown as surface model. Interfaces between subunits are indicated in faint slate, cross-linked residues are in black.

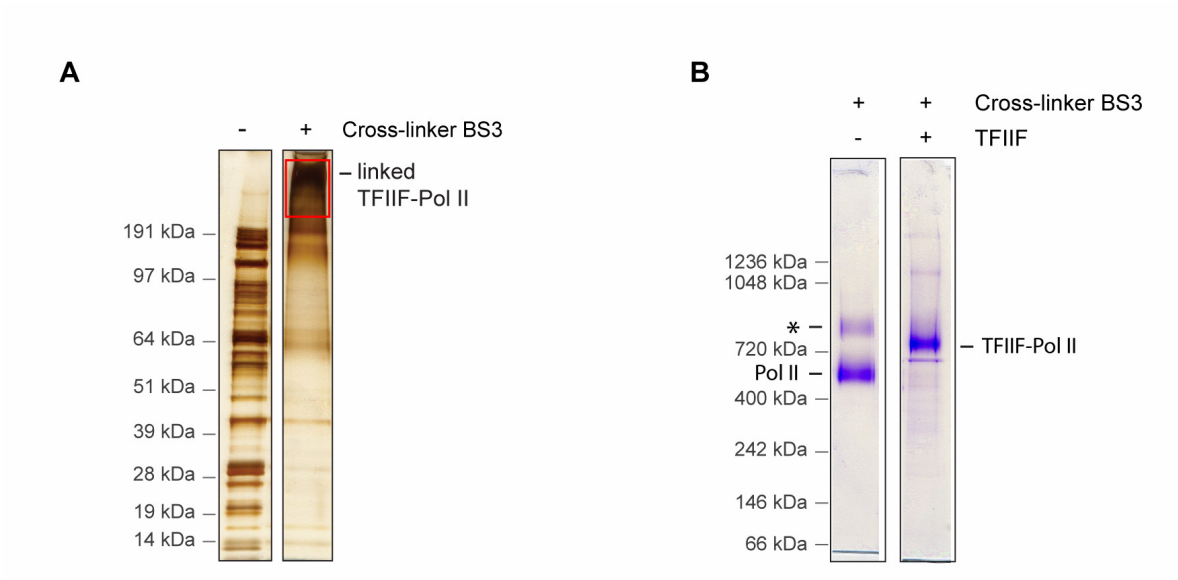


Figure 4.4 - Cross-linking reaction of Pol II –TFFIIF complex

A. SDS-PAGE analysis of Pol II-TFIIF complex and BS^3 cross-linked Pol II-TFIIF complex. Cross-linked Pol II-TFIIF complex was excised from the SDS-PAGE gel and analyzed (red box).

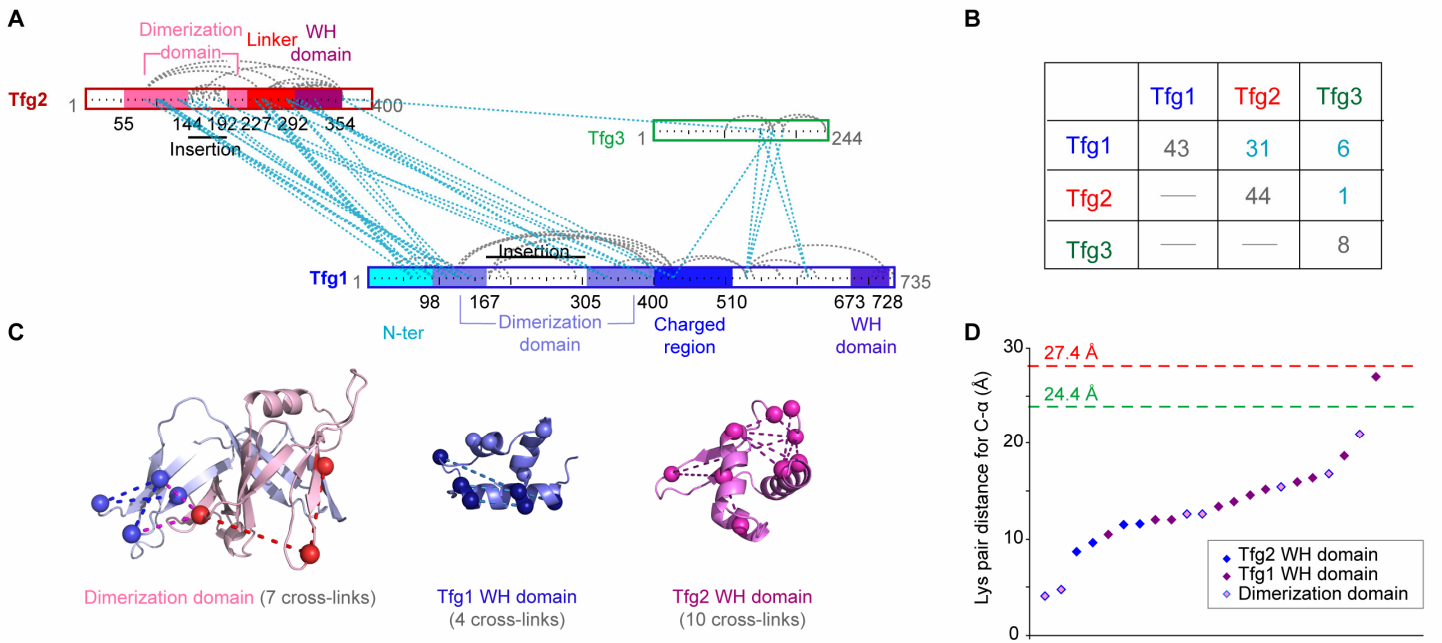
B. Native gel electrophoresis of BS3 cross-linked Pol II-TFIIF complex with BS3 cross-linked Pol II complex for comparison. The native gel shows absence of a dimeric complex for BS3 cross-linked Pol II-TFIIF complex. The band with indicated with asterisk is likely corresponds to Pol II with CTD in different phosphorylation state or dimer of the Pol II complex.

The data presented in this figure has been published (Chen *et al.*, 2010)

4.4.2 Yeast TFIIF domain structures

Cross-links observed in TFIIF provided insight into the structure of yeast TFIIF. 133 unique cross-links were detected from all 3 subunits - 95 within subunits and 38 between subunits (Figure 4.5 A). The 38 inter-protein cross-links revealed interactions between subunits in TFIIF. Tfg2 was extensively cross-linked to Tfg1 towards its N-terminus; the dimerization region and the neighbouring linker region in Tfg2 was cross-linked to the dimerization region, N-terminal segment and the charged region of Tfg1. The extensive cross-linking in the Tfg1-Tfg2 dimerization domain suggested that yeast Tfg1 and Tfg2 have similar interaction as their human homologs. There were relatively few cross-links involving Tfg3. Six linkages were detected from the C-terminal region of Tfg3 to the charged region of Tfg1 and the segment between the charged region and the C-terminal Winged Helix (WH) domain. A single cross-link was detected from the Tfg3 C-terminal region to the WH domain of Tfg2.

Homology models of the yeast Tfg1-Tfg2 dimerization domain, the Tfg1 WH domain and the Tfg2 WH domain were built by Patrick Cramer group (as described in (Chen *et al.*, 2010)). 22 cross-links were displayed in these homology models (seven in the dimerization domain, dour in the Tfg1 WH domain and eleven in the Tfg2 WH domain).The cross-link data matched well to the homology models on residue proximity: The C-α distance between cross-linked residues in the homology models for 21 of total 22 cross-links were shorter than 24.4Å, and all 22 were below 27.4 Å (Figure 4.5 B). Therefore, the homology models built based on sequence alignment were validated by the experimental data. Furthermore, the structural conservation between the yeast TFIIF and its human homolog as shown in the dimerization domain and Tfg1 and Tfg2 WH domains, suggests the functional similarity between them.



	Tfg1	Tfg2	Tfg3
Tfg1	43	31	6
Tfg2	—	44	1
Tfg3	—	—	8

Figure 4.5 - Cross-links observed within TFIIF and structures of TFIIF domains.

A. Cross-links between TFIIF subunits (blue) and within TFIIF subunits (grey) displayed in the schematic representation of TFIIF subunits and domains.

B. Summary of cross-links observed within FIIF. Number of cross-links within subunits and between subunits is listed.

C. Observed cross-links are displayed in the TFIIF homology models of the Tfg1 winged helix domain, the Tfg2 winged helix domain, and the dimerization domain of Tfg1 (light blue) and Tfg2 (pink). Cross-links: dashed lines; C-α of linked Lysine residues: spheres.

D. Cross-linking data confirmed domain modelling of yeast sequences onto the human homolog crystal structures. 21 of 22 cross-links shown in C are within cross-linker determined linking limit (24.4Å), and all 22 fit in cross-linking limit when consider the mobility of atoms in structures.

Part of the data presented in this figure has been published (Chen *et al.*, 2010)

4.4.3 Location of TFIIF on Pol II

Interactions between TFIIF and Pol II were revealed by 91 cross-links between them. These cross-links involved all three subunits of TFIIF and five subunits of Pol II whereas the majority of linkages were observed between Tfg1 and Tfg2 and the two largest Pol II subunits Rpb1 and Rpb2 (Figure 4.6). There were some Pol II residues that were cross-linked to more than one TFIIF domains, which implied spatial proximity between these TFIIF domains. This implication is supported by cross-links observed between TFIIF domains. Occasionally, linkages were observed to form closed networks of the type A-B-C-A. For example, Tfg1 K394 in the dimerization domain was cross-linked to Tfg1 K426 in the charged region, while they were both cross-linked to Rpb2 K228. Overall the cross-linkages between TFIIF and Pol II sketched the location of domains in TFIIF subunits on the surface of the Pol II structure (Figure 4.7).

Tfg1 consists of 735 residues. The N-terminal region was cross-linked to Rpb2 at the External1 domain (Figure 4.1), and the following dimerization domain was cross-linked to the Pol II lobe domain. The dimerization domain in Tfg1 included two parts of sequence with an insertion in between. Only the C-terminal to the insertion was observed cross-linked to Pol II. The following charged region was cross-linked the Rpb1 jaw at the downstream end of the cleft. Downstream to the charged region, the cross-links between Tfg1 and Pol II could be detected until residue K537 which was cross-linked to Rpb2 region close to where the dimerization domain was located, near the top of the cleft. Finally Tfg1 K614 was cross-linked to Rpb1 K49 on the Pol II Clamp, proximal to the Rpb4/7 sub-complex and the attachment point of the linker to the CTD of Rpb1 (Spahr *et al.*, 2009), which is consistent with the Tfg1 density observed on the Clamp as previously observed by electro microscopy (Chung *et al.*, 2003).

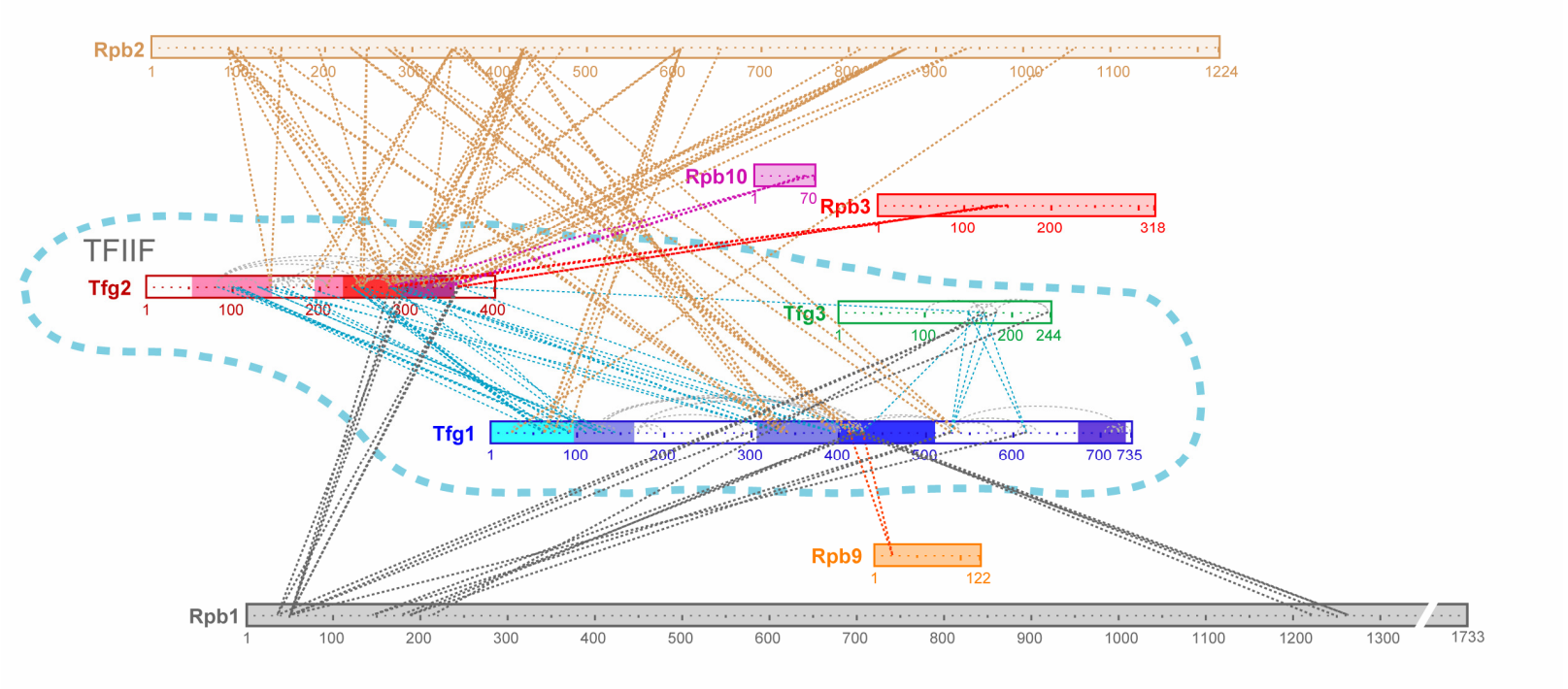


Figure 4.6 - Cross-links between Pol II and TFIIF

The TFIIF subunits and involved Pol II subunits (coded in canonical colours) are shown in bars while the observed cross-linkages from TFIIF to Pol II subunits are displayed in dashed lines and colour coded by the respective Pol II subunit. Links between TFIIF subunits (blue) and within TFIIF subunits (grey). The colour coding of domains in TFIIF followed Figure 4.4. The data presented in this figure has been published (Chen *et al.*, 2010)

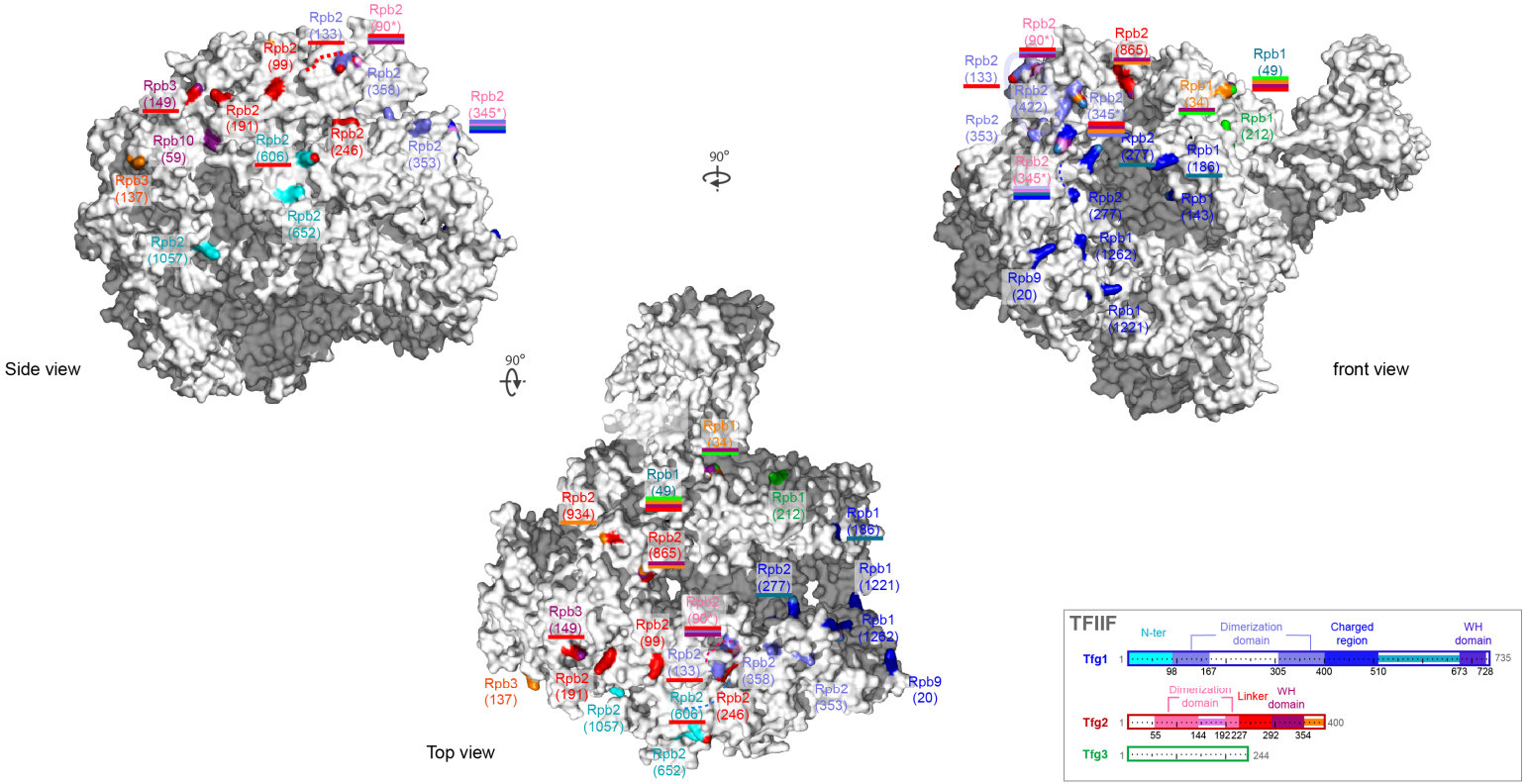


Figure 4.7 - Cross-linking footprints of TFIIF subunits on the surface of Pol II structure.

High confidence TFIIF cross-linking sites on Pol II surface (PDB|1WCM) are coloured and labelled according to the cross-linked TFIIF domains/regions. (For residues that cross-linked to multi TFIIF domains/regions, the coloured strokes are displayed additional to the coloured labels). Three views (side, top and front) are used. The colour code of TFIIF domains/regions are shown at the lower right corner.

The data presented in this figure has been published (Chen *et al.*, 2010)

Furthermore, another general transcription factor, TFIIE, which was located on the clamp, was reported to bind Rap74 *in vitro* (Maxon *et al.*, 1994; Chen *et al.*, 2007). There were no cross-links obtained between the Tfg1 WH domain and Pol II, however the cross-links observed within the Tfg1 WH domain indicated that this domain is cross-linked and detectable by our analysis. The absence of cross-link might be due to the lack of stable contact due to the high mobility of the Tfg1 WH domain. This hypothesis was supported by a later study by Madler *et al.* (Madler *et al.*, 2010). The failure of an amine-reactive cross-linker DSS (with NHS ester as its functional group) to specifically cross-link low affinity proteins (with K_d >> 25 μM) suggested the existence of a lower stability threshold for structures to be captured by cross-linking

C-terminal to the Tfg1-Tfg2 dimerization domain that was located on Pol II lobe, the Tfg2 domains could be unambiguously positioned along the Pol II protrusion across the Rpb2 side of the molecule according to the cross-links detected between Tfg2 and Pol II. The WH domain ended up close to the path of upstream DNA. However, the C-terminal region of Tfg2 was not restricted to only this location. Some additional cross-links were also detected from the Tfg2 linker region, WH domain and the C-terminal segment to the Pol II Wall and Clamp. The Pol II residues involved in these linkages were cross-linked to at least two domains in the Tfg2 C-terminal region. Hence, these additional cross-links revealed a dynamic binding patch of Tfg2 C-terminal region on Pol II Wall and Clamp (Figure 4.8) which is agrees with the additional density detected by EM at this region from the TFIIF bound Pol II (Chung *et al.*, 2003). This alternative binding pattern on the Pol II Wall and Clamp might be caused by absence of DNA or other transcription factors and it does not agree with in the preinitiation complex structure because it would block the path of DNA (Kostrewa *et al.*, 2009). The functionally non-essential subunit of TFIIF, Tfg3, showed few connections to Pol II with only 3 cross-linkages observed. However they still indicated the specific contact between Tfg3 and Pol II.

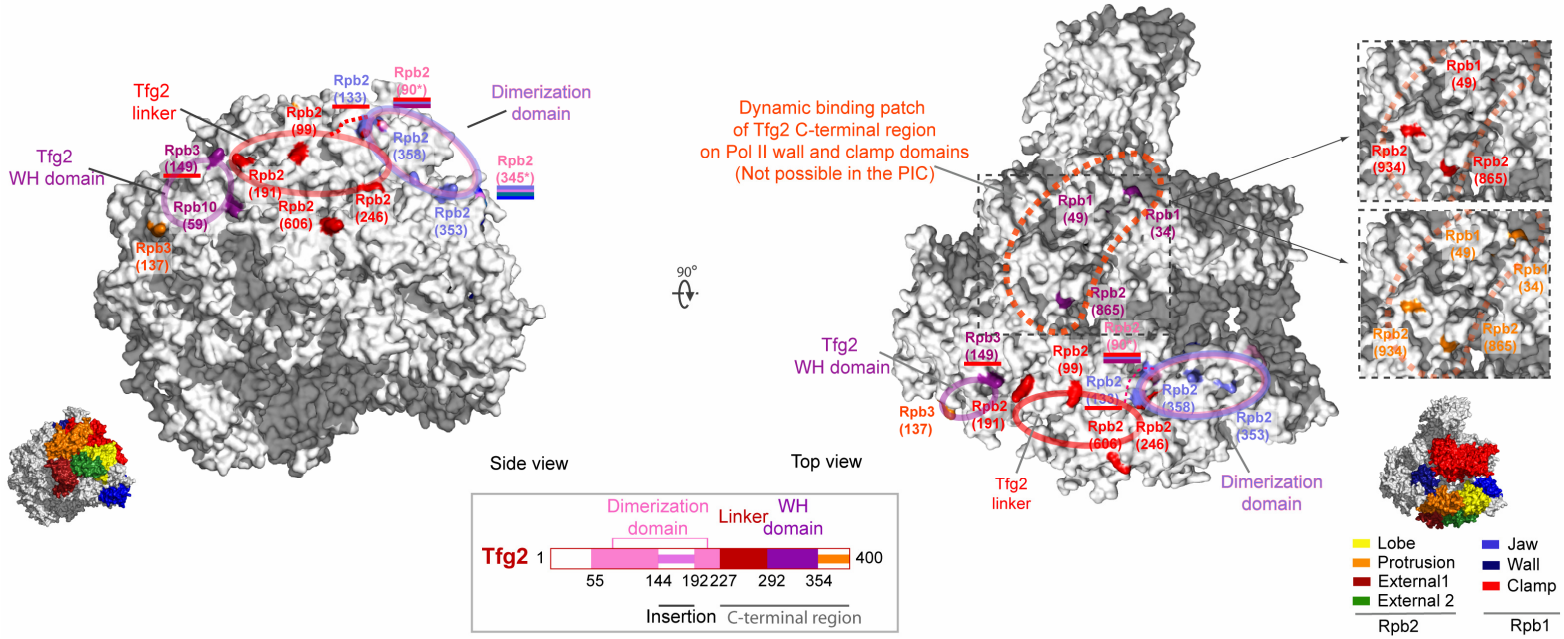


Figure 4.8 - Alternative position of Tfg2 C-terminal region (linker, WH domain and C-terminal) on the Pol II surface.

The residues on Pol II surface (PDB|1WCM) that cross-linked to Tfg2 are coloured and labelled according to the Tfg2 domains/regions they linked to (the colour code of Tfg2 domains/regions shown on the bottom). The top view and side view are displayed. As a reference, important Pol II domains are highlighted in the canonical colours in the smaller views. The unambiguous binding sites of the Tfg2 domains/regions suggest that Tfg2 extends from the dimerization domain on the lobe along the protrusion. While the Tfg2 WH domain is also cross–linked to the wall and clamp (Rpb2 K865, Rpb1 K49 and Rpb1 K34) (only visible in the top view) and shares these linkage sites with the linker and C-terminal Tfg2 domains (in inset up and inset down). This cross-linking footprint pattern of Tfg2 suggests that besides a well-defined binding position over the entire protrusion, the C-terminal region of Tfg2 has an alternative and dynamic binding patch on the Pol II wall and clamp. However this binding patch is not possible in the preinitiation complex, because it collides with predict DNA path and other general transcription factors. The data presented in this figure has been published (Chen *et al.*, 2010)

These three cross-links were detected from the C-terminal of Tfg3 to the Pol II Clamp close to the Rpb4/Rpb7 subunits. The shared linkage sites on Pol II (Rpb1 K49) and the number of cross-links between Tfg3 C-terminal and Tfg1C-terminal region indicate close proximity and suggested a possible involvement of Tfg3 in the function of the Tfg1 C-terminal region. A single cross-link between Tfg3 and Tfg2 might occur only due to the dynamic binding of Tfg2 C-terminal region on the Pol II Clamp.

In summary, the cross-linking footprints of TFIIF subunits on the surface of the Pol II structure illustrate the location of TFIIF in complex with Pol II. Except for the dynamic binding patch of Tfg2 WH domain, the location of TFIIF subunits fit the known structure of the Pol II initiation complex (Figure 4.9), and this allows for a better understanding of the TFIIF function during transcription.

4.4.4 Possible conformational changes of Pol II in the Pol II-TFIIF complex

Comprehensive analysis with 200 μg of material resulted in identification of 189 high confidence unique cross-links from Pol II core, within and between all twelve subunits. Comparing to the 108 linkages observed from the 30 μg Pol II sample, 58 cross-links were common in both analyses while 50 were unique to the Pol II sample and 131 were unique to the Pol II-TFIIF sample. Interestingly, although there were a large number of distinct cross-links between the two samples, the network between subunits established by cross-links was not significantly different between Pol II alone and PoI II in complex with TFIIF (Figure 4.10 A). This observation suggests that major parts of the PoI II architecture remain the same after TFIIF binding. However the cross-link data also showed some differences between Pol II and Pol II-TFIIF (Figure 4.10 B). The unique linkages in Pol II-TFIIF sample included 18 over-length linkages with respect to the cross-linking limit of BS^3 (28.3Å ~50.3 Å, in average 37.7 Å). 15 of these linkages were cross-linked to the Clamp domain (Figure 4.10 B).

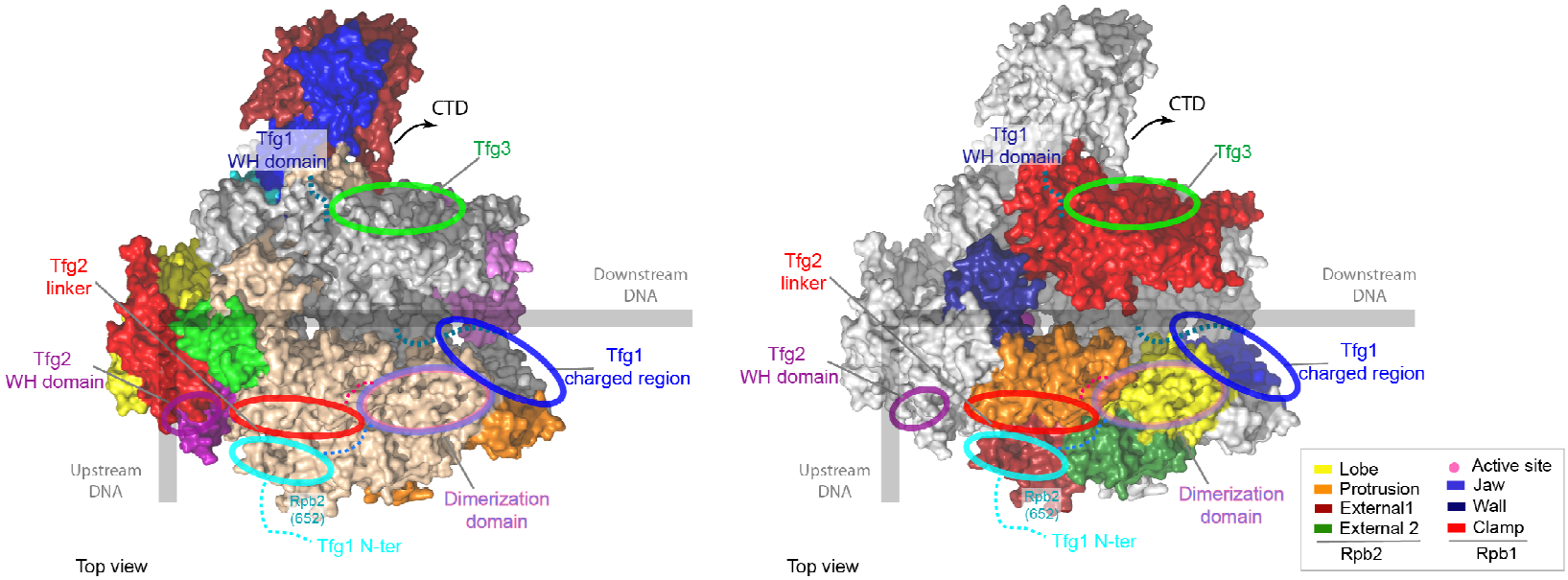


Figure 4.9 - Architecture of Pol II-TFIIF in preinitiation complex.

The TFIIF regions are indicated on Pol II structure (PDB|1WCM) in the top view, the path of the DNA in the preinitiation complex with a closed promoter conformation is indicated as a thick grey line (Kostrewa *et al.*, 2009). Pol II subunits (left) and important domains (right) are highlighted in canonical colours for reference. The point of attachment of the linker to the Rpb1 CTD predicted from *Schizosaccharomyces pombe* Pol II structure is shown as an arrow (Spahr *et al.*, 2009).

The data presented in this figure has been published (Chen *et al.*, 2010)

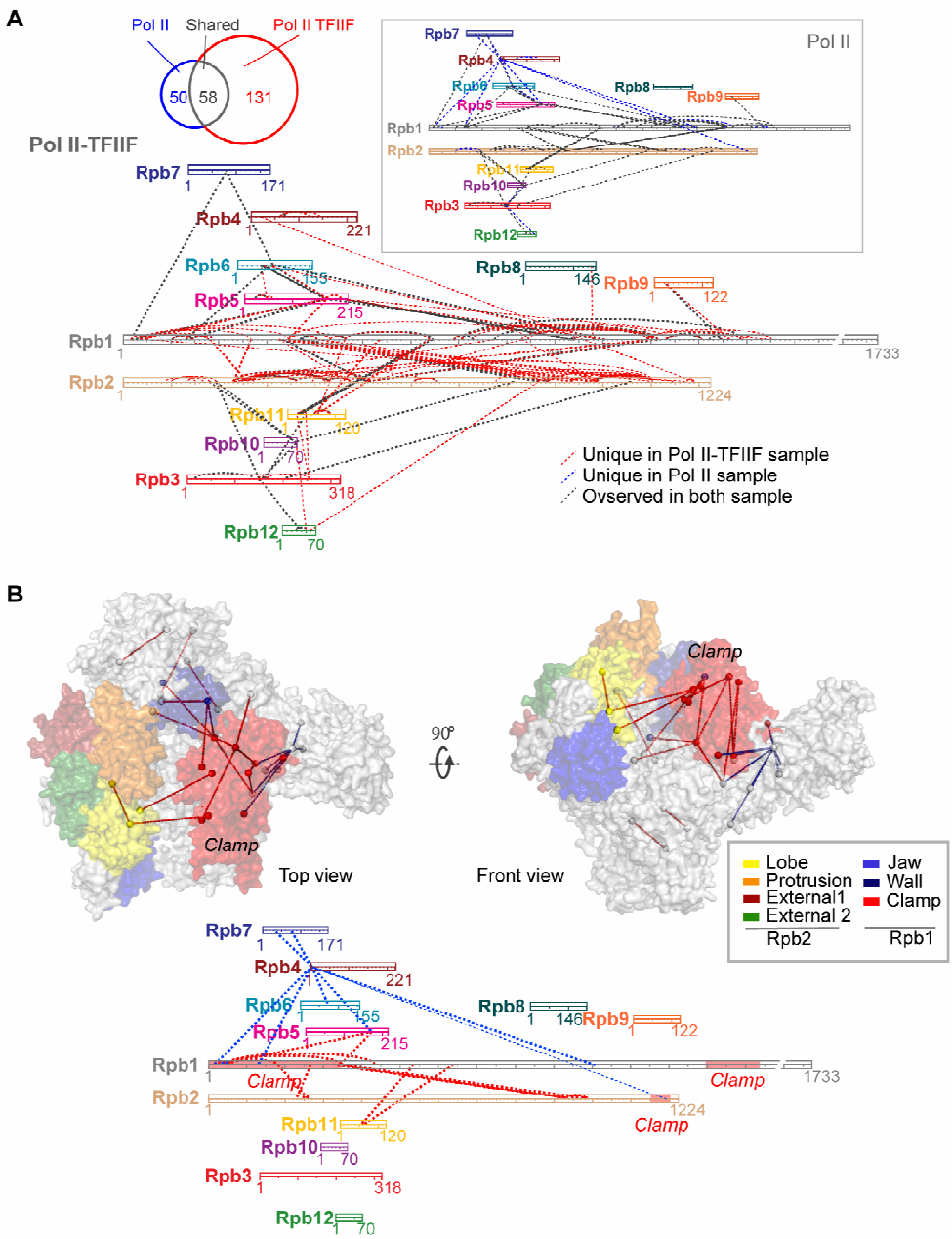


Figure 4.10 - Cross-links within Pol II observed in Pol II-TFIIF complex.

A. Cross-linkages observed within Pol II from the Pol II-TFIIF sample. The cross-linkages observed from Pol II sample are shown in the inset. (Pol II subunits: coloured bar; cross-linkages: dashed lines, Pol II-TFIIF sample unique in red, Pol II sample unique in blue, common ones in gray. Venn diagram shows number of cross-links in each group).

B. Cross-linkages that suggest potential difference between Pol II alone and Pol II-TFIIF samples on Pol II conformation are displayed in crystal structure (PDB|1WCM) (up) and in primary sequences (down) of Pol II. The Pol II structure shown in with semi-transparent surface and the domains highlighted in canonical colours (top and front); Pol II subunits indicated in coloured bars, the sequences involved in the Clamp are highlighted in red ; cross-linkages: dashed lined, Pol II unique in blue and Pol II- TFIIF unique in red).

To satisfy these cross-links, a movement of the Clamp towards the Rpb2 side of the cleft was required, which suggested that binding of TFIIF induced a more compact conformation around the cleft comparing to what was shown in the Pol II crystal structure PDB|1WCM (Armache *et al.*, 2003). This conformational change might have also caused the absence of eight cross-links from the N-terminal of Rpb4 to the surrounding subunit surfaces (Rpb1, Rpb2, Rpb5, Rpb6, Rpb7) in Pol II-TFIIF since the Rpb4/7 heterodimer is protruded from the base of the Clamp and the Rpb4 N-terminal region is in close proximity to the Clamp. On the contrary, the other 42 unique cross-links observed in the Pol II sample can be clustered with common or Pol II-TFIIF unique cross-links in terms of their locations in Pol II structure.

4.5 Discussion

4.5.1 Architecture of the Pol II-TFIIF complex and TFIIF functions

Knowing the location of transcription factors on Pol II structure advanced the knowledge of the three dimensional structure of the Pol II preinitiation complex, which is fundamental for understanding the mechanism of transcription initiation. This work positioned TFIIF subunits on the Pol II surface in the unit of domains, which allowed for further understanding of the TFIIF functions during transcription initiation.

TFIIF binding was reported to prevent Pol II from non-specific DNA binding (Conaway and Conaway, 1990; Killeen and Greenblatt, 1992). The majority of the Pol II molecule is negatively charged and only the deep hidden active centre, the active cleft, the wall and vicinity have a positively charged surface, which enables the non-specific binding to the negatively charged DNA backbones. This non-specific binding can be suppressed by either stabilizing the close state of the clamp or temporarily occupying the exposed positively charged Pol II surface. The cross-links observed within Pol in the Pol II-TFIIF sample suggested a more close up state for the clamp as compared to the observed structure

in the free Pol II complex. The dynamic binding patch of the Tfg2 C-terminal region that was located on the Wall and Clamp can partially cover the cleft and interrupt DNA binding. Moreover the cross-links between the downstream cleft and the Tfg1 charged region placed this predominantly negatively charged region on the jaw which could repel the DNA and it has been shown that the binding of TFIIF but not Rap30 alone can release the Pol II from non-specific DNA association (Killeen and Greenblatt, 1992). The bacterial initiation factor σ70 N-terminal region 1.1 with a conserved negatively charged surface is also consistently located in the downstream cleft (Murakami *et al.*, 2002).

TFIIF is required for the recruitment of Pol II to the promoter DNA, and it functions at least partly through the binding to the promoter DNA upstream of the start site via the C-terminal of the Tfg2 WH domain (Flores *et al.*, 1991). This domain was implicated for DNA binding, and the Tfg2 homolog Rap30 alone was sufficient for recruitment and assembly of the preinitiation complex at promoter DNA. Moreover, presence of whole TFIIF factor was shown to stabilize the formed complex (Tan *et al.*, 1994; Kamada *et al.*, 2001). As shown by this work, the Tfg2 WH domain was in fact positioned on the Pol II protrusion where the upstream DNA passes. The dynamic binding of Tfg2 C-terminal region on Pol II suggested the flexibility of the WH domain that permits accommodation of different promoters. On the other hand, the binding to the promoter DNA might also stabilize the localization of the Tfg2 WH domain on the Protrusion.

TFIIF is also reported to play a role in transcription start site utilization (Ghazy *et al.*, 2004; Freire-Picos *et al.*, 2005). The special constraints carried by cross-links between the TFIIF dimerization and Pol II structure enabled me and my collaborators to position and orient the yeast TFIIF dimerization domain homology model on the surface of the Pol II structure at the lobe (Chen *et al.*, 2010). In combination with previous reported evidence revealed by photoreactive amino acid cross-linking study (Chen *et al.*, 2007), this domain was assumed to act as a "lid" with two overlapping positions. During initial transcription it

slides into the cleft with the template DNA passing through underneath (Chen *et al.*, 2010). This closed dimerization domain is likely to stabilize an open promoter complex (Kostrewa *et al.*, 2009) and contribute to the correct transcription start site setting. The amino acid mutations in yeast Tfg1-Tfg2 dimerization domain, which was defined according to sequence alignment with the human homolog at residues Tfg1 E346A, W350A and tfg2 L59K caused upstream shift of initiation and decreased TFIIF binding affinity to Pol II (Ghazy *et al.*, 2004; Khaperskyy *et al.*, 2008), and the mutation of the nearby dimerization domain residue G363 suppressed the downstream shift of start site induced by mutations in Rpb1 and TFIIB (Freire-Picos *et al.*, 2005). The mutations in the Pol II lobe destabilized the TFIIF binding and caused a similar defect on start site selection (Chen *et al.*, 2007). Moreover, the deletion of the Pol II subunit Rpb9 also conferred an upstream shift of the transcription start site and this effect was assumed to be associated with the impaired TFIIF-Pol II interaction (Ziegler *et al.*, 2003).

The Tfg1 C-terminal WH domain was reported to interact with the CTD phosphatase Fcp1 (Chambers *et al.*, 1995; Kobor *et al.*, 2000), this interaction is possibly associated with the functional stimulation of Fcp1 (Chambers *et al.*, 1995). There was no cross-link detected between the Tfg1 WH domain and Pol II, however the cross-link to the nearest residue to this domain pointed the C-terminal of Tfg1 to the Clamp and close to the CTD linker revealed in *S. pombe* Pol II structure (Spahr *et al.*, 2009). The Tfg3 subunit of TFIIF was also positioned in the same location on the Pol II surface. Additionally, the cross-links between Tfg1 and Tfg3 indicated connections between C-terminals of these two subunits. Therefore, the Tfg1 WH domain, Tfg3, and Pol II CTD are likely to be in close proximity, and might be related to the function of the Tfg1 WH domain. Even though the high order sequence repeat of "YSPTSPS" in the CTD protected it from trypsin digestion and therefore detection in our mass spectrometric analysis, the structural disorder of the Pol II CTD

domain (Spahr *et al.*, 2009) is consistent with the high mobility of the Tfg1 WH domain in respect to the core Pol II in the crystal structure.

4.5.2 Study architectures of large multi-protein complexes using 3D proteomics

Higher-order multi-protein complexes are often not amenable to high resolution structure determinations such as X-ray crystallography or NMR due to their size, stability and limit on sample amount, homogeneity, *etc*. The inherent advantages of 3D proteomics (Chapter 1) determined its great potential on analyzing large and fragile multi-protein complexes. In this study, I demonstrated the application of 3D proteomics on the 513 kDa 12-subunit Pol II complex and the 670 kDa 15-subunit Pol II-TFIIF complex. The results extended the structural understanding of the Pol II complex from the crystallized 12-subunit core to a 15-subunit complex with peripheral transcription factor IIF.

Firstly, the advanced analytical workflow of 3D proteomics allowed for efficient detection and high confidence identification of cross-linked peptides from these large multi-protein complexes. In this study, 2320 high resolution MS^2 spectra of cross-linked peptides were identified and validated with high confidence. These cross-linked peptides derived from combinations of 421 unique peptide sequences, ranged from 3 to 38 amino acids. Regardless the structural information carried by the cross-links, this dataset *per se* is the biggest pool of cross-linked peptide spectra, and it was obtained from native multi-protein complexes. It provided a valuable resource for informatics applications conducted by other members of the Rappsilber group, such as statistical studies on cross-linked peptide spectrum features (Lutz Fischer) and development of our search algorithm and scoring system (Salman Tahir and Lutz Fischer).

Consistency between the crystal structure and cross-link data not only proved that 3D proteomics is capable of analyzing large multi-protein complexes, but also supported the

strategy to study protein-protein interactions between subunits based on cross-links detected between them.

Furthermore, making use of the crystal structure of the Pol II core complex, 3D proteomics data revealed binding position of TFIIF in a 3D fashion on the surface of the Pol II structure. The homology model of the TFIIF dimerization domain, which was validated by cross-link data was docked on the surface of the crystal structure of Pol II core complex based on distance constraints carried by a series of cross-links between them (Chen *et al.*, 2010). This resulted in a high resolution model of the Pol II-TFIIF complex and suggests a promising role for 3D proteomics in structural modelling of large high-order protein complexes. The potential contributions of 3D proteomics in structural biology were more comprehensively discussed by Juri Rappsilber (Rappsilber, 2011).

Finally, 3D proteomics analysis of Pol II-TFIIF also reflected the dynamic aspects of Pol II-TFIIF interactions and the possible conformational change of Pol II due to TFIIF binding. However, in order to draw conclusions about these conformational changes, and direct further study on the functional meaning of them, a proper conformational comparison between Pol II structures with and without TFIIF binding is required. An approach that is able to perform this overall conformational comparison for such large and delicate complexes is yet to be developed at the time when these analyses were conducted. Consequently I chose an adequate model system and tested the possibility of following conformational changes by 3D proteomics. This is the content of next chapter.

Chapter 5

QUANTITATIVE 3D PROTEOMICS DETECTED CONFORMATIONAL DIFFERENCES BETWEEN C3 AND C3B IN SOLUTION AND GAVE INSIGHT TO THE CONFORMATION OF SPONTANEOUSLY HYDROLYZED C3

5.1 Summary

As part of exerting their functions proteins or protein complexes frequently change their conformations. Here I present quantitative 3D proteomics as a tool to quantitatively measure the differences between protein conformations. I applied this approach to detect in solution the conformational differences between the key complement system component C3 and its active form C3b. Isotope labelled cross-linkers introduced a mass difference to cross-linking products from different conformations. The identified and quantified cross-links revealed the structural differences and similarities between C3 and C3b, which confirmed previous observations made by X-ray crystallography. Additionally, the spontaneous hydrolysis C3 analogue C3(H_2O), was detected from both C3 and C3b samples. Based on the clustered cross-links and the crystal structures of C3 and C3b, the conformation of C3(H_2O) presents certain similarities to both C3 and C3b, however also has unique structural features. The C3b-like conformation in C3(H_2O) may explain the functional similarity between C3b and C3(H_2O). Cross-link data also provided additional dynamics to the static crystal structures and contradicted a false C3b crystal structure. In conclusion, in this study I demonstrated that quantitative 3D proteomics is a valuable tool for conformational analysis of proteins and protein complexes.

5.2 Introduction

The function of protein requires protein conformational changes. These conformational changes can be induced by changes in environment such as alterations of temperature, pH and salt concentration, affects of biomolecular interactions such as residue modifications or binding of other molecules. The transformation between conformations can significantly affect a protein's binding ability and affinity to different biomolecules such as substrates, interaction partners, receptors or ligands and is often crucial for signal transduction, protein activation and regulation, and assembly of protein complexes and other macromolecules. Hence studying protein conformational changes is of great interest when wanting to understand the function of a protein or protein complex. High resolution structural analysis techniques including X-ray crystallography, nuclear magnetic resonance (NMR) and Electron microscopy (EM) can resolve detailed differences between protein conformations (Ishima and Torchia, 2000). However the applications are often restricted by the size of the proteins and protein complexes, the amount or homogeneity of the protein samples. The solution scattering using both X-ray and neutrons (Perkins and Sim, 1986), can reveal differences in size and shape at medium resolution. The spectroscopic technique circular dichroism (Isenman, 1983) provides information about conformational rearrangements to secondary structure; and Raman scattering may detect conformational transitions of chromophoric groups (Stryer, 1981), Other optical techniques like fluorescence polarization (FP) that detects the molecule volume changes (Stryer, 1981), and dual polarization interferometry (DPI) that measures the thickness and density of an immobilized protein layer (Swann *et al.*, 2004), also reflect protein conformational differences. Fluorescence resonance energy transfer (FRET), luminescence resonance energy transfer (LRET) and electron transfer (ET) can detect the conformational change induced shift in distance between the two labels (a donor and acceptor), requiring insertion of these labels in the protein molecule at the desired locations (Heyduk, 2002; Yang *et al.*, 2003). In addition,

mass spectrometry based hydrogen-deuterium exchange methodology (Englander *et al.*, 2003; Winters *et al.*, 2005) and chemical and immune probing (Isenman, 1983; Hack *et al.*, 1988) have also been reported to reveal protein conformational differences by detecting surface accessibility changes. These techniques managed to directly/indirectly detect protein conformational differences, however with limits on both technique and applications (Salafsky, 2006). Unfortunately, none of these techniques are suitable for studying large and fragile multi-protein complexes such as Pol II-TFIIF.

Here I develop quantitative 3D proteomics by introducing stable isotope labelling based quantitation to the cross-linking protocol. Previously, 3D proteomics has been used to reflect conformational changes of single proteins through observation of certain cross-links in one conformation or the other (Bhat *et al.*, 2005) However, in complex systems, it is not rigorous to draw conclusions just based on the appearance or absence of certain cross-linked peptides due to variation between individual mass spectrometry analyses. For example, as presented in Chapter 4, 80% of unique Pol II cross-links observed from the 30 μg Pol II sample can be clustered together with unique Pol II cross-links obtained from the Pol II-TFIIF sample in terms of their distribution in Pol II structure and did not reflect conformational differences. Therefore, the isotope labelling was introduced to achieve the quantitative comparison of different conformations. Mass spectrometry is able to accurately read out the signal ratios for isotopic analogues due to their identical chemistry when analyzed in the same experiment. The use of isotope labelling for quantitative proteomics at peptide level is well established for proteins and their modifications (Gygi *et al.*, 1999; Ong *et al.*, 2002; Mann, 2006). This study extends the use of isotope labelling for quantitation to cross-links by using isotope labelled cross-linkers. The isotope labelled cross-linker and its unlabelled analogue have identical cross-linking reactivity. The cross-linking products of differently labelled cross-linkers will only be resolved in mass spectrometry analysis by mass. In quantitative 3D proteomics analysis, two different conformations of a protein are

cross-linked with differently labelled cross-linkers (often labelled and non-labelled cross-links) separately and then equal amounts are mixed before mass spectrometric analysis. When a particular cross-link is equally possible in both conformations, the cross-linked peptides will generate a signature-like doublet signal in mass spectra. The ratio between the heavy and light signals in doublet signals will quantitatively reflect the overall difference between the two conformations projected on cross-links. Cross-links that are only possible in either conformation will be detected as singlet signals in MS. As an example to test the validity of the this approach, the deuterated amine-reactive cross-linker Bis[Sulfosuccinimidyl] 2,2,7,7-suberate-d4 (BS^3-d4, Thermo Scientific) and its 4 Da lighter unlabelled analogue Bis[Sulfosuccinimidyl] suberate-d0 (BS^3-d0, Thermo Scientific) (Figure 1.6) were applied to quantitatively detect, in solution, the conformational differences between the complement component 3 (C3) and its active form C3b.

Complement component 3 (C3) is a pivotal protein in the complement system. The complement system has a major role in mammalian innate and adaptive immunity, defending the host against bacterial infection, disposing cell debris, bridging innate and adaptive immunity and also enhancing the adaptive immune response. The complement system can be activated through three pathways: the classical, lectin and alternative pathways. All three pathways form C3 convertase and converge at the activation of C3 (Walport, 2001; Walport, 2001; Carroll, 2004). Moreover, the hydrolyzed C3 analogue is response for the formation of the C3 convertase in the alternative pathway (Charles A Janeway, 2001). During activation, C3 is cleaved by C3 convertase to C3b whereupon the internal thioester bond in C3 is activated and allows C3b to covalently attach to the hydroxyl groups on surrounding pathogenic, or apoptotic surfaces (Law *et al.*, 1979; Law and Dodds, 1997). The bound C3b can form C3 convertase (C3bBb) to amplify the complement response. The bound C3b can also form C5 convertase to induce the formation of the membrane attack complex by binding

to factor B (Vogt *et al.*, 1978; Fishelson *et al.*, 1984). Alternatively, it will be further cleaved to proteolytic fragments mediated by factor I and other cofactors (Ross *et al.*, 1982).

To achieve the transition from C3 to C3b, a conformational rearrangement is required for C3 to expose the thioester moiety in the TED domain as well as hidden and cryptic binding sites for different cell-surface receptors. This conformational change has been intensely studied, due to its central importance in the regulation and function of complement (Isenman *et al.*, 1981; Perkins and Sim, 1986; Hack *et al.*, 1988; Alsenz *et al.*, 1990; Winters *et al.*, 2005; Nishida *et al.*, 2006; Gros *et al.*, 2008). X-ray crystallographic data provide high resolution structures for both C3 and C3b. Native C3 (187kDa), including a β-chain (residues 1-645) and an α-chain(residues 650-1641), consists of 13 domains: 5 macroglobulin domains MG1-MG5 and the LNK domain form the β-chain; the ANA (anaphylatoxin), MG7, CUB, TED (thioester-containing), MG8 and C345C domains construct the α-chain; the MG6 consists of parts of both the β-chain and the α-chain. The ANA domain is connected to the MG6 domain α-chain part by α'NT segment and the Anchor segment sits in between the MG8 and the C345C domains (Janssen *et al.*, 2005). In C3b, the ANA domain is cleaved and the structure of the α-chain displays significant rearrangements from the C3 structure (Abdul Ajees *et al.*, 2006; Janssen *et al.*, 2006; Wiesmann *et al.*, 2006). In my results, the quantitative cross-link data revealed the conformational differences between C3 and C3b to be consistent to X-ray crystallographic characterization. Moreover the spontaneously hydrolyzed C3 C3(H_2O) was detected from both C3 and C3b samples. The domain architecture of C3(H_2O) was proposed based on the clustered cross-links and the crystal structures of C3 and C3b.

5.3 Quantitative 3D proteomics analysis of C3 and C3b samples

5.3.1 Cross-linking of C3 and C3b

In order to quantitatively compare the conformational differences between C3 and C3b in solution, an experiment scheme (Figure 5.1 A) was designed using stable isotope labelled cross-linkers. The purified C3 sample was cross-linked in solution with unlabelled amine specific cross-linker Bis(Sulfosuccinimidyl) suberate (BS^3-d0, Thermo scientific) while the C3b sample was cross-linked with BS^3-d4, a 4 Da heavier deuterated BS^3 analogue. As shown in Figure 5.2 A, after cross-linking reaction, bands corresponding to two individual chains of both C3 and C3b disappeared in SDS PAGE, while new bands appeared with molecular weight matched to molecules with connected α and β-chains and even higher order oligomers of the proteins. Cross-linker BS^3-d0 and BS^3-d4 are supposed to have identical reactivity. The isotope effect of the deuterated cross-linker may result in differences in retention times compared with its non-deuterated counterpart in reverse phase chromatography; however the cross-linking reaction *per se* is not supposed to be affected (Lee *et al.*, 2007). The similar pattern of cross-linking products on the SDS-PAGE gel (Figure 5.2A) indicated that BS^3-d0 and BS^3-d4 can cross-link both C3 and C3b samples with similar efficiency and products. As seen in figure 5.2B, higher order oligomer of C3 and C3b can form after cross-linking reaction. Only the monomer of cross-linked C3 and C3b were used for subsequent conformational analysis (Figure 5.2 B).

BS^3-d0 cross-linked C3 sample and BS^3-d4 cross-linked C3b sample were mixed with equal molar amount (Sample 1 and 2, *2.3.1*). Theoretically, cross-links formed in both conformations will give doublet signals with 4 Da mass difference in mass spectrometric spectra while the unique cross-links in either conformation will show as singlet signals and the mass of cross-linker indicate from which sample the identified cross-linked peptides are derived (Figure 5.1). An additional reverse labelled experiment (C3-BS3-d4/C3b-BS3-d0) was conducted parallel in order to reduce any systematic error (Sample 1R and 2R, *2.3.1*).

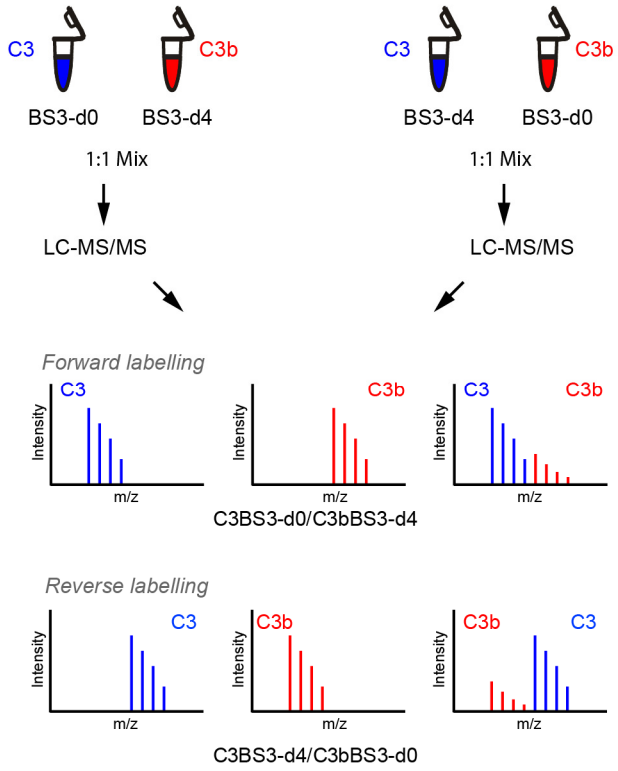


Figure 5.1 - The experimental scheme of quantitative 3D proteomics analysis of C3 and C3b conformational changes in solution.

Moreover, it particularly increases the identification specificity of cross-links that are observed as singlet signals. The SCX chromatogram of Sample 1 and Sample 1R (Figure 5.2 C) and the total ion chromatography (TIC) of SCX fraction 12 from Sample 1 and Sample 1R (Figure 5.2 D) indicated the reproducibility of the forward and reverse labelled experiments.

5.3.2 Identification and quantitation of cross-linked peptides

After LC –MS/MS analysis, 84 spectra were identified as cross-links in Sample1 and Sample 1R, which corresponded to 45 cross-linked peptides and gave rise to 35 unique cross-links. All these 35 cross-links had data in both Sample 1 and 1R. From Sample2 and 2R, 23 linkage pairs were identified based on 36 cross-linked peptides identified from 97 spectra. Again, all linkages had data in both Sample 2 and 2R. Possibly as a result from different fractionation methods the two datasets did not overlay completely. However 15 unique cross-linkages were observed in both datasets. Subsequently, quantitation was carried out at cross-linkage level. As expected, both singlet and doublet signals for cross-linked peptides were observed (Figure 5.3C). For cross-links that have signals derived from both C3 and C3b samples, the C3/C3b signal ratios were calculated as described in *2.3.5*. . This process resulted in 42 quantified unique cross-links (Table S4). Plotting the C3/C3b ratio in the forward labelled samples for each cross-link against its C3/C3b ratio in the paired reverse labelled samples in an x-y scatterplot (Figure 5.3A) showed the C3/C3b ratios were correlated between the forward labelled and the reverse labelled samples (R^2=0.96). For the 15 cross-links observed in both datasets, the C3/C3b ratio showed consistency between these datasets (the pairwise Pearson correlation coefficient of Sample 1 and 2, Sample 1R and 2R are 0.97 and 0.93, indicating that data from two compared experiments are similar).

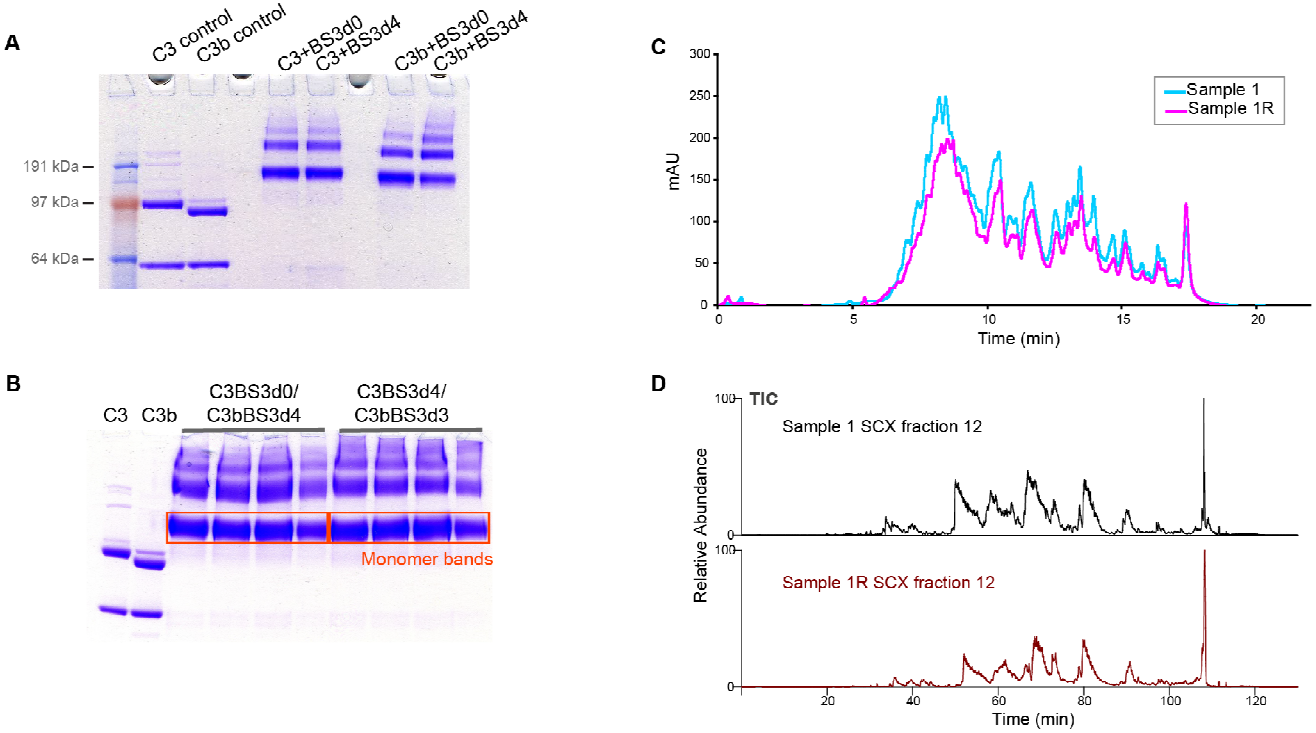


Figure 5.2 - Cross-linking of the C3 and C3b samples

A. Cross-linker BS^3-d0 and BS^3-d4 cross-link both C3 and C3b with similar efficiency and yields for cross-linking products.

B. The monomer of mixed cross-linking products that is isolated for conformational analysis.

C. SCX chromatogram of Sample 1 and Sample 1R.

D. Total ion chromatogram of SCX fraction 12 of Sample 1 and Sample 1R.

Interestingly, for those cross-links that were observed as doublets, the signals from the C3 sample and the C3b sample were not always close to 1:1 as expected. Instead, the C3 to C3b signal ratio of these cross-links ranged from 0.4 to over 10 and they were not evenly distributed (Figure 5.3B, C). Modelling normal distribution into the data resulted in six clusters (sub-population) and one cross-link falling between cluster 4 and 5 remained ambiguous (Figure 5.3B). Cross-links in Cluster 1 and Cluster 6 were observed as singlet signals. The absence of cross-links in either conformation could result from significant conformational differences. However it is also possible that the cross-linkage occurred in both conformations, but because the intensity of cross-linked peptide signal was so low that the less intense signal fell out of detection limit of the instrument. This possibility has been eliminated by the signal to noise ratios of these unique cross-links, since they are far beyond the range of C3/C3b signal ratios of detected doublets (the noise intensity was defined to 1E4 as the average intensity level of stochastic peaks detected in the mass spectrometer).

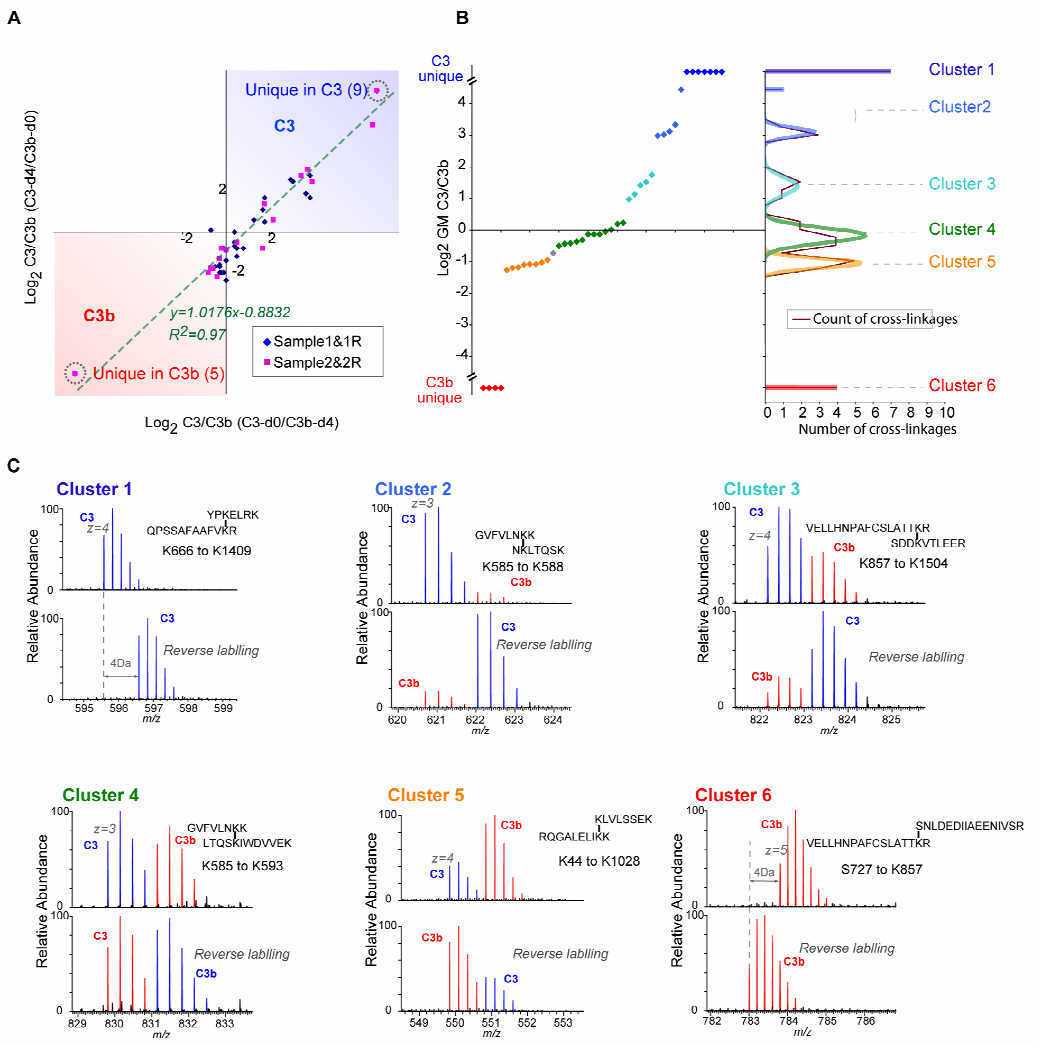


Figure 5.3 - Quantitation of cross-links

A. Scatterplot of the logarithm of C3 to C3b ratio (Log2 C3/C3b) for all cross-linkage pairs observed in Sample 1 and 1R: (blue rhombus); Sample 2 and 2R (magenta square). The upper right quadrant represents C3 conformation and the lower left quadrant represents C3b conformation. The 10 unique cross-links in C3 and 5 in C3b were also plotted and marked. The linear tend line of all data pointS is shown as a green dashed line with equation and R^2 labelled.

B. The Log2 value of geometric mean of C3 to C3b ratios detected for each cross-linkage from 2 datasets were sorted and plotted. 42 cross-links were grouped in 6 clusters and coloured. The unique cross-links in C3 and C3b: Class 1 and Class 6. The distribution of Log_2C3/C3b value for cross-links with signal from both conformations were modelled into baseline separated natural distributions which were defined as the Class 2-5. The single data point at 4.43 was combined in Class 2. The data point at -0.7 (Gray) could not be unambiguously assigned to either Class 4 or Class 5, therefore left out.

C. Example MS^1 spectra of cross-linked peptides from each quantitative cluster shown in B. The spectra from both forward labelled (top) and reverse labelled (bottom) experiment for each cross-linked peptide were shown.

5.3.3 Quantified cross-links suggested differences between C3 and C3b samples

Noticeably, the 42 cross-links that were quantitatively clustered were also not evenly distributed among C3/C3b domains (Figure 5.4).

- Cluster 1: the seven cross-links in this cluster were unique in the C3 sample. These cross-links were mainly observed in the α-chain, around the ANA, MG7, MG8 and TED domains.
- Cluster 2: five cross-links in this cluster were significantly enriched in the C3 sample (with geometric mean of C3 to C3b ratio 7.9 to 21.6). These cross-links were predominantly distributed around the ANA and MG8 domains.
- Cluster 3: the five cross-links in this cluster (with geometric mean of C3 to C3b ratio 1.9 to 3.4 average 2.6) were observed from the, MG7 domain and vicinity.
- Cluster 4: twelve cross-links in this cluster were observed nearly equal in C3 and C3b samples (with geometric mean of C3 to C3b ratio 0.6 to 1.1 and in average 0.9). These cross-links were spread across both the α-chain and the β-chain.
- Cluster 5: the eight cross-links in this cluster were enriched in the C3b sample (with geometric mean of C3 to C3b ratio 0.4 to 0.5). Half of these linkages were between the TED domain and the MG1 domain. The rest of the cross-links involved the MG7, MG8, C345C domains and the Anchor segment.
- Cluster 6: the four cross-links in this cluster were unique in the C3b sample Two of them were cross-linked to the N-terminal of the C3b α'-chain in C3b and two of them were between the MG7 and MG8 domains.

Cross-link data reflected the sequence difference between C3 and C3b. Cross-links involving the C3-specific ANA domain were observed only in the C3 sample (Cluster 1), and cross-links to the C3b α'-chain N-terminal were only observed in the C3b sample (Cluster 6). Interestingly, in Cluster 2, two linkages to the residues in the C3-specific ANA

domain were also detected in the C3b sample. According to the C3/C3b ratio (~9), there were around 10 % of the C3b sample gave rise to cross-links representing same structural feature as C3. Estimated from the ANA domain peptides identified in C3b sample by mass spectrometry, there was about 10%contamination with C3 sequences in the C3b sample (*Appendix 1*), the C3b signal detected in these linkages were very possible derived from this impurity in the C3b sample. While in Cluster 3 and 5, cross-links were rather enriched in either C3 or C3b sample. All cross-links in the aforementioned five clusters (Cluster 1, 2, 3, 5 and 6) indicated differences between the C3 and C3b samples. These cross-links were concentrated in the α-chain of C3/C3b implying major conformational rearrangements in this part of the protein. In contrast cross-links in Cluster 4 appeared to have near equal observation from both samples suggesting at least some structural similarity between the C3 and C3b.

5.4 Quantitative cross-link data is in agreement with the crystal structures of C3 and C3b

5.4.1 Cross-linking data and the crystal structures agreed on residue proximity

In order to further understand and interpret the quantitative cross-link data, I compared the cross-link data with the well accepted crystal structure of C3 (PDB|2a73) and C3b (PDB|2i07). The two structures contain 12 domains in common, with C3 having an additional ANA domain. The 42 quantified cross-links were displayed using Pymol on both the C3 and the C3b structures. Residues that are absent in the crystal structure, which was the case for 3 residues were substituted with the nearest residues in the sequences up to a maximum distance of 5 residues. Hence, 37 cross-links were visualized in both structures while 5 were displayed only in C3 due to the missing ANA domain in the C3b structure. As was the case for the Pol II analysis (see Chapter 4), cross-link data and X-ray structure data were first compared on residue proximity.

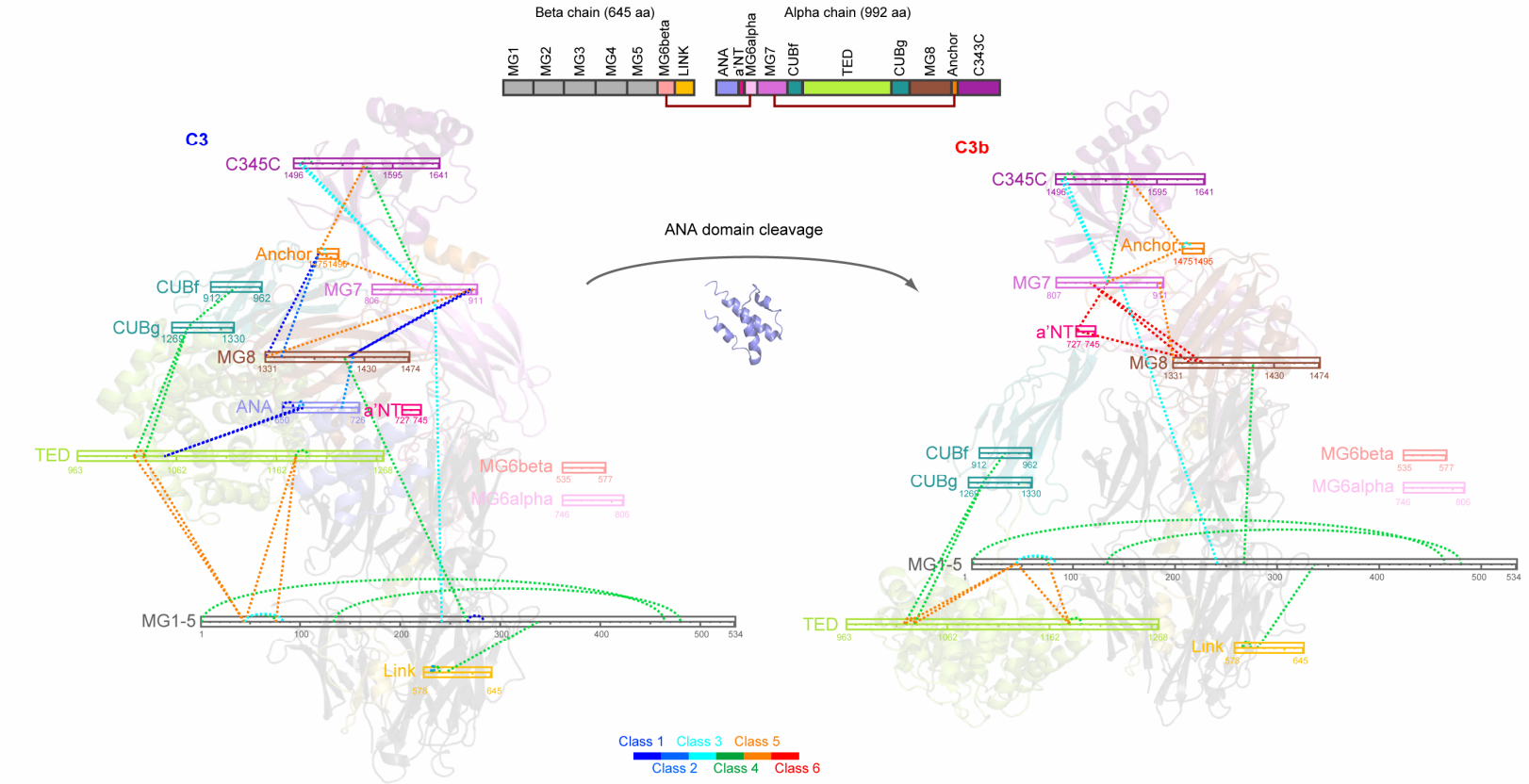


Figure5.4 - Cross-links observed in C3 and C3b samples

Quantified cross-linkages (dashed lines) mapped onto a schematic representation of C3 and C3b domains (bar). Cross-links were colour coded by the quantitation categories. Domains were coloured by the domains definition according to Janssen (Janssen *et al.*, 2005). The cross-links maps were overlaid on top of views of protein crystal structures, PDB 2a73 for C3 and PDB|2i07 for C3b.

The cross-linking limit of BS^3 is 24.4 Å between α-carbon atoms of two residues ignoring the coordinate error for mobile surface residues in the crystal structures (Chapter 4). In total, 61 C-α distances of cross-linked lysine pairs were measured from both C3 and C3b structures for 13 unique cross-links and 24 common linkages. 50 (82%) of them are within the 24.4Å limit indicating agreement between the cross-link data and the crystal structures. Eleven cross-links that did not fall bellow the distance threshold will be further discussed in *5.7* and *5.8*.

5.4.2 Cross-link data confirmed in solution the structural similarities and differences between C3 and C3b characterized by protein crystallography

In the six cross-link clusters, Cluster 1, 4 and 6 were expected. Cross-links in Cluster 1 and 6 were observed as singlet signals and implied significant conformational differences. Firstly, seven cross-links in Cluster 1 reflected C3-specific structural features. Three of them involve the C3 unique ANA domain (Figure 3.5B). Among them two between the ANA and TED domain also confirmed the location of the TED domain in C3 conformation distal to the MG1 end of molecule as observed in the C3 crystal structure. The other three cross-links in this cluster are only possible in the C3 structure because in the C3b structure the paired residues are too far away to be cross-linked by BS^3 (in average 50% longer then the cross-linking limit) (Figure 3.5 C). The linkage between $K267^{MG3}$ and $K283^{MG3}$ reflected the unique solvent accessibility of cross-linked residues in C3 conformation, which will be further discussed in next session. Secondly, four cross-links in Cluster 6 indicated specific C3b structural features. Two cross-links were linked to the N-terminal of C3b α' chain that becomes cross-linkable after the cleavage of the ANA domain (Figure 3.5 B). The other two cross-links formed between residues pairs that only get close enough for cross-linking in the C3b structure (Figure 3.5C).

In contrast, twelve cross-links in Cluster 4 appeared to have near equal observation from both samples (with average C3/C3b ratio 0.9, ranged from 0.6~1.1) implying similarities between the C3 and C3b structures. These cross-links were rather spread in C3/C3b structures, five within domains and seven between domains. Most (nine) involved amino acid pairs whose distances are virtually identical in the crystal structures of C3 and C3b (with less than 1Å distance variation between the two structures). Two linkages, in both crystal structures, vastly exceeded the distance that the cross-linker is able to bridge indicating dynamic aspects of the TED domain, as will be discussed below. Cross-links in this cluster reflected both structural conservation of domain conformations and the resembling domain architectures between C3 and C3b (Figure 5.5A). In summary quantitative cross-link data revealed conformational differences and similarities between C3 and C3b which are in agreement with previous observations by X-ray crystallography.

As shown in Figure 5.6, the schematic domain architectures C3 and C3b were built from the counterpart crystal structure based on the proximity between residues indicated by observed cross-links. The β-chain that includes the MG1-5 region and the LNK domain kept the similar structure in both models, while the domains in the α-chain (mainly the CUB, TED, C345C, MG7 and MG8 domains) had major rearrangements between the two conformations.

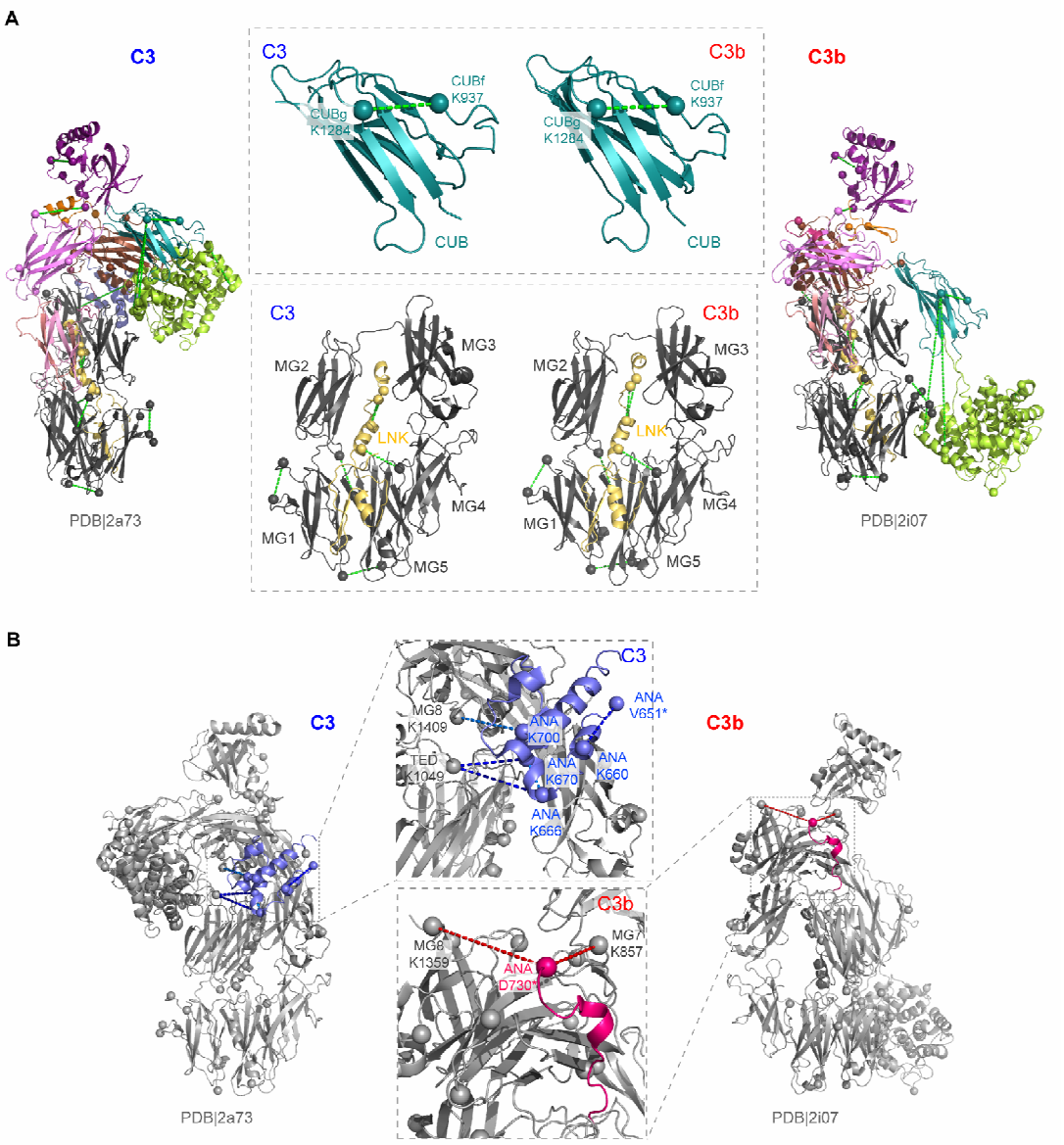


Figure 5.5 - Quantitative cross-link data reflects similarities and differences between C3 and C3b

Cross-linking data were annotated using crystal structures (PDB|2a73 for C3, PDB|2i07 forC3b). Cross-links were displayed by creating connections (shown as dashed lines) between α-carbons (sphere) of cross-linked residues. Cross-links and domains were coloured coded as defined in the figure 5.4. The substitutes for the missing residues in crystal structures for display were marked by asterisks in the labels

A. Cross-links common to C3 and C3b are displayed in the two crystal structures showing the similarity between the conformations. The cross-links in the CUB domain as well the MG1-5 and LNK domains were highlighted in the extracted magnified regions from both structures.

B. Conformation-specific cross-links, observed from the ANA domain in C3 and from the N-terminal of the α'-chain in C3b, reflect the cleavage of the ANA domain from C3 to C3b. The ANA domain in C3 and the α'NT domain in C3b were presented in the magnified regions

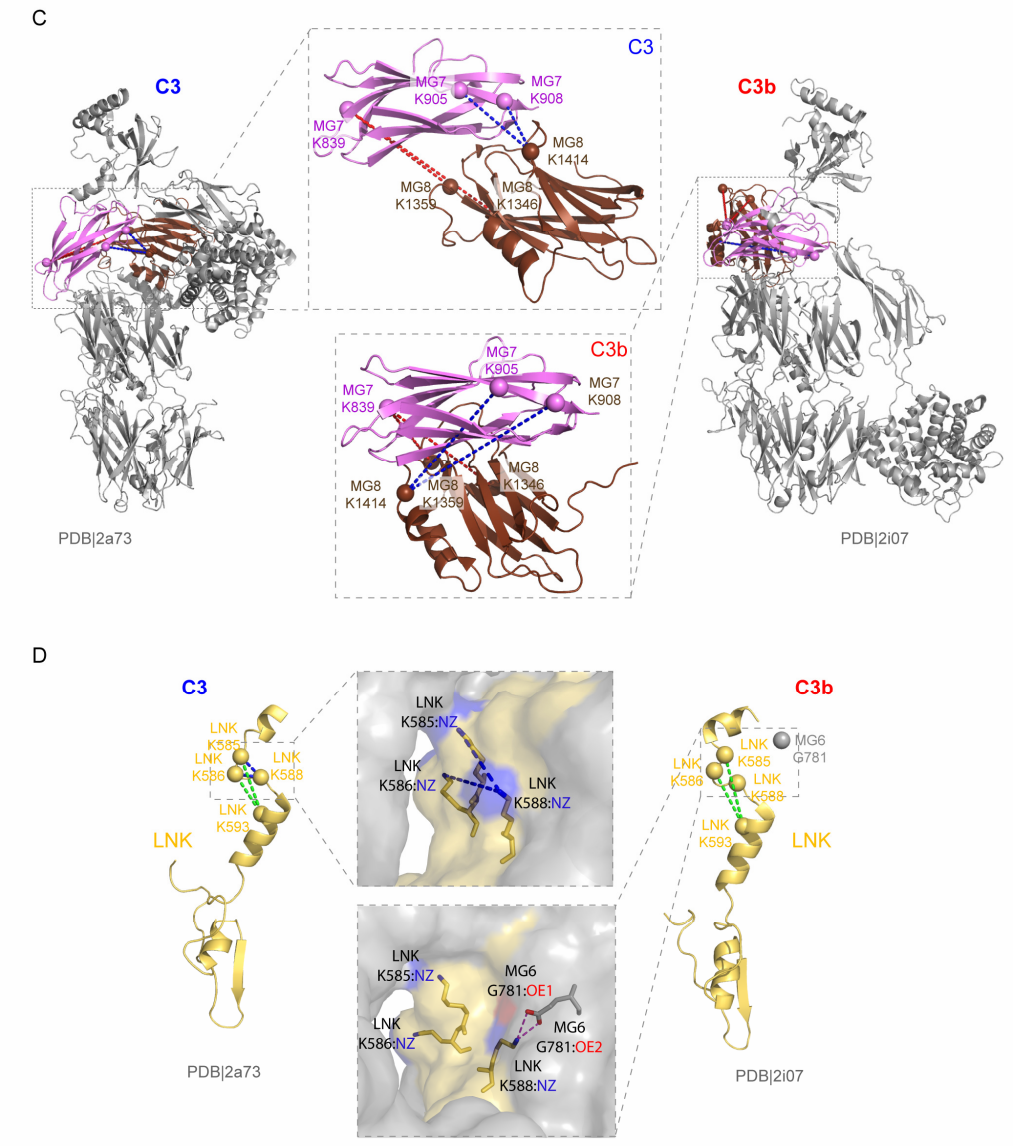


Figure 5.5 continued - Cross-link data reflects similarities and differences between C3 and C3b.

C. A series of cross-links between the MG7 and MG8 domains with proximity changes demonstrated the conformational rearrangement between these domains. The insets show the magnified views of the MG7 and MG8 domains from C3 and C3b structures.

D. The hydrogen bonds between the NZ atom of residue $K588^{LNK}$ and the OE1 and OE2 atoms in residue $G781^{MG6}$ shown in the C3b crystal structure could affect the reactivity of K588 to cross-linkers and be responsible for the absence of cross-links from K588 to K585 and K586 in C3b. (Displayed in the magnified view: the protein molecules in solvent accessible surfaces; residues K585, K586, K588 and G781 are represented by sticks with NZ atoms highlighted in blue and OE1 and OE2 in red). The cross-links $K585^{LNK}$-$K593^{LNK}$ and $K586^{LNK}$-$K593^{LNK}$ in the same region were observed near equally from both the C3 and C3b samples.

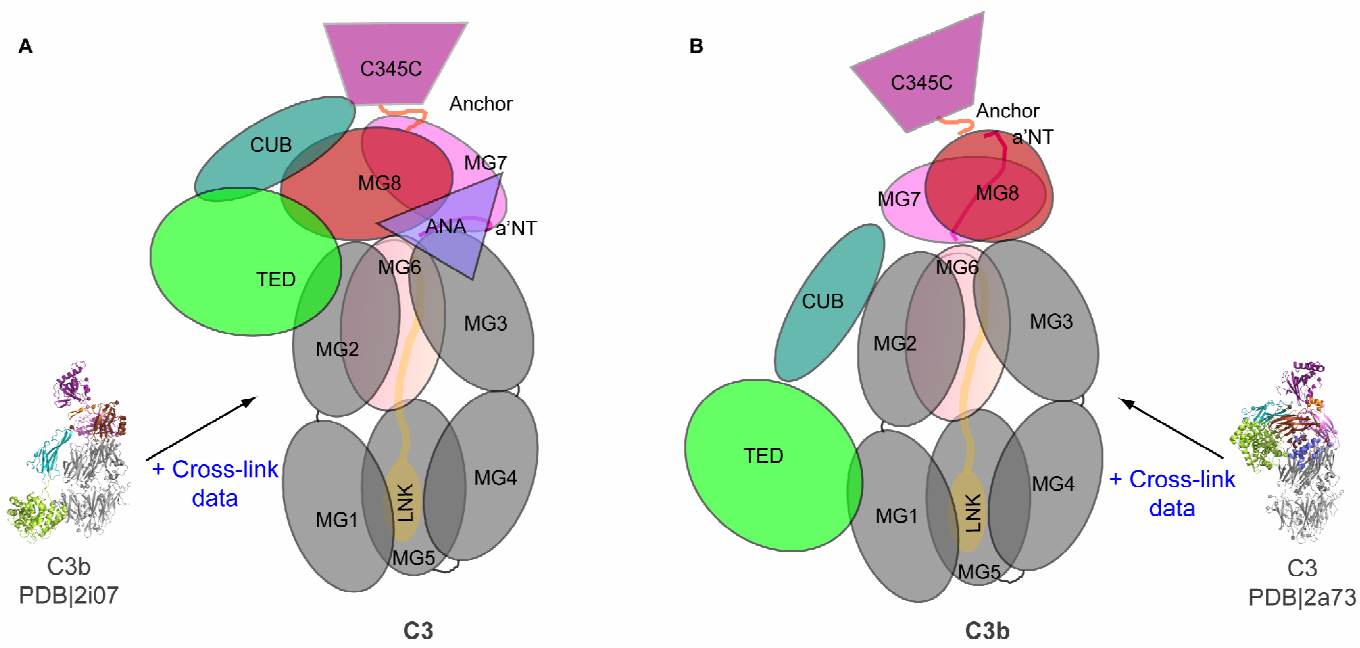


Figure 5.6 - Domain architectures of C3 and C3b as derived from cross-link data

A. The schematic domain architecture of C3 built from the C3b crystal structure PDB|2i07 based on C3-specific cross-links that suggest C3-specific structural features in respect to C3b (Cluster 1 and 2).

B. The schematic domain architecture of C3n built from the C3 crystal structure PDB|2a73 based on C3b-specific cross-links that suggest C3-specific structural features in respect to C3 (Cluster 5 and 6)

.

5.5 Quantitative cross-link data uncovered hydrolyzed C3 in the presence of C3 and C3b

In Cluster 5, four cross-links, $K44^{MG1}$-$K1181^{TED}$, $K75^{MG1}$-$K1181^{TED}$, $K44^{MG1}$-$K1019^{TED}$ and $K44^{MG1}$-$K1028^{TED}$, that indicated the close proximity of the TED and MG1 domains, were also observed in the C3 sample, although with no as intense as the signals detected in the C3b sample. Two cross-links between the TED and ANA domains and linkages between the ANA and MG8 domains (Cluster 1) indicated the TED domain location distal to the MG1 end and exclusively in the C3 conformation. Conflicting evidence for the TED domain location in C3 suggested conformational heterogeneity of the C3 sample (Figure 5.7). Beside the four cross-links between the TED and MG1 domains, the other three cross-links in Cluster 5, the $K908^{MG7}$-$K1475^{Anchro}$, $K908^{MG7}$-$K1331^{MG8}$ and the $K1475^{Anchor}$-$K1567^{C345C}$, implied the co-occurrence of TED domain migration and a rearrangement of the domains to a C3b-like conformation. This observation raises two possibilities: either there was some contaminating C3b in the C3 sample, or there is an alternative C3 conformation. Since some C3b specific features, for example the cross-links to the α'-chain N-terminal were not detected in the C3 sample, the possibility of C3b contamination the C3 sample was excluded (Figure 5.8). Interestingly, this second conformer was not resolved from C3 by SDS-PAGE, before and after cross-linking, which suggested that it has a similar molecular weight and α- and β-chain composition to C3. An important clue came from the fact that in aqueous environments, the thioester in the C3 TED domain can be spontaneously hydrolyzed with low rate (*t1/2*~200h) (Nishida *et al.*, 2006). Moreover, the C3b-like C3(H_2O) structural features have been observed previously by numerous biophysical techniques including EM, where the TED domain is proximal to the MG1 domain (Nishida *et al.*, 2006). C3b-like data in the C3 sample are in full agreement with the presence of contaminating C3(H_2O).

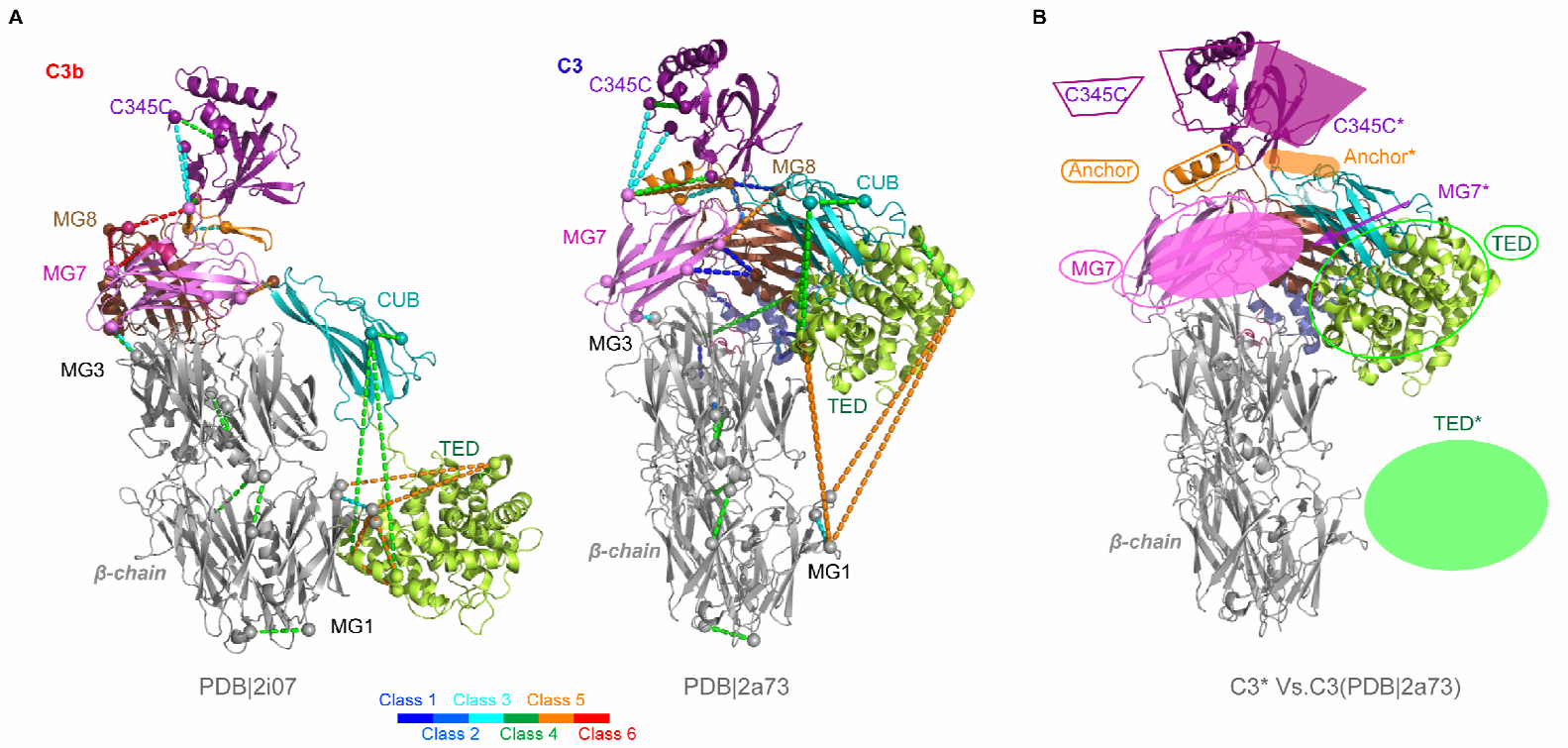


Figure 5.7 - Quantitative cross-link data suggested an alternative conformation existed in the C3 sample

A. Cluster 5 cross-linkages indicated an alternative conformation in C3 sample that exhibited C3b-like structural features in α-chain domains and conflicted with the C3 unique features (Cluster 1); this conformation also distinguished from C3b by its unique features (Cluster 3) and C3b unique features (Cluster 6); Moreover, this conformation kept similarity to C3 structure in α-chain (Cluster 2) and the β-chain domain architecture remain conserved among three conformations (Cluster 4). (The involved cross-linkages were displayed in C3 and C3b crystal structures as dotted lines and colour coded according to clusters.)

B. The alternative conformation in C3 sample marked as C3* was displayed on top of the C3 crystal structure (PDB|2a73) with the MG7, C345C, TED domains and Anchor presented in the C3b-like alternative position and marked MG7*, C345* , TED* and Anchor*.

A similar case was uncovered in the C3b sample. As mentioned previously, there was 10%-15% contamination with C3 sequence in C3b. In Cluster 2 the C3-specific linkages in the ANA domain were consequently observed in the C3b sample with around 10% of the signal intensity seen in the C3 sample. However, cross-links in Cluster 1were not detected for the C3b sample. These cross-links included 2 linkages between the TED domain and the ANA domain that are considered as a signature feature of the C3 conformation indicating the distal TED domain position to the MG1 domain. Therefore, the contamination in the C3b sample is more likely C3(H_2O) which has identical sequence to C3 and likely to be co-purified with C3b due to their similarities on physical and chemical properties.

Since the C3 and C3b sample were mixed in equal amounts for quantitation, equal total intensity for these two samples was assumed. Based on the ratios of Clusters 2 and 5 and the previously determined 10-15% C3(H_2O) in C3b sample (*Appendix.1*), C3(H_2O) contributed 30-50% protein mass to the C3 sample. Having C3(H_2O) as a shared component in both C3 and C3b samples also explained the existence of Cluster 3 as linkages specific to C3(H_2O). The C3/C3b signal ratios of cross-links in Cluster 3 were 2-3.5 which is in agreement with the estimated portion of C3(H_2O) in the C3 and C3b samples. In summary based on the C3/C3b signal ratio of cross-links and their distribution in the crystal structures of C3 and C3b, the structural features revealed by the six clustered cross-links were interpreted as follows: C3-specific features (Cluster 1, 7 linkages), features shared by C3 and C3(H_2O) (Cluster 2, 5 linkages), C3(H_2O)-specific features(Cluster 3, 5 linkages), common features of C3, C3b and C3(H_2O) (Cluster 4, 12 linkages), features shared by C3b and C3(H_2O) (Cluster 5, 8 linkages) and C3b specific features (Cluster 6, 4 linkages) (Table 5.1).

The assignment of linkages into clusters assumes that the influence of conformational changes on cross-links works in an ‘on/off’ manner. However, in reality, the yield of a cross-link may also be affected by the conformational changes in less dramatic

ways. This explains why the signal ratios of cross-links that are shared by C3, C3b, and C3(H_2O) (Cluster 4) were not always 1:1 between the C3 and C3b samples. Depending on the extent of the conformation effect, for individual cross-links it may lead to misassignment to clusters. In fact, I believe this to be true, having observed this for two linkages: $K267^{MG3}$-$K1409^{MG8}$ and $K44^{MG1}$-$K82^{MG1}$. According to its ratio, cross-link $K267^{MG3}$-$K1409^{MG8}$ falls onto the boundary of cluster 4 near to cluster 5. However, according to the crystal structure, this cross-link is possible only in C3b and not C3. Hence it is likely a misassigned member of Cluster 5. Similarly, the ratio of linkage $K44^{MG1}$-$K82^{MG1}$falls in Cluster 3 and was assigned as a C3(H_2O)-specific feature. However, from the crystal structures of C3 and C3b one would have expected this linkage to be possible also in these proteins. In fact, this is the case for a very similar linkage, $K44^{MG1}$-$K75^{MG1}$ (Cluster 4). Consequently, these two cross-links (5% of total 41 clustered cross-links) have to be regarded as false cluster assignment. Nevertheless, this study demonstrated that quantitative 3D proteomics overall is able to reliably reveal conformational differences. The quantitative cross-link data gave new insights into the conformation of C3(H_2O) as well as C3 and C3b.

Table 5.1 - Interpretation of clustered cross-links

Quantified linkage groups	Structural feature	Estimated C3/C3b ratio[2]	Observed C3/C3b ratio[1]	Involved linkages
Cluster 1	**C3 specific**	**C3 unique**	**C3 unique**	K650(ANA)-K660(ANA), K666(ANA)-K1049(TED), K670(ANA)-K1049(TED); K905(MG7)-K1414(MG8), K908(MG7)-K1414(MG8); K267(MG3)-K283(MG3), K1331(MG8)-K1475(Anchor)
Cluster 2	**Common in C3&C3 (H2O)**	**8.00**	**8.7 (7.9-10.1)**	K666(ANA)-K670(ANA), K700(ANA)-K1409(MG8); K585(Link)-K588(Link), K586(Link)-K588(Link); K1346(MG8)-K1475(Anchor)
Cluster 3	**C3(H2O) specific**	**3.20**	**2.6 (1.9-3.4)**	K241(MG3)-K869(MG7),K857(MG7)-K1504(C345C), K857(MG7)-K1500(C345C);K1475(Anchor)-K1482(Anchor); K44(MG1)-K82(MG1)
Cluster 4	**Common in C3, C3b &C3 (H2O)**	**1.00**	**0.9 (0.6-1.1)**	K586(Link)-K593(Link), K585(Link)-K593(Link); K267(MG3)-K1409(MG8), K133(MG2)-K480(MG5), K1(MG1)-K464(MG5), K337(MG4)-K600(Link); K1019(TED)-K1284(CUB2), K1029(TED)-K1284(CUB2), K1181(TED)-K1193(TED); K857(MG7)-K1567(C345C); K1504(C345C)-K1513(C345C); K267(MG3)-K1409(MG8);
Cluster 5	**Common in C3b and C3(H_2O)**	**0.40**	**0.5 (0.4-0.5)**	K44(MG1)-K1181(TED),K44(MG1)-K1028(TED), K44(MG1)-K1019(TED), K75(MG1)-K1181(TED); K908(MG7)-K1331(MG8), K857(MG7)-K1475(Anchor), K1475(Anchor)-K1567(C345C); K1019(TED)-K1029(TED)
Cluster 6	**C3b specific**	**C3b unique**	**C3b unique**	K727(α'-NT)-K857(MG7), K727(α'-NT)-K1359(MG8); K839(MG7)-K1359(MG8), K839(MG7)-K1346(MG8)

[1] The observed C3/C3b ratios in each cluster were the average of the geometric means (GM) of all observed ratios for each individual cross-linkage. The range for GM of all linkages in each cluster is listed in the parenthesis.

[2] Based on the hypothesized sample composition: Total signal from both C3 and C3b samples were assumed equal, and 10% to 15% C3(H_2O) in the C3b sample, 30% to 50% estimated from observed C3/C3b ratio, the estimate ratios were calculated with equal C3 and C3b total signals, 40% C3(H_2O) in the C3 sample and 12.5% C3(H_2O) in the C3b sample.

5.6 Domain architecture of C3(H_2O)

According to the structural features implied by cross-link clusters and based on the existing crystal structures of C3 and C3b, a schematic domain architecture for C3(H_2O) is proposed (Figure 5.8). The hydrolysis of the thioester in the TED domain causes relocation of the TED and CUB domains, and the incline of the C345C domain to a C3b-like conformation (Cluster 5 cross-links). The movement of these domains consequently induced the conformational rearrangements of the connected MG7, MG8 domains and the Anchor segment to some C3b-like features (Cluster 5). However, due to the the ANA domain remaining in C3(H_2O), the structural rearrangements in the α-chain cannot reach the identical conformation as in C3b. Instead, C3(H_2O)-specific conformation formed around the MG7 domain (Cluster 3) and some C3-link conformational features remained around the ANA and MG8 domains (Cluster 2). In contrast to the highly dynamic α-chain, the β-chain of the protein keeps similar architecture in C3, C3b and C3(H_2O).

Noticeably, there were some details of conformational difference between C3, C3b and C3(H_2O) revealed by cross-links. Two cross-links in Cluster 2 ($K585^{LNK}$-$K588^{LNK}$ and $K586^{LNK}$-$K588^{LNK}$) connected residues close in sequence and with almost no proximity change (~1Å) between C3 and C3b in the crystal structures. They were not observed in C3b while two other cross-links ($K585^{LNK}$-$K593^{LNK}$ and $K586^{LNK}$-$K593^{LNK}$) in the same region were equally detected in all three conformations (Cluster 4). The absence of the two Cluster 2 cross-links might due to the fact that in the C3b crystal structure, the nitrogen atom (NZ) of K588 side chain forms hydrogen bond with two Oxygen atoms (EO1 or EO2) in residue G781 (Figure 5.5D). The formation of hydrogen bonds could have affected the reactivity of K588 to the cross-linker. Given that these two cross-links were only absent in C3b, the formation of the hydrogen bonds is likely induced by ANA domain cleavage. In another case, the linkage between $K267^{MG3}$ and $K283^{MG3}$ was defined as C3 exclusive feature and was not detect from C3(H_2O).

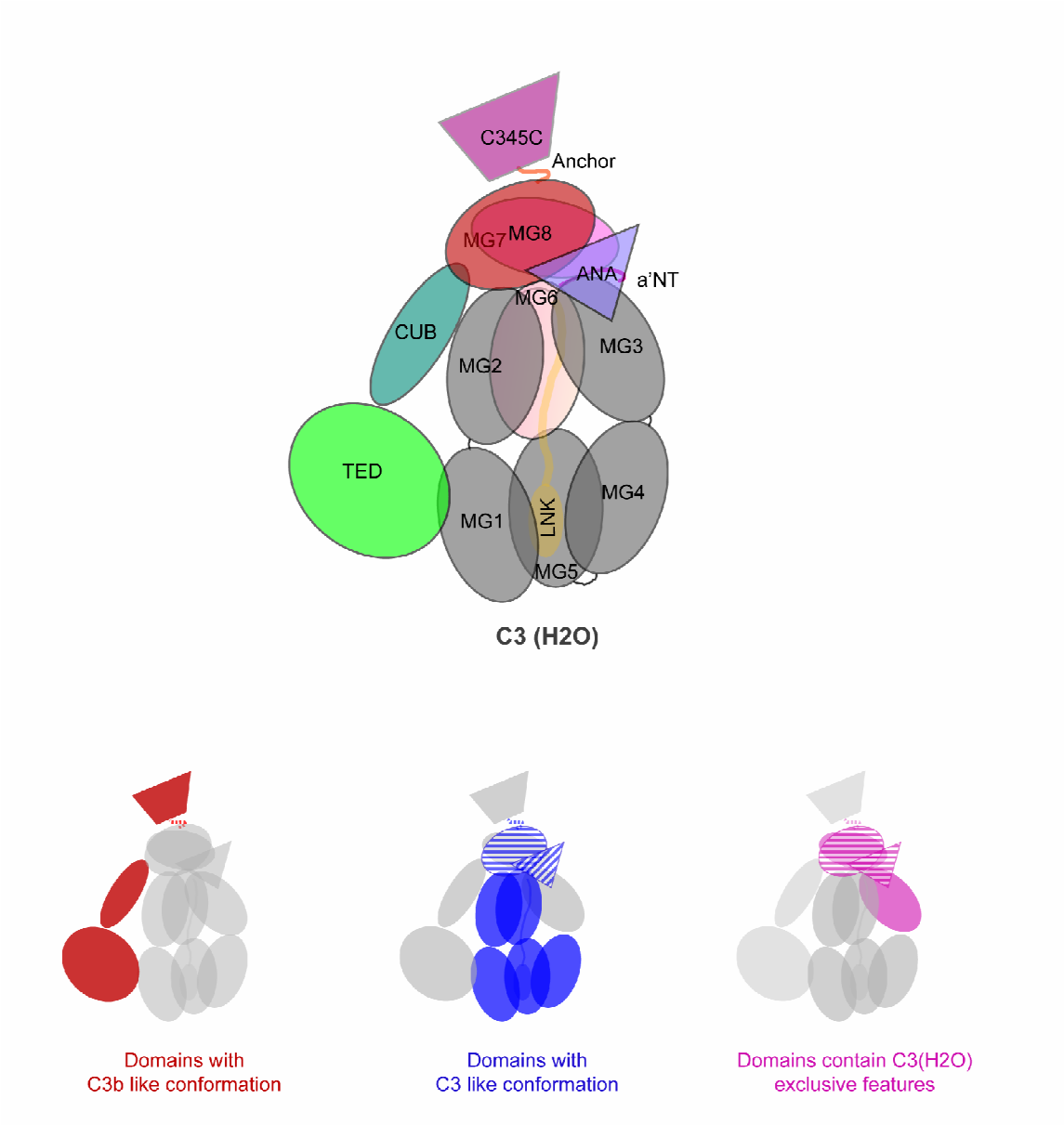


Figure 5.8 - Domain architecture of C3(H_2O)

The schematic domain architecture of C3 (H2O) is proposed based on the quantified cross-linking data and crystal structures of C3 and C3b (PDV|2a73 and PDB|2i07). It combines both C3-like and C3b-like conformational features. In the bottom panel, the C3 conformation like domains (blue), the C3b conformation like domains (red) and the domains that displayed exclusive C3(H_2O) features (magenta) are highlighted. The MG8 and ANA domains present C3-like structural features but also exhibit unique features in C3(H_2O) and are therefore displayed in coloured stripes. The flexible Anchor segment which presented specific features in all 3 conformations was displayed as a dashed stroke.

Significant solvent accessibility changes for peptides containing both resides were previously reported (Winters *et al.*, 2005). The absence of this linkage in C3b inferred similar solvent accessibility change occurred during C3 to C3b transition. In summary, in this study, quantitative 3D proteomics data distinguished C3, C3b and C3(H_2O) conformationally. Not only the differences on primary sequences and residue proximity can be detected by cross-links, but also changes on surface accessibility and residue reactivity will be reflected by the yield of certain cross-links.

5.7 Flexibility of the TED domain in C3b and C3(H_2O).

As discussed in *5.4.1*, the majority of cross-links in this study are in agreement with the crystal structures. However, I also observed 6 cross-links that conflict with the crystal structures with distance far greater (38~64 Å) than the given limit for cross-linking (24Å). All of these involved the TED domain. The mobility of a single residue could not explain the observation of these cross-links.

In C3b and C3(H_2O) conformations, four cross-links were observed from the TED to two neighbouring residues K44 and K75 in the MG1 domain. Two cross-links ($K44^{MG1}$-$K1019^{TED}$ and $K44^{MG1}$-$K1028^{TED}$) are in agreement with the C3b crystal structure. The other two cross-links to $K1181^{TED}$ that is located on the other side of the TED domain are beyond the cross-linking limit in the crystal structure. It is impossible to obtain these four cross-links simultaneously in a homogenous static structure. However the proximity required for these cross-links can be fulfilled with different TED domain positions when assuming flexibility of this domain in solution. Such a domain movement may also explain the other two over-length cross-links in the crystal structure ($K1019^{TED}$-$K1284^{CUB}$ and $K1029^{TED}$-$K1284^{CUB}$). Therefore in this analysis, multiple positions of the TED domain were captured by chemical cross-linking and detected as an overlaid image by mass spectrometry. This is in

agreement with the existence of multiple positions of the TED domain as seen by EM for C3b and C3(H_2O) (Nishida *et al.*, 2006).

Interestingly, the two cross-links between the TED and CUB domains were observed also in C3 and also there conflict with the crystal structure. This may indicate that the TED domain also has certain mobility in C3, however this mobility is restricted, since the hydrophobic interface between the TED and MG8 domain need to remain in order to protect the thioester in the TED domain from hydrolysis. In fact, the spontaneous hydrolysis of this bond to form C3(H_2O) may be a consequence of this TED mobility in C3.

5.8 Cross-link data contradicts a false C3b crystal structure

24 cross-links that represent C3b structural features (Clusters 4, 5 and 6) were also used as a reference to compare two crystal structures of C3b, PDB|2i07 and PDB|2hr0. When aligning these two crystal structures in Pymol, a visual comparison suggested that the major differences were the conformation of the CUB domain and the orientation and location of the TED domain. These differences are reflected in the proximity of a set of cross-links within the CUB domain and between the TED and MG1 domains. A cross-link was observed between $K937^{CUB}$ and $K1284^{CUB}$ as a common feature of C3, C3b and C3(H_2O) (Cluster 4). This cross-link indicates the close proximity between these two residues and the conserved conformation of the CUB domain (Figure 5.9). Therefore this evidence speaks against the unfolded structure in the 2hr0 structure. Furthermore, observation of four cross-links between the TED and MG1 domain indicated that the TED domain is located near to the MG1 domain (average C-α distance between cross-linked residues are 25 Å) than far away from it (average C-α distance between cross-linked residues are 57 Å). Consequently, the 2i07 structure for C3b was supported by cross-link data against the 2hr0 structure, which has been suggested to be fraudulent (Borrell, 2009)

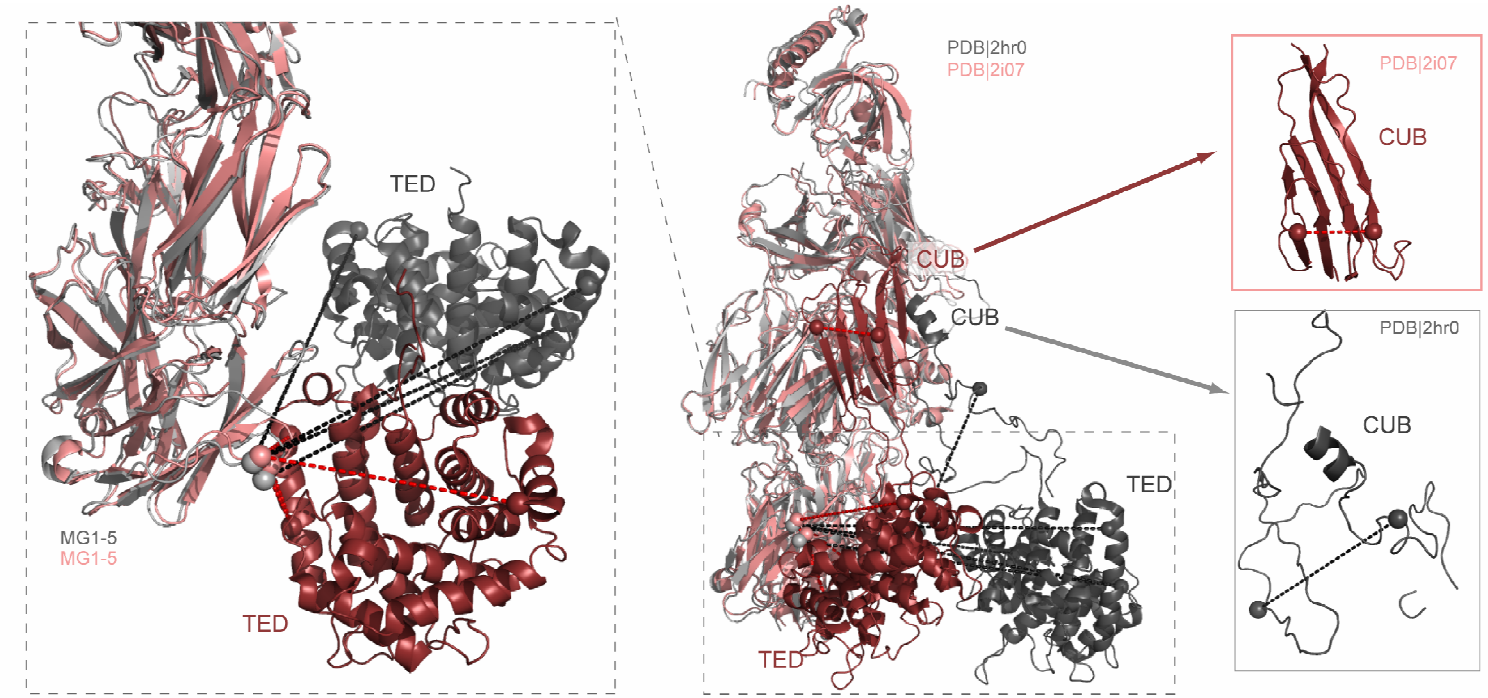


Figure 5.9 - Cross-link data contradicts a fraudulent C3b crystal structure.

Cross-linking data distinguished the C3b crystal structure 2i07 (light salmon) from the crystal structure 2hr0 (light grey). The two crystal structures are aligned. The CUB domain construction and the TED domain location carry the major differences between the two structures. These two domains are highlighted by more colour density and cross-links within the CUB domain and between the TED and MG1 domains are displayed in both structures (red for 2i07 and black for 2hr0). As shown in the magnified view, the distance constraint carried by cross-links supported the compacted CUB domain structure and proximal position of the TED domain to the MG1 domain in structure 2i07 contrary to the 2hr0 structure.

5.9 Discussion

5.9.1 C3b-like functional domain arrangement and the function of C3(H_2O)

In this work, quantitative 3D proteomic was applied to study in solution the conformational difference between C3 and C3b. The quantified cross-links uncovered the presence of C3(H_2O) both C3 and C3b samples. C3(H_2O), also called C3b-like C3, is a C3 analogue induced by spontaneous hydrolysis of the thioester bond in the TED domain. Functionally, C3(H2O) is responsible to the basal level activation of the complement alternative pathway. It can bind plasma protein factor B and form C3(H_2O)Bb, a fluid-phase C3 convertase which can cleave C3 to C3b(Charles A Janeway, 2001). In this way the alternative pathway and other two complement activation pathways converge at the central activation and amplification step of the complement system (Gros *et al.*, 2008). Like C3, C3(H_2O) is also susceptible to cleavage and inactivation by factor I in presence of factor H (Pangburn *et al.*, 1981) (Hack *et al.*, 1990). Previous studies revealed apparent conformational differences between C3(H_2O) and native C3(Isenman *et al.*, 1981; Hack *et al.*, 1988; Winters *et al.*, 2005; Nishida *et al.*, 2006), however the structure of C3(H_2O) remains unresolved. In this analysis quantitative cross-link data in combination with the crystal structures of C3 and C3b revealed that C3(H_2O) adopts the C3b domain arrangement for the TED CUB and C345C domain. This result is in agreement with previous studies. Analysis using hydrogen/deuterium exchange in combination with mass spectrometry revealed significant conformational difference among the CUB and TED domains between C3(H_2O) and native C3 (Winters *et al.*, 2005). The C3b-like TED domain location in C3(H_2O) was also confirmed by electron microscopy observations (Nishida *et al.*, 2006) and X-ray scattering data (Li *et al.*, 2010). X-ray crystallographic data on C3b in complex with factor B showed that factor B interaction sites on α'NT, CUB and C345C domains around the MG7 domain were exposed in C3b after the activation induced structural rearrangement from C3. The C3b-like conformation for the C345C and CUB domains would allow the similar binding of

factor B to C3(H_2O) and eventually form the C3 convertase C3(H_2O)Bb (Janssen *et al.*, 2006; Forneris *et al.*, 2010). As part of complement regulation, the degradation of the C3b CUB domain by factor I is assumed to be oriented by complement regulator factor H through its binding to the TED domain (Lambris *et al.*, 1988) (Janssen *et al.*, 2006). Due to the C3b-like conformation in these domains, this regulation mechanism could also be applied to C3(H_2O). Hence, the functional similarity between C3(H_2O) and C3b maybe explained by the structural similarity between them. It also allowed me to further hypothesize hypotheses on the structural similarity between the C3 convertase C3(H_2O)Bb and C3bBb which would structurally unify the complement activation pathways.

5.9.2 Outlook for quantitative 3D proteomics

Quantitative 3D proteomics expanded the 3D proteomics methodology towards quantitative analyses to study protein dynamics. Stable isotope labelling was introduced *via* the cross-linker. Other quantitative approaches, such as SILAC, could also be used. Distinct signal patterns in the MS^1 spectra derived from isotope labelling may also increase the specificity of identification of cross-linked peptides particularly in complex systems. The potential of obtaining dynamics in structural studies of high-order, large protein complexes, or even at the proteome level, using 3D proteomics makes this methodology an especially valuable tool. However, to achieve such applications, an automated computational tool for cross-link quantitation is required. Although none is available currently, integration of the established quantitation software to cross-linked peptides search algorithms would greatly increase progress in this area.

Chapter 6

STRUCTURAL ANALYSIS OF TAGGED PROTEIN COMPLEXES BY 3D PROTEOMICS

6.1 Summary

In this chapter, I present the application of 3D proteomics on the structural analysis of endogenous tagged protein complexes. Single-step purified complexes are cross-linked and digested directly on the affinity beads used for isolation, providing increased sensitivity through minimized sample handling. Charge-based cross-linked peptide enrichment was applied to further improve detection of cross-linked peptides. The occurrence of cross-links between complexes was monitored by a SILAC-based control. Cross-links observed for low micro-gram amounts of starting material provided insights into structural organization of *S. cerevisiae* Mad1-Mad2 complex and Ndc80 complex.

6.2 Introduction

3D proteomics started to show its ability to reveal organization of multi-protein complexes after a lengthy technical development process (as introduced and demonstrated in previous chapters). However, the studies thus far had been carried out on highly purified protein complexes only (Sinz, 2006; Maiolica *et al.*, 2007; Bohn *et al.*, 2010; Chen *et al.*, 2010; Leitner *et al.*, 2010). Will this preference for high purity material become a restriction on the application of 3D proteomics? Can 3D proteomics become an easy and general tool for thousands of molecular biologists to gain some level of structural information about their protein complexes? In this work, I explored the possibility of obtaining structural information from low micro-gram amounts of single-step affinity-purified endogenous protein complexes.

The use of protein tags for the affinity isolation and identification of interaction partners has become a central tool for studying the molecular details of cellular processes (Terpe, 2003). Affinity purification combined with mass spectrometry (AP-MS) has been used to characterize many protein complexes and large protein-interaction networks (reviewed by (Gingras *et al.*, 2007)). With the AP-MS method, protein complexes are isolated from a cell lysate by affinity purification, and the components of these protein complexes are then identified by MS. The protein list generated by the AP-MS technique often includes purification background and will not reveal the relationship between proteins in the list. Application of chemical cross-linking to the purified complexes can capture the physical proximity between interacting proteins, and identification of cross-linked peptides will reveal low resolution topology of purified proteins and an interaction network between these proteins.

However, the analysis of single-step purified endogenous protein complexes, in contrast to highly purified material, is complicated by the limited amount of protein that can be isolated, and the high background of the purification. This further compromises the

already inefficient detection of cross-linked peptides. To improve the sensitivity of analysis, an on-beads process was developed to reduce sample loss; and charge-based cross-linked peptide enrichment was applied to enhance the detection of cross-linked peptides. This customized workflow was applied to the affinity-purified *S. cerevisiae* endogenous Mad1-Mad2 complex and Ndc80 complex.

6.3 Cross-linking analysis of tagged endogenous protein complexes

6.3.1 'On-beads' cross-linking and digestion procedure

A customized 3D proteomics analysis procedure was developed using on-beads processing (Figure 6.1). The tagged protein complex is captured by affinity beads from a cell lysate. After the wash step, the buffer is exchanged to the cross-linking buffer, and the complex is cross-linked on-beads. This approach places no restriction on the choice of cross-linker. The composition of cross-linking buffer can vary according to the protein complex and purification protocol. The basic requirements for cross-linking buffers are 1) within optimal pH range for cross-linking reaction; 2) does not contain components that can interfere with the cross-linking reaction. An example is Tris buffer, which can react with amine specific cross-linkers and will significant reduce the yield of cross-links on proteins. Fulfilling the above two requirements, the cross-linking buffer should ensure the native structure of the target protein complex. After the cross-linking reaction, the protein complex is proteolytically digested and peptides are released from the affinity beads. The peptide mixture is fractionated to enrich for cross-linked peptides using the SCX-Stage-Tip method, which can handle low micro-gram amounts of material. The fractionated sample is then analyzed by LC-MS/MS. Charge selection of precursor ions is employed to direct the fragmentation onto the higher charged (>2+) ions by excluding the singly and doubly charge ions for fragmentation.

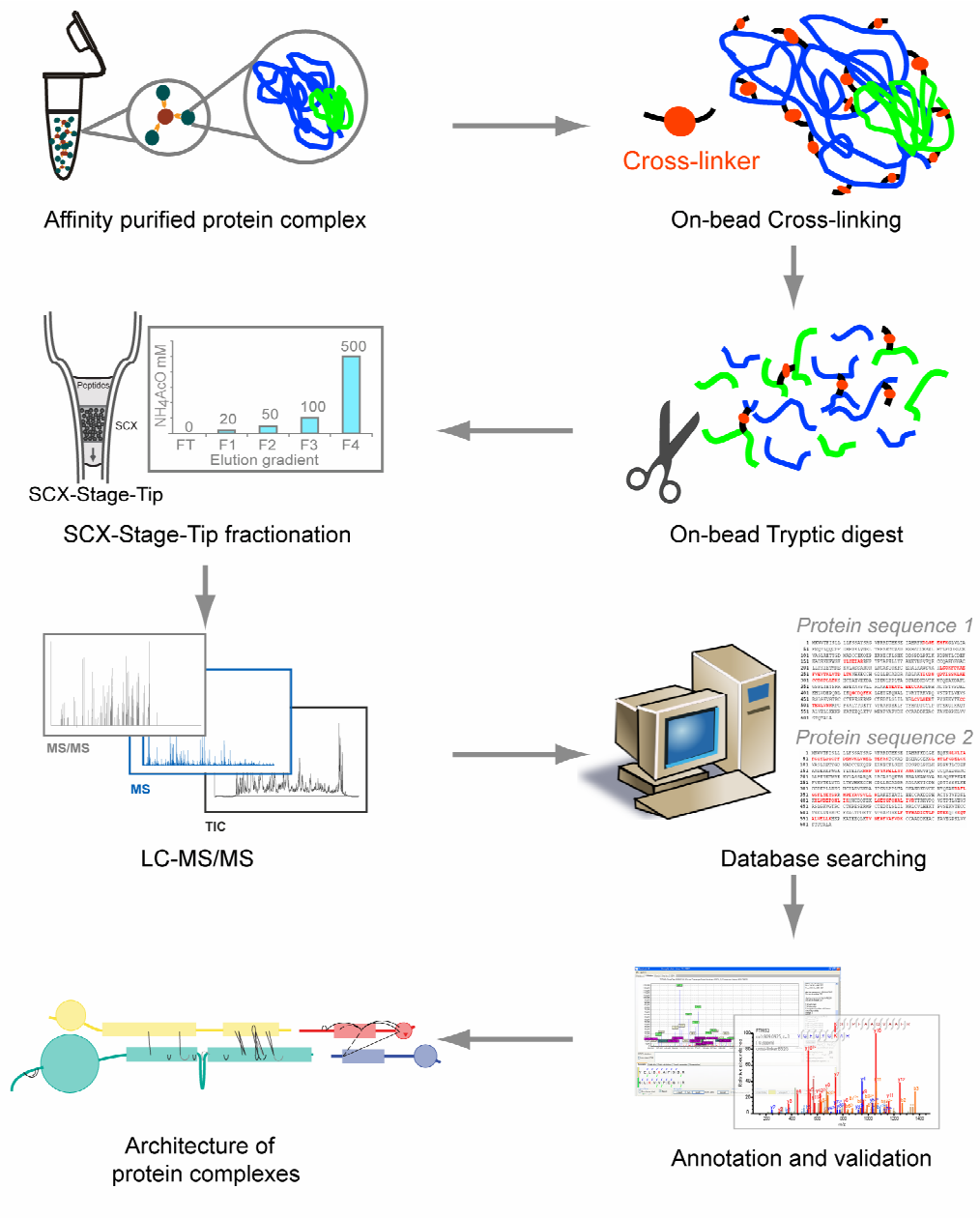


Figure 6.1 - Workflow of the 'on-beads' process for 3D proteomics analysis

The affinity purified protein complex is cross-linked on affinity beads, followed by 'on-beads' digestion. The peptide mixture is then fractionated using SCX-Stage-tip with a salt gradient for enrichment of cross-linked peptides. The peptide fractions are then analyzed by LC-MS/MS. The cross-linked peptides are identified by searching the LC-MS/MS data against a database containing protein sequences. The identified cross-linked peptide spectra are manually validated. The observed cross-links reveal architecture of the purified complex.

'On-beads' treatment has a number of advantages:

1) sample loss due to sample handling and exposure to surfaces is minimized;
2) experimental time is minimized, helping to maintain the integrity of the complex;
3) buffer exchange can be easily implemented;
4) existing isolation protocols need minimal adaptation.

The database search for cross-linked peptides can be performed on two levels. For well defined protein complexes, searching against protein sequences of the known components of a complex is sufficient to deliver a list of identified cross-linked peptides that can reveal the organization of the target complex, in terms of protein folding and protein-protein interactions (Maiolica *et al.*, 2007). A database containing only component protein sequences leads to limited search space and computational complexity. However, one-step affinity-purified complexes can be accompanied by a rather complex purification background due to specific or non-specific binding to target protein complex components or the affinity matrix. The composition of an affinity-purified complex sample can be determined through standard shotgun proteomics analysis procedures. Given that cross-linking analysis captures close proximity between proteins, searching against sequences of proteins identified in the purified complex sample can potentially reveal unknown specific interactions and discover new complex components. This process is tempting for studies of less defined protein complexes. However to achieve successful database searches at this comprehensive level, a search algorithm that can handle relatively large databases is required.

To evaluate the entire approach, the on-beads cross-linking analysis was applied to two endogenously TAP-tagged complexes from *S. cerevisiae*, the Mad1-Mad2 complex and the Ndc80 complex. Mad1-Mad2 (~220kDa) is a tetrameric complex, containing two copies of Mad1 and two copies of Mad2 (Maiolica *et al.*, 2007). The Ndc80 complex (~180kDa)

forms from four proteins: Ndc80, Nuf2, Spc24 and Spc25 (Wei *et al.*, 2005). The complexes were purified in a single step from a cell lysate using procedures compatible with large-scale complex pull-down analyses (Sjaak van der Sar, unpublished). Each captured complex was incubated with the amine-reactive cross-linker BS^2G (Thermo Scientific) and then digested with trypsin. The supernatant was fractionated using SCX-StageTips and analyzed by LC-MS/MS. To evaluate the specificity of cross-linked peptide identification in this process, a 1:1 mixture of non-labelled BS^2G-d0 and its deuterated analogue BS^2G-d4 were used for cross-linking the Mad1-Mad2 complex.

6.3.2 SILAC control experiments

On-bead cross-linking will to some extent minimize artificial aggregation as a side reaction of cross-linking because protein complexes are immobilized. However, whether complexes will be completely isolated from each other depends on spacing of the affinity groups on beads and the dimension of complexes. In this study, both complexes were purified through C-terminal TAP-tag using immobilized IgG. Each IgG on Dynabeads is spaced about 4nm apart, assuming equal and complete binding. Keeping in mind that the IgG molecule is about 15nm long, the possibility of contacts between captured protein complexes cannot be excluded. Cross-links between complexes would provide incorrect structural information for the arrangement of protein components. To monitor the occurrence of cross-links between complexes, an additional control experiment was designed using SILAC labelling (Figure 6.2). In the control experiment, the protein complexes were purified from 1:1 lysine$^{13}C_6$-labelled (isotopically heavy) and non-labelled (isotopically light) *S. cerevisiae* cells. The purified complexes were cross-linked and analyzed following the procedure described in Figure 6.1.

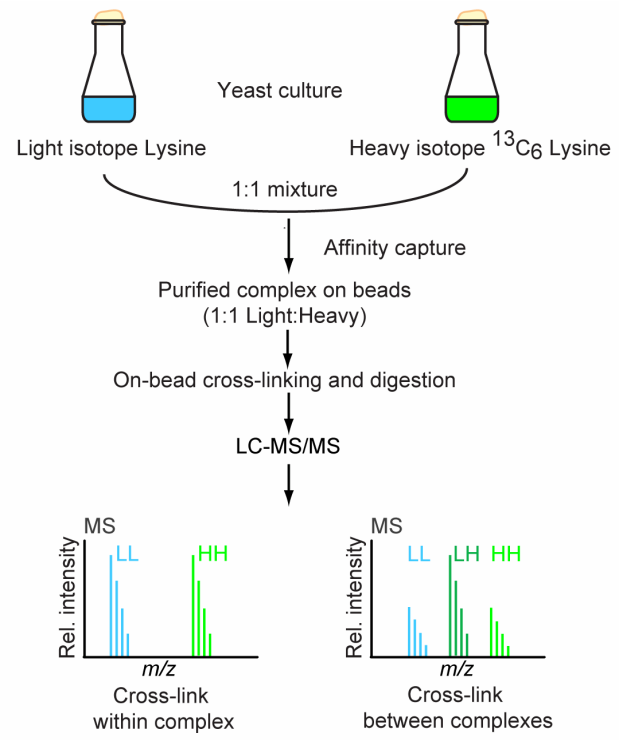


Figure 6.2 - Scheme of SILAC control experiment for monitoring the occurrence of inter-complex cross-links.

Only non-labelled cross-linker BS^2G-d0 was used to cross-link SILAC labelled material. For cross-links occurring within complexes, cross-linked peptides will appear as doublet signals in MS^1 spectra showing only light-light (LL) or only heavy-heavy (HH) peptide combinations. If a cross-link is formed due to the proximity between complexes, the cross-linked peptide(s) would have light-light, light-heavy, heavy-light, and heavy-heavy combinations. The distribution of these combinations will be reflected in the MS signals of cross-linked peptides as triplets with a pattern of 1:2:1 (LL:LH,HL:HH) (Figure 6.2). Additionally, SILAC signals of cross-linked peptides can also provide evidence for identification specificity, which will be discussed in *6.4.2*.

6.4 Cross-links observed from low microgram amounts of endogenous protein complexes

6.4.1 Composition of purified tagged protein complex samples

Parallel to the cross-linked peptides identification, the LC-MS/MS data of cross-linked complexes were also analyzed using standard proteomics database searching tools, to identify linear peptides and cross-linker modified peptides. This process gave rise to a list of identified proteins in the purified complex samples and provided an overall view on the purification background of sample preparation, and sample complexity for the cross-linking analysis.

There were six samples involved in this study: 12 μg of Mad1-Mad2 complex was divided into 4 μg and 8 μg samples after cross-linking for the subsequent analysis; 3.5 μg Mad1-Mad2 complex purified from 1:1 mixture of SILAC labelled and non-labelled cells (with <95% SILAC incorporation rate); 5 μg Mad1-Mad2 complex purified from 1:1 mixture of SILAC labelled and non-labelled cells, where the SILAC labelled cells had an improved $^{13}C_6$ incorporation rate (~98%); 9 μg of Ndc80 complex; 5 μg Ndc80 complex purified from 1:1 mixture of SILAC labelled and non-labelled cells. The analysis of linear and modified

peptides was applied to 4 sample preparations, named as Mad1-Mad2 (data from the 4 μg sample and the 8 μg samples were combined), Mad1-Mad2 SILAC (the 5 μg complex with ~98% incorporation rate for SILAC labelled material was used), Ndc80 and Ndc80 SILAC (Table 6.1). These data were searched against the *Saccharomyces* Genome Database using Mascot, and the final identification lists were generated using MaxQuant software with less than 1% false discovery rate (FDR) at both peptide and protein level. In the protein list, every protein was returned with an intensity that is estimated from the MS signal intensity of identified peptides from that protein. Even though these intensity values are not accurate quantification, they provide clues as to the abundance distribution among identified proteins. In both complex preparations, the components of target complexes were always ranked at the top in terms of intensity. There were 193 proteins identified from the Mad1-Mad2 sample and 61% of the summed intensity of all 193 *S. cerevisiae* proteins identified was assigned to Mad1 and Mad2. In the SILAC labelled Mad1-Mad2 sample, 67 proteins were identified and 11 of them were quantified by MaxQuant with at least two non-modified peptides detected with SILAC pair signals, which indicated high specificity of protein identification. These 11 proteins contributed 97% of the total intensity of all identified *S. cerevisiae* proteins and 84% were assigned to the Mad1-Mad2 complex. In the Ndc80 sample, there were 387 *S. cerevisiae* proteins identified, while in the SILAC labelled sample 480 proteins were identified with 154 quantified. The ten most intense proteins contributed 65% of total intensity. For both complexes, except for the complex component proteins, the top ten proteins lists were not consistent between the SILAC and non SILAC samples which suggested that these proteins might be co-purified due to non-specific effects. For safety, the subsequent database search for cross-links was not only conducted against a database containing only the sequences of the tagged complexes, but also against databases that

Table 6.1 - Composition of affinity-purified protein complex samples

Sample	Identified protein	Protein intensity distribution [%]		Identified IgG peptides	IgG sequence coverage [%]
		Target complex[1]	Ten most intense [2]		
Mad1-Mad2	193	61	88	6	27.2
Mad1-Mad2 SILAC	67 (11)[3]	84	96	9	22.9
Ndc80	378	49	65	10	23.8
Ndc80 SILAC	480 (154)	44	64	10	30.1

[1] Percentage of summed intensity of all protein components in the target complexes relative to the total intensity of all identified proteins in the sample.

[2] Percentage of summed intensity of the ten most intense identified proteins relative to the total intensity of all identified proteins in the sample.

[3] In the SILAC labelled samples, the number of proteins quantified by MaxQuant are listed in parentheses.

contained the ten most intense *S. cerevisiae* proteins identified from the purified complex samples (Table S5 and Table S6).

The identified linear peptides also reflected on-beads digestion efficiency. From the purified samples, the target complex proteins were identified with 60% sequence coverage on average. In the database search, the maximum missed cleavages allowed was set to 6, whereas none of the identified peptides was observed with more than 2 missed cleavages, and the majority of peptides had none or only one. Hence, the on-beads digestion was efficient. One major concern about the on-beads digestion is that in the sample, IgG is more abundant than any other protein. Therefore large amounts of IgG peptides could expand the dynamic range of the peptide mixture, and affect the detection of low abundant species, such as cross-linked peptides. Interestingly, there were no more than 10 IgG peptides identified from all samples and these peptides covered about 20% of the whole IgG sequence. The intensity of identified IgG peptides was similar to the peptides identified from targeted protein complexes. This could be partially attributed to the fact that in the on-beads digestion process, the reduction and alkylation treatment for cysteine residues was shifted after proteolytic cleavage by trypsin. Therefore, the highly compact structure of IgG was maintained and to some extent prevented complete digestion of IgG molecules. In this particular study, the reduction and alkylation treatment was skipped since all together cysteines make up less than 1% of the target complex proteins.

6.4.2 Identification of cross-linked peptides from affinity-purified complex samples

The database searches for cross-linked peptides were conducted at two levels. The LC-MS/MS data of cross-linked samples was searched against sequences of the known complex components first. The identified cross-linked peptide spectra were all validated manually following the criteria described in *3.5.1*. There were 316 cross-linked peptide spectra identified for the Mad1-Mad2 complex which gave rise to 50 unique cross-links. 32 of them were supported by at least one high confidence fragmentation spectrum and were validated as high confidence cross-links. The cross-links discussed here were generated from four different samples: Mad1-Mad2 4 μg sample, Mad1-Mad2 8 μg sample, 3.5 μg SILAC Mad1-Mad2 sample (<95% SILAC incorporation) and 5 μg SILAC Mad1-Mad2 sample (>98% SILAC incorporation). Due to the use of isotope labelled cross-linker and SILAC labelled samples, the identified cross-linked peptides were further validated independent of the fragmentation spectra. In non-SILAC samples, proteins were cross-linked with a 1:1 mixture of BS^2G-d0 and BS^2G-d4 cross-linkers. In MS^1 spectra all cross-linker containing species should show as doublets with 4 Da difference (Figure 6.3A). While in SILAC labelled samples, the *m/z* distance between SILAC doublet signals in the MS spectrum implies the number of lysine residues in the detected species, hence the number of lysine residues in the matched cross-linked peptide sequence should agree with the number expected from SILAC signals (Figure 6.3B). All 50 cross-links from the Mad1-Mad2 complex, including the low confidence ones, indeed conformed to the expected isotope shift. For the Ndc80 complex, 101 cross-linked peptide spectra were identified and validated from the 9 μg non-SILAC sample and the 5 μg SILAC sample. These identified spectra gave rise to 35 unique cross-links, among which 25 were supported by high confidence fragmentation spectra. Six cross-links that were only identified by low confidence fragmentation spectra were undetected, hence lacking SILAC support, and were discarded in the structural analysis.

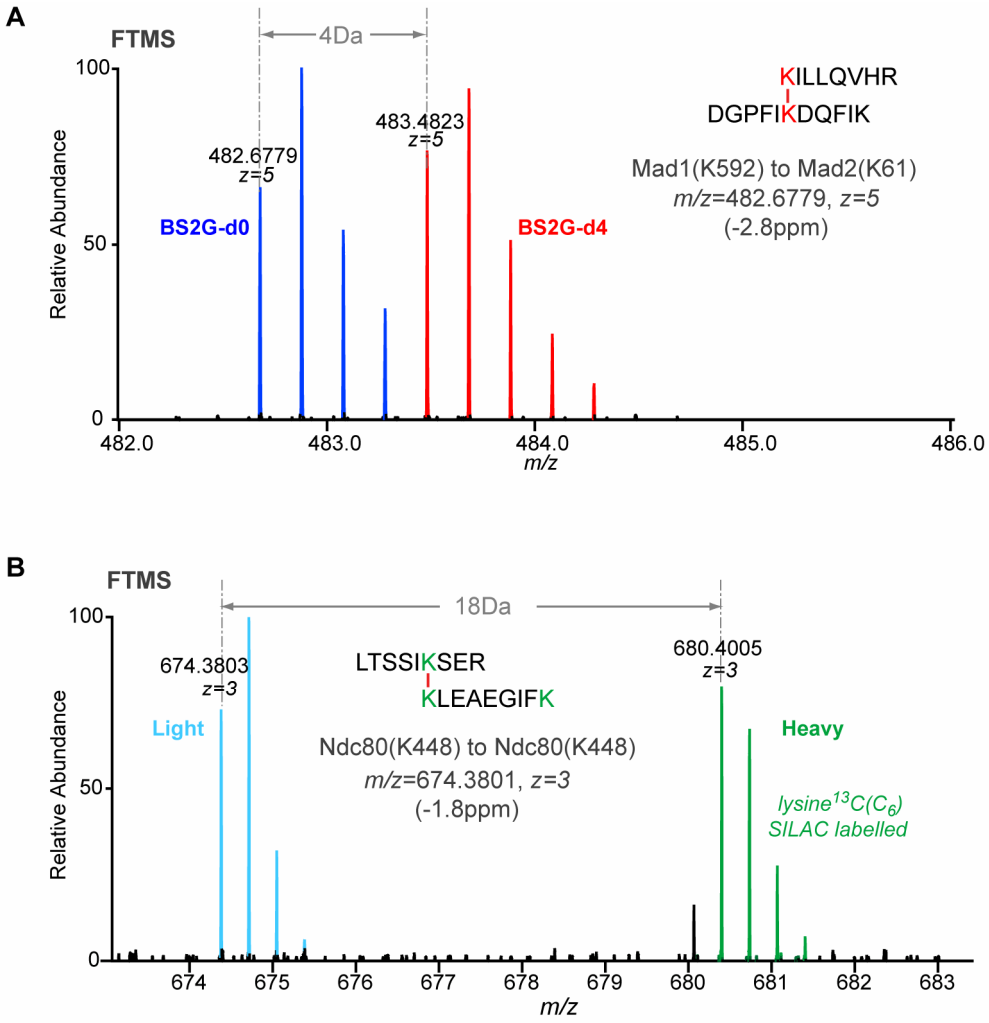


Figure 6.3 - Validation of cross-linked peptide identification in MS1 spectra

A. MS spectrum of a cross-linked peptide that was cross-linked with 1:1 mixture of BS^2G-d0 and BS2G-d4 cross-linkers. 4 Da difference between the doublet signals indicated the detected species contains one cross-linker.

B. MS spectrum of a cross-linked peptide LTSSIK(xl)SER-K(xl)LEAEGIFK observed from the lysine $^{13}C_6$ SILAC labelled sample. 18 Da distance between SILAC doublet signals is in agreement with 3 lysine residues in the identified peptide sequences.

The non-SILAC data of two complexes were also searched against the sequences of the ten most intense proteins that were identified from their purifications. This did not result in any additional identification of cross-linked peptides that passed manual validation. Consequently, there was no specific interaction between the target complexes and co-purified proteins detected by cross-linking analysis in this study.

The SILAC signal patterns of all identified cross-linked peptides were checked to monitor the occurrence of inter-complex cross-links. The MS signals of cross-linked peptides that were not identified in SILAC labelled samples were looked up in the raw data, based on the location in SCX Stage-Tip fractions, high accuracy mass (<6 ppm) and retention time. However, this process did not work for all cases. In summary, 39 of 50 cross-links from Mad1-Mad2 complex, and 25 of 35 cross-links from Ndc80 complex were confirmed to be cross-links within complex. There was no SILAC evidence for the rest of the cross-links. There was no cross-linked peptide detected with triplet SILAC signal. This supports the notion that cross-linking on-beads does not lead to protein aggregation.

It has been shown in Chapter 3 that the number of identified cross-linked peptide MS^2 spectra and unique cross-links were closely related to the amount of material analyzed. In order to figure out how sample amount affected the identification of cross-links at this low micro-gram range, 12 μg of cross-linked Mad1-Mad2 complex sample was divided into a 4 μg sample and an 8 μg sample. These two samples were analyzed in parallel through SCX-StageTip fractionation, LC-MS/MS and database searching. Here I compare the cross-link data from these two samples on three levels: identified spectra, cross-linked peptides, and unique cross-links (Table 6.2). These two datasets showed high reproducibility for relatively abundant cross-links. Increased sample amount gave rise to more identified spectra with better quality and new cross-linkages. However the detection of low abundance cross-links was not solely related to the sample amount

Table 6.2 - Influence of sample amount on cross-linking detection

	Sample	Total	Common	Unique
spectra	**4 µg**	**78 (34)** [1]	**67 (27)**	**11 (7)**
	8 µg	**128 (59)**	**103 (51)**	**25 (8)**
cross-linked peptides	**4 µg**	**43 (21)**	**33 (14)**	**10 (7)**
	8 µg	**52 (26)**	**33 (19)**	**19 (7)**
cross-linkages	**4 µg**	**35 (17)**	**26 (11)**	**9 (6)**
	8 µg	**40 (18)**	**26 (11)**	**14 (7)**

[1] High confidence observations are shown in parentheses. High confidence cross-linked peptides and high confidence cross-linkages are supported by at least one high confidence fragmentation spectrum.

and these species contribute most to the variation between datasets. 67 (85%) of 78 cross-linked peptide spectra identified in the 4 μg sample were generated from peptide combinations that were also identified in the 8 μg sample. These cross-linked spectra corresponded to 33 cross-linked peptides and 26 unique linkages. The increase in sample amount resulted in the average spectra number *per* linkage increasing from 2.5 to 4. The larger sample amount also resulted in an 88% increase for high confidence spectra for these cross-links. However, this increase did not cause any confidence change for corresponding linkages. On the contrary, the small variation between these two samples at spectra level was amplified at the linkage level. 25 unique spectra in the 8 μg sample brought in 14 new cross-linkages. Surprisingly, 9 cross-linkages that were detected in the 4 μg sample were no longer identified when the sample amount was increased. A noteworthy point is that 90% of these 9 cross-links were identified by a single spectrum, which implied their low abundance in the sample. The visibility of these low abundance signals could also be affected by chromatography and timing for fragmentation *etc*. Apparently, the 2-fold increase in the sample amount did not counteract these effects. This analysis suggests that replica analysis can be more valuable for increasing the data than increasing the starting material is.

6.5 Organization of the Mad1-Mad2 complex

Mad1 and Mad2 are proteins involved in the spindle assembly checkpoint, which targets to the anaphase-promoting complex (APC) (Hardwick, 1998; Shah and Cleveland, 2000). APC inhibition requires a direct interaction between Mad2 and an APC positive regulator Cdc20 (Li *et al.*, 1997; Fang *et al.*, 1998; Hwang *et al.*, 1998; Kallio *et al.*, 1998; Kim *et al.*, 1998). Mad1 has been shown to be a competitive inhibitor of the Mad2-Cdc20 complex and the Mad1-Mad2 complex was proposed to act as a regulated gate to control Mad2 release for Cdc20 binding (Sironi, Mapelli *et al.* 2002). The Mad1-Mad2 complex is a 220 kDa heterodimeric tetramer consisting of two copies of Mad1 and Mad2. An uninterrupted

coiled-coil was predicted for the N-terminal segment of human Mad1 (Berger *et al.*, 1995). The crystal structure of human Mad2 in complex with the Mad1 C-terminal segment, revealed an interaction region between Mad1 and Mad2 (Sironi *et al.*, 2002). The structure of yeast Mad1-Mad2 complex remains unclear. In this study, I applied 3D proteomics analysis to obtain an insight into the structure of the endogenous Mad1-Mad2 complex from *S. cerevisiae*. The Mad1-Mad2 complex was cross-linked after being purified through tagged Mad1, and in total 50 cross-links were identified (Table S7).

Eight of these cross-links connected Mad1 peptides with identical or overlapping sequences and six of them were confirmed to be within the complex by SILAC signals (Figure 6.4B). This proved that there are two copies of Mad1 in the complex and these linkages cross-linked between them. The homo-dimeric cross-links span the whole Mad1 sequence, from residues 95 to 600, which confirmed the predicted uninterrupted parallel coiled-coil structure along the Mad1 chain (Lupas *et al.*, 1991) (Figure 6.5A).

At the end of the coiled-coil region, there were two cross-links that were observed between Mad1 and Mad2 and these have been validated as intra-complex cross-links. Mapping these cross-links to the *S.cerevisiae* homology model of the human Mad1-Mad2 interacting region (Paul McLaughlin, unpublished), $K61^{Mad2}$ was cross-linked to $K592^{Mad1}$ which are positioned beside the Mad2 binding motif (580-591) (Luo *et al.*, 2002) (Figure6.5B). Another cross-link was from $K61^{Mad2}$ to $K649^{Mad1}$, a residue outside the homology model. Interestingly, $K649^{Mad1}$ was also cross-linked to $K592^{Mad1}$. The triangle linkage between these three residues indicated a spatial proximity between $K649^{Mad1}$ and the crystallized Mad1-Mad2 interacting region (Figure 6.5B). To satisfy this proximity, a fold back of the Mad1 molecule at the

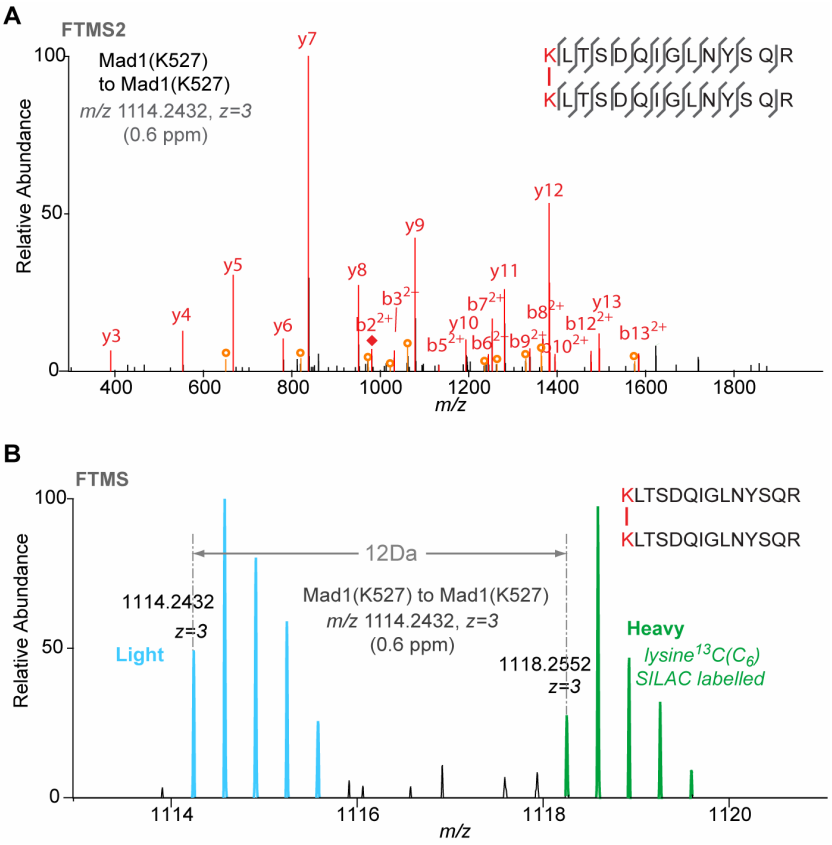


Figure 6.4 - Spectra of cross-links between Mad1 molecules in the Mad1-Mad2 complex

A. High resolution fragmentation spectrum of cross-linked peptides K(xl)LTSDQIGLNYSQR-K(xl)LTSDQIGLNYSQR. It corresponds to linkage between $K527^{Mad1}$ to $K527^{Mad1}$. The overlaid sequence in two cross-linked peptides indicates that this cross-link occurred between Mad1 molecules. B. The MS spectrum of the same cross-linked peptide (as shown in A) in the SILAC control experiment. The doublet signals indicated that this cross-linkage occurred within the Mad1-Mad2 complex.

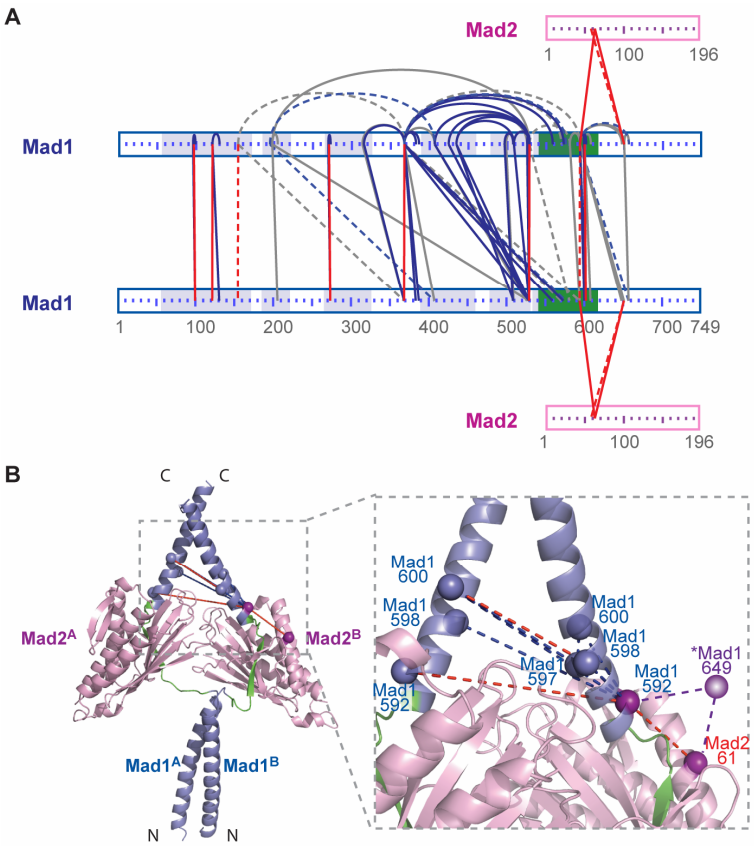


Figure 6.5 - Organization of the *S. cerevisiae* Mad1-Mad2 complex

A. Cross-link map for the Mad1-Mad2 complex. Mad1 and Mad2 are shown as bars. In Mad1, faint blue shades indicate coiled-coil domain predicted by COILS (Lupas *et al.*, 1991). Green shades indicate the sequence included in the *S. cerevisiae* homology model of the Mad1-Mad2 interaction region. High confidence cross-links occur between proteins: red; high confidence cross-links that can be either between or within Mad1 molecules: blue; low confidence cross-links: grey; cross-links within complex supported by SILAC signals: continuous line; cross-links without SILAC evidence: dashed line.

B. Cross-links displayed in the homology model of yeast Mad1-Mad2 interacting region.

Mad1 in faint blue, Mad2 in pink, green highlights the Mad2 binding motif in Mad1 (sphere for C-α atom of linked lysine residues). Cross-links occur between proteins: red dashed line; cross-links can be either between proteins or within a protein: blue dashed line. Residue Mad1 K649 that is outside of the homology model but cross-linked to residues in the model is shown by a purple sphere.

C-terminal region is required. This observation matched to the predicted structure for human Mad1-Mad2 complex, where an intra-molecular anti-parallel coiled-coil at the C-terminal of Mad1 was speculated based on the crystal structure (Sironi *et al.*, 2002).

For the 31 intra-complex cross-links between Mad1 peptides, it is hard to distinguish whether they are within or between molecules. Even for the three links that can be displayed in the homology model, the proximity of two Mad1 molecules, indicated by observation of two cross-links between them, suggested that these linkages could occur both within and between Mad1 molecules (Figure 6.5B). Nevertheless, some cross-links of Mad1 peptides that bridged long distance residues in sequence, no matter whether they are within Mad1 or between Mad1 molecules, indicates the possibility of a folded Mad1 state.

In summary, 3D proteomics data indicated that two Mad1 molecules in the Mad1-Mad2 complex form a parallel coiled-coil structure. Cross-links between Mad1 and Mad2 positioned Mad2 to the Mad2 binding motif in Mad1, and supported the speculation of a short intra-molecule anti-parallel coiled-coil structure at the C-terminal of Mad1. Moreover, the long distance cross-links suggested a possible folded state of Mad1.

6.6 Cross-link data revealed a conserved loop region in Ndc80.

The four protein Ndc80 complex forms from Ndc80, Nuf2, Spc24 and Spc25 with a 1:1:1:1 stoichiometry. It is an essential kinetochore component and plays a crucial role in proper chromosome alignment and segregation during mitosis (Ciferri *et al.*, 2007). The Ndc80 complex is conserved from yeast to humans, and Hec1, the human homologue of Ndc80, can substitute functionally for yeast Ndc80 (Zheng *et al.*, 1999). All four proteins in the complex are predicted to contain both a coiled-coil domain and a globular region. Scanning force microscopy and electron microscopy studies on the reconstituted human Ndc80 complex and its yeast homologue, indicated that the complex is ~57 nm long with an elongated dumbbell shape. Two sub-complexes form between Ndc80 and Nuf2 as well as

Spc24 and Spc25 through parallel heterodimeric coiled-coils and these two sub-complexes interact via C- and N-terminal portions of the Hec1-Nuf2 and Spc24-Spc25 coiled-coils. Two globular regions, that contain the N-terminal heads of Ndc80 and Nuf2 at one end and the globular C-terminal heads of Spc24 and Spc25 at the opposite end, form the heads of the dumbbell (Ciferri *et al.*, 2005; Wei *et al.*, 2005). 3D proteomics analysis of recombinant human Ndc80 complex has already described the internal architecture of the complex (Maiolica *et al.*, 2007) and assisted the crystallographic structure determination of a truncated version of the complex (Ciferri *et al.*, 2005). For the yeast complex structure, only the C-terminal heads of Spc24-Spc25 and the globular domain of Ndc80 have been determined (Wei *et al.*, 2006; Wei *et al.*, 2007).

In this study I applied 3D proteomics analysis to study the internal architecture of the endogenous Ndc80 complex from *S. cerevisiae*. The Ndc80 complex was purified *via* C-terminal tagged Ndc80. In total, 35 unique cross-links were identified. Except for six that were only identified with low confidence MS^2 spectra, 29 cross-links were used for structural analysis and named X1 to X29 for the convenience of description (Table S8). 16 cross-links were detected within proteins from Ndc80, Nuf2 and Spc24. 13 cross-links were detected between proteins, 11 in the Ndc80-Nuf2 sub-complex and two in the Spc24-Spc25 sub-complex. However, the interaction between these two sub-complexes was not reflected by cross-links. This is in contrast to the human Ndc80 complex analysis where the two sub-complexes were joined by several cross-links (Maiolica *et al.*, 2007).

In the Ndc80-Nuf2 complex, eight cross-links in Ndc80 (X2-X10) and one in Nuf2 (X11) were observed within the predicted coiled-coil domains. The maximum distance between bridged residues in these cross-links was 11 amino acids. Assuming an α-helix structure and calculating from Cα, this distance along sequence indicates a 16.5 Å shift along the helix, taking in mind the angle between residues, the spatial distance approaches the cross-linking limit of cross-linker BS^2G (7.7 Å spacer and 2 x 6 Å lysine side chains). This

observation supported the predicted coiled-coil structures (Lupas *et al.*, 1991). Moreover, the eight cross-links in Ndc80 distributed from residues 404 to 648 in the sequence, indicating an elongated Ndc80 molecule and agreed with observations made by rotary shadowing electron microscopy (Wei *et al.*, 2005).

Eleven cross-links (X17-X26 and X29) observed between Ndc80 and Nuf2 were all distributed along the predicted coiled-coil region. In Nuf2 these cross-links were located between residue 220 and residue 415; and in Ndc80 between residue 377 and residue 627 (Figure 6.6). It is puzzling that the span of these cross-links in Ndc80 was 55 residues more than that of Nuf2. Looking in more detail, these cross-links can be divided into three clusters and two individual cross-links, according to their location in the sequence. From the C-terminal of the coiled-coil region, a cluster around 627^{Ndc80} (X25, X26), the cluster around 600^{Ndc80} (X22-X24, X29) and a cluster around 580^{Ndc80} (X19-X21) showed consistent distances in both Nuf2 and Ndc80. However, between the cluster around 580^{Ndc80} (X19-X21) and the cross-link at 425^{Ndc80}(X18), the distance is different in Nuf2 and Ndc80 (~155 and ~90 residues). Further towards the N-terminus, between X18 and X17 at 377^{Ndc80}, the distance in Ndc80 and Nuf2 are back to near equal again. The inconsistent distance in Ndc80 and Nuf2 between cross-links X18 and X19 suggested a non-coiled coil interruption in the Ndc80 chain coiled-coil region with about 60 residues between 425^{Ndc80} and 577^{Ndc80}. This inference is supported by a 65 amino acid probability drop between 455^{Ndc80} and 520^{Ndc80} in the coiled-coil region prediction for Ndc80. A similar non-coiled-coil segment in the Ndc80 coiled-coil region was also reported in human Ndc80 complex, revealed by 3D proteomics analysis (Maiolica *et al.*, 2007). The conservation of this insertion in the Ndc80 coiled-coil region implied functional involvement of this structural feature, which needs to be further addressed. Subtraction of the 60 amino acid insertion from the Ndc80 sequence meant the distance between cross-links observed between 220^{Nuf2} and 415^{Nuf2}, 337^{Ndc80} and 627^{Ndc80} represented a consistent set. This confirmed the predicted coiled-coil structure in

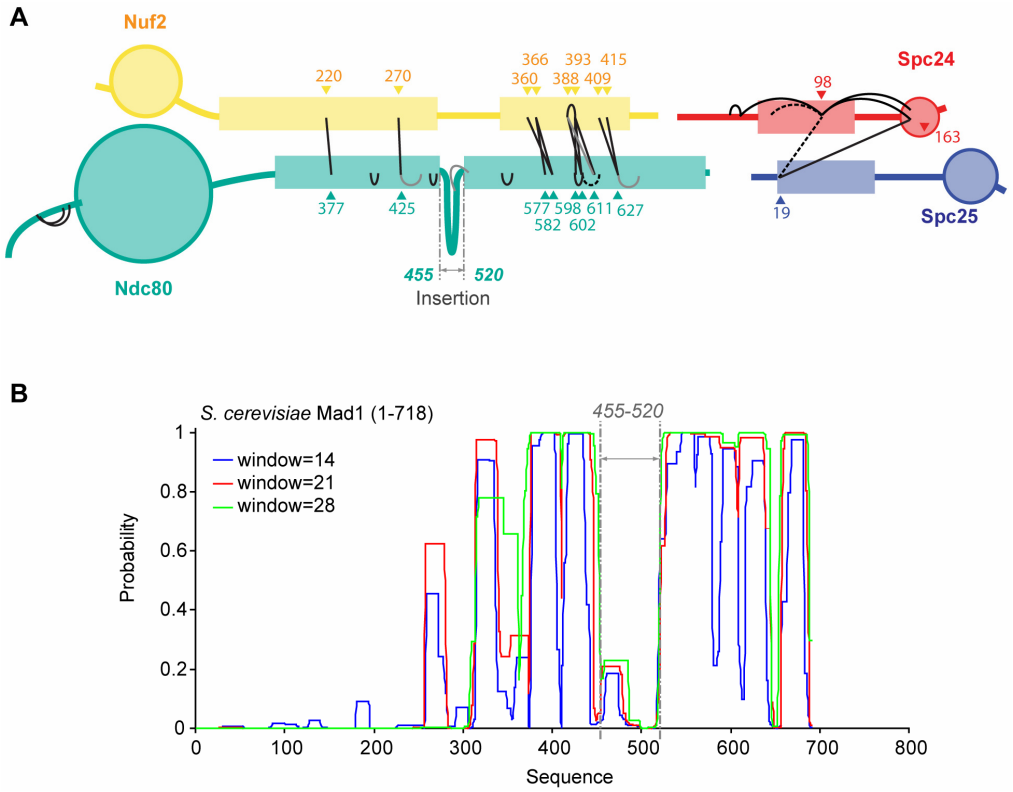


Figure 6.6 - Internal architecture of the *S. cerevisiae* Ndc80 complex

A. Ndc80 complex model built based on cross-linking data and previous electron microscopy observations (Wei, Sorger *et al.* 2005). High confidence cross-links: black; Low confidence cross-links: gray; Cross-links within complex with SILAC signal confirmation: lines; Cross-links without SILAC evidence: dashed lines.

B. Coiled-coil prediction of *S. cerevisiae* Ndc80 from COILS server (Lupas *et al.*, 1991).

this region and suggested a minimal range for the coiled-coil region in Ndc80 and Nuf2 with experimental evidence.

In the Spc24-Spc25 sub-complex, there were only two cross-links between Spc24 and Spc25. The Spc25 N-terminal residue 19^{Spc25} was cross-linked to 98^{Spc24} in the middle of Spc24 and 163^{Spc24} in the C-terminal globular region of Spc24. These cross-links indicated a Spc24-Spc25 arrangement that conflicts with the structure revealed in the previous studies on both the yeast and human Ndc80 complex (Ciferri *et al.*, 2005; Wei *et al.*, 2005; Wei *et al.*, 2006; Maiolica *et al.*, 2007; Ciferri *et al.*, 2008). These studies suggested that Spc24 and Spc25 form an elongated parallel coiled-coil through their N-terminal regions and form an integrated globular domain *via* C-terminal regions. A cross-link between 98^{Spc24} and 163^{Spc24} suggested their spatial proximity. A cross-link between 98^{Spc24} and 205^{Spc24} further confirmed that 98^{Spc24} is close to the globular domain. Furthermore, two cross-links from 98^{Spc24} to 62^{Spc24} and 42^{Spc24} indicated that 98^{Spc24} was also close to the N-terminal of Spc24 (Figure 6.6). This series of long distance cross-links in Spc24 suggested a folded state of Spc24 which also explained the linkages between a Spc25 N-terminal residue and residues that are in the middle and C-terminal of Spc24. Given that Spc24 and Spc25 were co-purified by tagged Ndc80, it is reasonable to assume that the interaction between Spc24 and Spc25, as well as the interaction between the Spc24-Spc25 sub-complex and Ndc80-Nuf2 sub-complex were not broken down. Therefore, I hypothesize that the folding observed here occurred on top of Spc24-Spc25 dimerization structure and this folding, to some extent, could have prevented the coiled-coil region in the Spc24-Spc25 dimer as well as the interaction between the two sub-complexes from chemical cross-linking detection. Since there is no information on the conformation of the Ndc80 complex *in vivo*, it is not clear whether this folding is a natural occurrence, or induced artificially during purification and cross-linking, or even by the Ndc80 C-terminal tag, even if it did not affect cell growth.

In summary, 3D proteomics analysis on *S. cerevisiae* endogenous Ndc80 complex confirmed a predicted coiled-coil structure form between Ndc80 and Nuf2, and also revealed a ~60 amino acid non-coiled-coil interruption in the Ndc80 chain in the coiled-coil region, which is conserved in the human Ndc80 complex. A series of long distance cross-links in Spc24 suggested an unexpected folded structure in the Spc24-Sp25 sub-complex.

6.7 From AP-MS to AP-3DMS

Isolating protein complexes using protein tags and affinity purification has become a well established technique in molecular biology. In recent years, a combination of affinity purification and mass spectrometric identification of purified proteins (AP-MS) have greatly advanced the understanding of protein complex compositions (Gingras *et al.*, 2007). In this study, I demonstrated that coupling 3D proteomics with the AP-MS analytical pipe-line, not only identified the components of two tagged yeast protein complexes, but also revealed the spatial arrangement of the complexes. Cross-links can be detected and identified from low microgram amounts of single-step affinity purified protein complexes. Integrating 3D proteomics with AP-MS can provide additional structural insights into affinity purified protein complexes. Moreover, application of AP-3DMS in protein interaction network studies can be expected to provide further evidence on spatial contacts of interacting proteins.

Here I discuss the possibility of establishing AP-3DMS as a routine analytical procedure for studying affinity-purified protein samples. Experimentally, AP-3DMS requires an additional cross-linking step and enrichment for cross-linked peptides. Efficient integration of the analytical workflows of affinity purification and 3D proteomics, such as the on-beads procedure described in this chapter, can improve the sensitivity of analysis and reduce analysis time. Mass spectrometers that are frequently used for standard AP-MS analysis are normally compatible with 3DMS analysis (as listed in *1.4.1*). Computationally,

linear peptides can be analyzed using standard AP-MS procedures to identify the composition of protein complexes. However, cross-linked peptides need to be identified using specialized search algorithms or computational strategies. Currently, several search algorithms have been available for identification of cross-linked peptides from samples with complexity like affinity-purified complexes (Chapter 1). Therefore, there is no major technical difficulty in integrating 3D proteomics with AP-MS approaches. However, a scoring system that can distinguish true and false identifications still needs to be developed to allow for applications by researches with no mass spectrometric expertise.

Chapter 7

SUMMARY AND PERSPECTIVE

7.1 Summary

In this thesis I presented my work on improving the analytical workflow for 3D proteomics and developing applications using this advanced workflow.

Firstly, I presented an advanced analytical workflow for 3D proteomics that was developed and evaluated using a cross-linked peptide library. This cross-linked peptide library provided a large dataset of cross-linked peptides, which facilitated the development of a charge based enrichment strategy for cross-linked peptides and the optimization of experimental settings. Over one thousand high quality MS^2 spectra of cross-linked peptides led to insights into fragmentation behaviours of cross-linked peptides and fundamentally supported the development of a search algorithm for cross-linked peptides.

Using this workflow, 3D proteomics was applied to analyze the 530 kDa 12-subunit Pol II complex. The consistency of cross-link data and the X-ray crystallographic data validated 3D proteomics as a sensible tool for studying large multi-protein complexes. A subsequent study on the 670 kDa 15-subunit Pol II-TFIIF complex revealed interactions between Pol II and its general transcription factor TFIIF. Cross-links between TFIIF and Pol II positioned TFIIF on the surface of the Pol II core crystal structure and allowed for further understanding of the TFIIF functions in Pol II initiated transcription. Furthermore, comparison of the cross-link data obtained from the Pol II and the Pol II-TFIIF samples suggests that using 3D proteomics analysis it is possible to reveal the structural dynamics of protein complexes.

I further developed a quantitative 3D proteomics approach to study protein conformational changes. I introduced isotope labelling based quantitation into the 3D proteomics analytical workflow and applied this approach to study the conformational differences between the complement protein C3 and its active from C3b in solution. The results confirmed previous observations by crystallography, and proved the ability of quantitative 3D proteomics for detecting conformational differences. Moreover, the quantitative cross-link data revealed hydrolysis of C3 in both C3 and C3b samples. The architecture of hydrolyzed C3 was proposed based on quantified cross-links and the crystal structures of C3 and C3b. This application suggested that combining quantitative 3D proteomics data and static structures can be a new way to study dynamics of proteins and protein complexes.

In the end, 3D proteomics analysis was coupled to the analytical pipeline for affinity purification of protein complexes. An on-beads procedure was applied to increase the sensitivity of analysis. The results showed that cross-links can be detected and identified from low micro-gram amount of single step affinity-purified protein complexes in mixture with over a hundred co-purified proteins. These cross-links provided insights into the architecture of the *S. cerevisiae* endogenous Mad1-Mad2 complex and Ndc80 complex.

In summary, the advanced analytical workflow of 3D proteomics described in this thesis allowed for applications of 3D proteomics on large multi-protein complexes and protein complexes in a complex protein mixture background. The combination of 3D proteomics data and crystallographic data in modelling the structures of the Pol II-TFIIF complexes and C3 (H2O), demonstrated the potential of 3D proteomics in integrated structural analysis. Moreover, the combination of 3D proteomics with quantitative proteomics and affinity purification in the applications, exemplified the possibility of integrating 3D proteomics with other structural biology and proteomics techniques.

7.2 Perspective

As summarized above and discussed in *1.6*, in the past five years, the 3D proteomics field has advanced significantly. The overall improvement of analytical workflow, extensively reduced the impacts of technical limitations on the applications of 3D proteomics. A wide range of applications indicated an evolution from the proof of principle study, to the question driven investigation. As drafted in Figure 7.1, various applications of 3D proteomics and its versatile combination with other techniques are expected in the near future. However, to become a generally applied technique, two major technical breakthroughs are still needed for 3D proteomics. Firstly, although several current available enrichment approaches are able to improve the visibility of low abundance cross-linked peptides in mass spectrometric analysis, the efficiency is largely limited by sample complexity. Better cross-linking yield and efficient purification schemes are expected to principally enhance the detection of cross-linked peptides. Secondly, the lack of automated verification and validation tools for cross-linked peptide identification is still a big obstacle for large scale applications. Progress has been made by current attempts following the successful experiences for linear peptides in standard proteomics studies.

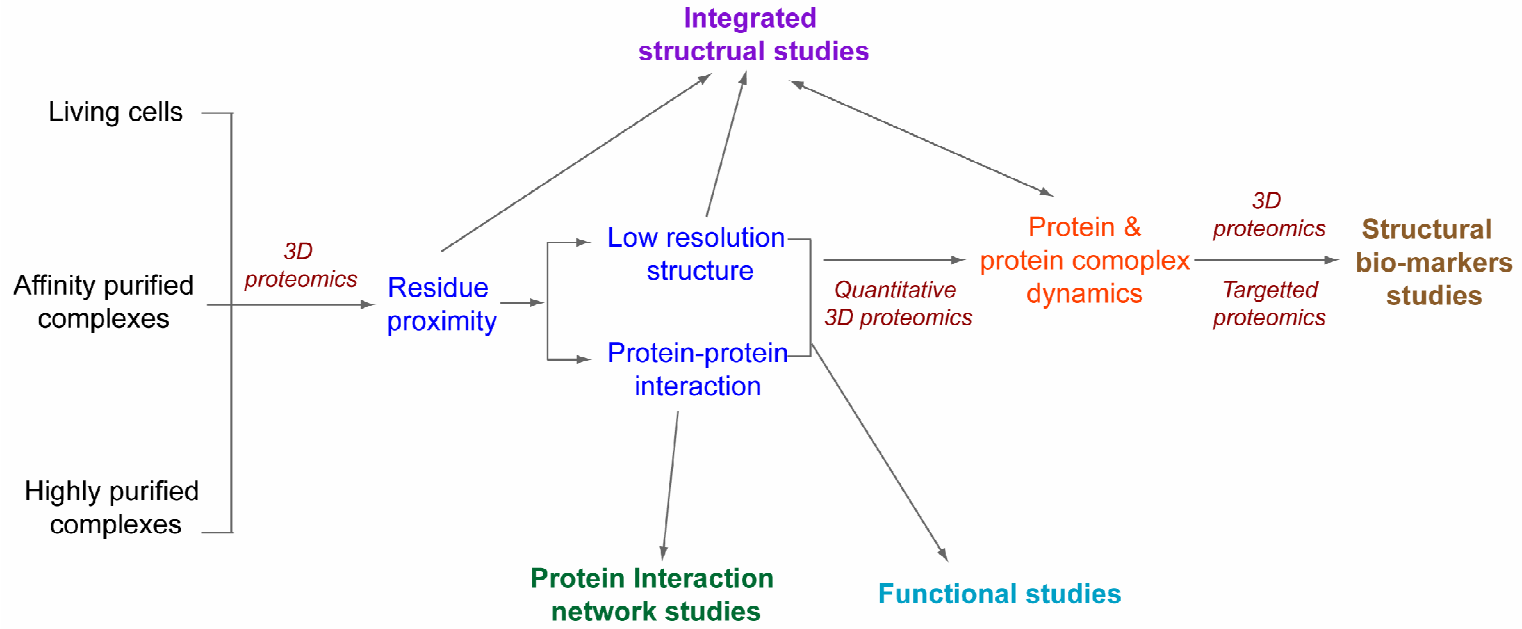


Figure 7.1 - Draft of expected versatile applications of 3D proteomics in near future.

APPENDIX

A.1 Observation of C3 contamination in the C3b sample

A.1.1 Detection of C3 contamination

A.1.1.1 Experimental procedure

A.1.1.1.1 Denaturing gel electrophoresis

5 pmol of C3 and C3b (Complement Technology, Inc) samples were separated on the NuPAGE 4-12% Bis-Tris gel using MES running buffer. The gel was fixed in 50% methanol, 5% acetic acid and the stained using colloidal blue kid follow the manufacture instruction (Invitrogen).

A.1.1.1.2 Mass spectrometric analysis

The C3b sample were cross-linked with cross-linker BS^3-d0 and BS^3-d4 (Thermo scientific), as described in *2.3.1*. 5pm of equal amount mixture of BS^3-d0 and BS^3-d4 cross-linked C3b sample was concentrated on the NuPAGE 4-12% Bis-Tris by just run the sample into the gel for about 5mm, then the gel was fixed in 50% methanol, 5% acetic acid and stained using colloidal blue kid. The protein was in-gel reduced/alkylated and digested using trypsin as described in *2.2.4*. After digestion, the peptides mixture was acetified using 0.1% TFA and cleaned up using C18-StageTips (Rappsilber *et al.*, 2003). Then the eluted peptides sample was prepared for mass spectrometric analysis follow the procedure (*2.1.3.1*).

The LC-MS/MS analysis was conducted as same as for the linear peptides sample in *2.1.3.2*. The MS^2 spectra peak list was generated from the raw data files using MaxQuant (version 1.0.11.2) Quant module (Cox and Mann, 2008) using default setting. Subsequently, the peak list was searched against SwissProt database using Mascot. The search parameters were set as listed in Table 2.3, except the hydrolyzed and ammonia reacted cross-linker (BS^3-d0 and BS^3-4) on lysine and protein N-termini were set as variable modifications in

addition. The returned peptide candidates with Mascot score above 25 were accepted as true matches. The protein candidates identified with more than 2 valid unique peptide matches were accepted as protein identifications.

A.1.1.2 Results

The only visible impurity in the C3b sample on the gel image was a band aligned with the 97 kDa marker and interestingly exhibited the same electrophoretic mobility as the α-chain of C3 (Figure A1). The mass spectrometric analysis identified 17 proteins from the C3b sample (Table A1). Although C3b shares all its sequence with C3 and they are both return as "C3 precursor" in database searches when searching against SwissProt, the identification of 5 peptides in the C3 unique C3a fragment (the ANA domain in structure) indicated the existence of C3 in the C3b sample (TableA2). All these five peptides were identified with Mascot score above 25. Moreover three of these five identified C3a peptides were modified by hydrolyzed and ammonia reacted cross-linkers. Since the C3b sample was a 1:1 mixture of BS^3-d0 and BS^3-d4 cross-linking products, the paired signals with 4 Da mass difference for cross-linker modified peptides were expected. The detection of these doublets signals for these three C3a peptides further validated them as true hits. Since mass spectrometric analysis is not quantitative, it is hard to judge the abundance of C3 contamination. However, according to the both the molecular weigh and the huge abundance deference to C3b hit (reflected on the number of identified peptide spectra) none of identified protein other than C3 could be responsible for the observed impurity band in the gel image. Based on the density of stain of gel bands, the abundance of the C3 contamination can be very roughly estimated to about one tenth of C3b.

A.1.2 Quantisation of C3 contamination

A.1.2.1 Experimental procedure

Both C3b and C3 samples were cross-linked with cross-linker BS^3-d0 and BS^3-d4 (Pierce, Thermo scientific), as described in *2.3.1*. Equal amount of BS3-d0 cross-linked C3 and BS3-d4 cross-linked C3b were mixed as forward labelled sample, while the BS3-d0 cross-linked C3b and BS^3-d4 cross-linked C3 were 1:1 mixed as reverse labelled sample. Both samples were analysis following the identical procedure as for the C3b sample in *A.1.1.2.* The quantitation was conducted for the identified peptides within C3 specific ANA domain (650-748) what were modified by hydrolyzed and ammonia reacted cross-linkers (BS^3-d0 and BS^3-d4). A C3 to C3b intensity ratio was read out for these peptides in the same way for quantifying the cross-linked peptides as described in *2.3.5*

A.1.2.2 Results

In total eight C3a peptides were quantified at MS^1 level (Figure A1.2). The C3 to C3b intensity ratio ranged from 4.9 to 23.4 with an average of 11.1. Therefore the abundance of C3 in C3b can be estimated to around 10% which agreed to the judgement from the gel image.

A.1.3 Discussion

The C3 hydrolysis analogue C3(H_2O) can arise spontaneously in the aqueous environment. C3 and C3(H_2O) are identical on sequence. Therefore the above analyses does not exclude the probability that the detected contamination signals in the C3b sample was C3(H_2O).

Table A1.1 - Identified C3a peptides from the C3b sample

Peptide sequence	Start residue	End residue	Identified spectra	BS3d4/d4 modification	Doublets signal
YPKELR	664	669	3	yes	yes
SVQLTEKR	650	657	6	yes	yes
FISLGEACKK	691	700	4	yes	yes
MDKVGKYPK	658	666	1	no	N/A
VFLDCCNYITELR	701	713	1	no	N/A

Table A1.2 - Proteins identified from the C3b sample using Mascot

Accession number	Protein description	Protein mass	Mascot score	Identified spectra
CO3_HUMAN	Complement C3 precursor	188569	62444	2824
K1C9_HUMAN	Keratin, type I cytoskeletal 9	62320	619	30
K1C10_HUMAN	Keratin, type I cytoskeletal 10	59703	468	12
ATPE_HAEIE	ATP synthase epsilon chain	15581	244	15
TRYP_PIG	Trypsin precursor	25078	135	11
ALBU_HUMAN	Serum albumin precursor	71317	126	16
SIA4A_PIG	CMP	40086	95	5
AMYB_HORSP	Beta-amylase precursor	59886	70	2
CASB_BOVIN	Beta-casein precursor	25148	66	2
MIAA_RHIME	tRNA delta(2)	33174	60	5
QUEF_DESPS	NADPH	31636	59	6
NCOA7_MOUSE	Nuclear receptor coactivator 7	106918	54	54
K1C14_HUMAN	Keratin, type I cytoskeletal 14	51875	49	2
TAF2_YEAST	Transcription initiation factor TFIID subunit 2	162851	45	8
SEPA_EMENI	Cytokinesis protein sepA	197577	45	15
DPOE_CANGA	DNA polymerase epsilon catalytic subunit A	257874	43	6
RL35_STRA1	50S ribosomal protein L35	7763	40	3

Table A1.3 - Quantitation of cross-linker modified C3a peptides

Peptide sequence	Start residue	End residue	Modification	Modified residue	Labelling	C3/C3b ratio	Identified spectra
SVQLTEKR	672	679	BS3-d0 (NH2) (K)	656	C3-BS3d0/ C3b-BS3d4	6.9	1
SVQLTEKR	672	679	BS3-d0 (OH) (K)	656	C3-BS3d0/ C3b-BS3d4	5	1
SVQLTEKR	672	679	BS3-d4 (NH3) (K)	656	C3-BS3d4/ C3b-BS3d0	4.9	3
YPKELR	686	691	BS3-d0 (OH) (K)	666	C3-BS3d0/ C3b-BS3d4	16.3	1
YPKELR	686	691	BS3-d4 (NH3) (K)	666	C3-BS3d4/ C3b-BS3d3	9.2	2
YPKELR	686	691	BS3-d4 (H2O) (K)	666	C3-BS3d4/ C3b-BS3d5	7.1	4
FISLGEACKK	713	722	BS3-d0 (NH2) (K)	699	C3-BS3d0/ C3b-BS3d4	23.4	1
FISLGEACKK	713	722	BS3-d4 (NH3) (K)	699	C3-BS3d4/ C3b-BS3d9	16.1	7

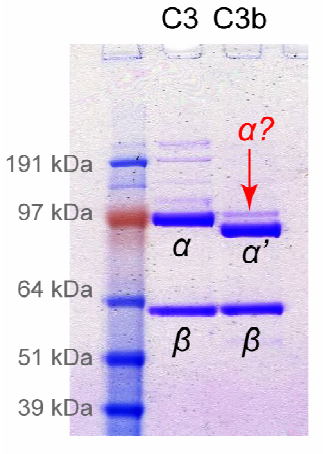


Figure A.1.1 - SDS-PAGE Gel image of the C3 and C3b samples

The band corresponds to C3 α-chain, β-chain, and C3b α'-chain, β-chain were marked. The impurity band that has same electrophoretic mobility was pointed out with red arrow.

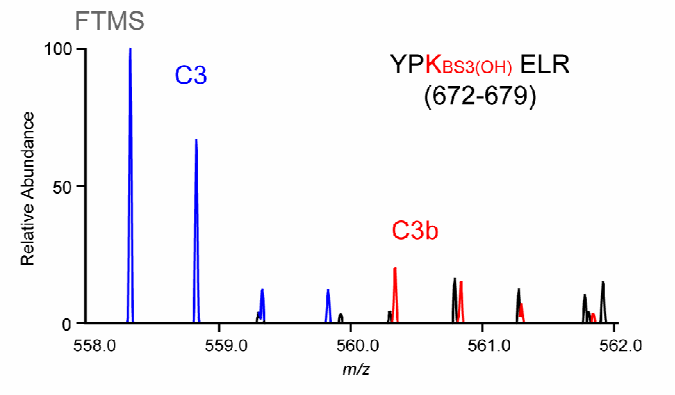


Figure A.1.2 - An example MS^1 spectrum of C3a peptide

The MS spectrum of hydrolyzed BS3-d0/d4 modified C3 specific peptide YPKELR (672-679) detected in the 1:1 C3-BS3d0 and C3b-BS3d4 mixture. The ratio between the signal of BS3-d0 modified peptide from the C3 sample (blue) and the signal of BS3-d4 modified peptide from the C3b sample (red) enable the quantitation of C3 contamination in the C3b sample.

A.2 Supplementary figures

Figure S1 - Mass accuracy of Orbitrap measurement

Figure S2 - Inconsistency between crystallographic and cross-link data on Pol II complex.

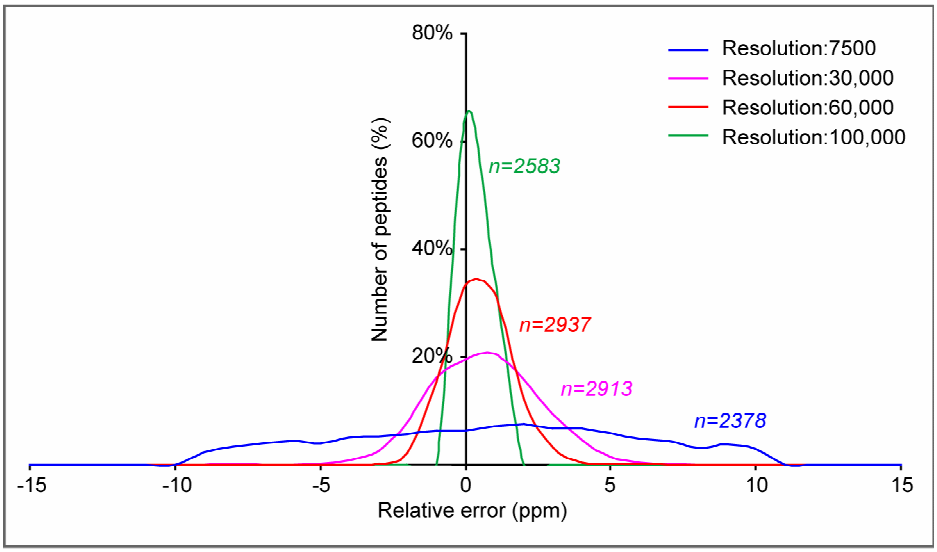


Figure S1 - Mass accuracy of Orbitrap mass analyzer at different resolutions

The mass accuracy of Orbitrap mass analyzer in the LTQ-Orbitrap mass spectrometer is reflected by the mass errors of identified peptides. 1µg of trypsin digested *E.coli* extract samples were analyzed by LC-MS/MS as described in *2.1.3.2* with no charge exclusion. A series of resolution settings were applied to the Orbitrap MS measurements. Peptides were identified through Mascot search as described in *2.1.4*, peptides with higher than 25 Mascot score were accepted. The mass error of identified peptides from in each acquisition were plotted to indicate the mass accuracy of Orbitrap mass analyzer at different resolutions.

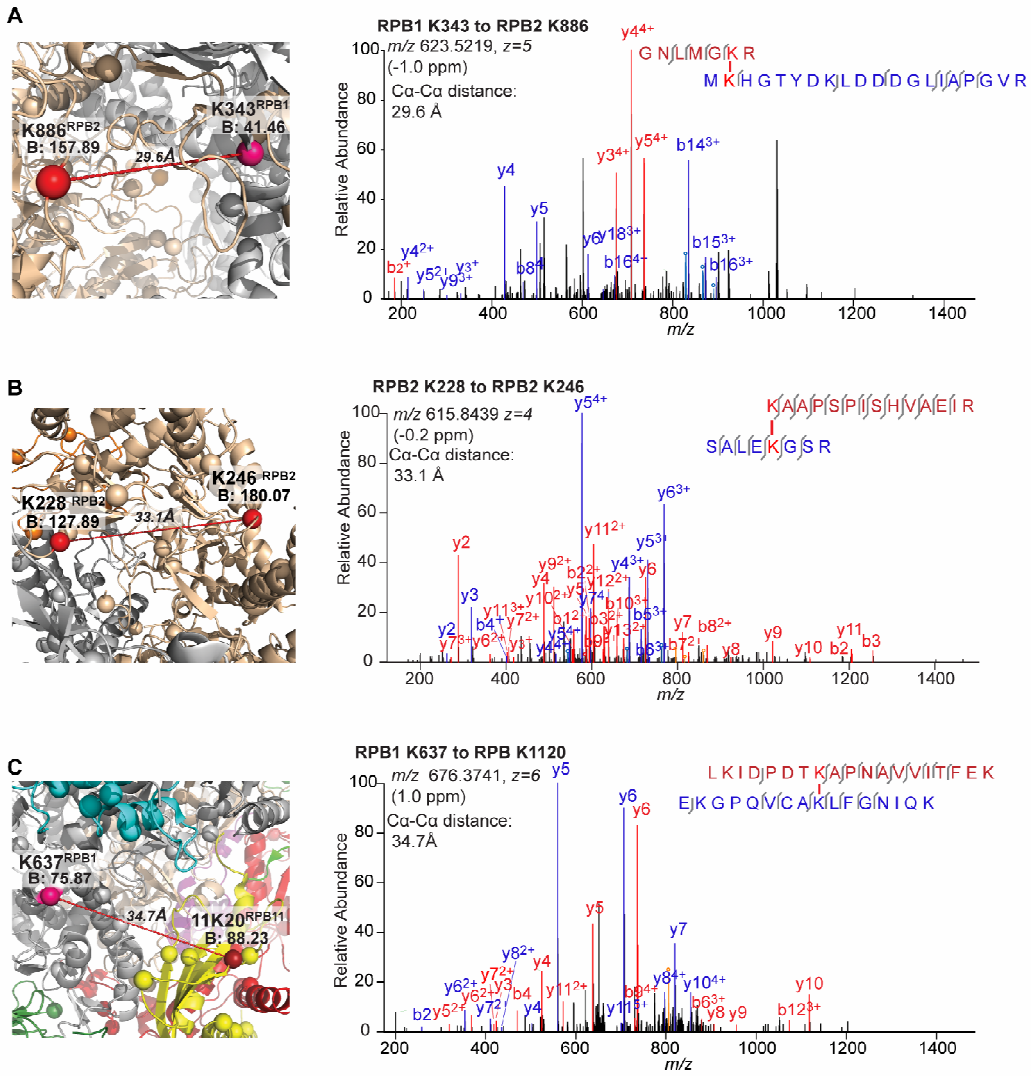


Figure S2 - Inconsistency between crystallographic and cross-link data on the Pol II complex.

Five high confidence cross-links (A, B, C, D, and E) were observed in Pol II complex with over cross-linking limit length in Pol II crystal structure (1WCM). Magnified view of cross-links in Pol II crystal structure (PDB|1WCM) are displayed (left) Alpha-carbons of linked residues are highlighted by coloured sphere (hot pink for Rpb1, red for Rpb2 and firebrick for Rpb11), B-factor for linked residues are displayed under the residue label. The annotated high resolution fragmentation spectra (on the right side) of cross-linked peptides corresponding to the linkage provide the experimental evidence for the cross-link data.

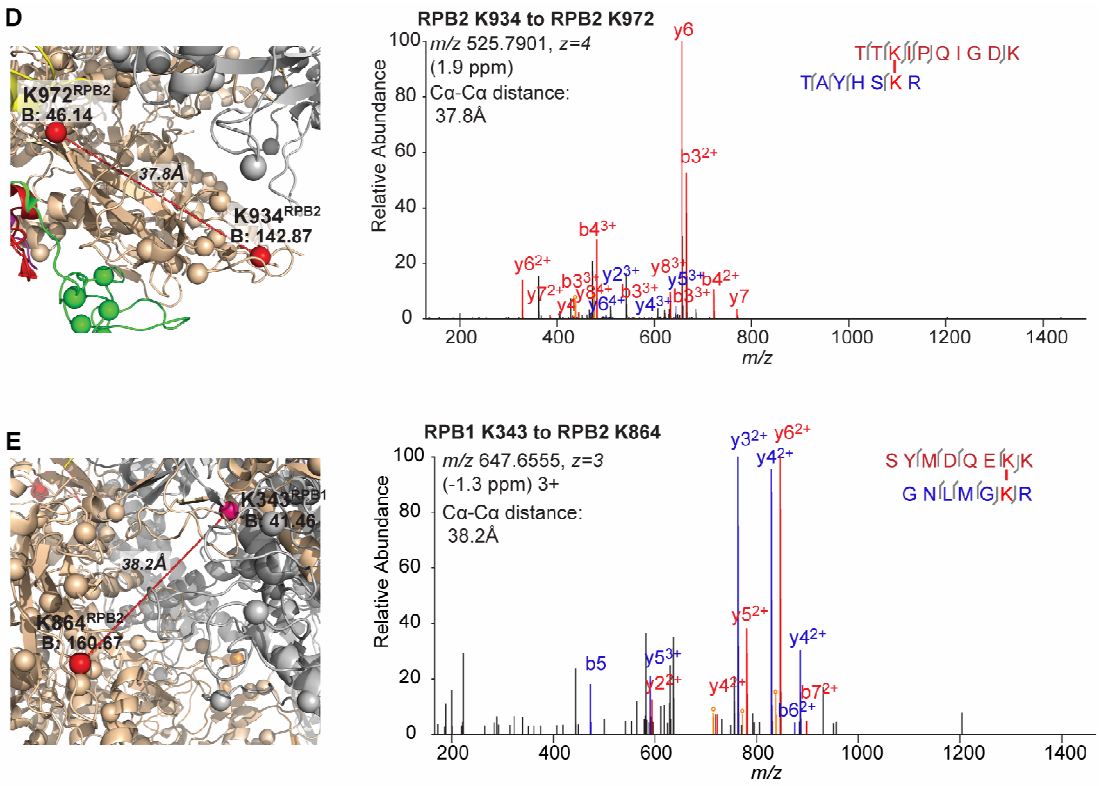


Figure S2 continued - Inconsistency between crystallographic and cross-link data on Pol II complex. (Continued)

A.3 Supplementary Tables

Table S1 - List of 49 synthetic peptides

Table S2 - List of high confidence cross-links observed from Pol II complex sample

Table S3 List of high confidence cross-links observed from Pol II-TFIIF complex sample

(A) Cross-links within Pol II

(B) Cross-links within TFIIF

(C) Cross-links between Pol II and TFIIF

Table S4 - Quantified cross-linkages in conformational comparison of C3 and C3b by quantitative 3D proteomics

Table S5 - Ten most intense protein identified from affinity purified *S. cerevisiae* Mad1-Mad2 complex

Table S6 - Ten most intense protein identified from affinity purified *S. cerevisiae* Ndc80 complex

Table S7 - List of cross-links observed from affinity purified *S. cerevisiae* endogenous Mad1-Mad2 complex

Table S8 - List of cross-links observed from affinity purified *S. cerevisiae* endogenous Ndc80 complex

Table S1 - List of 49 synthetic peptides

SEQUENCE	MASS	*E.coli* Ribosomal Protein	Start residue	End residue	Peptide length
NKAAR	558.3238	S20	69	73	5
KVGLR	571.3806	S9	115	119	5
KDVAR	587.3392	S15	73	77	5
LKLSR	615.4068	S4	9	13	5
YGVKR	621.3599	S12	117	121	5
QSIKR	630.3813	S2	109	113	5
TVKGGR	616.3657	S5	24	29	6
KSSAAR	618.3450	S9	13	18	6
EVNKAA	630.3337	S16	77	82	6
AMEKAR	704.3640	S5	63	68	6
QSMKAR	719.3749	S14	4	9	6
KEALMR	746.4109	S2	131	137	6
KPNSALR	784.4556	S12	44	50	7
MEGTFKR	867.4273	S4	178	184	7
LKAFDHR	885.4821	S10	10	16	7
RPQFSKR	917.5196	S9	124	129	7
LLDYLKR	919.5492	S15	66	72	7
GDKSMALR	876.4488	S7	112	119	8
VKDLPGVR	882.5288	S12	87	94	8
AIQSEKAR	901.4982	S20	10	17	8
ILFVGTKR	932.5808	S2	67	74	8
GEDVEKLR	944.4928	S3	81	88	8
TKSWTLVR	989.5659	S17	70	77	8
TKHAVTEAS	942.4771	S6	92	100	9
GIKVEVSGR	943.5451	S3	148	156	9
MAEANKAFA	951.4485	S7	144	152	9
VGFGYGKAR	953.5084	S5	46	54	9
ELAKASVSR	959.5401	S3	46	54	9
QKVHPNGIR	1047.5938	S3	3	11	9
GQKVHPNGIR	1104.6153	S3	2	11	10
LQEKLIAVNR	1182.7085	S5	11	20	10
EFYEKPTTER	1298.6143	S21	36	45	10
ANLTAQINKLA	1155.6612	S20	69	86	11
VKAALELAEQR	1226.6984	S4	155	165	11
MEKSIVVAIER	1273.7065	S17	17	27	11
IVIERPAKSIR	1280.7929	S3	55	65	11
EGTDLFLKSGVR	1320.7038	S4	15	26	12
LKAGVHFGHQTR	1349.7317	S2	10	21	12
WNAVLKLQTLPR	1437.8457	S14	42	52	12
KSTPFAAQVAAER	1374.7256	S11	57	68	13
AELKAIISDVNAR	1398.7832	S14	25	36	13
QLVSHKAIMVNGR	1451.8032	S4	116	127	13
NYITESGKIVPSR	1462.7781	S18	31	42	13
DEVAKFVVEGDLR	1475.7621	S13	58	69	13
EFADNLDSDFKVR	1554.7315	S3	28	39	13

YTAAITGAEGKIHR	1486.7893	S6	25	36	14
FTVLISPHVNKDAR	1595.8785	S10	17	28	14
LIDQATAEIVETAKR	1656.9047	S10	49	60	15
VFIKPGNGKIVINQR	1681.9993	S9	19	30	15

Table S2 - List of high confidence cross-links observed from the Pol II complex sample

Type	Protein 1	Linked residue 1	Protein 2	Linked residue 2	Confidence	Distance 1WCM (Å)
Intra-protein	Rpb1	Lys 34	Rpb1	Lys 49	High	13.7
Intra-protein	Rpb1	Lys 101	Rpb1	Lys 143	High	8.0
Intra-protein	Rpb1	Lys 101	Rpb1	Lys 176	High	10.5
Intra-protein	Rpb1	Lys 143	Rpb1	Lys 187	High	N/A
Intra-protein	Rpb1	Lys 368	Rpb1	Lys 461	High	9.3
Intra-protein	Rpb1	Lys 372	Rpb1	Lys 403	High	8.7
Intra-protein	Rpb1	Lys 461	Rpb1	Lys 637	High	25.7
Intra-protein	Rpb1	Lys 461	Rpb1	Lys 644	High	16.6
Intra-protein	Rpb1	Lys 688	Rpb1	Lys 728	High	16.3
Intra-protein	Rpb1	Lys 689	Rpb1	Lys 705	High	19.2
Intra-protein	Rpb1	Lys 689	Rpb1	Lys 728	High	14.8
Intra-protein	Rpb1	Lys 705	Rpb1	Lys 1093	High	16.0
Intra-protein	Rpb1	Lys 705	Rpb1	Lys 1132	High	14.2
Intra-protein	Rpb1	Lys 705	Rpb1	Lys 1286	High	16.0
Intra-protein	Rpb1	Lys 773	Rpb1	Lys 1093	High	12.4
Intra-protein	Rpb1	Lys 773	Rpb1	Lys 1112	High	18.2
Intra-protein	Rpb1	Lys 830	Rpb1	Lys 1093	High	19.1
Intra-protein	Rpb1	Lys 830	Rpb1	Lys 1102	High	10.6
Intra-protein	Rpb1	Lys 830	Rpb1	Lys 1112	High	17.7
Intra-protein	Rpb1	Lys 934	Rpb1	Lys 941	High	11.1
Intra-protein	Rpb1	Lys 938	Rpb1	Lys 991	High	15.6
Intra-protein	Rpb1	Lys 1093	Rpb1	Lys 1102	High	17.8
Intra-protein	Rpb1	Lys 1093	Rpb1	Lys 1112	High	11.3
Intra-protein	Rpb1	Lys 1102	Rpb1	Lys 1112	High	12.0
Intra-protein	Rpb1	Lys 1132	Rpb1	Lys 1205	High	9.9
Intra-protein	Rpb1	Lys 1132	Rpb1	Lys 1286	High	13.0
Intra-protein	Rpb1	Lys 1217	Rpb1	Lys 1246	High	N/A
Intra-protein	Rpb1	Lys 1221	Rpb1	Lys 1246	High	N/A
Intra-protein	Rpb1	Lys 1221	Rpb1	Lys 1262	High	16.3
Intra-protein	Rpb1	Lys 1246	Rpb1	Lys 1262	High	N/A
Intra-protein	Rpb2	Lys 99	Rpb2	Lys 164	High	15.8
Intra-protein	Rpb2	Lys 99	Rpb2	Lys 191	High	19.2
Intra-protein	Rpb2	Lys 228	Rpb2	Lys 246	High	33.1
Intra-protein	Rpb2	Lys 228	Rpb2	Lys 257	High	10.4
Intra-protein	Rpb2	Lys 228	Rpb2	Lys 270	High	15.2
Intra-protein	Rpb2	Lys 228	Rpb2	Lys 277	High	21.5
Intra-protein	Rpb2	Lys 228	Rpb2	Lys 507	High	N/A
Intra-protein	Rpb2	Lys 228	Rpb2	Lys 510	High	15.9
Intra-protein	Rpb2	Lys 246	Rpb2	Lys 358	High	14.6
Intra-protein	Rpb2	Lys 246	Rpb2	Lys 426	High	19.8
Intra-protein	Rpb2	Lys 345	Rpb2	Lys 353	High	13.8
Intra-protein	Rpb2	Lys 358	Rpb2	Lys 422	High	20.8
Intra-protein	Rpb2	Lys 422	Rpb2	Lys 426	High	6.0
Intra-protein	Rpb2	Lys 426	Rpb2	Lys 471	High	N/A
Intra-protein	Rpb2	Lys 606	Rpb2	Lys 652	High	13.9

Intra-protein	Rpb2	Lys 864	Rpb2	Lys 934	High	15.3
Intra-protein	Rpb2	Lys 876	Rpb2	Lys 886	High	12.0
Intra-protein	Rpb2	Lys 934	Rpb2	Lys 972	High	37.8
Intra-protein	Rpb2	Lys 979	Rpb2	Lys 987	High	7.2
Intra-protein	Rpb2	Lys 979	Rpb2	Lys 1102	High	11.9
Intra-protein	Rpb3	Lys 15	Rpb3	Lys 137	High	14.7
Intra-protein	Rpb4	Lys 60	Rpb4	Lys 71	High	17.2
Intra-protein	Rpb4	Lys 60	Rpb4	Lys 121	High	14.6
Intra-protein	Rpb4	Lys 60	Rpb4	Lys 142	High	11.9
Intra-protein	Rpb5	Lys 20	Rpb5	Lys 45	High	16.4
Intra-protein	Rpb5	Lys 161	Rpb5	Lys 171	High	11.5
Intra-protein	Rpb5	Lys 161	Rpb5	Lys 191	High	13.3
Intra-protein	Rpb5	Lys 166	Rpb5	Lys 171	High	8.8
Intra-protein	Rpb5	Lys 191	Rpb5	Lys 197	High	14.3
Intra-protein	Rpb6	Lys 123	Rpb6	Lys 129	High	8.6
Intra-protein	Rpb11	Lys 20	Rpb11	Lys 37	High	12.0
Intra-protein	Rpb11	Lys 55	Rpb11	Lys 84	High	12.3
Intra-protein	Rpb11	Lys 55	Rpb11	Lys 88	High	13.3
Intra-protein	Rpb12	Lys 28	Rpb12	Lys 49	High	14.8
Inter-protein	Rpb1	Lys 330	Rpb2	Lys 507	High	N/A
Inter-protein	Rpb1	Lys 332	Rpb2	Lys 507	High	N/A
Inter-protein	Rpb1	Lys 343	Rpb2	Lys 864	High	38.2
Inter-protein	Rpb1	Lys 343	Rpb2	Lys 886	High	29.6
Inter-protein	Rpb1	Lys 372	Rpb2	Lys 886	High	25
Inter-protein	Rpb1	Lys 1093	Rpb2	Lys 507	High	N/A
Inter-protein	Rpb1	Lys 1102	Rpb2	Lys 507	High	N/A
Inter-protein	Rpb1	Lys 1112	Rpb2	Lys 507	High	N/A
Inter-protein	Rpb1	Lys 15	Rpb4	N-terminus	High	N/A
Inter-protein	Rpb1	Lys 129	Rpb4	N-terminus	High	N/A
Inter-protein	Rpb1	Lys 1003	Rpb4	N-terminus	High	N/A
Inter-protein	Rpb1	Lys 15	Rpb5	Lys 171	High	21.2
Inter-protein	Rpb1	Lys 129	Rpb5	Lys 161	High	22.1
Inter-protein	Rpb1	Lys 129	Rpb5	Lys 171	High	15.1
Inter-protein	Rpb1	Lys 934	Rpb5	Lys 201	High	21.6
Inter-protein	Rpb1	Lys 938	Rpb5	Lys 201	High	15.6
Inter-protein	Rpb1	Lys 1003	Rpb5	Lys 166	High	12.6
Inter-protein	Rpb1	Lys 129	Rpb6	Lys 67	High	N/A
Inter-protein	Rpb1	Lys 1003	Rpb6	Lys 76	High	13.6
Inter-protein	Rpb1	Lys 15	Rpb7	Lys 73	High	23
Inter-protein	Rpb1	Lys 1246	Rpb9	Lys 20	High	N/A
Inter-protein	Rpb1	Lys 637	Rpb11	Lys 20	High	34.7
Inter-protein	Rpb1	Lys 637	Rpb11	Lys 26	High	21.6
Inter-protein	Rpb1	Lys 644	Rpb11	Lys 20	High	26.5
Inter-protein	Rpb1	Lys 644	Rpb11	Lys 26	High	18.4
Inter-protein	Rpb2	Lys 191	Rpb3	Lys 149	High	13.4
Inter-protein	Rpb2	Lys 1057	Rpb3	Lys 199	High	13.1
Inter-protein	Rpb2	Lys 1188	Rpb4	N-terminus	High	N/A
Inter-protein	Rpb2	Lys 177	Rpb10	Lys 68	High	N/A
Inter-protein	Rpb2	Lys 191	Rpb10	Lys 68	High	N/A
Inter-protein	Rpb2	Lys 813	Rpb10	Lys 59	High	14.1
Inter-protein	Rpb3	Lys 149	Rpb10	Lys 68	High	N/A
Inter-protein	Rpb3	Lys 154	Rpb10	Lys 68	High	N/A

Inter-protein	Rpb3	Lys 160	Rpb11	Lys 37	High	21.3
Inter-protein	Rpb3	Lys 154	Rpb12	Lys 62	High	12.1
Inter-protein	Rpb4	N-terminus	Rpb5	Lys 171	High	N/A
Inter-protein	Rpb4	N-terminus	Rpb6	Lys 72	High	N/A
Inter-protein	Rpb4	N-terminus	Rpb7	Lys 29	High	N/A
Inter-protein	Rpb4	N-terminus	Rpb7	Lys 73	High	N/A
Inter-protein	Rpb5	Lys 166	Rpb6	Lys 67	High	N/A
Inter-protein	Rpb5	Lys 171	Rpb6	Lys 46	High	N/A
Inter-protein	Rpb5	Lys 171	Rpb6	Lys 67	High	N/A
Inter-protein	Rpb5	Lys 171	Rpb6	Lys 72	High	16
Inter-protein	Rpb6	Lys 72	Rpb7	Lys 73	High	18

Table S3 - List of high confidence cross-links observed from the Pol II-TFIIF complex sample

(A) Cross-links within Pol II

Type	Protein1	Linked residue 1	Protein 2	Linked residue 2	Confidence	distance in 1WCM (Å)
Intra-protein	Rpb1	15	Rpb1	34	High	33.6
Intra-protein	Rpb1	15	Rpb1	49	High	38.9
Intra-protein	Rpb1	49	Rpb1	176	High	41.7
Intra-protein	Rpb1	101	Rpb1	143	High	8.0
Intra-protein	Rpb1	101	Rpb1	176	High	10.5
Intra-protein	Rpb1	143	Rpb1	101	High	8.0
Intra-protein	Rpb1	143	Rpb1	129	High	20.8
Intra-protein	Rpb1	143	Rpb1	186	High	15.1
Intra-protein	Rpb1	143	Rpb1	187	High	N/A
Intra-protein	Rpb1	265	Rpb1	34	High	17.4
Intra-protein	Rpb1	343	Rpb1	49	High	33.9
Intra-protein	Rpb1	343	Rpb1	403	High	25.4
Intra-protein	Rpb1	368	Rpb1	461	High	9.3
Intra-protein	Rpb1	372	Rpb1	403	High	8.7
Intra-protein	Rpb1	403	Rpb1	343	High	25.4
Intra-protein	Rpb1	431	Rpb1	49	High	42.3
Intra-protein	Rpb1	431	Rpb1	343	High	26.1
Intra-protein	Rpb1	461	Rpb1	431	High	23.3
Intra-protein	Rpb1	637	Rpb1	461	High	25.7
Intra-protein	Rpb1	689	Rpb1	728	High	14.8
Intra-protein	Rpb1	705	Rpb1	688	High	22.2
Intra-protein	Rpb1	705	Rpb1	689	High	19.2
Intra-protein	Rpb1	705	Rpb1	1286	High	N/A
Intra-protein	Rpb1	773	Rpb1	1093	High	N/A
Intra-protein	Rpb1	773	Rpb1	1112	High	N/A
Intra-protein	Rpb1	830	Rpb1	1093	High	N/A
Intra-protein	Rpb1	830	Rpb1	1102	High	N/A
Intra-protein	Rpb1	830	Rpb1	1112	High	N/A
Intra-protein	Rpb1	843	Rpb1	343	High	17.6
Intra-protein	Rpb1	843	Rpb1	1102	High	N/A
Intra-protein	Rpb1	938	Rpb1	941	High	5.5
Intra-protein	Rpb1	1092	Rpb1	830	High	N/A
Intra-protein	Rpb1	1093	Rpb1	705	High	N/A
Intra-protein	Rpb1	1093	Rpb1	728	High	N/A
Intra-protein	Rpb1	1093	Rpb1	1102	High	N/A
Intra-protein	Rpb1	1102	Rpb1	1093	High	N/A
Intra-protein	Rpb1	1112	Rpb1	830	High	N/A
Intra-protein	Rpb1	1112	Rpb1	1093	High	N/A
Intra-protein	Rpb1	1112	Rpb1	1102	High	N/A
Intra-protein	Rpb1	1132	Rpb1	705	High	N/A
Intra-protein	Rpb1	1132	Rpb1	1205	High	N/A
Intra-protein	Rpb1	1132	Rpb1	1286	High	N/A
Intra-protein	Rpb1	1205	Rpb1	1132	High	N/A
Intra-protein	Rpb1	1221	Rpb1	1246	High	N/A

Intra-protein	Rpb1	1262	Rpb1	1246	High	N/A
Intra-protein	Rpb1	1286	Rpb1	1093	High	N/A
Intra-protein	Rpb1	1350	Rpb1	1093	High	N/A
Intra-protein	Rpb1	1350	Rpb1	1290	High	N/A
Intra-protein	Rpb11	37	Rpb11	20	High	12.0
Intra-protein	Rpb11	55	Rpb11	88	High	13.3
Intra-protein	Rpb11	84	Rpb11	55	High	12.3
Intra-protein	Rpb11	88	Rpb11	55	High	13.3
Intra-protein	Rpb12	28	Rpb12	49	High	14.8
Intra-protein	Rpb2	148	Rpb2	87	High	N/A
Intra-protein	Rpb2	164	Rpb2	133	High	9.6
Intra-protein	Rpb2	177	Rpb2	99	High	9.2
Intra-protein	Rpb2	191	Rpb2	99	High	19.2
Intra-protein	Rpb2	227	Rpb2	507	High	N/A
Intra-protein	Rpb2	228	Rpb2	246	High	33.1
Intra-protein	Rpb2	228	Rpb2	270	High	15.2
Intra-protein	Rpb2	228	Rpb2	471	High	N/A
Intra-protein	Rpb2	228	Rpb2	507	High	N/A
Intra-protein	Rpb2	228	Rpb2	510	High	15.9
Intra-protein	Rpb2	246	Rpb2	358	High	14.6
Intra-protein	Rpb2	246	Rpb2	426	High	19.8
Intra-protein	Rpb2	246	Rpb2	471	High	N/A
Intra-protein	Rpb2	257	Rpb2	228	High	10.4
Intra-protein	Rpb2	257	Rpb2	246	High	28.8
Intra-protein	Rpb2	257	Rpb2	270	High	5.1
Intra-protein	Rpb2	257	Rpb2	507	High	N/A
Intra-protein	Rpb2	270	Rpb2	228	High	15.2
Intra-protein	Rpb2	358	Rpb2	246	High	14.6
Intra-protein	Rpb2	358	Rpb2	344	High	N/A
Intra-protein	Rpb2	358	Rpb2	426	High	23.0
Intra-protein	Rpb2	358	Rpb2	471	High	N/A
Intra-protein	Rpb2	393	Rpb2	622	High	13.9
Intra-protein	Rpb2	418	Rpb2	246	High	10.6
Intra-protein	Rpb2	426	Rpb2	246	High	19.8
Intra-protein	Rpb2	426	Rpb2	422	High	6.0
Intra-protein	Rpb2	426	Rpb2	471	High	N/A
Intra-protein	Rpb2	451	Rpb2	865	High	20.3
Intra-protein	Rpb2	471	Rpb2	423	High	N/A
Intra-protein	Rpb2	507	Rpb2	471	High	N/A
Intra-protein	Rpb2	606	Rpb2	649	High	21.2
Intra-protein	Rpb2	649	Rpb2	606	High	21.2
Intra-protein	Rpb2	775	Rpb2	979	High	13.6
Intra-protein	Rpb2	775	Rpb2	987	High	12.7
Intra-protein	Rpb2	775	Rpb2	1102	High	N/A
Intra-protein	Rpb2	864	Rpb2	934	High	15.3
Intra-protein	Rpb2	865	Rpb2	886	High	22.5
Intra-protein	Rpb2	865	Rpb2	934	High	16.3
Intra-protein	Rpb2	876	Rpb2	886	High	12.0
Intra-protein	Rpb2	886	Rpb2	876	High	12.0
Intra-protein	Rpb2	886	Rpb2	934	High	12.0
Intra-protein	Rpb2	886	Rpb2	1102	High	N/A
Intra-protein	Rpb2	914	Rpb2	886	High	11.4

Intra-protein	Rpb2	934	Rpb2	864	High	15.3
Intra-protein	Rpb2	972	Rpb2	965	High	22.4
Intra-protein	Rpb2	972	Rpb2	1102	High	N/A
Intra-protein	Rpb2	979	Rpb2	934	High	45.4
Intra-protein	Rpb2	979	Rpb2	987	High	7.2
Intra-protein	Rpb2	979	Rpb2	1102	High	N/A
Intra-protein	Rpb2	987	Rpb2	1102	High	N/A
Intra-protein	Rpb2	1102	Rpb2	507	High	N/A
Intra-protein	Rpb2	1102	Rpb2	886	High	N/A
Intra-protein	Rpb2	1102	Rpb2	934	High	N/A
Intra-protein	Rpb2	1102	Rpb2	972	High	N/A
Intra-protein	Rpb2	1102	Rpb2	979	High	N/A
Intra-protein	Rpb2	1102	Rpb2	987	High	N/A
Intra-protein	Rpb2	1174	Rpb2	1188	High	N/A
Intra-protein	Rpb2	1188	Rpb2	1174	High	N/A
Intra-protein	Rpb3	15	Rpb3	137	High	14.7
Intra-protein	Rpb3	149	Rpb3	161	High	37.0
Intra-protein	Rpb3	165	Rpb3	253	High	16.8
Intra-protein	Rpb4	121	Rpb4	60	High	14.6
Intra-protein	Rpb4	142	Rpb4	60	High	11.9
Intra-protein	Rpb5	20	Rpb5	45	High	16.4
Intra-protein	Rpb5	45	Rpb5	20	High	16.4
Intra-protein	Rpb5	56	Rpb5	45	High	12.1
Intra-protein	Rpb5	171	Rpb5	161	High	11.5
Intra-protein	Rpb5	171	Rpb5	201	High	26.7
Intra-protein	Rpb5	197	Rpb5	201	High	13.3
Inter-protein	Rpb1	461	Rpb11	62	High	20.2
Inter-protein	Rpb1	533	Rpb11	55	High	33.6
Inter-protein	Rpb1	637	Rpb11	20	High	34.7
Inter-protein	Rpb1	637	Rpb11	26	High	21.6
Inter-protein	Rpb1	637	Rpb11	62	High	28.3
Inter-protein	Rpb1	644	Rpb11	20	High	26.5
Inter-protein	Rpb1	644	Rpb11	26	High	18.4
Inter-protein	Rpb1	49	Rpb2	1174	High	N/A
Inter-protein	Rpb1	180	Rpb2	257	High	39.5
Inter-protein	Rpb1	265	Rpb2	228	High	47.2
Inter-protein	Rpb1	323	Rpb2	471	High	N/A
Inter-protein	Rpb1	323	Rpb2	934	High	35.5
Inter-protein	Rpb1	330	Rpb2	507	High	N/A
Inter-protein	Rpb1	332	Rpb2	507	High	N/A
Inter-protein	Rpb1	343	Rpb2	864	High	38.2
Inter-protein	Rpb1	343	Rpb2	886	High	29.6
Inter-protein	Rpb1	343	Rpb2	934	High	38.0
Inter-protein	Rpb1	343	Rpb2	972	High	34.4
Inter-protein	Rpb1	343	Rpb2	987	High	32.2
Inter-protein	Rpb1	343	Rpb2	1102	High	N/A
Inter-protein	Rpb1	343	Rpb2	1148	High	N/A
Inter-protein	Rpb1	372	Rpb2	886	High	25.0
Inter-protein	Rpb1	403	Rpb2	886	High	23.4
Inter-protein	Rpb1	403	Rpb2	1102	High	N/A
Inter-protein	Rpb1	773	Rpb2	227	High	20.7
Inter-protein	Rpb1	830	Rpb2	510	High	16.6

Inter-protein	Rpb1	830	Rpb2	987	High	23.4
Inter-protein	Rpb1	1093	Rpb2	227	High	N/A
Inter-protein	Rpb1	1093	Rpb2	228	High	N/A
Inter-protein	Rpb1	1093	Rpb2	507	High	N/A
Inter-protein	Rpb1	1102	Rpb2	507	High	N/A
Inter-protein	Rpb1	1102	Rpb2	510	High	N/A
Inter-protein	Rpb1	1112	Rpb2	227	High	N/A
Inter-protein	Rpb1	1112	Rpb2	228	High	N/A
Inter-protein	Rpb1	1112	Rpb2	507	High	N/A
Inter-protein	Rpb1	15	Rpb5	171	High	21.2
Inter-protein	Rpb1	34	Rpb5	171	High	50.3
Inter-protein	Rpb1	129	Rpb5	171	High	15.1
Inter-protein	Rpb1	343	Rpb5	171	High	37.5
Inter-protein	Rpb1	934	Rpb5	201	High	21.6
Inter-protein	Rpb1	938	Rpb5	201	High	15.5
Inter-protein	Rpb1	1003	Rpb5	166	High	N/A
Inter-protein	Rpb1	1003	Rpb5	197	High	N/A
Inter-protein	Rpb1	1350	Rpb5	201	High	N/A
Inter-protein	Rpb1	1003	Rpb6	46	High	N/A
Inter-protein	Rpb1	1003	Rpb6	76	High	N/A
Inter-protein	Rpb1	15	Rpb7	73	High	23.0
Inter-protein	Rpb1	977	Rpb8	136	High	13.7
Inter-protein	Rpb1	1246	Rpb9	20	High	N/A
Inter-protein	Rpb1	1262	Rpb9	20	High	N/A
Inter-protein	Rpb10	68	Rpb12	49	High	N/A
Inter-protein	Rpb2	177	Rpb10	68	High	N/A
Inter-protein	Rpb2	191	Rpb10	68	High	N/A
Inter-protein	Rpb2	813	Rpb10	59	High	14.1
Inter-protein	Rpb2	864	Rpb12	58	High	22.0
Inter-protein	Rpb2	865	Rpb12	58	High	24.2
Inter-protein	Rpb2	1057	Rpb3	199	High	N/A
Inter-protein	Rpb2	1188	Rpb4	17	High	N/A
Inter-protein	Rpb3	154	Rpb10	68	High	N/A
Inter-protein	Rpb3	160	Rpb11	37	High	21.3
Inter-protein	Rpb3	253	Rpb11	18	High	13.9
Inter-protein	Rpb3	253	Rpb11	37	High	15.0
Inter-protein	Rpb3	149	Rpb12	37	High	16.2
Inter-protein	Rpb5	45	Rpb6	46	High	N/A
Inter-protein	Rpb5	166	Rpb6	46	High	N/A
Inter-protein	Rpb5	171	Rpb6	46	High	N/A
Inter-protein	Rpb5	171	Rpb6	70	High	N/A

(B) Cross-links within TFIIF

Type	Protein 1	Linked residue 1	Domain	Protein 2	Linked residue 2	Domain	Confidence
Intra-protein	Tfg1	Lys 23	N-terminal region	Tfg1	Lys 61	N-terminal region	High
Intra-protein	Tfg1	Lys 23	N-terminal region	Tfg1	Lys 126	Dimerization domain	High

Intra-protein	Tfg1	Lys 60	N-terminal region	Tfg1	Lys 72	N-terminal region	High
Intra-protein	Tfg1	Lys 60	N-terminal region	Tfg1	Lys 89	N-terminal region	High
Intra-protein	Tfg1	Lys 60	N-terminal region	Tfg1	Lys 108	Dimerization domain	High
Intra-protein	Tfg1	Lys 61	N-terminal region	Tfg1	Lys 72	N-terminal region	High
Intra-protein	Tfg1	Lys 61	N-terminal region	Tfg1	Lys 89	N-terminal region	High
Intra-protein	Tfg1	Lys 61	N-terminal region	Tfg1	Lys 108	Dimerization domain	High
Intra-protein	Tfg1	Lys 72	N-terminal region	Tfg1	Lys 89	N-terminal region	High
Intra-protein	Tfg1	Lys 72	N-terminal region	Tfg1	Lys 91	N-terminal region	High
Intra-protein	Tfg1	Lys 120	Dimerization domain	Tfg1	Lys 394	Dimerization domain	High
Intra-protein	Tfg1	Lys 120	Dimerization domain	Tfg1	Lys 400	Dimerization domain	High
Intra-protein	Tfg1	Lys 124	Dimerization domain	Tfg1	Lys 394	Dimerization domain	High
Intra-protein	Tfg1	Lys 126	Dimerization domain	Tfg1	Lys 142	Dimerization domain	High
Intra-protein	Tfg1	Lys 126	Dimerization domain	Tfg1	Lys 426	Charged domain	High
Intra-protein	Tfg1	Lys 161	Dimerization domain	Tfg1	Lys 169	Insertion	High
Intra-protein	Tfg1	Lys 169	Insertion	Tfg1	Lys 184	Insertion	High
Intra-protein	Tfg1	Lys 169	Insertion	Tfg1	Lys 195	Insertion	High
Intra-protein	Tfg1	Lys 169	Insertion	Tfg1	Lys 289	Insertion	High
Intra-protein	Tfg1	Lys 169	Insertion	Tfg1	Lys 416	Charged domain	High
Intra-protein	Tfg1	Lys 169	Insertion	Tfg1	Lys 426	Charged domain	High
Intra-protein	Tfg1	Lys 333	Dimerization domain	Tfg1	Lys 426	Charged domain	High
Intra-protein	Tfg1	Lys 394	Dimerization domain	Tfg1	Lys 400	Dimerization domain	High
Intra-protein	Tfg1	Lys 394	Dimerization domain	Tfg1	Lys 426	Charged domain	High
Intra-protein	Tfg1	Lys 400	Dimerization domain	Tfg1	Lys 416	Charged domain	High
Intra-protein	Tfg1	Lys 400	Dimerization domain	Tfg1	Lys 426	Charged domain	High
Intra-protein	Tfg1	Lys 411	Charged domain	Tfg1	Lys 416	Charged domain	High
Intra-protein	Tfg1	Lys 411	Charged domain	Tfg1	Lys 426	Charged domain	High
Intra-protein	Tfg1	Lys 411	Charged domain	Tfg1	Lys 527		High
Intra-protein	Tfg1	Lys 415	Charged domain	Tfg1	Lys 426	Charged domain	High
Intra-protein	Tfg1	Lys 416	Charged domain	Tfg1	Lys 426	Charged domain	High

Intra-protein	Tfg1	Lys 426	Charged domain	Tfg1	Lys 527		High
Intra-protein	Tfg1	Lys 527		Tfg1	Lys 537		High
Intra-protein	Tfg1	Lys 527		Tfg1	Lys 574		High
Intra-protein	Tfg1	Lys 527		Tfg1	Lys 575		High
Intra-protein	Tfg1	Lys 527		Tfg1	Lys 614		High
Intra-protein	Tfg1	Lys 527		Tfg1	Lys 724	WH domain	High
Intra-protein	Tfg1	Lys 604		Tfg1	Lys 625		High
Intra-protein	Tfg1	Lys 704		Tfg1	Lys 719	WH domain	High
Intra-protein	Tfg1	Lys 711	WH domain	Tfg1	Lys 719	WH domain	High
Intra-protein	Tfg1	Lys 712	WH domain	Tfg1	Lys 720	WH domain	High
Intra-protein	Tfg1	Lys 719	WH domain	Tfg1	Lys 724	WH domain	High
Intra-protein	Tfg1	Lys 720	WH domain	Tfg1	Lys 733		High
Intra-protein	Tfg2	Lys 80	Dimerization domain	Tfg2	Lys 245	Linker	High
Intra-protein	Tfg2	Lys 80	Dimerization domain	Tfg2	Lys 279	Linker	High
Intra-protein	Tfg2	Lys 80	Dimerization domain	Tfg2	Lys 319	WH domain	High
Intra-protein	Tfg2	Lys 94	Dimerization domain	Tfg2	Lys 99	Dimerization domain	High
Intra-protein	Tfg2	Lys 99	Dimerization domain	Tfg2	Lys 127	Dimerization domain	High
Intra-protein	Tfg2	Lys 142	Dimerization domain	Tfg2	Lys 147		High
Intra-protein	Tfg2	Lys 142	Dimerization domain	Tfg2	Lys 156		High
Intra-protein	Tfg2	Lys 142	Dimerization domain	Tfg2	Lys 172		High
Intra-protein	Tfg2	Lys 142	Dimerization domain	Tfg2	Lys 186		High
Intra-protein	Tfg2	Lys 142	Dimerization domain	Tfg2	Lys 206	Dimerization domain	High
Intra-protein	Tfg2	Lys 147		Tfg2	Lys 163		High
Intra-protein	Tfg2	Lys 147		Tfg2	Lys 164		High
Intra-protein	Tfg2	Lys 148		Tfg2	Lys 163		High
Intra-protein	Tfg2	Lys 148		Tfg2	Lys 185		High
Intra-protein	Tfg2	Lys 163		Tfg2	Lys 359		High
Intra-protein	Tfg2	Lys 172		Tfg2	Lys 186		High
Intra-protein	Tfg2	Lys 179		Tfg2	Lys 186		High
Intra-protein	Tfg2	Lys 179		Tfg2	Lys 206	Dimerization domain	High
Intra-protein	Tfg2	Lys 245	Linker	Tfg2	Lys 279	Linker	High
Intra-protein	Tfg2	Lys 245	Linker	Tfg2	Lys 290	Linker	High
Intra-protein	Tfg2	Lys 245	Linker	Tfg2	Lys 319	WH domain	High
Intra-protein	Tfg2	Lys 249	Linker	Tfg2	Lys 279	Linker	High
Intra-protein	Tfg2	Lys 249	Linker	Tfg2	Lys 357		High
Intra-protein	Tfg2	Lys 279	Linker	Tfg2	Lys 279	Linker	High
Intra-protein	Tfg2	Lys 279	Linker	Tfg2	Lys 286	Linker	High
Intra-protein	Tfg2	Lys 279	Linker	Tfg2	Lys 297	WH domain	High
Intra-protein	Tfg2	Lys 279	Linker	Tfg2	Lys 316	WH domain	High
Intra-protein	Tfg2	Lys 279	Linker	Tfg2	Lys 319	WH domain	High
Intra-protein	Tfg2	Lys 279	Linker	Tfg2	Lys 341	WH domain	High

Intra-protein	Tfg2	Lys 279	Linker	Tfg2	Lys 357		High
Intra-protein	Tfg2	Lys 279	Linker	Tfg2	Lys 359		High
Intra-protein	Tfg2	Lys 286	Linker	Tfg2	Lys 319	WH domain	High
Intra-protein	Tfg2	Lys 290	WH domain	Tfg2	Lys 319	WH domain	High
Intra-protein	Tfg2	Lys 296	WH domain	Tfg2	Lys 335	WH domain	High
Intra-protein	Tfg2	Lys 296	WH domain	Tfg2	Lys 357		High
Intra-protein	Tfg2	Lys 296	WH domain	Tfg2	Lys 359		High
Intra-protein	Tfg2	Lys 297	WH domain	Tfg2	Lys 335	WH domain	High
Intra-protein	Tfg2	Lys 297	WH domain	Tfg2	Lys 359		High
Intra-protein	Tfg2	Lys 319	WH domain	Tfg2	Lys 330	WH domain	High
Intra-protein	Tfg2	Lys 330	WH domain	Tfg2	Lys 342	WH domain	High
Intra-protein	Tfg2	Lys 335	WH domain	Tfg2	Lys 357		High
Intra-protein	Tfg2	Lys 335	WH domain	Tfg2	Lys 359		High
Intra-protein	Tfg2	Lys 342	WH domain	Tfg2	Lys 359		High
Intra-protein	Tfg2	Lys 356		Tfg2	Lys 359		High
Intra-protein	Tfg3	Lys 100		Tfg3	Lys 166		High
Intra-protein	Tfg3	Lys 149		Tfg3	Lys 161		High
Intra-protein	Tfg3	Lys 149		Tfg3	Lys 166		High
Intra-protein	Tfg3	Lys 149		Tfg3	Lys 181		High
Intra-protein	Tfg3	Lys 161		Tfg3	Lys 166		High
Intra-protein	Tfg3	Lys 166		Tfg3	Lys 181		High
Intra-protein	Tfg3	Lys 166		Tfg3	Lys 240		High
Intra-protein	Tfg3	Lys 181		Tfg3	Lys 240		High
Inter-protein	Tfg1	Lys 60	N-terminal region	Tfg2	Lys 235	Linker	High
Inter-protein	Tfg1	Lys 60	N-terminal region	Tfg2	Lys 245	Linker	High
Inter-protein	Tfg1	Lys 60	N-terminal region	Tfg2	Lys 249	Linker	High
Inter-protein	Tfg1	Lys 61	N-terminal region	Tfg2	Lys 99	Dimerization domain	High
Inter-protein	Tfg1	Lys 61	N-terminal region	Tfg2	Lys 235	Linker	High
Inter-protein	Tfg1	Lys 61	N-terminal region	Tfg2	Lys 245	Linker	High
Inter-protein	Tfg1	Lys 61	N-terminal region	Tfg2	Lys 249	Linker	High
Inter-protein	Tfg1	Lys 72	N-terminal region	Tfg2	Lys 235	Linker	High
Inter-protein	Tfg1	Lys 89	N-terminal region	Tfg2	Lys 94	Dimerization domain	High
Inter-protein	Tfg1	Lys 89	N-terminal region	Tfg2	Lys 99	Dimerization domain	High
Inter-protein	Tfg1	Lys 89	N-terminal region	Tfg2	Lys 103	Dimerization domain	High
Inter-protein	Tfg1	Lys 89	N-terminal region	Tfg2	Lys 235	Linker	High
Inter-protein	Tfg1	Lys 91	N-terminal region	Tfg2	Lys 103	Dimerization domain	High
Inter-protein	Tfg1	Lys 108	Dimerization domain	Tfg2	Lys 147		High
Inter-protein	Tfg1	Lys 108	Dimerization domain	Tfg2	Lys 156		High

Inter-protein	Tfg1	Lys 108	Dimerization domain	Tfg2	Lys 279	Linker	High
Inter-protein	Tfg1	Lys 108	Dimerization domain	Tfg2	Lys 319	WH domain	High
Inter-protein	Tfg1	Lys 124	Dimerization domain	Tfg2	Lys 127	Dimerization domain	High
Inter-protein	Tfg1	Lys 125	Dimerization domain	Tfg2	Lys 127	Dimerization domain	High
Inter-protein	Tfg1	Lys 126	Dimerization domain	Tfg2	Lys 127	Dimerization domain	High
Inter-protein	Tfg1	Lys 126	Dimerization domain	Tfg2	Lys 279	Linker	High
Inter-protein	Tfg1	Lys 142	Dimerization domain	Tfg2	Lys 80	Dimerization domain	High
Inter-protein	Tfg1	Lys 333	Dimerization domain	Tfg2	Lys 142	Dimerization domain	High
Inter-protein	Tfg1	Lys 333	Dimerization domain	Tfg2	Lys 179		High
Inter-protein	Tfg1	Lys 333	Dimerization domain	Tfg2	Lys 290	Linker	High
Inter-protein	Tfg1	Lys 333	Dimerization domain	Tfg2	Lys 297	WH domain	High
Inter-protein	Tfg1	Lys 394	Dimerization domain	Tfg2	Lys 127	Dimerization domain	High
Inter-protein	Tfg1	Lys 394	Dimerization domain	Tfg2	Lys 245	Linker	High
Inter-protein	Tfg1	Lys 394	Dimerization domain	Tfg2	Lys 279	Linker	High
Inter-protein	Tfg1	Lys 426	Charged domain	Tfg2	Lys 279	Linker	High
Inter-protein	Tfg1	Lys 426	Charged domain	Tfg2	Lys 357		High
Inter-protein	Tfg2	Lys 359		Tfg3	Lys 166		High
Inter-protein	Tfg1	Lys 426	Charged domain	Tfg3	Lys 166		High
Inter-protein	Tfg1	Lys 527		Tfg3	Lys 149		High
Inter-protein	Tfg1	Lys 527		Tfg3	Lys 166		High
Inter-protein	Tfg1	Lys 527		Tfg3	Lys 181		High
Inter-protein	Tfg1	Lys 614		Tfg3	Lys 149		High
Inter-protein	Tfg1	Lys 614		Tfg3	Lys 166		High

(C) Cross-links between Pol II and TFIIF

Type	Protein 1	Linked residue 1	Domain	Protein 2	Linked residue 2	Confidence
Inter-protein	Tfg1	Lys 23	N-terminal region	Rpb2	Lys 1057	High
Inter-protein	Tfg1	Lys 60	N-terminal region	Rpb2	Lys 606	High
Inter-protein	Tfg1	Lys 61	N-terminal region	Rpb2	Lys 606	High
Inter-protein	Tfg1	Lys 61	N-terminal region	Rpb2	Lys 652	High
Inter-protein	Tfg1	Lys 72	N-terminal region	Rpb2	Lys 606	High
Inter-protein	Tfg1	Lys 89	N-terminal region	Rpb2	Lys 606	High
Inter-protein	Tfg1	Lys 333	Dimerization domain	Rpb2	Lys 87	High

Inter-protein	Tfg1	Lys 333	Dimerization domain	Rpb2	Lys 133	High
Inter-protein	Tfg1	Lys 333	Dimerization domain	Rpb2	Lys 358	High
Inter-protein	Tfg1	Lys 333	Dimerization domain	Rpb2	Lys 422	High
Inter-protein	Tfg1	Lys 333	Dimerization domain	Rpb2	Lys 426	High
Inter-protein	Tfg1	Lys 340	Dimerization domain	Rpb2	Lys 87	High
Inter-protein	Tfg1	Lys 340	Dimerization domain	Rpb2	Lys 358	High
Inter-protein	Tfg1	Lys 394	Dimerization domain	Rpb2	Lys 228	High
Inter-protein	Tfg1	Lys 394	Dimerization domain	Rpb2	Lys 353	High
Inter-protein	Tfg1	Lys 411	Charged region	Rpb9	Lys 20	High
Inter-protein	Tfg1	Lys 411	Charged region	Rpb2	Lys 270	High
Inter-protein	Tfg1	Lys 411	Charged region	Rpb2	Lys 344	High
Inter-protein	Tfg1	Lys 416	Charged region	Rpb1	Lys 186	High
Inter-protein	Tfg1	Lys 416	Charged region	Rpb2	Lys 270	High
Inter-protein	Tfg1	Lys 416	Charged region	Rpb2	Lys 277	High
Inter-protein	Tfg1	Lys 416	Charged region	Rpb1	Lys 1262	High
Inter-protein	Tfg1	Lys 426	Charged region	Rpb9	Lys 20	High
Inter-protein	Tfg1	Lys 426	Charged region	Rpb1	Lys 143	High
Inter-protein	Tfg1	Lys 426	Charged region	Rpb1	Lys 176	High
Inter-protein	Tfg1	Lys 426	Charged region	Rpb2	Lys 228	High
Inter-protein	Tfg1	Lys 426	Charged region	Rpb1	Lys 1221	High
Inter-protein	Tfg1	Lys 426	Charged region	Rpb1	Lys 1246	High
Inter-protein	Tfg1	Lys 426	Charged region	Rpb1	Lys 1262	High
Inter-protein	Tfg1	Lys 527		Rpb1	Lys 143	High
Inter-protein	Tfg1	Lys 527		Rpb1	Lys 186	High
Inter-protein	Tfg1	Lys 527		Rpb2	Lys 277	High
Inter-protein	Tfg1	Lys 537		Rpb2	Lys 344	High
Inter-protein	Tfg1	Lys 537		Rpb2	Lys 426	High
Inter-protein	Tfg1	Lys 614		Rpb1	Lys 49	High
Inter-protein	Tfg2	Lys 142	Dimerization domain	Rpb2	Lys 87	High
Inter-protein	Tfg2	Lys 142	Dimerization domain	Rpb2	Lys 148	High
Inter-protein	Tfg2	Lys 179	Insertion	Rpb2	Lys 344	High
Inter-protein	Tfg2	Lys 186	Insertion	Rpb2	Lys 344	High
Inter-protein	Tfg2	Lys 206	Dimerization domain	Rpb2	Lys 344	High
Inter-protein	Tfg2	Lys 235	Linker	Rpb2	Lys 606	High
Inter-protein	Tfg2	Lys 245	Linker	Rpb2	Lys 87	High
Inter-protein	Tfg2	Lys 245	Linker	Rpb2	Lys 246	High
Inter-protein	Tfg2	Lys 245	Linker	Rpb2	Lys 606	High
Inter-protein	Tfg2	Lys 249	Linker	Rpb2	Lys 246	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb1	Lys 49	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb10	Lys 68	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb2	Lys 99	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb2	Lys 133	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb3	Lys 149	High

Inter-protein	Tfg2	Lys 279	Linker	Rpb2	Lys 191	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb2	Lys 426	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb2	Lys 445	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb2	Lys 471	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb2	Lys 864	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb2	Lys 865	High
Inter-protein	Tfg2	Lys 279	Linker	Rpb2	Lys 934	High
Inter-protein	Tfg2	Lys 290	WH domain	Rpb1	Lys 49	High
Inter-protein	Tfg2	Lys 290	WH domain	Rpb2	Lys 87	High
Inter-protein	Tfg2	Lys 290	WH domain	Rpb3	Lys 149	High
Inter-protein	Tfg2	Lys 290	WH domain	Rpb2	Lys 426	High
Inter-protein	Tfg2	Lys 290	WH domain	Rpb2	Lys 865	High
Inter-protein	Tfg2	Lys 297	WH domain	Rpb1	Lys 34	High
Inter-protein	Tfg2	Lys 297	WH domain	Rpb1	Lys 49	High
Inter-protein	Tfg2	Lys 297	WH domain	Rpb2	Lys 426	High
Inter-protein	Tfg2	Lys 316	WH domain	Rpb2	Lys 426	High
Inter-protein	Tfg2	Lys 319	WH domain	Rpb2	Lys 87	High
Inter-protein	Tfg2	Lys 319	WH domain	Rpb2	Lys 344	High
Inter-protein	Tfg2	Lys 319	WH domain	Rpb2	Lys 426	High
Inter-protein	Tfg2	Lys 330	WH domain	Rpb2	Lys 864	High
Inter-protein	Tfg2	Lys 341	WH domain	Rpb2	Lys 87	High
Inter-protein	Tfg2	Lys 342	WH domain	Rpb10	Lys 59	High
Inter-protein	Tfg2	Lys 342	WH domain	Rpb2	Lys 426	High
Inter-protein	Tfg2	Lys 348	WH domain	Rpb10	Lys 59	High
Inter-protein	Tfg2	Lys 348	WH domain	Rpb2	Lys 426	High
Inter-protein	Tfg2	Lys 348	WH domain	Rpb2	Lys 813	High
Inter-protein	Tfg2	Lys 348	WH domain	Rpb2	Lys 864	High
Inter-protein	Tfg2	Lys 357		Rpb1	Lys 34	High
Inter-protein	Tfg2	Lys 357		Rpb1	Lys 49	High
Inter-protein	Tfg2	Lys 357		Rpb3	Lys 137	High
Inter-protein	Tfg2	Lys 357		Rpb2	Lys 426	High
Inter-protein	Tfg2	Lys 357		Rpb2	Lys 864	High
Inter-protein	Tfg2	Lys 357		Rpb2	Lys 865	High
Inter-protein	Tfg2	Lys 357		Rpb2	Lys 934	High
Inter-protein	Tfg2	Lys 359		Rpb1	Lys 49	High
Inter-protein	Tfg2	Lys 359		Rpb3	Lys 137	High
Inter-protein	Tfg2	Lys 359		Rpb2	Lys 865	High
Inter-protein	Tfg3	Lys 166		Rpb1	Lys 212	High
Inter-protein	Tfg3	Lys 181		Rpb1	Lys 34	High
Inter-protein	Tfg3	Lys 181		Rpb1	Lys 49	High
Inter-protein	Tfg3	Lys 240		Rpb1	Lys 49	High

Table S4 - Quantified cross-linkages in conformational comparison of C3 and C3b by quantitative 3D proteomics

Linkage site 1	Domain 1	Linkage site 2	Domain 2	Category	C3/C3b signal ratio						C-α distance (Å)	
					Sample 1	Sample 1R	Sample 2	Sample 2R	GM[1]	Log_2GM	PDB|2a73	PDB|2i07
Ser727	α'NT	K857	MG7	Cluster 6	<0.02 [3]	<0.01[3]	<0.005[3]	<0.005[3]	C3b	C3b	48.3*[2] (L729*-K857)	16.5* (D730*-K857)
K839	MG7	K1359	MG8	Cluster 6	N/A	N/A	<0.01[3]	<0.02[3]	C3b	C3b	26.3	23.0
K727	α'NT	K1359	MG8	Cluster 6	N/A	N/A	<0.02[3]	<0.01[3]	C3b	C3b	18.3* (L729*-K1359)	26.1* (D730*-K1359)
K839	MG7	K1346	MG8	Cluster 6	N/A	N/A	<0.02[3]	<0.01[3]	C3b	C3b	41.4	31.7
K44	MG1	K1181	TED	Cluster 5	0.8	0.2	0.6	0.3	0.4	-1.3	68.7	38.3
K44	MG1	K1028	TED	Cluster 5	0.9	0.2	0.8	0.2	0.4	-1.2	47.7	18.7
K44	MG1	K1019	TED	Cluster 5	1.0	0.2	N/A	N/A	0.4	-1.2	54.1	20.0
K1475	Anchor	K1567	C345C	Cluster 5	0.8	0.3	0.6	0.3	0.5	-1.1	13.8	8.2
K908	MG7	K1331	MG8	Cluster 5	0.7	0.3	N/A	N/A	0.5	-1.1	23.3	22.2
K1019	TED	K1029	TED	Cluster 5	0.8	0.3	0.7	0.3	0.5	-1.1	15.5	15.4
K75	MG1	K1181	TED	Cluster 5	0.9	0.3	N/A	N/A	0.5	-1.0	58.2* (G73*-K1181)	39.6* (R80*-K1181)
K857	MG7	K1475	Anchor	Cluster 5	0.8	0.3	N/A	N/A	0.5	-1.0	35.2	21.3
K586	LNK	K593	LNK	undefined	1.0	0.4	0.8	0.4	0.6	-0.7	9.8	12.1
K267	MG3	K1409	MG8	Cluster 4	N/A	N/A	0.9	0.6	0.7	-0.5	35.0	17.3
K337	MG4	K600	LNK	Cluster 4	0.9	0.6	N/A	N/A	0.7	-0.4	13.9	14.4
K585	LNK	K593	LNK	Cluster 4	1.0	0.6	1.0	0.5	0.7	-0.4	12.4	14.2
K1504	C345C	K1513	C345C	Cluster 4	1.4	0.4	N/A	N/A	0.8	-0.4	13.4	14.0
K1029	TED	K1284	CUB2	Cluster 4	0.6	1.0	N/A	N/A	0.8	-0.3	44.3	64.2
K133	MG2	K480	MG5	Cluster 4	1.3	0.6	N/A	N/A	0.9	-0.1	14.6	14.2
K1181	TED	K1193	TED	Cluster 4	1.4	0.6	N/A	N/A	0.9	-0.1	20.1	20.1
K857	MG7	K1567	C345C	Cluster 4	1.3	0.6	N/A	N/A	0.9	-0.1	23.9	19.1
K937	CUB	K1284	CUB	Cluster 4	1.3	0.7	N/A	N/A	1.0	-0.1	14.5	14.8

Ser1	MG1	K464	MG5	Cluster 4	N/A	N/A	1.4	0.7	1.0	0.0	14.0	14.2
K1019	TED	K1284	CUB2	Cluster 4	1.3	1.0	N/A	N/A	1.1	0.2	33.8	57.7
K44	MG1	K75	MG1	Cluster 4	1.7	0.6	3.3	0.6	1.2	0.2	13.0* (K44-E73*)	12.6* (K44-R80*)
K1475	Anchor	K1482	Anchor	Cluster 3	2.5	1.5	N/A	N/A	1.9	1.0	15.1	11.4
K44	MG1	K82	MG1	Cluster 3	3.4	1.4	N/A	N/A	2.2	1.1	13.3	13.7
K241	MG3	K869	MG7	Cluster 3	N/A	N/A	4.6	1.5	2.7	1.4	5.6	9.6
K857	MG7	K1504	C345C	Cluster 3	3.1	2.2	3.6	2.7	2.9	1.5	23.8	23.5
K857	MG7	K1500	C345C	Cluster 3	3.4	3.3	N/A	N/A	3.4	1.8	20.2	17.6
K666	ANA	K670	ANA	Cluster 2	13.2	3.8	11.1	6.9	7.9	3.0	8.3	N/A
K1346	MG8	K1475	Anchor	Cluster 2	7.9	6.2	15.5	5.5	8.1	3.0	12.1	12.8
K585	LNK	K588	LNK	Cluster 2	8.9	5.5	13.6	8.5	8.7	3.1	6.2	6.3
K586	LNK	K588	LNK	Cluster 2	14.9	6.8	N/A	N/A	10.1	3.3	5.4	5.5
K700	ANA	K1409	MG8	Cluster 2	14.8	3.3	109.5	40.0	21.6	4.4	25.6	N/A
K1331	MG8	K1475	Anchor	Cluster 1	>1.5[3]	>16[3]	N/A	N/A	C3	C3	16.2	26.6
K267	MG3	K283	MG3	Cluster 1	>2520[3]	>1370	>730.[3]	>520[3]	C3	C3	14.4	14.4
S650	ANA	K660	ANA	Cluster 1	>85[3]	>43[3]	N/A	N/A	C3	C3	15.4* (V651*-K660)	N/A
K666	ANA	K1049	TED	Cluster 1	>72[3]	>85[3]	N/A	N/A	C3	C3	20.0	N/A
K670	ANA	K1049	TED	Cluster 1	>34[3]	>83[3]	>117[3]	>103[3]	C3	C3	15.6	N/A
K905	MG7	K1414	MG8	Cluster 1	N/A	N/A	>83[3]	>43[3]	C3	C3	18.5	29.8
K908	MG7	K1414	MG8	Cluster 1	>87[3]	>50[3]	N/A	N/A	C3	C3	13.4	35.2

[1]. GM stands for geometric mean

[2]. the asterisk indicated the substitution of the observed residue with the nearest residue present in the crystal structure. Both the substitute residues *per se* and the C-α distance that was measured from the substitute residue in the crystal structure were marked with asterisk.

[3]. for the conformation specific cross-linkages (Cluster 1 and Cluster 6) that was with either heavy or light signal not detected, the C3/C3b ratio was defined based on the maximum peptide to noise ratio detected in the cross-linked peptides corresponding to these cross-linkages. The noise level was defined as 1E4, the average intensity of the sdochastic peaks in the raw data.

Table S5 - Ten most intense protein identified from the affinity purified *S. cerevisiae* Mad1-Mad2 complex

Intensity rank	SGD protein IDs	Protein name	Identified peptides	Sequence Coverage [%]	Mol. Weight [kDa]	relative intensity[%]
1[1]	**YJL030W**	**MAD2**	**16**	**55.1**	**22.3**	**39.4%**
2	**YGL086W**	**MAD1**	**88**	**55.1**	**87.7**	**21.5%**
3	YIL149C	MLP2	69	27.8	195.1	16.9%
4	YKR095W	MLP1	37	14.9	218.5	4.8%
5	YJR123W	RPS5	3	17.3	25.0	1.4%
6	YMR230W	RPS10B	1	17.1	12.7	1.0%
7	YAL005C	SSA1	9	19.8	69.7	1.0%
8	YIL002W-A	Putative protein	3	20.3	7.7	0.7%
9	YLR153C	ACS2	3	7.5	75.5	0.7%
10	YMR276W	DSK2	5	13.4	39.3	0.6%

[1]. Mad1-Mad2 complex component proteins are coloured in red.

Table S6 - Ten most intense protein identified from the affinity purified *S. cerevisiae* Ndc80 complex

Intensity rank	Protein SGD IDs	Protein name	Identified peptides	Sequence Coverage [%]	Mol. Weight [kDa]	relative intensity[%]
1[1]	**YIL144W**	**NDC 80**	**124**	**74.2**	**80**	**17.9%**
2	**YOL069W**	**NUF2**	**54**	**57**	**53**	**14.2%**
3	**YER018C**	**SPC25**	**18**	**42.5**	**25**	**10.0%**
4	**YMR117C**	**SPC24**	**38**	**84.5**	**25**	**7.3%**
5	YGL008C	PMA1	21	16.3	100	5.2%
6	YGR192C	TDH3	24	58.7	36	2.9%
7	YAL034W-A	MTW1	13	41.9	33	2.3%
8	YAL038W	CDC19	17	40.2	55	1.9%
9	YNL209W	SSB2	16	36.7	67	1.7%
10	YOL086C	ADH1	9	26.7	37	1.3%

[1]. Ndc80 complex component proteins are coloured in red.

Table S7 - List of cross-links observed from the affinity purified *S. cerevisiae* endogenous Mad1-Mad2 complex

Type	Protein 1	Link position 1	Protein 2	Link position 2	Confidence	SILAC labelled validation	Number of identified spectra
intra- or inter-protein	Mad1	94	Mad1	97	high	within complex	1
inter-protein	Mad1	95	Mad1	97	high	within complex	1
inter-protein	Mad1	119	Mad1	119	high	within complex	1
intra- or inter-protein	Mad1	119	Mad1	128	high	within complex	8
inter-protein	Mad1	152	Mad1	152	high	N/A	3
intra- or inter-protein	Mad1	152	Mad1	366	low	N/A	2
intra- or inter-protein	Mad1	193	Mad1	405	high	N/A	1
intra- or inter-protein	Mad1	196	Mad1	203	low	within complex	2
intra- or inter-protein	Mad1	196	Mad1	527	low	within complex	8
intra- or inter-protein	Mad1	268	Mad1	271	high	within complex	9
inter-protein	Mad1	269	Mad1	271	high	within complex	3
intra- or inter-protein	Mad1	312	Mad1	366	low	within complex	1
intra- or inter-protein	Mad1	314	Mad1	366	high	within complex	5
inter-protein	Mad1	366	Mad1	366	high	within complex	1
intra- or inter-protein	Mad1	366	Mad1	381	high	within complex	21
intra- or inter-protein	Mad1	366	Mad1	385	high	within complex	1
intra- or inter-protein	Mad1	366	Mad1	405	low	within complex	1
intra- or inter-protein	Mad1	366	Mad1	527	high	within complex	2
intra- or inter-protein	Mad1	366	Mad1	557	high	within complex	17
intra- or inter-protein	Mad1	366	Mad1	568	low	N/A	2
intra- or inter-protein	Mad1	366	Mad1	571	high	within complex	9
intra- or inter-protein	Mad1	366	Mad1	574	high	N/A	2
intra- or inter-protein	Mad1	366	Mad1	592	low	N/A	1
intra- or inter-protein	Mad1	405	Mad1	524	low	within complex	6
intra- or inter-protein	Mad1	405	Mad1	525	low	N/A	1
intra- or inter-protein	Mad1	405	Mad1	527	high	within complex	2

intra- or inter-protein	Mad1	423	Mad1	527	high	within complex	4
intra- or inter-protein	Mad1	433	Mad1	507	high	within complex	1
intra- or inter-protein	Mad1	433	Mad1	527	high	within complex	3
intra- or inter-protein	Mad1	495	Mad1	506	high	within complex	2
intra- or inter-protein	Mad1	497	Mad1	507	low	within complex	3
intra- or inter-protein	Mad1	507	Mad1	527	high	within complex	3
intra- or inter-protein	Mad1	524	Mad1	527	high	within complex	9
intra- or inter-protein	Mad1	525	Mad1	527	high	within complex	8
inter-protein	Mad1	527	Mad1	527	high	within complex	14
intra- or inter-protein	Mad1	527	Mad1	579	low	N/A	1
intra- or inter-protein	Mad1	579	Mad1	592	low	within complex	6
inter-protein	Mad1	592	Mad1	592	high	N/A	1
intra- or inter-protein	Mad1	592	Mad1	598	high	within complex	34
intra- or inter-protein	Mad1	592	Mad1	600	high	within complex	20
intra- or inter-protein	Mad1	592	Mad1	646	low	within complex	1
intra- or inter-protein	Mad1	592	Mad1	649	low	within complex	22
intra- or inter-protein	Mad1	592	Mad1	655	high	N/A	1
intra- or inter-protein	Mad1	597	Mad1	600	high	within complex	8
inter-protein	Mad1	598	Mad1	600	high	within complex	20
intra- or inter-protein	Mad1	598	Mad1	606	low	within complex	9
intra- or inter-protein	Mad1	649	Mad1	655	low	within complex	2
inter-protein	Mad1	592	Mad2	61	high	within complex	15
inter-protein	Mad1	649	Mad2	56	low	N/A	7
inter-protein	Mad1	649	Mad2	61	low	within complex	1

Table S8 - List of cross-links observed from the affinity purified *S. cerevisiae* endogenous Ndc80 complex

Name	Type	Protein 1	Link position 1	Protein 2	Link position 2	Confidence	SILAC labelled validation	Number of identified spectra
X1	intra-protein	Ndc80	48	Ndc80	67	high	within complex	1
X2	intra-protein	Ndc80	48	Ndc80	69	high	within complex	3
X3	intra-protein	Ndc80	404	Ndc80	409	high	within complex	1
X4	intra-protein	Ndc80	425	Ndc80	432	low	within complex	2
X5	intra-protein	Ndc80	445	Ndc80	448	high	within complex	5
X6	intra-protein	Ndc80	513	Ndc80	523	low	within complex	6
X7	intra-protein	Ndc80	548	Ndc80	554	high	within complex	1
X8	intra-protein	Ndc80	598	Ndc80	602	high	within complex	15
X9	intra-protein	Ndc80	602	Ndc80	613	high	N/A	1
X10	intra-protein	Ndc80	627	Ndc80	638	low	within complex	3
X11	intra-protein	Nuf2	388	Nuf2	393	high	within complex	2
X12	intra-protein	Spc24	32	Spc24	42	high	within complex	2
X13	intra-protein	Spc24	42	Spc24	98	high	within complex	1
X14	intra-protein	Spc24	62	Spc24	98	high	N/A	1
X15	intra-protein	Spc24	98	Spc24	163	high	within complex	3
X16	intra-protein	Spc24	98	Spc24	205	high	within complex	1
X17	inter-protein	Ndc80	377	Nuf2	220	high	within complex	1
X18	inter-protein	Ndc80	425	Nuf2	270	high	within complex	1
X19	inter-protein	Ndc80	577	Nuf2	366	high	within complex	2
X20	inter-protein	Ndc80	582	Nuf2	360	high	within complex	2
X21	inter-protein	Ndc80	582	Nuf2	366	high	within complex	1
X22	inter-protein	Ndc80	602	Nuf2	388	high	within complex	2
X23	inter-protein	Ndc80	602	Nuf2	393	high	within complex	1
X24	inter-protein	Ndc80	611	Nuf2	393	high	within complex	3
X25	inter-protein	Ndc80	627	Nuf2	409	high	within complex	1
X26	inter-protein	Ndc80	627	Nuf2	415	high	within complex	5
X27	inter-protein	Spc24	98	Spc25	19	high	N/A	1

X28	inter-protein	Spc24	163	Spc25	19	high	within complex	1
X29	inter-protein	Ndc80	611	Nuf2	388	low	within complex	1
N/A	intra-protein	Ndc80	48	Ndc80	344	low	N/A	2
N/A	intra-protein	Ndc80	338	Ndc80	344	low	N/A	1
N/A	intra-protein	Ndc80	349	Ndc80	359	low	N/A	2
N/A	intra-protein	Ndc80	404	Ndc80	408	low	N/A	1
N/A	intra-protein	Ndc80	554	Ndc80	566	low	N/A	2
N/A	intra-protein	Spc24	163	Spc24	184	low	N/A	1

A.4 Publications

1. Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry.
Chen ZA, Jawhari A, Fischer L, Buchen C, Tahir S, Kamenski T, Rasmussen M, Lariviere L, Bukowski-Wills JC, Nilges M, Cramer P, Rappsilber J.
EMBO J. 2010 Feb 17; 29(4):717-26. Epub 2010 Jan 21.

2. A genetic engineering solution to the "arginine conversion problem" in stable isotope labelling by amino acids in cell culture (SILAC).
Bicho CC, de Lima Alves F, Chen ZA, Rappsilber J, Sawin KE.
Mol Cell Proteomics. 2010 Jul; 9(7):1567-77. Epub 2010 May 10.

3. The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics.
Ohta S, Bukowski-Wills JC, Sanchez-Pulido L, Alves Fde L, Wood L, Chen ZA, Platani M, Fischer L, Hudson DF, Ponting CP, Fukagawa T, Earnshaw WC, Rappsilber J.
Cell. 2010 Sep 3;142(5):810-21.

CITED LITERATURES

Abdul Ajees, A., K. Gunasekaran, J. E. Volanakis, S. V. Narayana, G. J. Kotwal and H. M. Murthy (2006). "The structure of complement C3b provides insights into complement activation and regulation." Nature **444**(7116): 221-5.

Aebersold, R. and M. Mann (2003). "Mass spectrometry-based proteomics." Nature **422**(6928): 198-207.

Alber, F., S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B. T. Chait, M. P. Rout and A. Sali (2007). "Determining the architectures of macromolecular assemblies." Nature **450**(7170): 683-94.

Alsenz, J., J. D. Becherer, B. Nilsson and J. D. Lambris (1990). "Structural and functional analysis of C3 using monoclonal antibodies." Curr Top Microbiol Immunol **153**: 235-48.

Anderson, G. A., N. Tolic, X. Tang, C. Zheng and J. E. Bruce (2007). "Informatics strategies for large-scale novel cross-linking analysis." J Proteome Res **6**(9): 3412-21.

Armache, K. J., H. Kettenberger and P. Cramer (2003). "Architecture of initiation-competent 12-subunit RNA polymerase II." Proc Natl Acad Sci U S A **100**(12): 6964-8.

Armache, K. J., S. Mitterweger, A. Meinhart and P. Cramer (2005). "Structures of complete RNA polymerase II and its subcomplex, Rpb4/7." J Biol Chem **280**(8): 7131-4.

Asturias, F. J. (2004). "Another piece in the transcription initiation puzzle." Nat Struct Mol Biol **11**(11): 1031-3.

Back, J. W., L. de Jong, A. O. Muijsers and C. G. de Koster (2003). "Chemical cross-linking and mass spectrometry for protein structural modeling." J Mol Biol **331**(2): 303-13.

Back, J. W., V. Notenboom, L. J. de Koning, A. O. Muijsers, T. K. Sixma, C. G. de Koster and L. de Jong (2002). "Identification of cross-linked peptides for protein interaction studies using mass spectrometry and 18O labeling." Anal Chem **74**(17): 4417-22.

Berger, B., D. B. Wilson, E. Wolf, T. Tonchev, M. Milla and P. S. Kim (1995). "Predicting coiled coils by use of pairwise residue correlations." Proc Natl Acad Sci U S A **92**(18): 8259-63.

Bhat, S., M. G. Sorci-Thomas, E. T. Alexander, M. P. Samuel and M. J. Thomas (2005). "Intermolecular contact between globular N-terminal fold and C-terminal domain of ApoA-I stabilizes its lipid-bound conformation: studies employing chemical cross-linking and mass spectrometry." J Biol Chem **280**(38): 33015-25.

Bich, C., S. Maedler, K. Chiesa, F. DeGiacomo, N. Bogliotti and R. Zenobi (2010). "Reactivity and applications of new amine reactive cross-linkers for mass spectrometric detection of protein-protein complexes." Anal Chem **82**(1): 172-9.

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

Bohn, S., F. Beck, E. Sakata, T. Walzthoeni, M. Beck, R. Aebersold, F. Forster, W. Baumeister and S. Nickell (2010). "Structure of the 26S proteasome from Schizosaccharomyces pombe at subnanometer resolution." Proc Natl Acad Sci U S A **107**(49): 20992-7.

Borrell, B. (2009). "Fraud rocks protein community." Nature **462**(7276): 970.

Brunner, J. (1993). "New photolabeling and crosslinking methods." Annu Rev Biochem **62**: 483-514.

Burton, Z. F., M. Killeen, M. Sopta, L. G. Ortolan and J. Greenblatt (1988). "RAP30/74: a general initiation factor that binds to RNA polymerase II." Mol Cell Biol **8**(4): 1602-13.

Bushnell, D. A., C. Bamdad and R. D. Kornberg (1996). "A minimal set of RNA polymerase II transcription protein interactions." J Biol Chem **271**(33): 20170-4.

Bushnell, D. A. and R. D. Kornberg (2003). "Complete, 12-subunit RNA polymerase II at 4.1-A resolution: implications for the initiation of transcription." Proc Natl Acad Sci U S A **100**(12): 6969-73.

Carroll, M. C. (2004). "The complement system in regulation of adaptive immunity." Nat Immunol **5**(10): 981-6.

Chambers, R. S., B. Q. Wang, Z. F. Burton and M. E. Dahmus (1995). "The activity of COOH-terminal domain phosphatase is regulated by a docking site on RNA polymerase II and by the general transcription factors IIF and IIB." J Biol Chem **270**(25): 14962-9.

Chang, Z., J. Kuchar and R. P. Hausinger (2004). "Chemical cross-linking and mass spectrometric identification of sites of interaction for UreD, UreF, and urease." J Biol Chem **279**(15): 15305-13.

Charles A Janeway, J., Paul Travers, Mark Walport, Mark J Shlomchik. (2001). Immunobiology: the Immune system in Health and Disease. New York, Garland Science.

Chen, H. T., L. Warfield and S. Hahn (2007). "The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex." Nat Struct Mol Biol **14**(8): 696-703.

Chen, Z. A., A. Jawhari, L. Fischer, C. Buchen, S. Tahir, T. Kamenski, M. Rasmussen, L. Lariviere, J. C. Bukowski-Wills, M. Nilges, P. Cramer and J. Rappsilber (2010). "Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry." EMBO J **29**(4): 717-26.

Chowdhury, S. M., X. Du, N. Tolic, S. Wu, R. J. Moore, M. U. Mayer, R. D. Smith and J. N. Adkins (2009). "Identification of cross-linked peptides after click-based enrichment using sequential collision-induced dissociation and electron transfer dissociation tandem mass spectrometry." Anal Chem **81**(13): 5524-32.

Chu, F., P. R. Baker, A. L. Burlingame and R. J. Chalkley (2010). "Finding chimeras: a bioinformatics strategy for identification of cross-linked peptides." Mol Cell Proteomics **9**(1): 25-31.

Chu, F., J. C. Maynard, G. Chiosis, C. V. Nicchitta and A. L. Burlingame (2006). "Identification of novel quaternary domain interactions in the Hsp90 chaperone, GRP94." Protein Sci **15**(6): 1260-9.

Chu, F., S. O. Shan, D. T. Moustakas, F. Alber, P. F. Egea, R. M. Stroud, P. Walter and A. L. Burlingame (2004). "Unraveling the interface of signal recognition particle and its receptor by using chemical cross-linking and tandem mass spectrometry." Proc Natl Acad Sci U S A **101**(47): 16454-9.

Chung, W. H., J. L. Craighead, W. H. Chang, C. Ezeokonkwo, A. Bareket-Samish, R. D. Kornberg and F. J. Asturias (2003). "RNA polymerase II/TFIIF structure and conserved organization of the initiation complex." Mol Cell **12**(4): 1003-13.

Ciferri, C., J. De Luca, S. Monzani, K. J. Ferrari, D. Ristic, C. Wyman, H. Stark, J. Kilmartin, E. D. Salmon and A. Musacchio (2005). "Architecture of the human ndc80-hec1 complex, a critical constituent of the outer kinetochore." J Biol Chem **280**(32): 29088-95.

Ciferri, C., A. Musacchio and A. Petrovic (2007). "The Ndc80 complex: hub of kinetochore activity." FEBS Lett **581**(15): 2862-9.

Ciferri, C., S. Pasqualato, E. Screpanti, G. Varetti, S. Santaguida, G. Dos Reis, A. Maiolica, J. Polka, J. G. De Luca, P. De Wulf, M. Salek, J. Rappsilber, C. A. Moores, E. D. Salmon and A. Musacchio (2008). "Implications for kinetochore-microtubule attachment from the structure of an engineered Ndc80 complex." Cell **133**(3): 427-39.

Clegg, C. and D. Hayes (1974). "Identification of Neighbouring Proteins in the Ribosomes of Escherichia coli

A Topographical Study with the Cross-Linking Reagent Dimethyl Suberimidate." European Journal of Biochemistry **42**(1): 21-28.

Collins, C. J., B. Schilling, M. Young, G. Dollinger and R. K. Guy (2003). "Isotopically labeled crosslinking reagents: resolution of mass degeneracy in the identification of crosslinked peptides." Bioorg Med Chem Lett **13**(22): 4023-6.

Conaway, J. W. and R. C. Conaway (1990). "An RNA polymerase II transcription factor shares functional properties with Escherichia coli sigma 70." Science **248**(4962): 1550-3.

Cox, J. and M. Mann (2008). "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification." Nat Biotechnol **26**(12): 1367-72.

Cramer, P. (2007). "Finding the right spot to start transcription." Nat Struct Mol Biol **14**(8): 686-7.

DeLano, W. L. (2002). The PyMOL Molecular Graphics System, DeLano Scientific, San Carlos, CA, USA.

Dihazi, G. H. and A. Sinz (2003). "Mapping low-resolution three-dimensional protein structures using chemical cross-linking and Fourier transform ion-cyclotron resonance mass spectrometry." Rapid Commun Mass Spectrom **17**(17): 2005-14.

Dimova, K., S. Kalkhof, I. Pottratz, C. Ihling, F. Rodriguez-Castaneda, T. Liepold, C. Griesinger, N. Brose, A. Sinz and O. Jahn (2009). "Structural insights into the calmodulin-Munc13 interaction obtained by cross-linking and mass spectrometry." Biochemistry **48**(25): 5908-21.

Dvir, A., J. W. Conaway and R. C. Conaway (2001). "Mechanism of transcription initiation and promoter escape by RNA polymerase II." Curr Opin Genet Dev **11**(2): 209-14.

Englander, J. J., C. Del Mar, W. Li, S. W. Englander, J. S. Kim, D. D. Stranz, Y. Hamuro and V. L. Woods, Jr. (2003). "Protein structure change studied by hydrogen-deuterium exchange, functional labeling, and mass spectrometry." Proc Natl Acad Sci U S A **100**(12): 7057-62.

Fang, G., H. Yu and M. W. Kirschner (1998). "The checkpoint protein MAD2 and the mitotic regulator CDC20 form a ternary complex with the anaphase-promoting complex to control anaphase initiation." Genes Dev **12**(12): 1871-83.

Fishelson, Z., M. K. Pangburn and H. J. Muller-Eberhard (1984). "Characterization of the initial C3 convertase of the alternative pathway of human complement." J Immunol **132**(3): 1430-4.

Flores, O., H. Lu, M. Killeen, J. Greenblatt, Z. F. Burton and D. Reinberg (1991). "The small subunit of transcription factor IIF recruits RNA polymerase II into the preinitiation complex." Proc Natl Acad Sci U S A **88**(22): 9999-10003.

Flores, O., E. Maldonado and D. Reinberg (1989). "Factors involved in specific transcription by mammalian RNA polymerase II. Factors IIE and IIF independently interact with RNA polymerase II." J Biol Chem **264**(15): 8913-21.

Forneris, F., D. Ricklin, J. Wu, A. Tzekou, R. S. Wallace, J. D. Lambris and P. Gros (2010). "Structures of C3b in complex with factors B and D give insight into complement convertase formation." Science **330**(6012): 1816-20.

Forwood, J. K., A. S. Thakur, G. Guncar, M. Marfori, D. Mouradov, W. Meng, J. Robinson, T. Huber, S. Kellie, J. L. Martin, D. A. Hume and B. Kobe (2007). "Structural basis for recruitment of tandem hotdog domains in acyl-CoA

thioesterase 7 and its role in inflammation." Proc Natl Acad Sci U S A **104**(25): 10382-7.

Freire-Picos, M. A., S. Krishnamurthy, Z. W. Sun and M. Hampsey (2005). "Evidence that the Tfg1/Tfg2 dimer interface of TFIIF lies near the active center of the RNA polymerase II initiation complex." Nucleic Acids Res **33**(16): 5045-52.

Fu, C. Y., C. Uetrecht, S. Kang, M. C. Morais, A. J. Heck, M. R. Walter and P. E. Prevelige, Jr. (2010). "A docking model based on mass spectrometric and biochemical data describes phage packaging motor incorporation." Mol Cell Proteomics **9**(8): 1764-73.

Funk, J. D., Y. A. Nedialkov, D. Xu and Z. F. Burton (2002). "A key role for the alpha 1 helix of human RAP74 in the initiation and elongation of RNA chains." J Biol Chem **277**(49): 46998-7003.

Gaiser, F., S. Tan and T. J. Richmond (2000). "Novel dimerization fold of RAP30/RAP74 in human TFIIF at 1.7 A resolution." J Mol Biol **302**(5): 1119-27.

Gao, Q., S. Xue, C. E. Doneanu, S. A. Shaffer, D. R. Goodlett and S. D. Nelson (2006). "Pro-CrossLink. Software tool for protein cross-linking and mass spectrometry." Anal Chem **78**(7): 2145-9.

Gardner, M. W. and J. S. Brodbelt (2008). "Impact of proline and aspartic acid residues on the dissociation of intermolecularly crosslinked peptides." J Am Soc Mass Spectrom **19**(3): 344-57.

Garrett, K. P., H. Serizawa, J. P. Hanley, J. N. Bradsher, A. Tsuboi, N. Arai, T. Yokota, K. Arai, R. C. Conaway and J. W. Conaway (1992). "The carboxyl terminus of RAP30 is similar in sequence to region 4 of bacterial sigma factors and is required for function." J Biol Chem **267**(33): 23942-9.

Gaucher, S. P., M. Z. Hadi and M. M. Young (2006). "Influence of crosslinker identity and position on gas-phase dissociation of Lys-Lys crosslinked peptides." J Am Soc Mass Spectrom **17**(3): 395-405.

Gertz, M., H. Seelert, N. A. Dencher and A. Poetsch (2007). "Interactions of rotor subunits in the chloroplast ATP synthase modulated by nucleotides and by Mg2+." Biochim Biophys Acta **1774**(5): 566-74.

Ghazy, M. A., S. A. Brodie, M. L. Ammerman, L. M. Ziegler and A. S. Ponticelli (2004). "Amino acid substitutions in yeast TFIIF confer upstream shifts in transcription initiation and altered interaction with RNA polymerase II." Mol Cell Biol **24**(24): 10975-85.

Gingras, A. C., M. Gstaiger, B. Raught and R. Aebersold (2007). "Analysis of protein complexes using mass spectrometry." Nat Rev Mol Cell Biol **8**(8): 645-54.

Green, N. S., E. Reisler and K. N. Houk (2001). "Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers." Protein Sci **10**(7): 1293-304.

Groft, C. M., S. N. Uljon, R. Wang and M. H. Werner (1998). "Structural homology between the Rap30 DNA-binding domain and linker histone H5: implications for preinitiation complex assembly." Proc Natl Acad Sci U S A **95**(16): 9117-22.

Gros, P., F. J. Milder and B. J. Janssen (2008). "Complement driven by conformational changes." Nat Rev Immunol **8**(1): 48-58.

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol **17**(10): 994-9.

Hack, C. E., J. Paardekooper, A. J. Eerenberg, G. O. Navis, M. W. Nijsten, L. G. Thijs and J. H. Nuijens (1988). "A modified competitive inhibition radioimmunoassay for the detection of C3a. Use of 125I-C3 instead of 125I-C3a." J Immunol Methods **108**(1-2): 77-84.

Hack, C. E., J. Paardekooper and F. Van Milligen (1990). "Demonstration in human plasma of a form of C3 that has the conformation of "C3b-like C3"." J Immunol **144**(11): 4249-55.

Hahn, S. (2004). "Structure and mechanism of the RNA polymerase II transcription machinery." Nat Struct Mol Biol **11**(5): 394-403.

Hardman, M. and A. A. Makarov (2003). "Interfacing the orbitrap mass analyzer to an electrospray ion source." Anal Chem **75**(7): 1699-705.

Hardwick, K. G. (1998). "The spindle checkpoint." Trends Genet **14**(1): 1-4.

Henry, N. L., A. M. Campbell, W. J. Feaver, D. Poon, P. A. Weil and R. D. Kornberg (1994). "TFIIF-TAF-RNA polymerase II connection." Genes Dev **8**(23): 2868-78.

Heyduk, T. (2002). "Measuring protein conformational changes by FRET/LRET." Curr Opin Biotechnol **13**(4): 292-6.

Huang, B. X. and H. Y. Kim (2006). "Interdomain conformational changes in Akt activation revealed by chemical cross-linking and tandem mass spectrometry." Mol Cell Proteomics **5**(6): 1045-53.

Huang, B. X., H. Y. Kim and C. Dass (2004). "Probing three-dimensional structure of bovine serum albumin by chemical cross-linking and mass spectrometry." J Am Soc Mass Spectrom **15**(8): 1237-47.

Huermanson, G. (1996). Bioconjucate techniques. San Diego, CA:Academic Press.

Hwang, L. H., L. F. Lau, D. L. Smith, C. A. Mistrot, K. G. Hardwick, E. S. Hwang, A. Amon and A. W. Murray (1998). "Budding yeast Cdc20: a target of the spindle checkpoint." Science **279**(5353): 1041-4.

Iglesias, A. H., L. F. Santos and F. C. Gozzo (2009). "Collision-induced dissociation of Lys-Lys intramolecular crosslinked peptides." J Am Soc Mass Spectrom **20**(4): 557-66.

Iglesias, A. H., L. F. Santos and F. C. Gozzo (2010). "Identification of cross-linked peptides by high-resolution precursor ion scan." Anal Chem **82**(3): 909-16.

Isenman, D. E. (1983). "Conformational changes accompanying proteolytic cleavage of human complement protein C3b by the regulatory enzyme factor I and its cofactor H. Spectroscopic and enzymological studies." J Biol Chem **258**(7): 4238-44.

Isenman, D. E., D. I. Kells, N. R. Cooper, H. J. Muller-Eberhard and M. K. Pangburn (1981). "Nucleophilic modification of human complement protein C3: correlation of conformational changes with acquisition of C3b-like functional properties." Biochemistry **20**(15): 4458-67.

Ishihama, Y., J. Rappsilber and M. Mann (2006). "Modular stop and go extraction tips with stacked disks for parallel and multidimensional Peptide fractionation in proteomics." J Proteome Res **5**(4): 988-94.

Ishima, R. and D. A. Torchia (2000). "Protein dynamics from NMR." Nat Struct Biol **7**(9): 740-3.

Izban, M. G. and D. S. Luse (1992). "Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates." J Biol Chem **267**(19): 13647-55.

Jacobsen, R. B., K. L. Sale, M. J. Ayson, P. Novak, J. Hong, P. Lane, N. L. Wood, G. H. Kruppa, M. M. Young and J. S. Schoeniger (2006). "Structure and dynamics of dark-state bovine rhodopsin revealed by chemical cross-linking and high-resolution mass spectrometry." Protein Sci **15**(6): 1303-17.

Janssen, B. J., A. Christodoulidou, A. McCarthy, J. D. Lambris and P. Gros (2006). "Structure of C3b reveals conformational changes that underlie complement activity." Nature **444**(7116): 213-6.

Janssen, B. J., E. G. Huizinga, H. C. Raaijmakers, A. Roos, M. R. Daha, K. Nilsson-Ekdahl, B. Nilsson and P. Gros (2005). "Structures of complement component C3 provide insights into the function and evolution of immunity." Nature **437**(7058): 505-11.

Jin Lee, Y. (2008). "Mass spectrometric analysis of cross-linking sites for the structure of proteins and protein complexes." Mol Biosyst **4**(8): 816-23.

Jurneczko, E. and P. E. Barran (2011). "How useful is ion mobility mass spectrometry for structural biology? The relationship between protein crystal structures and their collision cross sections in the gas phase." Analyst **136**(1): 20-8.

Kallio, M., J. Weinstein, J. R. Daum, D. J. Burke and G. J. Gorbsky (1998). "Mammalian p55CDC mediates association of the spindle checkpoint protein Mad2 with the cyclosome/anaphase-promoting complex, and is involved in regulating anaphase onset and late mitotic events." J Cell Biol **141**(6): 1393-406.

Kamada, K., J. De Angelis, R. G. Roeder and S. K. Burley (2001). "Crystal structure of the C-terminal domain of the RAP74 subunit of human transcription factor IIF." Proc Natl Acad Sci U S A **98**(6): 3115-20.

Kamada, K., R. G. Roeder and S. K. Burley (2003). "Molecular mechanism of recruitment of TFIIF- associating RNA polymerase C-terminal domain phosphatase (FCP1) by transcription factor IIF." Proc Natl Acad Sci U S A **100**(5): 2296-9.

Kang, S., L. Mou, J. Lanman, S. Velu, W. J. Brouillette and P. E. Prevelige, Jr. (2009). "Synthesis of biotin-tagged chemical cross-linkers and their applications for mass spectrometry." Rapid Commun Mass Spectrom **23**(11): 1719-26.

Khaperskyy, D. A., M. L. Ammerman, R. C. Majovski and A. S. Ponticelli (2008). "Functions of Saccharomyces cerevisiae TFIIF during transcription start site utilization." Mol Cell Biol **28**(11): 3757-66.

Killeen, M. T. and J. F. Greenblatt (1992). "The general transcription factor RAP30 binds to RNA polymerase II and prevents it from binding nonspecifically to DNA." Mol Cell Biol **12**(1): 30-7.

Kim, S. H., D. P. Lin, S. Matsumoto, A. Kitazono and T. Matsumoto (1998). "Fission yeast Slp1: an effector of the Mad2-dependent spindle checkpoint." Science **279**(5353): 1045-7.

Kim, T. K., T. Lagrange, Y. H. Wang, J. D. Griffith, D. Reinberg and R. H. Ebright (1997). "Trajectory of DNA in the RNA polymerase II transcription preinitiation complex." Proc Natl Acad Sci U S A **94**(23): 12268-73.

Kitatsuji, C., M. Kurogochi, S. Nishimura, K. Ishimori and K. Wakasugi (2007). "Molecular basis of guanine nucleotide dissociation inhibitor activity of human neuroglobin by chemical cross-linking and mass spectrometry." J Mol Biol **368**(1): 150-60.

Kluger, R. and A. Alagic (2004). "Chemical cross-linking and protein-protein interactions-a review with illustrative protocols." Bioorg Chem **32**(6): 451-72.

Kobor, M. S., L. D. Simon, J. Omichinski, G. Zhong, J. Archambault and J. Greenblatt (2000). "A motif shared by TFIIF and TFIIB mediates their interaction with the RNA polymerase II carboxy-terminal domain phosphatase Fcp1p in Saccharomyces cerevisiae." Mol Cell Biol **20**(20): 7438-49.

Kodadek, T., I. Duroux-Richard and J. C. Bonnafous (2005). "Techniques: Oxidative cross-linking as an emergent tool for the analysis of receptor-mediated signalling events." Trends Pharmacol Sci **26**(4): 210-7.

Kornberg, R. D. (1999). "Eukaryotic transcriptional control." Trends Cell Biol **9**(12): M46-9.

Kostrewa, D., M. E. Zeller, K. J. Armache, M. Seizl, K. Leike, M. Thomm and P. Cramer (2009). "RNA polymerase II-TFIIB structure and mechanism of transcription initiation." Nature **462**(7271): 323-30.

Krauth, F., C. H. Ihling, H. H. Ruttinger and A. Sinz (2009). "Heterobifunctional isotope-labeled amine-reactive photo-cross-linker for structural investigation of proteins by matrix-assisted laser desorption/ionization tandem time-of-flight and electrospray ionization LTQ-Orbitrap mass spectrometry." Rapid Commun Mass Spectrom **23**(17): 2811-8.

Krissinel, E. and K. Henrick (2007). "Inference of macromolecular assemblies from crystalline state." J Mol Biol **372**(3): 774-97.

Lambris, J. D., D. Avila, J. D. Becherer and H. J. Muller-Eberhard (1988). "A discontinuous factor H binding site in the third component of complement as delineated by synthetic peptides." J Biol Chem **263**(24): 12147-50.

Law, S. K. and A. W. Dodds (1997). "The internal thioester and the covalent binding properties of the complement proteins C3 and C4." Protein Sci **6**(2): 263-74.

Law, S. K., N. A. Lichtenberg and R. P. Levine (1979). "Evidence for an ester linkage between the labile binding site of C3b and receptive surfaces." J Immunol **123**(3): 1388-94.

Lee, T. I. and R. A. Young (2000). "Transcription of eukaryotic protein-coding genes." Annu Rev Genet **34**: 77-137.

Lee, Y. J., L. L. Lackner, J. M. Nunnari and B. S. Phinney (2007). "Shotgun cross-linking analysis for studying quaternary and tertiary protein structures." J Proteome Res **6**(10): 3908-17.

Leitner, A., T. Walzthoeni, A. Kahraman, F. Herzog, O. Rinner, M. Beck and R. Aebersold (2010). "Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics." Mol Cell Proteomics **9**(8): 1634-49.

Li, K., J. Gor and S. J. Perkins (2010). "Self-association and domain rearrangements between complement C3 and C3u provide insight into the activation mechanism of C3." Biochem J **431**(1): 63-72.

Li, Y., C. Gorbea, D. Mahaffey, M. Rechsteiner and R. Benezra (1997). "MAD2 associates with the cyclosome/anaphase-promoting complex and inhibits its activity." Proc Natl Acad Sci U S A **94**(23): 12431-6.

Luo, X., Z. Tang, J. Rizo and H. Yu (2002). "The Mad2 spindle checkpoint protein undergoes similar major conformational changes upon binding to either Mad1 or Cdc20." Mol Cell **9**(1): 59-71.

Lupas, A., M. Van Dyke and J. Stock (1991). "Predicting coiled coils from protein sequences." Science **252**(5010): 1162-4.

Madler, S., M. Seitz, J. Robinson and R. Zenobi (2010). "Does chemical cross-linking with NHS esters reflect the chemical equilibrium of protein-protein

noncovalent interactions in solution?" J Am Soc Mass Spectrom **21**(10): 1775-83.

Maiolica, A., D. Cittaro, D. Borsotti, L. Sennels, C. Ciferri, C. Tarricone, A. Musacchio and J. Rappsilber (2007). "Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching." Mol Cell Proteomics **6**(12): 2200-11.

Makarov, A., E. Denisov, A. Kholomeev, W. Balschun, O. Lange, K. Strupat and S. Horning (2006). "Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer." Anal Chem **78**(7): 2113-20.

Mann, M. (2006). "Functional and quantitative proteomics using SILAC." Nat Rev Mol Cell Biol **7**(12): 952-8.

Mann, M. and O. N. Jensen (2003). "Proteomic analysis of post-translational modifications." Nat Biotechnol **21**(3): 255-61.

Maxon, M. E., J. A. Goodrich and R. Tjian (1994). "Transcription factor IIE binds preferentially to RNA polymerase IIa and recruits TFIIH: a model for promoter clearance." Genes Dev **8**(5): 515-24.

McAlister, G. C., D. Phanstiel, C. D. Wenger, M. V. Lee and J. J. Coon (2010). "Analysis of tandem mass spectra by FTMS for improved large-scale proteomics with superior protein quantification." Anal Chem **82**(1): 316-22.

McCracken, S. and J. Greenblatt (1991). "Related RNA polymerase-binding regions in human RAP30/74 and Escherichia coli sigma 70." Science **253**(5022): 900-2.

Melcher, K. (2004). "New chemical crosslinking methods for the identification of transient protein-protein interactions with multiprotein complexes." Curr Protein Pept Sci **5**(4): 287-96.

Mueller, D. R., P. Schindler, H. Towbin, U. Wirth, H. Voshol, S. Hoving and M. O. Steinmetz (2001). "Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis." Anal Chem **73**(9): 1927-34.

Murakami, K. S., S. Masuda and S. A. Darst (2002). "Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 A resolution." Science **296**(5571): 1280-4.

Nechifor, R. and K. S. Wilson (2007). "Crosslinking of translation factor EF-G to proteins of the bacterial ribosome before and after translocation." J Mol Biol **368**(5): 1412-25.

Nessen, M. A., G. Kramer, J. Back, J. M. Baskin, L. E. Smeenk, L. J. de Koning, J. H. van Maarseveen, L. de Jong, C. R. Bertozzi, H. Hiemstra and C. G. de Koster (2009). "Selective enrichment of azide-containing peptides from complex mixtures." J Proteome Res **8**(7): 3702-11.

Nguyen, B. D., K. L. Abbott, K. Potempa, M. S. Kobor, J. Archambault, J. Greenblatt, P. Legault and J. G. Omichinski (2003). "NMR structure of a complex containing the TFIIF subunit RAP74 and the RNA polymerase II

carboxyl-terminal domain phosphatase FCP1." Proc Natl Acad Sci U S A **100**(10): 5688-93.

Nishida, N., T. Walz and T. A. Springer (2006). "Structural transitions of complement component C3 and its activation products." Proc Natl Acad Sci U S A **103**(52): 19737-42.

Nyarko, A., K. Mosbahi, A. J. Rowe, A. Leech, M. Boter, K. Shirasu and C. Kleanthous (2007). "TPR-Mediated self-association of plant SGT1." Biochemistry **46**(40): 11331-41.

Olsen, J. V., B. Macek, O. Lange, A. Makarov, S. Horning and M. Mann (2007). "Higher-energy C-trap dissociation for peptide modification analysis." Nat Methods **4**(9): 709-12.

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 (2009). "A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed." Mol Cell Proteomics **8**(12): 2759-69.

Ong, S. E., B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey and M. Mann (2002). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." Mol Cell Proteomics **1**(5): 376-86.

Orphanides, G. and D. Reinberg (2002). "A unified theory of gene expression." Cell **108**(4): 439-51.

Pangburn, M. K., R. D. Schreiber and H. J. Muller-Eberhard (1981). "Formation of the initial C3 convertase of the alternative complement pathway. Acquisition of C3b-like activities by spontaneous hydrolysis of the putative thioester in native C3." J Exp Med **154**(3): 856-67.

Parvin, J. D. and P. A. Sharp (1993). "DNA topology and a minimal set of basal factors for transcription by RNA polymerase II." Cell **73**(3): 533-40.

Pearson, K. M., L. K. Pannell and H. M. Fales (2002). "Intramolecular cross-linking experiments on cytochrome c and ribonuclease A using an isotope multiplet method." Rapid Commun Mass Spectrom **16**(3): 149-59.

Perkins, S. J. and R. B. Sim (1986). "Molecular modelling of human complement component C3 and its fragments by solution scattering." Eur J Biochem **157**(1): 155-68.

Petrotchenko, E. V., V. K. Olkhovik and C. H. Borchers (2005). "Isotopically coded cleavable cross-linker for studying protein-protein interaction and protein complexes." Mol Cell Proteomics **4**(8): 1167-79.

Petrotchenko, E. V., J. J. Serpa and C. H. Borchers (2011). "An isotopically coded CID-cleavable biotinylated cross-linker for structural proteomics." Mol Cell Proteomics **10**(2): M110 001420.

Petrotchenko, E. V., K. Xiao, J. Cable, Y. Chen, N. V. Dokholyan and C. H. Borchers (2009). "BiPS, a photocleavable, isotopically coded, fluorescent cross-linker for structural proteomics." Mol Cell Proteomics **8**(2): 273-86.

Pierce, P. (2003/2004). Applications handbook and catalog. Rockford, IL.

Pimenova, T., A. Nazabal, B. Roschitzki, J. Seebacher, O. Rinner and R. Zenobi (2008). "Epitope mapping on bovine prion protein using chemical cross-linking and mass spectrometry." J Mass Spectrom **43**(2): 185-95.

Rani, P. G., J. A. Ranish and S. Hahn (2004). "RNA polymerase II (Pol II)-TFIIF and Pol II-mediator complexes: the major stable Pol II complexes and their activity in transcription initiation and reinitiation." Mol Cell Biol **24**(4): 1709-20.

Rappsilber, J. (2011). "The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes." J Struct Biol **173**(3): 530-40.

Rappsilber, J., Y. Ishihama and M. Mann (2003). "Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics." Anal Chem **75**(3): 663-70.

Rappsilber, J., M. Mann and Y. Ishihama (2007). "Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips." Nat Protoc **2**(8): 1896-906.

Rappsilber, J., S. Siniossoglou, E. C. Hurt and M. Mann (2000). "A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry." Anal Chem **72**(2): 267-75.

Reinberg, D., G. Orphanides, R. Ebright, S. Akoulitchev, J. Carcamo, H. Cho, P. Cortes, R. Drapkin, O. Flores, I. Ha, J. A. Inostroza, S. Kim, T. K. Kim, P. Kumar, T. Lagrange, G. LeRoy, H. Lu, D. M. Ma, E. Maldonado, A. Merino, F. Mermelstein, I. Olave, M. Sheldon, R. Shiekhattar, L. Zawel and et al. (1998). "The RNA polymerase II general transcription factors: past, present, and future." Cold Spring Harb Symp Quant Biol **63**: 83-103.

Richard S. Johnson, S. A. M., Klaus Biemann, John T. Stults, J. Throck Watson (1987). "Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine." Anal. Chem. **59**(21): 2621-2625.

Rinner, O., J. Seebacher, T. Walzthoeni, L. N. Mueller, M. Beck, A. Schmidt, M. Mueller and R. Aebersold (2008). "Identification of cross-linked peptides from large sequence databases." Nat Methods **5**(4): 315-8.

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

Ross, G. D., J. D. Lambris, J. A. Cain and S. L. Newman (1982). "Generation of three different fragments of bound C3 with purified factor I or serum. I.

Requirements for factor H vs CR1 cofactor activity." J Immunol **129**(5): 2051-60.

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

Salafsky, J. S. (2006). "Detection of protein conformational change by optical second-harmonic generation." J Chem Phys **125**(7): 074701.

Sali, A., R. Glaeser, T. Earnest and W. Baumeister (2003). "From words to literature in structural proteomics." Nature **422**(6928): 216-25.

Schilling, B., R. H. Row, B. W. Gibson, X. Guo and M. M. Young (2003). "MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides." J Am Soc Mass Spectrom **14**(8): 834-50.

Schmidt, A., S. Kalkhof, C. Ihling, D. M. Cooper and A. Sinz (2005). "Mapping protein interfaces by chemical cross-linking and Fourier transform ion cyclotron resonance mass spectrometry: application to a calmodulin / adenylyl cyclase 8 peptide complex." Eur J Mass Spectrom (Chichester, Eng) **11**(5): 525-34.

Schulz, D. M., S. Kalkhof, A. Schmidt, C. Ihling, C. Stingl, K. Mechtler, O. Zschornig and A. Sinz (2007). "Annexin A2/P11 interaction: new insights into annexin A2 tetramer structure by chemical crosslinking, high-resolution mass spectrometry, and computational modeling." Proteins **69**(2): 254-69.

Seebacher, J., P. Mallick, N. Zhang, J. S. Eddes, R. Aebersold and M. H. Gelb (2006). "Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing." J Proteome Res **5**(9): 2270-82.

Shah, J. V. and D. W. Cleveland (2000). "Waiting for anaphase: Mad2 and the spindle assembly checkpoint." Cell **103**(7): 997-1000.

Silva, R. A., L. A. Schneeweis, S. C. Krishnan, X. Zhang, P. H. Axelsen and W. S. Davidson (2007). "The structure of apolipoprotein A-II in discoidal high density lipoproteins." J Biol Chem **282**(13): 9713-21.

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

Singh, P., S. A. Shaffer, A. Scherl, C. Holman, R. A. Pfuetzner, T. J. Larson Freeman, S. I. Miller, P. Hernandez, R. D. Appel and D. R. Goodlett (2008). "Characterization of protein cross-links via mass spectrometry and an open-modification search strategy." Anal Chem **80**(22): 8799-806.

Sinz, A. (2006). "Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions." Mass Spectrom Rev **25**(4): 663-82.

Sinz, A. (2010). "Investigation of protein-protein interactions in living cells by chemical crosslinking and mass spectrometry." Anal Bioanal Chem **397**(8): 3433-40.

Sironi, L., M. Mapelli, S. Knapp, A. De Antoni, K. T. Jeang and A. Musacchio (2002). "Crystal structure of the tetrameric Mad1-Mad2 core complex: implications of a 'safety belt' binding mechanism for the spindle checkpoint." EMBO J **21**(10): 2496-506.

Sopta, M., Z. F. Burton and J. Greenblatt (1989). "Structure and associated DNA-helicase activity of a general transcription initiation factor that binds to RNA polymerase II." Nature **341**(6241): 410-4.

Spahr, H., G. Calero, D. A. Bushnell and R. D. Kornberg (2009). "Schizosacharomyces pombe RNA polymerase II at 3.6-A resolution." Proc Natl Acad Sci U S A **106**(23): 9185-90.

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

Stryer, L. (1981). "Rapid motions in protein molecules." Biochem Soc Symp(46): 39-55.

Suchanek, M., A. Radzikowska and C. Thiele (2005). "Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells." Nat Methods **2**(4): 261-7.

Sun, T., A. Bollen, L. Kahan and R. Traut (1974). "Topography of ribosomal proteins of the Escherichia coli 30S subunit as studied with reversible cross-linking reagent methyl 4-mercaptobutyrimidate." Biochemistry **13**(11): 2334-2340.

Sun, Z. W. and M. Hampsey (1995). "Identification of the gene (SSU71/TFG1) encoding the largest subunit of transcription factor TFIIF as a suppressor of a TFIIB mutation in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **92**(8): 3127-31.

Sutherland, B. W., J. Toews and J. Kast (2008). "Utility of formaldehyde cross-linking and mass spectrometry in the study of protein-protein interactions." J Mass Spectrom **43**(6): 699-715.

Swann, M. J., L. L. Peel, S. Carrington and N. J. Freeman (2004). "Dual-polarization interferometry: an analytical technique to measure changes in protein structure in real time, to determine the stoichiometry of binding events, and to differentiate between specific and nonspecific interactions." Anal Biochem **329**(2): 190-8.

Sydow, J. F., F. Brueckner, A. C. Cheung, G. E. Damsma, S. Dengl, E. Lehmann, D. Vassylyev and P. Cramer (2009). "Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA." Mol Cell **34**(6): 710-21.

Tan, S., T. Aso, R. C. Conaway and J. W. Conaway (1994). "Roles for both the RAP30 and RAP74 subunits of transcription factor IIF in transcription

initiation and elongation by RNA polymerase II." J Biol Chem **269**(41): 25684-91.

Tan, S., R. C. Conaway and J. W. Conaway (1995). "Dissection of transcription factor TFIIF functional domains required for initiation and elongation." Proc Natl Acad Sci U S A **92**(13): 6042-6.

Tan, S., K. P. Garrett, R. C. Conaway and J. W. Conaway (1994). "Cryptic DNA-binding domain in the C terminus of RNA polymerase II general transcription factor RAP30." Proc Natl Acad Sci U S A **91**(21): 9808-12.

Taverner, T., N. E. Hall, R. A. O'Hair and R. J. Simpson (2002). "Characterization of an antagonist interleukin-6 dimer by stable isotope labeling, cross-linking, and mass spectrometry." J Biol Chem **277**(48): 46487-92.

Terpe, K. (2003). "Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems." Appl Microbiol Biotechnol **60**(5): 523-33.

Trester-Zedlitz, M., K. Kamada, S. K. Burley, D. Fenyo, B. T. Chait and T. W. Muir (2003). "A modular cross-linking approach for exploring protein interactions." J Am Chem Soc **125**(9): 2416-25.

Trnka, M. J. and A. L. Burlingame (2010). "Topographic studies of the GroEL-GroES chaperonin complex by chemical cross-linking using diformyl ethynylbenzene: the power of high resolution electron transfer dissociation for determination of both peptide sequences and their attachment sites." Mol Cell Proteomics **9**(10): 2306-17.

Tyree, C. M., C. P. George, L. M. Lira-DeVito, S. L. Wampler, M. E. Dahmus, L. Zawel and J. T. Kadonaga (1993). "Identification of a minimal set of proteins that is sufficient for accurate initiation of transcription by RNA polymerase II." Genes Dev **7**(7A): 1254-65.

Vasilescu, J., X. Guo and J. Kast (2004). "Identification of protein-protein interactions using in vivo cross-linking and mass spectrometry." Proteomics **4**(12): 3845-54.

Vila-Perello, M., M. R. Pratt, F. Tulin and T. W. Muir (2007). "Covalent capture of phospho-dependent protein oligomerization by site-specific incorporation of a diazirine photo-cross-linker." J Am Chem Soc **129**(26): 8068-9.

Vogt, W., G. Schmidt, B. Von Buttlar and L. Dieminger (1978). "A new function of the activated third component of complement: binding to C5, an essential step for C5 activation." Immunology **34**(1): 29-40.

Waanders, L. F., R. Almeida, S. Prosser, J. Cox, D. Eikel, M. H. Allen, G. A. Schultz and M. Mann (2008). "A novel chromatographic method allows on-line reanalysis of the proteome." Mol Cell Proteomics **7**(8): 1452-9.

Walport, M. J. (2001). "Complement. First of two parts." N Engl J Med **344**(14): 1058-66.

Walport, M. J. (2001). "Complement. Second of two parts." N Engl J Med **344**(15): 1140-4.

Wang, B. Q. and Z. F. Burton (1995). "Functional domains of human RAP74 including a masked polymerase binding domain." J Biol Chem **270**(45): 27035-44.

Wei, R. R., J. Al-Bassam and S. C. Harrison (2007). "The Ndc80/HEC1 complex is a contact point for kinetochore-microtubule attachment." Nat Struct Mol Biol **14**(1): 54-9.

Wei, R. R., J. R. Schnell, N. A. Larsen, P. K. Sorger, J. J. Chou and S. C. Harrison (2006). "Structure of a central component of the yeast kinetochore: the Spc24p/Spc25p globular domain." Structure **14**(6): 1003-9.

Wei, R. R., P. K. Sorger and S. C. Harrison (2005). "Molecular organization of the Ndc80 complex, an essential kinetochore component." Proc Natl Acad Sci U S A **102**(15): 5363-7.

Wiesmann, C., K. J. Katschke, J. Yin, K. Y. Helmy, M. Steffek, W. J. Fairbrother, S. A. McCallum, L. Embuscado, L. DeForge, P. E. Hass and M. van Lookeren Campagne (2006). "Structure of C3b in complex with CRIg gives insights into regulation of complement activation." Nature **444**(7116): 217-20.

Winters, M. S., D. S. Spellman and J. D. Lambris (2005). "Solvent accessibility of native and hydrolyzed human complement protein 3 analyzed by hydrogen/deuterium exchange and mass spectrometry." J Immunol **174**(6): 3469-74.

Wittelsberger, A., B. E. Thomas, D. F. Mierke and M. Rosenblatt (2006). "Methionine acts as a "magnet" in photoaffinity crosslinking experiments." FEBS Lett **580**(7): 1872-6.

Wong, S. (1991). Chemistry of protein conjugation and cross-linking. Boca Raton, CRC Press Inc.

Woychik, N. A. and M. Hampsey (2002). "The RNA polymerase II machinery: structure illuminates function." Cell **108**(4): 453-63.

Yan, F., F. Y. Che, D. Rykunov, E. Nieves, A. Fiser, L. M. Weiss and R. Hogue Angeletti (2009). "Nonprotein based enrichment method to analyze peptide cross-linking in protein complexes." Anal Chem **81**(17): 7149-59.

Yan, Q., R. J. Moreland, J. W. Conaway and R. C. Conaway (1999). "Dual roles for transcription factor IIF in promoter escape by RNA polymerase II." J Biol Chem **274**(50): 35668-75.

Yang, H., G. Luo, P. Karnchanaphanurach, T. M. Louie, I. Rech, S. Cova, L. Xun and X. S. Xie (2003). "Protein conformational dynamics probed by single-molecule electron transfer." Science **302**(5643): 262-6.

Ye, X., P. K. O'Neil, A. N. Foster, M. J. Gajda, J. Kosinski, M. A. Kurowski, J. M. Bujnicki, A. M. Friedman and C. Bailey-Kellogg (2004). "Probabilistic cross-

link analysis and experiment planning for high-throughput elucidation of protein structure." Protein Sci **13**(12): 3298-313.

Yonaha, M., T. Aso, Y. Kobayashi, H. Vasavada, Y. Yasukochi, S. M. Weissman and S. Kitajima (1993). "Domain structure of a human general transcription initiation factor, TFIIF." Nucleic Acids Res **21**(2): 273-9.

Young, M. M., N. Tang, J. C. Hempel, C. M. Oshiro, E. W. Taylor, I. D. Kuntz, B. W. Gibson and G. Dollinger (2000). "High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry." Proc Natl Acad Sci U S A **97**(11): 5802-6.

Zhang, C. and Z. F. Burton (2004). "Transcription factors IIF and IIS and nucleoside triphosphate substrates as dynamic probes of the human RNA polymerase II mechanism." J Mol Biol **342**(4): 1085-99.

Zhang, H., X. Tang, G. R. Munske, N. Tolic, G. A. Anderson and J. E. Bruce (2009). "Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry." Mol Cell Proteomics **8**(3): 409-20.

Zheng, L., Y. Chen and W. H. Lee (1999). "Hec1p, an evolutionarily conserved coiled-coil protein, modulates chromosome segregation through interaction with SMC proteins." Mol Cell Biol **19**(8): 5417-28.

Zhou, M. and C. V. Robinson (2010). "When proteomics meets structural biology." Trends Biochem Sci **35**(9): 522-9.

Ziegler, L. M., D. A. Khaperskyy, M. L. Ammerman and A. S. Ponticelli (2003). "Yeast RNA polymerase II lacking the Rpb9 subunit is impaired for interaction with transcription factor IIF." J Biol Chem **278**(49): 48950-6.