

NMR spectroscopy to study dynamic protein structures

Dissertation

Zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich Biochemie, Chemie und Pharmazie

der Goethe-Universität

in Frankfurt am Main

von

Nina Kubatova

aus Odessa (Ukraine)

Frankfurt am Main

2019

(D30)

Vom Fachbereich Biochemie, Chemie und Pharmazie der Goethe-Universität als Dissertation
angenommen.

Dekan: Prof. Dr. Clemens Glaubitz

1. Gutachter: Prof. Dr. Harald Schwalbe
2. Gutachter: Prof. Dr. Clemens Glaubitz

Datum der Disputation: 11.12.2019

Meiner Familie.

Table of Contents

Summary	vi
Zusammenfassung.....	xii
List of Abbreviations.....	xix
Chapter I	
Methodological aspects of protein NMR spectroscopy.....	1
1. A brief history of NMR.....	2
2. Speed up the measurement time	3
2.1. Fast-pulsing techniques.....	4
2.2. Non-uniform sampling	6
2.2.1. NUS: general idea.....	6
2.2.2. Reconstitution approaches	7
2.3. Targeted Acquisition approach	9
Chapter II	
Rapid NMR and biophysical characterization of small proteins	15
1. General introduction.....	16
1.1. Nomenclature diversity.....	16
1.2. Historical neglect of small proteins.....	17
1.3. Identification of sORFs coding small proteins.....	18
1.3.1. Computational approaches.....	18
1.3.2. Experimental biochemical approaches	18
1.3.2.1. Ribosome profiling	18
1.3.2.2. Mass spectrometry.....	20
1.4. Functional studies	21
1.4.1. Structure-function paradigm.....	21
1.4.2. Structural studies of small proteins	23
1.5. Rapid secondary structure determination with NMR.....	26
1.6. SPP 2002 “Small Proteins in Prokaryotes, an Unexplored World”	26
1.7. Motivation and aim.....	28

2.	Material and Methods	29
2.1.	Biochemical Methods	29
2.1.1.	Peptide preparation	29
2.1.1.1.	Solid phase synthesis	29
2.1.1.2.	Reversed-phase HPLC	29
2.1.2.	Small protein preparation	29
2.1.2.1.	Design of the fusion protein	29
2.1.2.2.	Plasmids	30
2.1.2.3.	Protein expression and purification	32
2.1.2.3.1.	Test expression	32
2.1.2.3.2.	SDS-PAGE	32
2.1.2.3.3.	Protein expression	32
2.1.2.3.4.	Protein purification	33
2.2.	Biophysical Methods	34
2.2.1.	UV/Vis spectrometry	34
2.2.2.	Circular dichroism spectroscopy (CD)	34
2.3.	Nuclear magnetic resonance spectroscopy	35
2.3.1.	Fast secondary structure screening of small proteins	35
2.3.2.	Structural analysis of folded small proteins	36
2.3.2.1.	Assignment strategy	36
2.3.2.2.	Non-uniform sampling (NUS) and Targeted Acquisition (TA)	36
2.3.2.3.	FLYA automated assignment	36
2.3.2.4.	Temperature series	37
2.3.2.5.	Dynamical studies of P17 small proteins	37
2.3.2.5.1.	Hydrogen-deuterium exchange NMR experiments	37
2.3.2.5.2.	Heteronuclear relaxation experiments	37
2.3.2.6.	Chemical shift perturbations of I58P P17 mutant	38
2.3.2.7.	Structure calculation	38
3.	Results and discussions	39
3.1.	Screening overview	39
3.2.	Sequence to structure analysis	40
3.2.1.	Fast online structure prediction tools	40
3.2.2.	Prediction of dynamic transition	44
3.3.	Cooperation with Elena Evguenieva-Hackenberg group (University of Giessen)	45

3.3.1.	Structural characterization of P1, P3 and P5 peptides.....	45
3.3.2.	Structural characterization of P20 peptide	46
3.3.2.1.	Investigation of peptide-ligand interaction.....	46
3.2.2.2.	Structure calculation of peptide P20.....	49
3.4.	Cooperation with Ruth Schmitz-Streit group (University of Kiel)	50
3.5.	Cooperation with Wolfgang Streit group (University of Hamburg)	54
3.6.	Cooperation with Jörg Soppa (University of Frankfurt)	55
3.6.1.	Structural characterization of zinc finger proteins.....	55
3.6.2.	Structural characterization of P19	57
3.6.2.1.	NMR backbone assignment and TALOS prediction.....	58
3.6.2.2.	Influence of the zinc ions on the folding pathway	59
3.6.2.3.	Discussion.....	61
3.7.	Cooperation with Anita Marchfelder (University of Ulm).....	61
3.7.1.	Structural characterization of P18 and P17 small proteins.....	61
3.7.2.	Screening of folding conditions of P18.....	63
3.7.3.	Structural characterization of P17 small protein	65
3.7.3.1.	NMR spectroscopy of P17	65
3.7.3.1.1.	NMR backbone assignment and TALOS prediction.....	65
3.7.3.1.2.	Side chain assignment of P17.....	66
3.7.3.1.3.	$^3J_{\text{HNHA}}$ coupling constants.....	66
3.7.3.1.4.	Intramolecular hydrogen bond formation	66
3.7.3.1.5.	Hydrogen-deuterium exchange NMR experiments	68
3.7.3.1.6.	Heteronuclear relaxation experiments	69
3.7.3.1.7.	Structure calculation of P17	71
3.7.3.1.8.	Structure comparison with homologous proteins	74
3.7.3.2.	Dimer formation.....	78
3.7.3.2.1.	NMR spectroscopy of mutants.....	80
3.7.3.2.2.	Concentration dependent dimer formation	83
3.7.3.3.	Discussion.....	84
3.8.	Timesaving methods	84
3.8.1.	Automated FLYA assignment: proof of principle	84
3.8.2.	Targeted acquisition technique.....	85
4.	Conclusions and perspectives	87

Chapter III

Light dynamics of retinal disease relevant G90D bovine rhodopsin mutant	89
1. General introduction	90
1.1. Sense of sight and structure of the eye.....	90
1.2. Phototransduction.....	91
1.3. Retinal regeneration.....	94
1.4. Rhodopsin.....	96
1.4.1. Rhodopsin structure in the dark state.....	96
1.4.2. Rhodopsin structure in the light state.....	97
1.4.3. Photocycle	99
1.5. Retinal dystrophy diseases	101
1.5.1. Retinitis pigmentosa (RP)	101
1.5.2. Treatment options for RP	102
1.5.3. CSNB disease classification	103
1.5.4. Mode of inheritance of CSNB associated mutants.....	105
1.5.5. Treatment options for CSNB.....	105
1.5.6. Rhodopsin mutations associated with CSNB.....	106
1.6. State of the art.....	107
1.6.1. Development of rhodopsin structures	107
1.6.2. Crystal structure of constitutively active rhodopsin mutants	109
1.6.3. G90D Mutation	111
1.7. Motivation and aims.....	113
2. Material and Methods.....	114
2.1. Constructs.....	114
2.2. Cell culture protocol	114
2.3. Rhodopsin purification	115
2.3.1. Retinal purification	115
2.3.2. Antibody preparation	116
2.3.3. Rhodopsin reconstitution of wild type rhodopsin ^[179]	117
2.3.4. Rhodopsin reconstitution of stabilized mutants ^[218]	118
2.4. Sample preparation	118
2.5. NMR measurements.....	119
3. Results and discussions	120

3.1.	Retinal purification.....	120
3.2.	Rhodopsin purification.....	121
3.3.	Liquid state NMR experiments.....	125
3.4.	Photo flash and ultrafast kinetic experiments.....	131
3.5.	Solid state NMR experiments.....	133
3.5.1.	Dark state experiments.....	134
3.5.2.	Light state experiments.....	136
4.	Discussions and perspectives.....	140
5.	Conclusion:.....	142
	References.....	145
	Appendix A1: Plasmids.....	152
	Appendix A2: Buffers and media.....	155
	Appendix A3: Purification strategy of peptides and small proteins.....	157
	Appendix A4: Overview of small proteins screened in this study.....	160
	Appendix A5: NMR spectrometer.....	164
	Appendix A6: NMR experiments.....	165
	Appendix A7: Prediction of dynamic transition.....	167
	Appendix A8: NMR secondary structure analysis of small proteins.....	168
	Appendix A9: Rhodopsin sequences.....	182
	Appendix A10: Cell culture media and buffers.....	183
	Appendix A11: Rhodopsin purification.....	185
	Appendix A12: Liquid state NMR experiments.....	189
	Acknowledgments.....	190
	Publications.....	192
	Conference contributions.....	193
	Curriculum vitae.....	194

Summary

This doctoral thesis deals with the structural and dynamical NMR characterization of biomolecules, covering a broad range of proteins, from small peptides to large GPCRs proteins. This work consists of two projects, which are presented in chapter II and III. Chapter II is focused on the structural screening of peptides and small proteins ranging from 14 to 71 amino acids, while chapter III describes the structure and light dynamics of the disease relevant rhodopsin G90D mutant. The main method used to investigate both types of proteins is NMR spectroscopy. Both chapters comprise individual general introduction, materials and methods, results and discussion sections, and a final conclusion paragraph.

'Chapter I: Methodological aspects of protein NMR spectroscopy' presents an overview of different NMR methods developed for the rapid characterization of protein structure and dynamics. Multidimensional NMR, which is routinely used in structural biology, is indispensable for protein structure determination in solution. However, detailed information with resolution at the atomic level is time consuming and requires weeks of expensive measurement time, followed by the manual data analysis. Therefore, the development of time-saving NMR techniques is highly required for screening studies of a large amount of proteins, and can be also helpful for studying unstable biomolecules, as their short lifetime often restricts the experimental procedure.

This chapter covers the two main approaches to accelerate a multidimensional NMR experiment: fast-pulsing techniques that aim to reduce the duration of an individual measurement, and non-uniform sampling technique (NUS), which was developed to reduce the overall number of increments in virtual time domains. A combination of both approaches, fast-pulsing and non-uniform sampling, allows speeding up the measurement time by 2-3 orders of magnitude. Furthermore, recently developed software called TA (targeted acquisition) combines various time-saving approaches, including fast-pulsing, non-uniform sampling and targeted acquisition. Targeted acquisition algorithm records a set of multidimensional NMR spectra in semi-interleaved incremental mode. This provides the ability to monitor the quality of the recorded spectra in real-time and therefore enables the completion of the experiments after the desired quality is achieved. Using this approach will greatly reduce the measurement time without losing important structural information. The implemented automated FLYA assignment further contributes to the rapid and simplified readout of the chemical shift assignment progress of the TA program. During this doctoral dissertation, the scientific collaboration with the TA software developer Prof. Vladislav Orekhov (Sweden) took place, and resulted in the successful establishing of this new NMR technology in the Schwalbe laboratory. TA is now routinely applied in Prof. Schwalbe group for the structure elucidation of small proteins.

‘Chapter II: Rapid NMR and biophysical characterization of small proteins’ describes the structural analysis of peptides and small proteins, which were recently identified within the framework of the Priority Program (SPP 2002). Due to technical limitations in detections of small systems and strict assumptions concerning the smallest size of the gene that can be translated, small open reading frames (sORFs) were excluded from the automated gene annotation for a very long time. Thanks to the newly developed computational and experimental approaches, the ability to identify and detect the small proteins consisting of less than approximately 70 amino acids sparked a growing scientific interest by microbiologist. In the past years, hundreds of new short protein sequences were discovered. Although some peptides were found to be involved in diverse essential biological processes, the functional elucidation of a large number of recently discovered peptides and small proteins remains a challenging task. It is well established that the structure of proteins is often linked to their function. However, the size of small constructs often restricts the possible diversity of secondary structure elements that might be adopted by a protein. Furthermore, as was shown for intrinsic disordered proteins (IDPs), the absence of a well-defined three-dimensional structure does not necessarily mean lack of function. Moreover, peptides, which are initially unstructured in the isolated form can fold in a stable structured conformation upon interaction with their biological partners. Solution state NMR spectroscopy is perfectly amenable for the structural characterization of systems of this size. It provides a rapid readout about the conformational state of small peptides unambiguously, distinguishing between folded, molten globule and unstructured conformations.

During this doctoral thesis the workflow protocol for fast screening of peptides and small proteins was established and applied to 20 candidates ranging from 14 to 71 amino acids, which were identified and selected by six microbiological groups, all members of the Priority Program on small proteins (SPP2002) funded by the German research foundation (DFG). The screening protocol includes sample preparation and biochemical characterization. Peptides containing less than 30 amino acids were synthesized by solid phase synthesis (SPPS), while small proteins containing more than 30 amino acids were heterologously expressed in *E. coli*.

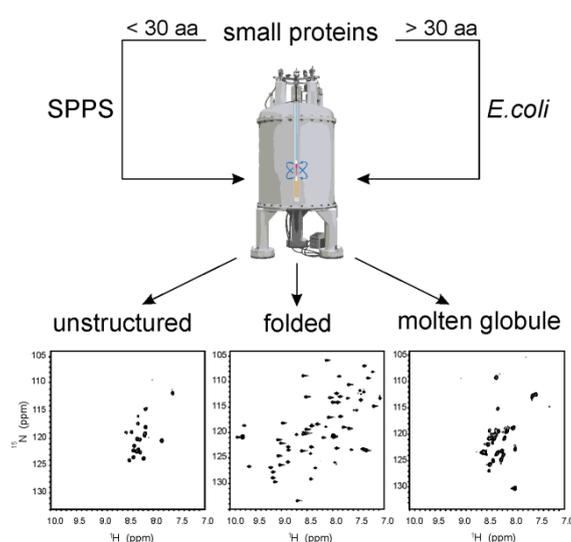


Figure I. Workflow for NMR structural investigations of peptides and small proteins.

Summary

After protein purification the constructs were screened for the secondary structure using CD and 1D ^1H NMR spectroscopy. The clear structural readout was conducted based on 2D $^1\text{H}^{15}\text{N}$ HSQC NMR spectra. The linewidth and dispersion of the backbone amide signals provide a structural fingerprint of the protein and allow unambiguous identification of their conformational state. For the most promising candidates TALOS chemical shift analysis was performed.

In total, 20 different constructs ranging from 14 to 71 amino acids were screened during this doctoral thesis. Due to issues with stability and high hydrophobicity, five small proteins were excluded from the fast secondary structure screening. For the remaining 15 candidates, fast secondary structure screening analysis was conducted. Eight peptides exist in an unstructured state, four small proteins adopt a molten globule conformation and three proteins were found to be structured. Two small proteins P17 and P19 were extensively studied for their structural and dynamic properties.

For the 60 residues long small protein P17, a high-resolution NMR solution structure was calculated. According to relaxation analysis and intermolecular NOE contacts, the small protein forms a symmetric dimer, the formation of which was found to be concentration dependent. Further mutagenesis studies performed on the dimer interface reported dimerization process to be essential for the protein stability and most likely for its functionality.

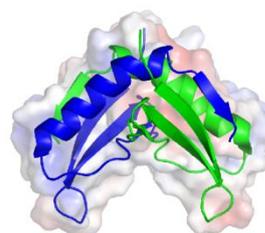


Figure II. NMR solution structure of P17 small protein (pdb code: 6QZZ).

The structural analysis of zinc-finger protein P19 shows a temperature dependent dynamic equilibrium of two long-lived states. Despite the presence of two putative zinc-binding sites in the sequence, only one metal ion is bound into the protein structure. Furthermore, it was shown that the incorporation of the zinc ion is essential for proper protein folding and its removal leads to the reversible unfolding of the protein.

All constructs, screened during this doctoral thesis, were investigated in their isolated form. A clear correlation between the size and the secondary structure elements adopted by the biomolecules was identified. All peptides (14-31 residues) exist in an unstructured state, while small proteins (38-71 residues) adopt folded or partially folded conformations. Bioinformatic sequence-to-structure prediction tools, performed within the scientific collaboration with Prof. Monika Fuxreiter (Hungary), show the potential of unstructured peptides to fold upon complex formation. Taken together, all these findings open up a new horizon for the small proteins research field, which should in future include interaction studies with potential binding partners and further elucidation of their functional role in the cellular processes.

The second part of the thesis described in '**Chapter III: Light dynamics of retinal disease relevant G90D bovine rhodopsin mutant**' is focused on investigation of the single point mutation G90D that leads to disruption of the retinal binding pocket of rhodopsin, resulting in an impaired visual cycle.

Rhodopsin belongs to the most studied GPCRs (G-protein coupled receptor) membrane protein family in the human genome. Despite that, the photocycle mechanism still remains not fully understood. Furthermore, very little is known about the influence of the particular disease relevant mutations on the photocycle. Therefore, any additional structural or dynamic information concerning the role of the mutation in the photoreceptor cells will fulfill the general description of the process and, possibly, will contribute to future development of the disease treatment.

Basically, the absorption of a photon initiates the signaling cascade in retinal rods that makes vision possible. Disruptions at any step of the photocycle, which represents a highly complex process, can impair the visual cycle and therefore lead to numerous visual disorders. A large number of diverse mutations are known to cause retinal dystrophy. Depending on the protein encoded by the mutated gene, a particular pathway in the visual cycle may be affected. In particular, mutations in the *RHO* gene, which encodes the rhodopsin protein, cause disruption of the phototransduction. The particular amino acid G90 represents a unique position in the protein sequence. Depending on the nature of the mutated amino acid it can either lead to the most common human inherited retinal dystrophy night blindness disease called *retinitis pigmentosa* (PR) or congenital stationary night blindness (CSNB). This study is particularly focused on the non-progressive inherited CSNB retinal disease, which is characterized by reduction of dim and night vision, problems with adaptation to darkness and in some cases loss of the general visual acuity. G90D is a constitutively active mutation, which affects the retinal binding position, resulting in the constant basal activity of the protein. Although the crystal structure of the light active state of the G90D mutant exists, the low resolution of the retinal binding pocket does not provide information about the exact retinal conformation at atomic level. Moreover, the lack of the crystals in the dark (ground) state reflects the crucial impact of the mutation on the binding pocket environment, leaving the molecular mechanism unclear. Different hypotheses have been proposed to explain the high level of the basal activity of the G90D mutant. One of them is based on the spontaneous isomerization of the chromophore; another suggests the pre-active dark state of the protein to be the reason for night blindness. However, there is no unambiguous interpretation of the molecular interaction mechanism driving this night blindness phenotype. Such investigation is needed in order to shed some light on the pathogenic mechanism driven by the G90D mutant.

Summary

To fulfil the missing structural information regarding the exact retinal conformation and the mutation causing structural rearrangements of the binding pocket of the G90D mutant, a combination of liquid and solid state NMR experiments was conducted. Furthermore, kinetic studies performed by flash photolysis and ultrafast absorption experiments gave a unique information about the light activation process of the G90D mutant.

The investigations performed during this doctoral thesis were based on the analysis of three different rhodopsin constructs: wild type (wt), stabilized wild type (N2C/D282C) and the disease relevant stabilized (N2C/D282C) G90D mutant. The double mutation forms an additional disulfide bond in the extracellular side, which was designed to increase thermal stability of the protein. So far, recombinant rhodopsin could be successfully crystallized only in presence of this disulfide bridge. Liquid state NMR experiments, which use chemical shift resonances of the five tryptophan residues as reporter signals, show no influence of the stabilized double mutation on the protein conformation in the inactive dark state. However, the light active conformation is indeed stabilized and remains so for several days.

The stabilizing mutations affect the whole kinetic process leading to an overall prolonged lifetime of the intermediate states. Thus, the formation of the early photoproduct bathorhodopsin takes four times longer in the stabilized construct compared to the unmodified wild type. The stabilizing effect of the N2C/D282C mutation is further reflected on the decay of the intermediate states. The Meta II decay and Meta III formation is increased by 50% and Meta III decay occurs even four times faster, the covalently bound retinal is stabilized and its hydrolysis is further delayed.

Liquid state NMR experiments report structural similarity between the dark and the light state of the G90D mutant, indicating the pre-active dark state formation. In contrast to wild type, the G90D mutant shows structural heterogeneity. This was observed by an additional tryptophan signal detected by liquid state NMR in the dark state. Remarkably, this signal still remains after illumination in Meta II state. This structural heterogeneity was further confirmed by DNP-enhanced solid state NMR experiments. The protein Schiff base and retinal C14 carbon represent the mixture of minor populations. Furthermore, it was shown that all protein populations include 11-cis attached retinal only, the minor conformation of which is attributed to the slightly different steric position of the 11-cis retinal in the binding pocket.

Illumination of the protein induces 11-cis to all-trans retinal isomerization, resulting in a series of intermediate photoproducts. The active Meta II state of rhodopsin wild type is characterized by deprotonated Schiff base and all-trans covalently bound retinal. In this thesis it was reported that the Meta II conformation of the G90D mutant exists in the protonated form. That was shown by solid state NMR spectroscopy and confirmed by flash photolysis. With solid state NMR equipment it was possible to cryo-trap the early photostate bathorhodopsin of the G90D mutant. The minor traces of the light

active Meta II state were identified in the batho conformation, pointing towards the previously observed pre-active dark state formation. Furthermore, for both, batho and Meta II light intermediate states, a protonated Schiff base signal was detected.

Eventually, unique information about the mutation effect on the rhodopsin photocycle was obtained by the kinetic experiments. The early processes of the photocycle were observed with ultrafast absorption measurements, which describe kinetics during the first minute. In contrast to the wild type, the photorhodopsin conformation was detectable for the G90D mutant, indicating the delayed retinal isomerization. Further performed flash photolysis experiments comprise kinetics up to two hours and describe the evolution of the light active Meta II and Meta III intermediate states. The formation of the Meta III state and decay of the protonated Meta II state take place with a similar rate as that of the stabilized wild type, while Meta III decay of G90D mutant is 25% faster.

Taken together, the results produced during this doctoral thesis indicate the heterogeneous pre-active conformation of the G90D ground state. Furthermore, only one single 11-cis bound retinal isomer could be observed, indicating the absence of spontaneous retinal isomerization. In contrast to the wild type, the light active Meta II intermediate of the G90D mutant was found in the protonated state. This mutation has a significant effect on the photocycle, leading to the absolutely unique kinetic profile. Thus, the light-induced retinal isomerization is delayed, while the retinal release is faster and, likely, partially occurs even before the intermediate states Meta II and Meta III are formed. All of these results support the pre-active ground state theory and provide new data points for a better understanding of the mechanism, which leads to the disruption of the visual cycle.

Zusammenfassung

Im Rahmen dieser Doktorarbeit wurde die Struktur und Dynamik von Biomolekülen mittels NMR-Spektroskopie untersucht. Dabei wurde ein breites Spektrum von Systemen, von kleinen Peptiden bis hin zu großen GPCR Proteinen (G-Protein-gekoppelte Rezeptoren), abgedeckt. Diese Arbeit besteht aus zwei Projekten, die in Kapitel II und III vorgestellt sind. Kapitel II befasst sich mit dem strukturellen Screening von Peptiden und kleinen Proteinen, während Kapitel III die Struktur und die Lichtdynamik der Rhodopsin G90D Mutante, die maßgeblich für Nachtblindheit verantwortlich ist, beschreibt. Für die beiden Themengebiete wurde NMR-Spektroskopie als Hauptmethode eingesetzt. Beide Kapitel enthalten jeweils eine allgemeine Einführung, ein Material und Methoden Kapitel, Ergebnisse und Diskussionsabschnitte sowie eine abschließend Zusammenfassung.

„Kapitel I: Methodische Aspekte der Protein-NMR-Spektroskopie“ stellt einen Überblick über verschiedene NMR Methoden dar, die für die schnelle Charakterisierung der Struktur und Dynamik von Proteinen entwickelt wurden. Die multidimensionale NMR-Spektroskopie, die routinemäßig in der Strukturbiologie verwendet wird, ist für die Bestimmung der Proteinstruktur in Lösung unersetzlich. Detaillierte strukturelle Informationen mit atomarer Auflösung sind jedoch zeitaufwändig und erfordern Wochen teurer Experimentzeit, gefolgt von der nicht weniger aufwändigen manuellen Datenanalyse. Daher ist die Entwicklung zeitsparender NMR Techniken für das Screening einer großen Menge von Proteinen sowie für die Untersuchung von instabilen, kurzlebigen Biomolekülen sehr wichtig.

In diesem Kapitel sind die beiden Hauptansätze zur Beschleunigung eines mehrdimensionalen NMR Experiments beschrieben. Diese umfassen die *fast-pulsing* Techniken, mit denen die Dauer einer einzelnen Messung durch die optimierte Pulssequenz verringert werden kann. Des Weiteren wird auf die *non-uniform sampling* Technik (NUS) eingegangen, bei der die Gesamtzahl der Messungen reduziert wird. Eine Kombination dieser beiden Ansätze, kann die experimentelle Zeit um 2-3 Größenordnungen verkürzen. Eine moderne Software namens TA (*targeted acquisition*) kombiniert verschiedene bereits existierende, zeitsparende Methoden, wodurch die Experimentaufnahme- und Analysezeit deutlich reduziert werden kann. Die Software vereint *fast-pulsing* und *non-uniform sampling* Techniken in ein gemeinsames Programm. Darüber hinaus ermöglicht sie eine zielgerichtete Aufnahme (*targeted acquisition*) von NMR Experimenten zur Resonanzzuordnung. Der TA Algorithmus erlaubt die Aufnahme einer Reihe von mehrdimensionalen NMR Spektren im semi-interleaved Modus. Dies ermöglicht eine Qualitätskontrolle der aufgenommenen Spektren in Echtzeit und führt dazu, dass die Experimente zu einem Zeitpunkt beendet werden können, an dem die gewünschte Qualität erreicht ist. Durch diesen Ansatz wird die Messzeit ohne Qualitätsverlust der strukturellen Daten erheblich verkürzt. Das in die

Software integrierte, automatisierte Programm zur Zuordnung chemischer Verschiebungen (FLYA) trägt weiterhin zur schnellen und vereinfachten Analyse der Daten bei.

Im Rahmen dieser Dissertation fand die wissenschaftliche Zusammenarbeit mit dem TA-Softwareentwickler Prof. Vladislav Orekhov (Schweden) statt, die zur erfolgreichen Etablierung der neuen NMR-Technologie im Schwalbe-Labor geführt hat. Die TA Technik wird nun routinemäßig zur Strukturcharakterisierung kleiner Proteine eingesetzt.

„Kapitel II: Zeiteffiziente NMR-spektroskopische und biophysikalische Charakterisierung kleiner Proteine“ beschreibt die Strukturanalyse von Peptiden und kleinen Proteinen, die kürzlich im Rahmen des Schwerpunktprogramms (SPP 2002) identifiziert wurden. Aufgrund technischer Einschränkungen und falscher Annahmen bei kleinen Proteine und Peptiden in der Vergangenheit, wurden kleine offene Leserahmen (sORFs) für eine sehr lange Zeit von der automatisierten Gen-Annotation übersehen. Neu entwickelte bioinformatische und experimentelle Methoden ermöglichen heute die Identifizierung und den Nachweis von kleinen Proteinen mit weniger als 70 Aminosäuren. So wurden in den letzten Jahren Hunderte von neuen kurzen Proteinsequenzen entdeckt. Obwohl nachgewiesen wurde, dass einige Peptide an mehreren essentiellen biologischen Prozessen beteiligt sind, bleibt die funktionelle Aufklärung einer großen Anzahl kürzlich identifizierter Peptide und kleiner Proteine eine herausfordernde Aufgabe. Grundsätzlich geht man davon aus, dass die Struktur der Proteine oft mit deren Funktion verbunden ist. Bei kleinen Proteinen jedoch wird die Vielfalt von Sekundärstrukturelementen häufig durch die Größe der Proteine begrenzt. Das Paradebeispiel sind die intrinsisch ungeordneten Proteine (IDPs), die trotz des Fehlens einer genau definierten dreidimensionalen Struktur an einem breiten Spektrum von Funktionen beteiligt sind. Darüber hinaus können Peptide, die in isolierter Form unstrukturiert sind, eine stabile strukturierte Konformation durch Wechselwirkung mit ihren biologischen Partnern annehmen. Die *NMR-Spektroskopie in Flüssigkeiten* eignet sich hervorragend für die strukturelle Charakterisierung von Systemen dieser Größe. Dabei kann man eindeutig und schnell zwischen gefalteten, dynamisch wechselnden (molten globule) und unstrukturierten Konformationen unterscheiden.

Während dieser Doktorarbeit wurde ein Protokoll für das schnelle Screening von Peptiden und kleinen Proteinen erstellt und auf 20 Modellsysteme angewendet. Die Systeme wurden von sechs kollaborierenden Gruppen aus der Mikrobiologie identifiziert und ausgewählt. Alle diese Forschungsgruppen sind Mitglieder des Schwerpunktprogramms (SPP2002), das von der Deutschen Forschungsgemeinschaft (DFG) gefördert wird. Das Screening-Protokoll umfasst die Probenvorbereitung und die biochemische Charakterisierung der Systeme. Peptide mit weniger als 30 Aminosäuren (AS)

Zusammenfassung

wurden durch Festphasensynthese (SPPS) synthetisiert, während kleine Proteine, die aus mehr als 30 Aminosäuren bestehen, in *E. coli* heterolog exprimiert wurden.

Nach der Protein Aufreinigung wurden die Konstrukte unter Verwendung von CD- und $1D\ ^1H$ NMR Spektroskopie auf die Sekundärstruktur untersucht. Die weitere Strukturanalyse erfolgte anhand von $2D\ ^1H,^{15}N$ HSQC NMR-Spektren. Die Linienbreite und Dispersion der Rückgratamid-Signale liefern dabei einen strukturellen Fingerabdruck des Proteins und ermöglichen eine eindeutige Identifizierung ihres Konformationszustands. Für die vielversprechenden Kandidaten wurde die TALOS-Analyse der chemischen Verschiebung durchgeführt.

Im Rahmen dieser Doktorarbeit wurden insgesamt 20 verschiedene Konstrukte im Bereich von 14 bis 71 Aminosäuren untersucht. Aufgrund ihrer Stabilität und Hydrophobizität wurden fünf kleine Proteine von dem Sekundärstruktur-Screening ausgeschlossen. Für die verbleibenden 15 Kandidaten wurde eine zeiteffiziente Sekundäranalyse durchgeführt. Acht Peptide existieren in einem unstrukturierten Zustand, vier kleine Proteine nehmen eine *molten globule* Konformation an und drei Proteine sind strukturiert. Für die zwei kleinen strukturierenden Proteine P17 und P19 wurde eine detaillierte Analyse bezüglich Struktur und Dynamik durchgeführt.

Für das kleine Protein P17 wurde eine hochaufgelöste NMR-Struktur berechnet. Gemäß der Relaxationsanalyse und den intermolekularen NOEs bildet das kleine Protein ein symmetrisches Dimer, dessen Entstehung konzentrationsabhängig ist. Mutagenese Studien, die an der Dimer-Grenzfläche durchgeführt wurden, wiesen darauf hin, dass der Dimerisierungsprozess entscheidend für die Proteinstabilität und höchstwahrscheinlich für die Funktionalität ist.

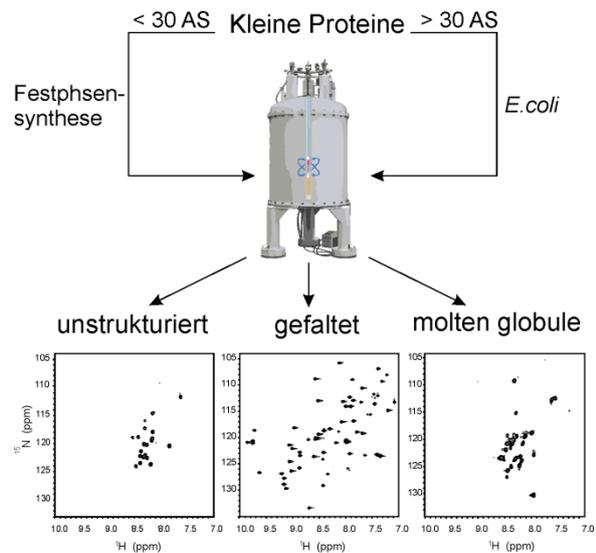


Abbildung I. Protokoll für NMR-Struktur-Untersuchungen von Peptiden und kleinen Proteinen.

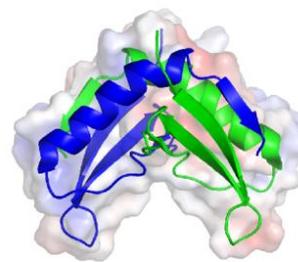


Abbildung II. NMR Struktur des kleinen Proteins P17 (pdb code: 6Q2Z).

Die Strukturanalyse des Zinkfinger-Proteins P19 zeigte ein temperaturabhängiges dynamisches Gleichgewicht zweier langlebiger Zustände. Trotz zweier Zinkbindungsstellen in der Sequenz, ist nur ein Metallion in die Proteinstruktur eingebunden. Weiterhin wurde gezeigt, dass der Einbau von Zink-Ionen für die korrekte Proteinfaltung notwendig ist und dessen Entfernung zur reversiblen Entfaltung des Proteins führt.

Alle im Rahmen dieser Dissertation analysierten Konstrukte wurden isoliert untersucht. Es wurde eine eindeutige Korrelation zwischen der Größe und den Sekundärstrukturelementen der Biomoleküle festgestellt. Somit existieren alle Peptide (14-31 Aminosäuren) in einem unstrukturierten Zustand, während kleine Proteine (38-71 Aminosäuren) eine gefaltete oder teilweise gefaltete Konformation einnehmen. Bioinformatische Sekundärstrukturvorhersagen wurden in wissenschaftlicher Zusammenarbeit mit Prof. Monika Fuxreiter (Ungarn) durchgeführt und zeigten ein gesteigertes Potenzial der unstrukturierten Peptide sich bei einer Komplexbildung zu falten und eine definierte Konformation anzunehmen.

Zusammengenommen eröffnen all diese Erkenntnisse einen neuen Horizont für das Forschungsfeld der kleinen Proteine, das zukünftig Interaktionsstudien mit potenziellen Bindungspartnern sowie eine weitere Aufklärung ihrer funktionellen Rolle in den zellulären Prozessen umfassen sollte.

Der zweite Teil dieser Doktorarbeit ist im "**Kapitel III: Lichtdynamik der krankheitsbedingten G90D Mutation des Rhodopsins**" beschrieben und befasst sich mit der biophysikalischen Untersuchung dieser Mutante mit Hilfe der NMR Spektroskopie.

Rhodopsin gehört zu den am besten untersuchten GPCRs Membranproteinen. Dennoch ist der Mechanismus des Photozyklus immer noch nicht vollständig aufgeklärt und die Auswirkungen von krankheitsbedingten Mutationen des Rhodopsins sind noch wenig untersucht. Daher werden dringend strukturelle bzw. dynamische Informationen dieses Modell-GPCRs bezüglich der Rolle von Mutationen in Photorezeptorzellen benötigt, um den gestörten Prozess der Lichtsignaltransduktion bei Krankheiten zu verstehen und möglicherweise neue Ansätze zur zukünftigen Krankheitsbehandlung zu erhalten.

Grundsätzlich löst die Absorption von Photonen eine Signalkaskade in der Netzhaut aus, die das Sehen ermöglicht. Der Photozyklus ist ein sehr komplexer, mehrstufiger Prozess und kleinste Beeinträchtigungen dieses Sehzyklus können zu zahlreichen Sehbehinderungen führen.

Eine Vielzahl unterschiedlicher Mutationen, die eine Netzhautdystrophie verursachen, ist bekannt. Abhängig vom Protein, für das das mutierte Gen kodiert, kann ein bestimmter Pfad im Sehzyklus beeinflusst werden. Insbesondere führen Mutationen im *RHO* Gen, das Rhodopsin kodiert, zu einer Störung der Phototransduktion. Die Aminosäure G90 stellt hierbei eine einzigartige Position in der

Rhodopsin-Proteinsequenz dar. Abhängig von der Art der an dieser Position mutierten Aminosäure kann dies entweder zur Nachtblindheit, *retinitis pigmentosa* (PR) oder zur *congenital stationary night blindness* (CSNB) führen. Bei letzterer handelt es sich um eine nicht progressiv, vererbte Netzhauterkrankung des Zentralnervensystems, die durch eine Abnahme des Dunkel- und Nachtsichtvermögens, Anpassungsprobleme an die Dunkelheit und in einigen Fällen einen Verlust der Sehschärfe gekennzeichnet ist. G90D ist eine konstitutiv aktive Mutation, die die Position der Netzhauttasche beeinflusst und zu einer konstanten Basalaktivität führt. Obwohl die Kristallstruktur des photoaktiven Zustands der G90D Mutante bereits existiert, kann durch deren geringe Auflösung keine Aussage über die genaue Konformation des Retinals getroffen werden. Das publizierte Ergebnis, dass diese Mutation sich auch im Gegensatz zum Rhodopsin Wild Typ nicht im dunklen (Grund-) Zustand kristallisiert lässt, spiegelt den entscheidenden Effekt der Mutation auf die Bindungstaschenumgebung wider. Der genaue molekulare Mechanismus der Krankheit blieb durch die fehlende Elektronendichte unklar. Daher wurden mehrere Hypothesen für die experimentell beobachtete hohe Basalaktivität der G90D Mutante vorgeschlagen. Eine davon beruht auf einer spontanen Isomerisierung des Chromophors; eine Andere schlägt einen prä-aktiven Dunkelzustand des Proteins als Ursache für die Nachtblindheit vor. Die Datenlage ermöglichte jedoch keine eindeutige Erklärung des molekularen Mechanismus.

Um die fehlenden Strukturinformationen bezüglich der exakten Konformation des Retinals und der mutationsinduzierten strukturellen Veränderungen der Bindungstasche der G90D Mutante zu erhalten, wurde eine Kombination von Flüssig und Festkörper NMR Experimenten durchgeführt. Weiterhin lieferten kinetische Studien inklusive Flash-Photolyse und ultraschneller Absorption neue Informationen über den Prozess der Lichtaktivierung der G90D Mutante.

Die in dieser Dissertation durchgeführten Untersuchungen basieren auf einer Analyse von drei verschiedenen Rhodopsinkonstrukten: Wild Typ (wt), stabilisierter Wild Typ (N2C/D282C) und krankheits-assoziierte stabilisierte (N2C/D282C) G90D Mutante; im Folgenden als „G90D Mutante“ bezeichnet. Die Doppelmutation (N2C/D282C) bildet eine zusätzliche Disulfidbindung in der extrazellulären Domäne, die für die erhöhte thermische Stabilität des Proteins sorgt. Bisher konnte rekombinantes Rhodopsin nur mit dieser Disulfidbrücke erfolgreich kristallisiert werden. Flüssigkeit NMR Experimente, die die chemische Verschiebung der fünf Tryptophan Aminosäuren als Reportersignale nutzen, zeigten keinen signifikanten Einfluss der stabilisierten Doppelmutation auf die Konformation des Proteins im inaktiven, dunklen Zustand. Die photoaktive, beleuchtete Konformation wird jedoch durch die Doppelmutation stabilisiert, so dass ihre Lebensdauer auf mehrere Tage verlängert wird und nicht wie der Wild Typ nach Beleuchtung präzipitiert.

Stabilisierende Mutationen beeinflussen den gesamten kinetischen Prozess und führen zu einer Verlängerung der Lebensdauer der Zwischenzustände. Die Bildung eines frühen Photoprodukts (Bathorhodopsin) des stabilisierten Wild Typs dauert viermal länger als beim unmodifizierten Konstrukt. Der stabilisierende Effekt der N2C/D282C Mutation beeinflusst auch den Zerfall der Meta II und Meta III Zwischenzustände. Somit wird der Zerfall von Meta II und die Bildung von Meta III um 50% beschleunigt, und die Geschwindigkeit des Zerfalls von Meta III wiederum wird vervierfacht. Das kovalent gebundene Retinal wird stabilisiert und dessen Hydrolyse wird weiterhin verzögert.

Die NMR Experimente in Lösung zeigen strukturelle Ähnlichkeiten zwischen den Dunkel- und Lichtaktiven Zuständen der G90D Mutante, was auf die Bildung eines prä-aktiven Dunkelzustands hinweist. Im Gegensatz zum Wild Typ weist die G90D Mutante eine strukturelle Heterogenität auf. Diese Mutation zeigt ein zusätzliches Tryptophan Signals, welches interessanterweise auch nach der Beleuchtung im Meta II Zustand erhalten bleibt. Diese strukturelle Heterogenität wurde durch Festkörper NMR Experimente untermauert. Sowohl die Schiff'sche Base des Proteins als auch das C14 Kohlenstoff Atom des Retinals sind an der Bildung einer zusätzlichen kleineren Proteipopulation im Dunkelzustand beteiligt. Hinsichtlich der Konformation des Retinals konnte geklärt werden, dass alle beobachteten Proteipopulationen nur das *11-cis*-Retinal enthalten, dessen geringe konformationelle Heterogenität auf die unterschiedliche Position in der Bindungstasche zurückzuführen ist.

Die Beleuchtung des Proteins induziert die *11-cis* zu *all-trans* Retinal Isomerisierung, die zur Entstehung einer Reihe von Zwischenprodukten führt. Der aktive Meta II Zustand des Wild Typs ist durch eine deprotonierte Schiff'sche Base und ein vollständig kovalent gebundenes *all-trans* Retinal charakterisiert. Im Rahmen dieser Arbeit wurde die protonierte Form des Meta II Zustandes der G90D Mutante nachgewiesen. Dieser Zustand wurde durch Festkörper NMR Spektroskopie beobachtet und durch Flash-Photolyse bestätigt. Weiterhin konnte die vergleichsweise kurzlebige Bathorhodopsin Konformation unter Verwendung einer Kryoanlage ausgefroren und mittels Festkörper NMR analysiert werden. Interessanterweise wurden hierbei kleinere Spuren des photoaktiven Meta II Zustands in der Batho Konformation identifiziert. Diese wiesen auf die zuvor beobachtete Bildung des prä-aktiven Dunklen Zustands hin. Zusätzlich wurde die protonierte Schiff'sche Base sowohl für den Batho als auch für den Meta II Zustand detektiert.

Die einzigartige Information bezüglich der Auswirkung der Mutation auf den Rhodopsin-Photozyklus konnte in Kinetik-Experimenten beobachtet werden. Frühe Photozyklusprozesse (Kinetik innerhalb der ersten Minute) wurden in ultraschnellen Absorptionsmessungen charakterisiert. Im Gegensatz zum Wild Typ zeigte die G90D Mutante eine Photorhodopsin Konformation, die auf eine deutlich verlangsamte Retinalisomerisierung hindeutet. Die weitere Entwicklung der lichtaktiven Zwischenzustände von Meta II

Zusammenfassung

und Meta III wurde anhand der Flash Photolyse Experimente (Kinetik bis zu zwei Stunden) beschrieben. Die Bildung des Meta III-Zustands und der Zerfall des protonierten Meta II Zustands erfolgen mit der gleichen Geschwindigkeit, wie dies für den stabilisierten Wild Typ beobachtet wurde. Dagegen zerfällt der Meta III Zustand der G90D Mutante um 25% schneller als der stabilisierte Wild Typ.

Zusammengenommen berichten die Ergebnisse dieser Arbeit eine heterogene prä-aktive Konformation des Grundzustands der G90D Mutante. Außerdem wurde nur ein *11-cis* gebundenes Retinalisomer beobachtet, was die spontane Retinalisomerisierung ausschließt. Im Gegensatz zum Wild Typ wurde das lichtaktive Meta II Zwischenprodukt der G90D Mutante im protonierten Zustand detektiert. Diese Mutation hat signifikanten Einfluss auf den Photozyklus und führt zu einem einzigartigen kinetischen Profil. Somit wird die lichtinduzierte Isomerisierung des Retinals verzögert, während dessen Freisetzung schneller und wahrscheinlich teilweise sogar vor Beginn der Bildung der Zwischenzustände Meta II und Meta III stattfindet. All diese Ergebnisse unterstützen die Theorie eines prä-aktiven Grundzustands und liefern neue Datenpunkte für ein besseres Verständnis des Mechanismus, der zu einer Veränderung des Sehzyklus führt.

List of Abbreviations

2D	Two dimensional
a.u.	Arbitrary units
aa	Amino acids
AB	Antibody
ABCR	(ATP)-binding cassette transporter
<i>ampR</i>	Ampicillin
AMPs	Antimicrobial peptides
ARIA	Ambiguous restraint for iterative assignment
ATP	Adenosine triphosphate
ATP	Adenosine triphosphate
BEST	Band selective short transient
BK	Bradykinin
BMRB	Biological Magnetic Resonance data bank
BMRZ	Biomolecular Magnetic Resonance
BR	Bradykinin G-protein-coupled receptor
BSA	Bovine serum albumin
CD	Circular dichroism
cGMP	Cyclic guanosine monophosphate
CNG	Cyclic-nucleotide-gated
COSY	Correlated spectroscopy
CRBP1	Cellular retinol-binding protein-1
Cryo	Cryogenic
CS	Compressed Sensing
CSNB	Congenital stationary night blindness
CSP	Chemical shift perturbation
CYANA	Combined assignment and dynamics algorithm for NMR applications
DAKD	desArg10-kallidin
DDM	n-Dodecyl β -D-maltoside
DFG	Deutsche Forschungsgemeinschaft
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNP	Dynamic nuclear polarization
DOPC	1,2-dioleoyl-sn-glycero-3-phospholone
DQF	Double quantum-filtered
DQ-SQ	Double quantum-single quantum
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid
DTT	Dithiothreitol
DUF	Domain with unknown function
EDTA	Ethylenediaminetetraacetic acid
eq	Equivalent
ERG	Electroretinography
ESI	Electrospray-ionization
EX4	Exendin-4
FDA	Food and Drug Administration
FID	Free induction decay
Fmoc	N-(9-fluorenyl) methoxycarbonyl
FPLC	Fast protein liquid chromatography
FTT	Fast Fourier Transformation
GARP	Globally optimized alternating phase rectangular pulse
GC	Guanylate cyclase
GDP	Guanosine diphosphate

List of Abbreviations

GMP	Guanosine monophosphate
GPCR	G-protein coupled receptor
GST	Glutathione S-transferase
GTP	Guanosine triphosphate
GTP	Guanosine triphosphate
H/D	Hydrogen/Deuterium
HD	Hyperdimensional
HDQC	Heteronuclear double quantum coherence
hetNOE	Heteronuclear NOE
His-tag	Polyhistidine tag
HMQC	Heteronuclear multiple quantum coherence
HPLC	High pressure liquid chromatography
HSQC	Heteronuclear single quantum coherence
<i>HVO</i>	<i>Haloferax volcanii</i>
IDPs	Intrinsically disordered proteins
IHT	Iterative hard thresholding
IMAC	Immobilized metal ion affinity chromatography
IPTG	Isopropyl b-D-1-thiogalactopyranoside
IRBP	Inter photoreceptor retinoid-binding protein
IRD	Inherited retinal dystrophy
IRLS	Iterative re-weighted least squared
IST	Iterative soft thresholding
Kn	Kanamycin
LB	Lysogeny Broth
LCA	Leber congenital amaurosis
LDMs	Lifetime density maps
lncRNA	Long non-coding RNA
LRAT	Lecithin retinol acyl transferase
lw	Linewidth
MALDI	Matrix-assisted laser desorption/ionization
MAS	Magic angle spinning
MBP	Maltose-binding protein
MDD	Multidimensional Decomposition
mdeg	Millidegree
Miz-1	Myc-interacting zinc finger protein 1
mRNA	Messenger RNA
MS	Mass spectrometry
MWCO	Molecular weight cutoff
NADPT	Nicotinamide adenine dinucleotide phosphate
NCKX	Na ⁺ /Ca ²⁺ , K ⁺ exchanger
ncRNAs	Non-coding RNA
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser enhancement spectroscopy
NRPs	Non-ribosomally synthesized peptides
NS	Number of scans
NUS	Non-uniform sampling
NusA	N- utilization substance A
OD	Optical density
OG	Octyl β-D-glucopyranoside
ORFs	Open reading frames
PAGE	Polyacrylamide gel electrophoresis
PDB	Protein data bank
PDE	Phosphodiesterase
pl	Isoelectric point

pSB	Protonated Schiff base signal
RDH	Retinol dehydrogenase
Ribo-Seq	Ribosome sequencing
RK	Rhodopsin kinase
RMSD	Root mean square deviation
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RP	Retinitis pigmentosa
RPE	Retinal pigment epithelium
rpm	Rounds per minute
RPs	Ribosomally synthesized peptides
rreB	Ribosome- regulated expression in Bradyrhizobiaceae
rreR	Ribosome- regulated expression in Rhodobacter
rRNA	Ribosomal RNA
SB	Schiff base
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
sf	SOFAST, Band selective optimized flip angle short transient
SOFAST	Selectively optimized flip-angle short transient
sORFs	Small open reading frames
SP	Small proteins
SPPS	Solid phase peptide synthesis
SUMO	Small ubiquitin-related modifier
TA	Targeted aquisition
TALOS	Torsion angle likelihood obtained from shift and sequence similarity
Tc	Tetracycline
TEMPO	(2,2,6,6-Tetramethylpiperidin-1-yl)oxyl
TEV	Tabacco etch virus
Thr	thrombin
TRIS	Tris(hydroxymethyl-)aminoethan
tRNA	Transfer RNA
TROSY	Transverse relaxation optimized spectroscopy
TRX	Thioredoxin
UV	Ultraviolet
v/v	Volume per volume
Vis	Visible light
w/v	Weight per volume
WaterGATE	Water suppression by gradient tailored excitation
wt	Wild type
ZF	Zinc finger

Chapter I

Methodological aspects of protein NMR spectroscopy

1. A brief history of NMR

The phenomenon of nuclear magnetic resonance spectroscopy (NMR) discovered in 1946 by Felix Bloch and Edward Purcell (Nobel Prize in Physics, 1952) marked the beginning of a new era of spectroscopy. Initially, NMR was used to analyze small organic molecules but already in 1957, the first attempts to measure the spectrum of a protein were made. The first NMR spectrum was acquired on ribonuclease T1 and consisted of only four poorly resolved signals ^[1]. 20 years later, a breakthrough in the NMR field was made by implementing the Fourier-transform technique into the NMR spectroscopy methodology ^[2]. It laid the foundation to expand NMR from 1D to multidimensional spectroscopy and initiated extensive spectroscopic investigations of more complex biomolecules such as proteins and RNAs. Richard Ernst, Kurt Wüthrich, Peter Bachmann und Kuniaki Nagayama were the pioneers of the two dimensional NMR field in the mid-70s ^[3]. A plethora of different two dimensional experiments were invented ^[4] but major breakthrough was achieved by the development of the correlated spectroscopy (COSY) and nuclear Overhauser enhancement spectroscopy (NOESY). Due to this progress the first 3D NMR solution structure of a small protein (57 amino acids long proteinase inhibitor) solved in 1985 by Kurt Wüthrich (Nobel Prize in Chemistry, 2002) and coworkers ^[5] and a structure of a DNA binding protein solved in the group of Kaptein ^[6] were published. That allowed determining structures of proteins in solution and not in the crystals, as it is usually done by X-ray crystallography. Since then, NMR became the second method after X-ray crystallography for protein structure determination and represents an absolutely unique technology for studying dynamic systems.

The next phase in the history of NMR development was associated with the discovery of the heteronuclear spectroscopy. Isotopic labeling of a sample with the non-radioactive but magnetically active and low-abundant ¹⁵N and ¹³C nuclei triggered the development of experiments, in which the carbon, nitrogen and protons are correlated ^[7]. Development of multidimensional spectroscopy, where each axis encodes information about the corresponding nucleus, allowed to increase the size limitation of proteins amenable for NMR structural analysis, and to perform structural characteristic on the atomic level. Recently, specific labeling schemes and special pulse sequences were applied on large complexes of up to mega Dalton ^[8]. Such studies significantly extend the limits of the liquid state NMR methodology. However, NMR still remains extremely challenging and extraordinary in the field of structural biology ^[9].

Another innovative approach for NMR structural analysis is based on the increase of dimensionality of the spectra. High dimensional NMR experiments (4D-6D) were developed for structural characterization of complex systems ^[10]. The main application of these innovative experiments was found for intrinsically disordered proteins (IDPs), the spectra of which are typically characterized by low signal dispersion

compared to spectra of folded proteins ^[11]. High-dimension technique allows resolving the signals and performing the assignments of complex systems. However, the price for the resolution is the unreasonable experimental time needed for recording the nD spectra (**Figure 1.1A**). Without additional algorithms that accelerate the acquisition of these experiments, their application remains impractical. In addition, development of time-saving NMR techniques is often driven by limited intrinsic stability of biological samples. The set of experiments must be performed during the lifetime of the protein, which in some cases can significantly complicate the procedure. Furthermore, screening studies within the framework of structural genomic research would also benefit from accelerating the experimental time required for data acquisition. Therefore, subsequent efforts of various research groups in recent years have focused on the development of new techniques, aiming to reduce the experimental and computational time of multidimensional NMR experiments.

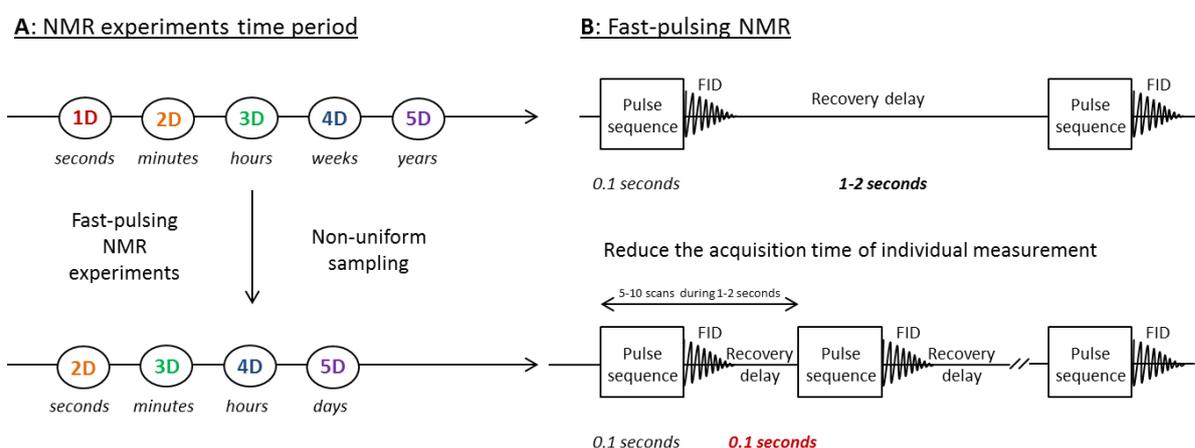


Figure 1.1. Accelerating NMR experiments. A – Experimental duration of standard and optimized NMR experiments. B – Fast-pulsing NMR methods.

2. Speed up the measurement time

Resolution in multidimensional spectra is determined by the increments recorded in the indirect dimension. Thus, the experimental time is directly correlated with the number of measurements and duration of one individual measurement. It becomes clear that there are two basic approaches to speed up the nD NMR experiments ^[12]:

- reduce the acquisition time of individual measurement and/or
- reduce the number of points (increments) recorded in the indirect dimension.

The first method includes the development of new fast-pulsing techniques, which allow to reduce the waiting time between two measurements and to increase the repetition rate of recording the data points (**Figure 1.1B**). The second method uses sparse data sampling, in which only a part of the data points is recorded in the indirect dimension. The missing data points are subsequently reconstructed

with mathematical algorithms and the full sampled spectrum is back-calculated. By combining both of these approaches, fast-pulsing and non-uniform sampling (NUS), it is possible to accelerate the measurement time by 2-3 orders of magnitude (**Figure 1.1A**).

2.1. Fast-pulsing techniques

Following the definition by Ernst, a 2D NMR experiment consists of four main blocks:

- Preparation (recovery delay)
- Excitation and evolution
- Mixing
- Detection

The recovery delay is determined mainly by the spin-lattice relaxation time T_1 (longitudinal spin relaxation), which describes the relaxation of the excited nuclear spins back to their Boltzmann thermal equilibrium. For the backbone amide protons this value is about one second. If all protons of a protein are uniformly excited, the required recovery time is significantly longer than the combined length of the pulse sequence and detection.

Longitudinal relaxation includes proton-proton dipolar interaction and hydrogen exchange between labile protons and solvent protons. Protons located in the core of a well-folded rigid protein undergo the cross- and auto-relaxation processes, resulting in polarization transfer within the dipolar lattice network. In contrast, flexible and unstructured proteins or small peptides cannot benefit from this type of relaxation enhancement.

Another way to reduce the recovery delay and to increase the signal-to-noise (S/N) ratio is to make use of unselective Ernst angle excitation^[13]. Equation 1 demonstrates the correlation between the recovery delay (T_{rec}), the longitudinal spin relaxation (T_1) and the effective flip angle (β_{Ernst}) and represents a general condition for maximum sensitivity.

$$\cos \beta_{Ernst} = e^{-\frac{T_{rec}}{T_1}} \quad (1)$$

According to this mathematical relation, when using a hard pulse with a length of 90° , the recovery delay should be at least four times longer than the longitudinal relaxation time. The correlation between the flip angle and the recovery time allows optimization of the pulse length ($>90^\circ$) dependent on the desired magnetization recovery efficiency. Decreasing the flip angle will result in incomplete magnetization recovery of spins. Due to the complex magnetization transfer path caused by the non 90°

pulse, this type of fast-pulsing technique is mainly used in HMQC-based (heteronuclear multiple-quantum correlation) NMR experiments, in which the first hard pulse is followed by a single 180° ^1H pulse and the relaxation disorders are therefore avoided ^[14]. SOFAST-HMQC (selectively optimized flip-angle short transient) is a second generation of HMQC experiment, which combines the concept of Ernst flip angle with selective excitation ^[15,16]. This experiment is particularly useful for the systems with short sample life times or limited sample concentrations and is successfully applied to the real-time kinetic studies of protein folding or binding processes ^[17]. SOFAST-HMQC can be performed in dual mode, either with the goal of i) high repetition rate experiments, which is achieved by minimizing the recovery delay up to milliseconds, or ii) high sensitivity, which is gained by optimizing the correlation between flip angle and recovery time (equation 1). SOFAST-HMQC experiments are the most sensitive and rapid 2D NMR experiments.

In contrast to Ernst flip angle based experiments, another fast-pulsing method called BEST (band-selective excitation short-transient) found wide application in multidimensional protein NMR field ^[18]. Similarly to SOFAST-HMQC, BEST approach is based on different relaxation behavior of selectively and non-selectively excited groups of spins. Selectively excited specific clusters of spins (e.g. amide, imino, methyl protons etc.) are subjected to a more rapid relaxation mechanism than the unselectively excited compounds. This effect is caused by partial absorption of magnetization by not selected spins, providing an alternative relaxation path for the selectively excited ones. This spectroscopic trick reduces the longitudinal relaxation time and speeds up the recycle delay by three to five times. Most of the triple-resonance experiments used for the protein assignment were developed using amide proton optimized BEST-type sequences. These experiments are part of the pulse sequence library implemented in the Bruker TopSpin software.

The second main approach, aimed at reducing the measurement time and making high dimensional NMR experiments more convenient for the general use, is based on minimization of acquired data points in the indirect dimension. The so-called sparse or non-linear sampling and corresponding mathematical processing algorithms are discussed in the next chapter.

2.2. Non-uniform sampling

2.2.1. NUS: general idea

The non-uniform sampling (NUS) approach was first applied on 2D NMR spectra ^[19] but nowadays it found widespread use in multidimensional NMR spectroscopy ^[20]. The main goal of this method is to reduce the experimental time and/or improve the spectral resolution during a certain measurement time. Since the NMR structure determination of proteins is based on a series of 3D NMR experiments, the acquisition of which is very time consuming, the NUS method become a routine in all NMR structural biology groups.

The main difference to the conventional methods is the acquisition scheme. In multidimensional spectra the main time limitation is determined by the increments in the indirect dimensions. In the direct dimension the acquisition time (t_2) determines the resolution and the experimental time, while in the indirect dimension the resolution is determined by the evolution time (t_1). In turn the measurement time is directly correlated with the increments recorded in the indirect dimension. The experimental time required for sufficient resolution in the indirect dimension increases proportionally to the number of increments. Thus, two increments in the indirect dimension would take twice as long as a 1D experiment.

The basic idea of NUS is to acquire 1D data at each evolution time t_1 but only for certain increments, while skipping the rest of them. The omitted increments are back-calculated during processing. The resolution is therefore not affected by the changes in the increment pattern. This allows to accelerate the experiment time without loss of the spectral resolution or to improve the resolution in a given amount of measurement time (**Figure 1.2**). Of course, NUS is not the magic tool and the price for all the benefits is a reduced signal-to-noise ratio. The latter can be manipulated using an optimized sampling scheme that takes advantage of the relaxation-based signal decay. In this case, the increments are not chosen randomly like it was proposed by Rovnyak *et al.* ^[21] but in accordance to the exponential signal attenuation ^[19], more dense at the beginning and less at the end. Another sampling approach uses a radial ^[22] /concentric pattern ^[23,24], and more modern methods are based on the Poisson distribution ^[25].

Reduction of the experimental time is not the only advantage of NUS sampling. The fact that resolution is not directly dependent on the increments allows to immensely increase the spectral resolution by keeping the machine time constant. This method becomes particularly useful in the metabolomics field, where the assignment of low abundance of metabolites is the object of study ^[26,27].

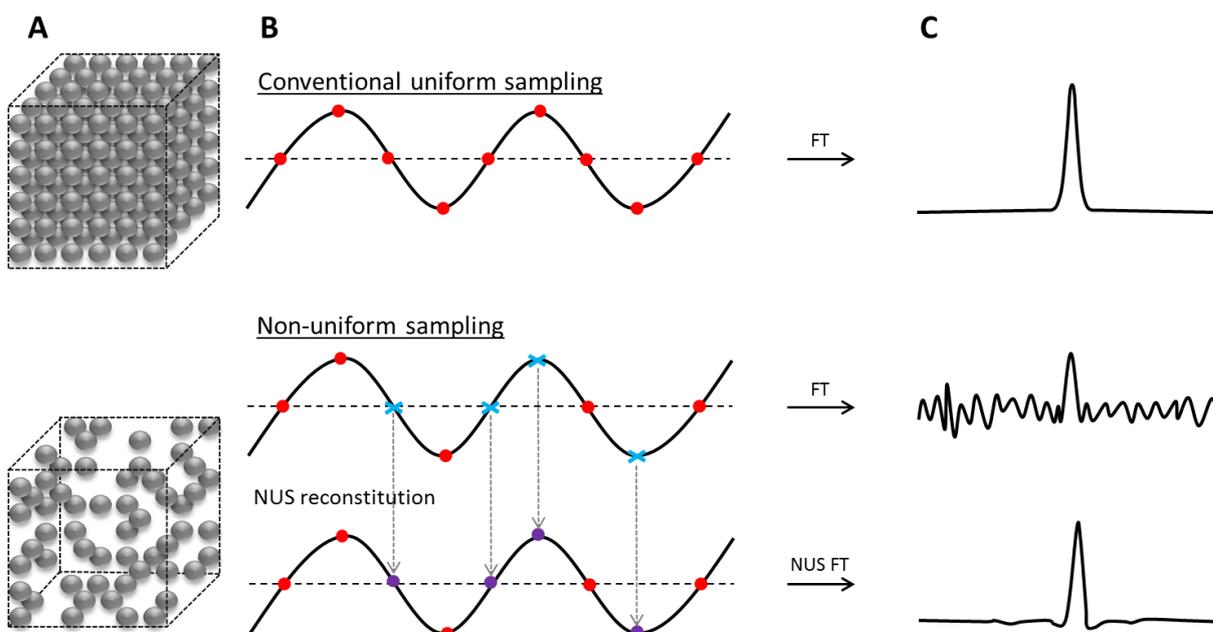


Figure 1.2. Schematic representation of different sampling approaches. Upper panel – Conventional method, spectrum is recorded with 100% points. Lower panel – Non-uniform sampling, missing points are reconstituted with mathematical algorithms. **A** – 3D sampling scheme. **B** – Reconstitution of spectrum recorded with uniform and non-uniform sampling. **C** – Fourier transform of uniform versus non-uniform sampling schedules. Adapted from Deladlio, 2017 ^[28].

2.2.2. Reconstitution approaches

The main challenge of NUS is the impossibility of using the Fast Fourier Transformation (FFT), which is the key algorithm in NMR spectroscopy. Lack of data points in the acquisition leads to random signal artefacts when using FFT ^[29]. Therefore, development of the methods that interpolate the missing data points and reconstruct the spectra is currently a growing research field. To describe the main challenge arising in the processing of the undersampled spectra, the linear algebra principles used in NMR spectroscopy will be simplified below.

A set of NMR data can be represented as a combination of linear equations (equation 2), where F is an $A \times B$ inverse FT matrix, \mathbf{a} and \mathbf{b} describe vectors of frequency and time domain with the length A and B , respectively. According to the Nyquist-Shannon theory, $A = B$ ^[30]. Mathematically, this results in a unique solution of equation 2 and is typically used in conventional methods.

$$F\mathbf{a} = \mathbf{b} \quad (2)$$

This also implies that every single point encodes significant information and is crucial for the final quality of the spectrum. However, this assumption is overestimated and usually data points that do not encode the signal can be filtered out without loss of important information. This assumption of sparse spectra is the main key point of the NUS-based methods. Mathematically, this results in the infinite solution

number of equation 2 and represents the main challenge of the reconstitution algorithms^[31]. Thus, all developed reconstitution methods use additional assumptions such as minimal power, maximum entropy etc. The easiest way of reconstruction is the principle of minimal power. It assumes that all missing points do not code important spectral information and thus may be replaced by zero. A generated “full” data set can be further processed with discrete FT. Unfortunately, each signal produced by this algorithm is accompanied by artifacts. Increasing number of signals lead to proportionally increasing amount of artifacts and thus the main challenge of the method is the distinction between real peaks and mathematically generated fake peaks. The method developed to clean up the reconstituted spectra was simply called CLEAN^[24]. It uses mathematical tools to predict the artifact pattern that could be further subtracted from the reconstituted spectrum.

The first approach which completely avoids the classical Fourier transformation was based on the concept of maximum entropy^[19]. In the past decades, researchers put a lot of effort into developing new FT-free NUS reconstitution algorithms, and now two main methods can be distinguished. Both approaches, Compressed Sensing (CS)^[29,31–33] and Multidimensional Decomposition (MDD)^[34–36], have found a widespread use in multidimensional NMR spectroscopy and represent a balance between spectra quality and processing time with their own advantages and disadvantages.

The mathematical background of CS theory was described in 2004 by Doroto^[32]. More recently, Bostock and Nietlispach have published a review with detailed theoretical explanation of newly developed CS-based reconstitution methods including an overview of their application^[29].

The CS approach is based on the assumption of sparseness of the spectra. According to CS, the best solution of equation 2 should contain the highest amount of zeros. However, searching for this solution comes at the expense of considerable amounts of computational time. Another issue arising from the CS assumption is hidden in the definition of the absolute sparseness of the spectrum. In reality, signals are never strictly sparse and typically the peak intensity attenuation can be described by Lorentzian distribution. If the number of points encoding the “true” signal is much smaller compared to the points coding noise, the CS method can be successfully applied^[37]. But in case of signals with different intensities, which are in particular common for molten globule state or for a combination of different populated conformations, the signals with low intensities may be assigned to noise and filtered out. Moreover, NOESY-based NMR experiments, where the signal intensity encodes crucial structural information, could be falsely reconstructed. In order to solve this problem, different modifications of the CS approach were developed. Now, based on the minimization protocol, CS methods can be divided into iterative soft thresholding (IST)^[38] (convex minimization) and iterative hard thresholding (IHT) (nonconvex minimization), which includes a new version called iterative re-weighted least squared

(IRLS). Both reconstitution algorithms, IST and IRLS, were verified on 2D $^1\text{H}^{15}\text{N}$ HSQC and 2D $^1\text{H}^1\text{H}$ NOESY NMR experiments ^[31]. Kazimierczuk and Orekhov have demonstrated that IST method is less time consuming and provides an excellent result for the spectra with high signal-to-noise ratio, while being enormously expensive in case of NOESY-based NMR experiments. The advantage of IST algorithm is typically used in reconstruction of 3D NMR experiments used for protein assignment. IRLS in turn allows accurate reconstitution of the signals with different intensities and different shapes in reasonable time and is therefore used for processing NOE-based NMR experiments.

Another very popular reconstitution approach, called Multidimensional Decomposition (MDD), is based on a completely different mathematical assumption than CS theory. The theory states that every signal in a multidimensional spectrum can be described as a combination of 1D projections from all dimensions. This allows to present a matrix F by a sum of tensor products of 1D vectors, where matrix F describes the undersampled spectrum (input) and 1D vectors represent line shapes (output) ^[39]. This assumption is true for most NMR experiments, where one cross peak corresponds to one component in each dimension (backbone NMR experiments). If several cross peaks are generated by one component, like in NOESY-type experiments, the MDD main assumption is no longer valid and additional algorithms are needed ^[35]. In contrast to CS approach, MDD method is much less time consuming and nowadays routinely used for 3D based NMR experiments.

Typically, the processing of the undersampled spectra requires specific software such as NMRPipe or MddNMR. But the CS and MDD algorithms have now also been implemented in the TopSpin package provided by Bruker, allowing processing of NMR spectra recorded in NUS mode. This extremely simplifies the scientific workflow and made NUS a routine approach in studying the protein structure and dynamics by NMR spectroscopy.

2.3. Targeted Acquisition approach

Recently, the group of Vladislav Orekhov developed a new software that speeds up experimental time and simplifies the data analysis process ^[39-41]. The method was called Targeted Acquisition (TA) in accordance with the main approach, which is used in the software for the recording of the spectra ^[42]. Originally this software was developed for the backbone assignment of IDPs and was verified on the example of five 6-13 kDa cytosolic domains from B- and T-cell receptors ^[41].

Basically, the TA approach combines several already existing time-saving methods (as shown below) in one program, which is very user-friendly and does not require any special knowledge about each contained unit:

- targeted acquisition ^[42]
- non-uniform sampling with multidimensional decomposition reconstitution ^[39]
- hyperdimensional (HD) spectrum algorithm ^[40]
- automated chemical shift assignment with FLYA ^[43]

The spectra are recorded in NUS mode and the reconstruction is performed using the MDD approach, which is automatically implemented in the TA software. Each multidimensional spectrum can be represented by a model of multidimensional decomposition (MDD), which assumes that M-dimensional matrix can be described as a sum of tensor products of one-dimensional vectors.

$$S_{HNCO} = \sum a F^H \otimes F^N \otimes F^{CO^{-1}} \quad (3)$$

S_{HNCO} is a model spectrum, which is given by a sum of products of the signal amplitude a and three one-dimensional shapes, in this case represented by amide proton, amide nitrogen and carbonyl carbon F^H , F^N , F^{CO} , respectively. Symbol \otimes specifies the product operation.

Typically, for protein backbone assignment 3D NMR experiments are recorded as one set with several shared dimensions, allowing to combine them in a hyper-dimensional spectrum ^[40]. To describe the main principle used in HD algorithm, a set of six 3D NMR experiments for a protein backbone assignment including e.g. HNCO, HN(co)CA, HNCA, HN(co)CACB, HNCACB, HN(ca)CO is used. All these 3D experiments have the same 2D $^1H^{15}N$ HSQC dimension, which forms a connecting bridge between the experiments. The third individual (not-shared) dimension from each experiment contributes to the HD spectrum one additional indirect (carbon) dimension, resulting in a nine-dimensional HD spectrum. Thus, the HD spectrum describes the correlation between direct and indirect dimensions for one spin system.

$$\begin{cases}
 S_{HNCO}: \sum a F^H \otimes F^N \otimes F^{CO^{-1}} \\
 S_{HN(co)CA}: \sum a F^H \otimes F^N \otimes F^{CA^{-1}} \\
 S_{HNCA}: \sum a F^H \otimes F^N \otimes F^{CA CA^{-1}} \\
 S_{HN(co)CACB}: \sum a F^H \otimes F^N \otimes F^{CA^{-1} CB^{-1}} \\
 S_{HNCACB}: \sum a F^H \otimes F^N \otimes F^{CA CB CA^{-1} CB^{-1}} \\
 S_{HN(ca)CO}: \sum a F^H \otimes F^N \otimes F^{CO CO^{-1}}
 \end{cases}$$

$$H = \sum A \otimes F^H \otimes F^N \otimes F^{CO^{-1}} \otimes F^{CA^{-1}} \otimes F^{CA CA^{-1}} \otimes F^{CA^{-1} CB^{-1}} \otimes F^{CA CB CA^{-1} CB^{-1}} \otimes F^{CO CO^{-1}} \quad (4)$$

H is a nine-dimensional HD spectrum, which is built of six low-dimensional spectra. HN and N are two shared dimensions, while all carbon dimensions belong to non-shared individual frequencies (equation 4 highlighted in red). A represents a shape matrix, which describes amplitude of the signals resulted from each experiment $A = (a^{HNCO}, a^{HN(co)CA}, a^{HNCA}, a^{HN(co)CACB}, a^{HNCACB}, a^{HN(ca)CO})$.

In order to simplify and speed up the calculation, the TA software uses the co-processing method of the experiments recorded in one set. For this the MDD model is built based on the most sensitive experiment in the set, which is the HNCO experiment. Two shared dimensions, amide proton and nitrogen, are therefore defined as known and are further used to co-process all other experiments in a set. Using fixed HN and N frequencies allows filtering out the noise-peaks and thus reduces artifacts. The created HD model described by a matrix of nine 1D line shapes needs dramatically less data space compared to the fully processed same set of experiments and is much more convenient for the further analysis. However, the accuracy of the method completely depends on the MDD model, generated from the HNCO shape line pattern. Therefore, high signal dispersion in the 2D $^1\text{H}^{15}\text{N}$ HSQC is crucial for the quality of the algorithm.

A set of NMR experiments, which are used to generate the HD spectrum is not strictly defined in the TA program and may be chosen individually by a user. Furthermore, there is no need to create a new pulse sequence for the TA set, all experiments can be selected from the Bruker pulse sequence library. Additional reduction of NMR measurement time can be achieved by using the 3D BEST-based NMR experiments, which are also included in TopSpin software. After a HD spectrum is created, the peak list can be subsequently analyzed by the automated resonance assignment FLYA ^[43], which is a part of automated structure calculation program CYANA ^[44-46]

The main idea, which makes this program so unique, is hidden behind the targeted acquisition algorithm. A set of NMR experiments is recorded in semi-interleaved incremental NUS mode. That means that at first all experiments are recorded with e.g. 1% NUS, corresponding to one step of acquisition. On the second step all experiments are recorded with the same amount of NUS, and so forth. The spectra are recorded and co-processed in parallel and can be visualized after every step in a real-time manner. Typically, the TA approach is completed when convergence of the number of signals found in individual experiments has been reached and does not increase in the next iteration. After the required quality of the spectra is achieved and the expected amount of signals is detected, the acquisition can be terminated. The ability to monitor the spectra in real-time and to finish the experiments at any step allows optimization of the required acquisition steps (NUS amount), which will greatly reduce the measurement time. The implemented automated FLYA assignment can be performed after each step of acquisition, which provides the demonstrative readout of the assignment progress.

TA software consists of two modules:

- i) spectrometer module, which includes experimental setup and data acquisition, and
- ii) linux operated terminal module with processing and analysis blocks.

The first module is responsible for the selection and optimization of the set of NMR experiments, which are run in the Bruker TopSpin software. The TA script begins with the recording of the 3D HNCO root experiment with relatively high NUS amount, which is further used for generating the MDD model. Other experiments are run in interleaved manner afterwards and can be co-processed and visualized in real-time. For co-processing of the combined set of 3D experiments the HNCO-based MDD model is used. The TA analysis and processing script requires additional Python libraries and the MddNMR software and can be operated in the terminal window.

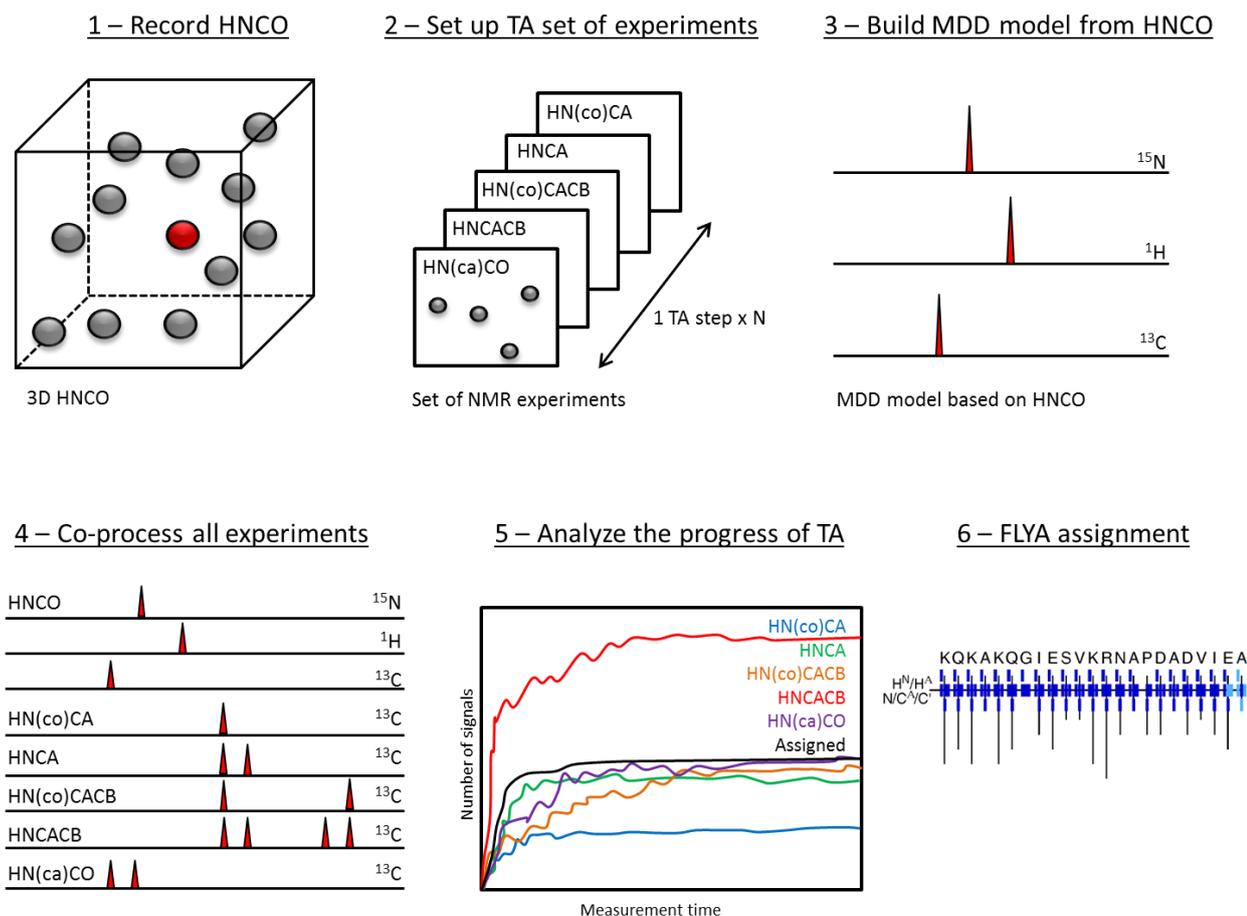


Figure 1.3. Schematic representation of the main steps of the TA algorithm.

TA algorithm includes following steps (**Figure 1.3**):

1. Create an MDD model based on 3D HNCO experiment
2. Co-process a set of TA spectra using MDD model
3. Automated peak picking of the TA processed spectra after acquisition step
4. FLYA assignment after each acquisition step
5. Graphical output of assignment and the amount of signal evolution as a function of recorded acquisition steps.

All processing steps are encoded by an individual command and can be performed separately with subsequent monitoring of each step or all together within a single command. The TA program is not restricted to use all of the processing steps described above. After the number of signals reaches steady state and the acquisition is terminated, the experiments can be reconstructed and merged for further visualization and analysis.

TA is a unique program that provides real-time analysis of individual NMR experiments recorded within a TA data set. This allows to control the required measurement time according to the desired quality of the spectra. The implementation of TA software can reduce the measurement time by approximately a factor of 20 compared to the spectra recorded with conventional methods. The ability to monitor the spectra quality in real-time allows to reduce the NUS amount and terminate the acquisition after the desired amount of signals is reached. Typically, 3D NMR experiments for backbone assignment are recorded with 25% NUS amount, TA technique shows that already 10-20% NUS, depending on the sample conditions, is sufficient for the same spectra quality. The main requirement for the successful application of this method is signal sparseness in the 3D root experiment (e.g. HNCO). If this condition is fulfilled, spectroscopists can take full advantage of this method.

In the framework of scientific collaboration with Swedish NMR center in particular with Prof. Vladislav Orekhov, the TA software was successfully installed on the 800 MHz spectrometer, which is a part of the Center for Biomolecular Magnetic Resonance (BMRZ) equipment and is now routinely used in Prof. Schwalbe group for the structural analysis of small proteins.

Chapter II

Rapid NMR and biophysical characterization of small proteins

The following colleagues contributed to this chapter: Dr. Henry Jonker provided help with structure calculation of small protein P17, Dennis Pyper in context of his master thesis performed a titration series of small protein P19 and purified small protein P8. Dr. Monika Fuxreiter from University of Debrecen (Hungary) performed Espritz NMR prediction analysis of all screened peptides and small proteins.

1. General introduction

Small open reading frames (sORFs) were excluded from the genome annotation for a very long time. This has led to the disregard of small proteins (SP) encoded by sORFs. Recently, computational and experimental progress in the identification of proteins of this type has made it possible to recognize the important role of small proteins in many essential cellular processes, giving rise to a new era in the study of small proteins with a focus on their identification and functional characterization. Currently, hundreds of previously unknown small peptides have been identified. However, the functional characterization of a rapidly growing number of the new small proteins remains challenging and is a key point of current research topics. Structure is often linked to function. Thus, the investigation of this well-known protein structure-function paradigm can now be extended to the set of proteins that are significantly smaller than the previously characterized proteome. In that sense, the characterization of structure and dynamics of small proteins is a very important part of a puzzle for studying general biological pathways.

1.1. Nomenclature diversity

In recent years, the expansion of small proteins research field has grown exponentially, encompassing more and more scientific research groups interested in the identification and functional characterization of small proteins. So far, there is no universally accepted nomenclature. Commonly, the size of the amino acid sequence determines the name, distinguishing between proteins, small proteins and peptides. The limitations of these definitions are, however, vague. Historically, it is considered that proteins consist of 50 to 1000 amino acids (aa), while peptides are in the smaller range, covering 2 to 50 aa. Peptides can be classified into oligopeptides (2-20 aa) and polypeptides (20-50 aa). However, in order to emphasize the difference in the length between large proteins and short peptides, some researchers use the term *peptides* instead of *oligopeptides* and *small proteins* instead of *polypeptides*. Even more, according to another definition, peptides are the truncated products of the partial hydrolysis of a protein^[46]. Typically, the cutoff between peptides and proteins is approximately 50-100 aa. In most studies proteins containing less than 100 aa are defined as small proteins^[47]. Nevertheless, other cutoffs like 200 aa^[48] or 85 aa^[49] are also known. In 2014, the pioneer of the research field small proteins, Gisela Storz, set the upper limit to 50 amino acids (aa), defining them as small proteins that are encoded by small open reading frames (ORFs)^[46]. Since then, the maximal cutoff size varies in the range up to 100 aa. The diversity of names includes the terms: *very small proteins*^[50], *short proteins*^[51], *microproteins*^[52] / μ -*proteins*^[50], *miniproteins*^[53], *peptides*^[46] and *micropeptides* / μ -*peptides*^[54]. Even within the same nomenclature the definition may vary. For example, the upper limit of the term μ -*protein* defined by Nagel *et al.* is 70 aa, while Baumgartner *et al.* use the cutoff of 80 aa for the same name^[55].

All amino acid sequences screened during this doctoral thesis (14-71 aa) fulfil the definition of small proteins. But in order to emphasize the impact of the size on the structural behavior of the protein and to notice the type of synthesis used, the terms *peptides* and *small proteins* are implemented in this study. Peptides consist of less than ca. 30 aa and are synthesized by solid phase peptide synthesis, while small proteins contain more than 30 aa and are differentially expressed. This definition is convenient for the particular screening study but does not pretend to propose general uniform usage.

1.2. Historical neglect of small proteins

Historically, it was assumed that mainly large proteins are encoded in the genome and the short genome sequences do not carry genetic information (non-coding), allowing their exclusion from the genome annotation. This assumption decreased the theoretical number of open reading frames and simplified the genome annotation procedure. An arbitrarily defined cutoff of 100 codons was applied as a decisive filter criterion for the functionality of the gene. Genes consisting of more than 300 nucleotides, which correspond to proteins longer than 100 aa, were decided to be included in the annotation profile, while ORFs with less codons were neglected. Noteworthy, the definition of 100 codons is not strict, making a general cutoff for genomes from all domains of life not possible^[56]. Thus, the cutoff used in bacteria is shorter than 300 nucleotides, while the cutoff used for the annotation of Haloarchaea was defined as 100 nucleotides, or 34 aa for a protein. The cutoff limitation determined by the genetic sequence data GenBank is, however, 200 nucleotides, which correspond to a minimal length of 66 aa for a protein^[46]. Despite the diversity in the minimal cutoff definition, the size of 100 codons is most commonly found in the literature.

The technical methods used for the detection of the proteins also did not give the small proteins a chance to be identified. For example, in case of co-purification with another protein, the typically used SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) does not show a characteristic band for proteins smaller than 10 kDa. Without using special gels, adjusted for the short length, these small proteins would simply run off the gel. If the small proteins were not lost during the purification and reached the phase of mass spectrometry detection, they were potentially ignored because of lack in the genome annotation. Consequently, the small proteins were overlooked in the gene annotation and, as a result, in the general biological studies.

After some sORFs encoding small proteins were identified in several organisms and their functional importance had been shown^[57,58], the confidence in previous assumption, regarding the cutoff of 100 aa, was undermined.

1.3. Identification of sORFs coding small proteins

The largest challenge in genome annotation is distinguishing between small coding and large non-coding ORFs. The high abundance of non-coding ORFs and their random distribution throughout the whole genome makes the identification and filtering of the not-translated sequences difficult. Since all mRNA transcripts include multiple ORFs, which differ in their lengths, the assumption that the longest ORFs encode functional proteins and the remaining ORFs may be neglected, leads to the loss of the small proteins during the sequencing. Thus, it is challenging to identify sORFs, which encode small proteins, among a high number of long non-coding RNA (lncRNA). The revision of the previously used cutoff of 100 codons as the detection limit for coding ORFs immediately increases the number of possible ORFs in the genome, increasing the complexity of the analysis and requiring the development of new approaches for their identification.

1.3.1. Computational approaches

Computational approaches for predicting the putative coding sORFs are based on differences in the level of conservation between the coding and non-coding sequences in the genome. The fact that coding sequences are highly conserved compared to ncRNAs facilitates the development of these methods. The prediction of the protein coding potential can be complemented by the evolutionary developed synonymous codon substitutions^[59]. Although the size of the small proteins limits the possible nucleotide changes, this nucleotide composition analysis was included in the computational approach, which led to an accuracy improvement of the sORFs identification^[54].

1.3.2. Experimental biochemical approaches

Besides computational methods the experimental approaches such as ribosomal profiling and mass spectrometry are often used for the identification of coding sORFs in the genome sequence.

1.3.2.1. Ribosome profiling

Ribosome profiling, also known as ribosome foot printing or Ribo-Seq, is a relatively new method that allows the monitoring of the protein translation *in vivo* and is particularly powerful for the identification of sORFs. By using this method, it is possible to identify which proteins are produced within the cell. The technique is based on the deep-sequencing of ribosome-protected mRNA fragments (RPFs) and works as follows: during the translation the ribosome covers short mRNA fragments (ca. 30 nucleotides) protecting them from digestion by RNases. Using the inhibitor (typically cycloheximide) the translation can be stopped and the ribosome position can be trapped at any given moment of time. The mRNA regions, which are not protected by ribosomes are degraded by the enzyme nuclease. The remaining fragments, also called monosome, consist of ribosome, RPFs and translated protein in case of prolonged

translation inhibition. The mRNA-ribosome complexes are subsequently isolated and purified (typically by the sucrose gradient). From the extracted RNA, also called ribosome footprints, the DNA library is generated, which is further used for the deep sequencing analysis. Deep sequencing simultaneously performs millions of sequencing reactions and provides information about the abundance of RNA fragments in the transcript. Based on the speed of the movement of the ribosome, this analysis can give an estimation of the rate of the protein synthesis. The higher amount of RPFs corresponds to the slower motion of the ribosome on this transcript. Additionally, the localization of the ribosome on the mRNA during the translation can be performed on the sub-codon level, which allows identifying the new coding sORFs in the sequence and performs a “snap-chat” of the translational process^[60,61].

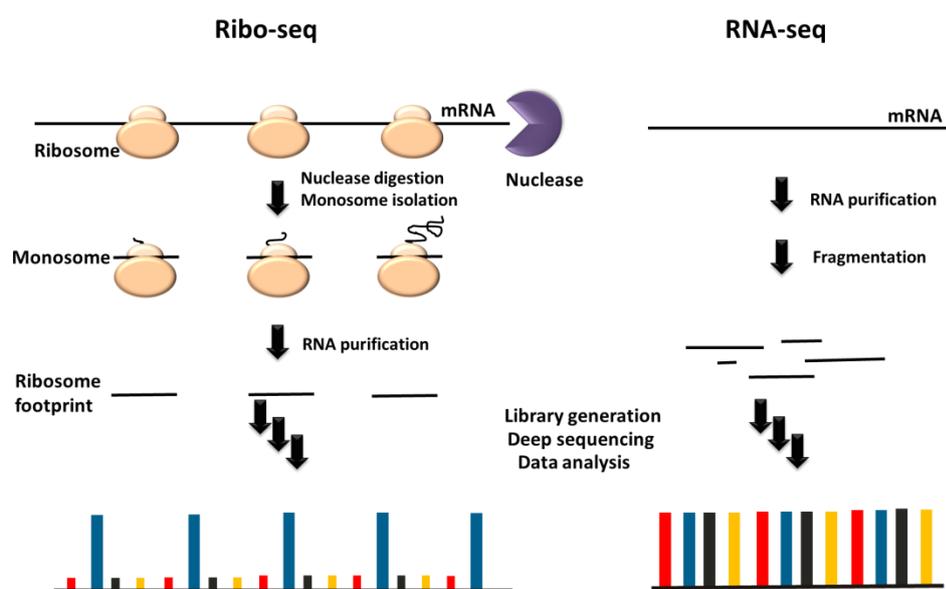


Figure 2.1. Experimental biochemical methods based on sequencing of the mRNA template for the identification of sORFs. Left panel – ribosome profiling; right panel – RNA sequencing approach. Adapted from Hsu, 2016^[62].

The advantage of ribosome profiling in comparison with its predecessor, total mRNA sequencing, which was typically used prior to the discovery of the method, is in using the ribosome as a protector for the mRNA (**Figure 2.1**). The RNA-Seq is based on the total random fragmentation of the mRNA template, which results in less accurate fragment abundance compared to Ribo-Seq. More importantly, RNA-Seq uses size-based filtering of the RNA fragments which results in the loss of sORFs during the genome annotation^[60].

Ribosomal profiling approach has contributed to the discovery of a big number of small proteins in all domains of life. New sequences were extensively studied in organism such as yeast^[61,64], plant^[64], insect^[65], zebrafish^[66], mouse^[61], humans^[67] etc. Besides the identification of new sequences, the improvement and further development of the methods is a subject of research by many scientific groups. The main issue in identifying sORFs using the Ribo-Seq approach is the prediction of the coding potential of the identified mRNA fragments. The mRNA position recognized by the ribosome does not

necessarily implement the functional protein encoded by this sequence. Recently in 2019, Gisela Storz and colleagues have improved the ribosome profiling method by combining it with additional information about the initiations site. This allowed the increase in the prediction rate of the small protein expressed in the cell (38 out of 41 tested genes were experimentally expressed in *E. coli*)^[68].

1.3.2.2. Mass spectrometry

Mass spectrometry (MS) is the second highly important method for the identification of small proteins routinely applied in peptidomics research. In contrast to the sequencing-based methods, MS detects proteins, which have already been translated; thus, avoiding the coding potential issue. This powerful technique contributes to the ever-growing number of newly found proteins but a high percentage of identified segments remain unassigned. This reflects the lack of annotation for coding sORFs in the genomic studies.

Initially, the experimentally obtained data were compared with the existing annotated database. This procedure did not allow the identification of new sORFs and had to be optimized for the identification of small proteins. An experimental breakthrough was achieved by Slavoff *et al.* by combining experimental MS data with massively parallel RNA sequencing approach^[69]. In this case, the obtained data set from the MS experiments for a particular transcriptome was compared with all possible theoretical proteins that can be translated by this RNA. This allowed the identification of small proteins, which were previously not annotated in the global database. Typically, the match between the MS detected protein with the predicted ORF or translated protein should occur multiple times but in case of small proteins, the size significantly reduces the abundance of the match^[70]. Nevertheless, this advanced method was applied to the human genome and 86 new coding sORFs were identified.

Despite all the practical advantages, MS still has to deal with some challenges regarding the identifications of small proteins. Thus, due to the small size these proteins may be lost during the sample purification or some unstable proteins may not survive until detection and get degraded during the purification. Further, MS is fundamentally not quantitative. Still, MS is the most important experimental proof of existence for small proteins translated in the cell.

Computational and ribosome profiling approaches contributed to the discovery of hundreds of previously unknown small proteins. However, these methods do not answer the question of whether the new identified proteins are indeed being translated in the cell and are biologically functional. Therefore, often the results obtained using computational and Ribo-Seq approaches are further validated by mass spectrometry. The combination of all these approaches is successfully used by various research groups and hundreds of newly identified small proteins have been already identified in all domains of life.

1.4. Functional studies

Once the sORFs have been identified via computational and experimental ribosome profiling experiments, and the existence of translated small proteins in the cell was confirmed by mass spectrometry analysis, the other important questions arise, whether these small proteins are biologically active and what their physiological role is. The fact that the protein was translated in the cell does not necessarily imply that it has an important biological function. To analyze and determine the cellular functional activity of newly identified small proteins, each candidate has to be investigated individually. Since there are hundreds of proteins of this type, this implies a huge amount of work for microbiologists.

Nowadays, the most common method to analyze the biological impact of the protein is the loss-of-function approach ^[71]. Typically, knockout or untargeted mutagenesis approaches are used to remove the gene of interest from the system and investigate the resulting phenotype. The difference in the phenotype between the mutant and the wild type should provide the information about possible biological roles of the protein. Nevertheless, very often these experiments fail. One reason could be that small proteins often act as a part of biologically active large protein complexes.

Small proteins were found to be involved in various regulatory cellular pathways, encompassing processes such as stress response ^[73-75], DNA repair ^[74], RNA decapping ^[63], Ca²⁺ homeostasis ^[75], metabolism ^[76], cell division and cell death ^[77]. Most of these biological functions involve complex formation. Thus, identification of the potential interaction partners is critical for the further elucidation of the biological relevance of small proteins.

Most of the small proteins with known functions (hormones, neurotransmitters, antioxidants, toxins, antibiotics etc.) belong to the cleaved proteolytic products from large proteins, which are biologically active.

1.4.1. Structure-function paradigm

According to the structure-function paradigm, native tertiary structure of the protein defines its biological activity. The protein folding pathway is, however, a complex process, which includes numerous intermediate conformations on the way to the native functional state. Funnel energy landscape theory describes the transformation of unfolded proteins into a low-energy native state (**Figure 2.2A**). Subsequently formed intermediate states exist in conformational equilibria and adopt partially structured folds with a fluctuating amount of residual structures. These folding intermediates are referred to molten globule state and can evolve into fully or partially folded conformation. In some cases, if the kinetic barrier between folding intermediates and the native structure is high enough, the protein can be trapped in the kinetic stable conformation and cannot reach the thermodynamic stable,

low-energy state. Incorrect folding is often the cause of protein dysfunction and aggregation that promote some diseases (Alzheimer disease, Parkinson disease, diabetes etc.) [78].

Recently, the initial structure-function paradigm was completely revised. These radical changes were triggered by the extensive studies of the intrinsically disordered proteins (IDPs) [79,80] which showed regulatory functions of IDPs [81]. Since IDPs have function without structure [82], the new paradigm was significantly expanded [83].

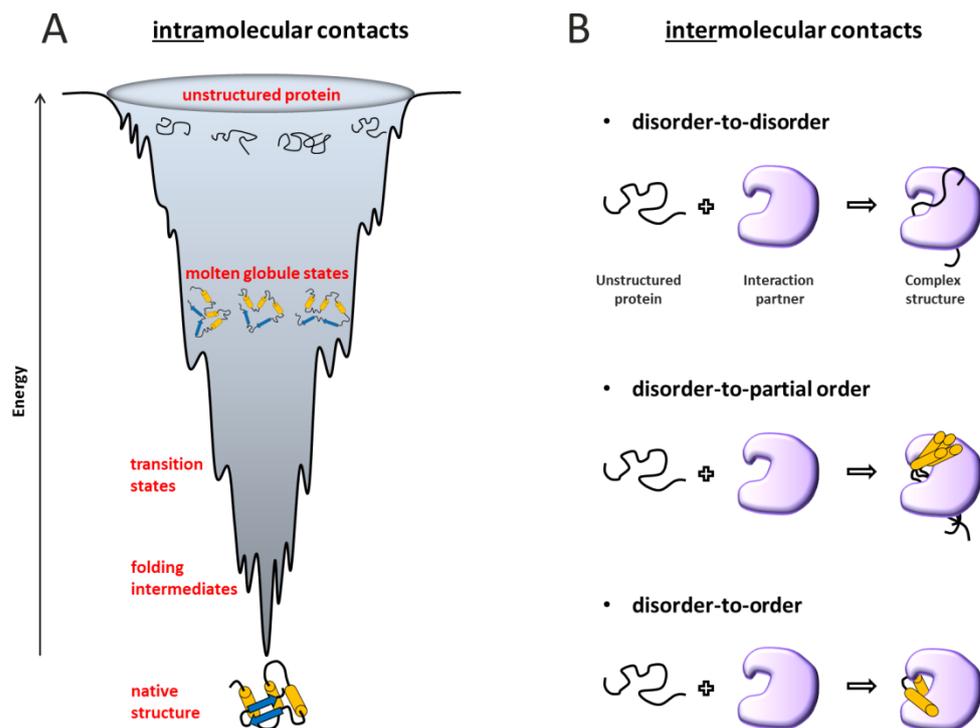


Figure 2.2. Structural transition pathways of a protein. A - Folding funnel energy landscape. Adapted from Wolynes, 1995 [84]. B - Structural transition of IDPs upon interaction with biological targets. Adapted from Miscei, 2017 [85].

Typically, flexible IDPs include short amino acid sequences with binding activity for specific interaction partners and may undergo structural rearrangements upon complex formation. Regarding the conformational changes induced by the physiological targets, these transformations can be divided into three groups: i) disorder-to-order, ii) disorder-to-partial order and iii) disorder-to-disorder (**Figure 2.2B**) [85]. During the disorder-to-order transition, interaction with binding partners induces the transformation of IDP into a well-defined tertiary structure. Moreover, folding may comprise either the whole domain or only the binding pocket of IDP, while stabilizing the remaining segments but not folding them completely. Disorder-to-partial order transition occurs when an IDP adopts a conformational mixture, similar to the molten globule state. Typically, in the biological processes these IDPs are parts of a multiple complex, requiring several ligands to adopt persistent tertiary fold. The disorder-to-disorder transition path describes structural disorder of IDPs, which can be preserved in the

bound form. Structural diversity, resulting from disorder-to-partial order and disorder-to-disorder transitions, lead to polymorphism, which might be regulated by the cellular conditions ^[85].

In a nutshell, structure of a protein formed by intramolecular interactions or by complex formation remains the central key to the elucidation of its functions.

1.4.2. Structural studies of small proteins

The size of small proteins restricts the possible diversity of secondary structure elements that might be adopted by a protein. The smaller the amino acid sequence, the lower the probability of peptides and small proteins to adopt tertiary structure. Similar to IDPs, the small proteins may need a biological interaction partner to adopt a folded conformation and become functionally active as a part of a large complex. The principle of structural transitions originally developed for IDPs can be applied to the unfolded peptides and small proteins, whose size is below the definition of IDPs (50 aa ^[81]). There are numerous examples that folding can be induced by the interaction with biological targets. Upon interaction with the human bradykinin G-protein-coupled receptor (B₂R), the nine amino acids long neuropeptide bradykinin (BK), a drug target for cardiovascular regulation (BK), undergoes a disorder-to-order transition. Unlike BK, its analog peptide desArg10-kallidin (DAKD) does not adopt a persistent tertiary fold upon binding to its receptor B₁R ^[86] (**Figure 2.3A**). Here, a disorder-to-disorder transition takes place. Interestingly, both complexes are biologically active signal transducers. This example indicates that absence of a structure does not necessarily mean lack of function.

Another specific class of peptides, which fold upon interaction with targets, is membrane peptides. These peptides comprise antimicrobial peptides (AMPs), cell-penetrating peptides, channel forming peptides and amyloid peptides and are extensively studied in the context of diseases and medical therapeutics ^[87]. Typically, they form a well-defined structure in detergent environment, which mimics the biological membrane but stay unfolded in water. For example, in 2019, NMR structures of two individual peptides from two-peptide bacteriocin plantaricin S have been published in Nature by Ekblad *et al.* Similar to all bacteriocins, these peptides adopt α -helical structures in DPC micelles, while being unstructured in aqueous solutions ^[88] (**Figure 2.3B**). Interestingly, for maximum antimicrobial activity, both of these peptides, Pls- α (27 aa) and Pls- β (26 aa), must be present in a 1:1 ration, while separately they are not of a biological importance.

In the framework of completed genome annotation, and in particular the identification of coding sORFs, it has been found that membrane small proteins are well presented in this field. In 2008, Storz and coworkers showed that half of newly identified small proteins from *E. coli* (16-50 aa) are predicted to be single-transmembrane proteins ^[89].

But what about peptides and small proteins in isolated form? Are they able to fold without biological targets, and what is the smallest size required for autonomously adopted secondary structure? The native folded conformation of proteins is stabilized by thousands of interactions including hydrophobic, electrostatic, van der Waals and hydrogen bonding formation. The size of peptides and small proteins restricts the number of possible interactions. The smallest folded systems are using additional stabilizing agents such as disulfide and metal chelation crosslinks as adjuvants for adopting a folded conformation.

NMR structures of the two smallest short peptides consisting of 21 and 23 amino acids have been published 20 years ago. The authors were interested in the smallest size that still adopts a native-like folded conformation. They designed short peptides with two disulfide bridges, removing the third disulfide bond from the original trypsin inhibitor protein, and could show that the cysteine-stabilized β -sheet motif still remains in the structure^[90] (structures are not in the pdb database). Both peptides were the subject of a theoretical structural study and do not imply any biological activity. One interesting example of a cysteine rich peptide with high biological impact was published by Ohki *et al.* The scientists have performed the first NMR structure of plant peptide hormone, which is a regulator of stomatal development in plants (**Figure 2.3C**). The 46 aa long peptide consists of three disulfide bonds, which are essential for proper folding. Mutation experiments led to unfolding and loss of biological activity of the peptide^[91].

Another common way to stabilize the small proteins is coordinating metal ions. Zinc finger (ZF) proteins contain in particular zinc ion and are the most abundant proteins in eukaryotic genome. The structural fold occurring in nature is very diverse and can be generally divided into eight classes^[92]. According to structural diversity, zinc fingers are involved in a wide range of function mediating interactions with DNA (DNA recognition), RNA (RNA packaging), proteins (protein folding, apoptosis) and membrane (lipid binding)^[93]. The stabilizing effect of the metal ion is exemplarily shown on putative ZFs of the transcription factor Miz-1 (Myc-interacting zinc finger protein 1), which controls the expression of cell cycle regulator genes. The ZF adopts the most common $\beta\beta\alpha$ fold, which includes the coordination of a zinc ion by two cysteines and two histidine (Cys₂His₂) (**Figure 2.3D**). Folding of ZP is zinc dependent and the removal of the metal ion unfolds the protein completely^[94].

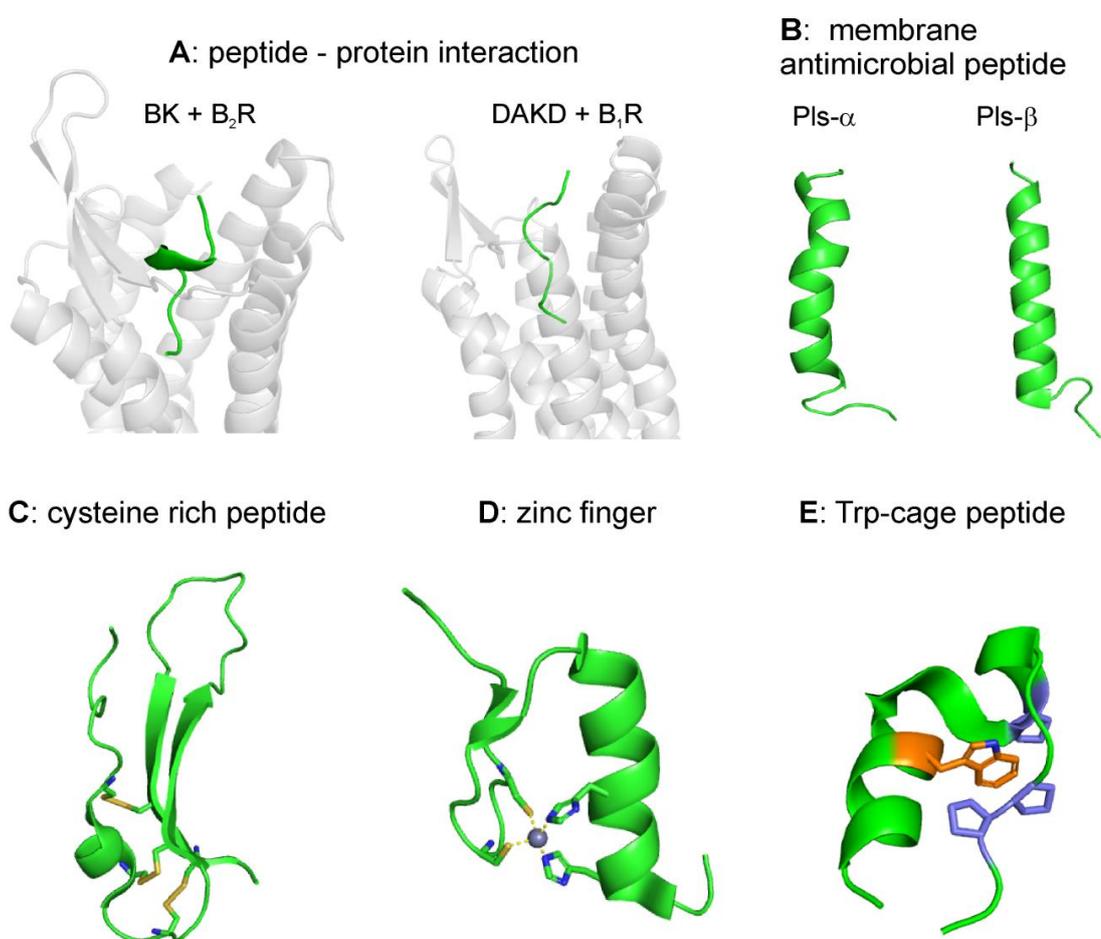


Figure 2.3. Structural fold diversity of small proteins. **A** – Interaction with large biological target. **Left** – disorder-to-order transition. BK folds upon interaction with B₂R (pdb code: 6f3v). **Right** – disorder-to-disorder transition. DAKD does not fold upon interaction with B₁R (pdb code: 6f3y)^[86]. **B** – Membrane AMPs. **Left** - Pls-α (pdb code: 6gn7). **Right** - Pls-β (pdb code: 6go0)^[88]. **C** – Cysteine-rich plant peptide hormone (pdb code: 2liy). Sulfide bonds are shown as sticks and highlighted in orange^[91]. **D** – Zinc finger protein from the transcription factor Miz-1 (pdb code: 2lvr)^[94]. Zinc atom is shown as gray sphere; cysteines and histidine, which coordinate metal ion are shown as sticks. **E** – Tryptophan caged truncated mutant of peptide exendin-4 (EX4) (pdb code: 2ll5)^[95]. Tryptophan and prolines are shown as sticks, tryptophan is colored orange and prolines are highlighted in blue. The figure was generated by PyMol.

Besides the stabilization linkage via a disulfide bond formation and metal ion incorporation, the tryptophan-cage fold was reported as well. The first smallest peptide (20 aa) was designed by truncation and mutation of an originally 39 aa long peptide exendin-4 (EX4) and appeared to be more folded in water solution than its predecessor. This was achieved by using the hydrophobic effect of the tryptophan side chain, which was enclosed between two proline rings^[95] (**Figure 2.3E**). Tryptophan-cage small proteins fold already with 18 to 20 residues in the sequence, thus being the smallest peptides adopting a stable secondary structure.

1.5. Rapid secondary structure determination with NMR

Nuclear magnetic resonance (NMR) spectroscopy is the method of choice in the structural study of small proteins. Their small size finally becomes an advantage in determining the structure by solution state NMR spectroscopy. This powerful approach provides rapid readout concerning the conformational state adopted by peptides and small proteins, and can precisely distinguish between folded, partially folded or molten globule state as well as unstructured (random coil) states ^[96] (**Figure 2.4**). Furthermore, NMR spectroscopy is potentially powerful for elucidating the structure and dynamics of proteins in isolated form and within complexes with their biological targets ^[99,100], which becomes especially important in context of peptides and small proteins.

1.6. SPP 2002 “Small Proteins in Prokaryotes, an Unexplored World”

Within the framework of new Priority Program with the title of „Small Proteins in Prokaryotes, an Unexplored World” the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) started a big project with a global goal of identification and functional characterization of the prokaryotic proteome. In particular, this program supports the investigation of ribosomally synthesized (RPs) small proteins in prokaryotes. Therefore, such classes of small proteins as non-ribosomally synthesized peptides (NRPs) and peptides originated from truncation of bigger proteins are out of scope of this program. 30 research groups from fields such as microbiology, infection biology, plant physiology, chemistry, biochemistry, genetics, genomics as well as applied bioinformatics are involved in this scientific network and created 26 individual and several interdisciplinary collaboration projects. A variety of projects encompass studies of different organisms such as archaea, proteobacteria, cyanobacteria, *Bacillus subtilis*, phages, human microbiome relates, pathogenesis related and membrane small proteins. Furthermore, the computational and experimental biochemical methods are presented by four bioinformatics groups that use ribosomal profiling and peptidomics to identify coding sORFs. Last but not least, Prof. Schwalbe group provides the scientific community with structural information about small proteins of interest.

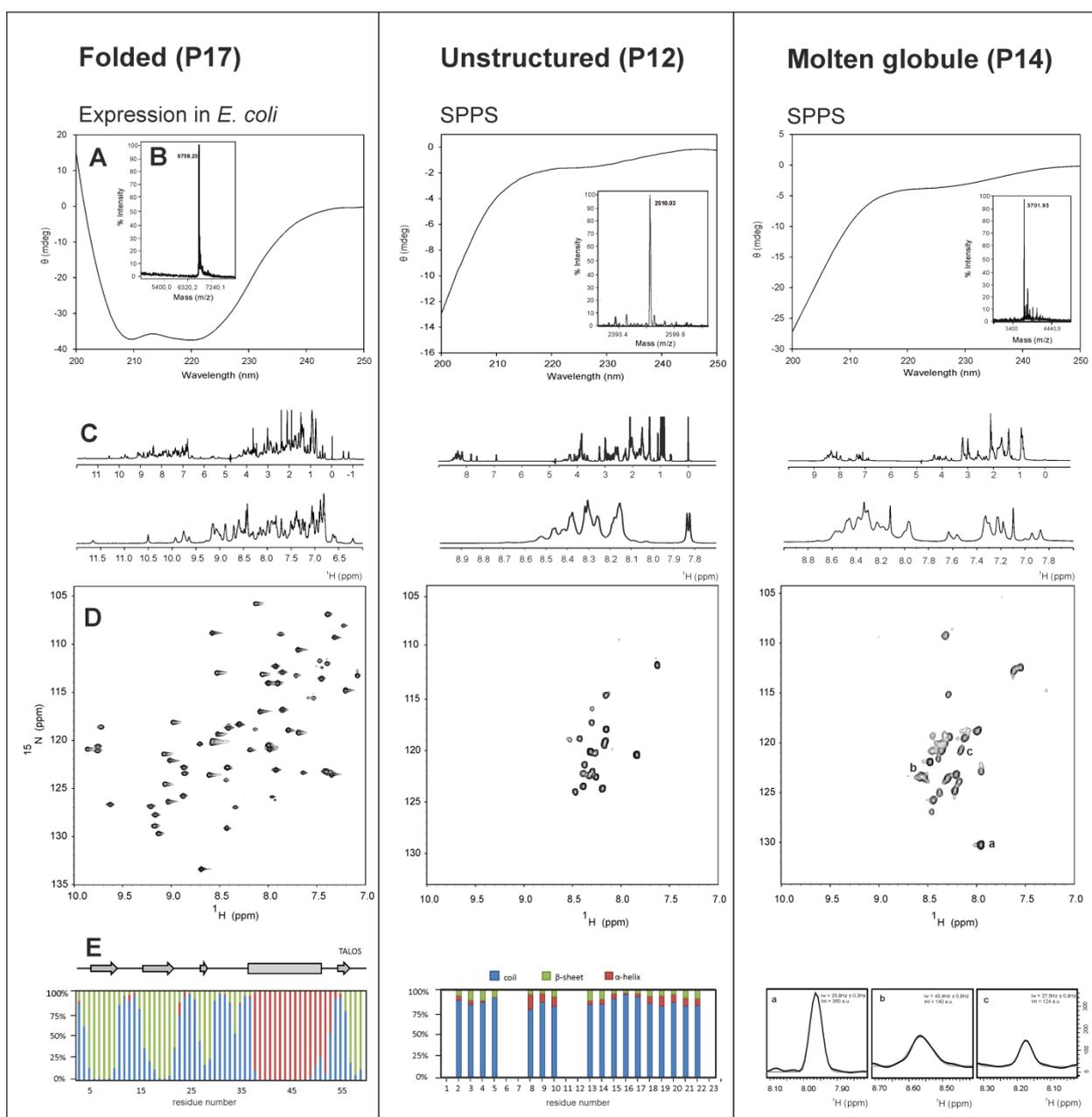


Figure 2.4. Overview of possible folds of small proteins. Left panel: P17 small protein from *Haloferax volcanii*; middle panel: P12 small protein from *Methanosarcina mazei*; right panel: P14 small protein from *Sinorhizobium fredii*. **A** - CD-spectra in phosphate buffer at pH 7. **B** - matrix-assisted laser desorption/ionization (MALDI) mass analysis of the purified small protein. **C** - 1D ^1H -NMR spectra with amide proton region enlarged. **D** - 2D ^1H - ^{15}N HSQC spectra at 600 MHz, 298 K. **E** - left and middle panels: TALOS secondary structure prediction of the residues which are classified as “good”; right panel: expanded regions from 2D ^1H - ^{15}N HSQC spectrum for three signals showing differences in linewidth and intensity of the NMR signals, characteristic for molten globule type conformational behavior. These results agree perfectly with calculations by the Espritz method^[99] in free form propensity of structure P17 (73.3% folded), P12 (0% structured) and P14 (61.3% molten globule).

1.7. Motivation and aim

In the framework of the Priority Program (SPP 2002), the main goal of this part of the doctoral thesis is to conduct a structural analysis of small proteins, which were identified by collaborative groups. This information will provide an additional piece of the puzzle on the way to the global functional elucidation of small proteins. For this, a protocol for fast NMR structure screening for those small proteins has to be established. This workflow should include the sample preparation optimized for the system of this size and the determination of the structural conformation adopted by the system. Since the structural analysis has to be performed on a big set of small proteins, the main requirement for this screening procedure is rapidity and precision. The accuracy of the sample preparation can be achieved by using the SUMO-fusion strategy, which allows producing proteins with exactly the desired sequence, making it perfectly amenable for the small size protein. After peptides and small proteins have been produced and the purity confirmed by mass spectrometry (MALDI) the first structural readout of the candidates is performed by CD and 2D NMR spectroscopy. At this stage, it should be possible to distinguish between structured, partially structured or molten globule state and unstructured (random coil) conformations (**Figure 2.4**). For the small proteins, which adopt well-defined tertiary folds, the NMR solution structure calculation should be performed. This procedure is quite time-consuming and requires a lot of machinery time, as well as manual involvement. To accelerate the analysis, new techniques such as NMR non-uniform sampling in combination with targeted acquisition and subsequent automated resonance assignment should be implemented in the routine laboratory workflow.

2. Material and Methods

2.1. Biochemical Methods

2.1.1. Peptide preparation

2.1.1.1. Solid phase synthesis

The peptides consisting of less than 30 amino acids were synthesized by solid phase peptide synthesis using standard Fmoc (fluorenylmethyloxycarbonyl) chemistry. The C-terminal residue was manually attached to a 2-chlorotrityl chloride resin. The synthesis was performed in a 0.1 mM scale using Fmoc-amino acids purchased from Merck.

2.1.1.2. Reversed-phase HPLC

Peptides were further purified by reversed-phase high-performance liquid chromatography (HPLC). Water-soluble peptides were dissolved in 0.1% TFA water solution and purified using a peptide optimized acetonitrile gradient. The peptides were eluted at specific acetonitrile concentrations.

For the non-water-soluble peptide P20, 70% formic acid (5 mg/ml) solution was used. Purification occurred using a gradient from 30 to 50% acetonitrile with 0.1% TFA. The pure peptide fraction eluted at 38% acetonitrile with 0.1% TFA.

Products were characterized using mass spectrometry (MALDI) and analytical HPLC. The purity of the produced peptides was determined by analytical HPLC and was greater than 96%. The preparative and analytical chromatograms with corresponding MALDI spectra are shown in **Appendix A3, Figure A1**.

2.1.2. Small protein preparation

2.1.2.1. Design of the fusion protein

The expression of small proteins is usually performed with an N-terminal SUMO (small ubiquitin-related modifier) fusion protein. This increases the solubility and expression yield of the target protein and concomitantly decreases its degradation. Typically used fusion motifs (such as GST (glutathione S-transferase), NusA (N-utilization substance A), MBP (maltose-binding protein) and TRX (thioredoxin)) require a classical specific protease cleavage site, leading to the inclusion of non-native amino acids downstream from the cleavage site after protease digestion.

In contrast to other commonly used proteases, which need a specific amino acid sequence for the recognition site, the SUMO protease recognizes the tertiary structure of the SUMO tag and cleaves the tag without any remaining residues ^[102,103]. Since native folding motifs can easily be perturbed by

additional residues, this unique property is highly important for the study of small proteins. The SUMO fusion protein is supplemented with an N-terminal hexa-histidine tag (His₆-tag), which simplifies the tandem Ni-NTA affinity chromatography purification ^[102].

Small proteins, which were not possible to express with SUMO tag, were supplemented only with N-terminal His₆-tag, linked to the protein sequence via a thrombin cleavage site.

For small proteins, which involve Zn-binding motifs, the Ni-NTA purification strategy was changed to an N-terminal GST fusion protein. Ni²⁺ and Zn²⁺ metal ions could compete at the metal binding site of the target protein and potentially destabilize the protein structure ^[103]. Therefore, Zn-binding protein purification was performed using a glutathione column ^[104] followed by size-exclusion chromatography to remove the cleaved tag.

2.1.2.2. Plasmids

The gene sequences were purchased from GenScript (USA) and cloned into a pE-SUMO vector (**Figure 2.5A**) or into pGEX-CS (**Figure 2.5B**) (modified pGEX with an engineered TEV (*tobacco etch virus*) cleavage site). All plasmids carry the gene for ampicillin (*ampR*).

For the zinc binding protein P19 sequence, the pGEX-CS vector with a TEV protease cleavage site was used as expression vector. Ligation was performed using the *NcoI* and *BamHI* restriction sites. The palindromic recognition sequence of 5'-site *NcoI* restriction enzyme (CCATGG) determines the reading frame (ATG for methionine) adding two nucleotides upstream and one nucleotide downstream of the start codon of P19. Therefore, the translation product contains an additional undesired amino acid. Glycine is placed after methionine and alanine occurs before methionine. The TEV cleavage (Glu-Asn-Leu-Tyr-Phe-Gln-(Gly)...) occurs between the glutamine and the glycine resulting in one additional amino acid (glycine) added on the N-terminus of the target protein. Therefore, the translation product contains additional undesired amino acids.

The vector (pET28a) containing small protein P8 sequence was provided by the collaborative group of Schmitz-Streit. The plasmid was implemented with kanamycin gene resistance (*Kn*) and included thrombin (Thr) cleavage recognition site (Leu-Val-Pro-Arg-Gly-Ser). The cleavage occurs between arginine and glycine leading to two additional amino acids (glycine and serine) added on the N-terminus of the protein sequence (**Figure 2.5C**).

Overview of the plasmids used in this study is summarized in **Table 1**.

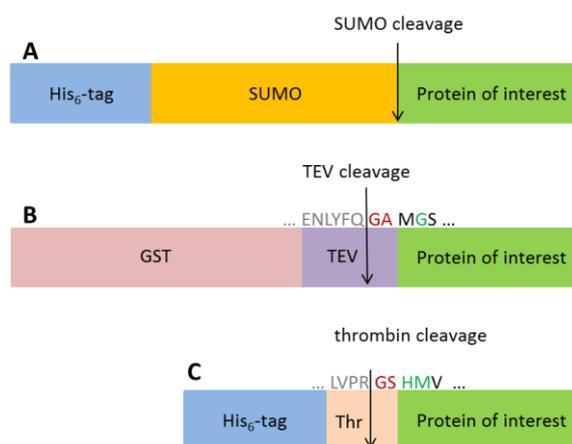


Figure 2.5. General protein design. **A** – Proteins expressed with an N-terminal SUMO fusion protein in combination with a polyhistidine tag. **B** – Zinc binding protein P19 expressed with an N-terminal GST tag. **C** - Small protein P8 expressed with N-terminal His₆-tag. The protease cleavage recognition site and its cleavage position are shown above. Additional amino acids resulting from the protease cleavage are indicated in red, while the amino acid from *NcoI* codon translation is highlighted in green.

Table 1. Overview of the plasmids used in this study.

	Vector	Antibiotic resistance	Fusion	Cleavage recognition site	Protease
All small proteins	pE-SUMO	<i>ampR</i>	His ₆ -tag - SUMO	tertiary structure of the SUMO tag	SUMO
P19	pGEX-CS	<i>ampR</i>	GST	TEV	TEV
P8	pET28a	<i>Kn</i>	His ₆ -tag	Thrombin	Thrombin

Plasmid amplification and transformation were performed using T7-Express High Efficiency *E. coli* cells from NEB® (*Ipswich, Massachusetts, USA*) following the instructions of the high efficiency transformation protocol of the manufacturer. The expression plasmids were used for protein expression.

Site-directed mutagenesis was performed for I58P and for the deletion mutant of small protein P17 (**chapter II 3.7.3.2**) using the QuikChange Lightning site-directed mutagenesis kit (*Agilent Technologies, USA*) and the following primers:

P17 I58P fwd 5'- CGGATGCGGATGTTCTCTGAAGCGTGA TCTAG -3'
 P17 I58P rev 5'- GCCTACGCCTACAAGGACTTCGCACT AGATC -3'
 P17 Δ4 fwd 5'- GCGCCGGATG CGGATTGAAT TGAAGCGTGA TCTAG -3'
 P17 Δ4 rev 5'- CGCGGCCTAC GCCTAACTTA ACTTCGCACT AGATC -3'

Successful mutagenesis was confirmed by DNA sequencing performed by Eurofins Genomics (Germany).

2.1.2.3. Protein expression and purification

The plasmids containing the protein sequences were transformed into T7-Express *E. coli* cells and expressed as SUMO- or GST-fusion proteins. The plasmids used in this thesis are listed in **Appendix A1**.

2.1.2.3.1. Test expression

Before large scale expression of the desired proteins, a test expression was performed. One clone of *E. coli* transformed cells was inoculated to LB medium supplemented with 100 µg/ml ampicillin and 2% glucose (w/v). After the optical density (OD₆₀₀) reached a value between 0.6 and 0.8, the culture was split into two batches. One was induced with IPTG (final concentration 1 mM) and another one was kept uninduced as a negative control. Cultures were incubated at 37°C with continuous mixing at 160 rpm for 4 hours and subsequently centrifuged at 14000 g for 1 minute. Cells were lysed by resuspending the pellets in SDS-loading buffer and heating them up to 95°C for 5 minutes. Expression was analyzed by SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) analysis. All used buffers are listed in **Appendix A2**.

2.1.2.3.2. SDS-PAGE

The SDS-PAGE was used for the separation of proteins. Reducing loading buffer (BME) was applied on 4-12% Bis-Tris NuPAGE gradient gels (novex by *Life Technologies*). The electrophoretic separation of proteins was performed in SDS running buffer at a constant voltage of 200 V for 40 minutes. Proteins were stained with Coomassie Brilliant Blue solution.

2.1.2.3.3. Protein expression

For NMR spectroscopy, ¹⁵N- and ¹⁵N/¹³C-labeling schemes were performed using M9 minimal medium containing ¹⁵N-labeled ammonium chloride (1 g/l) and ¹³C-labeled glucose (2 g/l) (*Cambridge Isotope Laboratories, Cambridge, MA, USA*). *E. coli* transformed cells were inoculated to 5 ml LB medium supplemented with 100 µg/ml ampicillin and 2% glucose (w/v) and incubated at 37°C with continuous mixing at 160 rpm for 4 hours. The LB culture was transferred into 200 ml M9 medium containing isotope labeling precursors and incubated at 3°C, 160 rpm for 12 hours. The preparative culture of 2 l M9 medium was inoculated with the pre-culture to a final OD₆₀₀ value of 0.1, followed by the incubation at 37°C, 120 rpm. After the OD₆₀₀ reached a value between 0.6 and 0.8, the culture was induced with 1 mM IPTG. The final preparative expression was performed at 24°C, 120 rpm for 12 hours. Cell growth was monitored by measuring the optical density every 30 minutes. After the OD₆₀₀ increase ended, cells were harvested at 5000 rpm for 15 minutes using a Beckmann centrifuge with a JLA 8.1000 rotor. The pellet was resuspended in buffer A (**Appendix A2**) supplemented with 10% glycerol (w/v) and one protease-inhibitor tablet (*cOmplete™, Roche, Germany*) per liter. Cells were stored at -80°C.

2.1.2.3.4. Protein purification

Cells were lysed mechanically using Microfluidics M-110P at 15000 PSI under continuous ice cooling followed by centrifugation at 4°C and 16000 rpm for 45 minutes using a Beckmann centrifuge with JA 20 rotor. The supernatant was further purified using immobilized metal ion affinity chromatography (IMAC) according to the fusion design.

SUMO-fusion protein purification

The purification of the SUMO-fusion proteins was accomplished by tandem Ni-NTA affinity chromatography. After centrifugation, the supernatant was loaded onto a 5ml HisTrap HP column (*GE Healthcare, USA*) and connected to the FPLC (*Äktapurifier™, GE Healthcare, USA*). Equilibration and loading of the protein was performed using buffer A, while the elution of the bound protein was achieved with the linear gradient of imidazole containing buffer B (500 mM imidazole) (**Appendix A2**). The high affinity of imidazole to the metal ions allows displacing the His-tagged protein from the column. Typically, a SUMO-fusion protein elutes at 33% of buffer B (165 mM imidazole). Fractions containing the fusion protein were analyzed by SDS-PAGE and confirmed by MALDI analysis. Cleavage was performed with SUMO protease during dialysis against imidazole in buffer C (**Appendix A2**) at 4°C for 8 hours. After cleavage, reverse Ni-NTA affinity chromatography was performed to separate the desired small protein from the remaining fusion tag. The affinity fusion tag (His₆-tag), which was attached to the N- terminus of the SUMO protein, was bound to the metal ions immobilized on the column resin, while the cleaved protein of interest was collected in the flow-through. The identification and the purity of produced protein was conducted with SDS-PAGE analysis and then confirmed with mass spectrometry (MALDI). Typically, the achieved purity of the protein was higher than 95%. Otherwise, additional size-exclusion chromatography (SEC) was applied. In this case, the protein was loaded onto a 320 ml HiLoad 26/600 Superdex 75 pg gel filtration column (*GE Healthcare, USA*) with an NMR running buffer (**Appendix A2**). Subsequently, the fractions were analyzed by SDS-PAGE and confirmed by MALDI. Finally, the protein was concentrated using centrifugal concentrator devices (*VivaSpin20, Sartorius, Germany, MWCO 3000*). Chromatograms with corresponding gels and MALDI spectra are shown in **Appendix A3, Figure A2**.

GST-fusion protein purification

The Zn²⁺ binding protein P19 with N-terminal GST fusion-tag was purified using glutathione-sepharose beads ^[104]. After centrifugation the supernatant was loaded onto a 5 ml GSTrap HP column (*GE Healthcare, USA*) and connected to the FPLC (*Äktapurifier™, GE Healthcare, USA*). Equilibration of the column and loading of the protein were performed by using buffer A, while the elution of the bound protein occurred with buffer B containing 10 mM glutathione without gradient application. Fractions

containing fusion protein were analyzed by SDS-PAGE and confirmed by MALDI analysis. Cleavage was performed with the TEV protease at 4°C for 8 hours. The cleaved fusion tag was removed by size-exclusion chromatography. The mixture of cleaved protein of interest and fusion protein was loaded onto a 320 ml HiLoad 26/600 Superdex 75 µg gel filtration column (*GE Healthcare, USA*) with NMR running buffer. Separation of the proteins was based on the difference in their size, which is correlated to the molecular weight. Subsequently the fractions were analyzed by SDS-PAGE and confirmed by MALDI analysis. Finally, the protein was concentrated using centrifugal concentrator devices (*VivaSpin20, Sartorius, Germany, MWCO 3000*). Chromatograms with corresponding gels and MALDI spectra are shown in **Appendix A3, Figure A3**.

2.2. Biophysical Methods

2.2.1. UV/Vis spectrometry

Protein concentration was determined using ultraviolet-visible spectrophotometry (UV/Vis) at 280 nm. The Lambert-Beer law describes the correlation between absorption **A** at a specific wavelength and extinction coefficient ϵ , the concentration **c** and the sample path length **d**.

$$A = \epsilon \cdot c \cdot d \quad (5)$$

The extinction coefficient for each protein was determined using the online software ProtParam provided on the ExPASy Bioinformatics Resource Portal^[105] and can be found in **Appendix A4, Table A1**. Concentrations of proteins containing aromatic residues (tyrosine and tryptophan) were measured using NanoDrop photometers (*NanoDrop One from Thermo Scientific, Germany*). The volume of the measured sample was 2 µl and the path length was 1 mm.

2.2.2. Circular dichroism spectroscopy (CD)

Circular dichroism spectroscopy (CD) is a powerful method in biochemistry, which allows rapid analysis of the secondary structure and folding properties of proteins in solution. The polarized light can be divided into two circular polarized parts rotating clockwise and counter-clockwise. By passing through the optical active sample these light components absorb differently. Measurement of the absorbance difference of these compounds as a function of wavelength results in a CD spectrum. Since different types of secondary structure elements in proteins absorb a characteristic wavelength range, it is possible to determine a secondary structure composition of proteins. CD profiles of proteins consisting of α -helix and/or β -sheet rich regions as well as unstructured elements are exemplified in **Figure 2.6**.

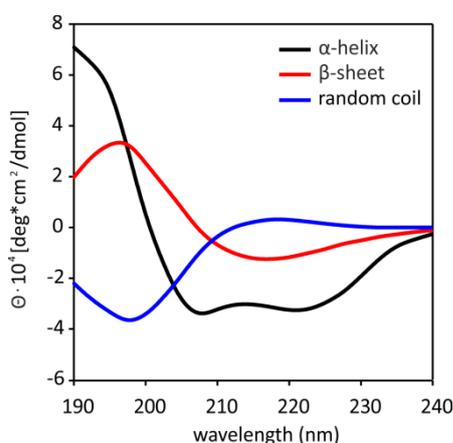


Figure 2.6. Graphical representation of CD profiles of proteins. Spectra were adapted from Reed and Reed, 1997^[106].

All measurements were carried out with a Jasco J-810 CD spectropolarimeter with constant influx of nitrogen gas to prevent ozone formation. The cuvette path length was 10 mm, the scanning speed was 50 nm/min, bandwidth 1 nm and a response time of 1 s. Every spectrum was background corrected with the corresponding buffer mixture. In each experiment, measurements were done at 20°C and 3 spectra were summed and averaged. The sample concentration of small proteins varied between 50 and 200 μM. All spectra were later smoothed and plotted using SigmaPlot (version 12.5). The units of the raw data (millidegree (mdeg)) were recalculated to molar ellipticity, which is given in deg·cm⁻¹·dmol⁻¹.

$$[\theta] = \frac{\theta}{10 \cdot c \cdot l} \quad (6)$$

Equation 6 describes the correlation between the ellipticity θ in mdeg, the sample concentration c in mol/l and the sample path length l in cm.

2.3. Nuclear magnetic resonance spectroscopy

2.3.1. Fast secondary structure screening of small proteins

The first secondary structure screening of small proteins includes 1D and 2D NMR experiments. For peptides that were synthesized via SPPS (solid phase peptide synthesis), 1D ¹H (zgesgp) and 2D natural abundance ¹H¹⁵N HSQC (hsqcetf3gpsi or hsqcftp3gpplhwg) experiments were recorded. For ¹⁵N labeled small proteins expressed in *E. coli* ¹⁵N decoupling was included in the 1D ¹H-NMR pulse sequence using a GARP4 decoupling sequence^[107]. NMR screening analysis was carried out on Bruker spectrometers ranging from 600 to 950 MHz and equipped with z-axis gradient ¹H{¹³C, ¹⁵N} triple resonance cryogenic probes at 298 K (**Appendix A5**).

The NMR buffer for the fast structure screening consists of 25-50 mM potassium phosphate (pH 7-8), 200-300 mM NaCl and 3 mM DTT or 5 mM 2-mercaptoethanol as a reducing agent for small proteins containing cysteines. All samples include 5% D₂O and 0.5 mM DSS as an internal reference for ¹H chemical shifts^[108]. ¹³C and ¹⁵N chemical shifts were calculated according to the ¹H frequency.

All NMR spectra were processed by using TopSpin version 3.2 (Bruker Biospin) and analyzed with SPARKY version 3.114^[109].

2.3.2. Structural analysis of folded small proteins

2.3.2.1. Assignment strategy

For three peptides (P4, P12 and P20) the complete resonance assignments were performed using natural abundance 2D NMR experiments: ¹H¹⁵N HSQC, ¹H¹³C HSQC, ¹H¹H COSY, ¹H¹H TOCSY and ¹H¹H NOESY.

Sequential backbone resonance assignments of P17 and P19 small protein were performed with BEST-TROSY-based HNCACB, HN(CO)CACB, HNCO and HNHA 3D NMR experiments. Side chain assignment included (H)CC(CO)NH, H(C)CH-TOCSY and ¹H¹H¹⁵N TOCSY-HSQC 3D NMR experiments. A set of NMR experiments for P19 was complemented with additional HNN and HNNH 3D NMR experiment. 2D ¹H¹H NOESY, 3D ¹H¹H¹⁵N NOESY-HSQC and 3D ¹H¹H¹³C NOESY-HSQC spectra were used for the NOE-based distance restraints.

Sample composition, parameters of the NMR experiments and spectrometer, which were used for the structural analysis of P17 and P19 small proteins are listed in **Appendix A5 and A6**.

2.3.2.2. Non-uniform sampling (NUS) and Targeted Acquisition (TA)

The set of spectra, recorded with NUS and targeted acquisition technique, include the following 3D NMR experiments: HNCO, HN(CO)CA, HNCA, HN(CO)CACB, HNCACB, HN(CA)CO. For P17 and P19 the BEST-TROSY-based versions of the NMR pulse sequences were used. The completed set of NMR experiments for the backbone assignments of P17 and P19 proteins were recorded with 6% and 17% NUS, respectively. This corresponds to the total measurement time of 4.5 and 46 hours. Spectra were processed in real-time using the TA script from V. Orekhov group (Goteborg, Sweden).

2.3.2.3. FLYA automated assignment

Automated chemical shift assignment was performed with automated FLYA assignment^[43] using the manually performed resonance assignment as a reference for the quality validation of the method.

2.3.2.4. Temperature series

The temperature series of P17 small protein was recorded from 278 to 318 K with 5 K increments using 2D $^1\text{H}^{15}\text{N}$ BEST-TROSY spectra measured on a 600 MHz spectrometer. The amide protons temperature coefficient of the P17 small protein was calculated from a linear fit of the chemical shift perturbation (CSP) as a function of the temperature^[110]. The temperature series of P19 zinc binding protein includes spectra recorded from 278 to 308 K with 10 K increments and was recorded on a 800 MHz spectrometer.

2.3.2.5. Dynamical studies of P17 small proteins

2.3.2.5.1. Hydrogen-deuterium exchange NMR experiments

Hydrogen-deuterium exchange NMR experiments were acquired at a 600 MHz spectrometer at 288 K. Therefore, the protein sample was lyophilized and dissolved in D_2O . A series of 2D $^1\text{H}^{15}\text{N}$ BEST-TROSY spectra were recorded immediately after dissolving the protein in D_2O . The first ten experiments were recorded with an experimental time of 3 minutes, respectively. Then, the experimental time was increased to 12 minutes and recorded 10 times. The next ten experiments were recorded with a duration time of 45 minutes. The final eight experiments were acquired with 90 minutes experimental time. The total experimental time was 22 hours.

2.3.2.5.2. Heteronuclear relaxation experiments

Heteronuclear ^{15}N relaxation experiments^[111,112] were performed for P17 wild type and I58P P17 mutant proteins at 298 K on a 600 MHz spectrometer. The wild type concentrations were 5 mM and 0.6 mM, while I58P mutant concentration was 0.6 mM. The T_1 longitudinal ^{15}N relaxation time was calculated based on a series of spectra with the different relaxation delays: 20, 50, 100, 150, 200, 250, 400, 600, 750, 1000, 1300, 1600, 2000 ms. The T_2 transverse ^{15}N relaxation time was recorded with the following delays: 16.96, 33.92, 67.84, 101.76, 135.68, 169.60, 203.52, 271.36 ms. The $\{^1\text{H}\}^{-15}\text{N}$ hetNOEs were calculated from the signal intensity ratio ($I_{\text{on}}/I_{\text{off}}$) obtained from the spectra recorded with and without saturation of amide protons. Recovery (d_1) and saturation delay were set to 5 seconds. The S^2 order parameter was calculated using the TENSOR2^[111] software and the average of the first three lowest energy NMR structure model calculated with CYANA 3.97^[44-46].

2.3.2.6. Chemical shift perturbations of I58P P17 mutant

Chemical shift perturbations of the I58P mutant compared to the P17 wild type were calculated using the following formula:

$$\Delta\delta = \sqrt{(\Delta\delta_H)^2 + (\alpha \cdot \Delta\delta_N)^2} \quad (7)$$

Equation 7 describes the combination of ^1H - and ^{15}N -CSPs of the backbone amides by correcting for the gyromagnetic ratio ($\alpha = 0.1$).

2.3.2.7. Structure calculation

Structure calculation of the small protein P17 (HVO_2922) was performed with CYANA 3.97^[44-46] and was followed by the refinement with ARIA/CNS^[113]. The input contained the manually assigned resonance list in combination with the unassigned NOESY peak list. NOESY assignment was performed using the automated NOE crosspeak assignment routine. Therefore, two 3D NOESY NMR experiments were used: i) the aliphatic 3D $^1\text{H}^1\text{H}^{13}\text{C}$ -NOESY-HSQC with a mixing time of 120 ms and ii) the amide 3D $^1\text{H}^1\text{H}^{15}\text{N}$ -NOESY-HSQC with a mixing time of 120 ms. Peak picking of the spectra was performed by using the restricted peak picking function in the Sparky 3.114 program^[109]. The structure calculation was complemented with a manually peak-picked 2D $^1\text{H}^1\text{H}$ -NOESY with a mixing time of 100 ms spectrum.

The chemical shift tolerances were set to 0.015 ppm for the bound protons, to 0.020 ppm for all other protons and to 0.20 ppm for the heavy atoms. The final structure calculation included the NOE data, hydrogen bond distances and dihedral angles, derived from the TALOS-N^[114] predictions. The structure calculation was performed using a standard protocol with 100 initial steps, 15000 refinement steps per iteration and 20 final structures. The last iteration included 200 structures, which were used for a refinement in explicit water with CNS 1.1^[115] using the ARIA 1.2 program^[116]. The final refinement also contains $^3J_{\text{HNHA}}$ coupling constants and relaxation data with T_1/T_2 values as well as diffusion anisotropy, anisotropy of 1.20, rhombicity of 0.60 and rotation correlation time τ_c of 8.63 ns.

3. Results and discussions

3.1. Screening overview

In total 20 different peptides and small proteins ranging from 14 to 71 amino acids were screened for the presence of secondary structure elements during this study. The amino acids sequences were provided by six academic collaborators, focused on the identification and functional characterization of small proteins in prokaryotic systems. An overview of the peptides and small proteins screened in this study, including, organism of origin, collaborating research group and the type of synthesis used for their production as well as the results of the structural analysis are listed in **Table 2**. Amino acid sequences of all investigated proteins and the additional information can be found in **Appendix A4**. Both, peptides and small proteins investigated during this doctoral thesis derive from different bacterial and archaeal organisms. Nevertheless, the screening method used in this study is not limited to the prokaryotic systems and can prospectively be applied to the eukaryotic small proteins.

Table 2. Overview of the results of the structural analysis of small proteins screened in this study.

Internal Name	N° amino acids	Organism	Research group	SPPS/ Expression	Structural analysis
P1	14	<i>Bradyrhizobium japonicum</i>	Elena Evguenieva-Hackenberg	SPPS	unstructured ^[117]
P2	23	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	SPPS	unstructured
P3	18	<i>Dinoroseobacter shibae</i>	Elena Evguenieva-Hackenberg	SPPS	unstructured ^[117]
P4	28	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	SPPS	unstructured
P5	27	<i>Sinorhizobium meliloti</i>	Elena Evguenieva-Hackenberg	SPPS	unstructured
P6	29	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	SPPS	unstructured
P7	53	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	Expression	degradation
P8	61	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	Expression	folded
P9	70	<i>Rhodobacter sphaeroides</i>	Gabriele Klug	Expression	degradation
P10	71	<i>Rhodobacter sphaeroides</i>	Gabriele Klug	Expression	degradation
P11	38	<i>Haloferax volcanii</i>	Jörg Soppa	Expression	molten globule
P12	23	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	SPPS	unstructured
P13	18	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	SPPS	n.a. *
P14	31	<i>Sinorhizobium fredii</i> NGR234	Wolfgang Streit	SPPS	molten globule ^[118]
P15	51	<i>Sinorhizobium fredii</i> NGR234	Wolfgang Streit	Expression	degradation
P16	43	<i>Sinorhizobium fredii</i> NGR234	Wolfgang Streit	Expression	molten globule ^[118]
P17	60	<i>Haloferax volcanii</i>	Anita Marchfelder	Expression	structured ^[134]
P18	61	<i>Haloferax volcanii</i>	Anita Marchfelder	Expression	molten globule
P19	59	<i>Haloferax volcanii</i>	Jörg Soppa	Expression	structured
P20	17	<i>Sinorhizobium meliloti</i>	Elena Evguenieva-Hackenberg	SPPS	unstructured

* - n.a. due to hydrophobicity

All peptides (14-31 aa) were synthesized by solid phase peptide synthesis (SPPS), while the small proteins (38-71 aa) were produced using *E. coli* expression system. The structural characterization of peptide P13 was hampered due to its extended hydrophobicity. Four small proteins (P7, P9, P10 and P15) were successfully expressed, however, degraded during the purification. In total, for 15 out of 20 peptides/proteins fast secondary structure screening analysis could be conducted. Eight peptides were found to be unstructured (P1-6, P12 and P20), four small proteins adopt a molten globule conformation (P11, P14, P16, and P18) and three proteins (P8, P17 and P19) are structured. For these three proteins structural and dynamical investigations were performed. For small protein P17 structure calculation was conducted.

3.2. Sequence to structure analysis

3.2.1. Fast online structure prediction tools

The first insight into the potential secondary structure adopted by the protein can be gained by using the sequence to structure analysis. For this purpose a number of available sequence-based web-tools such as pep-FOLD^[121,122], PEPstr^[121,122], PepLook^[121,122], I-Tasser^[123], Rosetta^[124], SWISS model^[127-131] etc. can be used.

In this thesis, pep-FOLD analysis^[121,122] was applied on peptides and SWISS model structure prediction^[127-131] was used for small proteins. The pep-FOLD approach utilizes the concept of structural alphabet, which means that the whole protein structure can be described by the ensemble of specific consecutive residues called “letters”. The peptide is built of a combination of such sets of four amino acids, which encode the structural information of the peptide^[130]. Since the first generation of pep-FOLD algorithm was limited to the length between 9 and 36 amino acids^[22], the second version, pep-FOLD2, has been extended to 52 residues with a minimal residue number of 25 amino acids^[119] and the third version, pep-FOLD3, covers small proteins in the range of 5 to 50 amino acids^[131]. This makes the pep-FOLD tool suitable for peptides but not for larger proteins.

The pep-FOLD3 sequence-based secondary structure prediction showed folded states with abundant α -helical regions for all screened peptides (**Figure 2.7**). In contrast, experimental data from the CD and NMR measurements reported a completely unstructured state for the same set of peptides. However, the lack of a well-defined three-dimensional structure does not necessarily implement the absence of the function. Moreover, initially unstructured peptides can fold in a stable structured conformation, which can be induced by the interaction with their biological partners. Therefore, the identification of such potential biological partners for unstructured proteins and detailed interaction studies are very important, however, beyond the scope of this study and can be subject to further investigations.

All screened peptides (14-31 aa) were shown to be predominantly unstructured, while the small proteins (38-61 aa) rather tend to form a structured or partially structured conformation (**Figure 2.8**). However, the small size of the peptides seems to be an important restricting factor during the formation of possible secondary structure elements.

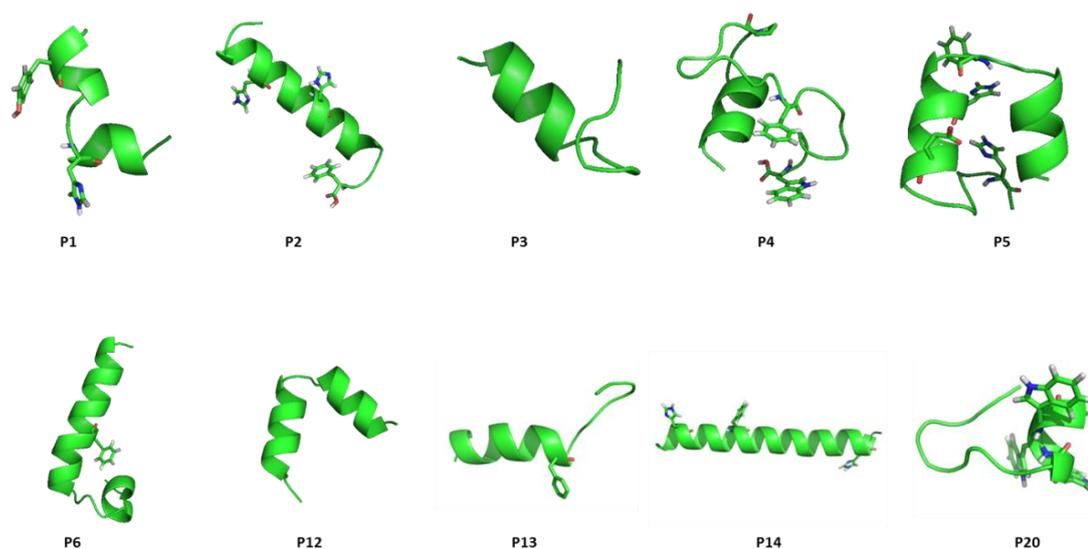


Figure 2.7. Sequence-based pep-FOLD3 secondary structure prediction of peptides synthesized with SPPS. Aromatic amino acids are highlighted as sticks. All peptides were determined as unstructured using CD and NMR spectroscopy. The figure was generated by PyMOL.

In contrast to secondary structure prediction tools available for peptides, the SWISS model approach can be applied to proteins containing more than 30 amino acids in the sequence. This prediction method is based on the modeling of the structure using a homologous structure template with sequence similarity to the protein of interest ^[23–27]. The prediction quality is strongly dependent on the sequence identity between target protein and the existing pdb structure.

The SWISS model structures of the screened small proteins (P16-P19) are shown in **Figure 2.8**. Experiments show that P17 and P19 are folded; P16 and P18 adopt molten globule states. The homology model of P16, which is based on the part of the homo-tetramer complex, shows low sequence identity (13.5%). The template used for the SWISS model structure of P18 protein is a homo-dimer complex. Contrary to the prediction, which indicates folding of P18, this protein adopts in isolation a mixture of different conformational states, suggesting lack of additional interaction partners, which would be associated with the right protein folding.

The SWISS homology model for P17 with the high sequence identity (39%) reports a symmetrical dimer formation. The predicted geometry and fold are in good agreement with the NMR solution structure of P17 ^[132]. Swiss model structure of the zinc binding protein P19 was defined using a catalytic part of DNA

polymerase subunit as a template. It shows 26% of sequence identity. Almost all templates used for the SWISS model structures of small proteins, excluding template for P17, are parts of big complexes, showing the important role of intra- and intermolecular interactions on the folding of the biological systems.

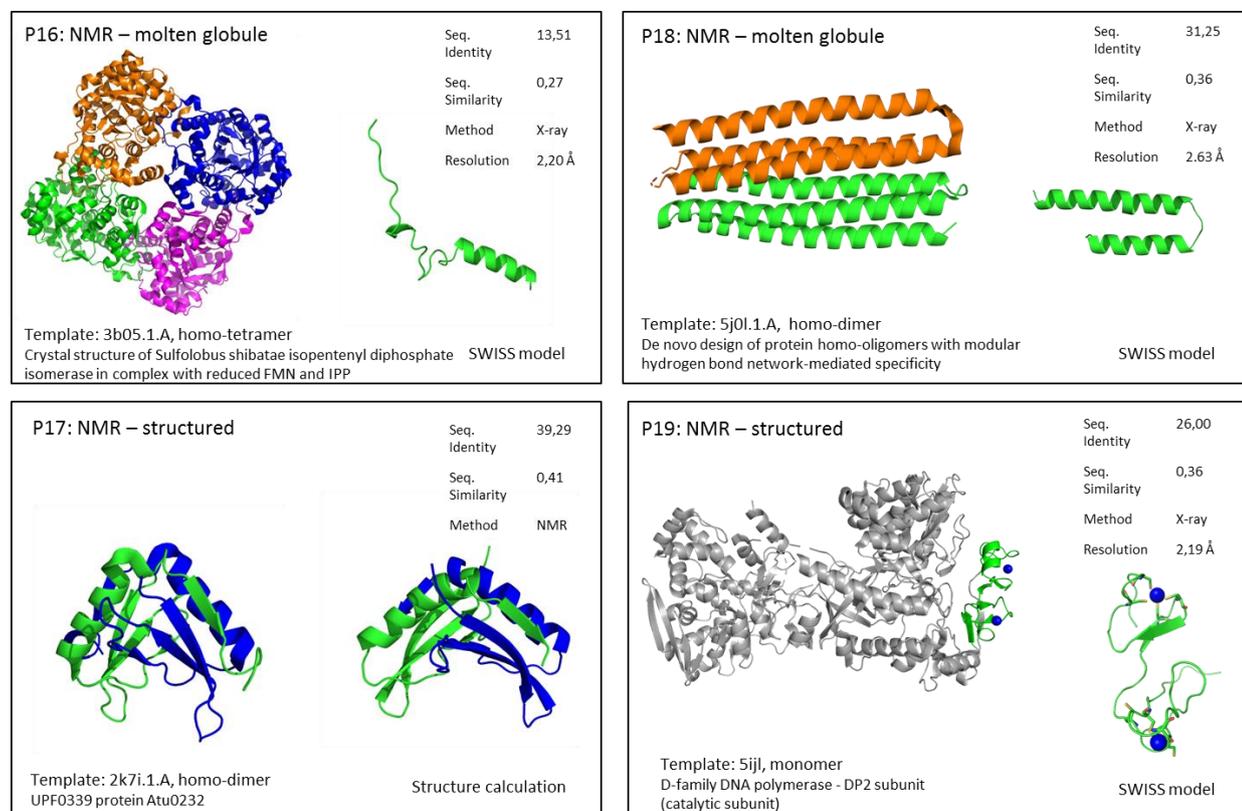


Figure 2.8. Sequence-based SWISS model secondary structure prediction of small proteins. Templates used for homology modeling, resulting SWISS models and the calculated NMR solution structure are shown. The figure was generated by PyMOL.

In addition to the SWISS model structure, the novel sequence-based computational method, called s2D, was applied to the screened small proteins. This method predicts the secondary-structure population for every amino acid in the sequence using the NMR chemical shift data base ^[133]. Originally, this method was developed for IDPs with a minimal length of 25 residues. Seven of twelve screened small proteins, which fulfill this restriction, were predicted to have more than 50% of the residues in the unstructured conformation (**Table 3**).

Overall, the comparison of the s2D prediction with the experimental NMR secondary structure analysis provides a good way for the validation of the s2D method. Thus, three (P11, P14, P16) out of four small proteins, which were found to exist in molten globule state, were predicted to be predominantly unstructured. Only one small protein, P18, which belongs to the *Haloferox volcanii* family, was predicted to adopt a folded conformation, as confirmed by further NMR-based analysis (**chapter II 3.7.2**). Unstructured peptides P4 and P6 with 28 and 29 aa, respectively, are very close to the method size limit.

Two small proteins (P19 and P17), which are extensively structurally characterized in **chapter II 3.6.2 and 3.7.3**, respectively, were also subjected to the prediction by s2D method. P19 was predicted to be unstructured to 95%. This dramatic difference to the experimental data can be explained through the nature of the protein. P19 is a Zn²⁺-binding protein with two zinc-binding motifs in the sequence. The presence of metal ions, which is essential for the proper folding, is not included in s2D prediction, resulting in the falsely predicted unfolded conformation. Similar divergence was observed for P17, where the experimental data and the s2D prediction show different folding properties. S2D predicts partial folding for P17, while the NMR data show a fully folded symmetrical dimer conformation. The intramolecular dimer interaction is not considered in the prediction analysis. Such inconsistency observed between the experimental data and the prediction makes folding comparison at least for those two proteins impossible. Four additional small proteins (P8, P9, P10 and P15) were predicted to consist of combined structured and unstructured regions. In general, the s2D method provides an important additional information source of sequence resolved secondary structure population within the protein. Nevertheless, the method needs experimental validation and should be used in combination with other complementary tools.

Table 3. S2D NMR chemical shift based statistical population prediction of small proteins screened in this study. Small proteins are combined in classes with respect to experimental secondary structure screening analysis. The preferred predicted secondary structural elements are highlighted in the table.

NMR analysis	ID	aa	Helix	Coil	Extended
folded	P19	59	0	0.95	0.05
	P17	60	0.23	0.48	0.28
	P8	61	0.13	0.66	0.21
molten globule	P14	31	0.32	0.64	0.03
	P11	38	0.29	0.71	0
	P16	43	0	0.86	0.14
	P18	61	0.71	0.29	0
unstructured	P4	28	0	1	0
	P6	29	0.62	0.38	0
degradation	P15	51	0.31	0.61	0.08
	P9	70	0.56	0.44	0
	P10	71	0.66	0.34	0

* - protein degradation under described conditions

3.2.2. Prediction of dynamic transition

Since structure prediction tools described above are strongly limited to the peptides and proteins in the isolated form and these macromolecules are known to be prone for interactions with biological partners, prediction of the potential ability to fold upon complex formation and identification of these regions in the protein sequence would provide an important additional information to the global secondary structure prediction analysis. The FuzPred method can be used to predict the probability to adopt a folded conformation upon binding of a given protein. The probability to fold in the free form was calculated by Espritz NMR tool ^[99] and meta approach ^[134] and was subsequently used as a reference to free form in FuzPred approach. In order to simplify the output of both tools, the possible conformational states are divided into two groups: i) structured and ii) unstructured.

Table 4. Espritz NMR predictions of peptides and small proteins screened in this study. Dynamic transition induced by interactions was computed by the FuzPred method with a reference to Espritz NMR free form. Small proteins are classified according to experimental secondary structure screening analysis. The preferred predicted secondary structural elements are highlighted in the table.

NMR analysis	ID	aa	Free form		Bound form	
			structured	disordered	structured	disordered
folded	P19	59	71.2	28.8	84.7	15.3
	P17	60	73.3	26.7	100	0
	P8	61	100	0	100	0
molten globule	P14	31	61.3	38.8	93.5	6.5
	P11	38	73.7	26.4	100	0
	P16	43	53.5	46.5	67.4	32.6
	P18	61	67.2	32.8	78.7	21.3
unstructured	P1	14	78.6	21.4	100	0
	P20	17	21.4	78.6	78.6	21.4
	P3	18	0	100	61.1	38.9
	P2	23	21.7	78.2	87	13
	P12	23	0	100	52.2	47.8
	P5	27	0	100	63	37
	P4	28	0	100	57.1	42.9
P6	29	79.3	20.7	100	0	
degradation	P15	51	31.4	68.7	82.4	17.6
	P7	53	73.1	26.9	92.3	7.7
	P9	70	81.4	18.6	87.1	12.9
	P10	71	74.6	25.3	83.1	16.9
n.a.	P13	18	88.9	11.1	100	0

* - n.a. due to hydrophobicity

Most of the small proteins analyzed by Espritz NMR method and meta approach, were predicted to be structured in the free form (**Table 4 and Appendix A7, Table A8**). This is in good agreement with the experimental NMR results showing structured and molten globule conformations of small proteins. Peptides with less than 25 amino acids in the sequence were predicted to be mostly unstructured, which is as well consistent with the experimental results. Small deviations between predictions and NMR data can be addressed to the original purpose of this prediction method, which was developed for proteins with IDP regions. Thus, short sequences are difficult to estimate and to validate.

In summary, the prediction of the dynamic transitions, induced by the complex formation with the hypothetical interaction partners, reports an overall increased tendency to fold upon interaction for all small proteins. Since the degraded small proteins follow the same trend, a possible increase of the protein stability upon complex formation can be assumed.

3.3. Cooperation with Elena Evguenieva-Hackenberg group (University of Giessen)

The group of Evguenieva-Hackenberg from Justus Liebig University of Giessen is focused on the annotation and functional investigation of conserved small proteins in alphaproteobacteria. In 2016 Hahn *et al.* published two newly discovered conserved small open reading frame (sORF), which encode a 14 amino acids long peptide, termed rreB, and its homolog, the 18 amino acids long peptide rreR. The rreB (ribosome-regulated expression in Bradyrhizobiaceae) (internal name P1) was found in the bacterial soybean symbiont *Bradyrhizobium japonicum* USDA 110, while rreR (internal name P3) occurs in *Rhodobacter*. These two previously not known peptides were targets of investigation during this work. Peptides P1, P3 and two additional peptides from *Sinorhizobium meliloty* (P5 and P20) were structurally characterized using CD and NMR spectroscopy. Although an additional sequence to structure analysis, using web tools described above, indicated presence of α -helical regions within those peptides, the experimental data showed completely unstructured states for all peptides (**Figure 2.9**).

3.3.1. Structural characterization of P1, P3 and P5 peptides

CD experiments indicate unstructured conformation for the peptides P1, P3 and P5 at all screened buffer conditions (**Figure 2.9A**). Similarly, the 1D ^1H -NMR spectra of those peptides (**Figure 2.9B**) show low signal dispersion and high signal line broadening of the amide protons, both typical NMR characteristics of unfolded conformations. These experimental data are in strong contrast to the sequence based prediction, which reports the presence of conformation with abundant α -helical secondary structure elements. Such different folding properties obtained for all screened peptides suggest that not only the amino acid composition of the peptides but also the dynamic environment, e.g. the presence of interacting proteins, dictates the pathway of folding.

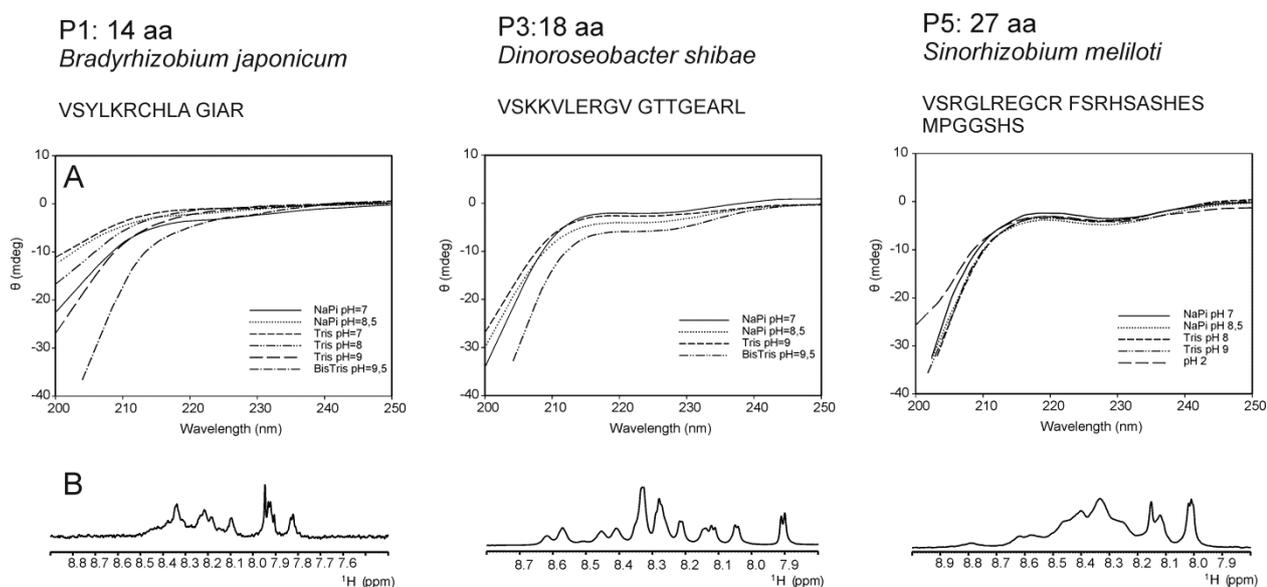


Figure 2.9. CD and NMR spectroscopy of peptides P1, P3 and P5. **A** - CD-spectra of 200 μM peptides in 25 mM different buffer systems as indicated in the legend. **B** - 1D ^1H -NMR spectra with the enlarged amide proton region of 1D ^1H -NMR spectra. All spectra were recorded at 298 K on the 700 MHz spectrometer in 25 mM sodium phosphate buffer (pH 7) containing 5% D_2O , 0.5 mM DSS as an internal reference and 0.2 mM (P1), 2 mM (P3) and 0.5 mM (P5) of peptides, respectively.

3.3.2. Structural characterization of P20 peptide

P20, encoded in trpL terminated attenuator RNA, was found to be functional in *Sinorhizobium meliloti*. This peptide induces the multidrug-efflux pump of *S. meliloti* and increases the minimal inhibitor concentration (mic) for some antibiotics. P20 is involved in the gene down-regulation, initiated by the exposure to antibiotics as tetracycline, chloramphenicol, erythromycin or rifampicin, which lead to increased protein expression, efflux activity and particularly to the multiresistance of *S. meliloti*. sRNA in combination with P20 and tetracycline (Tc) is capable for interaction with mRNA *in vivo*. Furthermore, the presence of tetracycline ligand is crucial for this interaction. To investigate the effect of those interactions on the structural properties of P20, both P20 in isolation and in complex with Tc were studied using NMR spectroscopy.

3.3.2.1. Investigation of peptide-ligand interaction

Due to the low water solubility of the peptide P20, NMR structural characterization was performed in DMSO-d_6 . P20 contains three tryptophan residues of which side chain ($\text{N}\epsilon\text{-H}\epsilon$) signals can be used as reporters for peptide-ligand interaction or protein aggregation. Tetracycline was found to be essential for proper cellular peptide functioning and in particular peptide-sRNA interaction, therefore the main interest was focused on peptide tetracycline interactions.

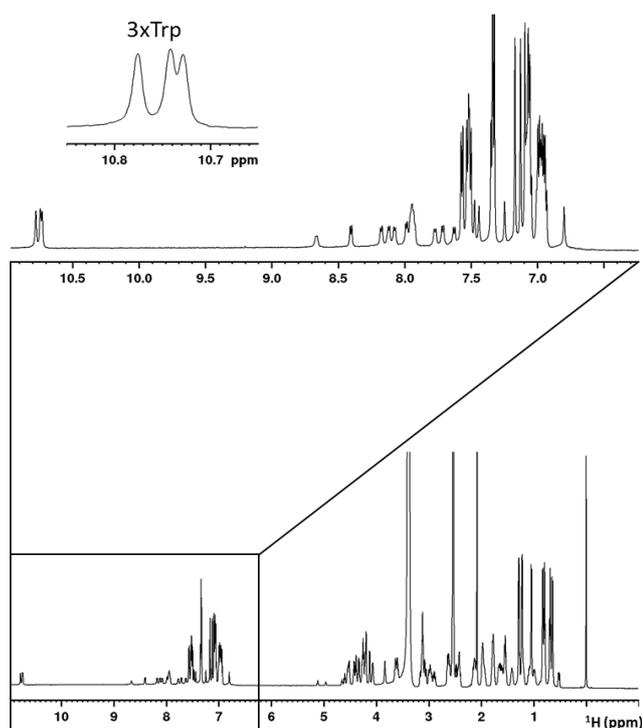


Figure 2.10. 1D ^1H -NMR spectrum of peptide P20. Spectral region corresponding to the backbone amide protons of P20 and the aromatic side chain, including region of the side chain H_ϵ proton are enlarged. 1D- ^1H -NMR spectrum (256 ns) was recorded at 298 K on the 600 MHz spectrometer. Sample contains 5 mM peptide in DMSO-d_6 and 0.5 mM DSS as an internal reference.

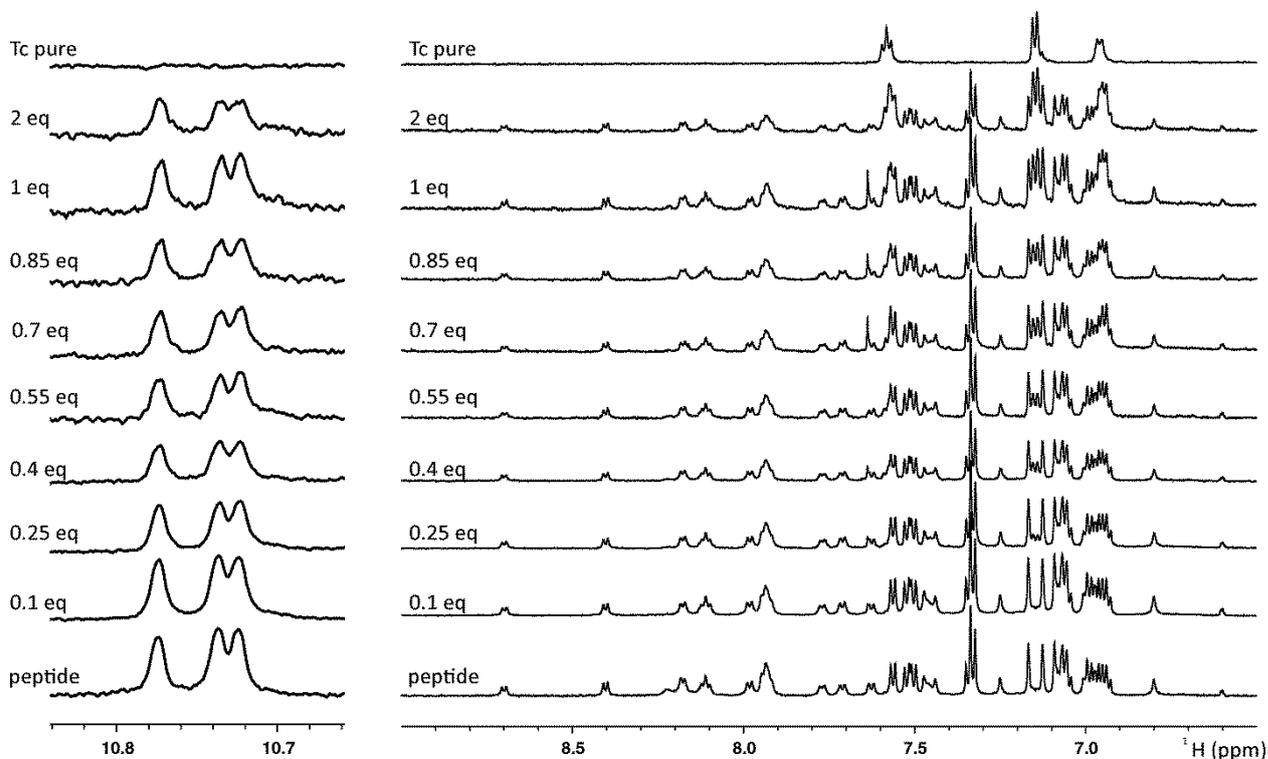


Figure 2.11. Titration studies of tetracycline and peptide P20. **Right panel** - enlarged region of the backbone amide protons of the 1D ^1H -NMR spectra of P20 (bottom), P20 and tetracycline (middle) and Tc (top). **Left panel** - enlarged region of the tryptophan side chain H_ϵ proton. All spectra were recorded at 298 K, 256 number of scans

on the 600 MHz spectrometer. Sample contains 5 mM peptide in DMSO- d_6 and 0.5 mM DSS as an internal reference.

NMR titration studies of P20 with tetracycline show CSPs for a set of backbone amide protons and the tryptophan side chain. These CSPs can result either from the interaction between P20 and tetracycline or from the reduction of the solubility of P20 due to the increasing amount of water during addition of the aqueous Tc solution. In order to avoid aggregation of P20, both, peptide and tetracycline were dissolved in DMSO- d_6 . No interaction between peptide and tetracycline could be observed (**Figure 2.11**).

Unfortunately, addition of pure water to the P20-DMSO- d_6 solution leads to CSPs (**Figure 2.12**), indicating water-induced aggregation of the peptide, excluding further interaction studies between P20 and RNA

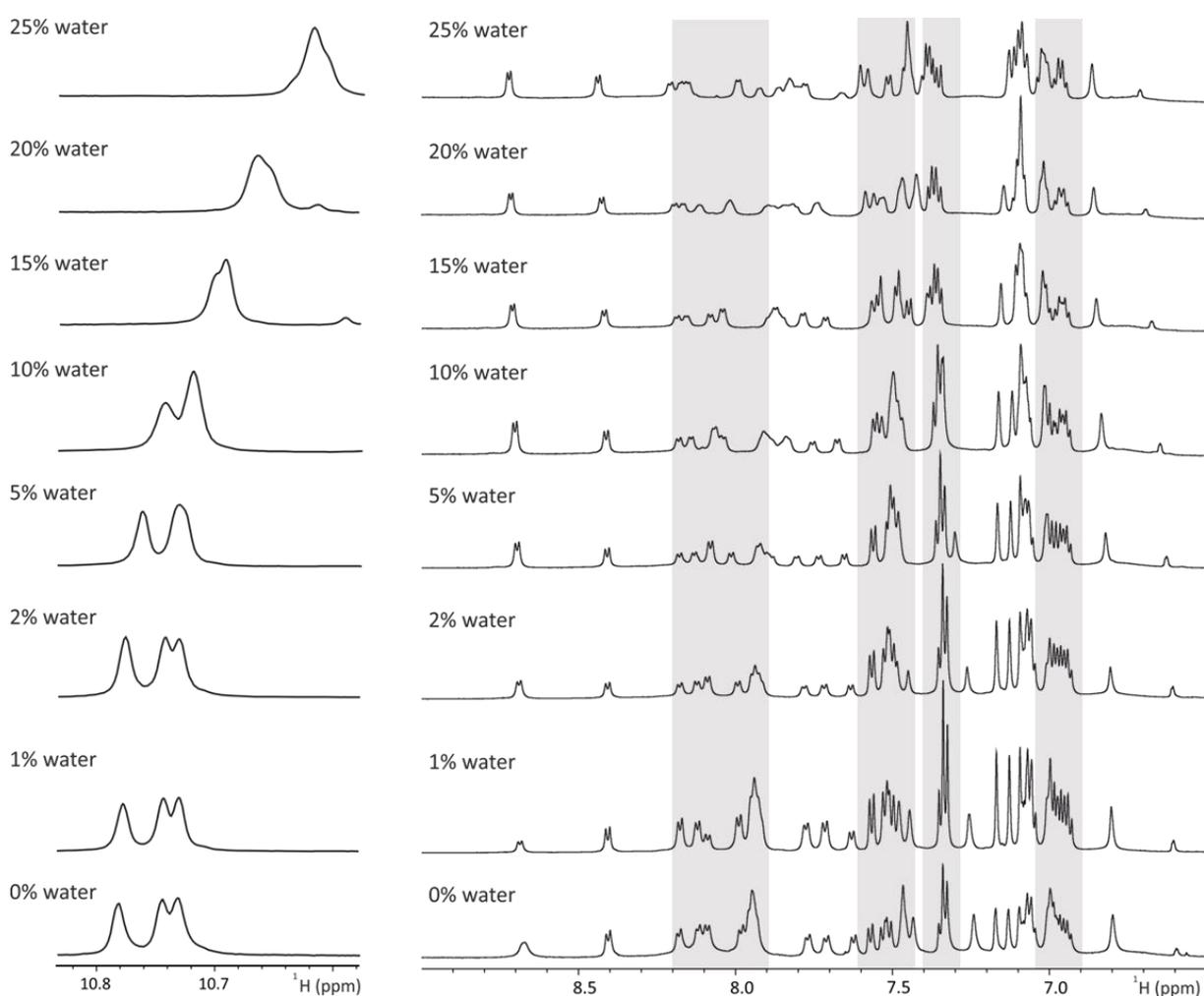


Figure 2.12. Titration studies of peptide P20 and water. Right panel - enlarged region of the backbone amide protons of the 1D ^1H -NMR spectra of P20 (bottom), P20 and water (middle). **Left panel** - enlarged region of the tryptophan side chain H_ϵ proton. All spectra were recorded at 298 K, 256 number of scans on the 600 MHz spectrometer. Sample contains 5 mM peptide in DMSO- d_6 and 0.5 mM DSS as an internal reference. Spectra regions which undergo significant chemical shift perturbations are highlighted in gray.

3.2.2.2. Structure calculation of peptide P20

The small size of P20 peptide allows to perform the full assignment based on natural abundance 2D NMR experiments. With 2D $^1\text{H}^{15}\text{N}$ HSQC, $^1\text{H}^{13}\text{C}$ HSQC (Appendix A8, Figure A4), $^1\text{H}^1\text{H}$ COSY, $^1\text{H}^1\text{H}$ TOCSY and $^1\text{H}^1\text{H}$ NOESY spectra recorded at 600 MHz, 298 K the backbone resonance assignment was completed to 98% and total resonance assignment to 92%.

The three-dimensional structure calculation of P20 was performed based on the 2D $^1\text{H}^1\text{H}$ NOESY cross-peak assignment using CYANA [44-46]. The manually assigned peak list was used as an input for the structure calculation. Figure 2.13 shows the NMR solution structure of P20 in DMSO- d_6 and the bundle of 10 lowest energy NMR structures (RMSD $0.33 \text{ \AA} \pm 0.14 \text{ \AA}$). The three-dimensional NMR structure of P20 shows an unstructured peptide with one C-terminally located α -helix formed by three tryptophan residues (W10, W11 and W12).

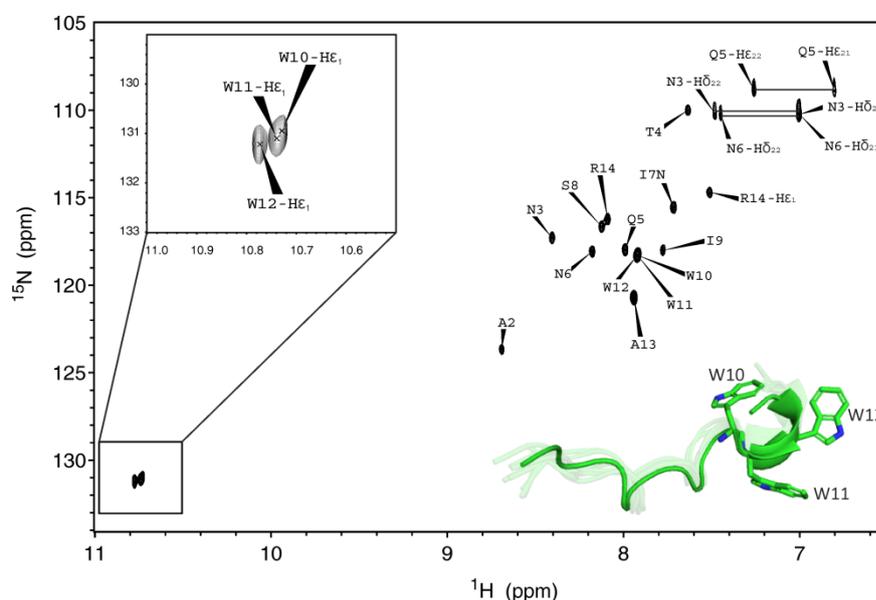


Figure 2.13. Structural characterization of peptide P20. Natural abundance 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum of peptide P20 (5 mM in DMSO- d_6) with the enlarged indole proton region (left corner, top) recorded at 298 K on the 600 MHz spectrometer. The backbone assignment of P20 is shown for the corresponding residue. Ribbon representation of the three-dimensional lowest energy NMR solution structure of P20 overlaid with the ensemble of the 10 best structures of P20 (right corner, bottom). Tryptophan residues (W10, W11 and W12) involved in the formation of the C-terminal α -helix are indicated as sticks. The figure was generated by PyMOL.

In general, presence of the partially folded regions within unstructured proteins may indicate a potential site of interaction with the biological partners. P20 is a mainly unstructured small protein containing a C-terminal located α -helix, which can play a role in the interaction with RNA, where the formation of protein-RNA complexes may potentially stabilize the structured conformation of the peptide. NMR titration studies of P20 and ligand (tetracycline) suggest no interaction. This indicates that Tc probably acts as an inhibitor during the interaction of P20 with RNA. However, the high hydrophobicity of P20 hampers further interaction studies of peptide and RNA under physiological conditions by NMR.

3.4. Cooperation with Ruth Schmitz-Streit group (University of Kiel)

The research group of Prof. Schmitz-Streit from the Christian-Albrechts-University of Kiel focuses on the investigation of small proteins from *Methanosarcina mazei*, which are known to be involved in the nitrogen regulation or biofilm formation. Around 70 small RNAs with sORF have been annotated within differential RNA-seq and term-seq analysis^[135,136]. Several peptides were found to be encoded by highly conserved spRNAs^[137]. It was hypothesized that, such peptides can be functional in *Methanosarcina mazei*. This study includes investigation of two peptides P8 and P12 (classified as protein 36 and protein 26, respectively, by Prasse *et al.*), which were assumed to be involved in the nitrogen regulation.

As a part of the academic collaboration, the secondary structure screening analysis was performed for five peptides and two small proteins originating from *Methanosarcina mazei*, which potentially play a role in either the nitrogen regulation or biofilm formation. Four of those seven constructs (P2, P4, P6 and P12) containing 18 to 23 amino acids in the sequence were successfully synthesized by SPPS. Only one peptide (P13) could not be synthesized by SPPS due to its high hydrophobicity. Two small proteins (P7 and P8), containing 53 and 61 residues in the sequence were attempted for the recombinant expression in *E. coli* cells as SUMO-fusion proteins. Unfortunately, small protein P7 degraded during the purification steps and P8 could not be expressed with SUMO-tag. Therefore, for P8 another fusion strategy was applied. A hexa-histidine tag was linked to the protein sequence via a thrombin cleavage recognition site without additional fusion tag. Purification included Ni-NTA affinity chromatography followed by size-exclusion chromatography to separate the small protein from the cleaved His-tag. This allowed to successfully perform the expression in *E. coli* cells and to obtain a stable protein.

Both CD and NMR based structural screening indicate unstructured states for all those peptides at different buffer conditions (**Figure 2.14**). Similar, as already observed for the peptides (P1, P3 and P5) provided by Elena Evguenieva-Hackenberg's group (**chapter II 3.3**), also here α -helical secondary structure elements were predicted, however could not be confirmed by the experimental data. The profile of the CD spectra for P2, P4, P6 and P12 recorded at different buffer conditions (**Figure 2.14B**) indicate random coil conformation. However, broad signal distribution of amide protons observed in 1D ^1H -NMR spectra of P4 and P12 might be interpreted as partially folded state (**Figure 2.14A**).

To investigate the initial structure of P4 and P12 peptides 2D natural abundance NMR experiments were performed. With 2D $^1\text{H}^{15}\text{N}$ HSQC, $^1\text{H}^{13}\text{C}$ HSQC, $^1\text{H}^1\text{H}$ COSY, $^1\text{H}^1\text{H}$ TOCSY and $^1\text{H}^1\text{H}$ NOESY it was possible to perform the backbone resonance assignment to 92% and 98% for peptide P4 and P12, respectively. For the P4 the degradation signals were observed in 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum, indicating the unstable form of the isolated peptide.

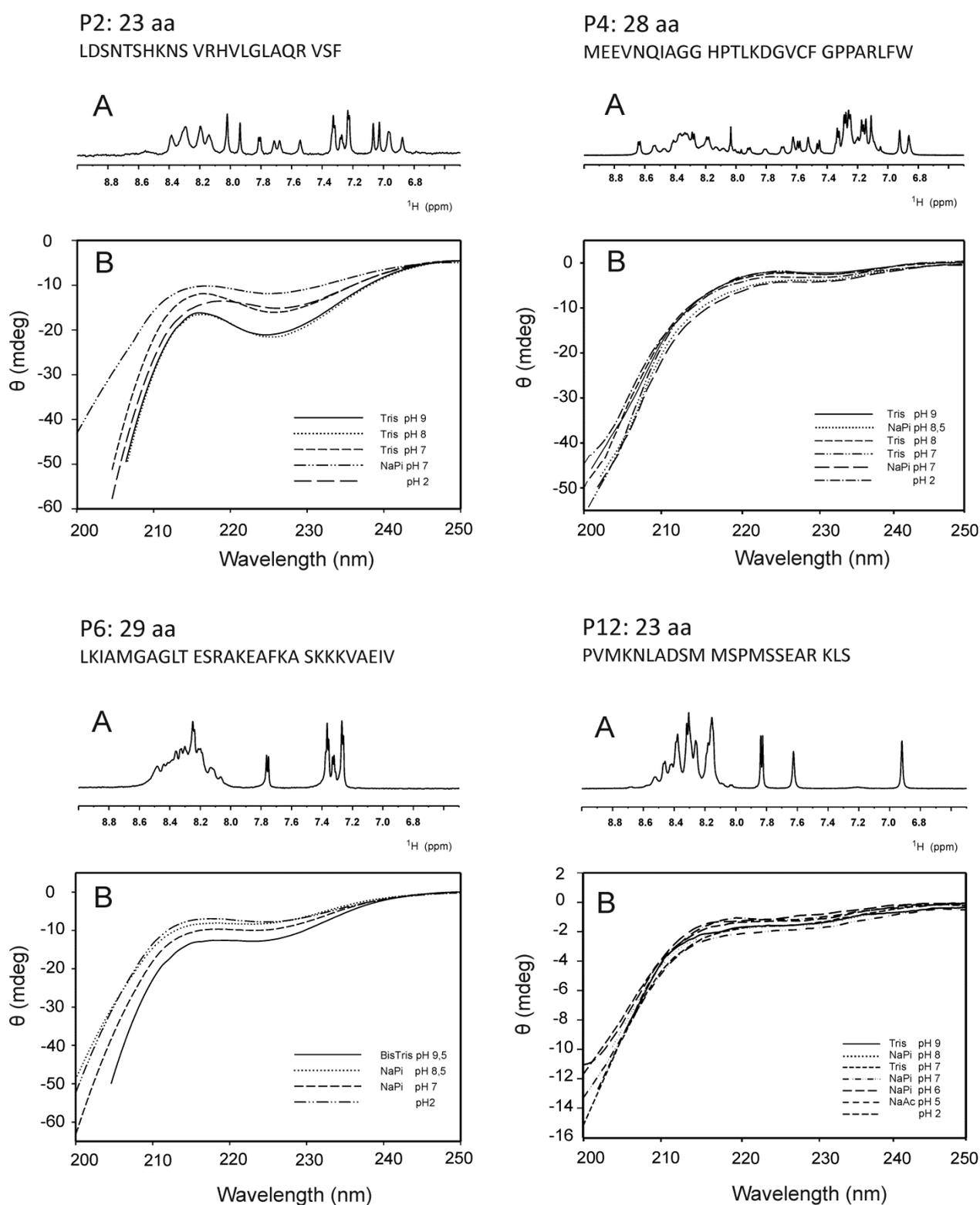


Figure 2.14. CD and NMR spectra of P2, P4, P6 and P12. **A** – 1D ^1H -NMR spectra of the enlarged amide proton region recorded at 298 K on the 700 MHz (P2, P4, P6) and 600 MHz (P12) spectrometer, (256 number of scans). NMR samples contain 0.2 mM (P2 and P5) and 0.5 mM (P6 and P12) peptides in 25 mM sodium phosphate buffer (pH 7), 5% D_2O and 0.5 mM DSS as an internal reference. **B** – CD spectra of 200 μM peptides (P2, P4 and P6) in different buffers (25 mM) as indicated in the legend. P12 contains 50 μM peptides in 50 mM different buffers as indicated in the legend.

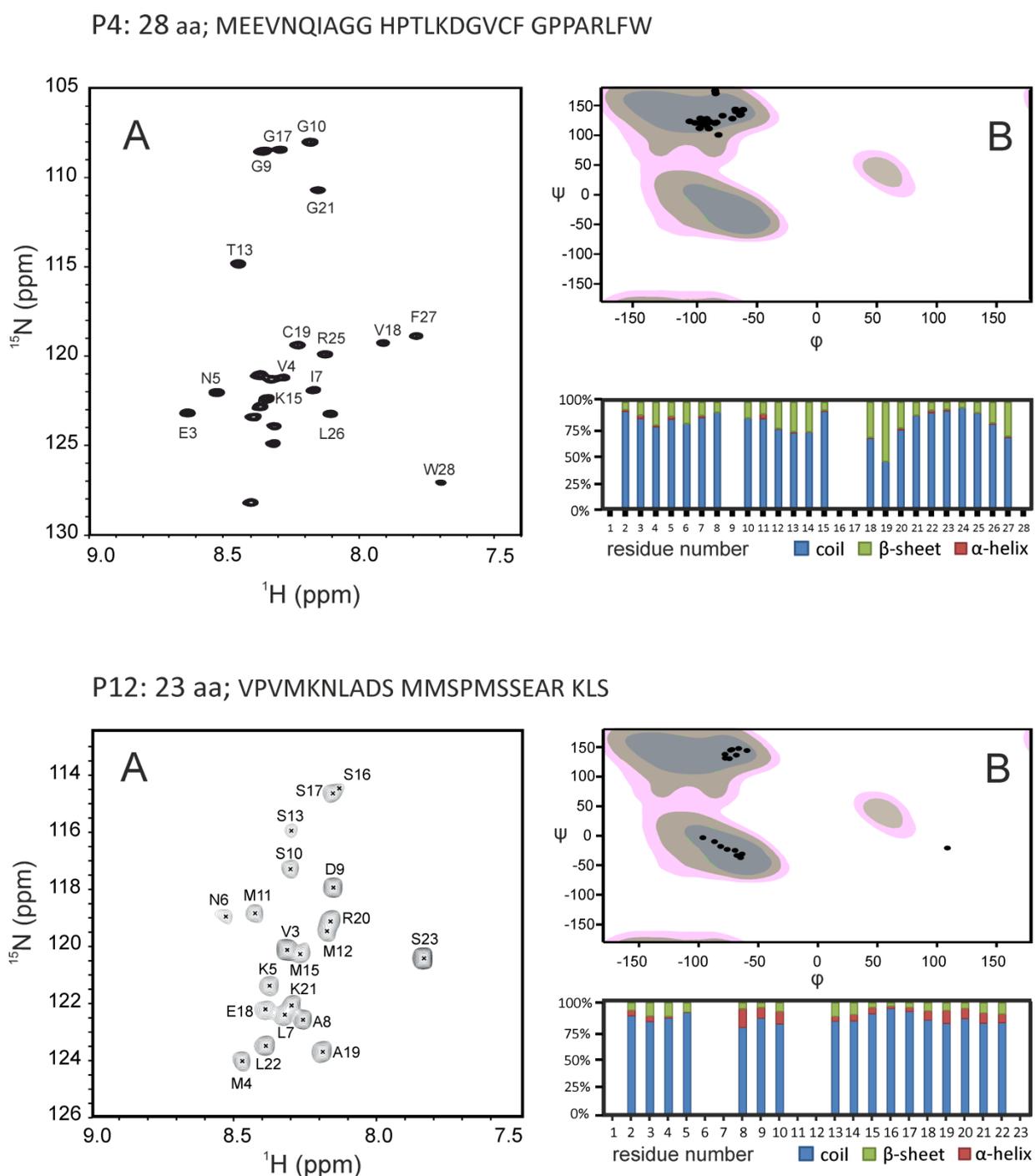


Figure 2.15. Secondary structure analysis of P4 (top) and P12 peptides (bottom). A – 2D SOFAST ^{15}N -NMR natural abundance spectrum recorded at 298 K. The amide cross-peaks are assigned to the corresponding residue. Peptides were dissolved in 25 mM sodium phosphate buffer (pH 7) containing 5% D_2O and 0.5 mM DSS. Spectrum of P4 was measured on a 800 MHz spectrometer with 2014 number of scans, while spectrum of P12 was measured on the 600 MHz spectrometer with 384 number of scans. B – TALOS prediction of residues which are classified as “good”. Top - Ramachandran plot of the predicted backbone torsion angles, bottom -secondary structure prediction as a function of residue number.

Backbone assignment was further used as an input for the site-specific secondary structure prediction TALOS^[114]. Such analysis predicts the secondary structure elements of the protein based on the experimentally determined chemical shift resonances. Both peptides were predicted to be completely unstructured (**Figure 2.15**).

P8: 61 aa;

VTIWEYDVKE IRFSEWSKAK EDLNNLGVVEG WELIKFSNEI DENG MVA AVF KRPVDYVDAA F

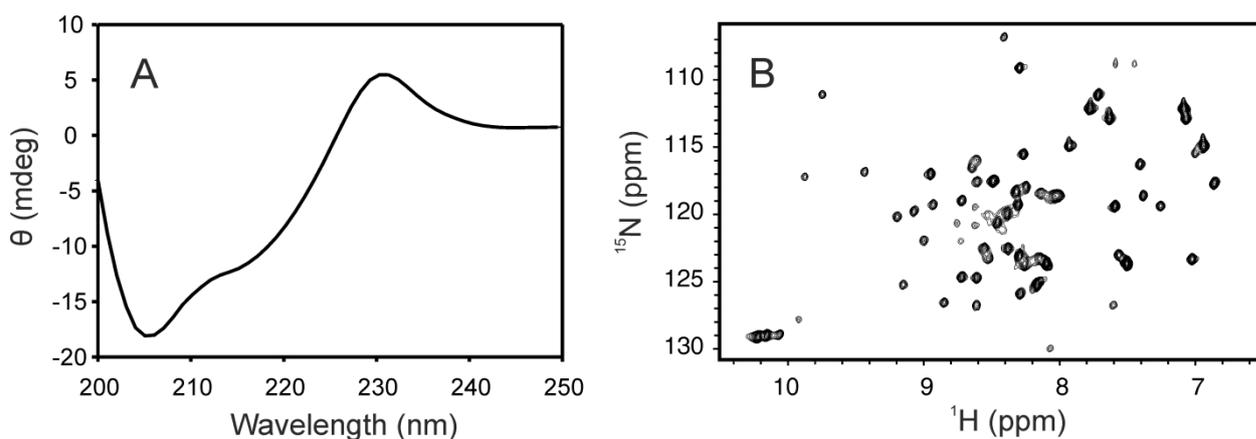


Figure 2.16. CD and NMR spectra of small protein P8. **A** – CD spectrum of 35 μM P8 small protein dissolved 25 mM sodium phosphate buffer (pH=8), 75 mM NaCl, recorded at 303K. **B** - 2D ^{15}N -NMR spectrum recorded at 303 K, on a 800 MHz spectrometer with 64 number of scans. Small protein was dissolved in 25 mM sodium phosphate buffer (pH=7) containing 300 MM NaCl, 5% D_2O and 0.5 mM DSS.

Additionally, small protein P8 was structurally characterized by CD and NMR spectroscopy (**Figure 2.16**). CD spectrum indicates a combination of α -helical and β -strand secondary structure elements. 54 amide cross-peaks were observed in 2D ^1H - ^{15}N HSQC spectrum, which corresponds to 92% of initially expected ones (59 expected signals). Broad signal distribution indicates folded β -sheet rich conformation, making small protein P8 a perfect candidate for further conformational analysis and structure calculation. This conformational analysis is, however, beyond the scope of this study and can be subjected to further investigations.

3.5. Cooperation with Wolfgang Streit group (University of Hamburg)

A symbiotic *Sinorhizobium fredii* (NGR234) is involved in the nitrogen-fixing nodules formation in the wide range of plants^[138]. Hundreds of peptides related to the infection process were found using cut-off 10-60 amino acids in the smallest replicon pNGR234a. Three of those peptides (P14, P15 and P16) were shown to increase the replicon copy number^[118]. Such increased copy number of plasmids leads to the low-level of the gene expression, including genes, which are responsible for the infection process.

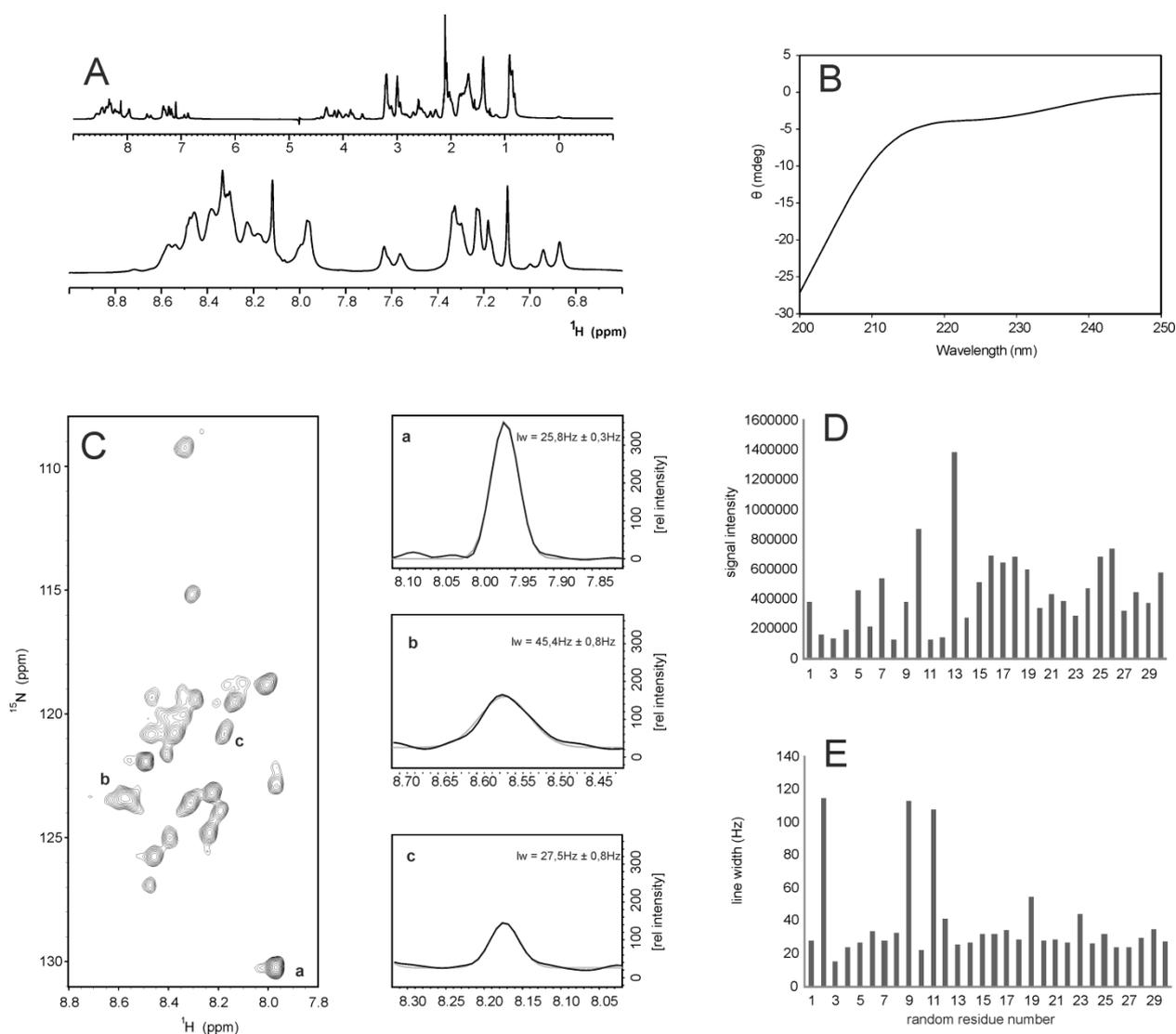


Figure 2.17. Secondary structure analysis of P14 peptide. A – 1D ^1H -NMR spectra of the enlarged amide proton region recorded at 298 K on a 600 MHz spectrometer. Peptide was dissolved in 25 mM sodium phosphate buffer (pH 7) containing 5% D_2O and 0.5 mM DSS. B – CD-spectrum of 100 μM peptide in 25 mM sodium phosphate buffer (pH 7). C – 2D natural abundance $^1\text{H}^{15}\text{N}$ HSQC spectrum (384 number of scans), at 298 K on a 600 MHz spectrometer. 1D-plane from 2D $^1\text{H}^{15}\text{N}$ HSQC of the corresponding signals (a, b, c). The resonance linewidths (lw) were obtained by deconvolution (shown in gray). D, E – graphical representation of signal intensity and line width from 2D $^1\text{H}^{15}\text{N}$ HSQC as a function of the random residue number^[118].

The 31 amino acid long peptide P14 was synthesized by SPPS. Small proteins P15 (51 amino acid chain) and P16 (43 amino acid chain) were expressed in *E. coli* cells as a SUMO-fusion. P14 and P16 could be successfully produced, while the P15 degraded during the purification indicating the unstable form of the isolated conformation.

Structural analysis of P14 and P16 using 2D $^1\text{H}^{15}\text{N}$ HSQC spectroscopy showed sets of broad signals with heterogeneous signal intensities, which are characteristic for the molten globule conformations. The different signal profiles observed in the 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum of P14 suggest rather a mixture of different conformations with the interconverting residual structures than a single favored folded conformation. The 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum of P16 resolves only 23 out of 41 expected cross-peaks (**Appendix A8, Figure A5**). The lack of expected signals indicates a conformational mixture rather than one preferred conformational state as well. The experimentally obtained data are in strong contrast to the predicted structures, where α -helical regions were defined for both peptides. The lack of experimental evidence for the presence of α -helical secondary structure does not necessarily exclude their existence.

3.6. Cooperation with Jörg Soppa (University of Frankfurt)

The scientific interest of the Soppa research group is focused on the zinc finger proteins (ZF) from the haloarchaeon *Haloferax volcanii*, containing two *C(P)XCG* zinc binding motifs. These proteins, containing less than 70 amino acids, are characterized by high amount of charged or hydrophilic residues. Moreover, they are potentially prone to post-translational modifications ^[50]. Generally, zinc finger proteins are known to bind DNA, RNA, proteins and small molecules. Typically, zinc ions stabilize the structure of the protein without being involved in direct interaction with the binding partners. The most typical structure motif adopted by the ZF includes two β -strand and one α -helical secondary structure elements.

3.6.1. Structural characterization of zinc finger proteins

Two zinc finger proteins P11 and P19, containing 38 and 59 amino acids, respectively, were structurally characterized by CD and NMR spectroscopy. CD spectra of P11, recorded at different pH values, show a combination of α -helical and random coil structural elements. The 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum of P11 clearly indicates a molten globule conformation (**Figure 2.18 A and B**). Addition of zinc ions and variation of the temperature do not significantly change the structural profile of the protein (**Appendix 8, Figure A6 and A7**).

In contrast, the 59 aa long small protein P19 was found to adopt a persistent structured conformation (**Figure 2.18 C and D**). CD spectrum of P19 indicates a combination of β -strand and α -helical secondary structure elements, which is consistent with the typical zinc finger structure. 71 amide cross-peaks were

observed in 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum of P19, which corresponds to 30% more signals than initially expected (55 expected signals). The additional peaks indicate presence of a second conformation. The analysis of the signal intensity defines two conformations with a population ratio of approximately 4 : 1. Detailed structural investigations of P19 are described in **chapter II 3.6.2**.

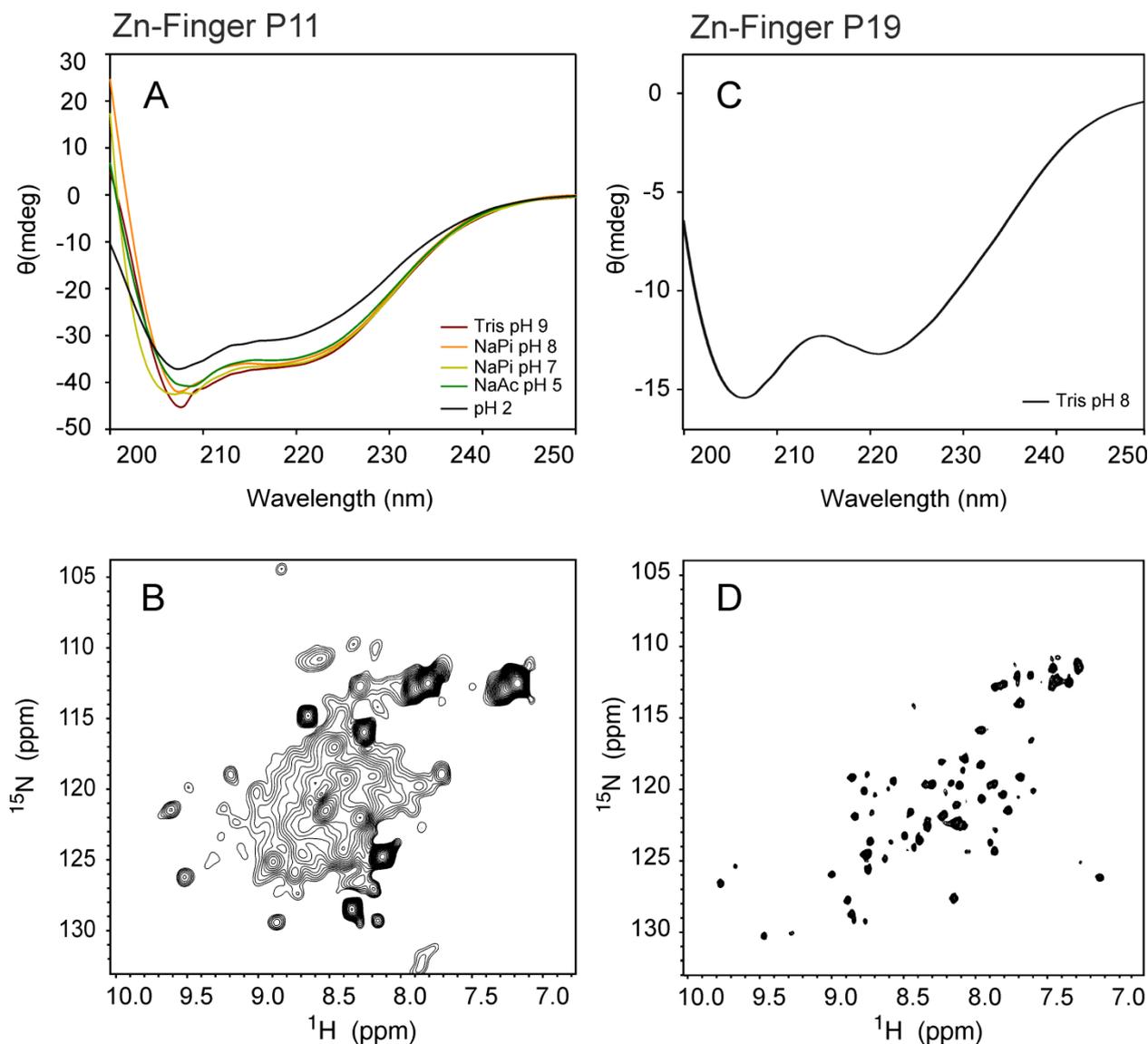


Figure 2.18. CD and NMR spectra of P11 and P19. **A, C** – CD-spectra in different buffers (50 mM) as indicated in the legend. **B, D** - 2D $^1\text{H}^{15}\text{N}$ HSQC spectra recorded at 800 MHz, 298 K. P11 contains 1mM protein, 50 mM sodium phosphate buffer pH=7.5, 100 mM NaCl, 3 mM DTT, 5% D_2O , 0.5 mM DSS; recorded at 600 MHz, 298 K. P19 small protein was dissolved in 25 mM Tris buffer pH=8, 200 mM NaCl, 100 μM ZnCl_2 , 5% D_2O , 0.5 mM DSS.

3.6.2. Structural characterization of P19

Small protein P19, which was heterologously expressed in *E. coli* (uniformly labeled M9 medium and LB medium), was structurally compared with the protein isolated from *Haloferax volcanii* in Prof. Soppa group. The 1D ^1H -NMR experiments show consistent chemical shifts for each small protein (**Figure 2.19**). The corresponding 2D $^1\text{H}^{15}\text{N}$ HSQC and 2D $^1\text{H}^{13}\text{C}$ HSQC spectra of both proteins can be found in **Appendix A8, Figure A10 - A13**.

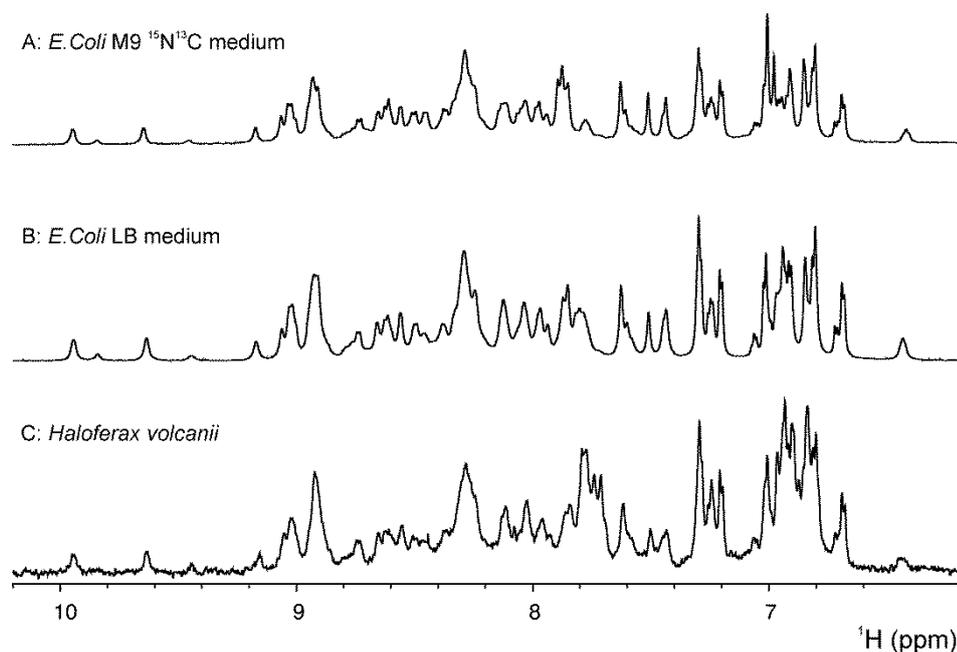


Figure 2.19. 1D ^1H spectra of P19 small protein recorded at 298 K. The samples contain 25 mM Tris buffer pH=7, 200 mM NaCl, 5% D_2O , 0.5 mM DSS. **A** – Small protein was expressed in *E. coli* using uniformly labeled M9 medium. 1.2 mM sample; 1D ^1H NMR experiment with ^{13}C , ^{15}N decoupling; 16 scans; 298K, 600 MHz. **B** – Small protein was expressed in *E. coli* using LB medium. 2.5 mM sample; 1D ^1H NMR experiment; 16 scans; 700 MHz. **C** – Small protein was isolated from *Haloferax volcanii*. 0.34 mM sample; 1D ^1H NMR experiment; 64 scans; 298K, 600 MHz.

Since the high salt conditions are typically required for bacterial growth of *HVO* (*Haloferax volcanii*), the influence of 1 M NaCl on the folding motif of P19 small protein was examined using 2D $^1\text{H}^{15}\text{N}$ HSQC spectra (**Appendix A8, Figure A14**). It was found that the high salt does not have an effect on the protein structure, allowing the NMR suitable low salt conditions to be transferred for the structural biology studies.

3.6.2.1. NMR backbone assignment and TALOS prediction

The ^1H , ^{15}N and ^{13}C chemical shift assignment of the P19 small protein was performed manually using standard double- and triple-resonance NMR experiments at 298 K. All collected NMR experiments are listed in **Appendix A6, Table A6**. The backbone assignment of the major conformation was completed to 97%. Approximately 22 residues show splitting pattern of the signals, which is referred to the minor conformation of P19 (**Figure 2.20B**).

Some of the signals, which correspond to the second conformation overlap with the major conformation. Furthermore, assignment of P19 based on 3D ^{15}N NOESY-HSQC and 3D ^{13}C NOESY-HSQC spectra, where the signal intensity relates directly to the distance restraints, is difficult. The automated restricted peak picking procedure by SPARKY 3.114^[109] is not possible. Therefore, elucidation of the folding conditions, where mainly one conformation is present, is highly required.

The ZZ-exchange NMR spectrum of P19 show no signals indicating the exchange between two observed conformational states. This suggests that the conformational exchange is in slow exchange regime and cannot be covered by the time scale (ms-s) of the experiment. Also, temperature modulations in the range from 278 to 318 K (**Appendix 8, Figure A8**) show no thermally induced conformational exchange. Nevertheless, formation of both conformations was found to be temperature dependent. At the low temperature (278 K) mainly major conformation is present, while the amide signals from the minor conformation are almost vanished. Temperature increase favors the minor conformation and decreases structural homogeneity. Thus, all NMR experiments needed for the chemical shift assignment and further structure calculation were repeated at 278 K. The experiments for the chemical shift assignment of the P19 small protein at 278 K were recorded with NUS and TA program (**chapter II 2.3.2.2**). The resonance assignment was performed using the automated FLYA method^[43] (**Appendix A8, Figure A9**) and is fully comparable with the manually achieved chemical shift assignment of the major conformation at 298 K.

Finally, the TALOS-N secondary structure prediction^[114] was performed for both major and minor conformations at 298 K and for single major conformation at 278 K (**Figure 2.20B**). The predicted secondary structure elements were compared with the SWISS model (**Figure 2.20C**). There are only minor differences between two observed populations, while the SWISS model is significantly less structured and differs completely from TALOS-N. The reason is the low sequence identity to the template (26%), which does not allow a prediction with high accuracy.

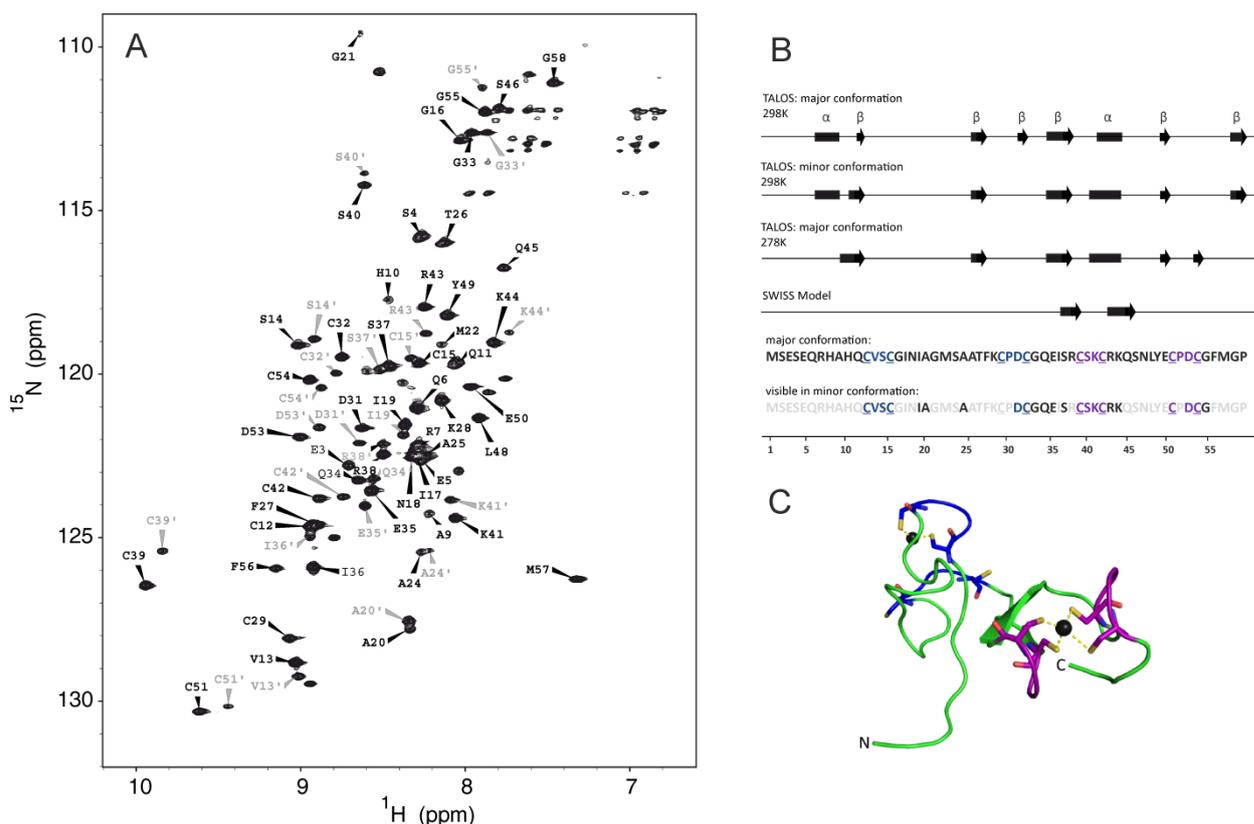


Figure 2.20. Backbone assignment of P19 small protein. **A** – 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum recorded at 800 MHz, at 298 K. Assignment of the major conformation is highlighted in black and the minor conformation is colored in grey. The sample contains 25 mM Tris buffer pH 8, 200 mM NaCl, 100 μM ZnCl_2 , 5% D_2O , 0.5 mM DSS. **B** - TALOS-N secondary structure prediction for major and minor conformations at 298 K, single major conformation at 278 K and secondary structure elements from the SWISS model. Secondary structure elements are shown as a function of the residue number. Color code for the protein sequence: blue - N-terminal zinc binding motif; purple - C-terminal zinc binding motif; black - protein sequence; gray - amide signals, for which the chemical shifts of the minor conformation are identical to the major conformation at 298 K. **C** – SWISS model structure based on the template with pdb code: 5IJL (D-family DNA polymerase - DP2 subunit catalytic subunit). Sequence identity is 26%. Zinc ions are shown as black spheres.

3.6.2.2. Influence of the zinc ions on the folding pathway

Typically, zinc ions serve as stabilization agents and induce folding of the zinc finger protein. Since P19 consists of two zinc binding motifs and NMR structural analysis showed two conformations of the protein, the following questions are attempted to be answered in this thesis: how zinc ions are involved in protein folding pathway and how many zinc ions are incorporated in the binding pockets.

In order to answer these questions, the zinc ions were first removed from the pure protein and then added again to the zinc free protein solution. The structural changes of the protein induced by the zinc ions were monitored by 2D $^1\text{H}^{15}\text{N}$ HSQC spectra. The complete removal of the ions was achieved by incubating the protein with tenfold excess of EDTA ($K_D = 10^{-6}$ M) and dialysis against zinc ions at high temperature (338 K) for four hours. The unfolded protein was subsequently titrated with Zn^{2+} ions

(Figure 2.21). Addition of 0.25 equivalents of Zn^{2+} ions already induces the refolding of the protein. Remarkably, both conformations refold simultaneously appearing with the same population ratio (4 : 1). 0.75 equivalent of Zn^{2+} are enough to completely refold the protein. Additional zinc ions do not change the fold of the protein any further. This indicates that both protein conformations incorporate only one zinc ion. The fact that only one zinc ion is bound to the protein was confirmed by the fluorescence experiments on the native *Haloferax volcanii* protein, performed by Sebastian Zahn from the Prof. Soppa group.

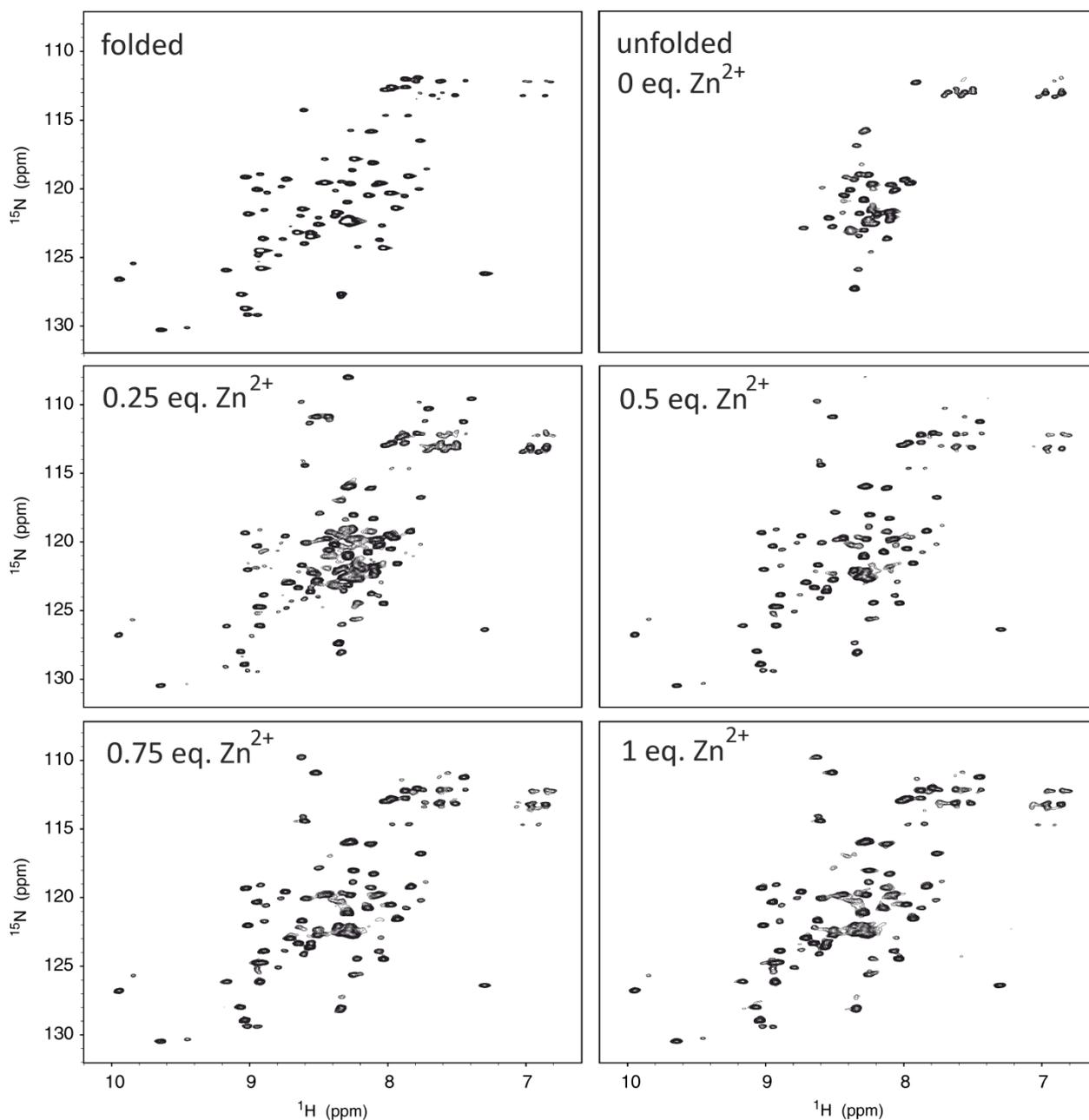


Figure 2.21. Zinc titration series of P19 small protein. 2D $^1H^{15}N$ HSQC spectrum recorded at 800 MHz, at 298 K. The sample contains 25 mM Tris buffer pH 7, 200 mM NaCl, 5% D_2O , 0.5 mM DSS.

3.6.2.3. Discussion

Despite the presence of two zinc binding motifs in the sequence, it has been shown, that only one zinc ion is bound to the protein. This might indicate that one binding pocket remains unoccupied, while the second one is incorporated with the metal ion. TALOS-N predictions reported more structural elements located in the C-terminal region compared to the flexible N-terminus. Even more, the amide signals from two cysteines C39 and C51 are shifted downfield, indicating the deshielding effect induced by the coordination of the metal ion. According to the SWISS model, the zinc ion is coordinated by four C-terminal cysteines. The second metal ion, which is located in the N-terminal binding pocket, is coordinated by only two cysteines, while the side chains of the other two cysteines are turned outside of the binding pocket. Taken together, the C-terminal zinc binding motif more likely contains the metal ion, while the N-terminal binding pocket remains unoccupied. To prove this assumption, further studies will focus on structure calculation of the P19 small protein. Moreover, mutational studies involving cysteine to alanine mutations will shed light on the zinc binding properties of this protein.

3.7. Cooperation with Anita Marchfelder (University of Ulm)

The research group of Prof. Marchfelder from the University of Ulm aims to identify the complete proteome of the halophilic archaeon *Haloferax volcanii* and to define the cellular functions of the selected small proteins. In order to identify the proteins overexpressed under different stress conditions, mass spectrometry was utilized. Some biological impact with an interesting cellular phenotype was observed.

This thesis comprises structural studies of two small proteins from *Haloferax volcanii* (P17 and P18). Both proteins were differentially expressed in response to changes in iron and salt concentrations, suggesting a stress-regulated expression. P17 is an iron regulated protein containing 60 amino acids in the sequence. P18 is a 61 amino acids long protein and it belongs to the high salt regulated proteins. Both proteins are conserved among Haloarchaea and contain domain with unknown function (DUF1508). For the structural characterization of P17 and P18 ¹⁵N, ¹³C-isotope labeled proteins were produced.

3.7.1. Structural characterization of P18 and P17 small proteins

Screening analysis of the 61 amino acids long high salt regulated protein P18 shows a mixture of different conformational states. CD spectra recorded at pH 9, 7.5 and 5 indicate a mixture of α -helical and β -sheet secondary structure elements in the protein. Additional reduction of the pH value to pH 2 lead to the partial unfolding of the protein (**Figure 2.22A left panel**), while all other pH values do not show significant impact on the secondary structure. The wide chemical shift distribution of amide protons observed in 1D ¹H-NMR experiment indicate a partially folded conformation of the protein. On

the other side the amide signals from the 2D $^1\text{H}^{15}\text{N}$ HSQC NMR experiment are characterized by different intensities and increased line broadening (Figure 2.22C left panel and Appendix A8, Figure A15). This provides a clear evidence for molten globule state adopted by the protein. NMR screening of folding conditions for P18 is shown in chapter II 3.7.2.

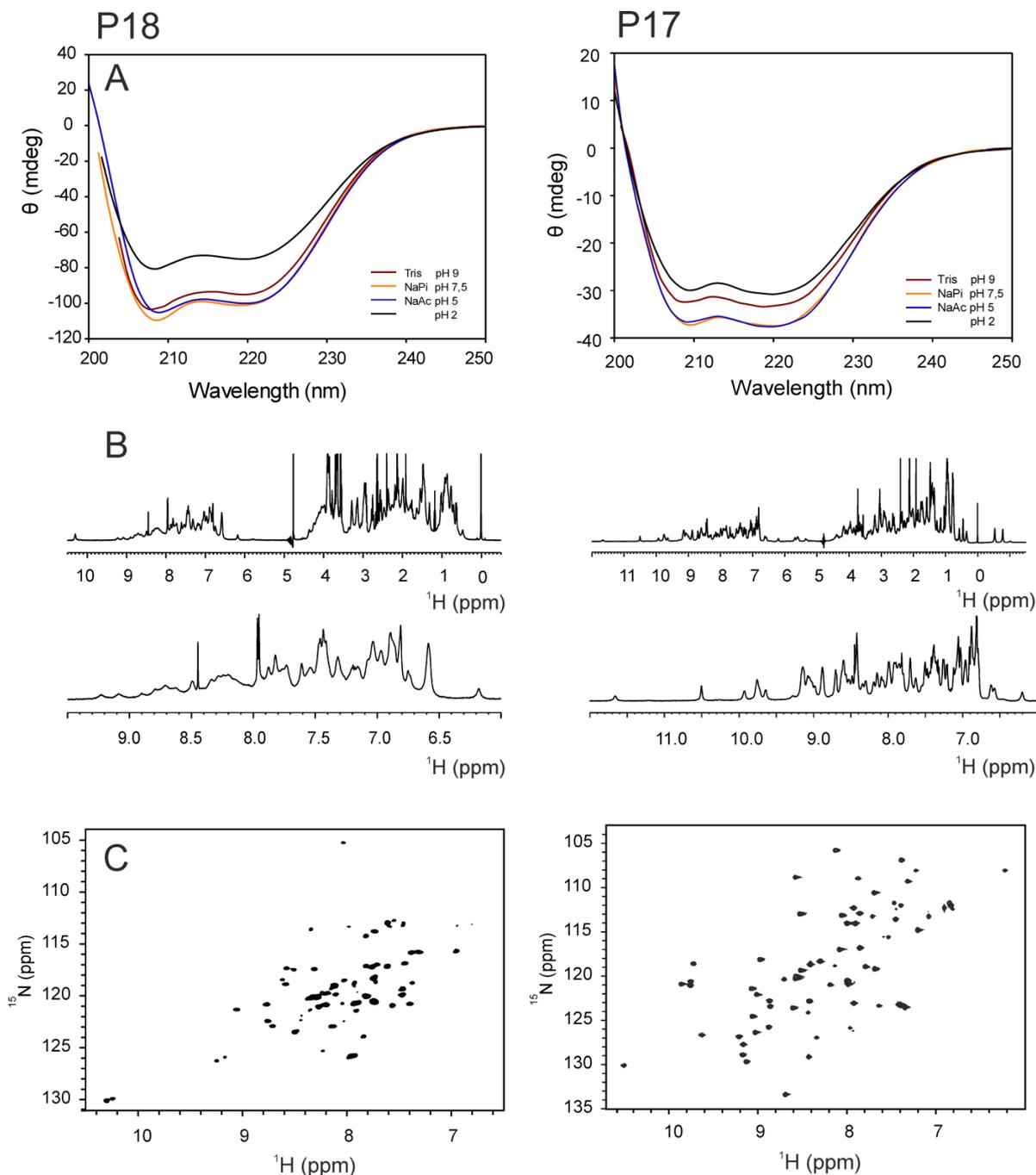


Figure 2.22. CD and NMR spectroscopy of P18 (left) and P17 (right) small proteins. A – CD-spectra of 100 μM (P18) and 50 μM (P17) protein in different buffers (50mM) as indicated in the legend. **B** – 1D ^1H -NMR spectra with amide proton region enlarged recorded with ^{15}N decoupling. **C** – 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum recorded at 600 MHz, 298 K; P18 contains 1 mM protein, 50 mM sodium phosphate buffer pH 7, 100 mM NaCl, 3 mM DTT, 5% D_2O , 0.5 mM DSS. P17 contains 1 mM protein, 50 mM sodium phosphate buffer pH 7.5, 100 mM NaCl, 5% D_2O , 0.5 mM DSS.

Small protein P17 was also investigated by CD and NMR spectroscopy. All CD spectra at different pH show the typical profile for folded proteins with specific minima at 208 and 222 nm. This indicates a mixture of α -helical and β -sheet secondary structure elements in the protein. The pH values of 7.5 and 5 show the same intensity profile, while increasing the pH value to pH 9 lead to the less pronounced CD profile and therefore partial loss of the folded conformation (**Figure 2.22A right panel**). The acidic pH 2 partially unfolds the protein resulting in the intensity decrease of the CD profile. All following experiments were performed in phosphate buffer at pH 7.5.

The 1D ^1H -NMR experiment showed a broad chemical shift distribution of backbone amide signals ranging from 10 to 7 ppm, indicating a folded protein conformation. The visible methyl signals at 0.5 and 0.3 ppm also provide evidence for the folded state (**Figure 2.22B right panel**). The 2D $^1\text{H}^{15}\text{N}$ HSQC NMR experiment is shown in **Figure 2.22C right panel**. Backbone amide signal dispersion indicated a folded conformation rich of β -sheet secondary structure elements. Similar peak intensities and line widths reflect the only one conformation adopted by the protein.

Both CD and NMR spectroscopy experiments indicate a persistent structure of the P17 protein, making it a perfect candidate for further conformational analysis and structure calculation. The structural analysis can be found in **chapter II 3.7.3**.

3.7.2. Screening of folding conditions of P18

In order to identify the most suitable folding conditions of protein P18, the screening analysis was performed. Therefore, the influence of temperature, pH, salt concentration and detergent was investigated by recording 2D $^1\text{H}^{15}\text{N}$ HSQC NMR experiments.

To investigate the influence of the pH on the preferable conformational state of the protein, three different pH values (6, 7, and 8) were monitored by NMR (**Figure 2.23A**). 57 amide signals are expected for the 61 aa long protein. Approximately this number of signals can be observed at pH 7. At pH 6 the number of amide signals is two times higher. Newly appeared signals with lower intensities indicate formation of a new conformation of the protein. Increasing the pH to more basic values (pH 8) has an opposite effect. Due to the faster amide proton-water exchange the number of visible signals in the 2D spectrum decreases, making these conditions not optimal for the NMR structure analysis.

The impact of salt concentration on the folding pathway and its stabilizing effect is widely used in protein folding investigations. Therefore, the effect of the different salt concentration was monitored by NMR (**Figure 2.23B**). The salt concentration was varied from 0 to 1 M, keeping the pH value constant for each measurement (pH 7). The salt effect on the protein conformation becomes visible by adding 100 mM NaCl, while the spectrum with 25 mM NaCl shows no significant difference to the protein

recorded without salt. At 100 mM salt concentration the 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum undergoes significant changes including signal loss and appearance of new signals. The line broadening of the signals decreases, indicating the salt stabilizing effect. Additional increase of salt concentration (up to 300 mM) leads to appearance of new signals that are less intense, indicating the formation of an additional minor conformation. This effect became stronger by increasing the salt concentration up to 500 mM, where a steady state was reached.

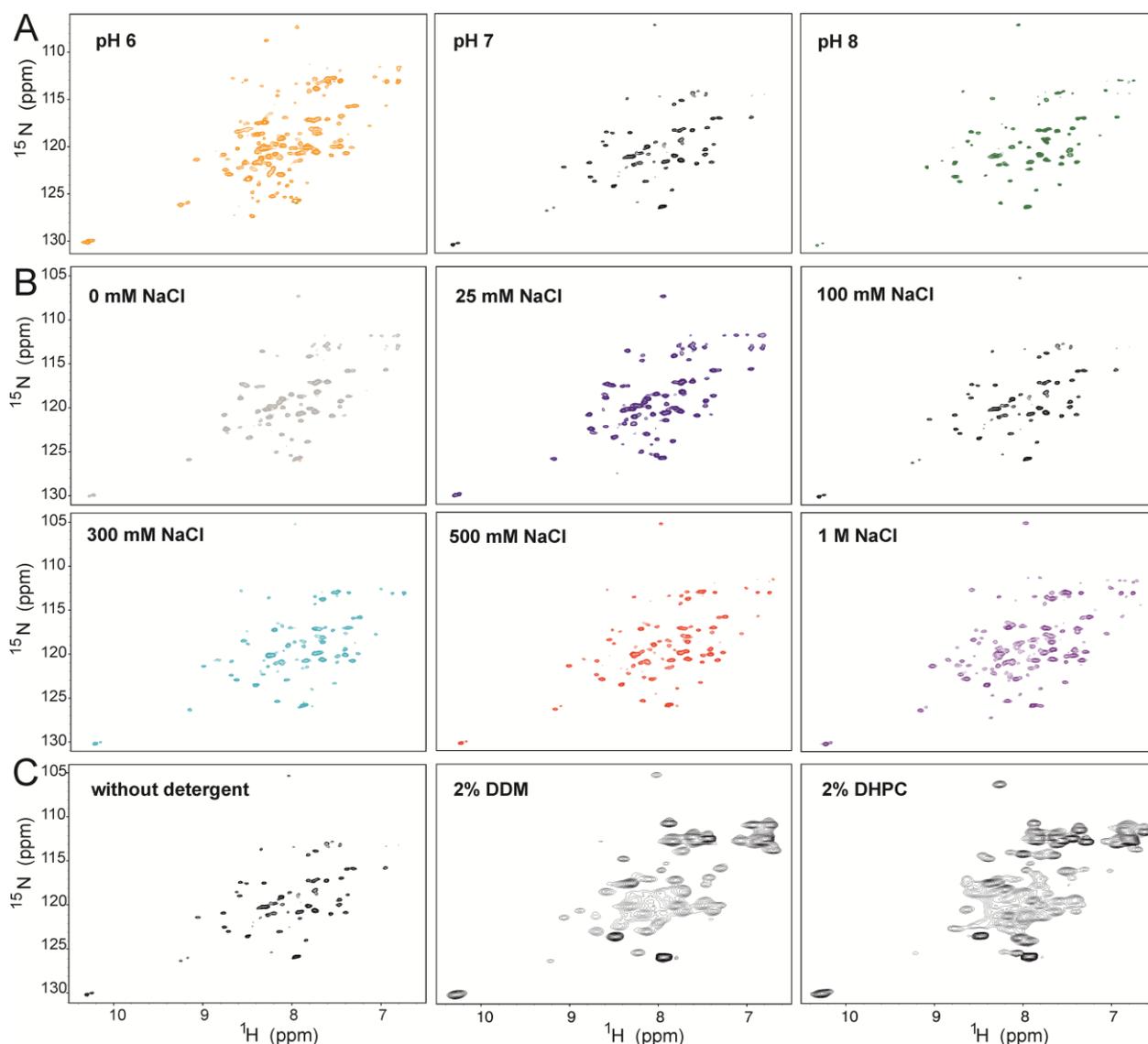


Figure 2.23. Screening of folding conditions of P18 small protein. **A, B** – A titration series of 2D $^1\text{H}^{15}\text{N}$ HSQC spectra acquired at 700 MHz, 298 K, 25 mM phosphate buffer, pH and NaCl concentration as indicated in the figure. **A** – a pH titration series acquired at 100 mM NaCl. **B** – a salt dependent titration series acquired at pH 7. **C** – 2D $^1\text{H}^{15}\text{N}$ HSQC spectra of P18 protein upon interaction with the lipids DHPC or DDM recorded at 600 MHz, 298 K. It contains 1 mM protein, 50 mM sodium phosphate buffer pH 7.5, 3 mM DTT, 5% D_2O , 0.5 mM DSS. The spectrum without detergent was recorded with 16 scans, while spectra with detergent were recorded with 256 scans.

The conformational mixture of the protein is clearly indicated by the well resolved tryptophan indole signals. With only one tryptophan in the protein sequence, the $^1\text{H}^{15}\text{N}$ HSQC NMR spectrum shows a doubling of the indole signal at all stress conditions. Different intensities in dependence of the conditions reflect the equilibrium change of the preferable conformation.

P18 can potentially belong to the membrane proteins. Thus, in order to investigate the influence of the lipid on the protein structure, the 2D $^1\text{H}^{15}\text{N}$ HSQC spectra were recorded after addition of DHPC and DDM detergents (**Figure 2.23C**). The full membrane mimetic screening requires a high number of different detergents for the simulation of the biological membrane. The choice of these detergents was made based on the commonly used detergents in Prof. Schwalbe group. The extremely increased line broadening followed by the bad resolution of the signals suggests that this strategy is not suitable for the particular small protein.

3.7.3. Structural characterization of P17 small protein

3.7.3.1. NMR spectroscopy of P17

3.7.3.1.1. NMR backbone assignment and TALOS prediction

The ^1H , ^{15}N and ^{13}C chemical shift assignment of the P17 protein was performed using standard double- and triple-resonance NMR experiments at 298 K. For the backbone assignment the following 3D NMR experiments were recorded: HNCACB, HN(CO)CACB, HNC(O) and HNHA. All measured NMR experiments are listed in **Appendix A6, Table A7**. The sequential assignment of the protein was performed by using a combination of HNCACB and HN(CO)CACB experiments (**Figure 2.24B**).

The 2D $^1\text{H}^{15}\text{N}$ BEST-TROSY spectrum, recorded at 600 MHz and 298 K, with the annotated assignment is shown in **Figure 2.24C** and the 2D $^1\text{H}^{13}\text{C}$ HSQC spectrum can be found in **Appendix A8, Figure A16**. All amide chemical shifts, except of not visible residues N2 and A36, could be assigned. The signals of the residues A41 and I45 overlap in the 2D $^1\text{H}^{15}\text{N}$ BEST-TROSY spectrum. However, with the use of 3D NMR spectra it was possible to resolve and unambiguously assign them.

In order to predict the secondary structure elements of the protein, the TALOS-N analysis ^[114] was performed. The backbone chemical shift assignment was used as an input. The predicted secondary structure elements include five folding motifs: four β -strands and one α -helix. The β -strand I involves residue H5 to V10, β -strand II is formed from residues Y16 to H22, small β -strand III includes two residues L28 and A29 and β -strand IV is built from residues V57 to E59. The α -helix involves residues K38 to P53 (**Figure 2.24A**).

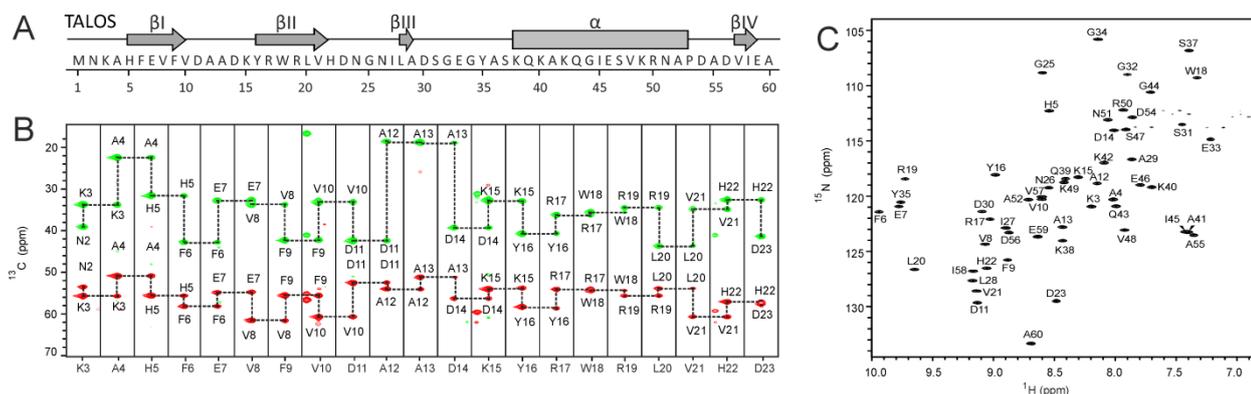


Figure 2.24. Backbone assignment of P17. **A** – TALOS-N secondary structure prediction is shown. Predicted secondary structure elements are shown as a function of the residue number. **B** – Sequential assignment of P17 protein for the residues K3 to D23. 3D HNCACB NMR spectrum was recorded at 700 MHz, 298 K. **C**^A are shown in red and **C**^B are highlighted in green. **C** – 2D ^{15}N BEST-TROSY spectrum of P17 small protein. Assignment is indicated in the picture. The experiment was recorded at 600 MHz, 298 K. Sample contains 5 mM protein, 50 mM sodium phosphate buffer pH 7.5, 100 mM NaCl, 5% D₂O, 0.5 mM DSS.

3.7.3.1.2. Side chain assignment of P17

The aliphatic side chain assignment of the 2D $^1\text{H}^{13}\text{C}$ HSQC was performed using 3D (H)CC(CO)NH and H(C)CH-TOCSY NMR experiments. To complete the assignment, homonuclear 2D $^1\text{H}^1\text{H}$ NOESY and heteronuclear 3D $^1\text{H}^{15}\text{N}$ TOCSY and $^1\text{H}^{13}\text{C}$ TOCSY in combination with 3D $^1\text{H}^1\text{H}^{15}\text{N}$ -NOESY-HSQC and 3D $^1\text{H}^1\text{H}^{13}\text{C}$ -NOESY-HSQC were recorded. With the 2D aromatic side chain (HB)CB(CGCDCE)HE and (HB)CB(CGCD)HD experiments it was possible to obtain a resonance assignment of 96.1%.

3.7.3.1.3. $^3\text{J}_{\text{HNHA}}$ coupling constants

The $^3\text{J}_{\text{HNHA}}$ coupling constants were calculated from 3D HNHA NMR experiment [139]. The determined values are in the range from 2 to 9 Hz, supporting the TALOS secondary structure prediction. Coupling constants with the value higher than 8 Hz correspond to the β -strand secondary structure elements, while the lower coupling constants (< 5 Hz) indicate the α -helical structure formation. The obtained coupling constants are in agreement with the predicted secondary structure elements, such as β -strand I (from residue H5 to F9), β -strand II (from the residue K15 to V21) and α -helix (from the residue K38 to R50) (Appendix A8, Figure A17).

3.7.3.1.4. Intramolecular hydrogen bond formation

The intramolecular hydrogen bond formation can be studied by monitoring the temperature induced amide protons chemical shift perturbations. A series of 2D $^1\text{H}^{15}\text{N}$ HSQC spectra was recorded for the temperature range from 278 to 318 K with 5 K increments. The temperature coefficients of the amide protons was calculated from a linear fit of a chemical shift perturbation as a function of the temperature [110]. The temperature coefficients with values more negative than -4.5 ppb/K are

characteristic for the rapidly exchangeable and therefore not hydrogen bonded amide protons, while protons with the temperature coefficient higher than -4.5 ppb/K are involved in hydrogen bonding ^[110].

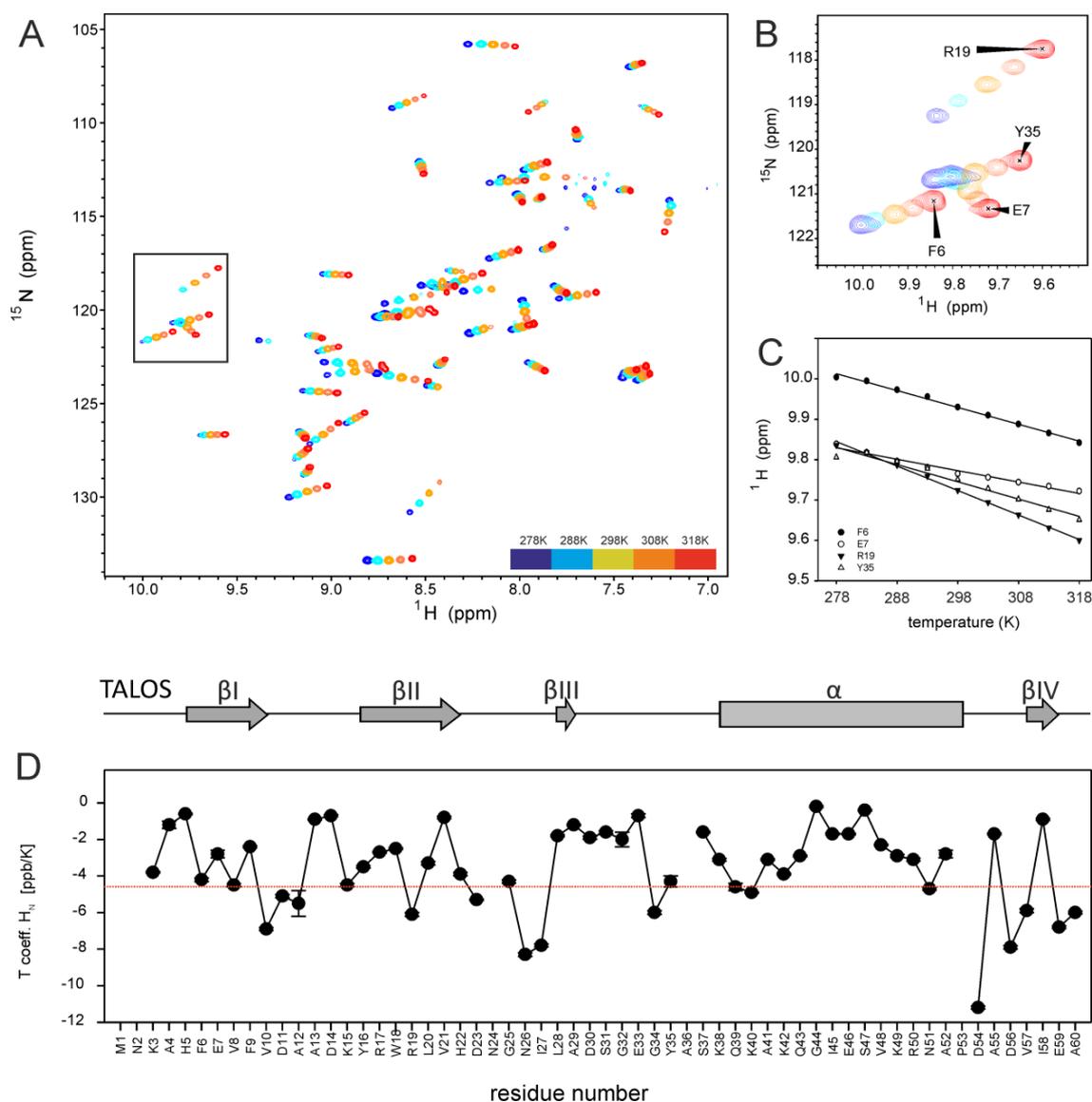


Figure 2.25. Amide proton temperature-dependency of P17 small protein. **A** – 2D $^1\text{H}^{15}\text{N}$ BEST-TROSY spectra recorded at 600 MHz, at different temperatures (color code is shown on the legend) using 50 mM phosphate buffer pH 7.5, 100 mM NaCl. **B** – Zoom in the region of F6, E7, R19 and Y35 residues, representing different hydrogen bond formation. **C** – Plot of the amide proton chemical shift (in ppm) as a function of temperature (in Kelvin). **D** – The calculated temperature coefficient (T_{coeff}) determined from a series of 2D $^1\text{H}^{15}\text{N}$ BEST-TROSY spectra (measured from 278 K to 318 K with 5 K increments). If the temperature coefficient is more negative than -4.5 ppb/K, highlighted as a red line, the amide proton exchanges rapidly and is not hydrogen bonded ^[110]. On the top the TALOS-N secondary structure prediction is shown.

The determined temperature coefficients are consistent with the TALOS predicted secondary structure elements β -strand I, β -strand II and α -helix. The obtained values for the residues from L28 to E33, located in the loop region between β -strand II and α -helix, are more negative than -4.5 ppb/K, indicating the residues to be involved in the intramolecular hydrogen bond formation (**Figure 2.25**).

3.7.3.1.5. Hydrogen-deuterium exchange NMR experiments

The additional structural information about the solvent accessibility of the amide protons and therefore the hydrogen bond formation was obtained from the hydrogen-deuterium exchange NMR experiments. For the 60 amino acids long protein 55 signals were observed in the 2D $^1\text{H}^{15}\text{N}$ BEST-TROSY NMR experiment. After the reconstitution with D_2O , followed by the immediate acquisition of time-dependent 2D NMR spectra, 18 amide signals disappear, indicating the fast hydrogen-deuterium exchange of these residues. 22 hours after reconstitution 21 signals are still remaining in the spectrum. The visible residues ranging from H5 to F9 and from K15 to V21 are consistent with the predicted β -strand I and β -strand II, respectively. Further, the α -helical part of the structure ranging from the residues Q39 to A52 is also stabilized through the hydrogen bonding and thus is in line with the TALOS analysis.

The amides forming the predicted β -sheet III (L28 and A29) followed by the next two residues D30 and S31 are very well protected and therefore not accessible for the solvent proton exchange. Although these residues are predicted to be in the flexible loop connecting β -sheet II and the α -helix, the temperature series results supported by hydrogen-deuterium analysis indicate the presence of persistent hydrogen bonds and strongly suggest their location in the rigid core of the protein.

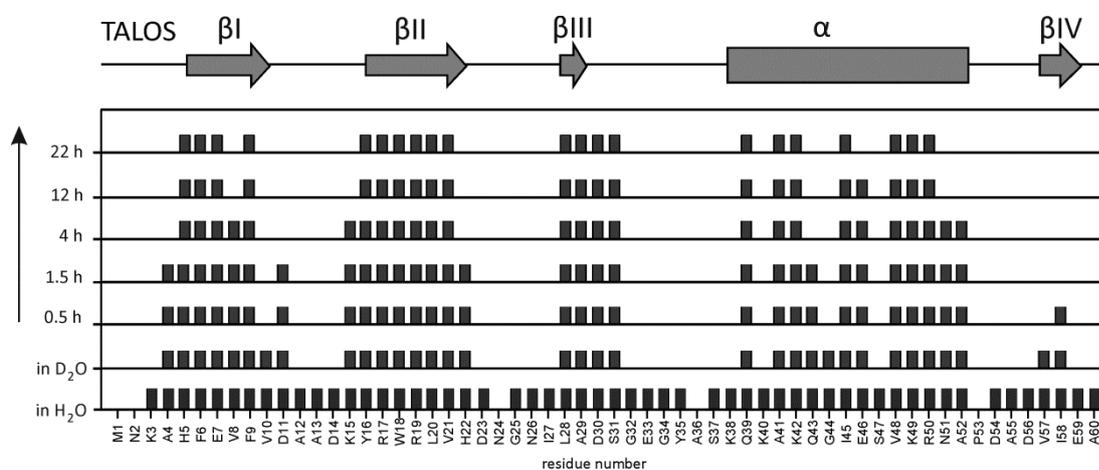


Figure 2.26. Hydrogen-deuterium exchange studies of P17 small protein. Bar chart representation of the visible amide signals based on a series of 2D $^1\text{H}^{15}\text{N}$ BEST-TROSY spectra recorded at 600 MHz, 288 K. On top, the TALOS-N secondary structure prediction is shown.

The amides V57 and I58, which are predicted to construct the β -strand IV, exchange faster with the solvent than the other secondary structure elements. The complete replacement against the deuterium protons is achieved after one hour of reconstitution, resulting in loss of the respective signals in the spectrum.

3.7.3.1.6. Heteronuclear relaxation experiments

Heteronuclear relaxation measurements provide values of T_1 , T_2 and $\{^1\text{H}\}$ - ^{15}N het-NOE relaxation parameters, which are then used to determine the Lipari-Szabo order parameter S^2 . S^2 was calculated based on these values using TENSOR2^[111] software. The N- and C-terminal residues are characterized by low hetNOE and order parameter, indicating high flexibility compared to the more rigid core of the protein (**Figure 2.27**).

The determined experimental rotational correlation time (τ_c) was compared to the predicted τ_c values from HydroNMR^[140] tool. Experimental correlation time is two times higher than the predicted ones, suggesting the protein dimer formation. This is confirmed by the exact match of the experimental data with the predicted correlation time for the dimer conformation. Furthermore, the numerous inter-monomer NOE distance restraints assigned for the structure calculation of a dimer complex support this result.

To investigate the influence of the protein concentration on the tendency to form a dimer complex, two samples with concentrations of 0.6 mM and 5.0 mM were analyzed. The correlation time for both concentrations are in the dimer regime but the slight decrease of the low concentrated sample (7.6 ns) against the high concentrated sample (8.7 ns) might indicate the concentration dependent dimer formation (**Table 5**).

Table 5. Experimental and predicted rotational correlation time (τ_c). For experimental data heteronuclear ^{15}N relaxation experiments ($\{^1\text{H}\}$ - ^{15}N het-NOE, T_1 and T_2) were performed for the protein at two different concentrations. Experimental rotational correlation times (τ_c) were determined using TENSOR2. Predictions of rotational correlation time (τ_c) for monomeric and dimeric protein form were obtained by HydroNMR^[140]. The comparison of experimental data with predicted values shows that the protein forms a dimer.

	NMR (TENSOR 2) 0.6 mM	NMR (TENSOR 2) 5 mM	Prediction (HydroNMR) monomer	Prediction (HydroNMR) dimer
τ_c (ns)	7.6	8.7	4.2	8.4

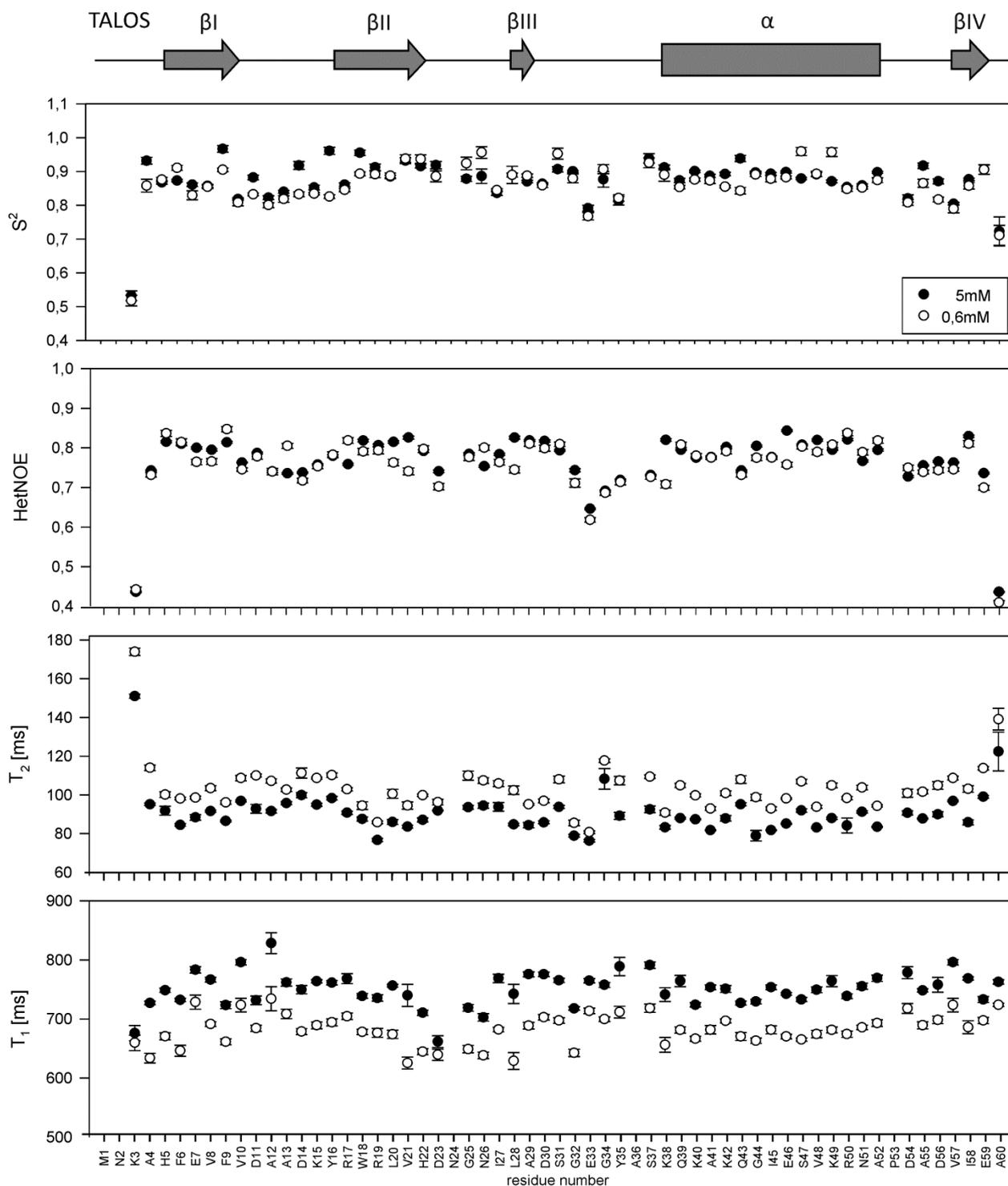


Figure 2.27. Heteronuclear relaxation data of P17 small protein. The determined order parameter (S^2) is shown as a function of the residue number. Analysis was performed using TENSOR2^[111] based on the presented experimentally measured relaxation parameters: $\{^1\text{H}\} - ^{15}\text{N}$ heteronuclear NOE, T_1 and T_2 . Relaxation experiments were performed at 600 MHz, 298 K for two different protein concentrations. 5.0 mM protein concentration is shown as filled circles and 0.6 mM concentration is indicated as unfilled circles. On top, the TALOS-N secondary structure prediction is shown.

3.7.3.1.7. Structure calculation of P17

The structure calculation of P17 was performed with fully automated NOESY cross-peak assignment using CYANA ^[44-46] and refined with ARIA/CNS ^[113] protocol. The backbone and side chain peak list, which was manually assigned to 96.1%, was used as an input for structure calculation. Since it has been shown that the protein adopts a symmetrical dimer conformation with one set of signals for one monomer, the identical peak list was defined for each monomer. The regular weak symmetry restraints were applied. The chemical shift tolerances for the bound and others protons were defined to be 0.015 and 0.020 ppm, respectively, while heavy atoms were set to 0.20 ppm. Unassigned NOE-based distance restraints were obtained from three-dimensional ¹³C and ¹⁵N NOESY NMR experiments by restricted peak picking procedure using SPARKY 3.114 ^[109]. For 3D ¹⁵N NOESY-HSQC the amide assignment was taken from 2D ¹H¹⁵N HSQC, while for 3D ¹³C NOESY-HSQC the 2D ¹H¹³C HSQC assignment was used. Both 3D heteronuclear NOESY-HSQC spectra were recorded with a mixing time of 120 ms. The refinement of the overlaid regions was performed manually. Further, a 2D ¹H¹H-NOESY spectrum with a mixing time of 100 ms was included in the structure calculation, for which the peaks were manually picked. The NOE data was supported by additional information about hydrogen bond distances and dihedral angles determined with the secondary structure prediction TALOS-N tool ^[114].

The structure calculation was performed with CYANA 3.97 using standard protocols with 100 initial and 20 final structures per iteration and 15000 refinement steps. In the last iteration, 200 structures were calculated, used as input for a refinement in explicit water with CNS 1.1 ^[115] using the ARIA 1.2 setup and protocols ^[116]. In the final refinement the ³J_{HNHA} coupling constants and the heteronuclear relaxation data, including T₁, T₂ relaxation time, and rotational correlation time τ_c of 8.63 ns, diffusion anisotropy, anisotropy of 1.20 and rhombicity of 0.60 were included.

The atomic coordinates of the final 20 ensemble of structures with lowest energy were deposited in the protein data bank with the PDB code 6Q2Z. The resonance assignment was deposited in the biological magnetic resonance bank (BMRB) with id code 34334. The structural statistics are represented in **Figure 2.28**. 87.3% restraints are in the most favored region and 12.7% restraints belong to additionally allowed Ramachandran regions. The Backbone RMSD of 0.27 Å and heavy atom RMSD of 0.60 Å were obtained from the ensemble of the 20 lowest energy structures.

Parameter	Value
Experimental restraints: *	
NOE distances	1170
intra residue (i-j=0)	252
sequential (i-j=1)	316
medium range (1<i-j<5)	206
long range (i-j>5)	242
dimer (inter-monomer)	154
Hydrogen bonds	24
intra	19
dimer (inter-monomer)	5
Backbone torsion angles (ϕ, ψ)	110
Diffusion anisotropy (T_1/T_2)	52
Scalar couplings ($^2J_{\text{HNHA}}$)	51
Ramachandran score (%): **	
most favoured	87.3
additionally allowed	12.7
generously allowed	0.0
disallowed	0.0
Average rmsd to mean (\AA): ***	
all atoms	0.78
backbone	0.27
heavy atom	0.60
* per monomer unit	
** for the ensemble of 20 structures, excluding glycine and prolines by using PROCHECK	
*** for the ensemble of 20 structures, including all residues by using MOLMOL	

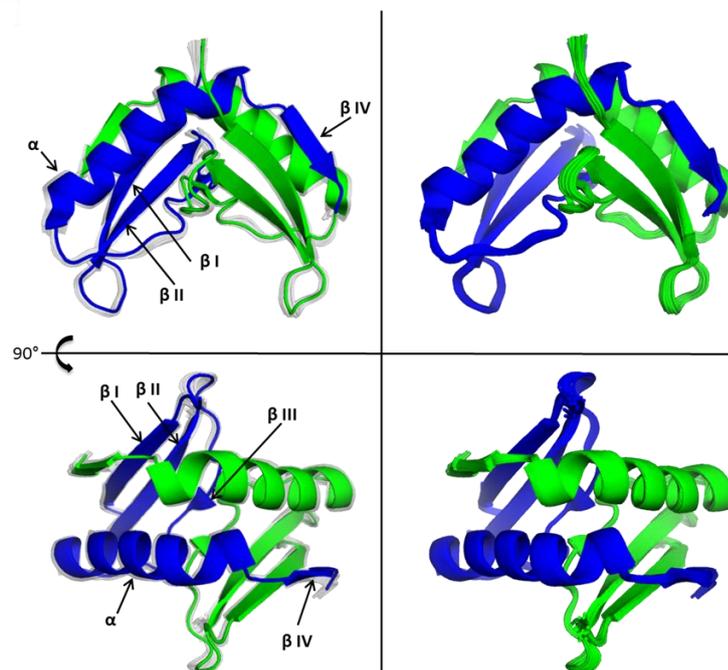


Figure 2.28. NMR solution structure of P17 small protein. Left – structural statistics for the ensemble of 20 best NMR structures of P17. Right – NMR solution structure of P17. On top – the ribbon representation of the lowest energy structure as a dimer (left) and the ensemble of the best 20 structures (right). On bottom – the same protein representation is rotated by 90° along x-axis. One monomer is shown in blue and another one is indicated in green. The secondary structure elements of one monomeric unit are indicated in the figure. The figure was generated by PyMOL.

Figure 2.28 shows the NMR solution structure of P17 as a dimer and the bundle of 20 best NMR structures with the lowest energy. Each monomer consists of four β -strands and one α -helix. The β -strand I involves residues A4 to V10, the β -strand II is built from residue K15 to H22, the small β -strand III in the middle of the structure has two residues A29 and D30 and the C-terminal β -strand IV is formed from D56 to E59. The α -helix consists of residues S37 to D54. The β -strand I undergoes intra-molecular hydrogen bonds with β -strand II, forming a twisted antiparallel intra-monomer β -sheet. The small β -strand III is located in the middle of the dimeric structure and is connected through flexible loop regions to β -strand II and α -helix. It forms an antiparallel intra-monomer β -sheet with β -strand II and an inter-monomer antiparallel β -sheet with the same β -strand III from the second monomer. The C-terminal β -strand IV forms an inter-monomer parallel β -sheet with the β -strand I of the second monomer. In the symmetrical dimer, interactions involve the β -strands III and IV.

The interface analysis of the dimer complex was performed by use of the online PISA tool ^[141]. 40 out of 60 residues (67%) from each monomer are involved in the dimer interface interaction. The protein dimeric complex has a solvent accessible surface area of 6650 \AA^2 , while solvent accessible surface area for one monomer unit is 4919 \AA^2 . The total surface area that is buried upon dimer formation is

calculated as the difference between total accessible surface area of each monomeric unit and interfacing area of the dimer, and has a value of 3188 Å². The dimer surface potential representation is shown in **Figure 2.29**.

The C-terminal β-strand IV includes two aspartates and one glutamate in the sequence and therefore is negatively charged. It undergoes the side electrostatic interactions with a positively charged area of the α-helix residing in the other monomer unit.

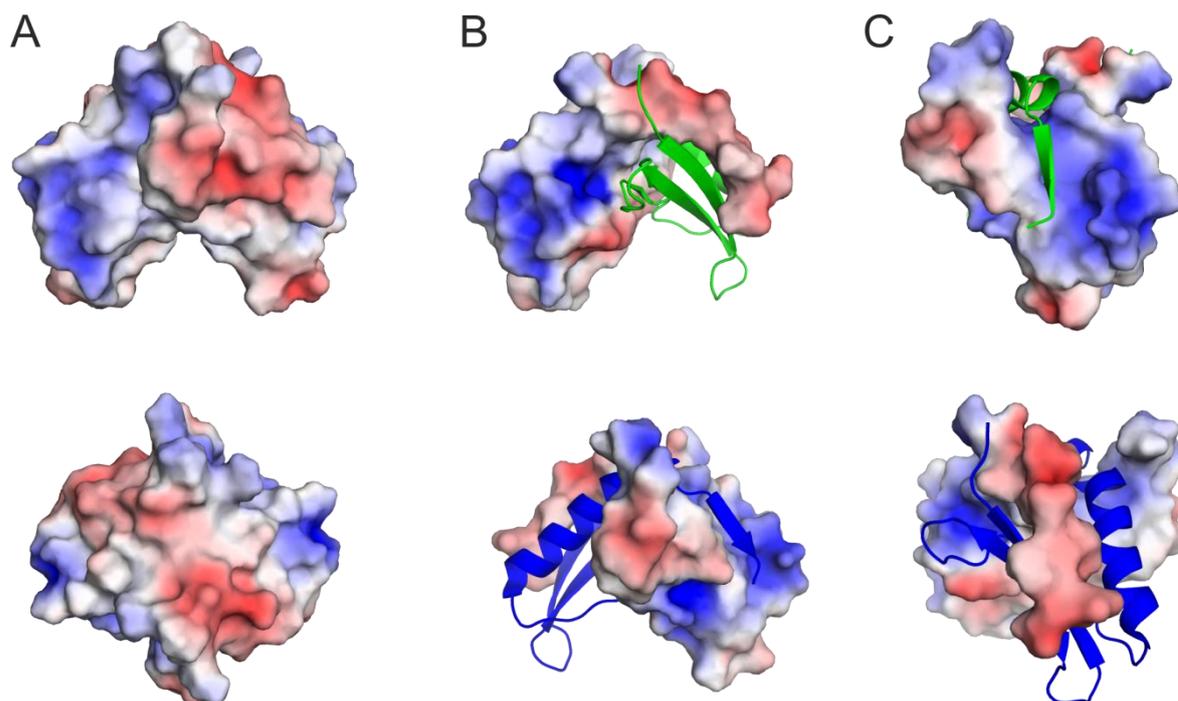


Figure 2.29. Charged surface potential representation of P17 small protein from different perspectives. Red: negatively charged, blue: positively charged, gray: no charge. **A** – Charge distribution of the dimer protein (on the top) rotated by 90° along x-axis (on the bottom). **B** – One monomer unit is shown as surface while the other monomer is represented as a ribbon (in green, on the top) and vice versa (in blue, on the bottom). **C** – The structures of the representation in B rotated by 90° along y-axis. The figure was generated by PyMOL using APBS electrostatic tool.

The residue E59 forms an intermolecular salt bridge with the α-helix residues K38 and K42 stemming from the other monomer, stabilizing dimer formation of the protein. On the other side the β-strand IV forms an intermolecular parallel β-sheet with β-strand I. The residues involved in hydrogen bonds formation are D56, I58 and A60 from the C-terminal β-strand IV and P6 and V8 from the β-strand I. Therefore, the β-strand IV plays an important role in dimer stabilization.

Residues I27, D30 and E33 from the flexible loop region in the center of the complex are also involved in dimer stabilization by constructing a network of intermolecular hydrogen bonds to the same region in the other monomer. The backbone amide of residue I27 forms an inter-monomer hydrogen bond to the sidechain carboxyl of residue E33 from the other monomer. The residue D30 from β-strand III is forming

a small inter-monomer antiparallel β -sheet with its symmetrical residue from the other monomeric unit (**Figure 2.30**).

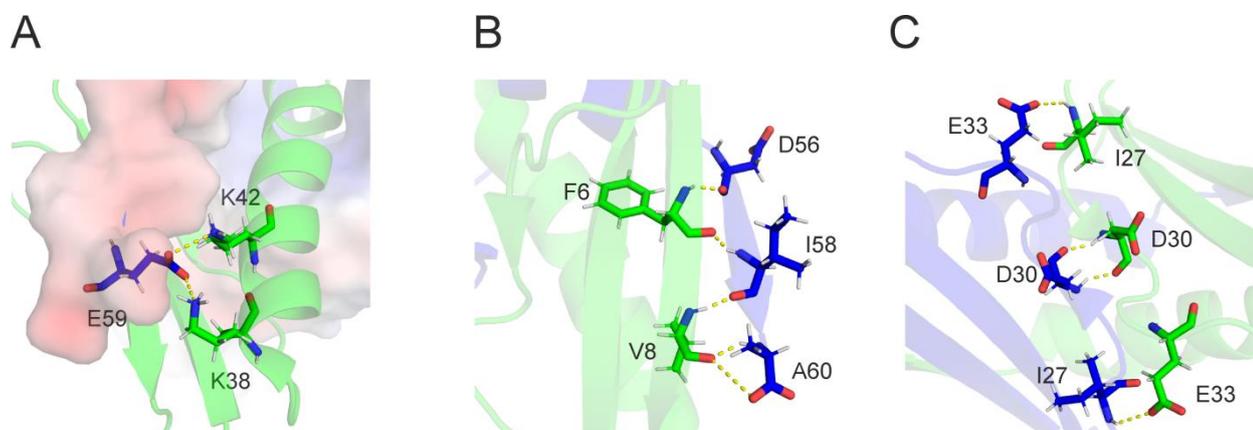


Figure 2.30. Inter-monomer interactions on the dimer interface of P17 small protein. **A** – Salt bridge formation between residue E59 from first monomer and K38, K42 from the second monomer. The involved residues are highlighted as sticks. The first monomer unit is shown as the surface potential while the second monomer is represented as a ribbon (green). **B** – Parallel β -sheet between β -strand IV from the first monomer and β -strand I from the second monomer. The residues involved in hydrogen bond formation are highlighted as sticks. **C** – Hydrogen bonds between residues I27, D30 and E33 within the center of the dimer. The figure was generated by PyMOL using APBS electrostatic tool.

3.7.3.1.8. Structure comparison with homologous proteins

The protein homology modeling analysis using SWISS model tool ^[125–129] was performed using the P17 NMR solution structure as a target. A number of PDB entries of structures with homology sequence part to the target sequence are listed in **Table A9** and displayed in **Figure 2.31**. The sequence identity and similarity in combination with the total coverage are directly related to the calculated structure folding similarity. The structures with the highest coverage ≥ 0.9 (2k7i, 3bid, 2k8e, 2k49) are shown in the first line of **Figure 2.31**. Although the origin of these proteins differs, they all belong to the UPF0339 protein family. The adopted secondary structure is similar to the target, showing the dimer-like fold. These proteins have already been targets of structural genomics consortium with unknown function.

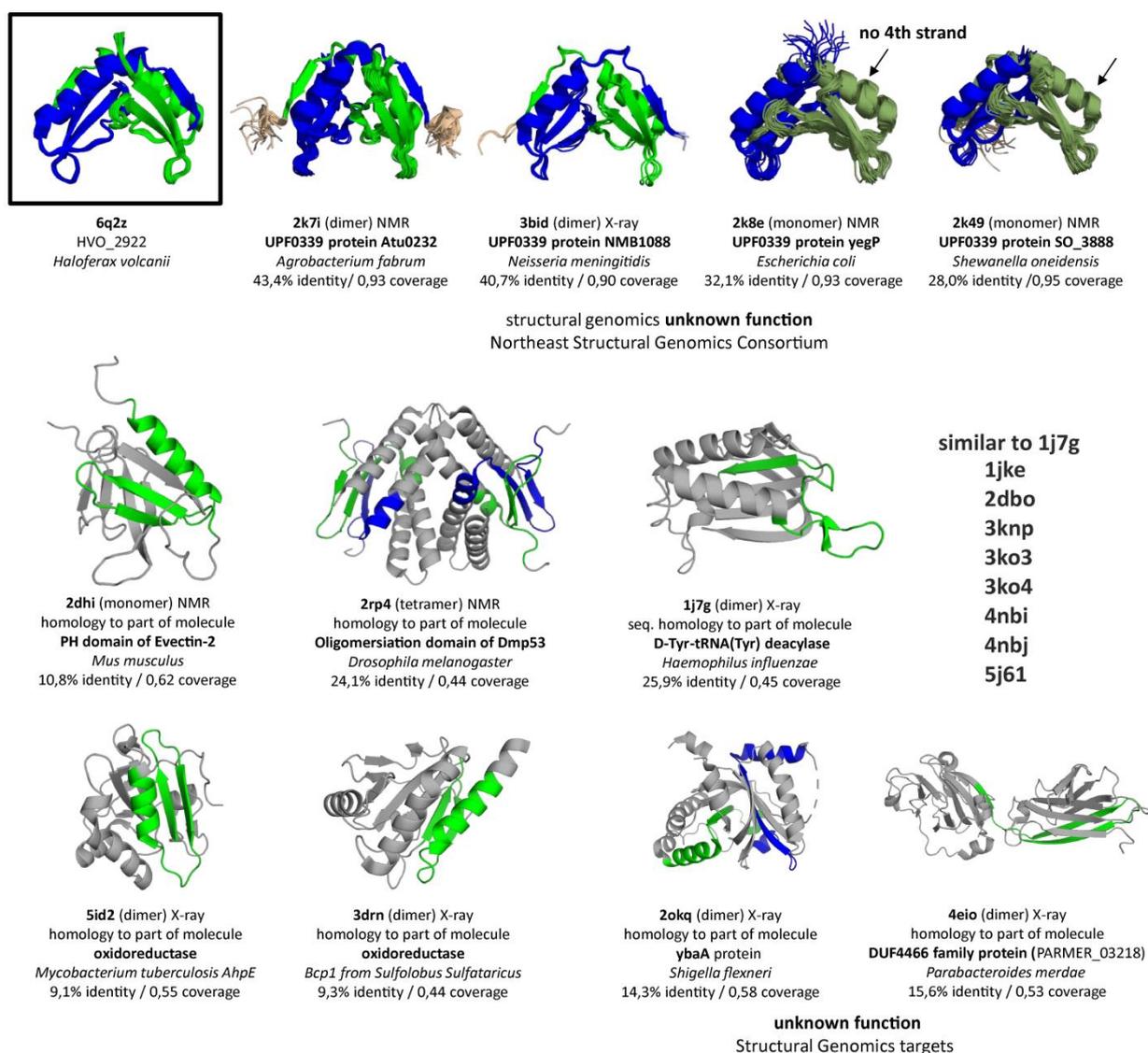


Figure 2.31. PDB structures with the homology part to P17 small protein (alternative name HVO_2922, PDB ID: 6Q2Z). Homology proteins delivered by sequence-based SWISS model secondary structure prediction^[125–129] of P17 is shown. The structure, pdb code, description of the protein, the organisms and the identity and coverage with P17 structure are shown for each SWISS model. The homology to part of the P17 is indicated with the same color code as the calculated structure, one monomer is indicated in blue, the other one is shown in green, the rest of the SWISS model structure which has no similarity with target is shown in gray.

Structures 2k7i and 3bid adopt a very similar dimeric conformation to the target P17 protein. The electrostatic profile of both structures, however, has less charged surface and smaller number of electrostatic interactions on the dimer interface compared to the P17 protein (**Figure 2.32 A-C**).

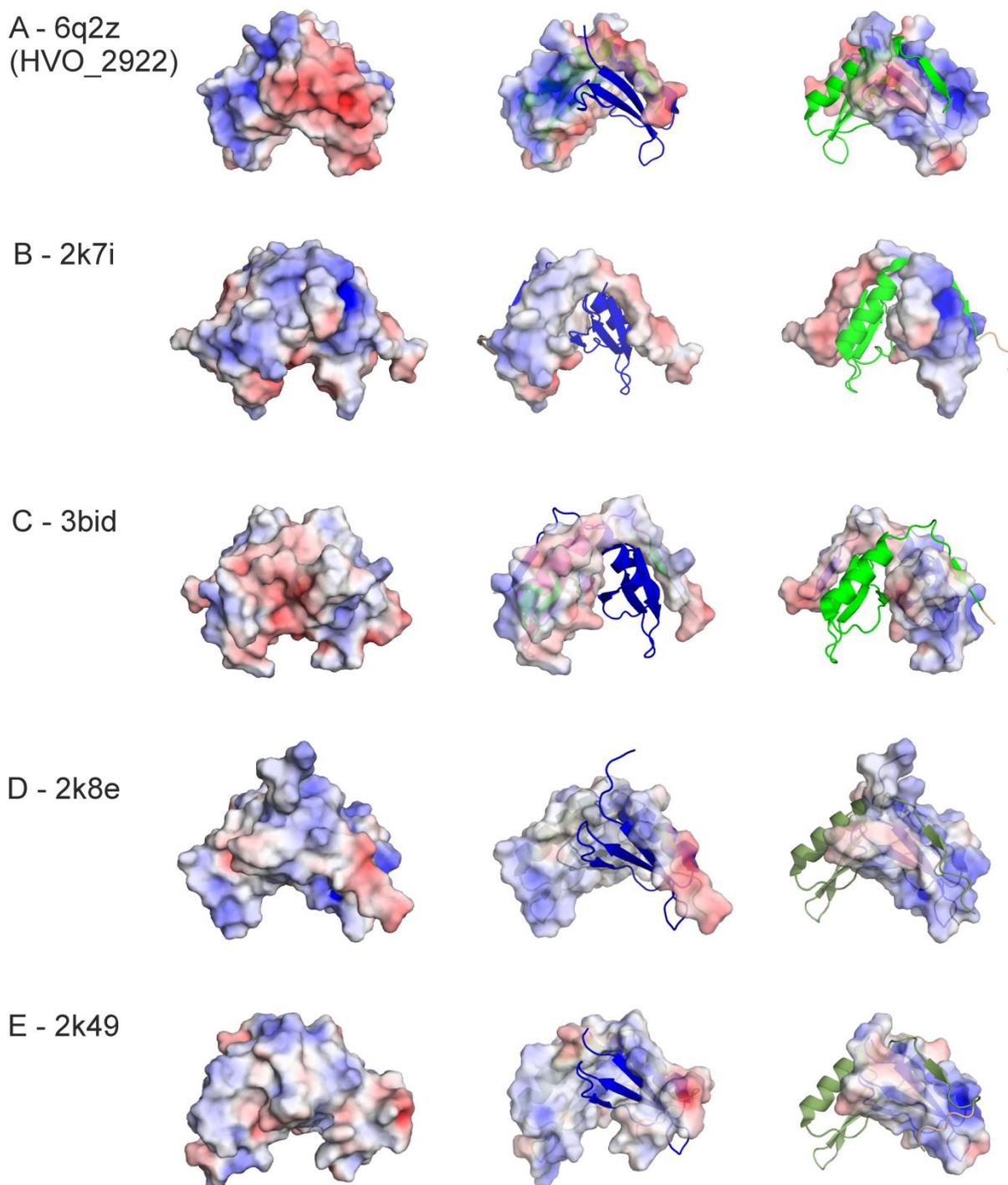


Figure 2.32. A - E – Four structures from the PDB (2K7I, 3BID, 2K8E and 2K49) with the highest sequence identity to the P17 protein (alternative name HVO_2922, PDB ID: 6Q2Z). Surface representation with respect to charges is shown. Red: negatively charged, blue: positively charged, gray: no charge. PDB codes are indicated in the figure. The first monomer unit is shown as the surface potential while the second monomer is represented as a ribbon and vice versa. The figure was generated by PyMOL using APBS electrostatic tool.

The 2k7i structure has a less negatively charged, extended and flexible C-terminus, with the additional residues connected to the β -strand IV. Compared to the P17 protein the 2k7i PDB structure has a higher number of residues, involved in the interface interaction (74% with 46 out of 62 for 2k7i and 6% with 40 out of 60 for P17). This results in the more compact dimeric structure with an increased interface area of 3636 \AA^2 (P17 - 3188 \AA^2). This is supported by the higher number of inter-monomer hydrogen bonds built at the dimer interface. It forms four symmetrical pairs in the center of the complex, while the P17 protein is stabilized only by two hydrogen bond pairs in the protein core between two monomeric units (**Figure 2.33**).

The 3bid structure has the second highest sequence identity and similarity to the P17 protein. The dimeric structure is less compact with the buried interface area of 2900 \AA^2 . Only 60% residues are involved in the interface interaction (35 out of 58). One inter-monomer salt bridge formed between the β -strand IV and the α -helix, and two hydrogen bonds in the middle of the complex are involved in the dimer formation and appear to be sufficient for its stabilization.

In all structures the β -strand IV forms an inter-monomer twisted parallel β -sheet with the β -strand I at the dimer interface. This allows assuming the essential role of these interactions for the dimeric complex formation.

In contrast to the PDB structures described above, the next two PDB structures (2k8e and 2k49) with the highest coverage to P17 are monomers. They consist of two repeated monomeric units, connected with each other via a flexible loop. This allows the protein to adopt a P17 similar dimer-like 3D structure. The N-terminal α -helix of the first unit is directly connected to the C-terminal β -strand I of the second unit, resulting in the lack of the N-terminal β -strand IV from the first repeat. In both monomeric structures (2k8e and 2k49) the electrostatic interactions between the C-terminal β -strand IV and the N-terminal α -helix are less intense. Nevertheless, the β -sheet between two repeats (C-terminal β -strand IV and N-terminal β -strand I) is still formed and stabilize the dimer-like structure. This suggests that the inter-monomer hydrogen bond formation has a higher structure stabilizing impact as the electrostatic interaction.

All other PDB structures can be classified in four groups: oxidoreductases, transcription factors, hydrolases and structural genomics with unknown function. Besides the structures listed above, the structural genomic targets with unknown functions further include a PH domain of Evec tin-2 (2dhi), ybaA protein (2okq) and DUF4466 family protein (4eio). To the oxidoreductases belong two structures 5id2^[142] and 3drn^[143], deacylases involve 1j7g^[144], 2dbo, 1jke^[145], 3knp^[146], 3ko3^[146], 3ko4^[146], 4nbj^[147], 4nbi^[147] and 5j61^[148] PDB entries and one structure (2rp4) reported to be part of an oligomerization domain of Dmp53^[149]. Even though their cellular functions are known, the low sequence similarity and

total coverage compared to the target P17 structure, do not allow a reliable prediction of the possible functional role of the P17 protein.

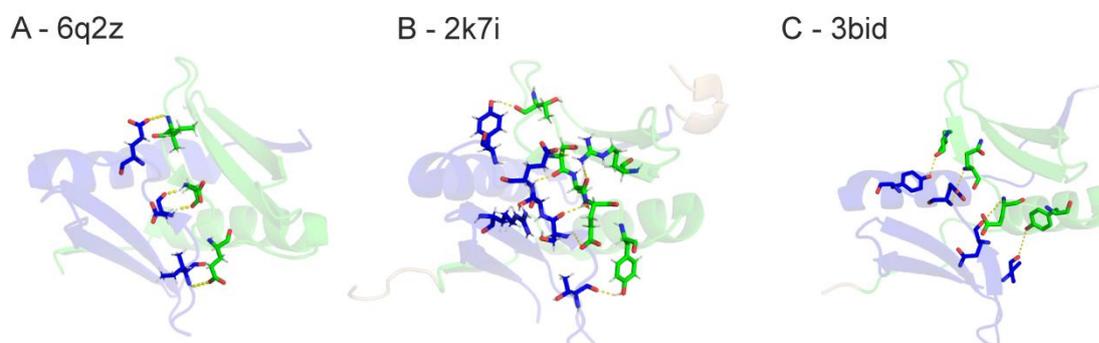


Figure 2.33. Dimer structures from the PDB (2K7I, 3BID) with the highest sequence identity to the P17 protein (alternative name HVO_2922, PDB ID: 6Q2Z). Surface cartoon representation is shown. One monomer is indicated in blue, another one is shown in green; the residues involved in the intermolecular hydrogen bonds in the middle of the complex between two units are highlighted as sticks. The figure was generated by PyMOL.

3.7.3.2. Dimer formation

Dimer formation of P17 was further confirmed by size-exclusion chromatography (SEC) and by Glutaraldehyde-Crosslinking approach performed in Prof. Marchfelder group ^[132]. Size-exclusion chromatography of the P17 protein indicates two prominent signals with retention volumes of 11.8 ml and 13.0 ml in the fraction 1 and 2, respectively (**Figure 2.34**). Calibration occurred according to gel filtration LMW protein calibration kit (Sigma). Retention volumes observed from peak absorbance at 280 nm were calculated to an apparent mass by a standard curve. The fraction 2 with the retention time of about 13.0 ml corresponds to the theoretical molecular weight of 15.9 kDa, which can be identified as the dimeric form of the protein. The theoretical mass in the fraction 1 is about 28.0 kDa, which reflects the oligomerization of the protein. SDS-PAGE of both fractions shows characteristic bands for monomer and dimer species, however, the denaturing conditions of the protein in the SDS-PAGE buffers reduce the relevance of this fact.

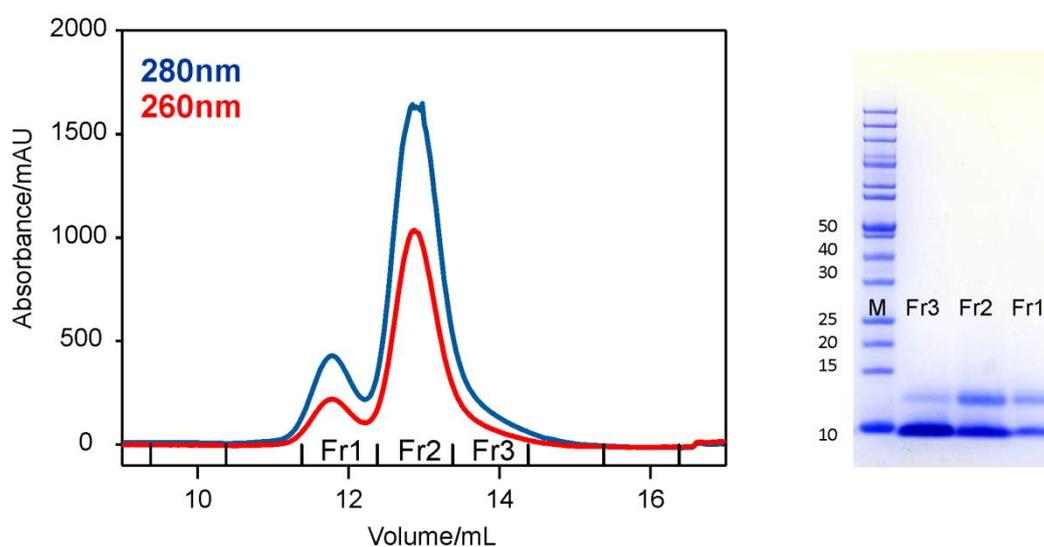


Figure 2.34. Size-exclusion purification of P17 small protein. **Left** – Analytical S75 Tricon size-exclusion chromatogram (SEC) (after NiNTA purification). The collected fractions (SEC) are shown on the x-axis. **Right** – the corresponding SDS-PAGE analysis. SDS-PAGE Marker (*PageRuler™ Unstained Protein Ladder, ThermoFisher Scientific*) (M) was used. Visualization was accomplished by coomassie staining.

To investigate the dimerization process of the P17 protein, site-specific mutagenesis and deletion experiments at the dimer interface were performed. As it has been shown in the SWISS model analysis, the negatively charged C-terminal β -strand IV of the first monomeric unit is involved into electrostatic interactions with the α -helix from the second unit and further it constructs an inter-monomer β -sheet with the β -strand I of the second unit. Based on the structural comparison and analysis of homologous proteins, the hydrogen bonds forming the inter-monomer β -sheet have a higher impact on the dimer formation than the electrostatic interactions.

The residue I58 in the center of the β -strand IV from the first monomer unit forms hydrogen bonds to the residues F6 and V8 of β -strand I from the second monomer unit. In order to disturb the monomer-monomer interaction and therefore dimerization of the protein, the I58 was mutated to a proline. This prevents the intermolecular hydrogen bond formation between I58 amide and F6 carbonyl from β -strand I of the second monomer unit.

In order to investigate the role of the whole β -strand IV on the protein dimerization, additional deletion mutant experiments were performed. For these the last four N-terminal residues from V57 to A60, which are involved in the β -strand IV, were removed.

3.7.3.2.1. NMR spectroscopy of mutants

The I58P mutation destabilizes the protein and leads to partial sample degradation. Two new formed tryptophan indole side chain signals with different intensities in the 2D $^1\text{H}^{15}\text{N}$ BEST-TROSY spectrum indicate the formation of additional conformational states. This is supported by newly appeared, broadened and poor resolved signals in the center of the spectrum, which are characteristic for a molten globule state (**Figure 2.35A**). The highest CSPs are observed for residues located in the dimer interface. A4 and F6 residues from the β -strand I are involved in the β -sheet with β -strand IV and therefore undergo mutation induced conformational changes. The other three residues D23, G25 and N26 with significantly high CSPs are involved in the loop region connecting β -strand II and β -strand III. The loop is located in the dimer interface and any conformational changes at this position lead to the dimer destabilization. For the further analysis heteronuclear relaxation experiments of the I58P mutant were performed (**Figure 2.35**). The calculated experimental correlation time for the I58P mutant (6.9 ns) is in the same dimer-like range as the correlation time for wild type with the same concentration (7.6 ns). The slight decrease in the τ_c value of the mutant compared to the wild type indicates stability decrease of the protein. The dimer complex formation is anyhow not completely inhibited.

The importance of the negatively charged β -strand IV on the dimeric structure was investigated by removing the last four residues, which are part of the inter-monomer β -sheet. During the purification steps, the cleaved deletion mutant was partially interacting with the SUMO-fusion by forming a complex. This indicates the striving of the target protein for an external stabilizing interaction partner. The isolated deletion mutant unfolds and aggregates (**Figure 2.35B**). This result confirms the suggestion that the β -strand IV plays an important role in the β -sheet formation and therefore dimerization of the protein.

However, these mutations seem to have no significant effect on the growth profile of the cells, which might indicate that dimerization is not essential for functionality^[132]. One possible scenario could be the binding of the protein to an unknown interaction partner followed by the dissolving of the dimeric structure or the monomer form of the protein might be functional as well.

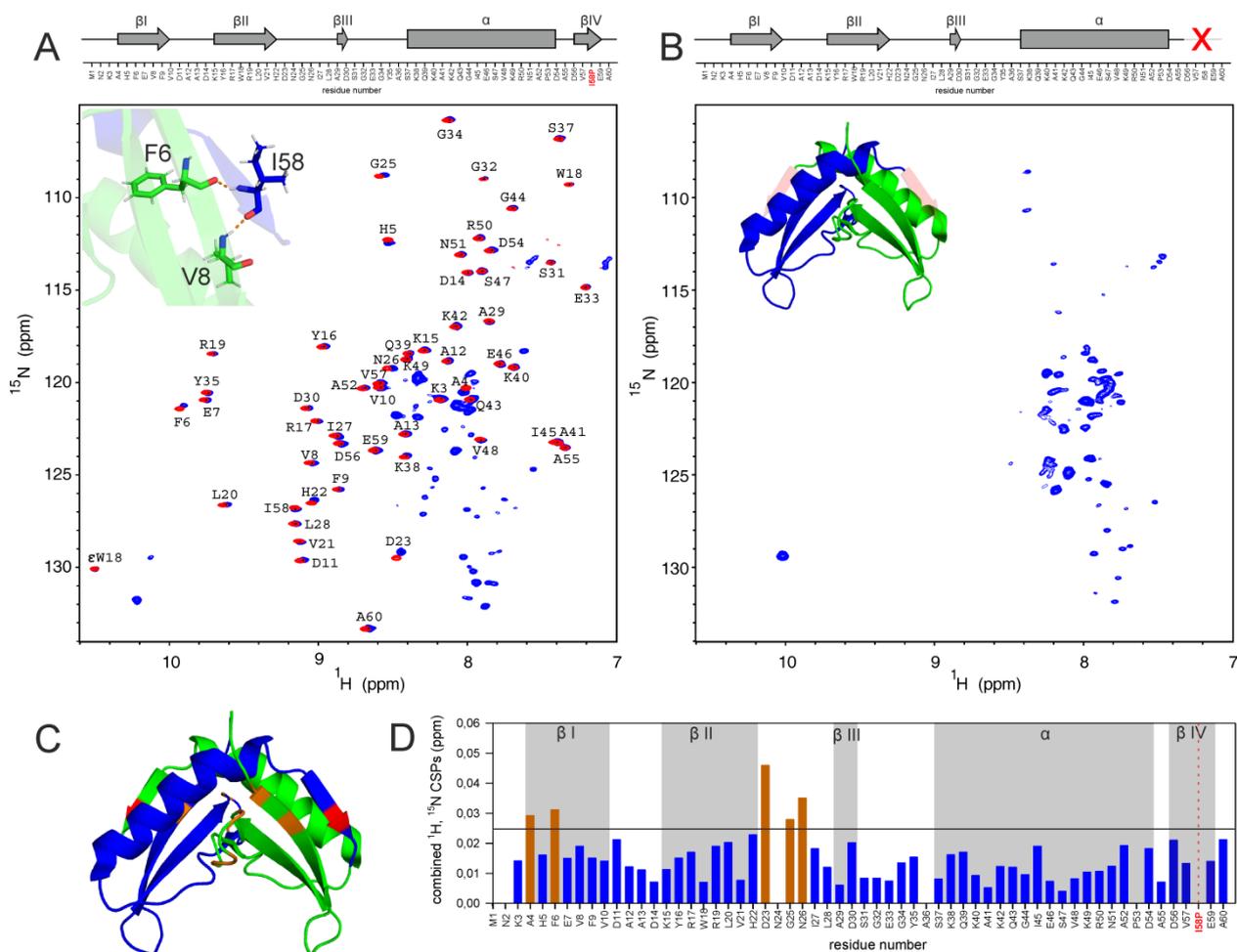


Figure 2.35. NMR structural investigations of mutants. **A** – I58P mutant. Interaction at dimer interface of wild type is shown. One monomer is indicated in green and the other is shown in blue. I58 is indicated in blue and F6 and V8 are colored in green. The figure was generated by PyMOL. On the top, a schematic representation of secondary structure elements based on CYANA structure calculation of wild type (mutated position is highlighted in red) is shown. On the bottom $^1\text{H}^{15}\text{N}$ BEST-TROSY spectra of the wild type P17 (red) and I58P mutant (blue) are overlaid. Both spectra were recorded at 600 MHz and 298 K. **B** – C-terminal deletion mutant. The wild type solution structure is shown. One monomer is indicated in green and the other is shown in blue. The last four deleted amino acids, which would form β -strand IV, are colored in red. The figure was generated by PyMOL. On the top, a schematic representation of secondary structure elements based on CYANA structure calculation of wild type (mutated position is highlighted in red) is shown. On the bottom $^1\text{H}^{15}\text{N}$ BEST-TROSY spectrum recorded at 600 MHz and 298 K is displayed. **C** v Mapping of CSPs of I58P mutant on the wild type solution structure. CSPs > 0.025 ppm are colored in orange and the mutated amino acid I58P is colored in purple. The figure was generated by PyMOL. **D** – Combined ^1H and ^{15}N CSPs of the backbone amides in ^1H ppm as a function of the I58P mutant residue number¹.

¹ $\Delta\delta = \sqrt{(\Delta\delta_H)^2 + (\alpha \cdot \Delta\delta_N)^2}$, ($\alpha = 0.1$).

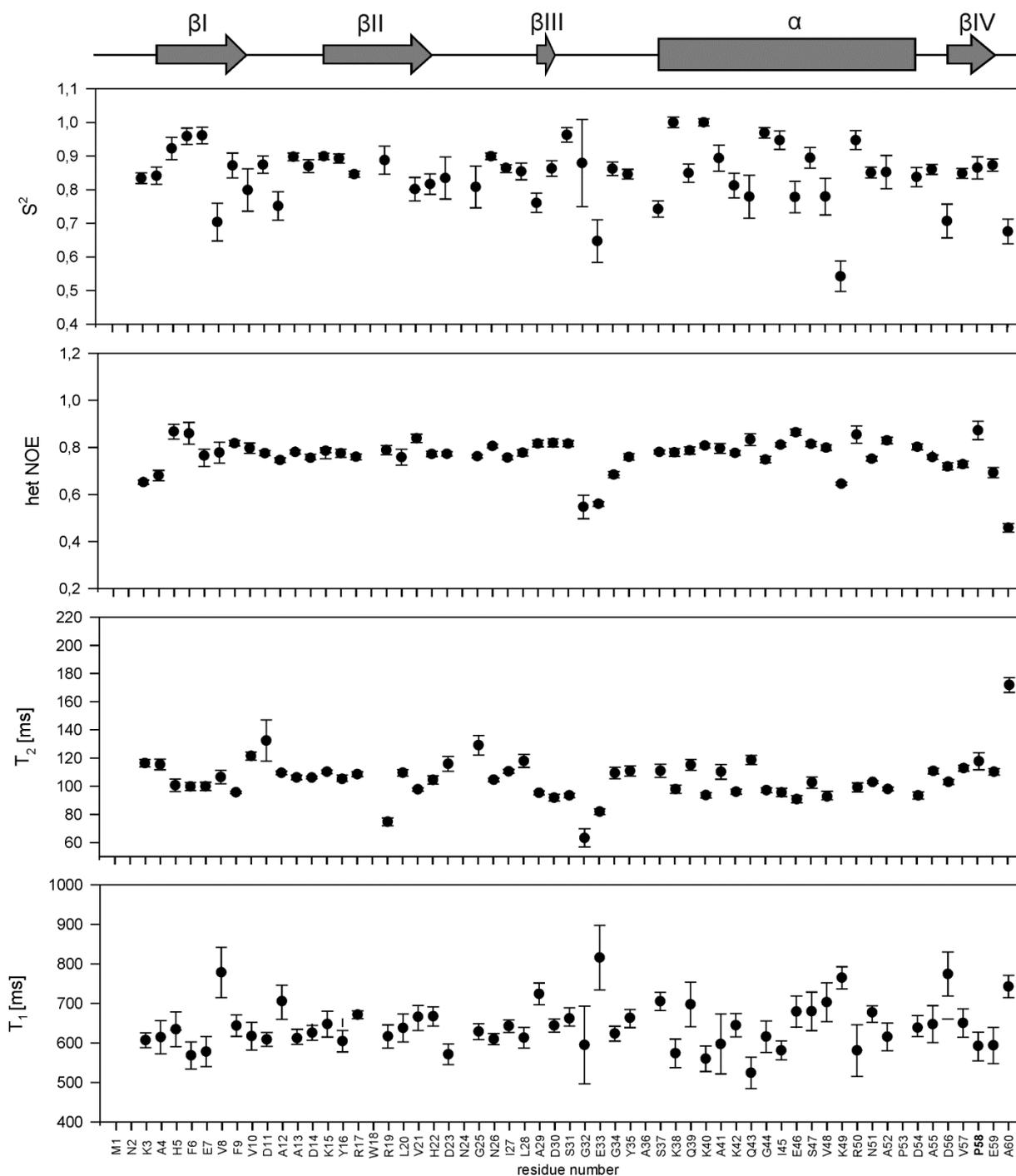


Figure 2.36. Heteronuclear relaxation data of I58P mutant. On the top, a schematic representation of secondary structure elements based on CYANA structure calculation is shown. Backbone dynamics of I58P mutant (0.6 mM) in 50 mM sodium phosphate buffer pH 7.5, 100 mM NaCl, 95% H₂O/ 5% D₂O measured at 600 MHz 298 K are shown and plotted as a function of the protein sequence. The determined order parameters (S^2) are shown on the top. Calculation was performed using TENSOR2^[111] software based on the experimentally measured relaxation parameters: ¹⁵N heteronuclear NOE, T_1 and T_2 . Shown is a plot of T_1 , T_2 relaxation times and $\{^1\text{H}\} - ^{15}\text{N}$ hetNOE as a function of residue number. The experimental rotational correlation time (τ_c) calculated using TENSOR2 program is 6.9 ns.

3.7.3.2.2. Concentration dependent dimer formation

The heteronuclear relaxation experiments of the wild type discussed in **chapter II 3.7.3.1.6** indicate a slight tendency of concentration dependent dimer formation. This is supported by the mass spectrometry experiments performed on the samples with two different concentrations. At low P17 protein concentration (15 μM) (**Figure 2.37B**) only one signal corresponding to the molecular weight of one monomer unit is observed in the MALDI spectrum, while the higher concentrated sample (80 μM) shows an additional signal at the molecular weight of two monomer units (**Figure 2.37C**). However, the 2D NMR experiments do not confirm this result. 2D ^1H - ^{15}N BEST-TROSY spectra of the 15 μM and 80 μM concentrated samples show only slight chemical shift perturbations in the N-terminal residues K3 and A4 (**Figure 2.37A**). Overall, no significant chemical shift changes of the amide protons could be observed at these sample conditions, suggesting preferable fully symmetrical dimer formation.

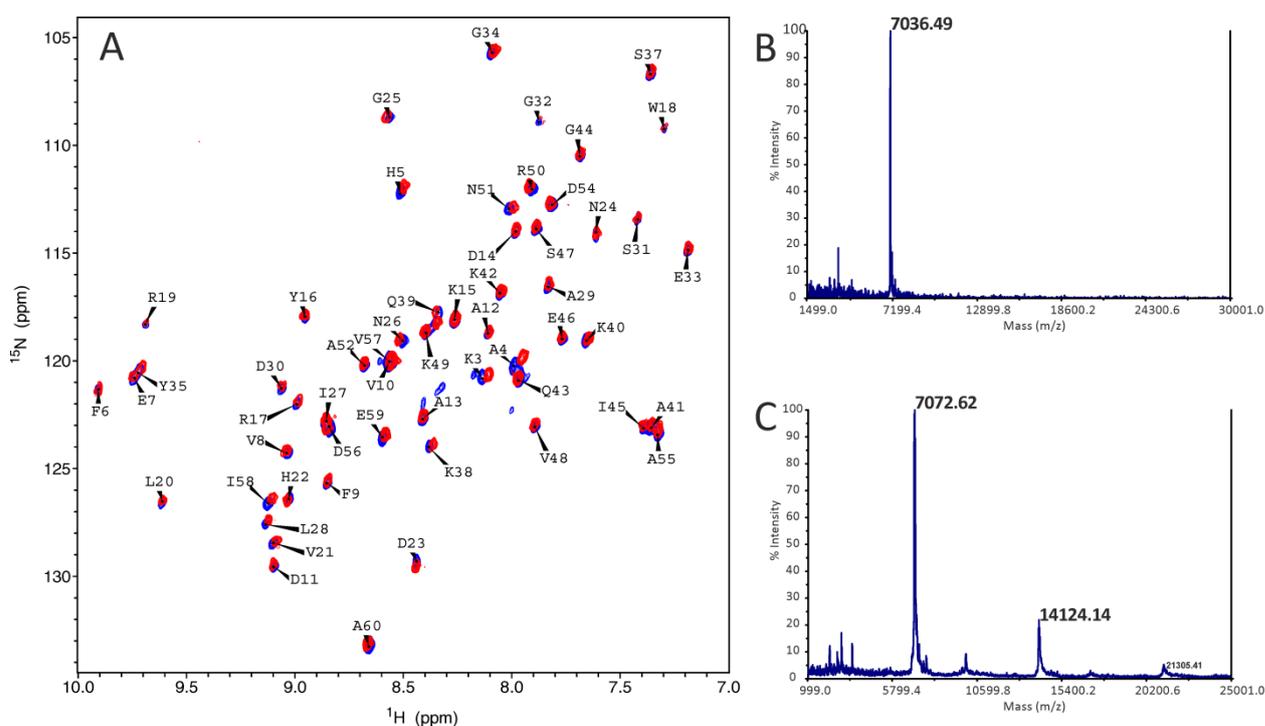


Figure 2.37. Concentration dependent dimer formation. **A** – 2D ^1H - ^{15}N BEST-TROSY NMR spectra of P17 small protein recorded at 600 MHz, 298 K. Shown is an overlay of two spectra of different concentrated samples. 15 μM protein is indicated in red and high concentrated protein (80 μM) is highlighted in black. **B**, **C** – MALDI spectra of P17 peptide at two different concentrations are displayed. **B** – Low concentrated sample (15 μM) shows only one signal corresponding to the monomer molecular weight. **C** – Protein concentration increase leads to a dimer formation, which is indicated as a second signal in MALDI spectrum at 80 μM protein concentration.

3.7.3.3. Discussion

The high-resolution NMR solution structure of the P17 protein was performed by using 3D heteronuclear NMR experiments. The bundle of 20 top-ranked structures with lowest energy were deposited in the protein data bank (PDB, code 6Q2Z). The resonance assignment was deposited in the biological magnetic resonance bank (BMRB, id code 34334). Based on the heteronuclear relaxation data supported by the intermolecular NOEs assigned for the dimer complex, it has been shown that the protein adopts a dimeric conformation. The dimer formation is characterized by high stability and is partially concentration dependent. The mutagenesis and deletion experiments at inter-monomer β -sheet dimer interface prove the essential role of the dimerization for the stability of the protein.

In order to clarify the biological function of the small protein and in particular the role of the dimer formation, biological experiments were conducted. The growth experiments of the *Haloferax volcanii* performed by Prof. Marchfelder group indicate a clear phenotype for the wild type P17 small protein. However, dimerization hampering mutants have no influence on the growth process of *HVO* ^[132]..

3.8. Timesaving methods

3.8.1. Automated FLYA assignment: proof of principle

In order to speed up the structural analysis of small proteins the automated FLYA assignment ^[43] method was applied on the P17 protein. Generally, the fully computational FLYA assignment is based on the peak lists from the multidimensional NMR spectra and allows rapid, automated chemical shift assignment, which can be used for further structure calculation by CYANA ^[44-46]. The same number of NMR experiments used for manual assignment was included for generation of the FLYA input data. The manually determined chemical shift assignment was used as a reference for FLYA automation. The color code in the graphical representation of the FLYA assignment output provides information about the quality of the assignment and the agreement with the assignment used as a reference. Most of the atoms (indicated in green) are in good agreement with the manually determined resonance assignment. The automated backbone assignment shows a high match with the reference ones, while the side chain resonances are more sensitive to false assignments and need to be verified manually. Since the backbone FLYA assignment provides a fast and high-quality result, it was used for further tests of timesaving structural analysis.

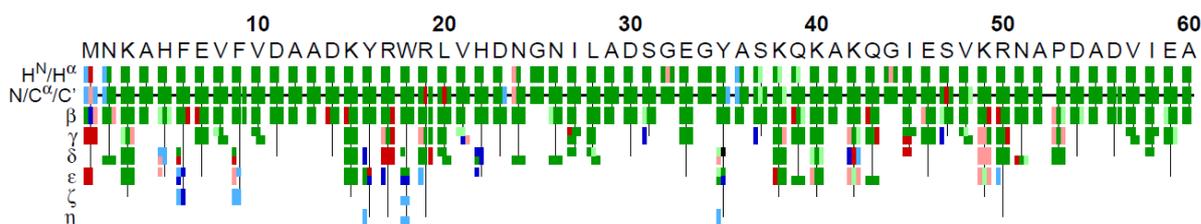


Figure 2.38. Graphical representation of automated completed FLYA assignment of P17 with a manually determined reference assignment. Green indicates the FLYA assignment in good agreement with the reference one, red shows atoms which do not agree with the reference, blue markers represent the FLYA assignment with no reference assignment used and black color is used for the atoms assigned in the reference but not assigned by FLYA. The assignment which is not determined to be strong is highlighted in the corresponding light colors.

3.8.2. Targeted acquisition technique

The second approach to reduce the time needed for the structural analysis is to speed up the experimental NMR measurement time. Non-uniform sampling (NUS) is already a standard method to record multidimensional NMR spectra. Its combination with targeted acquisition (TA) approach and Multi-Dimensional Decomposition (MDD) signal processing technique ^[39,40] provides the possibility to save measurement time without losing quality of the spectra. The comparison analysis of the quality of the recorded spectra was performed by the FLYA automated assignment.

The backbone assignment of P17 protein was performed using 2D $^1\text{H}^{15}\text{N}$ HSQC, 2D $^1\text{H}^{13}\text{C}$ HSQC and 3D heteronuclear HNHA, HNCOC, HN(CO)CACB, HNCACB experiments. The total experimental time using the conventional methods was 5 days. 93.15% of the backbone atoms were assigned with the score of 0.751 (Figure 2.39).

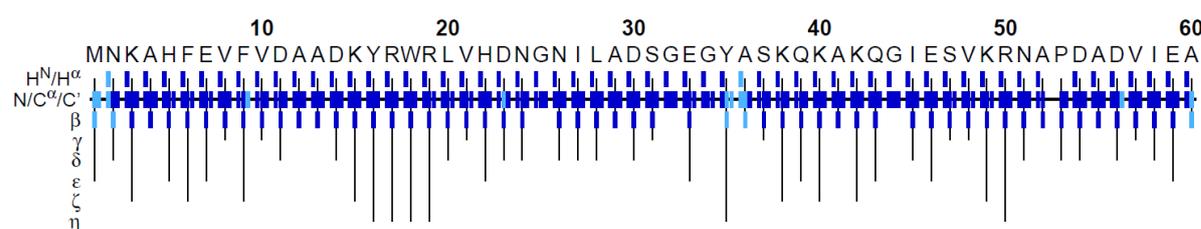


Figure 2.39. Automated backbone FLYA assignment of P17 based on spectra recorded with conventional 3D NMR methods. Total measurement time is 5 days. Assigned 93.15%, Score 0.751.

The set of spectra, recorded with NUS and targeted acquisition technique, include the following 3D heteronuclear NMR experiments: HNCOC, HN(CO)CA, HNCA, HN(CO)CACB, HNCACB, HN(CA)CO. Monitoring the number of signals appearing in real-time showed that after recording of 6 TA steps corresponding to 6% NUS amount, the peak amount and therefore the quality of the recorded spectra do not change significantly anymore. This allowed to stop the spectra recording after 4.5 hours of measurement time.

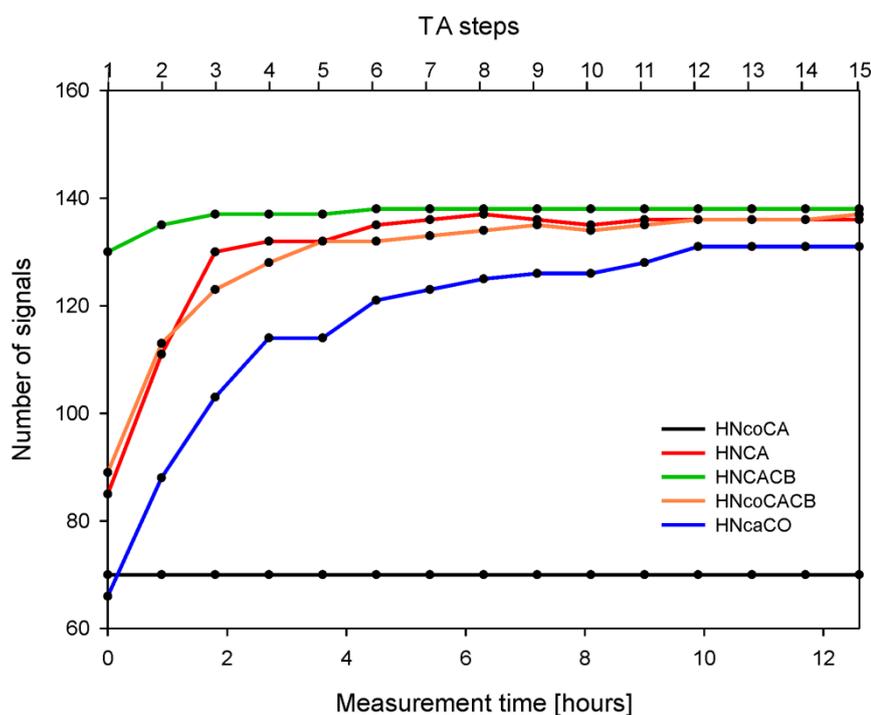


Figure 2.40. Progress of targeted acquisition against measurement time of P17 protein.

The resulting backbone FLYA assignment was performed to 90.29% with the score of 0.728. The highly reduced measurement time and recorded number of signals shows a strong potential of this method for the future structural analysis of small proteins.

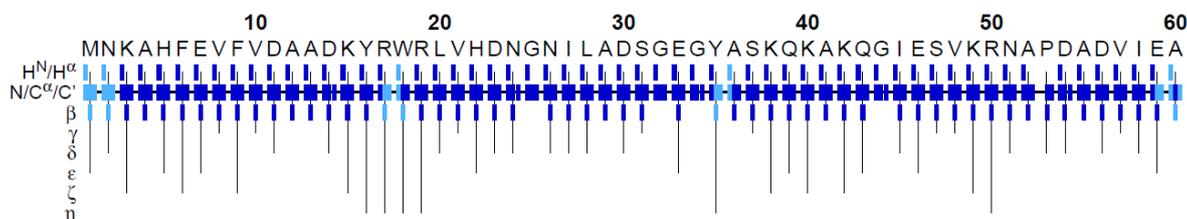


Figure 2.41. Automated backbone FLYA assignment of P17 based on spectra recorded with 6% NUS amount and processed with targeted acquisition (TA) technique. Total measurement time is 4.5 hours. Assigned 90.29%, Score 0.728.

The targeted acquisition in combination with the Multi-Dimensional Decomposition signal processing technique applied on non-uniform sampled data allows an enormous reduction of the NMR measurement time. The implemented automated FLYA assignment simplifies the evaluation of the spectra and speeds up the structural screening of small proteins.

4. Conclusions and perspectives

Increasing scientific interest in previously ignored coding small open reading frames (sORFs) accomplished with newly developed and optimized computational and experimental approaches allowed the identification of hundreds of small proteins. Elucidation of their functions in the cell is the next step in the global biological annotation. This task remains challenging and is currently the main subject of many research groups. In the context of structure-function-relation, determination of the conformational properties of small proteins provides additional important information. Solution NMR spectroscopy, as the most powerful tool for structural characterization of proteins, is perfectly amenable for conformational analysis of small proteins.

During this doctoral thesis, 20 small proteins from bacterial and archaeal organisms were structurally investigated. The workflow protocol for fast screening of small proteins including sample preparation and biochemical characterization was established and applied to these 20 candidates. Peptides and small proteins ranging from 14 to 71 amino acids were identified and selected by six microbiological groups, all members of the Priority Program on small proteins (SPP2002) funded by the German research foundation (DFG). Three possible conformations were observed during the structural screening: 1) random coil state – proteins that are entirely unstructured 2) molten globule state - proteins with fluctuating amount of residual structure but no defined tertiary fold and 3) structured proteins with clearly defined tertiary structure. Due to issues with stability and high hydrophobicity, five small proteins were excluded from the fast secondary structure screening. Among the remaining 15 candidates, 8 peptides adopted an unstructured fold, 4 small proteins were in the molten globule conformation and 3 small proteins were found to be structured. For the most promising candidates, screening of the folding conditions by NMR spectroscopy was performed and the optimal ones were used for further structural and dynamic characterization.

Two small proteins (P17 and P19), both from *H. volcanii*, were extensively investigated for their structural and dynamic behavior. For the 60 amino acids long small protein P17, a high-resolution NMR solution structure was calculated. The 20 top-ranked structures with lowest energy were deposited in the protein data bank (PDB, code 6Q2Z) and the NMR chemical shift assignment was deposited in the biological magnetic resonance bank (BMRB, id code 34334). According to the relaxation data and the intermolecular NOEs the protein forms a symmetric dimer. Mutational studies conducted on the dimer interface had the highest impact on dimerization and on the stability of the protein. In combination with a biological study of the growth trend of wild type and mutants in response to different conditions, performed in Prof. Marchfelder group, these data were recently published in the scientific journal *ChemBioChem* with the topic “Solution structure and dynamics of the small protein HVO_2922 from *Haloferax volcanii*”^[132].

NMR structural analysis of folded zinc-finger protein P19 shows a dynamic equilibrium of two long-lived states with a population ratio of 4:1 at room temperature (298 K). The structural heterogeneity is temperature dependent with a major conformation preferably stabilized at 278 K. Incorporation with zinc ions is crucial for the protein folding, removal of zinc with EDTA leads to the formation of an unfolded metal-free state. In addition, it was shown that despite the two zinc-binding motifs in the sequence, only one zinc ion is incorporated into the protein. Structure calculation and mutagenesis experiments in order to identify the correct binding pocket are a subject for further studies.

Furthermore, within the scientific collaboration with Vladislav Orekhov (Sweden), new NMR technology, which reduces the machine measurement time, was successfully implemented and is now routinely used in Prof. Schwalbe group for structural characterization of small proteins. This method is based on the non-uniform sampling (NUS) approach in combination with targeted acquisition (TA) and a new Multi-Dimensional Decomposition (MDD) signal processing technique and allows recording a set of 3D heteronuclear NMR experiments for the chemical shift assignment within hours. In combination with automated resonance assignment (FLYA) this approach gives an advantage for the structural screening of a large number of small proteins.

NMR structural screening of peptides and small proteins analyzed in this doctoral thesis was summarized and recently submitted as a review article in the *Angewandte Chemie* scientific journal. All peptides and small proteins, screened during this doctoral thesis, were investigated in their isolated form. All analyzed peptides (14-31 amino acids) were found to be unstructured, while small proteins (38-71 amino acids) adopt a folded or partially folded conformation. This clearly indicates the strong influence of the sequence length on the conformational behavior of the protein. The small size generally restricts the ability to adopt secondary structure. Bioinformatics prediction tools, performed in collaboration with Monika Fuxreiter (Hungary), reported the high potential of these peptides to adopt a folded conformation upon complex formation. Thus, the correct folding, which is likely needed for the biological activity of the peptides, is induced by the interaction with the bigger biomolecules (DNA, RNA, proteins etc.). This assumption leads to the next stage of the studies of small proteins, where the focus should be on the interaction with binding partners and in particular on the conformational changes induced by this interaction. The exploration and search for the potential interaction partners for the small proteins of interest is already the subject of different microbiology groups from the Priority Program on small proteins (SPP2002). In the following studies, NMR spectroscopy will be applied for the structural characterization of these biologically active complexes.

Chapter III

Light dynamics of retinal disease relevant G90D bovine rhodopsin mutant

The following colleagues contributed to this chapter: Dr. Jiafei Mao from Prof. Clemens Glaubitz group (Institute for Biophysical Chemistry, Goethe-University Frankfurt) performed solid state NMR experiments and Dr. Elias Eckert from Prof. Josef Wachtveitl group (Institute for Physical and Theoretical Chemistry, Goethe-University Frankfurt) performed flash photolysis and ultrafast absorption experiments. Detailed analysis of kinetic processes is described in Elias Eckert's doctoral thesis. Stably transfected HEK293 cells of the stabilized wild type and the stabilized G90D mutant were kindly provided by Prof. Joerg Standfuss (Paul Scherrer Institute, Switzerland).

1. General introduction

1.1. Sense of sight and structure of the eye

Sense of sight is the most important human sense, which has a direct impact on the quality of life. The process, which transforms the electromagnetic waves from the light into electrical signals occurs in the eye. In general, light enters the eye, gets focused by the cornea and lens, hits the back of the eye, which is called retina, and gets converted into a neural impulse, which is then transported via the optic nerve into the brain where the final image is formed (**Figure 3.1**).

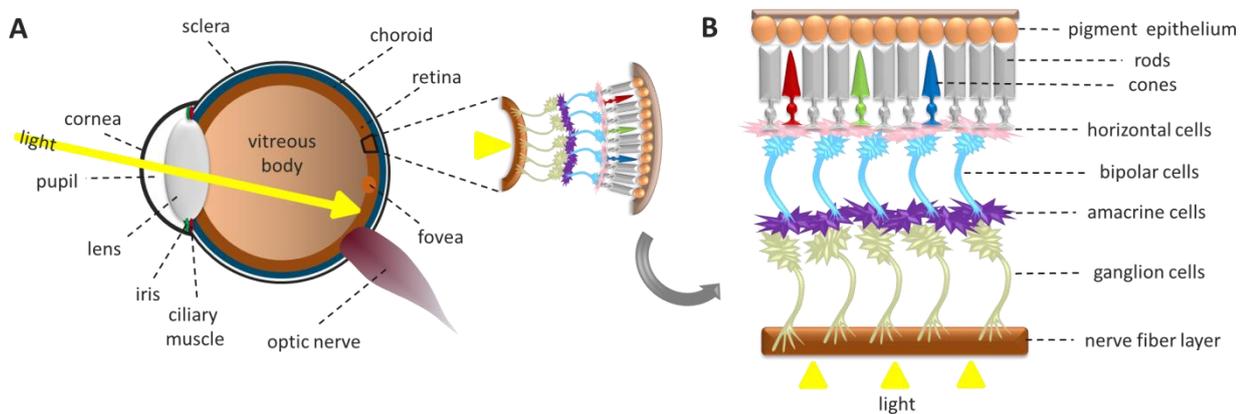


Figure 3.1. Schematic representation of the structure of the vertebrate eye. Adapted from Roy, 2010 ^[150].

The retina is composed of different cells, which respond to light. Photoreceptors are a special type of cells where the absorbed photons of light are transformed into the neural impulse. Originated from their shapes, the photoreceptors are divided into cones and rods. Rods are very sensitive to dim light and are responsible for night vision, while cones allow the perception of color and details in normal bright light conditions. Depending on the absorbed light cones are classified into red, green and blue types. There are around 120 million rods per retina in the human eye, mostly located on the periphery of the eye and only 6 to 7 million cones, many of them are focused in the center of the eye, called fovea. Both, rods and cones consist of inner segments, outer segments and synaptic ends. The inner segments include nuclei and mitochondria region, while the outer segments are filled with hundreds of membranous discs containing visual pigments. In rods, these discs are stacked on top of each other in a cylindrical shape and in case of cones they form a conical shape, which is shorter compared to rods (**Figure 3.2**). Prolonged form of the outer segment of the rods is responsible for higher amount of visual pigments, which absorb more photons and lead to 70 times higher sensitivity compared to cones. Additionally, the light-induced photo transduction process is faster in the rods compared to cones, which also contributes to the sensitivity of rod cells ^[151].

In the rods, the discs are located within the cytoplasm with no contact to the plasma membrane. While in the cones, the discs are invaginations of the plasma membrane. The main component of the membrane discs in the rods is the visual pigment rhodopsin. In the cones, three types of visual pigments called photopsin or iodopsin are present. Besides rhodopsin, the membrane of the discs contains other additional proteins called G protein (transducin) and phosphodiesterase (PDE), which are directly involved into the transduction cascade and are highly important for the visual photocycle.

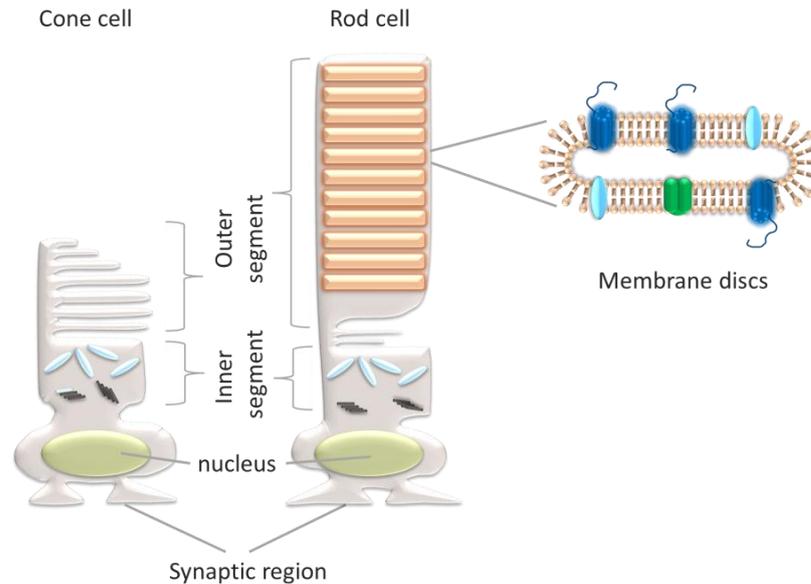


Figure 3.2. Schematic representation of rod and cone photoreceptors. Adapted from Ingram, 2016 ^[151].

1.2. Phototransduction

Phototransduction takes place in the outer segment of the photoreceptors. This process is responsible for converting light, an electromagnetic wave signal (stimulus), into a neural (electrical) impulse and therefore is initiated by the absorption of the light via rhodopsin. The membranous discs, located in the rod outer segment, contain high concentration of the visual pigment rhodopsin, which is specialized for light absorption (500 nm). In the dark, a small molecule called retinal is covalently bound to the rhodopsin (**Figure 3.5**). One photon of light induces the isomerization of the 11-cis retinal to the all-trans retinal, followed by conformational changes of rhodopsin. Light active rhodopsin (R^* or Meta II state) binds to the G protein transducin, which is attached to the inner layer of the plasma membrane. Once the G protein is activated, guanosine diphosphate (GDP) that is bound to the α G subunit of the G protein gets converted into guanosine triphosphate (GTP). After one phosphate is added to the α -subunit of the heterotrimeric transducin, the activated subunit releases the remaining two transducin subunits ($G\beta$ and $G\gamma$) and binds to the photoreceptor enzyme, called phosphodiesterase (PDE or PDE6). The PDE complex in turn consists of two catalytic subunits ($PDE\alpha$ and $PDE\beta$) and two inhibitor subunits ($PDE\gamma$). The GTP-

α G subunit binds to the PDE γ inhibitor and activates phosphodiesterase. This leads to the cyclic guanosine monophosphate (cGMP) hydrolysis and conversion into guanosine monophosphate (GMP). A single light activated rhodopsin molecule can catalyze hundreds of G proteins ^[152] and even though Yue *et al.* 2019 has shown that this value was overestimated as one rhodopsin molecule activates only 12 to 14 transducin-PDE effector complexes, the high ratio between G proteins to rhodopsin in the cell still leads to the efficient signal cascade ^[153]. In turn one active PDE enzyme hydrolyzes about one thousand cGMP molecules ^[154]. This significant signal amplification results in cGMP concentration reduction accompanied by GMP concentration increase, immediately affecting the cGMP-gated channels, finally resulting in their closure. Thus, in contrast to the first messenger, presented by the retinal, which activates the receptor at the beginning of the phototransduction process, cGMP acts as a second messenger in the downstream signaling. The cyclic-nucleotide-gated (CNG) membrane channels are regulated by the cGMP concentration in the cell. As long as cGMP molecule is bound, the channel is opened and the Na⁺ and Ca²⁺ ions influx into the cytosol takes place. Another Na⁺/Ca²⁺, K⁺ exchanger (NCKX) channel located in the membrane is not affected by the cGMP and therefore stays unaffected under the light illumination. It regulates the influx of sodium ions and efflux of potassium and calcium ions from the cytosol. The photo-induced decrease of the cGMP concentration forces the CNG channel to close, which in turn leads to reduction of Na⁺ and Ca²⁺ in the cell. In combination with the NCK exchanger channel, which moderates the efflux of Ca²⁺ ions, resulting in the calcium reduction in the cell. The constant ion current, which is present in the dark state with opened CNG gates, depolarize the membrane potential of the cell, while under the light illumination, the decreased concentration of calcium ions in the cell leads to the hyperpolarization of the membrane potential of the cell. In the dark, the photoreceptors are turned on, while the exposure to the light turns them off. The signal is transferred from the synaptic region of the photoreceptor via the bipolar cells and other intermediate cells (Müller cells, horizontal cells) to the retinal ganglion cells, which when activated, send the signals through their axon and the optic nerve to the brain. In the absence of light, the depolarized photoreceptor cells cause the generation of the action potential which is transmitted to the brain and is recognized as darkness. The hyperpolarization of the cells triggered by the light-induced-phototransduction disturbs the signal transmission and is identified by the brain as light. The photoreceptors are located in the outer layer of the retina, meaning that in order to reach the visual pigment and start the phototransduction cascade the light has to travel through the whole inner layer of nerve cells and then back to the nerve fiber layer to transfer the information into the brain ^[151,155–158].

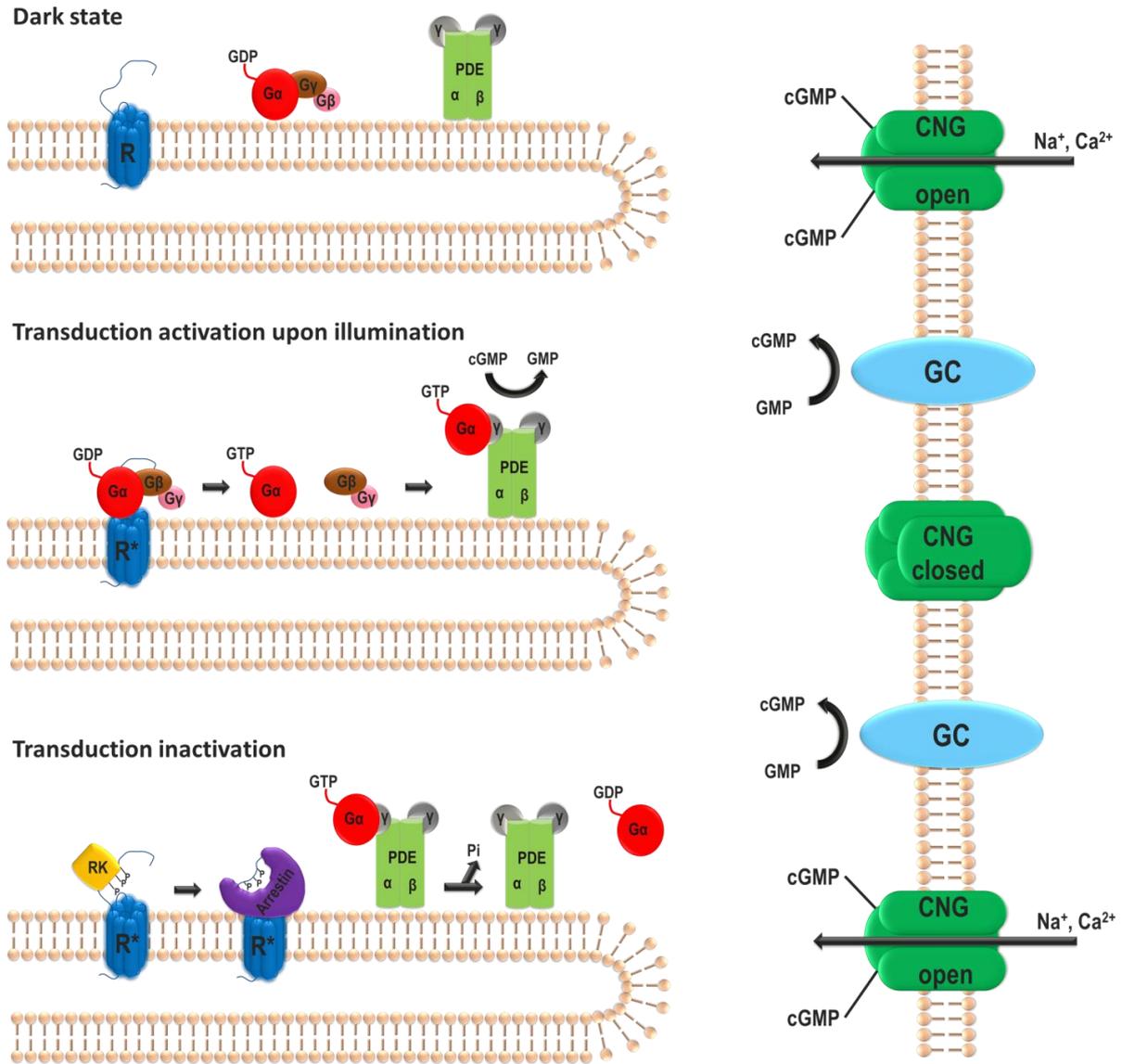


Figure 3.3. Schematic overview of the rhodopsin phototransduction cascade by photoreceptors located in the rod outer segment. Adapted after Bruns *et al.* 2005 ^[157].

For the recovery of the rod cells, two main proteins are needed: rhodopsin kinase (RK) and arrestin. Rhodopsin kinase phosphorylates the rhodopsin light-active Meta II state three times, providing higher affinity for arrestin binding. Once arrestin is bound to rhodopsin, the interaction with transducin is sterically hindered and does not take place any more. Thus, the phototransduction cascade is terminated and rhodopsin recovery can start. During this step all-trans retinal releases the protein, which then gets dephosphorylated by the phosphatase into the so called opsin protein. Incorporation with fresh 11-cis retinal induces a new photocycle again.

Besides rhodopsin, PDE complex also undergoes reactivation. GTP which is bound to the G α subunit autocatalytically dephosphorylates to GDP and triggers G α subunit release from the PDE complex. Thus, transducin recovers by forming a trimer complex again and free PDE γ subunits inhibit the PDE activity, resulting in the increase of the cGMP concentration in the cell. Excess of GMP in the cell is converted into cGMP by the catalytic reaction triggered by the enzyme guanylate cyclase (GC). Finally, CNG channels open, hyperpolarization of the photoreceptors disappears, and the recovery process is then considered completed.

1.3. Retinal regeneration

For the complete regeneration of the visual cycle, the retinal should also undergo the recovery process (**Figure 3.4**). Under illumination, 11-cis retinal bound to the rhodopsin isomerizes into all-trans conformation. This is accomplished with conformational changes of the rhodopsin, resulting in the light active Meta II state with the all-trans retinal attached to the protein. Due to thermally driven conformational degradation of rhodopsin, all-trans retinal releases the protein, converting it into opsin. Recovery of all-trans retinal into its 11-cis conformation is a complicated process, which involves different enzymes and cofactors. For this, retinal has to take a long journey starting from rod membrane disc, passing through the membrane into the retinal pigment epithelium (RPE), where the actual chemical recovery occurs, and then all the way back to the opsin protein embedded in the membranous discs ^[159,160]. After dissociation from the rhodopsin, all-trans retinal, which is located in the cytosolic space of the membranous disk, is transferred through the membrane via adenosine triphosphate (ATP)-binding cassette transporter (ABCR) ^[161].

In the next step, all-trans-retinol dehydrogenase (RDH) enzyme supported by the nicotinamide adenine dinucleotide phosphate (NADPT) cofactor reduces all-trans retinal to all-trans-retinol, also known as vitamin A ^[162]. Freshly reduced all-trans-retinol moves from the cytoplasmic space in the rod to the extracellular space between rod and retinal pigment epithelium. Binding to the inter photoreceptor retinoid-binding protein (IRBP) provides the entry to the RPE space, where all-trans-retinol gets bound by the cellular retinol-binding protein-1 (CRBP1). This exchange of the interaction partners is caused by the CRBP1 with three times higher affinity to the retinol compared to IRBP. The final all-trans-retinol to 11-cis retinal conformational transformation takes place in RPE and requires a high number of enzymes.

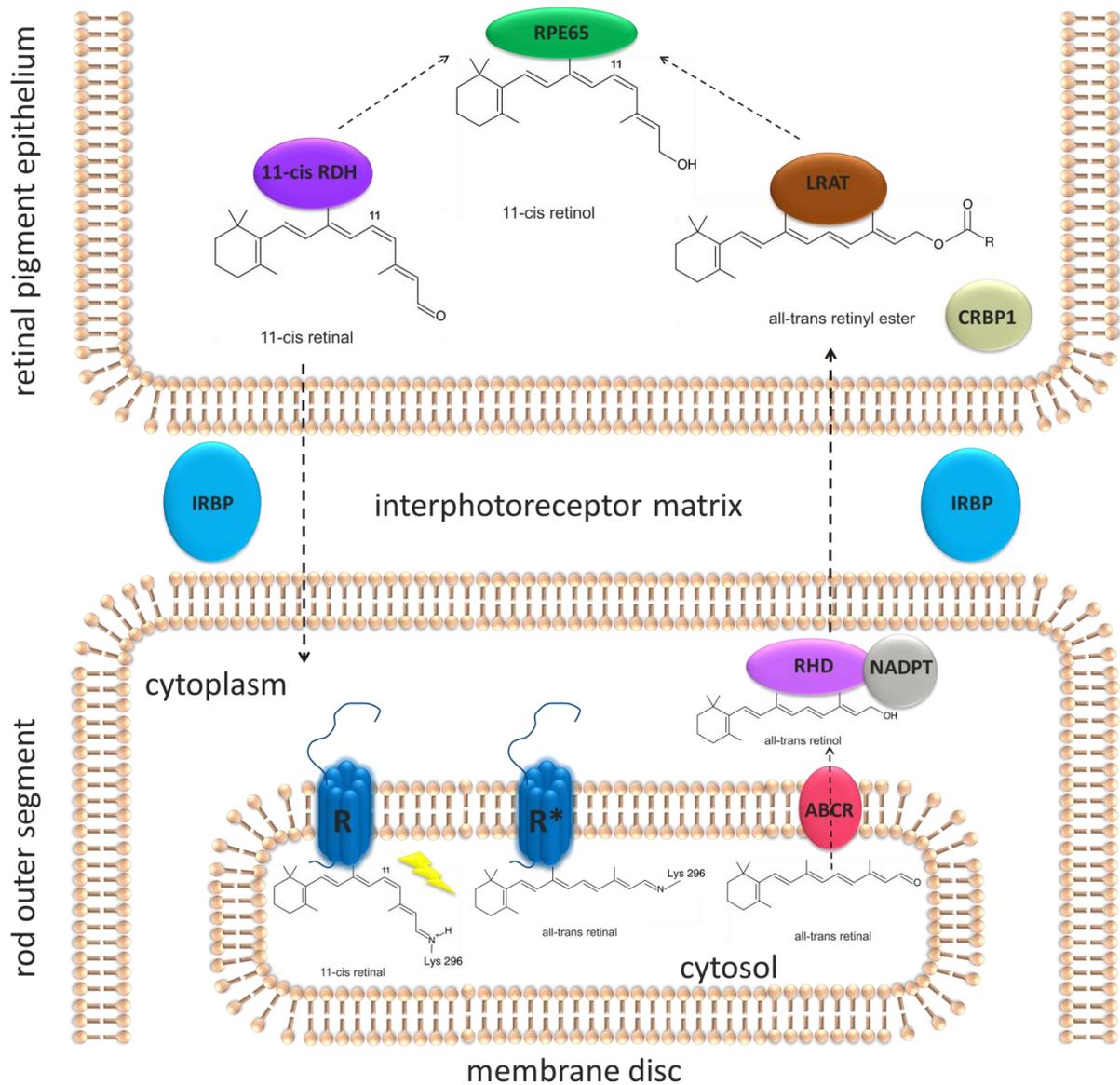


Figure 3.4. Schematic overview of the retinal regeneration in the visual cycle. Adapted after Maeda *et al.* 2013^[15].

The first one lecithin retinol acyl transferase (LRAT)^[165] is responsible for the transformation of all-trans-retinol to all-trans-retinyl ester (a stable storage form of vitamin A in the eye), which can isomerize to 11-cis-retinol^[166]. This reaction is catalytically stimulated by the second enzyme called isomerohydrolase (RPE65)^[166,167]. In the last step 11-cis-retinol is oxidized to 11-cis-retinal by 11-cis-specific retinol dehydrogenase (11-cis RDH)^[168]. The completely recovered 11-cis retinal travels back from RPE to the rod outer segment where it can reunite with opsin forming a visual pigment rhodopsin^[169].

1.4. Rhodopsin

1.4.1. Rhodopsin structure in the dark state

Rhodopsin is a light sensitive G-protein coupled receptor (GPCR) which is composed of two parts: the chromophore 11-cis retinal and the protein opsin. Bovine rhodopsin consists of 348 residues resulting in a molecular weight of approximately 40 kDa. The opsin is composed of seven transmembrane helices (H1-H7) connected through the loops in an alternating manner (**Figure 3.5**). One small helix H8 is located at the C-terminal cytoplasmic side. The protein has three loops in the extracellular region including the N-terminus which are named E1, E2 and E3 and three loops that are extended to the cytoplasm including C-terminus (C1, C2 and C3). Extracellular and cytoplasmic loops contain more hydrophilic residues while the transmembrane region contains mostly the hydrophobic residues, which allow the protein to be located in the nonpolar membrane of the cells.

Retinal is a ligand composed of a polyene chain and a β -ionone group. It is bound to the protein via a Schiff base bond between retinal carbon atom C15 and the ϵ -amino group of K296 residue. In the center of the transmembrane core, hydrophobic residues E122, M207, F208, H211, W265 and Y268 form a hydrophobic binding pocket for the retinal by stabilizing its nonpolar polyene chain. This binding pocket is already formed in the opsin and does not undergo significant changes during the photocycle. Negatively charged side chain of the E113 residue serves as a counter ion for the retinal Schiff base preventing the hydrolysis and release of the retinal in the active conformation.

The role of the secondary structure elements located in the extracellular region is particularly important for the retinal binding pocket and the sufficient phototransduction cascade. Two β -sheets in the second loop (E2) stabilized by the disulfide bond between C110 and C187 sterically prevent the retinal from its dissociation from the protein when rhodopsin is in the inactive state. Another interaction which influences the proper signaling pathway is the salt bridge formation between E247 and R135 also known as G-binding pocket. This prevents G protein (transducin) from binding to the rhodopsin in the inactive state ^[170,171]. Besides bound retinal, bovine rhodopsin includes several post translational modifications. Palmitoylation takes place at the C-terminal C322 and C323 and glycosylation occurs at the N-terminal N2 and N15.

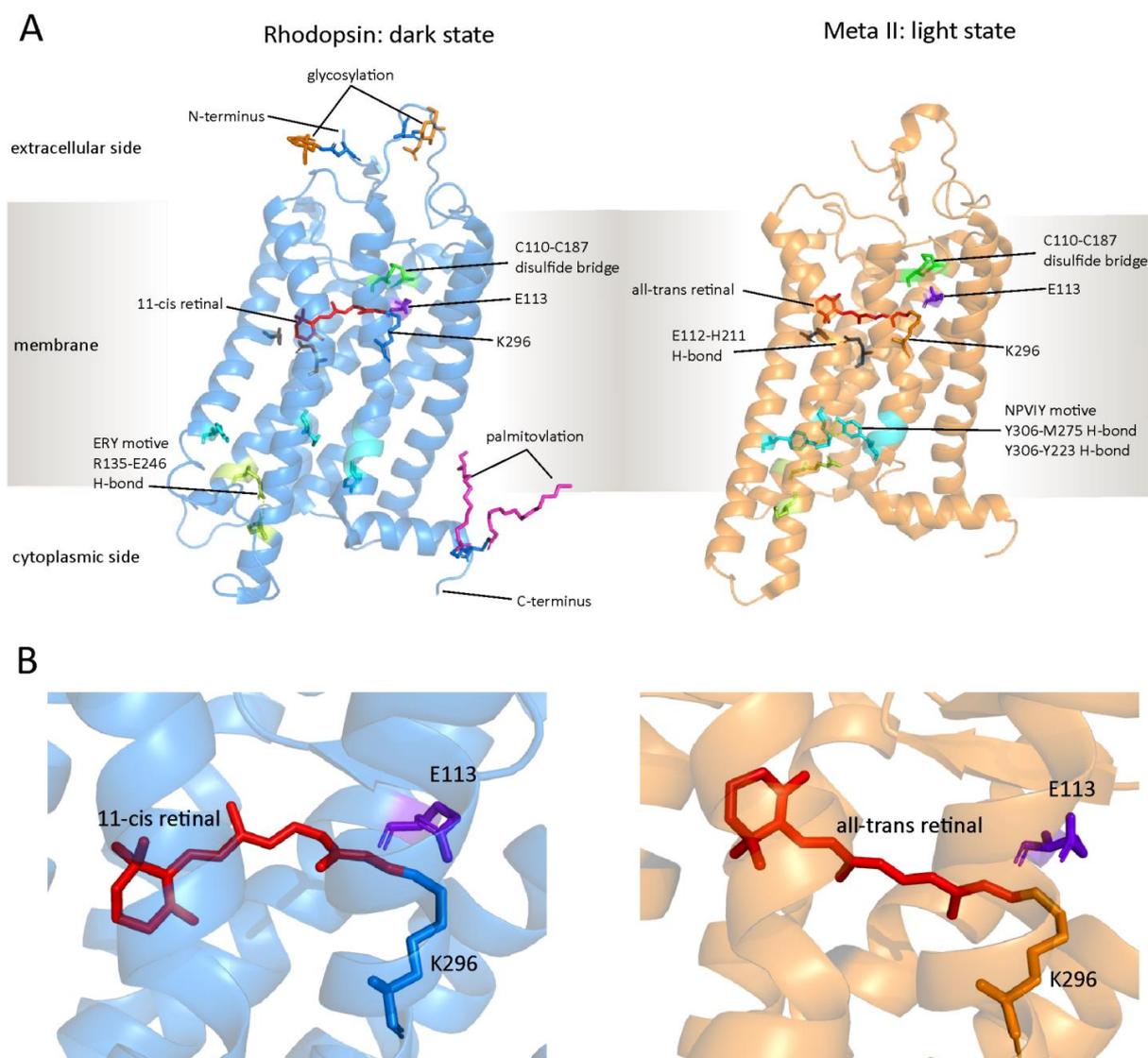


Figure 3.5. Crystal structure of bovine rhodopsin. On the left, rhodopsin is shown in the dark state (pdb code: 1U19), and the light active Meta II conformation is shown on the right (pdb code: 3PXO). Amino acid side chains and retinal are shown as sticks and are indicated in red, the counter ion is highlighted in purple. **A** – Posttranslational modifications are present only in the ground state structure and are highlighted in pink (palmitoylation) and orange (glycosylation). ERY, NPVIY motives and stabilizing H-bonds are indicated in the picture. **B** – Zoom into the binding pocket with retinal in 11-cis (left) and all-trans (right) conformation.

1.4.2. Rhodopsin structure in the light state

Light-induced isomerization of 11-cis retinal to all-trans retinal triggers rearrangements in the binding pocket, which lead to the conformational change of almost all secondary structure elements of rhodopsin (**Figure 3.6**). Proton transfer from positively charged protonated retinal Schiff base to the negatively charged counter ion E113 is a part of the activation switch^[172], resulting in an increased distance between the Schiff base and the side chain carboxyl group of E113^[173] in the light active Meta II state.

Light-induced conformational changes mostly affect the secondary structure elements responsible for the activation of G protein, which are located close to the cytoplasmic region. Structural rearrangements in the light active state affect H3 and H5 allowing new hydrogen bond formation between residues E122 (H3) and H211 (H5). This affects the movement of helices 3, 5 and 6, by turning H5 towards H6 and turning H6 in the opposite direction from H3^[173]. Thus, H6 and H5 extend to the cytoplasmic side increasing the binding interface of G protein transducin.

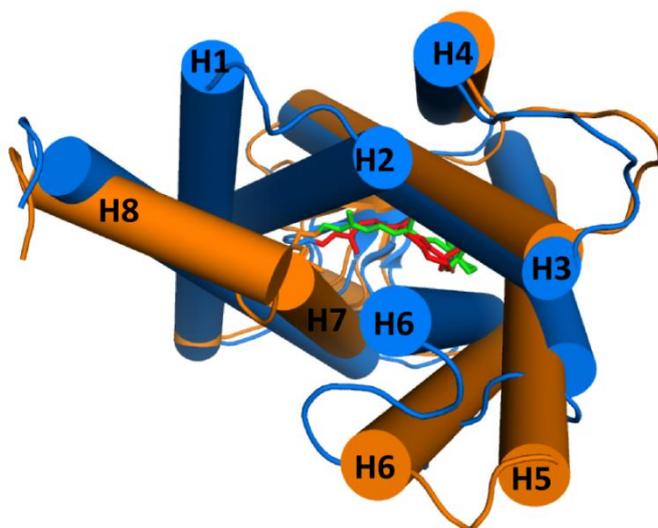


Figure 3.6. Superposition of rhodopsin and Meta II viewed from the cytoplasmic side. Structure of inactive rhodopsin in dark state is highlighted in blue (pdb code: 1U19) and light active Meta II (pdb code: 3PXO) is shown in orange. Amino acids side chains and retinal are shown as stick. 11-cis retinal is colored in red and all-trans retinal is shown in green. Adapted after Choe, 2011^[173].

Additional conformational changes affect the highly conserved amino acid sequence ERY (residues 134 - 136), which is located in the helix III and belongs to the “ionic lock”. Upon illumination, the salt bridge between residues R135 (H3) and E247 (H6) dissociates and the helices move away from each other. Breaking the “ionic lock” opens up the binding site of rhodopsin for the G protein transducin, and this makes it a key player in the activation of phototransduction cascade.

Another highly conserved motif NPVIY (302-306) is involved in important light-triggered conformational changes. The electrostatic interaction between the aromatic side of the residues Y306 from helix H7 and F313, which is located in the small C-terminal helix H8, holds H7 and H8 in close proximity to each other. Upon illumination, this interaction breaks resulting in Y306 rotation towards H6. The side chain interactions of Y306 with residues M257 (H6) and Y223 (H5) stabilizes helix 6 preventing it from the back movement towards H3^[174]. These interactions are responsible for irreversible light-induced conformational changes of active rhodopsin during the phototransduction activation process.

1.4.3. Photocycle

The retinal chromophore, which is a derivative of vitamin A, is a key player in the photocycle of rhodopsin. It is bound to the opsin via Schiff base with K296 and is stabilized by the counter ion E113. Upon light illumination 11-cis retinal undergoes cis-trans isomerization, which induces a conformational change of rhodopsin, resulting in several intermediate states of the photocycle. Meta II state with all-trans retinal bound is the light active conformation, which initiates the phototransduction cascade. Relaxation of the protein leads to the release of the all-trans retinal followed by its return to the apoprotein opsin.

Intermediate states adopted during the photocycle and in particular the retinal protonated or deprotonated conformation can be distinguished by UV/Vis spectroscopy. The absorption maximum of the protein in the ground state is around 500 nm, while each intermediate state has its characteristic absorption maximum (free retinal absorbs at 380 nm).

The first light-induced intermediate state, which arises in 200 fs is called photorhodopsin and has an absorption maximum of 570 nm ^[175]. It is followed by the stable rhodopsin conformation called bathorhodopsin. Bathorhodopsin absorbs at 543 nm and can be stabilized at -140°C (pdb code: 2G87). Retinal adopts an all-trans distorted conformation, which is expressed in 11-trans, 15-anti conformation.

Conversion from bathorhodopsin to lumirhodopsin occurs in the nanosecond regime and flows through the blue shifted intermediate state (477 nm), which could only be observed by time-resolved measurements. Lumirhodopsin in turn could be trapped at -40°C (pdb code: 2HPY) and is characterized by the absorption maximum of 497 nm.

Meta I state with an absorption maximum of 478 nm presents the first conformation in the photocycle with significant conformational changes of the secondary structure elements. It exists in the equilibrium with the light active Meta II state and can be observed by high pH or low temperature ^[176].

The light-induced photocycle from dark state rhodopsin to Meta I conformation includes several intermediate conformations and happens in the micro seconds range, while next photoconversion from Meta I into Meta II state takes milli seconds. Meta II state (pdb code: 3PXO) is achieved by deprotonation of the Schiff base leading to the salt bridge destruction, which results in an absorption maximum shift to 380 nm. The counter ion E113 is therefore protonated indicating the proton transfer in the hydrophobic binding pocket. This leads to the biggest conformational changes in the photocycle reflected by the opening of G protein binding site followed by the transducin activation.

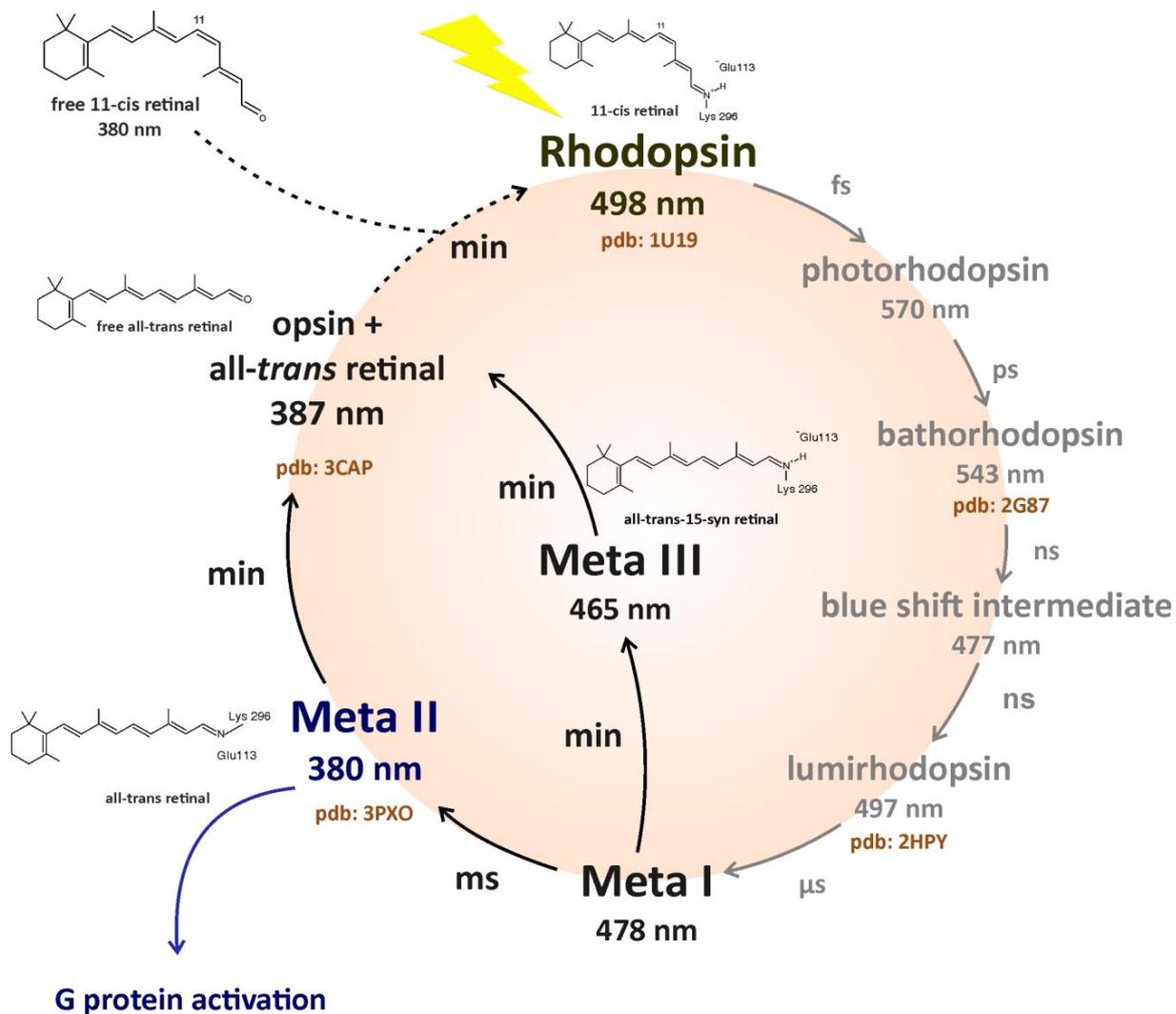


Figure 3.7. Photocycle of bovine rhodopsin. Adapted after Ernst, 2002 and Stehle, 2014^[176,177].

Meta II rhodopsin - transducin interaction occurs in 100 ms, while thermal relaxation of Meta II state into opsin and free retinal is in the minutes regime^[178]. Transducin activation is interrupted by rhodopsin kinase phosphorylation of rhodopsin enabling the binding of arrestin.

Rhodopsin decay to opsin and free retinal is mainly defined by the retinal hydrolysis and can go through two alternative pathways with fast and slow kinetics. The direct rhodopsin relaxation to opsin and free retinal via Meta II state is characterized by deprotonated all-trans retinal and protonated counter ion E113. With a duration of five minutes, this process is five times faster than the relaxation via Meta III state, which lasts for 25 minutes^[177]. In contrast to Meta II state, Meta III state is characterized by protonated retinal in all-trans-15-syn conformation with an absorption maximum of 465 nm. Since this

pathway is significantly slower than Meta II decay, Meta III state is considered as a storage conformation of inactive rhodopsin. Approximately 40% of the thermal relaxation takes place via a slow kinetic pathway Meta III state.

Finally, released free all-trans retinal undergoes retinal regeneration cycle in the retinal pigment epithelium, where it gets transferred into active 11-cis chromophore. The last one can bind to the opsin again and forms rhodopsin, which in turn can undergo the photocycle upon light photon absorption.

1.5. Retinal dystrophy diseases

1.5.1. Retinitis pigmentosa (RP)

Disruptions in the highly complex process of our sense of sight or photocycle can lead to diverse visual diseases, which have enormous impact on the quality of life. Retinitis pigmentosa (RP) affects around 1.5 million people in the world, circa 1 in 4000 patients, which makes it the most common inherited retinal dystrophy (IRD) disease. The classification of IRD is not straight forward, since some phenotypes may be the same for different diseases in their later stages. Generally, one could distinguish between progressive and stationary retinal diseases. Progressive retinal diseases encompass cone-rod dystrophy (cone degeneration triggers loss of rod functions), cone dystrophy, Leber congenital amaurosis (LCA) and macular dystrophy, while stationary retinal disease includes congenital stationary night blindness (CSNB). RP belongs to the progressive inherited diseases, in which degeneration of rod photoreceptors occurs first and then the cone photoreceptors get affected. Since rods are located in the periphery of the retina and they are responsible for the dim vision, their degeneration results in night blindness and tunnel vision, which are the first RP symptoms (**Figure 3.8**).

The cones are responsible for detailed vision and color recognition, therefore their degeneration leads to color loss and reduction of central vision. RP is a progressive disease, which can end up with a full blindness in the later stages of the disease affecting patients by the age of 40. RP encompasses high amount of mutations found in more than 80 different genes. Depending on the proteins encoded by the gene, which is affected by RP caused mutation, a particular pathway in the visual cycle may be affected. This includes phototransduction cascade, visual cycle including retinal development, ciliary structure and transport, inter photoreceptor matrix etc. ^[179]



Figure 3.8. Vision of patients affected by retinitis pigmentosa and congenital stationary night blindness. The figure was generated by Picasa 3.

1.5.2. Treatment options for RP

Nowadays, there are two available treatment approaches for the genetic causes leading to RP. The first one is gene-specific (in some cases mutation-specific) and the second is mutation independent. The main idea of the gene-specific approach is reproduction of the healthy wild type protein in the cell. Therefore, this method has to be applied at the earliest stages of RP disease, when the degeneration of the photoreceptors has not yet begun. The first clinical studies performed in 2008 were successful and in 2018 this resulted in the first US Food and Drug Administration (FDA) approved gene therapy for retinal dystrophy diseases ^[180].

For the disruption of the retinal recovery process in the visual cycle, the oral treatment with synthesized organic compound was successfully tested on mice. The patients with RPE65 and LRAT mutated genes were treated with 9-cis retinyl acetate which is the synthetically stable analog for the 11-cis retinal. The chromophore can bind to the opsin in the same way as biologically active 11-cis retinal and restore the photocycle ^[181]. Overall, the certain diet treatment may reduce the disease symptoms related to the vitamin E deficiency but it has no effect on the vitamin A based diseases such as RP ^[182].

Another mutation independent treatment is cell replacement therapy, where the cells are introduced to the vitreous body or subretinal space. Despite the fact that transplantation surgery is always risky, finding a donor cell is usually a difficult process and using the stem cells derived directly from the patient is highly expensive, the first transplantation of stem RPE and photoreceptor cells were already successfully performed in patients ^[183]. However, the clinical tests are still in the early phases.

One of the most futuristic treatment approaches at the last stage of RP are electronic implants. These implants are produced in California (epiretinal implant) and in Germany by Retina Implant AG in

Reutlingen (subretinal implant). An epiretinal electrical implant is connected to a camera, which is placed on the eyeglass, and stimulates the ganglion cells, while a subretinal implant stimulates the bipolar cells [184,185]. This treatment has shown an increased visual activity and overall very promising results, however, it requires a long and difficult recovery and adaptation period for the patients.

1.5.3. CSNB disease classification

CSNB is a non-progressive inherited retinal disorder, which was found to be genetically and clinically heterogeneous. First symptoms are reduction of dim and night vision, problems with adaptation to the darkness and in some cases loss of the general visual acuity. Nowadays, around 360 different mutations in 17 genes are known to cause CSNB. These genes encode proteins involved in the phototransduction cascade, in signaling transmission from photoreceptors to the bipolar cells and in visual cycle in particular retinal development in RPE (Figure 3.9).

Depending on the affected proteins, CSNB disease types with normal fundus appearance can be divided into two classes [186] (Figure 3.10). The first type, so called Riggs-type of CSNB, affects the photoreceptor, while the second (Schubert-Bornschein-type) impairs the bipolar cells function. Patients with Riggs-type phenotype have a normal day light vision including color and details but prolonged adaptation in the darkness. Single point mutations leading to this type of disease were identified in proteins involved in the phototransduction cascade such as transducin (alpha-subunit and GNAT1 subunit), phosphodiesterase (beta subunit PDE6B) and rhodopsin.

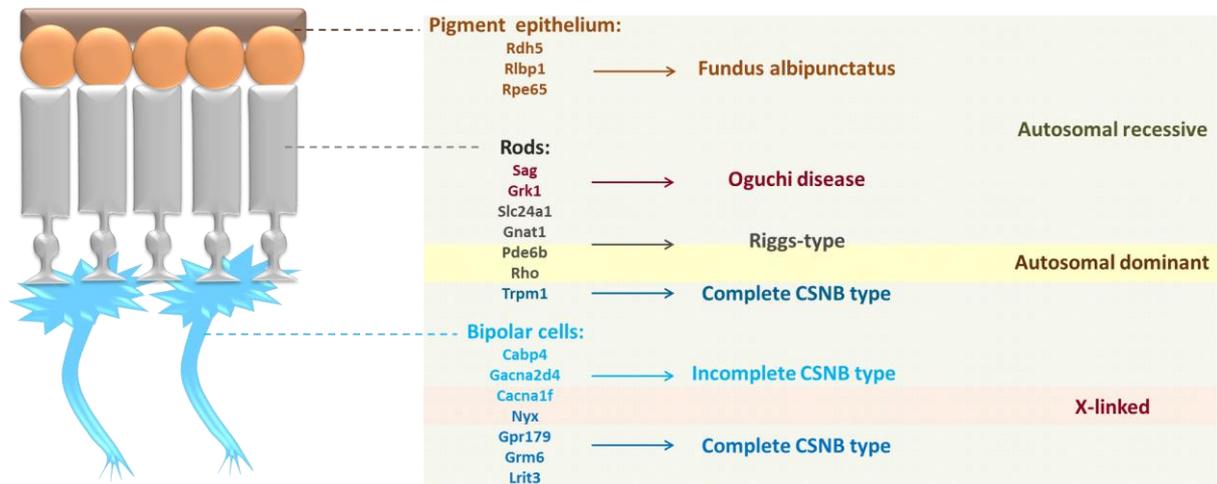


Figure 3.9. Localization of gene defects known to cause CSNB and mode of inheritance. Mutations in the genes, which were found in pigment epithelium (brown), rods (black) and bipolar cells (blue), and associated disease are shown with the same color-code. Inheritance mode for each gene and corresponding disease is highlighted in specific background color: green - autosomal recessive, yellow - autosomal dominant and pink - X-linked inheritance.

Schubert-Bornschein-type of CSNB affects post phototransduction process, in particular the signal transmission between photoreceptors and bipolar cells. Depending on the pathway dysfunction of the bipolar cells, the Schubert-Bornschein-type, can be subdivided into complete (cCSNB / CSNB1) and incomplete (icCSNB / CSNB2). Complete CSNB form leads to the postsynaptic dysfunction of the ON bipolar cells, while incomplete CSNB form influences the presynaptic process, by affecting both ON and OFF bipolar cells. Mutations causing the CSNB1 type are located directly in the bipolar cells, while CSNB2 type mutations occur in the voltage-dependent calcium channel (CACNA1F)^[187] or in the proteins involved in the calcium regulation (CABP4). Patients affected by CSNB1 usually have normal day vision but suffer from decreased visual acuity, myopia (shortsightedness) and in some cases nystagmus (involuntary eye movement). In contrast to CSNB1, people with CSNB2 phenotype almost do not feel the disease symptoms in the normal life, the vision in the darkness may be little or even completely not disturbed. Since CSNB2 has a very heterogeneous phenotype compared to CSNB1, in some cases also color vision may get affected^[188,189].

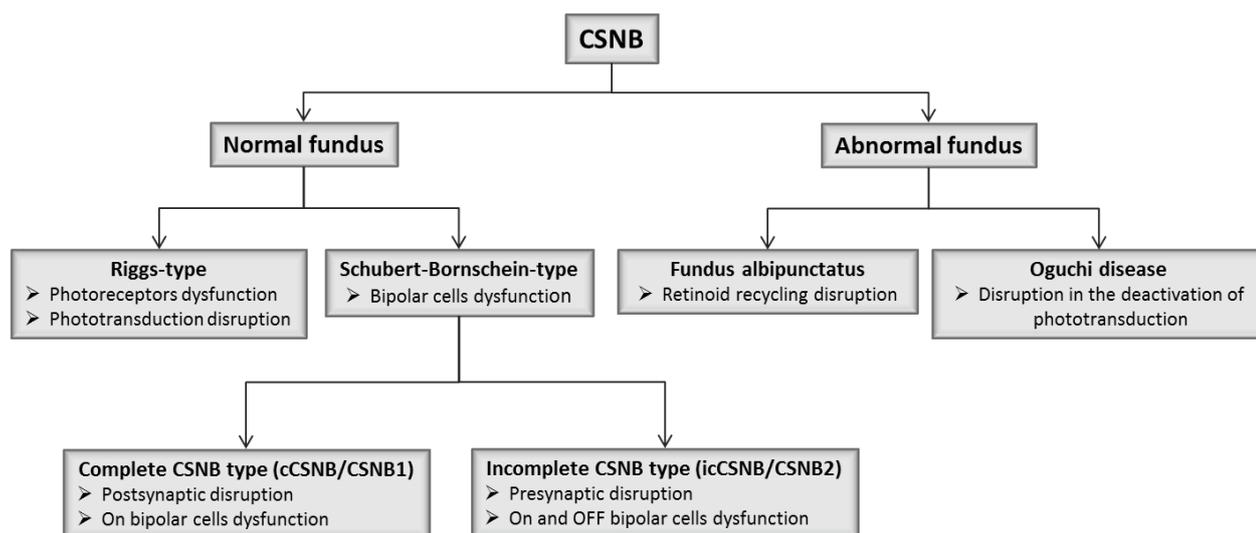


Figure 3.10. Classifications of CSNB disease types.

In case of abnormal fundus appearance CSNB disease types can be divided into fundus albipunctatus and Oguchi disease. Similar to all CSNB based diseases; patients with abnormal fundus appearance have typically normal day vision and impaired dark vision. In fundus albipunctatus this is caused mostly by the mutations in RDH5 gene, which encodes retinol dehydrogenase^[190]. 11-cis-specific retinol dehydrogenase is involved in the last step of retinal recovery process in visual cycle by oxidizing 11-cis-retinol to 11-cis-retinal in RPE. Oguchi disease involves mutations in arrestin and rhodopsin kinase proteins, which are involved in the deactivation of the phototransduction cascade^[191,192].

1.5.4. Mode of inheritance of CSNB associated mutants

CSNB is a very heterogeneous disease, with overlapping phenotype-genotype correlation (**Figure 3.9**). It can be inherited as an autosomal dominant, autosomal recessive or X-linked recessive disorder. X-linked mutations with 57.9% are the most common ones, followed by the autosomal recessive mutations with 40% and the autosomal dominant mutations encompass 2.1% of all cases ^[186].

The only one CSNB type with direct phenotype-genotype correlation is CSNB with abnormal fundus. Both, fundus albipunctatus and Oguchi disease are associated with autosomal recessive mutations in pigment epithelium and rods, respectively.

Schubert-Bornschein type of CSNB includes six autosomal recessive mutations (genes *Trpm1*, *Gpr179*, *Grm6*, *Lrit3* for CSNB1 type and *Cabp4* and *Gacna2d4* for CSNB2 type) and two X-linked mutations (genes *Nyx* and *Cacna1f* for CSNB1 and CSNB2 type, respectively). All genes are expressed in the bipolar cells (except *Trpm1*, which is expressed in the rods).

Around 25% of all RP cases account to autosomal dominant mutations. Riggs-type of CSNB is caused by mutations in the rods, which can be either autosomal recessively or autosomal dominant inherited. Autosomal dominant mutations are located in the *RHO* and *Pde6b* genes, which encode rhodopsin and phosphodiesterase, respectively. These proteins are involved in the rod photo transduction cascade.

1.5.5. Treatment options for CSNB

Currently, there are no preventive measures for this disease. Since the most common CSNB symptoms are poor dim light and night vision, these are often overlooked and not diagnosed by the patients. Therefore, a detailed clinical diagnosis is highly required for the possible further treatment. It includes electroretinography (ERG) to distinguish between four CSNB subtypes and gene-specific sequencing in order to identify the exact gene affected by the mutation.

CSNB exhibits an overlapping phenotype with visual diseases such as RP, progressive rod-cone dystrophy and acquired night blindness (vitamin A deficiency) but in contrast to these diseases, CSNB is non-progressive. Therefore, a patient's clinical history should be maintained over time.

So far, there is no treatment for CSNB affected patients. Similar to RP, gene therapy and photoreceptor transplantations are possible future cures, which are currently under development.

1.5.6. Rhodopsin mutations associated with CSNB

Mutations in the *RHO* gene, which encodes the rhodopsin protein, lead to the disruption of the phototransduction resulting in the impaired visual cycle. The majority of the single point mutations in the rhodopsin sequence (around 150) lead to RP and are accomplished with rod photoreceptors apoptosis. Since CSNB relevant mutations result in increased basal activity of receptors, they are called constitutively active mutations. In case of rhodopsin, where dark state is inactive, they shift the equilibrium between inactive and active state towards the light active conformation. Disruption of intramolecular interaction in the ground state can be achieved by affecting the retinal binding pocket, in particular Schiff base linkage to the chromophore or by impairing the transducin active side needed for the photo cascade process.

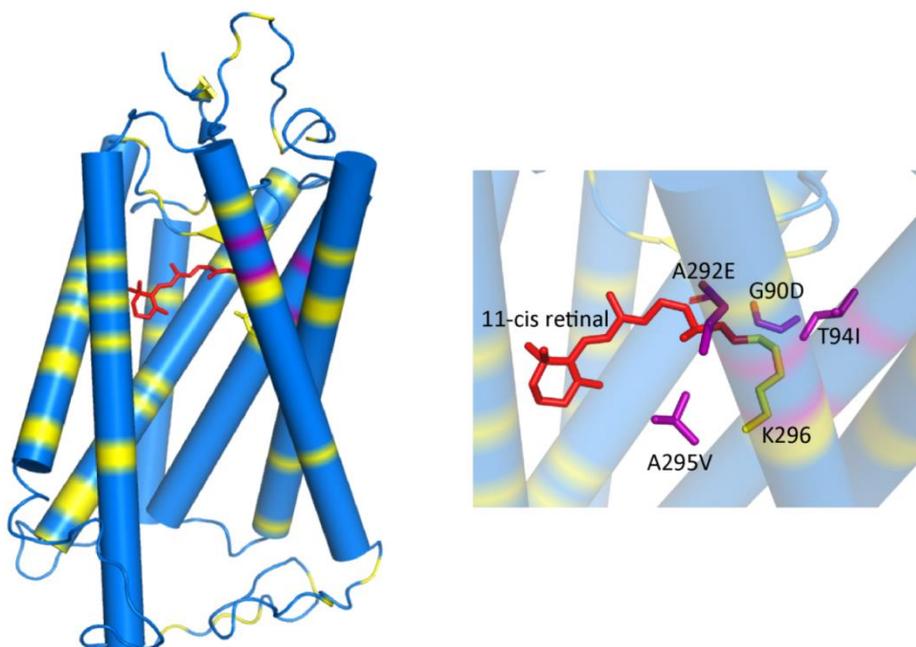


Figure 3.11. Distribution of RP and CSNB single point mutations in rhodopsin. RP mutations are shown in yellow and CSNB-causing mutants are highlighted in purple; 11-cis retinal, Schiff base K296 and CSNB associated mutants are shown as sticks. **Left** – cylindrical representation of rhodopsin in the ground state (pdb code: 1U19). **Right** – zoom in of the binding pocket with retinal in 11-cis. Four CSNB-causing mutations (G90D, T94I, A292E and A295V) are introduced by using PyMol.

Only four single point mutations in rhodopsin are known to cause CSNB, namely: G90D^[193], T94I^[194], A292E^[195] and A295V^[196] (**Figure 3.11**). In contrast to RP-associated mutations, which are spread throughout the whole protein sequence, the CSNB mutants are located in the retinal binding pocket.

G90 is a unique position; depending on the nature of the mutated amino acid it can either cause RP or CSNB. Thus, a G90V mutation lead to RP and G90D is associated with CSNB disease.

1.6. State of the art

1.6.1. Development of rhodopsin structures

In 2000 the first high resolution GPCR and concurrently first rhodopsin crystal structure was determined by Palczewski *et al.* The crystals were obtained by using the purified protein from bovine rod outer segment membranes in the ground apo state with a resolution of 2.8 Å^[170]. After this structural breakthrough, rhodopsin was extensively studied by many research groups. This resulted in an increase of the high-resolution crystal structures with more detailed description of the retinal binding pocket on the atomic level. The next crystal structures improved the original one by adding missing residues^[197] and the water molecules to the retinal binding pocket, which are involved in the rhodopsin activity regulation (pdb code: 1HZX^[198] and 1L9H^[199], respectively). Two years later, Okada *et al.* improved the resolution to 2.2 Å and could fully resolve the polypeptide chain (pdb code: 1U19). This structure includes analysis of the 11-cis retinal bound to K296 via a Schiff base and the whole retinal binding pocket at the atomic resolution^[200]. In further conformational comparison and model-based investigations of rhodopsin, this X-ray ground state structure is used as a reference.

Meanwhile, Yeagle and coworkers performed three-dimensional solution NMR structure of rhodopsin in the ground state (pdb code: 1JFP)^[197] and light active Meta II state (pdb code: 1LN6)^[201]. However, the protein structure was constructed from many short peptides, which represent isolated secondary structure elements of the protein. The main focus was the analysis and comparison of the cytoplasmic side of the protein, which was not well resolved in previous crystallographic studies, in the dark and light active state. The transmembrane domain in the inactive state was in good agreement with X-ray structures. Comparison of these two conformations gave an insight into the transducin binding site of the protein.

In 2006 Nakamichi and Okada published two X-ray structures of early photo intermediate states of rhodopsin: lumirhodopsin (pdb code: 2HPY)^[202] and bathorhodopsin (pdb code: 2G87)^[203]. Trapping of short-lived rhodopsin conformations was performed by the argon laser illumination of the crystals under a stream of cold nitrogen gas (105 K). The insight in the retinal binding pocket showed the distorted all-trans retinal conformation in the bathorhodopsin followed by the β-ionone ring rotation in the lumirhodopsin intermediate state.

The first crystal structure of light active Meta II state was performed in 2006 with a resolution of 4.15 Å (pdb code: 2I37)^[204]. Five years later Choe *et al.* was able to improve the resolution to 3.0 Å (pdb code: 3PXO)^[173]. Trapping of short lived Meta II state was achieved by soaking hundreds of grown opsin crystals with all-trans retinal in the dark assuming binding of all-trans retinal to the high concentrated

opsin crystals. This structure is used as a reference for all further structural comparison of light active Meta II rhodopsin conformation.

Table 6. Overview of rhodopsin structures.

year	pdb code	author	resolution	state	mutation	referenc
2000	1F88	Palczewski <i>et al.</i>	2.8 Å	ground state	wt	[170]
2001	1HZX	Teller <i>et al.</i>	2.8 Å	ground state	wt	[198]
2001	1JFP	Yeagle <i>et al.</i>	Solution NMR	ground state	wt	[197]
2002	1L9H	Okada <i>et al.</i>	2.6 Å	ground state	wt	[199]
2002	1LN6	Choi <i>et al.</i>	Solution NMR	Meta II	wt	[201]
2004	1GZM	Li <i>et al.</i>	2.65 Å	ground state	wt	[205]
2004	1U19	Okada <i>et al.</i>	2.2 Å	ground state	wt	[200]
2006	2HPY	Nakamichi <i>et al.</i>	2.8 Å	lumirhodopsin	wt	[202]
2006	2G87	Nakamichi <i>et al.</i>	2.8 Å	bathorhodopsin	wt	[203]
2006	2I37	Lodowski <i>et al.</i>	4.15 Å	Meta II	wt	[204]
2007	2J4Y	Standfuss <i>et al.</i>	3.4 Å	ground state	N2C/D282C	[206]
2008	3CAP	Park <i>et al.</i>	2.9 Å	opsin	N2C/D282C	[174]
2011	3PXO	Choe <i>et al.</i>	3.0 Å	Meta II	wt	[173]
2011	2X72	Standfuss <i>et al.</i>	3.0 Å	Meta II	E113Q, N2C/D282C	[207]
2012	4A4M	Deupi <i>et al.</i>	3.3 Å	Meta II	M257Y, N2C/D282C	[208]
2013	4BEZ	Singhal <i>et al.</i>	3.3 Å	Meta II	G90D, N2C/D282C	[209]
2016	5DYS	Singhal <i>et al.</i>	2.3 Å	Meta II	T94I, N2C/D282C	[210]
2017	5TE3	Gulati <i>et al.</i>	2.7 Å	opsin	N2C/D282C	[211]

Ligand free opsin is unstable under detergent solubilization and therefore its crystallization turned to be more challenging compared to rhodopsin. Only after the introduction of the stabilizing mutations, which form the additional disulfide bond formation (N2C/D282C) ^[212], it was possible to crystallize opsin. The protein was purified from bovine eyes rod outer-segment disc membranes extraction. The first X-ray structure with the resolution of 2.9 Å was published in 2008 by Ernst and coworkers (pdb code: 3CAP). The main changes were observed in the cytoplasmic side affecting the movement of helices H5 and H6. Also highly conserved ERY and NPVIY motives, which are responsible for light-induced structural rearrangements, undergo conformational changes compared to rhodopsin. Almost ten years later, the structure resolution was improved to 2.7 Å (pdb code: 5TE3) ^[211].

1.6.2. Crystal structure of constitutively active rhodopsin mutants

2007 Standfuss *et al.* published the first crystal structure of recombinant rhodopsin, expressed in mammalian cells. This was achieved by introducing the mutations in the extracellular side, which form an additional disulfide bond between the N terminus and the loop E3. Double mutant N2C/D282C increases the thermal stability of the protein in detergent but does not significantly affect its activity and structural conformation^[206]. Xie *et al.* showed that the double mutation has no significant impact on its transducin activation; the life time of the active Meta II state of the stabilized mutant (6.6 minutes) is in the same range as for the wild type (7.9 minutes)^[212].

Further structural investigations performed in Prof. Standfuss group were focused on the constitutively active rhodopsin mutants, which affect the retinal binding pocket moiety and lead to visual diseases such as retinitis pigmentosa (**chapter III 1.5.1**)^[213] and congenital stationary night blindness (**chapter III 1.5.3**)^[193]. All these mutants are less stable than the wild type rhodopsin and therefore carry the stabilizing double mutations N2C/D282C, which allows successful protein purification in detergent. Reconstitution with retinal was performed by adding 11-cis retinal during purification followed by selective illumination to trap the light active Meta II conformation^[207,208]. The determined structure is in good agreement with the Meta II conformation obtained by soaking the opsin crystals with all-trans retinal^[173].

Constitutively active rhodopsin mutations can be classified into two groups. First type affects the retinal binding pocket and disturbs retinal binding and/or retinal release. While the second type of mutations is located close to the cytoplasmic side and influences the transducin binding site, which is responsible for the G protein activation. The following crystal structures of constitutively active mutants were obtained in the light active Meta II state - E113Q (pdb code: 2X72)^[207], G90D (pdb code: 4BEZ)^[209] and T94I (pdb code: 5DYS)^[210] from the first class and M257Y (pdb code: 4A4M)^[208] from the second class.

The Schiff base linked K296 and its counter ion E113 are the two most important amino acids in the visual cycle (**Figure 3.12**). Mutations in both of them lead to constitutive activation which is directly related to the visual diseases^[213]. E113Q mutation breaks the salt bridge between counter ion E113 and K296 resulting in the proton transfer distortion from all-trans Schiff base bound retinal to negatively charged counter ion E113 and prolongs the Meta II life time by dramatically slowing down the retinal release. Meta II structure of the E113Q stabilized mutant was found to be highly similar to the light active state of the wild type. However, weak density of the K296 indicates that the retinal is not covalently bound to the protein as it is in the case of Meta II wild type. Besides that, the occupancy refinement of the retinal binding pocket showed a mixture of all-trans and other retinal populations for E113Q stabilized mutant in Meta II state^[207].

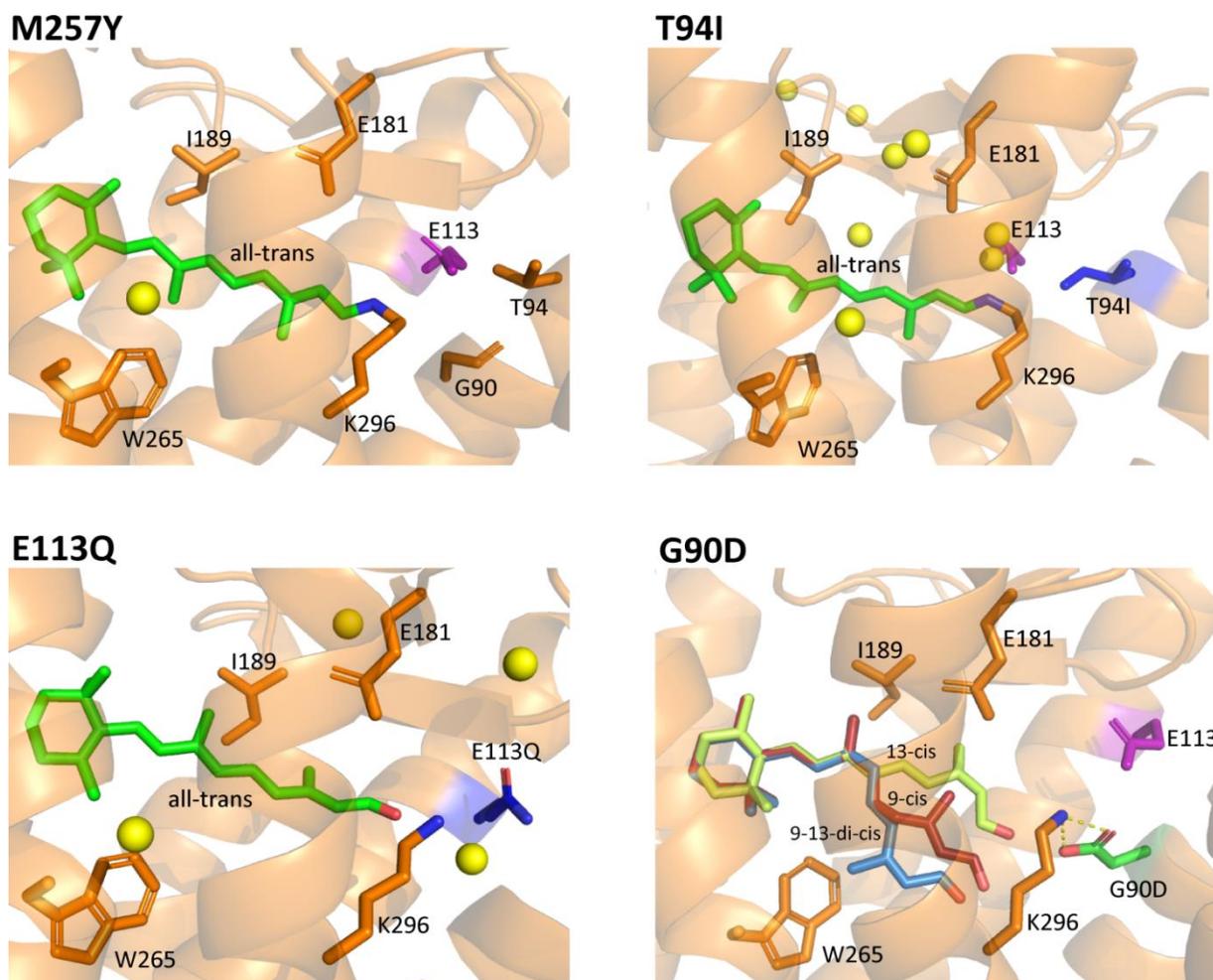


Figure 3.12. Comparison of retinal binding pocket of Light active Meta II state of constitutively active mutants. M257Y (pdb code: 4A4M^[208]), E113 (pdb code: 2X72^[207]), T94I (pdb code: 5DYS^[210]) and G90D (pdb code: 4BEZ^[209]). All-trans retinal is shown in green and cis-isomers are highlighted in yellow (13-cis), sky-blue (9-13-di-cis) and firebrick (9-cis). Water molecules from the pdb crystal structure are represented as yellow balls.

In contrast to E113, the crystal structure of another constitutively active M257Y mutant does not affect the retinal binding pocket showing only all-trans covalently bound chromophore conformation in the light active state. The basal activation is therefore achieved by the disruption of the ionic lock, which is responsible for the conformational rearrangements needed for transducin activation^[208].

Recently, Singhal *et al.* have published crystal structures of two constitutively active mutants G90D (pdb code: 4BEZ)^[209] and T94I (pdb code: 5DYS)^[210], which are known to cause congenital stationary night blindness. Both mutations are located in close proximity to the retinal binding pocket and thus, affect the chromophore attachment to the protein in the ground state. Charged G90D and hydrophobic T94I mutations reduce the interaction between the Schiff base and its counter ion, which might cause CSNB symptoms. For both stabilized mutants, it was not possible to crystallize rhodopsin in the ground state

but only in the light active Meta II conformation, which was obtained by selective illumination of 11-cis reconstituted rhodopsin for the G90D mutant ^[208,209] and by soaking opsin crystals with all-trans retinal in case of T94I mutant ^[210].

Despite the fact that both mutants cause one phenotype of disease, the influence on the retinal binding pocket differs significantly. The negatively charged aspartate in G90D mutation forms a salt bridge with K296, impairing the normal Schiff base formation between all-trans retinal and K296, resulting in the stabilization of Meta II conformation. The hydrophobic side chain of T94I mutant also prolongs the lifetime of rhodopsin active state but the retinal binding side is not affected showing all-trans retinal bound to the K296 in a similar way as observed in the wild type. T94I mutation affects the Schiff base hydrolyses needed for the retinal release from the protein. Thus, this process occurs faster in the dark state and is delayed in the light active Meta II conformation ^[214]. The solved crystal structure of the T94I mutant shows significant changes in the water network close to the retinal binding pocket compared to the wild type Meta II state. Also steric hindrance of the mutated amino acid might affect the Schiff base hydrolysis ^[210], resulting in disease relevant global changes of the visual cycle.

1.6.3. G90D Mutation

The G90 residue is located in a unique position. Depending on the nature of the mutated amino acid, it can lead to either RP or CSNB disease. The introduction of a branched-chain valine in place of G90 causes RP, while negatively charged aspartate leads to CSNB. The overall structure of the G90D mutant is very similar to the wild type, showing no significant structural changes. However, this constitutively active mutation affects the retinal binding position, resulting in the constant basal activity of the protein. Singhal *et al.* from Prof. Standfuss group in Paul Scherrer Institute performed a crystal structural analysis of the G90D mutant in order to explain the cause of CSNB disease on the molecular level.

The high structural similarity of the G90D and wild type in ligand free opsin conformation and light active state allows the authors to limit the impact of the G90D mutation to the chromophore binding pocket. Crystallization of G90D in the dark state was not successful. The authors address it to the conformational heterogeneity of the G90D mutant in the dark state. After the incorporation with 11-cis retinal, the absorption maximum of the protein, solubilized in detergent, shifts from 498 nm to 482 nm. This indicates mutation-induced Schiff base perturbation. Also, the 280 nm/480 nm ratio, which indicates the quality of retinal incorporation to the opsin, was in the range of 2.0 to 2.2, in contrast to wild type values, which are between 1.6-1.7. This was recognized as a mixture of opsin and rhodopsin and used as an explanation for the conformational heterogeneity which prevents the crystallization. FTIR experiments published in 1996 by Zvyaga *et al.* reported the protonated E113 counter ion in the

dark state of the G90D mutant, which is characteristic for Meta II conformation of wild type and thus propose the light active Meta II like state of the G90D mutant ^[215].

Singhal *et al.* performed crystallographic refinement of the electron density in the binding pocket of the light active G90D mutant. The modeling of different retinal isomers showed a heterogeneous, non-covalently bound population, which was determined by the authors as a mixture of different cis isomers.

Thermal stability studies ^[209] performed on wild type and G90D in presence and absence of retinal showed that in contrast to wild type, there is no significant effect of retinal binding on the stability of the G90D mutant ($T_m = 60^\circ\text{C}$). Wild type incorporated with 11-cis retinal showed the most stable form compared to opsin and all-trans bound conformations ($T_m^{11\text{-cis}} = 63^\circ\text{C}$ vs. $T_m^{\text{opsin}} = 52^\circ\text{C}$). While reconstitution of the G90D opsin with 11-cis or all-trans retinal did not show this stabilizing effect, suggesting a lower stability of the G90D in the ground state or disruption in the binding pocket, which perturb the normal retinal binding.

E113 is responsible for the stabilization of the ground state as a counter ion for the protonated Schiff base and is also crucial for the light-induced retinal release. The introduced charged glutamate side chain interacts with K296, perturbing the activation switch and therefore the proper retinal binding. The salt bridge between D90 and K296 stabilizes the opsin and prevents the Schiff base formation with the retinal. The rhodopsin ground state is thereby destabilized, favoring the light active conformation.

Increased basal activity, which is characteristic of all constitutively active mutants, is the main cause of CSNB retinal disease. Different explanations theories have been proposed: i) spontaneous isomerization of the chromophore, ii) pre-active dark state of the protein and iii) spontaneous activation of the opsin ^[209]. However, there is still no clear explanation of the molecular mechanism that causes this retinal disorder.

1.7. Motivation and aims

Crystal structures of the constitutively active G90D mutant are available for the ligand free opsin conformation and the light active state. So far, there is no X-ray structure of its dark conformation, nevertheless Singhal *et al.* reported structural heterogeneity addressed to the mixture of opsin and rhodopsin in the ground conformation and proposed that the ground state of the G90D mutant is destabilized due to the E113-K296 Schiff base disruption, which leads to the increased rate of retinal thermal isomerization ^[209]. Even though FTIR spectroscopy and spin-labelling experiments showed the pre-active Meta II like dark state of the G90D mutant, the lack of the structural data in the dark state keeps the question open as to how the retinal is bound to the opsin and what the influence of the mutation on the binding pocket environment is.

Although the light active conformation of the G90D mutant was shown to be stabilized and structurally very similar to the wild type Meta II conformation, the crystallographic refinement of the binding pocket has indicated a mixture of non-covalently bound retinal cis isomers. This leads to the question of how this mutation caused salt bridge destabilization, influences the binding pocket and what the actual active retinal conformation in the light state of the G90D mutant is. In order to answer these questions and to fulfill the lacking structural gaps concerning the exact retinal conformation and mutation-induced changes in the binding pocket in the dark and light active state of the G90D mutant, a combination of liquid and solid state NMR spectroscopic experiments have to be performed.

Since the active single point mutations in the rhodopsin results in the distortion of the visual cycle, characterization of the light-activation process and comparison with the wild type kinetics are needed for the mutation impact analysis. The photocycle, in particular the light-induced conformational changes, preferred pathway and dynamics are highly sensitive to the active mutation and therefore represent high biological importance. The impact of the G90D mutation on the photodynamic process can be analyzed by flash photolysis analysis.

The overall goal of this part of the thesis is to get insights into the mutation-caused conformational changes of the constitutively active G90D mutant, which lead to the disruption of the visual cycle and results in the CSNB disease.

2. Material and Methods

2.1. Constructs

Three different rhodopsin constructs are discussed in this doctoral thesis: wild type (wt), wild type containing a stabilizing mutation N2C/D282C^[206], and the disease relevant G90D mutant^[193], which also contains the stabilizing mutation (N2C/D282C)^[209]. All genes were cloned into the pACMV-tetO vector for tetracycline-inducible expression in HEK293 cells^[216]. For the wild type, the stably transfected cell line previously produced by Stehle *et al.*^[177] from Prof. Schwalbe group was used. Two other constructs (wt and G90D) carrying the stabilizing mutation (N2C/D282C) were kindly provided by Prof. Standfuss (Switzerland). Singhal *et al.*^[209] from Prof. Standfuss group performed site-directed mutagenesis to insert the G90D mutation into a pACMV-tetO vector containing a synthetic bovine opsin gene with stabilizing mutation N2C/D282C^[212] and a subsequent stable transfection of HEK293 cells. The sequences of all rhodopsin constructs used in this thesis are listed in **Appendix A9**.

2.2. Cell culture protocol

Stably transfected mammalian cells were stored in liquid nitrogen with 10% DMSO. After the cells were thawed, 1 ml of the maintenance medium (medium A) (**Appendix A10**) was added to the cryo tube, slowly mixed with the cells and subsequently transferred to a 15 ml falcon tube. The suspension was supplemented with additional 10 ml medium A and centrifuged at 160 g for 10 minutes at 4°C. The supernatant was discarded, the pellet was resuspended in 2 ml medium A and subsequently supplemented with additional 5 ml. Cells were grown in an incubator with 5% CO₂ at 37°C. After reaching high confluence, the cells were split and/or transferred from a small (T25) to a big cell culture flask (T175). For this, the medium was removed from the flask; the adherent cells were washed with PBS buffer and detached with 1X trypsin solution. The cell suspension was incubated for 3 minutes in an incubator with 5% CO₂ at 37°C. The cell solution was resuspended with additional 5 ml of medium A until the entire adherent cells were detached from the cell culture flask surface. Finally, the cells were supplemented with additional volume of cell medium and either transferred into a big flask for further expression or split into 143 cm² culture petri dishes for the large scale expression. The induction of the cells was conducted with medium B containing the desired isotope selective labeling scheme (**Appendix A10**).

NMR experiments were performed using selectively labeled amino acids for the induction medium (medium B). Medium B was manually prepared and included α,ϵ -¹⁵N labeled tryptophan for liquid state NMR experiments and ¹⁵N labeled lysine for solid state NMR experiments. The samples for the

photoflash analysis were prepared with natural abundance amino acids in the induction medium (medium C).

Two to three days after induction, after the cells were confluent, they were harvested (5000 g for 10 minutes at 4°C) and resuspended in PBS solution (pH 7.4). For liquid and solid state NMR experiments, approximately 150 to 200 petri dishes were required to achieve the minimal sufficient concentration of the sample (ca. 0.7 mg), while the photoflash experiments were performed with less concentrated sample with an approximate number of petri dishes ranging from 50 to 70 (ca. 80 µg). The large expression required for one NMR sample is time consuming and typically lasts approximately for two months. Due to the limited size of the used incubator in the laboratory only two constructs could simultaneously be expressed.

2.3. Rhodopsin purification

2.3.1. Retinal purification

The only commercially available isomer of the chromophore is the light stable all-trans retinal. The 11-cis isomer (**Figure 3.13**), which binds to the opsin in the dark state and triggers the photo transduction process, can be achieved by the isomerization of the all-trans isomer. For the liquid state NMR and photoflash analysis a commercially available all-trans retinal was purified, while for the solid state NMR experiments ¹³C selectively labeled all-trans retinal provided by Dr. Jiafei Mao from Prof. Glaubitz group was used. For the last one, two different labeling scheme were applied, the first one was ¹³C isotope labeled at the positions C12, C13 and C20 (12,13,20-¹³C), while the second one was ¹³C labeled at the C14 and C15 (14,15-¹³C) (**Figure 3.13**).

To perform the isomerization, the retinal (100 mg natural abundance, 5 mg ¹³C labeled) was dissolved in 50 ml and 20 ml ethanol, respectively, and illuminated with white light for 30 minutes under constant stirring on ice. Under illumination the retinal isomerizes into four isomers: all-trans, 9-cis, 11-cis and 13-cis, which can be separated by HPLC (high performance liquid chromatography). Therefore, the solvent of the mixture of isomers was exchanged against the retinal HPLC equilibration buffers (92% hexane and 8% diethyl ether). The initial ethanol solvent was evaporated under nitrogen gas and the remaining substance was dissolved in the HPLC equilibration buffer. After the retinal isomerization, all further actions should be performed under dark conditions. Separation of the retinal isomers was performed with Kromasil-SIL 10x250 mm HPLC column (*Merck*). The 11-cis isomer was collected and subjected to buffer exchange against ethanol according to the procedure described above. Remaining retinal isomer fractions were pooled for further recycled isomerization and stored at -80°C.

The isomerization of the retinal is irreversible, therefore, it is highly important to perform all experiments under dim light conditions and store the stock solution in a lightproof tube.

2.3.2. Antibody preparation

The Rho 1D4 monoclonal antibody (*University of British Columbia, Vancouver, Canada*) was coupled to activated Sepharose 4B (*GE Healthcare*) and used for further rhodopsin purification. Therefore, the isolated antibody (AB) was purified. AB solution was first dissolved with AB equilibration buffer (pH 7) with a 1:4 ratio and applied onto an antibody protein A column. The column was previously equilibrated with the same buffer. Optimal binding was achieved by a loading rate of approximately 1 ml/min at 4°C. The column was washed with the equilibration buffer until the yellow color originated from the antibody disappeared, indicating completed binding of the AB to the column. Elution of the antibody was carried out with the AB elution buffer (pH < 3) (**Appendix A10**). For the visualization of the fractions containing the AB, the protein A column was connected to the FPLC (*Äktapurifier™, GE Healthcare, USA*). Additionally, the fractions were pooled and the pH was restored up to the physiological conditions with AB pH adjusting buffer (pH 10-11). AB solution was subsequently dialyzed against AB dialysis buffer over night at 4°C

The next preparation step included AB coupling to the activated Sepharose 4B. Based on the protein size, the optimal ratio between the antibody and the sepharose beads for the rhodopsin (42 kDa) is 0.286 g beads for 6 mg of antibody. Therefore, the corresponding amount of the CNBr sepharose beads were swelled in approximately 20 ml of 1 mM hydrochloric acid (15 minutes at room temperature) and subsequently transferred into a sintered glass funnel, followed by an additional wash with 200 ml of 1 mM HCl and 20 ml coupling buffer. Without allowing the suspension to dry, it was transferred to the antibody solution and incubated at room temperature for two hours. To control the binding process of the AB to the beads, the suspension was passed through a gravity column. If the flow-through showed free antibody by ultraviolet-visible spectrophotometry (UV/Vis), the incubation was extended to another hour. After the coupling was completed, the antibody-beads solution was subjected to several washing steps:

- 20 ml of AB coupling buffer
- 20 ml of AB blocking buffer
- 10 ml of AB low pH buffer (2 times)
- 10 ml of AB coupling buffer (2 times)
- 20 ml of PBS.

The final purified antibody bound to the sepharose beads was either used for further rhodopsin purification by immunoaffinity chromatography or was stored for later experiments at -20°C.

2.3.3. Rhodopsin reconstitution of wild type rhodopsin ^[177]

Cell pellet was resuspended in PBS buffer, one protease-inhibitor tablet (*cOmplete™, Roche, Germany*) was added, the suspension was supplemented with 11-cis retinal to the final concentration of 20 µM. The cells were mechanically homogenized and incubated for two hours at 4°C before the second portion of 20 µM 11-cis retinal was added and the incubation was repeated for another two hours. The final concentration of 11-cis retinal was 40 µM. After adding the retinal, all further purification steps were performed under dim light conditions.

After the incorporation of the retinal into the opsin was finished, rhodopsin was extracted from the cellular membrane using DDM (n-dodecyl-β-maltose) as a detergent. Therefore, 1% w/v DDM was added to the rhodopsin solution and subsequently incubated for two hours at 4°C. The extracted rhodopsin was centrifuged at 48000 g for 30 minutes at 4°C and collected in the supernatant. The concentration of the protein was measured using UV/Vis spectroscopy at 500 nm.

Finally, immunoaffinity chromatography using 1D4 monoclonal antibodies, which are highly selective for the epitope of rhodopsin, was performed for rhodopsin purification. Antibodies coupled to resin were added to the DDM rhodopsin solution and rotated overnight at 4°C.

After rhodopsin was bound to the resin beads, the mixture was subjected to several washing steps:

- 50 ml of PBS with 0.05% DDM
- 50 ml of 20 mM phosphate buffer (pH 6.5) with 0.05% DDM
- 50 ml of 20 mM phosphate buffer (pH 6.5) with 0.05% DDM + 70 µM elution peptide.

The peptide, used to elute the protein, mimics the amino acid sequence from the rhodopsin epitope and therefore allows releasing the rhodopsin from the antibodies bound state. This nine residues long peptide (TETSQVAPA) was previously synthesized via solid phase synthesis in Prof. Schwalbe group. Elution buffer was added to the protein solution until no more rhodopsin was detected. The monitoring of the fractions was carried out using UV/Vis spectroscopy. The protein-containing fractions were pooled and concentrated to a final volume of approximately 1 ml (ca. 20 µM).

2.3.4. Rhodopsin reconstitution of stabilized mutants ^[217]

Rhodopsin constructs containing the stabilizing mutation were purified differently than the wild type. In contrast to the wild type, these stabilized mutant constructs were first extracted from the membrane with detergent and only after binding to antibodies, supplemented with 11-cis retinal. The incorporation with the retinal was performed overnight (vs. 4 hours for wt) at 4°C. The washing steps of the rhodopsin remain the same as those for the wild type, while the elution procedure was slightly changed. The rhodopsin-resin suspension was transferred into a 50 ml falcon tube and supplemented with 12 ml of solution buffer. The mixture was rotated for 45 minutes at 4°C and subsequently collected in one fraction. This step was repeated four times, after which rhodopsin was no longer detected.

2.4. Sample preparation

Liquid state NMR sample preparation

To prepare a liquid state NMR sample, a concentrated rhodopsin solution, which was initially obtained in an elution buffer, was subjected to buffer exchange with the liquid state NMR buffer (**Appendix A10**). The protein was concentrated to a total volume of 250 μ l and transferred to a Shigemi NMR tube. The concentration was in the range of 60-80 μ M (ca. 0.6-0.8 mg).

Photoflash

The samples for the flash photolysis experiments were prepared with the same buffer conditions as liquid state NMR experiments. The protein was concentrated to a total volume of 250 μ l with a concentration of 8 μ M (ca. 80 μ g).

Solid state NMR sample preparation ^[218,219]

For solid state NMR experiments proteins were reconstituted in lipids. Therefore, the rhodopsin purification strategy was modified. During the washing steps (described below) the detergent was exchanged from DDM to OG (octyl β -D-glucopyranoside):

- 50 ml of PBS with 0.05% DDM
- 50 ml of 2 mM phosphate buffer (pH 6.5) with 0.05% DDM
- 100 ml of 2 mM phosphate buffer (pH 6.5) with 1.46% OG
- 50 ml of 2 mM phosphate buffer (pH 6.5) with 1.46% OG + 150 μ M elution peptide.

The protein-containing fractions were pooled and concentrated to a final volume of 3 ml (ca. 0.7 mg).

DOPC (1,2-dioleoyl-sn-glycero-3-phospholone) lipids were used for the reconstitution of the protein. Therefore, a 100-fold molar excess of lipids were dissolved in degassed buffer D (4% OG) and sonicated for 10-15 minutes. After lipid suspension became a clear solution, it was added to rhodopsin and rotated for one hour at 4°C. When the reconstitution of rhodopsin into lipids was finished, the mixture was dialyzed against detergent using 200 volumes of buffer E. The molecular cut-off of the dialysis membrane of 7 kDa allows free OG detergent to diffuse through the membrane. The buffer was exchanged six times every four hours. The proteoliposomes were centrifuged at 52000 g for one hour at 4°C and carefully supplemented with 200 µl DNP juice. DNP juice contains a polarization agent, Amupol, which is used to enhance the DNP signal during solid state NMR measurements.

The mixture was incubated overnight at 4°C without applying any mechanical rotation. The excess of the DNP juice was removed and the remaining pellet subsequently transferred into a 3.2 mm sapphire NMR rotor. The sample was stored at -80°C.

2.5. NMR measurements

Liquid state NMR experiments were recorded at 600 MHz and 950 MHz spectrometers at 298 K. 1D x-filter and 2D SOFAST-HMQC experiments were applied to detect α, ϵ -¹⁵N tryptophan labeled indole resonances. NMR kinetics experiments on the stabilized G90D mutant were acquired at 800 MHz, 298 K, at different time intervals after illumination with emission wavelength of 514 nm of 300 mW for 0.5 second. Dark state measurements were performed under dim light conditions. All NMR spectra were processed and analyzed using TopSpin version 3.2 (Bruker Biospin).

DNP enhanced solid state NMR experiments were recorded at 400 DNP Bruker system including 400 MHz NMR spectrometer and 263 GHz Gyrotron as a microwave source. Dark state experiments were performed at 110 K.

3. Results and discussions

3.1. Retinal purification

Three all-trans retinal samples with different labeling schemes were purified as described in **chapter III 2.3.1**. The labeling scheme of the retinal is shown in **Figure 3.13A**: the first sample was not isotope labeled, while the other two were ^{13}C -isotope labeled at the positions C12, C13, C20 (12,13,20- ^{13}C) and C14, C15 (14,15- ^{13}C). Regardless of the labeling scheme, the samples were purified using the same HPLC method (**Figure 3.13B**). Four isomers (13-cis, 11-cis, 9-cis and all-trans) show different retention times and therefore, could be collected separately. Fractions containing the three other isomers (13-cis, 9-cis and all-trans) were pooled for further recycling.

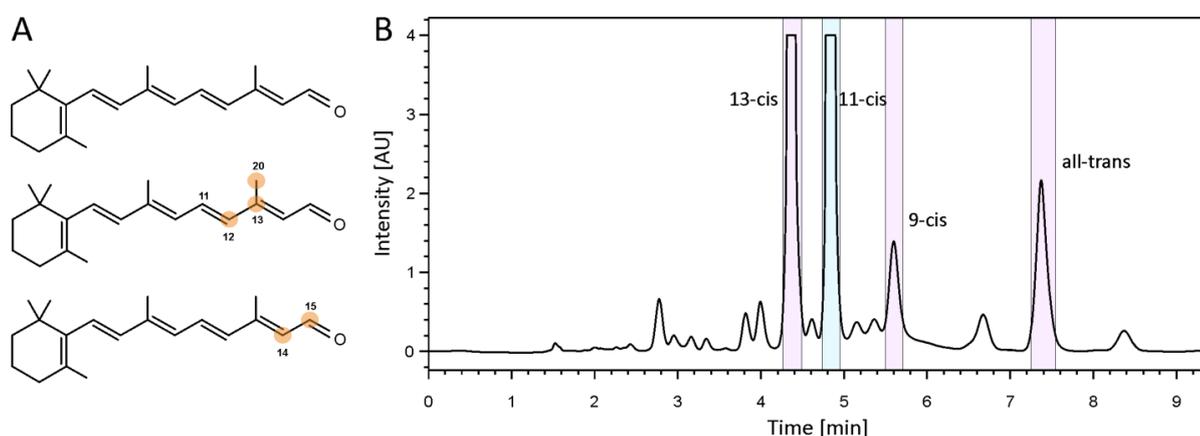


Figure 3.13. Purification of retinal labeled with different schemes. **A** – Retinal labeling schemes. Orange circles represent ^{13}C labeling positions. **B** – HPLC chromatogram of illuminated retinal. Four major signals correspond to the retinal isomers and can be isolated separately. 11-cis fraction is highlighted in blue, while other isomers fractions are colored in purple.

Depending on the labeling scheme of the all-trans retinal, different amounts of starting material were used. The first purification was performed on the unlabeled natural abundance retinal. Therefore, 100 mg of all-trans retinal was dissolved in 50 ml ethanol. For the second purification 20 mg of 12,13,20- ^{13}C all-trans retinal was used, while for the third one only 5 mg were purified. The purification yields and further details are listed in **Appendix 11, Table A10**.

3.2. Rhodopsin purification

This doctoral thesis comprises three different rhodopsin constructs: wild type (wt), stabilized wild type (N2C/D282C) and the disease relevant stabilized (N2C/D282C) G90D mutant. The double mutation N2C/D282C was first introduced by Standfuss *et al.* in 2007^[206] for the crystallization of rhodopsin, recombinantly expressed in mammalian cells. These mutations form an additional disulfide bond in the extracellular side between the N-terminus and the loop E3 (**Figure 3.16**), which increases the thermal stability of the protein without significantly affecting its activity and structural conformation^[212].

Rhodopsin sequence comprises five tryptophans (W35, W126, W161, W175, W265) (**Figure 3.16**), which were used as reporter signals. The chemical shift assignment of indole resonances for the wild type was carried out in earlier studies performed by Jochen Stehle^[177].

UV/Vis-absorption spectroscopy allows distinguishing between different conformations states adopted by rhodopsin. Thus, the absorption maximum at 500 nm corresponds to the dark state with 11-cis retinal bound to the K296 via Schiff base (SB). After illumination 11-cis retinal isomerizes into the all-trans conformation, which gives a rise to an absorption maximum at 380 nm and is called light active Meta II state. Remarkably, free retinal dissociated form rhodopsin absorbs at the same wavelength (380 nm) as Meta II, while opsin shows only absorption maximum at 280 nm, which is mainly caused by the absorption of aromatic amino acids.

Three constructs were expressed in tetracycline inducible stably transfected HEK293 cells as described in **chapter III 2.2**. A test expression of all three proteins is shown in **Appendix A11, Figure A18**. First, all constructs were purified according to the purification protocol of wild type, where 11-cis retinal was added before the solubilization. The UV/Vis- absorption spectroscopy in combination with 1D ¹H NMR experiments was used to quantify the quality of retinal reconstitution. All spectra were recorded under dim light condition and correspond to the rhodopsin dark state. In **Figure 3.14A left panel** 1D ¹H NMR spectra of the indole region of the three investigated constructs are displayed. The stabilized mutants show poorly resolved bulk of signals at 10.1 ppm, which indicates a retinal free opsin conformation. This is confirmed by the absorption spectra shown in **Figure 3.14B left panel**. Wild type profile shows an absorption maximum at 500 nm, which indicates the successful 11-cis retinal binding and a rhodopsin dark state formation, while the stabilized mutants do not contain any characteristic absorption maxima that could be referred to the rhodopsin formation. The absorption profiles of the wild type (rhodopsin) and the stabilized constructs (opsin) showed in **Figure 3.14B left panel** display the crucial effect of the retinal binding, which was performed according to the same protocol for all samples. For the wild type the retinal is successful bound, while the stabilized constructs do not contain any attached

chromophore. An additional disulfide bridge introduced into the extracellular domain impedes the retinal reconstitution and requires optimization in the rhodopsin purification strategy.

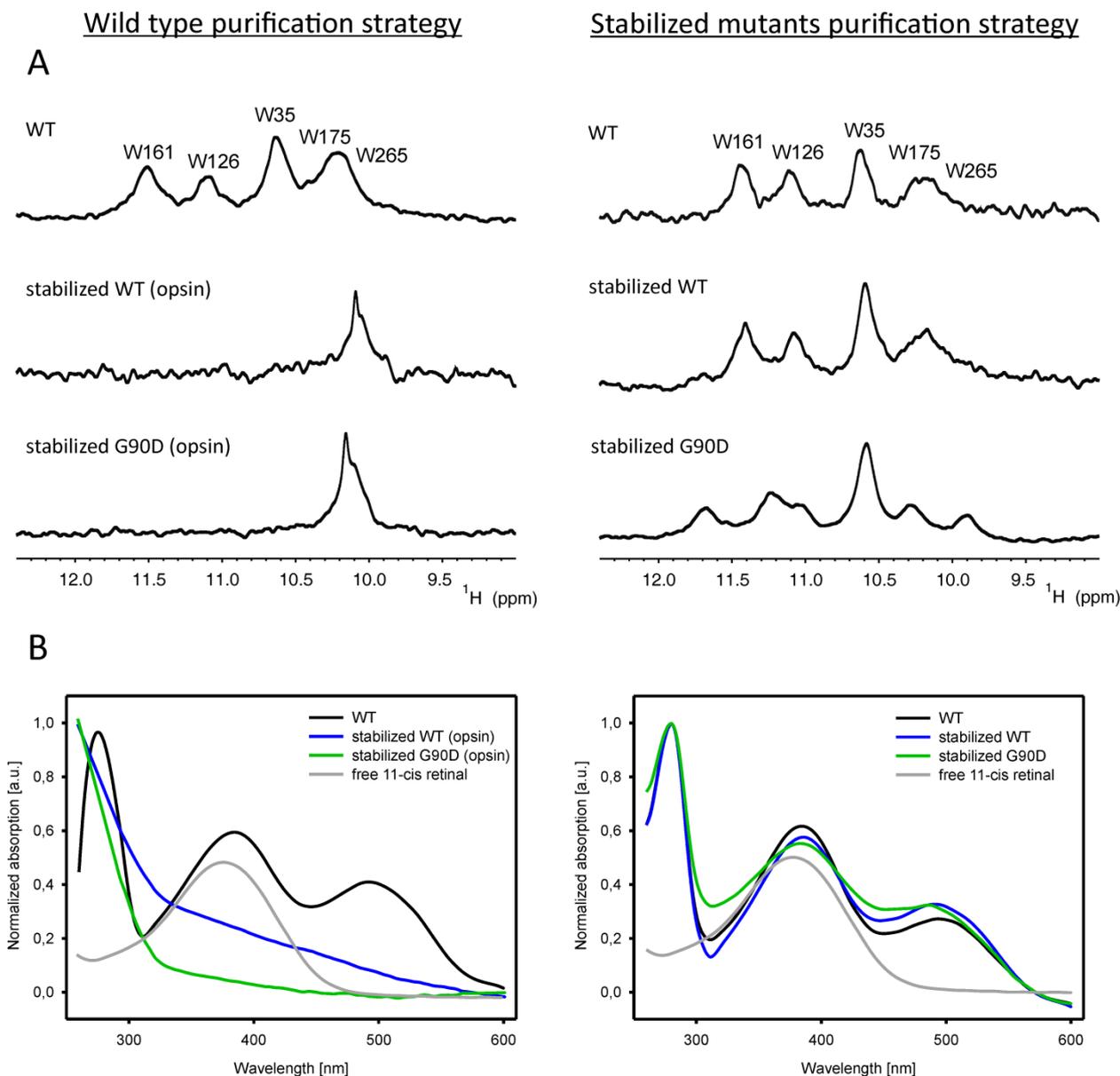


Figure 3.14. Retinal reconstitution. ^1H 1D NMR spectra of the indole region (**A**) and absorption spectra (**B**) recorded under dim light conditions of wild type, stabilized wild type and stabilized G90D mutant purified according to the wild type purification protocol (**left panel**) and the optimized protocol (**right panel**). The absorption spectra are normalized at 280 nm.

Therefore, the rhodopsin purification protocol was modified for the stabilized mutants. In contrast to the wild type purification, the retinal was added in excess after the solubilization and purification of opsin to improve the retinal incorporation into the protein. By this modification, both mutants show resolved indole signals in ^1H 1D NMR (**Figure 3.14A right panel**) indicating successful rhodopsin dark

state formation. The 1D ^1H NMR spectra of the amide region of the three constructs purified according to different strategies are shown in **Appendix A11, Figure A19**.

Wild type and stabilized wild type show similar signal profiles, while the G90D mutant resonance distribution differs from the wild type. The absorption spectra of the stabilized mutants also confirm the 11-cis chromophore incorporation. The absorption maximum of the G90D mutant is blue shifted to 490 nm compared to the wild type absorption at 500 nm, since this mutation causes the disruption of the retinal binding pocket as previously reported by the Standfuss laboratory ^[209]. The second absorption maximum at 380 nm, which is present in the spectrum of each construct can be referred to either i) the light active Meta II state, formed by partial illumination during purification or the performing of the absorption spectroscopy experiments, or ii) an excess of 11-cis free not bound retinal, that remained during the purification steps. In order to clarify the origin of the signal at 380 nm, size-exclusion chromatography (SEC) was performed on the G90D mutant.

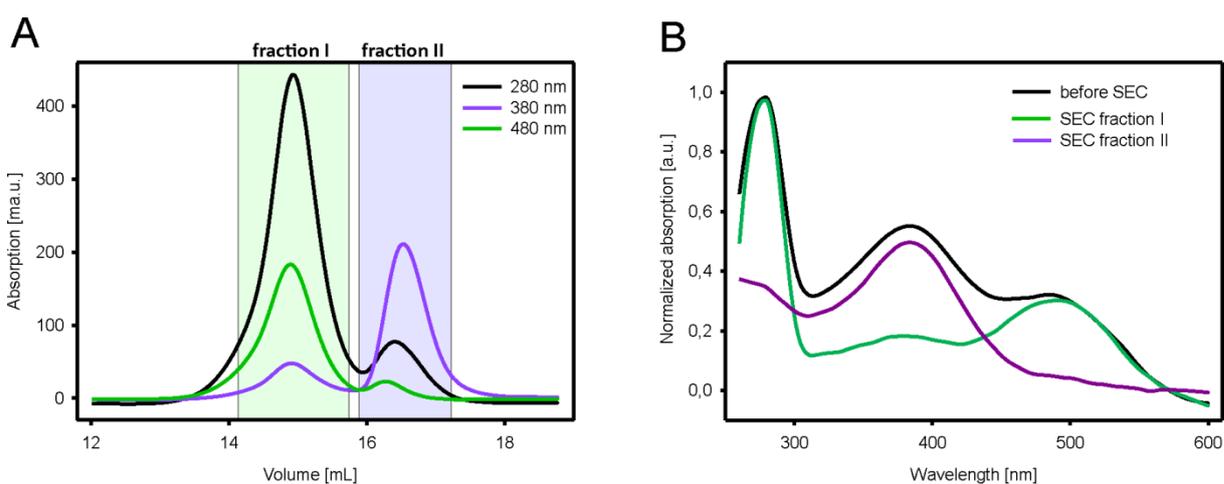


Figure 3.15. Size-exclusion chromatography (SEC) of the G90D mutant. **A** – Analytical chromatogram performed on Superdex 200 10/300 column. The G90D mutant was purified in 1% OG at 4°C under dim light conditions.² The collected fractions are highlighted in green (fraction I) and purple (fraction II). **B** – Absorption spectra of the collected fractions and protein before size-exclusion purification. The spectra are normalized at 280 nm.

SEC of the G90D mutant indicates two prominent signals with the retention volumes of 15.0 ml and 16.5 ml in the fraction I and II, respectively (**Figure 3.15**). Absorption profiles were first monitored at 280 nm, 380 nm and 480 nm and subsequently recorded under dim light condition to avoid illumination. The fractions were collected, concentrated and subjected to independent absorption spectroscopy measurements. According to the absorption spectra shown in **Figure 3.15B**, the first fraction (green) contains pure rhodopsin, while the second fraction comprises free retinal. The spectrum profile of the

² SEC was also performed in 0.05% DDM detergent and can be found in **Appendix A11, Figure A20**.

second fraction with only one absorption maximum at 380 nm and the lack of the protein signal at 280 nm confirm the presence of pure free retinal.

The SEC analysis enables unambiguous assignment of the free retinal and further confirms that no light active Meta state is formed via partial illumination during the purification process. Since the SEC purification results in a 50% loss of the protein yield and the additional free retinal does not affect the structure and kinetics of the mutants measured by NMR, SEC was not applied to these samples.

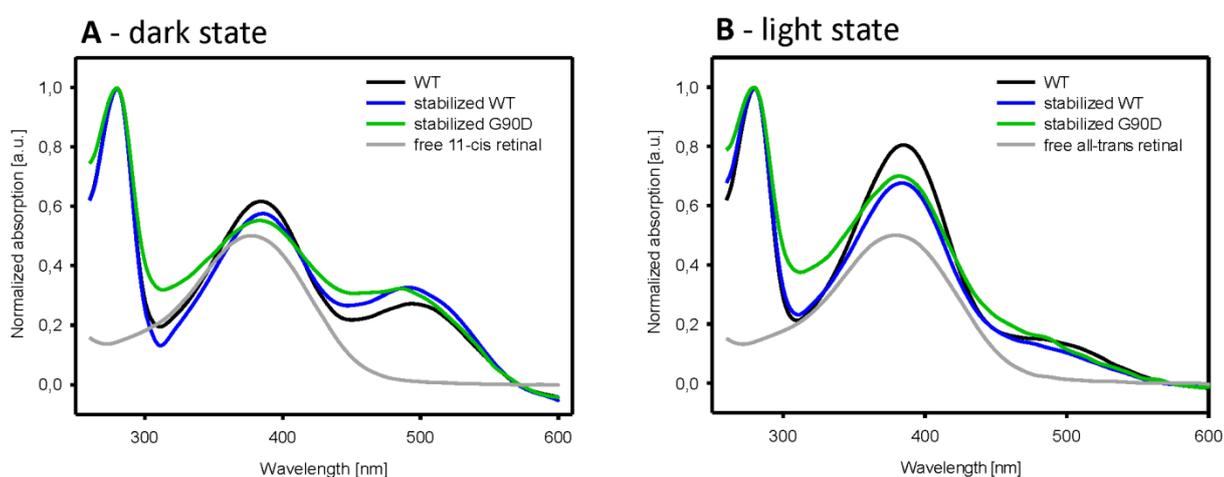


Figure 3.16. UV/Vis-absorption spectra. A – Absorption spectra of the three constructs and free 11-cis retinal recorded under dim light conditions. B – Absorption spectra of the three constructs and free all-trans retinal after illumination. The absorption spectra are normalized at 280 nm.

Light-driven retinal isomerization from 11-cis to all-trans can be explicitly monitored by the blue-shift of the signal in the absorption spectrum (Figure 3.16³). The absorption maximum at approximately 500 nm, which corresponds to the 11-cis bound dark state conformation, disappears after light exposure in all three constructs. Simultaneously, a new signal at 380 nm appears representing the Meta II formation. The concentrations of the wild types and the G90D mutant were calculated according to the light-induced absorption maximum drop at 500 nm and 490 nm, respectively. Since 11-cis and all-trans free retinal isomers absorb at the same wavelength as all-trans bound light active Meta II state (380 nm), it cannot be stated whether the retinal is still bound to the protein after illumination or not. Therefore, liquid state NMR experiments in combination with flash photolysis experiments were performed.

³ In **Appendix A11 Figure A21** the absorption spectra of the dark and the light state are shown separately for each construct.

3.3. Liquid state NMR experiments

Five tryptophan residues presented in the rhodopsin sequence were selectively isotope ^{15}N labeled and used as reporter signals in liquid state NMR experiments. Residues W126, W161 and W265 are highly conserved among GPCRs. They are located in the trans-membrane region and are involved in light-induced conformational changes ^[174]. Two other not conserved amino acids W35 and W175 are placed at the extracellular part of rhodopsin (**Figure 3.17**) and are not affected during the light exposure. W35 is located in H1, W126 – H3, W161 – H4, W265 – H6 and W175 is in E2. The tryptophan signals are sensitive to light-induced conformational changes of the protein, resulting in chemical shift perturbations (CSPs) of the light active conformation. Conformational changes were monitored by the 2D SOFAST-HMQC NMR experiments.

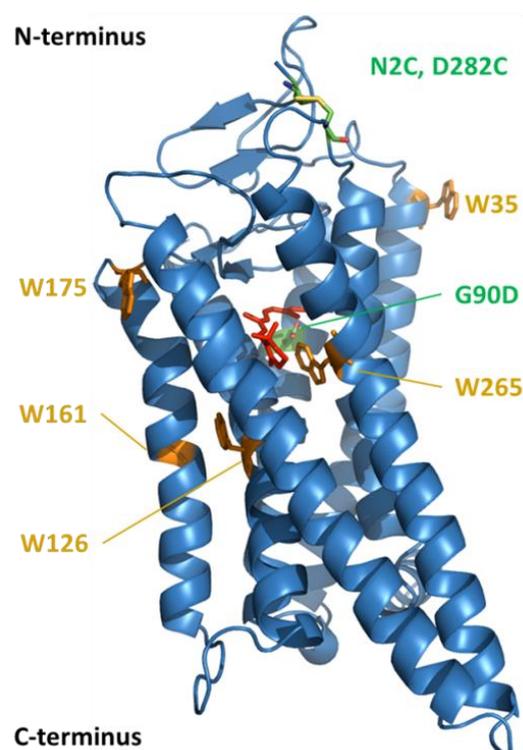


Figure 3.17. Crystal structure of the stabilized wild type in the dark state (pdb: 2j4y) with modeled G90D mutation. The five tryptophan residues are highlighted as orange sticks, mutations are colored in green and the retinal is shown in red.

Figure 3.18 shows 2D SOFAST-HMQC NMR spectra of wild type, stabilized wild type and stabilized G90D mutant acquired under dim light conditions. The indole resonances of stabilized and not-stabilized wild type show the same signal pattern, allowing to transfer the chemical shift assignment from the wild type to the stabilized one (**Figure 3.18A**). From the 2D experiments it can be inferred that the N2C/D282C double mutation does not affect the dark state structure of the wild type.

In contrast to the stabilized wild type, the G90D mutant shows significant CSPs of the W161 and W265 signals and, more importantly, an additional signal at 11.65 ppm (**Figure 3.18B**), indicating a conformational heterogeneity for one tryptophan signal. The negatively charged G90D mutation forms a salt bridge with K296, impairing the native Schiff base (SB) formation between the retinal and K296 ^[193]. Residue W265 is located in close proximity to the retinal binding pocket and is responsible for the stabilization of the dark inactive rhodopsin conformation. An additional signal at 11.65 ppm indicates

conformational heterogeneity of the G90D mutant and cannot be assigned without additional single point mutation experiments.

Temperature series of all constructs in the dark state performed in the range between 283 K and 303 K can be found in **Appendix A12, Figures A23 and A24**.

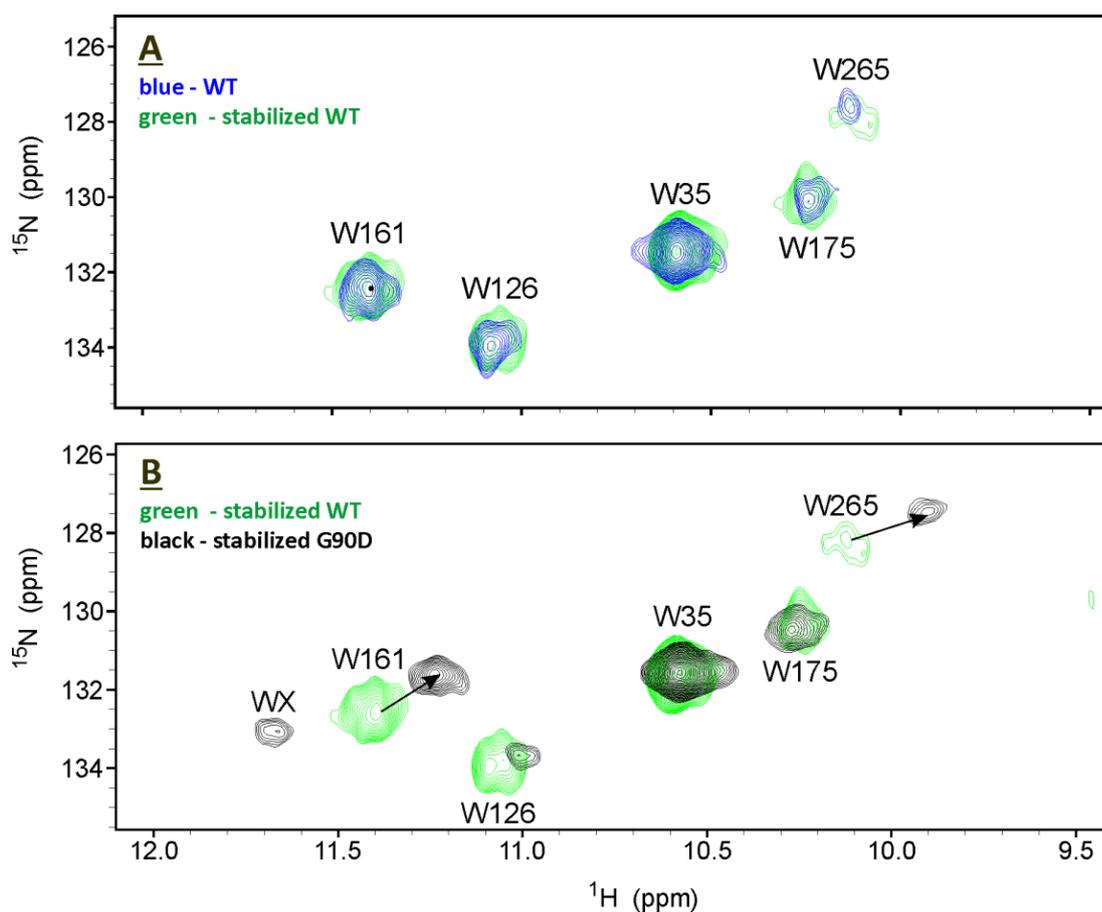


Figure 3.18. 2D SOFAST-HMQC NMR spectra of rhodopsin constructs in the dark state. All spectra were recorded at 950 MHz, 298 K under light dim conditions. **A** – The wild type is shown in blue and the stabilized wild type in green. **B** – The stabilized wild type is colored in green, the stabilized G90D mutant is shown in black. Arrows indicate CSPs of W161 and W265. An additional tryptophan signal in the G90D mutant is assigned as WX.

Light-induced conformational changes of all constructs were monitored by illuminating the sample and recording the 2D SOFAST-HMQC NMR experiments. Comparison of the dark and light state of the wild type, stabilized wild type and the G90D mutant are shown in **Figure 3.18**. Stehle *et al.* extensively studied kinetics of the light activation process of the wild type. Based on the light-induced time resolved CSPs of tryptophan indole resonances they assigned two photocycle pathways, Meta II and Meta III, which occur in parallel with a ratio of 3 : 1 ^[177].

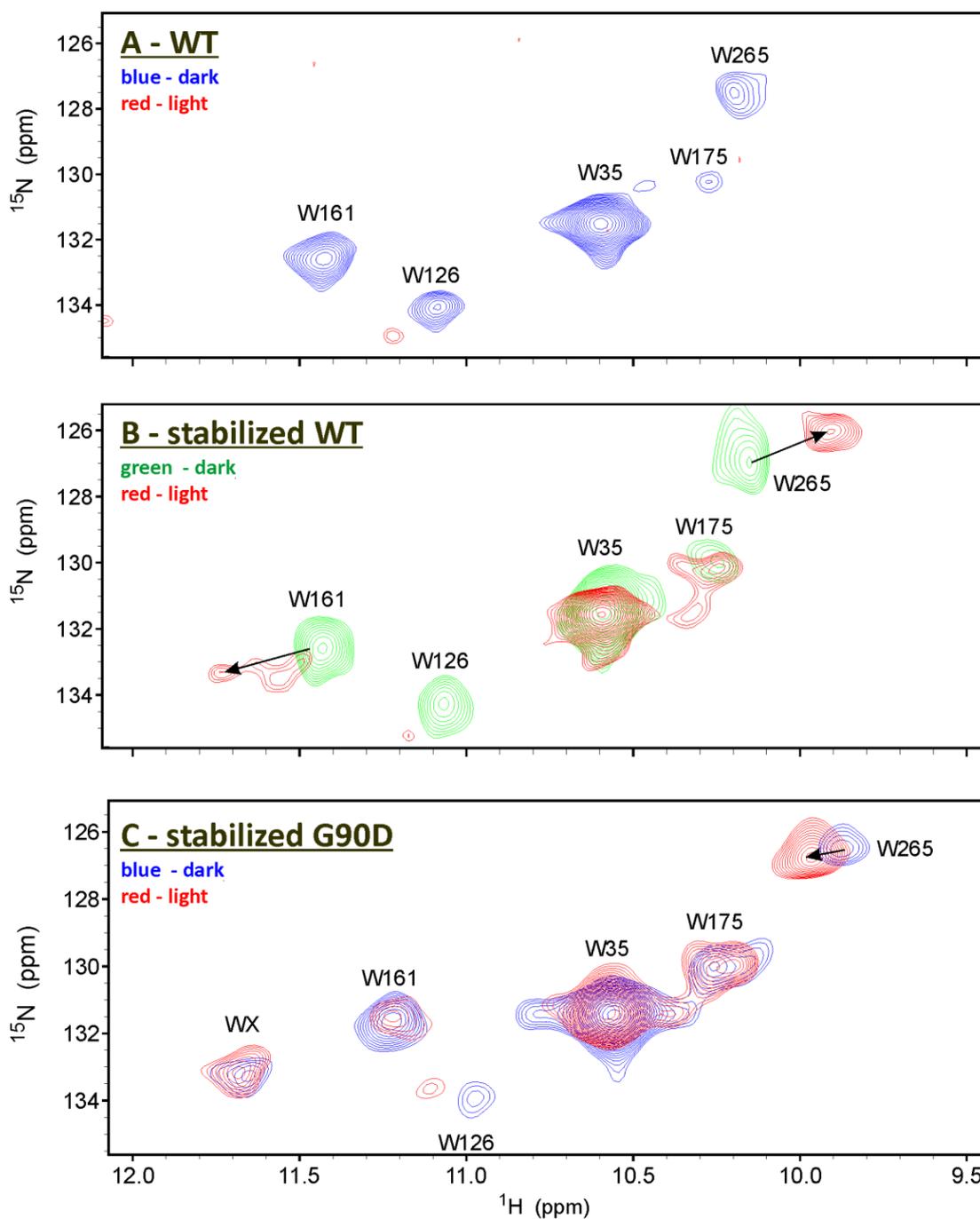


Figure 3.19. Comparison of the 2D SOFAST-HMQC NMR spectra of rhodopsin constructs in the dark and light state. All spectra were recorded at 600 MHz and 298 K. **A** – The wild type dark state is shown in blue, the light state is highlighted in red and is almost not visible compared to spectra shown in B and C, due to sample aggregation after light application. **B** – The stabilized wild type dark state is colored in green, the light state is shown in red. **C** – The stabilized G90D mutant dark state is shown in blue, the light state is pictured in red. Arrows indicate CSPs of W161 and W265. The additional tryptophan signal in the G90D mutant is assigned as WX.

The light-induced photocycle of the not-stabilized wild type is completed within one hour, resulting in opsin and free all-trans retinal ^[177]. Retinal release from the binding pocket is irreversible and leads to sample aggregation of opsin, accompanied with the decrease and vanishing of all tryptophan signal intensities. Due to the low sample concentration, all 2D spectra were recorded within several hours, which is beyond the photocycle regime. Thus, in **Figure 3.19A**, no signals corresponding to the light active wild type are visible. This is in contrast to previously published spectra by Stehle *et al.* ^[177] In these former studies, the high sample concentration allowed to run experiments for shorter times and signals for the light state could be recorded.

The stabilizing N2C/D282C mutation was originally developed to stabilize the protein by forming an additional disulfide bond in the extracellular side ^[206]. According to **Figure 3.18A**, this mutation does not significantly influence the conformation of rhodopsin in the dark state but stabilizes the light state. In contrast to the wild type construct, the stabilized sample remained stable for several days after illumination. This was also confirmed by mass spectrometry (MALDI), which a priori detects the light state. Thus, the stabilized constructs showed a MALDI signal that correspond to the expected molecular weight of 42 kDa, while wild type sample degraded upon light exposure (**Appendix A11, Figure A22**).

The tryptophan resonances of the stabilized wild type in the light state are in agreement with the chemical shift assignment of the wild type in the Meta II state trapped one minute after illumination ^[177]. Similar to wild type, W35 and W175 from the extracellular domain remain unaffected by the light exposure, showing no CSPs under light activation (**Figure 3.19B**). W126, W161 and W265, located in the transmembrane region in turn are highly sensitive to the light-induced conformational changes ^[174], resulting in significant CSPs for these signals. The observed downfield shift of the W161 signal and the upfield shift of the W265 signal have been previously reported for the light active state of the wild type ^[177]. W161 is located in the middle of H4, which is not directly involved in the light-induced conformational changes but due to the direct connection to the H3, which undergoes large structural rearrangements, it is co-affected during the light exposure and is involved in the Meta II state formation. W265 is located in the binding pocket and its interaction via the indole ring with the β -ionone group of the retinal is highly sensitive to the light-induced retinal isomerization, taking part in Meta II and Meta III conformations. W126 of the wild type, as a part of H3, undergoes large structural reorientations caused by light-induced retinal isomerization. In the non-stabilized wild type rhodopsin, this signal shows Meta II and Meta III kinetics with half-life times of 4.2 minutes and 24.3 minutes, respectively. After one minute, the signal undergoes a strong upfield shift, resulting in a high intensity signal, which completely disappears after 25 minutes ^[177]. For the stabilized wild type, the W126 signal is almost not visible upon light exposure. Overall, upon illumination the stabilized wild type is trapped in the light active Meta II

state with characteristic chemical shift perturbations of the most photosensitive W265, W161 and W126 signals.

In contrast to the stabilized wild type, the chemical shifts of the tryptophan indole signals of the G90D mutant do not undergo significant perturbations upon illumination, indicating no major movement of the helices (**Figure 3.19C**). Even W265, which is located in the binding pocket and is highly sensitive to the retinal conformation, shifts downfield (vs. upfield shift of wild type) much less compared to the stabilized wild type.

In order to observe the real-time changes of indole resonances induced by the light exposure of the stabilized G90D mutant and to compare and confirm with the previous results, time resolved NMR experiments of G90D kinetics were performed. Therefore, first, the sample was illuminated inside the NMR spectrometer using an argon ion laser for 0.5 seconds, and afterwards pseudo 2D NMR experiments were recorded. Due to the low signal-to-noise ratio, 16 successively recorded 1D planes were summed to represent the light state recorded 10 minutes after the illumination. The spectrum recorded at the end of the experiment corresponds to the 12 hours of measurement time. No significant difference between the dark state and the light state (recorded 10 minutes after the illumination) and the last 1D plane recorded after 12 hours of the experiment series could be observed (**Figure 3.20**).

Although in the dark state 11-cis retinal is bound to the Schiff base (according to absorption and 1D NMR spectroscopy), the structural geometry of the G90D mutant in the dark state is similar to the light active conformation. Taken together, the G90D mutation leads to the retinal binding pocket disruption, which results in the conformational heterogeneity and the pre-active dark state conformation. After illumination no significant conformational changes could be observed, suggesting that all-trans retinal is still incorporated in the binding pocket.

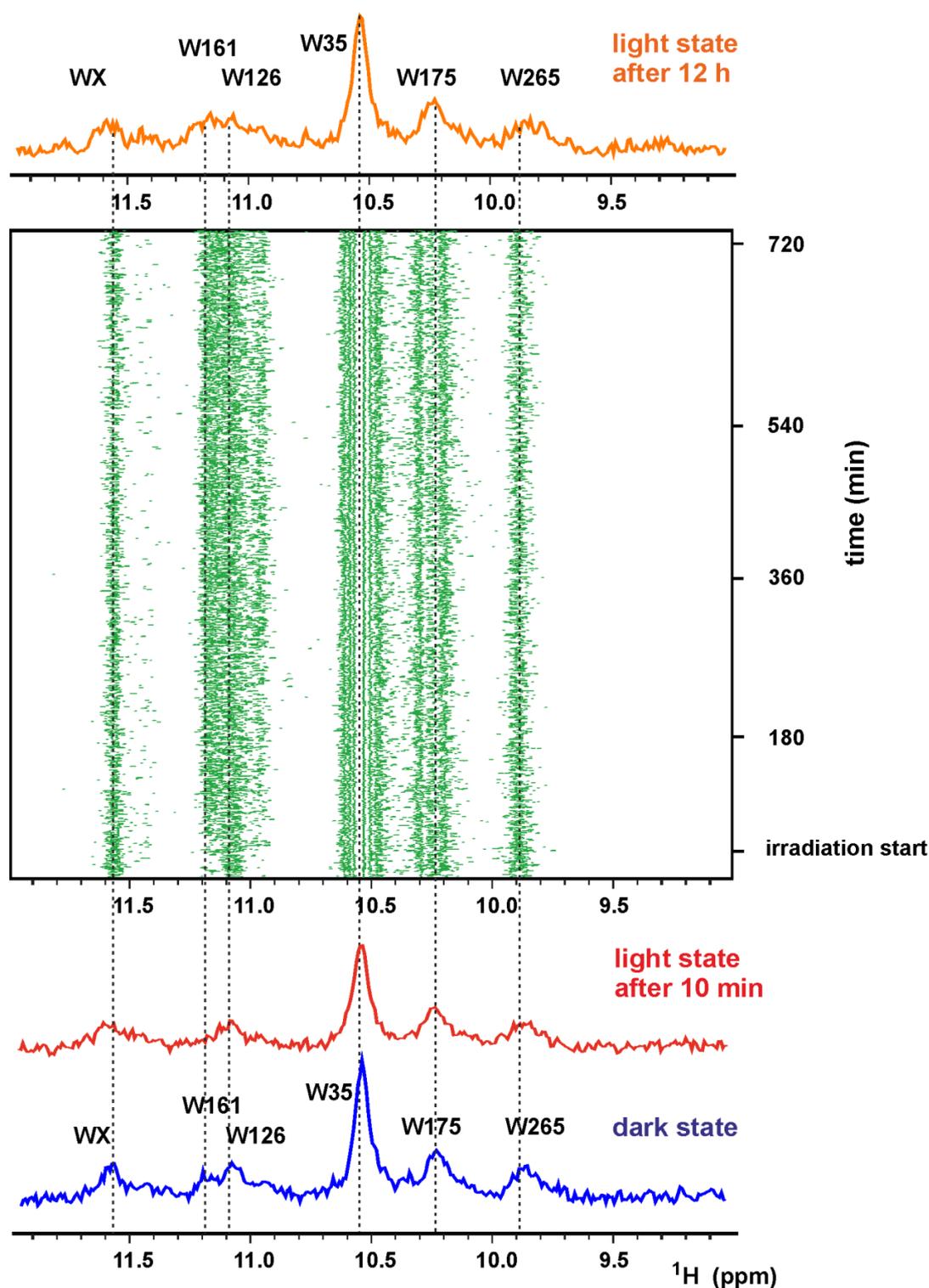


Figure 3.20. NMR kinetics of the stabilized G90D mutant. A series of 1D ^1H NMR spectra of the stabilized G90D mutant recorded at 800 MHz, 298 K, at different time intervals after illumination of 3 W for 0.5 seconds. The indole region of the spectrum is shown. The blue spectrum corresponds to the dark state recorded before the illumination, the red one represents the light active state after 10 minutes and the orange spectrum displays the indole resonances after 12 hours of experiment time. Due to the low signal-to-noise ratio, 16 1D planes recorded successively were summarized to represent red and orange spectra. No difference between red and orange spectra could be observed. The dash lines connect the chemical shift assignment of the indole tryptophan signals.

3.4. Photo flash and ultrafast kinetic experiments

The photocycle of the stabilized G90D mutant was also investigated by flash photolysis performed by Dr. Elias Eckert from Prof. Wachtveitl group. The samples were photo excited with a home-built flash photolysis setup with a single ns pulse at the wavelength of 500 nm and kinetics were measured over 120 minutes with an interval of 0.5 minutes. To reduce the external influence of the environment, the buffer conditions were kept identical to those used for liquid state NMR measurement. The results are presented in a lifetime density maps (LDMs) profile ^[220], which consists of four major and two minor distributions, called **C - D** and **A - B**, respectively (**Figure 3.21A**). Signature **A** represents free retinal absorption and **B** stands for the ground-state bleaching. Both belong to the long-lived non-decaying components, which are similar for each sample and do not carry any relevant photodynamic information. In contrast, short lifetime components **C** and **D** describe Meta II decay and Meta III formation, respectively. **E** is assigned to the longer lifetime Meta III decay and **F** corresponds to the free retinal formation ^[221]. The positive amplitude is highlighted in red and corresponds to the absorption decay, while the negative area is colored in blue and represents the absorption increase. Furthermore, in order to resolve early photo kinetic processes, additional ultrafast absorption experiments, comprising the first minute with the intervals of 100 ms were performed.

The signature patterns of the wild type and stabilized wild type are very similar but differ significantly in the lifetime scale. For both constructs, no early intermediate photorhodopsin state could be observed separately, which might be explained by the insufficient resolution of the measurements. The next photo intermediate state bathorhodopsin of the stabilized wild type is formed four times slower compared to the not-stabilized one. In contrast to the biexponential decay of wild type bathorhodopsin, for the stabilized wt only one process (1 ns) could be detected. Similarly to the wild type, the stabilized wild type forms a light active Meta II and Meta III states, which decay accompanied by Schiff base hydrolysis followed by the all-trans retinal release. However, due to the stabilizing mutations, the intermediates decay more slowly, the covalently bound retinal is stabilized and its hydrolysis is significantly delayed, which is in line with NMR observations. Meta II decay (signature **C**) and Meta III formation (signature **D**) take place within nine minutes, which is three minutes longer compared to the not-stabilized wild type, while further decay of Meta III state (signature **E**) of the stabilized wild type is delayed by a factor of four. A graphical representation of the analysis of LDMs and ultrafast experiments is shown in **Figure 3.21B**.

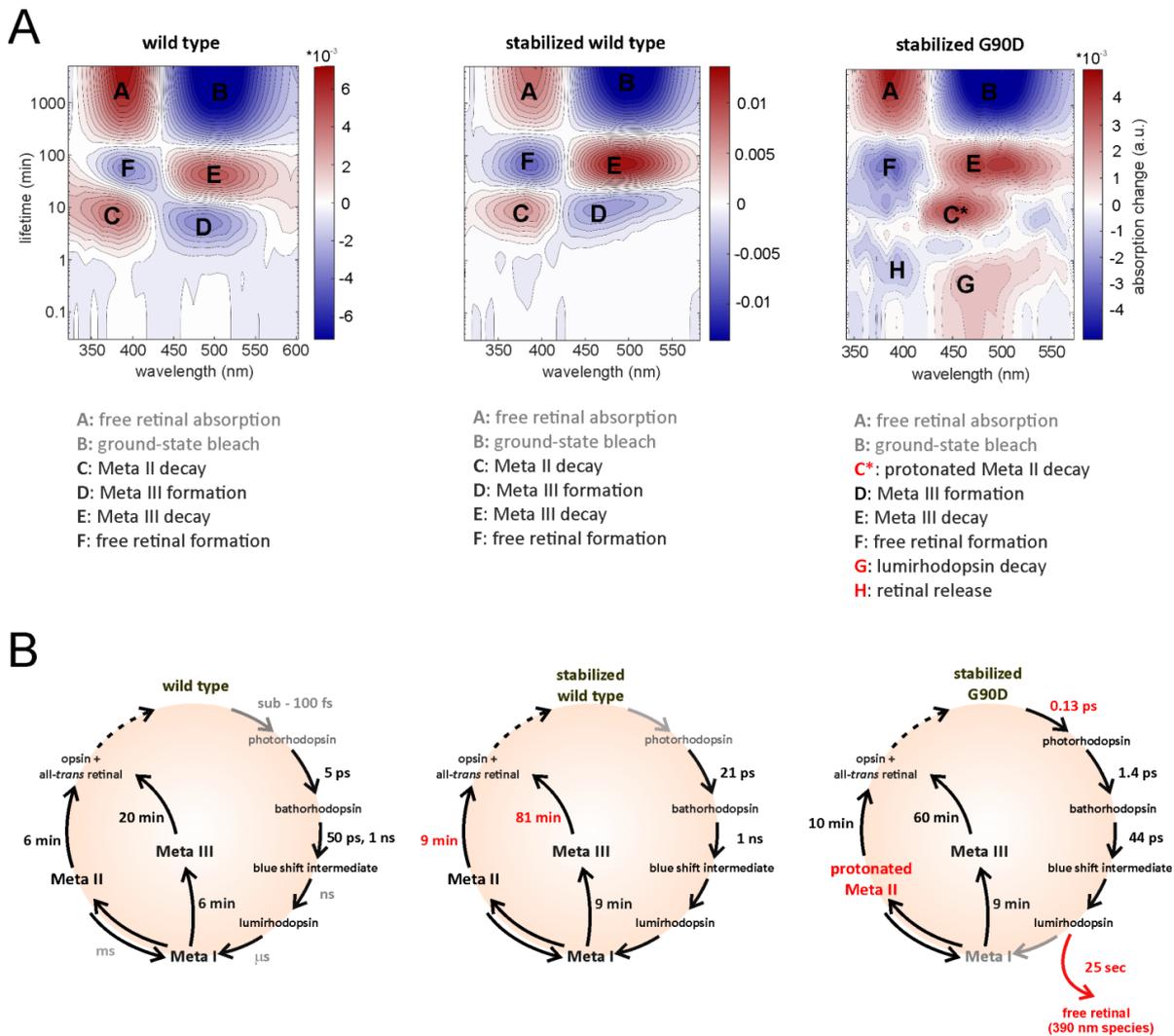


Figure 3.21. Photo flash kinetic analysis of three constructs. A – Lifetime density maps of the transient absorption data. Negative absorption decay is colored in blue and positive is highlighted in red. **B** – Schematic representation of the analysis of the lifetimes. Estimated lifetimes reported in the literature but not measured on these constructs are shown in grey.

The spectral differences between the stabilized wild type and the stabilized G90D mutant are clearly visible by the direct visual comparison of their signature pattern. In contrast to wild type, the early intermediate state photorhodopsin, which is formed within 0.13 ps, could be detected for the G90D mutant. The ability to observe this photo intermediate state indicates the delayed retinal isomerization caused by the G90D mutation. According to the UV/Vis and infrared studies of the G90D mutant performed by Fahmy *et al.*, E113 is protonated in the dark state, perturbing retinal binding and resulting in delayed primary photoproducts formation [222]. In turn the formation of the next photo intermediate state bathorhodopsin of the G90D mutant is 20 times faster than for the wild type, while its decay to the blue shifted intermediate state is significantly slower (44 ps for G90D vs. 1 ns for wt). At 0.4 minutes, a weak absorption decay at approximately 490 nm is observed [220]. Due to the specific absorption maximum, this signature is assigned to the photo intermediate state lumirhodopsin (signature **G**). This

conformation is accompanied by an absorption increase at 390 nm, which is referred to the partial retinal release (signature **H**) that takes place before the active intermediate states are formed. At 10 and 60 minutes additional absorption decays at 460 nm and 480-520 nm, respectively, are observed (signature **E** and **C***). Based on FTIR studies, Zvyaga *et al.* reported an intermediate species with an absorption maximum at 468 nm. They were characterized by a protonated Schiff base, that could activate transducin and were structural similar to the active Meta II state ^[215]. Therefore, the positive signature observed at 460 nm with the lifetime of 10 minutes (signature **C***) was assigned to the protonated Meta II decay. The second process at 60 minutes is the delayed Meta III decay (signature **E**), which evolves in opsin and free retinal. Absorption increase at 390 nm (signature **F**) could be interpreted as an accumulation of free retinal, which occurs even before the Meta III decay starts. Meta III formation and Meta II decay processes of the G90D mutant are in the same regime as for the stabilized wild type, while Meta III decay takes 20 minutes faster.

Typical signature characteristics for the deprotonated Meta II state formation (signature **C**) could not be detected. Due to the similar absorption maximum of inactive Meta I and active protonated Meta II state (478 nm and 460 nm, respectively) and poor resolution in the early photo kinetics, no Meta I state was detected. It is not proven whether Meta I state of the G90D mutant exists or the protonated Meta II conformation develops directly from lumirhodopsin.

Taken together, the N2C/D282C mutations stabilized the light active conformation of rhodopsin delaying the relaxation of its Meta states. The photocycle of the G90D mutant is unique, showing a protonated Meta II state and an untypical retinal behavior. Retinal isomerization is delayed, while the retinal release is faster, suggesting its release even before the active intermediates are formed.

3.5. Solid state NMR experiments

Liquid state NMR experiments reported that the G90D rhodopsin mutant shows a protein conformational heterogeneity and Meta II like pre-active state in the dark. In contrast to the stabilized wild type, no significant conformational changes could be detected after illumination. This suggests a fixed all-trans retinal position in the binding pocket in the light active conformation. If the retinal is covalently bound via the Schiff base to the protein or only sterically trapped in the binding pocket remains unclear. Photo flash kinetic experiments of the G90D mutant showed a protonated light active Meta II state and a unique retinal kinetics. Retinal isomerization and, therefore, the first transition to the photorhodopsin is delayed, while later retinal release is fast and occurs even before Meta products are formed. However, the detailed description of the retinal conformation in the dark and light state is not covered by this method.

Generally so far, very little is known about the retinal conformational changes induced by the G90D mutation. No dark state crystal structure of the G90D mutant is available and the poor resolution of the light state structure does not allow the retinal conformational analysis at the atomic level. Based on the crystallographic refinement, Singhal *et al.* proposed a mixture of different cis-isomers existing in the light state ^[209].

3.5.1. Dark state experiments

In order to provide insight into the exact retinal conformation of the G90D mutant, DNP enhanced solid state MAS (magic angle spinning) NMR experiments were performed. Therefore, retinal was ¹³C-isotope selectively labeled at the positions C14, C15 (14,15-¹³C) and C12, C13, C20 (12,13,20-¹³C). Furthermore, to conduct ¹⁵N NMR experiments, ¹⁵N lysine isotope labeling was performed.

DNP-enhanced double quantum-filtered (DQF) ¹³C spectrum of 14,15-¹³C labeled retinal in the dark state is shown in **Figure 3.22B**. The chemical shift assignment was performed according to the wild type rhodopsin resonance assignment published by Patel *et al.* ^[171] Two signals corresponding to C14 and C15 could be identified. C14 shows conformational heterogeneity, two populations can be observed. The major conformation resonates at 123 ppm, while the minor conformation is downfield shifted with the chemical shift of 125 ppm. C15 signal is homogeneous and resonates at 168 ppm.

To clarify, whether the retinal is bound to the Schiff base, a 1D ¹⁵N spectrum was recorded. Among 11 lysine residues in the rhodopsin sequence, the Schiff base (K296) has a characteristic chemical shift resonance. The broad ¹⁵N chemical shift distribution of protonated (156 ppm ^[218,223]) and deprotonated SB (315 ppm ^[223]) allows their unambiguous assignment. Similarly to the wild type ^[218,223], the 1D ¹⁵N spectrum shown in **Figure 3.22A** clearly indicates a protonated Schiff base signal (pSB) at 179.5 ppm and the absence of any peak in the deprotonated SB area. Downfield to the main pSB signal (179.5 ppm) a shoulder around 188.5 ppm was detected, which might indicate a minor protein conformation. A broad, intense signal around 120 ppm corresponds to the backbone of ¹⁵N lysine residues and other natural abundance amino acids.

To prove the conformational heterogeneity by ¹³C retinal and ¹⁵N protein spectra, a 2D ¹⁵N-¹³C TEDOR dipolar through-space correlation experiment was performed. Both major retinal signals (C14 and C15) correlate with the ¹⁵N protonated SB (**Figure 3.22C**). The minor conformation of retinal represented by C14 (125 ppm) also shows a strong cross peak with pSB. Moreover, the minor pSB signal (188.5 ppm) also correlates with retinal C14, C15 carbon atoms. The signal intensity is quite low and only visible for the major retinal conformation. However, the cross peak to the minor C14 conformation cannot be excluded and could be beyond the noise level.

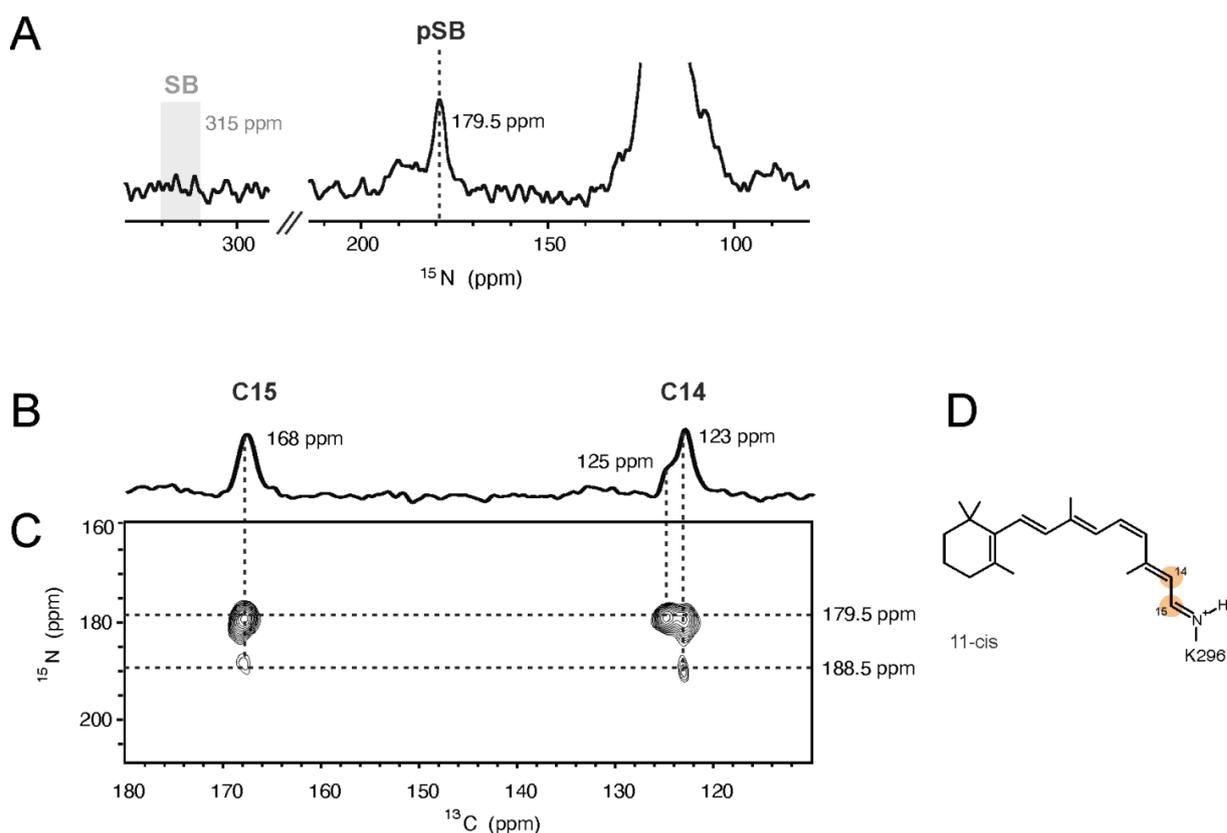


Figure 3.22. DNP-enhanced spectra of the G90D mutant in the dark state. **A** – DNP-enhanced, ^{15}N spectrum of the ^{15}N lysine isotope labeled G90D mutant. The signal at 179.5 ppm corresponds to the protonated Schiff base. The deprotonated Schiff base signal expected at 315 ppm could not be detected. **B** – DNP-enhanced double quantum-filtered (DQF) ^{13}C spectrum of 14,15- ^{13}C retinal. **C** – DNP-enhanced ^{15}N - ^{13}C TEDOR spectrum of the 14,15- ^{13}C retinal in the ^{15}N lysine G90D mutant. **D** – 11-cis retinal labeling schemes used in these experiments. Orange circles represent ^{13}C labeling positions.

To find out whether other retinal atoms are involved in the conformational heterogeneity, an additional retinal labeling scheme was performed. C12, C13 and C20 were chosen as the most light sensitive retinal carbons, which are in close proximity to the Schiff base. The chemical shift assignment was performed using double quantum-single quantum (DQ-SQ) correlation experiment (**Figure 3.23**) and are in agreement with Patel *et al.* [171] None of these signal profiles reflects structural heterogeneity. Therefore, the G90D resonances were compared with the stabilized wild type and a CSP analysis was performed. The overlay of the 2D DQ-SQ spectra of the stabilized G90D mutation with the stabilized wild type is shown in **Figure 3.23B and C**. In contrast to C13, whose chemical shift remains similar for both constructs, C12 and C20 show significant CSPs. C12 of the G90D mutant is 0.5 ppm and the C20 methyl group is 0.7 ppm downfield shifted.

To sum up, 11-cis retinal is bound to the Schiff base of the G90D mutant in the dark state. The observed conformational heterogeneity of the G90D mutant affects its Schiff base and the retinal C14 carbon. A minor protein conformation correlates with the C14 and C15 retinal atoms. The C14 retinal carbon signal

indicates a second minor population. C20 and C12 signals of the G90D mutant show small chemical shift perturbations compared to the wild type but no signs of a second retinal conformation.

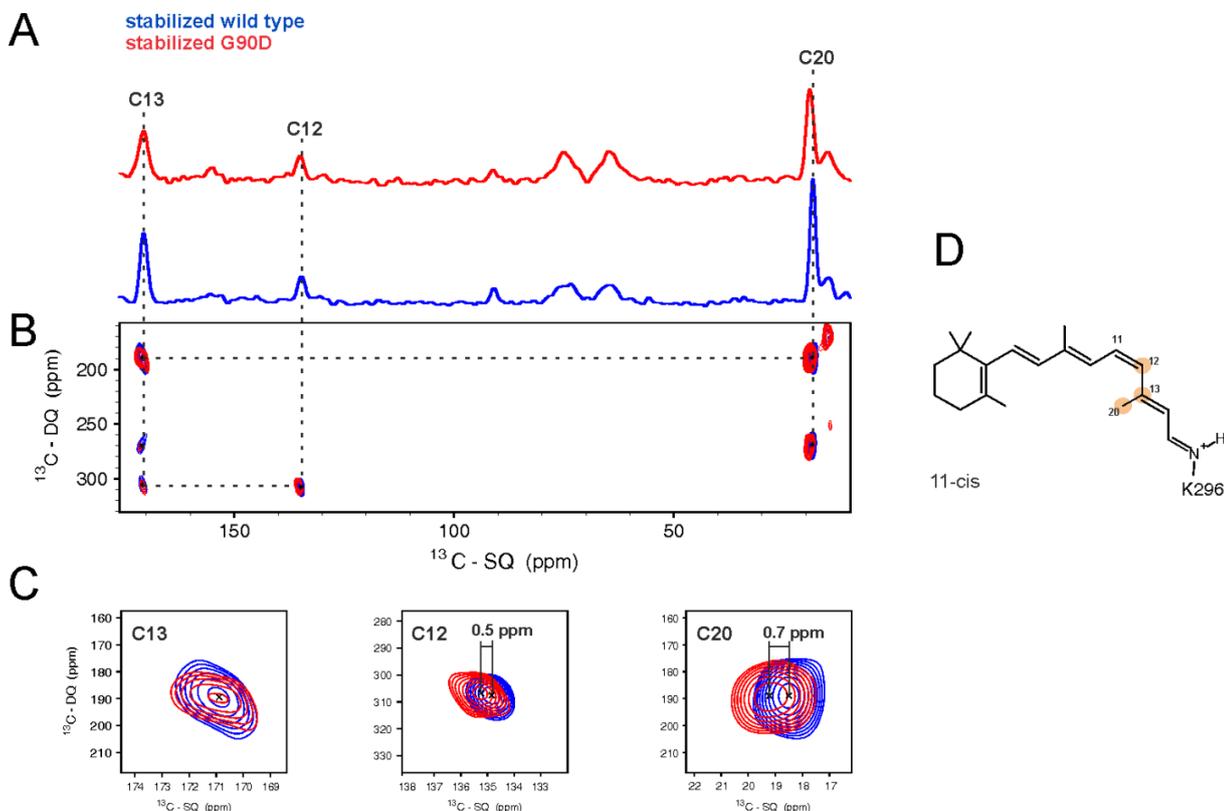


Figure 3.23. Comparison of DNP-enhanced spectra of the stabilized G90D mutant with stabilized wild type in the dark state. **A** – DNP-enhanced, double quantum-filtered (DQF) ^{13}C spectrum of 12,13,20- ^{13}C retinal in the stabilized wild type (blue) and stabilized G90D mutant (red). **B** – DNP-enhanced, double quantum-single quantum (DQ-SQ) spectra of 12,13,20- ^{13}C retinal. **C** – Zoom in of C12, C13 and C20 retinal signals from DQ-SQ spectra. **D** – 11-cis retinal labeling schemes used in these experiments. Orange circles represent ^{13}C labeling positions.

3.5.2. Light state experiments

Further characterization of the retinal conformational changes induced by illumination and the analysis of the impact of the G90D mutation on the binding pocket geometry would provide important insight into a disease modified photocycle. Therefore, DNP-enhanced MAS NMR experiments were performed on 14,15- ^{13}C labeled retinal, incorporated in the ^{15}N lysine stabilized G90D mutant. To generate photo intermediate state, first the sample was illuminated in situ in the MAS rotor under cryogenic conditions (100 K). The illumination was performed for two hours with blue light. Conduct of illumination at 100 K allowed trapping of the bathorhodopsin state [224] in the NMR experiments. Subsequently, the illuminated sample was warmed to room temperature, which corresponds to the conditions used for generating of the light active Meta II conformation by Patel *et al.* [171] 1D ^{15}N and ^{13}C 14,15-retinal DQF spectra of the dark state and the photo intermediate products of the G90D mutant are shown in **Figure 3.24**.

Similar to the dark state, the batho and Meta II states show characteristic protonated Schiff base signals at 179.5 ppm. For both intermediates the pSB signal profiles are broader than for the dark state, with an additional maxima shifted 3 ppm upfield (**Figure 3.24A**). However, the low signal-to-noise ratio does not allow unambiguous assignment of these shapes to the second conformational population. The signals in the protonated SB region (179.5 ppm) and the absence of any signals in the deprotonated SB area (around 315 ppm) indicate a protonated Meta II conformation, which is in perfect agreement with the flash photolysis analysis discussed above.

In **Figure 3.24B** the C14, C15 retinal carbon atoms are used as indicators for the conformational changes of the retinal under light exposure. DQF spectra of bathorhodopsin show a new C14 signal appearing at 119.8 ppm. This observed 3.2 ppm upfield shift is in perfect agreement with the wild type bathorhodopsin resonances (4.8 ppm upfield shift) reported by Concistre *et al.* ^[224] Less intense signal at 123 ppm corresponds to the residual dark state conformation that might remain due to an incomplete illumination process.

Carbon C14 in Meta II state of the G90D mutant shows conformational heterogeneity with residual signals corresponding to the dark state (123 ppm) and bathorhodopsin (119.8 ppm), which might be attributed to the mixture of different retinal conformations. This hypothesis would be in partial agreement with crystallographic refinement data of the G90D mutant, published by Singhal *et al.* ^[209] However, the reported mixtures of different cis-isomers could not be detected in our experiments. A new signal at 132.6 ppm provides evidence for Meta II formation and is in agreement with downfield shifted wild type Meta II C14 resonance published by Patel *et al.* ^[171] A closer examination of C14 DQF spectra shows that even in bathorhodopsin a low intensity Meta II signal around 120 ppm is visible. This was not observed for the wild type ^[225] and might indicate the presence of the light active state in earlier stages of the photocycle, which could possibly result from the pre-active conformation in the dark state.

The second retinal carbon signal C15 does not undergo significant perturbations, showing slight downfield shift, which occurs in the opposite direction than previously published for Meta II and batho wild type.

The observed ¹³C chemical shifts of retinal C14 and C15 carbons measured for dark state and light intermediates are summarized and compared with published wild type data in **Table 7**. Since the wild type retinal resonances of batho and Meta II states were published by different groups and are not absolutely identical, analysis of the chemical shift perturbations of the light intermediates was performed by calculating the difference between the dark states, which was recorded before each light experiment, and the corresponding photo product. Light-induced resonance changes are visualized in **Figure 3.25**.

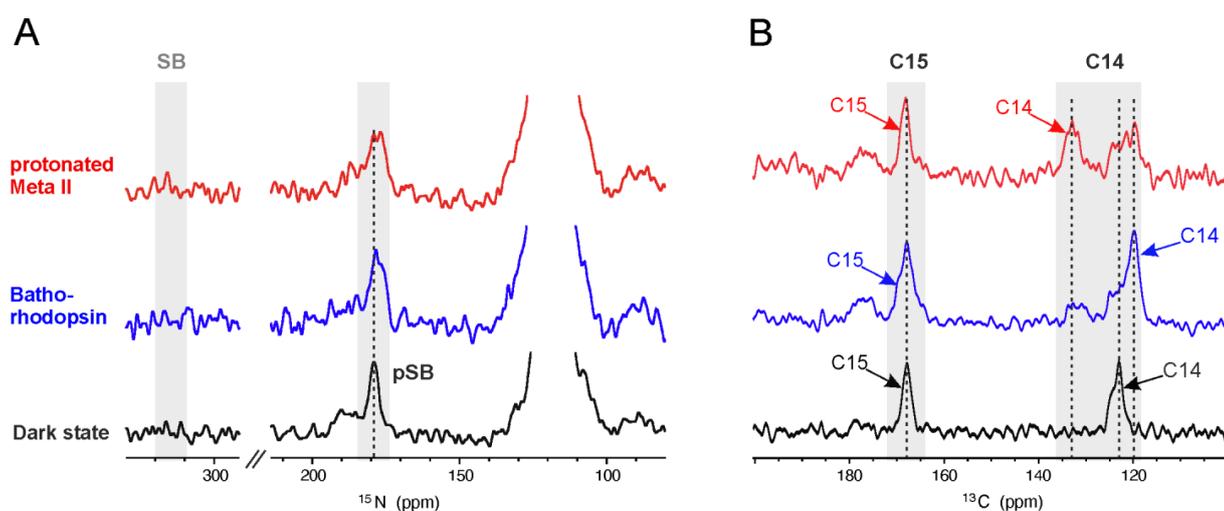


Figure 3.24. DNP-enhanced spectra of the G90D mutant under illumination. Three states were detected: dark state (black), bathorhodopsin (blue) and light active Meta II (red) as discussed in the main text. **A** – DNP-enhanced, ^{15}N spectrum of ^{15}N lysine isotope labeled G90D mutant. The signal at 179.5 ppm, which corresponds to the protonated Schiff base, is visible in every state. The deprotonated Schiff base signal expected at 315 ppm is not detected in any state. **B** – DNP-enhanced double quantum-filtered (DQF) ^{13}C spectrum of 14,15- ^{13}C retinal. C14 and C15 signal areas are highlighted in gray. For each signal corresponding to the respective state, the same color code was used.

Table 7. Chemical shifts of the G90D mutant observed for ^{13}C 14,15-retinal in the dark, bathorhodopsin and light active Meta II conformation. Resonances for the wild type (shown in gray) bathorhodopsin were taken from Concistre et al. ^[224] and for Meta II state from Patel et al. ^[171] Chemical shift perturbations are calculated as the difference between dark state and the corresponding intermediate conformation. Positive and negative sign reflects upfield and downfield shift of the light products, respectively.

	C14				C15			
	G90D, ppm	$\Delta\delta$, ppm	wt ^{[171][224]} , ppm	$\Delta\delta$, ppm	G90D, ppm	$\Delta\delta$, ppm	wt ^{[171][224]} , ppm	$\Delta\delta$, ppm
Meta II	132.6	-9.6	126.3 ^[171]	-4.7	168.1	- 0.4	162.4 ^[171]	2.7
batho	119.8	3.2	117.5 ^[224]	4.8	169.6	- 1.9	164.4 ^[224]	1.0
Dark state	123.0		121.6 ^[171] 122.3 ^[224]		167.7		165.1 ^[171] 165.4 ^[224]	

The retinal carbon C14 undergoes the largest light-induced conformational changes, while C15 appears to be less affected by light exposure. Comparison with the wild type light-induced retinal conformational changes is shown in **Figure 3.25**. C14 of G90D in bathorhodopsin state is in a similar

spectroscopic range as the wild type, while the G90D Meta II state formation induces significantly stronger deshielding effect compared to the wild type CSP. The small deshielding effect of C15, observed for the G90D mutant, can be attributed to the differences in the environment conditions, used for the NMR experiments. Thus, the Meta II state, reported by Patel *et al.* [171] was measured in detergent (DM), while Concistre *et al.* [224] reconstituted bathorhodopsin in a natural-composition lipid membrane. The G90D mutant was supplemented with DOPC lipids as described in **chapter III 2.4.3**.

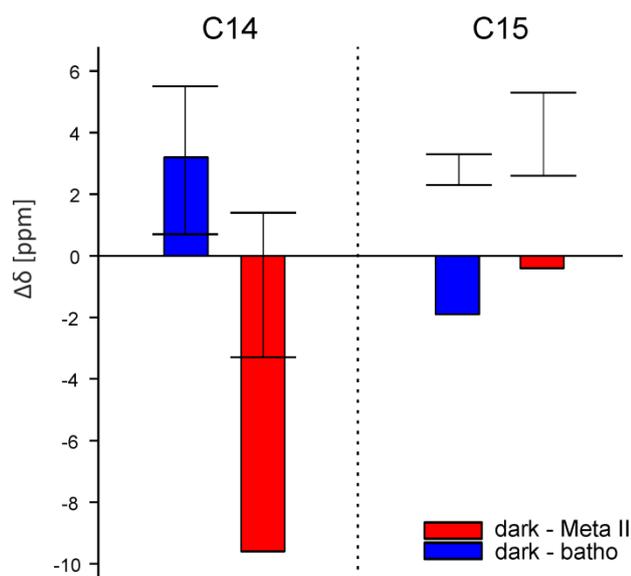


Figure 3.25. Graphical representation of light-induced chemical shift changes of the ^{13}C 14,15-retinal G90D mutant. Resonance differences between dark state and bathorhodopsin are shown in blue, while differences between dark state and Meta II conformation are colored in red. The black bars represent equivalently calculated retinal CSP, reported for the wild type [171] [224].

Summarizing, solid state NMR experiments, performed on the G90D mutant, confirm conformational heterogeneity observed by liquid state NMR in the dark state. Furthermore, it was shown, that the G90D point mutation causes a second 11-cis retinal population, which is not derived from the light intermediates. Both retinal populations are attached to the Schiff base, limiting the retinal heterogeneity to the slight changes in the steric distribution in the binding pocket.

Light cryotrapping experiments allowed detecting a bathorhodopsin state and light active Meta II state. The Meta II conformation was found to be protonated, which is consistent with the photo flash results. Moreover, a minor, light active state was detected in the bathorhodopsin conformation. This might result from the pre-active dark state conformation, observed by liquid state NMR, or from the perturbed photo cycle, where Meta II is partially formed during the early photocycle steps. However, the last hypothesis is not confirmed by ultrafast kinetic experiments.

4. Discussions and perspectives

Although rhodopsin is one of the most widely studied GPCRs, its photocycle is still not fully understood. This part of the doctoral thesis focuses mainly on the investigation of the G90D mutant, which causes congenital stationary night blindness (CSNB) in humans. The overall goal is to clarify the origin of the constitutively active nature of the mutant and to characterize its influence on the visual cycle. In particular, it is of interest to identify whether the increased basal activity and therefore developed retinal dystrophy is caused by the pre-active dark state or by the spontaneous retinal isomerization in the binding pocket. To answer these questions and to shed light on the mechanism of the photocycle of the G90D mutant, different complementary methods were applied.

As known, the rhodopsin photocycle is triggered by light-induced chromophore isomerization. In the dark inactive state, 11-cis retinal is covalently bound to K296 via a protonated Schiff base. An important stabilizing role plays the negatively charged E113 residue, which acts as a counter ion for the protonated Schiff base in the dark state. 11-cis to all-trans isomerization induces a series of intermediate photoproducts, resulting in an active Meta II conformation, which in turn is characterized by deprotonated Schiff base and all-trans covalently attached retinal. The proton transfer from the all-trans Schiff base bound retinal to the negatively charged counter ion E113 is a key process in the activation switch and retinal release. An alternative relaxation pathway of the Meta I state involves Meta III conformation, which is considered as a storage conformation of inactive rhodopsin and comprises protonated retinal in its all-trans-15-syn conformation. Both intermediates, Meta II and Meta III, subsequently decay to opsin and free all-trans retinal.

The constitutively active G90D rhodopsin mutant is known to cause congenital stationary night blindness (CSNB) disease. This single point mutation is located in the retinal binding pocket in close proximity to the residues K296 and E113. In 1994, Rao *et al.* showed that the charged D90 side chain interacts with the Schiff base K296, perturbing the salt bridge connection between E113 and K296^[193]. Residue E113 is, therefore, neutralized and exists in the protonated form in the dark state. Two years later Zvyaga *et al.* confirmed this suggestion and reported protonated active Meta II state of the G90D mutant^[215,222].

To provide atomic details of the retinal binding pocket of the G90D mutant and its disease related mechanism Singhal *et al.*^[209] aimed to crystalize this mutant in its dark and light conformation. Unfortunately, no crystals could be obtained for the dark state.

According to Singhal *et al.*^[209], the quality of the retinal incorporation rhodopsin formation, can be estimated from the maximum absorption ratio at 280 nm and 480 nm (A_{280}/A_{480}). For the stabilized wild

type this ratio was in the range between 1.6 to 1.7 and for the G90D mutant 2.0-2.2. Due to the higher values for the G90D mutant, the authors hypothesized a conformational heterogeneity consisting of opsin and rhodopsin, which is in line with their unsuccessful dark state crystallization trial. The NMR experiments performed during this doctoral thesis confirm impeded retinal binding for the constructs with additional stabilizing disulfide bridge. Thus, it was shown that for the successful retinal binding the wild type purification protocol has to be optimized for the constructs containing the stabilizing mutation. However, neither liquid nor solid state NMR revealed a retinal free opsin conformation. An additional tryptophan signal detected by liquid state NMR in combination with a second population of the Schiff base and attached retinal, identified by solid state NMR indicate the conformational heterogeneity of the system. However, this heterogeneity is caused by a second minor populated protein conformation with incorporated 11-cis retinal. Liquid state NMR results indicated no significant structural changes between dark and light active G90D conformations, pointing towards the pre-active dark state. These results confirm proposal of Zvyaga *et al.* ^[215] The protein conformational heterogeneity observed by liquid state NMR remains identical in the light active state, again questioning the reason for the lack of G90D crystals in the dark state suggested by Singhal *et al.* ^[209]

Furthermore, both dark state conformations include a protonated Schiff base and comprise a 11-cis retinal covalently bound to K296. Retinal conformational heterogeneity was observed only on the C14 carbon, and is referred to the slightly different steric position of 11-cis retinal in the binding pocket. These results are in agreement with the statement of Singhal *et al.*, which limits structural changes, induced by the G90D mutant, to the ligand binding pocket. This assumption allows us to conclude that the additional tryptophan signal in the liquid 2D NMR experiment, resulting from the second conformation (**Figure 3.17B**), should belong to the residue located in the close vicinity of the retinal binding pocket.

Steven Smith and coworkers investigated the impact of retinal isomerization on the rhodopsin activation process by solid state NMR experiments performed on the dark and light active wild type rhodopsin. 11-cis to all-trans retinal isomerization resulted in a large C20 methyl group rotation towards the extracellular loop E2 and the chromophore translation towards H5 ^[171]. The last motion leads to a H6 rotation, which is known to play a key role in the rhodopsin activation process. The residue W265 is located in H6 in the retinal binding pocket; its indole ring is embedded by 11-cis retinal and the K296 side chain, making it the most light-sensitive tryptophan in the rhodopsin sequence. Remarkably, in the 1990s, even before the first GPCR crystal structure was solved, W265 was found to be located close to the chromophore. Furthermore, single point mutation experiments showed W265 to be essential for retinal regeneration and transducing activation ^[226]. Solid state NMR experiments proved W265 to be involved in the rhodopsin activation switch and showed its stabilizing effect on the dark inactive

conformation. Thus, upon light-induced retinal β -ionone ring motion towards H5, the C20-W265 interaction of the dark state is destroyed and a new one with C19 methyl group of retinal is built ^[227]. Taking these results into account, it appears that the additional signal (WX) observed in liquid 2D NMR experiment at 11.65 ppm could belong to W265. However, the strong downfield shift and spectral vicinity to the W161 from the wild type requires additional experiments to fully prove this assumption. Single point mutation experiments (W265F and W161F) should provide unambiguous evidence and are the subject of further investigations.

In contrast to the G90D dark state, Meta II crystals of the stabilized G90D mutant were successfully grown and 3D X-ray structure was presented by Singhal *et al.* The low resolution (3.3 Å) does not provide atomic details but allowed the authors to conclude the overall structural similarity to the Meta II wild type conformation. The NMR experiments presented in this doctoral thesis are focused mainly on the retinal binding pocket and all conformational differences induced by the G90D mutation can be attributed to its structural rearrangements, which do not contradict the Standfuss data.

Due to its pure resolution, Standfuss and coworkers performed a crystallographic refinement of the electron density in the retinal binding pocket of the G90D Meta II state and modeled a heterogeneous mixture of non-covalently bound retinal populations, a mixture of cis-isomers, illustrated in the publication as 13-cis, 9-13-di-cis chromophores ^[209]. Solid state NMR and photo flash experiments, carried out within the framework of this thesis, clearly report a protonated G90D Meta II state. 1D ¹⁵N NMR detection demonstrates a protonated Schiff base signal of the G90D Meta II conformation, reflecting an all-trans retinal bound conformation. No free unbound cis-isomers could be detected in the light active Meta II state.

5. Conclusion:

The main goal of this part of the thesis was to characterize conformational changes of the constitutively active G90D mutant, which is responsible for the disruption of the visual cycle and results in the CSNB disease in humans. In order to fulfil the lacking structural gaps concerning the exact retinal conformation and mutation-induced changes in the binding pocket in the dark and light active state of the G90D mutant, a combination of liquid and solid state NMR spectroscopic experiments were performed. The analysis was accomplished by the comparison with the wild type data. Furthermore, kinetic studies of the light activation process of the G90D mutant were performed by flash photolysis and ultrafast absorption experiments.

Since the G90D construct includes N2C/D282C stabilizing double mutation, the impact of this modification was analyzed by comparison of the wild type with and without this additional disulfide

bridge. In total, three rhodopsin constructs were analyzed: wild type, stabilized wild type with N2C/D282C double mutation and the stabilized CSNB related G90D mutant. The stabilized wild type includes an additional disulfide bridge in the extracellular side, which aims to increase the thermal stability of rhodopsin. Opsin incorporation with 11-cis retinal was found to be impeded by the stabilized constructs, requiring an optimization of the reconstitution protocol. The stabilizing effect of the N2C/D282C mutation could be clearly observed. The dark state was not affected, while the illuminated sample remained stable and did not precipitate as the wt even after several days. For liquid state NMR five tryptophan residues were used as reporter signals. Three of them located in the transmembrane region (W265, W161 and W126) are sensitive to the light-induced conformational changes and show characteristic CSPs, equivalent to the Meta II state of the not-stabilized wild type. Two other tryptophan amino acids from the extracellular domain (W35 and W175) do not undergo any chemical shift perturbations and are resistant to the protein structural rearrangements.

Ultrafast absorption experiments, which describe the early photocycle process, detect a four times slower bathorhodopsin state formation of the stabilized wild type. Photo flash analysis also reports a stabilizing effect of N2C/D282C mutation. Thus, the decay of the intermediate states, Meta II and Meta III, is significantly delayed, the covalently bound retinal is stabilized and its hydrolysis is prolonged.

The constitutively active G90D mutation is located in the retinal binding pocket. Its negatively charged side chain perturbs the H-bond network around the Schiff base bound retinal and results in the crucial photocycle distortion.

Liquid state NMR experiments of the G90D mutant show conformational similarity of dark and light state, indicating the pre-active dark state conformation. Interestingly, conformational heterogeneity, observed for this active mutant in the dark state remains also in the Meta II conformation. Mixture of different protein populations in the dark state was confirmed by solid state NMR experiments. The structural heterogeneity includes the protein Schiff base and the retinal C14 carbon. However, the second retinal population in the dark state is not originated from the light active form and is attributed to the slight different steric position of the 11-cis retinal. Both populations are covalently bound to the Schiff base.

Upon illumination no significant conformational changes could be observed for the G90D mutant in the 2D liquid state NMR, suggesting retinal still to be incorporated in the binding pocket. Since both free retinal isomers (11-cis and all-trans) absorb at the same wavelength as all-trans bound light active Meta II state (380 nm), the conclusion whether retinal is still bound to the protein after illumination cannot be stated based on the absorption spectroscopy. Therefore, solid state NMR studies were conducted. First, it was shown that Meta II conformation of the G90D mutant exists in the protonated form. This result

was also observed by independent photo flash analysis. Secondly, photostate bathorhodopsin was successfully cryotrapped and analyzed. The minor Meta II state was identified in the batho conformation, suggesting a pre-active dark state, which was supported by the liquid state NMR. For all light intermediate states, batho and Meta II, a protonated Schiff base signal was observed.

Ultrafast kinetic measurement provided unique information about the G90D mutation modified photocycle. Thus, in contrast to the wild type, the early photorhodopsin state was detectable for the G90D mutant, indicating delayed retinal isomerization. The next transition to the bathorhodopsin occurred faster and relaxed slower, resulting in a longer lived batho state. Furthermore, intermediate Meta III formation and Meta II decay of the G90D mutant are in the same regime as for the stabilized wild type, while Meta III decay is 20 minutes faster.

To sum up, during this doctoral thesis important structural information regarding the G90D CSNB related mutant was gained. This mutation induced conformational heterogeneity in the dark state, as demonstrated by different methods, is interpreted as the pre-active dark state. The Schiff base was found to be protonated in all conformations, including the light active Meta II state. Furthermore, a unique photocycle with delayed retinal isomerization and accelerated retinal release was detected. These data support the pre-active ground state theory, used to explain the increased basal activity of the G90D mutant, excluding the spontaneous retinal isomerization, and provide an important piece of information for the detailed understanding of the molecular mechanism of night blindness disease.

References

- [1] M. Saunders, A. Wishnia, J. G. Kirkwood, *J. Am. Chem. Soc.* **1957**, *79*, 3289–3290.
- [2] E. D. Becker, **1993**, *65*, 295–302.
- [3] K. Nagayama, P. Bachmann, K. Wuthrich, R. R. Ernst, *J. Magn. Reson.* **1978**, *31*, 133–148.
- [4] W. P. Aue, E. Bartholdi, R. R. Ernst, *J. Chem. Phys.* **1976**, *64*, 2229–2246.
- [5] M. P. Williamson, T. F. Havel, K. Wüthrich, *J. Mol. Biol.* **1985**, *182*, 295–315.
- [6] R. Boelens, H. Vis, C. E. Vorgias, K. S. Wilson, R. Kaptein, *Biopolymers* **1997**, 553–559.
- [7] D. Marion, P. C. Driscoll, L. E. Kay, P. T. Wingfield, A. Bax, A. M. Gronenborn, G. M. Clore, *Biochemistry* **1989**, *28*, 6150–6156.
- [8] R. Sprangers, L. E. Kay, *Nature* **2007**, *445*, 618.
- [9] A. K. Mittermaier, L. E. Kay, *Trends Biochem. Sci.* **2009**, *34*, 601–611.
- [10] K. Kazimierczuk, J. Stanek, A. Zawadzka-kazimierczuk, **2013**, 3015–3025.
- [11] K. Grudziąż, A. Zawadzka-Kazimierczuk, W. Koźmiński, *Methods* **2018**, *148*, 81–87.
- [12] B. Brutscher, P. Schanda, *Encycl. Magn. Reson.* **2009**, 1–10.
- [13] R. R. Ernst, W. A. Anderson, *Rev. Sci. Instrum.* **1966**, *37*, 93–102.
- [14] A. Ross, M. Salzmann, H. Senn, *J. Biomol. NMR* **1997**, *10*, 389–396.
- [15] P. Schanda, B. Brutscher, *J. Am. Chem. Soc.* **2005**, *127*, 8014–8015.
- [16] P. Schanda, E. Kupče, B. Brutscher, *J. Biomol. NMR* **2005**, *33*, 199–211.
- [17] P. Schanda, V. Forge, B. Brutscher, *Proc. Natl. Acad. Sci.* **2007**, *104*, 11257–11262.
- [18] P. Schanda, H. Van Melckebeke, B. Brutscher, *J. Am. Chem. Soc.* **2006**, *128*, 9042–9043.
- [19] J. C. Barna, E. D. Laue, M. R. Mayger, J. Skilling, S. J. P. Worrall, *J. Magn. Reson.* **1987**, *73*, 69–77.
- [20] M. A. Zambrello, A. D. Schuyler, M. W. Maciejewski, F. Delaglio, I. Bezsonova, J. C. Hoch, *Methods* **2018**, *138–139*, 62–68.
- [21] D. Rovnyak, D. P. Frueh, M. Sastry, Z. J. Sun, A. S. Stern, J. C. Hoch, G. Wagner, , *J. Magn. Reson.* **2004**, *170*, 15–21.
- [22] R. Freeman, *J. Magn. Reson.* **2005**, *173*, 317–321.
- [23] R. Freeman, *J. Am. Chem. Soc.* **2003**, 13958–13959.
- [24] B. E. Coggins, P. Zhou, *J. Biomol. NMR* **2008**, 225–239.
- [25] S. G. Hyberts, K. Takeuchi, G. Wagner, **2010**, 2145–2147.
- [26] R. Saborano, Z. Eraslan, J. Roberts, F. L. Khanim, P. F. Lalor, M. A. C. Reed, U. L. Günther, *Sci. Rep.* **2019**, *9*, 1–13.
- [27] T. Von Schlippenbach, P. J. Oefner, W. Gronwald, *Sci. Rep.* **2018**, 1–10.
- [28] F. Delaglio, *EMBO summer course* **2017**.
- [29] M. Bostock, D. Nietlispach, *Concepts Magn. Reson. Part A.* **2018**, DOI 10.1002/cmr.a. 21438
- [30] H. Nyquist, *Proc. IEEE* **2002**, *90*, 280–305.
- [31] K. Kazimierczuk, V. Y. Orekhov, *Angew. Chem. Int. Ed* **2011**, 5556–5559.
- [32] D. L. Donoho, *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306.
- [33] D. J. Holland, M. J. Bostock, L. F. Gladden, D. Nietlispach, *Angew. Chem. Int. Ed.* **2011**, *50*, 6548–6551
- [34] V. Y. Orekhov, I. Ibraghimov, M. Billeter, *J. Biomol. NMR* **2001**, *20*, 49–60.
- [35] V. Y. Orekhov, I. Ibraghimov, M. Billeter, *J. Biomol. NMR* **2003**, *27*, 165–173.

References

- [36] V. Tugarinov, L. E. Kay, I. Ibraghimov, V. Y. Orekhov, *J. Am. Chem. Soc.* **2005**, *127*, 2767–2775.
- [37] A. Shchukina, P. Kasprzak, R. Dass, M. Nowakowski, *J. Biomol. NMR* **2017**, *68*, 79–98.
- [38] A. S. Stern, D. L. Donoho, J. C. Hoch, *J. Magn. Reson.* **2007**, *188*, 295–300.
- [39] V. Y. Orekhov, V. Jaravine, *Prog. Nucl. Magn. Reson. Spectrosc.* **2011**, *59*, 271–292.
- [40] V. Jaravine, A. V. Zhuravleva, P. Permi, I. Ibraghimov, V. Y. Orekhov, *J. Am. Chem. Soc.* **2008**, *130*, 3927–3936.
- [41] L. Isaksson, M. Mayzel, M. Saline, A. Pedersen, J. Rosenl ow, B. Brutscher, B. G. Karlsson, V. Y. Orekhov, *PLoS One* **2013**, *8*, 1–8.
- [42] V. A. Jaravine, V. Y. Orekhov, *J. Am. Chem. Soc.* **2006**, *128*, 13421–13426.
- [43] E. Schmidt, P. G ntert, *J. Am. Chem. Soc.* **2012**, *134*, 12817–12829.
- [44] A. Varmazyari, J. Khalili-Mahaleh, M. Roshdi, K. Kharazmi, *Adv. Environ. Biol.* **2012**, *6*, 587–592.
- [45] P. G ntert, *Prog. Nucl. Magn. Reson. Spectrosc.* **2003**, *43*, 105–125.
- [46] G. Storz, Y. I. Wolf, K. S. Ramamurthi, *Annu. Rev. Biochem.* **2014**, *83*, 753–77.
- [47] M. Su, Y. Ling, J. Yu, J. Wu, J. Xiao, *Front. Genet.* **2013**, *4*, 1–9.
- [48] X. Yang, T. J. Tschaplinski, G. B. Hurst, S. Jawdy, P. E. Abraham, P. K. Lankford, R. M. Adams, M. B. Shah, R. L. Hettich, E. Lindquist, et al., *Genome Res.* **2011**, *21*, 634–641.
- [49] P. Zuber, *Peptides* **2001**, *22*, 1555–1577.
- [50] Nagel, Machulla, Zahn, Soppa, *Genes* **2019**, *10*, 361.
- [51] M. C. Frith, A. R. Forrest, E. Nourbakhsh, K. C. Pang, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, T. L. Bailey, S. M. Grimmond, *PLoS Genet.* **2006**, *2*, 515–528.
- [52] S. Kent, *Curr. Opin. Biotechnol.* **2004**, *15*, 607–614.
- [53] E. G. Baker, G. J. Bartlett, K. L. Porter Goff, D. N. Woolfson, *Acc. Chem. Res.* **2017**, *50*, 2085–2092.
- [54] C. a. Makarewich, E. N. Olson, *Trends Cell Biol.* **2017**, *27*, 685–696.
- [55] D. Baumgartner, M. Kopf, S. Kl hn, C. Steglich, W. R. Hess, *BMC Microbiol.* **2016**, *16*, 1–15.
- [56] P. M. Harrison, A. Kumar, N. Lang, M. Snyder, M. Gerstein, *Analysis* **2002**, *30*, 1083–1090.
- [57] M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, *PLoS Biol.* **2007**, *5*, 1052–1062.
- [58] T. Kondo, Y. Hashimoto, K. Kato, S. Inagaki, S. Hayashi, Y. Kageyama, *Nat. Cell Biol.* **2007**, *9*, 660–665.
- [59] S. D. Mackowiak, H. Zauber, C. Bielow, D. Thiel, K. Kutz, L. Calviello, G. Mastrobuoni, N. Rajewsky, S. Kempa, M. Selbach, *Genome Biol.* **2015**, *16*, 1–21.
- [60] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. Weissman, *Science* **2009**, *324*, 218–324.
- [61] N. T. Ingolia, L. F. Lareau, J. S. Weissman, *Cell* **2011**, *147*, 789–802.
- [62] P. Y. Hsu, L. Calviello, H.-Y. L. Wu, F.-W. Li, C. J. Rothfels, U. Ohler, P. N. Benfey, *Proc. Natl. Acad. Sci.* **2016**, *113*, E7126–E7135.
- [63] J. E. Smith, J. R. Alvarez-Dominguez, N. Kline, N. J. Huynh, S. Geisler, W. Hu, J. Collier, K. E. Baker, *Cell Rep.* **2014**, *7*, 1858–1866.
- [64] T. Fujita, Y. Kurihara, S. Iwasaki, *Plant Cell Physiol.* **2019**, *0*, 1–10.
- [65] J. G. Dunn, C. K. Foo, N. G. Belletier, E. R. Gavis, J. S. Weissman, *Elife* **2013**, *2*, 1–32.
- [66] A. a. Bazzini, T. G. Johnstone, R. Christiano, S. D. MacKowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, *EMBO J.* **2014**, *33*, 981–993.
- [67] S. Lee, B. Liu, S. Lee, S.-X. Huang, B. Shen, S.-B. Qian, *Proc. Natl. Acad. Sci.* **2012**, *109*, E2424–E2432.
- [68] J. Weaver, F. Mohammad, A. R. Buskirk, G. Storz, *MBio* **2019**, *10*, DOI 10.1128/mBio.02819-18.
- [69] S. Slavoff, A. J. Mitchell, A. G. Schwaid, M. N. Cabili, J. Ma, J. Z. Levin, A. D. Karger, B. Budnik, J. L. Rinn, A. Saghatelian, *Nat. Chem. Biol.* **2013**, *9*, 59–64.

- [70] J. I. Pueyo, E. G. Magny, J. P. Couso, *Trends Biochem. Sci.* **2016**, *41*, 665–678.
- [71] B. E. Housden, M. Muhar, M. Gemberling, C. a. Gersbach, D. Y. R. Stainier, G. Seydoux, S. E. Mohr, J. Zuber, N. Perrimon, *Nat. Rev. Genet.* **2016**, *18*, 24–40.
- [72] A. Matsumoto, A. Pasut, M. Matsumoto, R. Yamashita, J. Fung, E. Monteleone, A. Saghatelian, K. I. Nakayama, J. G. Clohessy, P. P. Pandolfi, *Nature* **2017**, *541*, 228–232.
- [73] G. Storz, J. Vogel, K. M. Wassarman, *Mol. Cell* **2011**, *43*, 880–891.
- [74] S. Slavoff, J. Heo, B. Budnik, L. Hanakahi, A. Saghatelian, *J. Biol. Chem.* **2014**, *289*, 10950–10957.
- [75] E. G. Magny, E. G. Magny, J. I. Pueyo, F. M. G. Pearl, M. A. Cespedes, J. E. Niven, S. Bishop, J. P. Couso, *Science* **2013**, *1116*, 921–926.
- [76] C. Lee, J. Zeng, B. G. Drew, T. Sallam, A. Martin-Montalvo, J. Wan, S. J. Kim, H. Mehta, A. L. Hevener, R. De Cabo, et al., *Cell Metab.* **2015**, *21*, 443–454.
- [77] B. Guo, D. Zhai, E. Cabezas, K. Welsh, S. Nouraini, A. C. Satterthwait, J. C. Reed, *Nature* **2003**, *423*, 456–461.
- [78] P. Mendoza-Espinosa, V. García-González, A. Moreno, R. Castillo, J. Mas-Oliva, *Mol. Cell. Biochem.* **2009**, *330*, 105–120.
- [79] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, **2001**, *3263*, 26–59.
- [80] P. Tompa, *Trends Biochem. Sci.* **2002**, *27*, 527–533.
- [81] H. J. Dyson, P. E. Wright, *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.
- [82] E. Kidding, C., Parge, H., Knighton, D., Lewis, C., Pelletier, L., Tempczyk, A., Kallish, V., Tucker, K., Showalter, R., Moomaw, E., Gastinel, L., Habuka, N., Chen, X., Maldonado, F., Barker, J., Bacquet, R., Villafranca, *Lett. to Nat.* **1995**, *378*, 641–644.
- [83] A. K. Dunker, Z. Obradovic, *Nat. Biotechnol.* **2001**, *19*, 805–806.
- [84] P. G. Wolynes, J. N. Onuchic, D. Thirumalai, *Science* **1995**, *267*, 1619–1620.
- [85] M. Miskei, A. Gregus, R. Sharma, N. Duro, F. Zsolyomi, M. Fuxreiter, *FEBS Lett.* **2017**, *591*, 2682–2695.
- [86] L. Joedicke, J. Mao, G. Kuenze, C. Reinhart, T. Kalavacherla, H. R. a. Jonker, C. Richter, H. Schwalbe, J. Meiler, J. Preu, et al., *Nat. Chem. Biol.* **2018**, *14*, 284–290.
- [87] A. I. Herrera, J. M. T. and O. Prakash, *Curr. Protein Pept. Sci.* **2016**, *17*, 827–841.
- [88] B. Ekblad, P. E. Kristiansen, *Sci. Rep.* **2019**, *9*, 1–10.
- [89] M. R. Hemm, B. J. Paul, T. D. Schneider, G. Storz, K. E. Rudd, *Mol. Microbiol.* **2008**, *70*, 1487–1501.
- [90] A. Heitz, D. Le-Nguyen, L. Chiche, *Biochemistry* **1999**, *38*, 10615–10625.
- [91] S. Ohki, M. Takeuchi, M. Mori, *Nat. Commun.* **2011**, *2*, 512–517.
- [92] S. S. Krishna, I. Majumdar, N. V. Grishin, *Nucleic Acids Res.* **2003**, *31*, 532–550.
- [93] J. H. Laity, B. M. Lee, P. E. Wright, *Curr. Opin. Struct. Biol.* **2001**, *11*, 39–46.
- [94] M. Bédard, L. Maltais, M. E. Beaulieu, J. Bilodeau, D. Bernard, P. Lavigne, *J. Biomol. NMR* **2012**, *54*, 317–323.
- [95] J. W. Neidigh, R. M. Fesinmeyer, N. H. Andersen, *Nat. Struct. Biol.* **2002**, *9*, 425–430.
- [96] H. Schwalbe, K. M. Fiebig, M. Buck, J. a. Jones, S. B. Grimshaw, A. Spencer, S. J. Glaser, L. J. Smith, C. M. Dobson, *Biochemistry* **1997**, *36*, 8977–8991.
- [97] K. Wüthrich, *Angew. Chemie - Int. Ed.* **2003**, *42*, 3340–3363.
- [98] L. E. Kay, *J. Magn. Reson.* **2005**, *173*, 193–207.
- [99] I. Walsh, A. J. M. Martin, T. Di domenico, S. C. E. Tosatto, *Bioinformatics* **2012**, *28*, 503–509.
- [100] M. P. Malakhov, M. R. Mattern, O. a. Malakhova, M. Drinker, S. D. Weeks, T. R. Butt, *J. Struct. Funct. Genomics* **2004**, *5*, 75–86.
- [101] J. G. Marblestone, *Protein Sci.* **2006**, *15*, 182–189.

References

- [102] X. Zuo, M. R. Mattern, R. Tan, S. Li, J. Hall, D. E. Sterner, J. Shoo, H. Tran, P. Lim, S. G. Sarafianos, et al., *Protein Expr. Purif.* **2005**, *42*, 100–110.
- [103] R. J. Sundberg, *Chem. Rev.* **1974**, *74*, 471–517.
- [104] S. Harper, D. W. Speicher, *Curr. Protoc. Protein Sci.* **2008**, *2008*, 1–26
- [105] Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch, *The Proteomics Protocols Handbook*, Humana Press **2005**, 571–608.
- [106] J. Reed, T. A. Reed, *Anal. Biochem.* **1997**, *254*, 36–40.
- [107] P. B. Barker, X. Golay, D. Artemov, R. Ouwerkerk, M. A. Smith, A. J. Shaka, *Magn. Reson. Med.* **2001**, *45*, 226–232
- [108] D. S. Wishart, *Prog. Nucl. Magn. Reson. Spectrosc.* **2011**, *58*, 62–87.
- [109] T. D. Goddard, D. G. Kneller, SPARKY 3, University of California, san Francisco.
- [110] N. J. Baxter, M. P. Williamson, *J. Biomol. NMR* **1997**, *9*, 359–369.
- [111] P. Dosset, J. C. Hus, M. Blackledge, D. Marion, *J. Biomol. NMR* **2000**, *16*, 23–28.
- [112] N. Farrow, R. Muhandiram, S. M. Pascal, L. E. Kay, A. U. Singer, J. D. Forman-Kay, C. M. Kay, G. Gish, T. Pawson, S. E. Shoelson, *Biochemistry* **1994**, *33*, 5984–6003.
- [113] J. P. Linge, M. Habeck, W. Rieping, M. Nilges, *Bioinformatics* **2003**, *19*, 315–316.
- [114] Y. Shen, A. Bax, *J. Biomol. NMR* **2013**, *56*, 227–241.
- [115] A. T. Brünger, P. D. Adams, G. M. Clore, W. L. Delano, P. Gros, R. W. Grossekunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1998**, *54*, 905–921.
- [116] J. P. Linge, M. a. Williams, C. a. E. M. Spronk, A. M. J. J. Bonvin, M. Nilges, *Proteins Struct. Funct. Genet.* **2003**, *50*, 496–506.
- [117] J. Hahn, S. Thalmann, A. Migur, R. Freiherr von Boeselager, N. Kubatova, E. Kubareva, H. Schwalbe, E. Evguenieva-Hackenberg, *RNA Biol.* **2016**, *14*, 1353–1363
- [118] J. Grote, D. Krysiak, K. Petersen, S. Güllert, C. Schmeisser, K. U. Förstner, H. B. Krishnan, H. Schwalbe, N. Kubatova, W. R. Streit, *Front. Microbiol.* **2016**, *7*, 1858.
- [119] Y. Shen, J. Maupetit, P. Derreumaux, P. Tufféry, *J. Chem. Theory Comput.* **2014**, *10*, 4745–4758.
- [120] P. Thévenet, Y. Shen, J. Maupetit, F. Guyon, P. Derreumaux, P. Tufféry, *Nucleic Acids Res.* **2012**, *40*, 288–293.
- [121] S. Singh, H. Singh, A. Tuknait, K. Chaudhary, B. Singh, S. Kumaran, G. P. S. Raghava, *Biol. Direct* **2015**, *10*, 1–19.
- [122] Harpreet Kaur, Aarti Garg, G.P.S. Raghava, *Protein Pept. Lett.* **2007**, *14*, 626–631.
- [123] Y. Zhang, *BMC Bioinformatics* **2008**, *9*, 1–8.
- [124] C. a. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, *Methods Enzymol.* **2004**, *383*, 66–93.
- [125] N. Guex, M. C. Peitsch, T. Schwede, *Electrophoresis* **2009**, *30*, 162–173.
- [126] P. Benkert, M. Biasini, T. Schwede, *Bioinformatics* **2011**, *27*, 343–350.
- [127] M. Bertoni, F. Kiefer, M. Biasini, L. Bordoli, T. Schwede, *Sci. Rep.* **2017**, *7*, 1–15.
- [128] S. Bienert, A. Waterhouse, T. a. P. De Beer, G. Tauriello, G. Studer, L. Bordoli, T. Schwede, *Nucleic Acids Res.* **2017**, *45*, D313–D319.
- [129] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. a. P. De Beer, C. Rempfer, L. Bordoli, et al., *Nucleic Acids Res.* **2018**, *46*, W296–W303.
- [130] C. Camproux, R. Gautier, P. Tufféry, *J. Mol. Biol.* **2004**, *339*, 591–605.
- [131] A. Lamiable, P. Thévenet, J. Rey, M. Vavrusa, P. Derreumaux, P. Tufféry, *Nucleic Acids Res.* **2016**, *44*, W449–W454.
- [132] N. Kubatova, H. R. A. Jonker, K. Saxena, C. Richter, V. Vogel, S. Schreiber, H. Schwalbe, *ChemBioChem.* **2019**, DOI 10.1002/cbic.201900085.
- [133] P. Kukic, G. T. Heller, a K. Dunker, M. Bonomi, M. Fuxreiter, R. V Pappu, P. Sormanni, Z.

- Dosztanyi, M. M. Babu, V. N. Uversky, et al., *Nat. Chem. Biol.* **2017**, *13*, 339–342.
- [134] T. Ishida, K. Kinoshita, *Bioinformatics* **2008**, *24*, 1344–1348.
- [135] D. Jager, C. M. Sharma, J. Thomsen, C. Ehlers, J. Vogel, R. a. Schmitz, *Proc. Natl. Acad. Sci.* **2009**, *106*, 21878–21882.
- [136] D. Dar, D. Prasse, R. a. Schmitz, R. Sorek, *Nat. Microbiol.* **2016**, *1*, 1–9.
- [137] D. Prasse, J. Thomsen, R. De Santis, J. Muntel, D. Becher, R. a. Schmitz, *Biochimie* **2015**, *117*, 138–148.
- [138] D. Krysiak, J. Grote, M. R. Orbegoso, C. Utpatel, K. U. Förstner, L. Li, C. Schmeisser, H. B. Krishnan, W. R. Streit, *Appl. Environ. Microbiol.* **2014**, *80*, 5655–5671.
- [139] G. W. Vuister, A. Bax, *J. Am. Chem. Soc.* **1993**, *115*, 7772–7777.
- [140] J. García de la Torre, M. . Huertas, B. Carrasco, *J. Magn. Reson.* **2000**, *147*, 138–146.
- [141] E. Krissinel, K. Henrick, *J. Mol. Biol.* **2007**, *372*, 774–797.
- [142] A. Kumar, A. M. Balakrishna, W. Nartey, M. S. S. Manimekalai, G. Grüber, *Free Radic. Biol. Med.* **2016**, *97*, 588–601.
- [143] K. D’Ambrosio, D. Limauro, E. Pedone, I. Galdi, C. Pedone, S. Bartolucci, G. De Simone, *Proteins Struct. Funct. Bioinforma.* **2009**, *76*, 995–1006.
- [144] K. Lim, A. Tempczyk, N. Bonander, J. Toedt, A. Howard, E. Eisenstein, O. Herzberg, *J. Biol. Chem.* **2003**, *278*, 13496–13502.
- [145] M. L. Ferri-Fioni, E. Schmitt, J. Soutourina, P. Plateau, Y. Mechulam, S. Blanquet, *J. Biol. Chem.* **2001**, *276*, 47285–47290.
- [146] T. K. Bhatt, M. Yogavel, S. Wydau, R. Berwal, A. Sharma, *J. Biol. Chem.* **2010**, *285*, 5917–5930.
- [147] S. Ahmad, S. B. Routh, V. Kamarthapu, J. Chalissery, S. Muthukumar, T. Hussain, S. P. Kruparani, M. V. Deshmukh, R. Sankaranarayanan, *Elife* **2013**, *2013*, 1–18.
- [148] S. B. Routh, K. I. Pawar, S. Ahmad, S. Singh, K. Suma, M. Kumar, S. K. Kuncha, K. Yadav, S. P. Kruparani, R. Sankaranarayanan, *PLoS Biol.* **2016**, *14*, 1–22.
- [149] H. Der Ou, F. Löhr, V. Vogel, W. Mäntele, V. Dötsch, *EMBO J.* **2007**, *26*, 3463–3473.
- [150] K. Roy, L. Stein, S. Kaushal, *Hum. Gene Ther.* **2010**, *21*, 915–927.
- [151] N. T. Ingram, A. P. Sampath, G. L. Fain, *J. Physiol.* **2016**, *594*, 5415–5426.
- [152] N. T. Slater, J. Swann, **1984**, 9–11.
- [153] W. W. S. Yue, D. Silverman, X. Ren, R. Frederiksen, K. Sakai, T. Yamashita, Y. Shichida, M. C. Cornwall, J. Chen, K.-W. Yau, *Proc. Natl. Acad. Sci. U. S. A.* **2019**, DOI 10.1073/pnas.1817781116.
- [154] D. Raman, S. Osawa, E. R. Weiss, *Biochemistry* **1999**, *38*, 5117–5123.
- [155] V. Y. Arshavsky, T. G. Wensel, *Investig. Ophthalmol. Vis. Sci.* **2013**, *54*, 7725–7733.
- [156] X. Deupi, J. Standfuss, *Curr. Opin. Struct. Biol.* **2011**, *21*, 541–551.
- [157] M. E. Burns, V. Y. Arshavsky, *Neuron* **2005**, *48*, 387–401.
- [158] M. E. Burns, E. N. Pugh, *Physiology* **2010**, *25*, 72–84.
- [159] a. H. Bunt-Milam, *J. Cell Biol.* **2004**, *97*, 703–712.
- [160] Q. Wu, L. R. Blakeley, M. C. Cornwall, R. K. Crouch, N. Barbara, **2008**, *46*, 8669–8679.
- [161] X. Liu, P. Garriga, H. G. Khorana, *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 4554–4559.
- [162] K. Palczewski, F. Haeseleer, J. Huang, J. C. Saari, L. Lebioda, *J. Biol. Chem.* **2002**, *273*, 21790–21799.
- [163] D. Chatterjee, Doctoral Thesis **2016**.
- [164] A. Maeda, K. Palczewski, *Drug Discov. Today Dis. Model.* **2013**, *10*, e163–e172.
- [165] J. C. Saari, D. L. Bredberg, *J. Biol. Chem.* **1989**, *264*, 8636–8640.
- [166] G. Moiseyev, R. K. Crouch, P. Goletz, J. Oatis, T. M. Redmond, J. X. Ma, *Biochemistry* **2003**, *42*, 2229–2238.

References

- [167] T. M. Redmond, S. Yu, E. Lee, D. Bok, D. Hamasaki, N. Chen, P. Goletz, J. X. Ma, R. K. Crouch, K. Pfeifer, *Nat. Genet.* **1998**, *20*, 344–351.
- [168] H. J. Winkens, H. M. Van Vugt, L. M. De Leeuw, F. Deutman, C. G. G. Driessen, J. J. M. Janssen, P. M. Janssen, *Vision Res.* **2003**, *35*, S65.
- [169] M. C. Cornwall, E. Tsina, P. Ala-Laurila, R. K. Crouch, V. I. Govardovskii, Y. Koutalos, B. Wiggert, M. E. Estevez, S. a. Shukolyukov, A. V. Kolesnikov, *J. Gen. Physiol.* **2006**, *128*, 153–169.
- [170] K. Palczewski, T. Kumasaka, T. Hori, C. a Behnke, H. Motoshima, B. a Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, *Sci. (New York, NY)* **2000**, *289*, 739–745.
- [171] B. Patel, E. Crocker, M. Eilers, Hirshfeld, M. Sheves, S. O. Smith, *Proc. Natl. Acad. Sci.* **2004**, *101*, 10048–10053.
- [172] J.-M. Kim, C. Altenbach, M. Kono, D. D. Oprian, W. L. Hubbell, H. G. Khorana, *Proc Natl Acad Sci U S A* **2004**, *101*, 12508–12513.
- [173] H. W. Choe, Y. J. Kim, J. H. Park, T. Morizumi, E. F. Pai, N. Krauss, K. P. Hofmann, P. Scheerer, O. P. Ernst, *Nature* **2011**, *471*, 651–655.
- [174] O. P. Ernst, K. P. Hofmann, H.-W. Choe, J. H. Park, P. Scheerer, *Nature* **2008**, *454*, 183–187.
- [175] R. W. Schoenlein, L. a. Peteanu, R. a. Mathies, C. V. Shank, *Science* **1991**, *254*, 412–415.
- [176] O. P. Ernst, F. J. Bartl, *ChemBioChem* **2002**, *3*, 968–974.
- [177] J. Stehle, R. Silvers, K. Werner, D. Chatterjee, S. Gande, F. Scholz, A. Dutta, J. Wachtveitl, J. Klein-Seetharaman, H. Schwalbe, *Angew. Chemie - Int. Ed.* **2014**, *53*, 2078–2084.
- [178] A. Mendez, J. Lem, M. I. Simon, L.-W. Wu, D. a. Baylor, M. E. Burns, J. Chen, A. Roca, *Neuron* **2004**, *28*, 153–164.
- [179] S. H. Tsang, T. Sharma, *Adv. Exp. Med. Biol.* **2018**, *1085*, 125–130.
- [180] E. Waltz, *Nat. Biotechnol.* **2018**, *36*, 6–7.
- [181] H. P. N. Scholl, A. T. Moore, R. K. Koenekoop, Y. Wen, G. a. Fishman, L. I. Van Den Born, A. Bittner, K. Bowles, E. C. Fletcher, F. T. Collison, *PLoS One* **2015**, *10*, 1–16.
- [182] F. Badalà, K. Nouri-mahdavi, D. a Raoof, *Computer* **2008**, *144*, 724–732.
- [183] M. J. Seiler, R. B. Aramant, *Prog. Retin. Eye Res.* **2012**, *31*, 661–687.
- [184] L. da Cruz, J. D. Dorn, M. S. Humayun, G. Dagnelie, J. Handa, P. O. Barale, J. A. Sahel, P. E. Stanga, F. Hafezi, A. B. Safran, *Ophthalmology* **2016**, *123*, 2248–2254.
- [185] K. Stingl, R. Schippert, K. U. Bartz-Schmidt, D. Besch, C. L. Cottrill, T. L. Edwards, F. Gekeler, U. Greppmaier, K. Kiel, A. Koitschev, *Front. Neurosci.* **2017**, *11*, 445.
- [186] C. Zeitz, A. G. Robson, I. Audo, *Prog. Retin. Eye Res.* **2015**, *45*, 58–110.
- [187] K. M. Boycott, W. G. Pearce, N. T. Bech-Hansen, *Can. J. Ophthalmol.* **2000**, *35*, 204–213.
- [188] M. M. C. Bijveld, M. M. van Genderen, F. P. Hoeben, A. a. Katzin, R. M. a. van Nispen, F. C. C. Riemsdag, A. M. L. Kappers, *PLoS One* **2013**, *8*, E62927
- [189] M. M. C. Bijveld, R. J. Florijn, A. a. B. Bergen, L. I. Van Den Born, M. Kamermans, L. Prick, F. C. C. Riemsdag, M. J. Van Schooneveld, A. M. L. Kappers, M. M. Van Genderen, *Ophthalmology* **2013**, *120*, 2072–2081.
- [190] T. P. Dryja, *Am. J. Ophthalmol.* **2000**, *130*, 547–563.
- [191] S. Yamamoto, K. C. Sippel, E. L. Berson, T. P. Dryja, *Nat. Genet.* **1997**, *15*, 175–178.
- [192] S. Fuchs, M. Nakazawa, M. Maw, M. Tamai, Y. Oguchi, A. Gal, **1995**, *10*, 360–362.
- [193] V. R. Rao, G. B. Cohen, D. D. Oprian, *Nature* **1994**, *367*, 639–42.
- [194] N. Al-Jandal, G. J. Farrar, A. S. Kiang, M. M. Humphries, N. Bannon, J. B. C. Findlay, P. Humphries, P. F. Kenna, *Hum. Mutat.* **1999**, *13*, 75–81.
- [195] T. P. Dryja, E. L. Berson, D. D. Oprian, **1993**, *4*, 290–293.
- [196] C. Zeitz, A. K. Gross, D. Leifert, B. Kloeckener-Gruissem, S. D. McAlear, J. Lemke, J. Neidhardt, W. Berger, *Investig. Ophthalmol. Vis. Sci.* **2008**, *49*, 4105–4114.
- [197] P. L. Yeagle, G. Choi, a. D. Albert, *Biochemistry* **2001**, *40*, 11932–11937.

- [198] D. C. Teller, T. Okada, C. A. Behnke, K. Palczewski, R. E. Stenkamp, *Biochemistry* **2001**, *40*, 7761–7772.
- [199] T. Okada, J. Navarro, Y. Fujiyoshi, M. Silow, E. M. Landau, Y. Shichida, *Proc. Natl. Acad. Sci.* **2002**, *99*, 5982–5987.
- [200] T. Okada, M. Sugihara, A. N. Bondar, M. Elstner, P. Entel, V. Buss, *J. Mol. Biol.* **2004**, *342*, 571–583.
- [201] G. Choi, J. Landin, J. F. Galan, R. R. Birge, A. D. Albert, P. L. Yeagle, *Biochemistry* **2002**, *41*, 7318–7324.
- [202] H. Nakamichi, T. Okada, *Proc. Natl. Acad. Sci.* **2006**, *103*, 12729–12734.
- [203] H. Nakamichi, T. Okada, *Angew. Chemie - Int. Ed.* **2006**, *45*, 4270–4273.
- [204] D. Salom, D. T. Lodowski, R. E. Stenkamp, I. Le Trong, M. Golczak, B. Jastrzebska, T. Harris, J. a Ballesteros, K. Palczewski, *Proc. Natl. Acad. Sci.* **2006**, *103*, 16123–16128.
- [205] J. Li, P. C. Edwards, M. Burghammer, C. Villa, G. F. X. Schertler, *J. Mol. Biol.* **2004**, *343*, 1409–1438.
- [206] J. Standfuss, G. Xie, P. C. Edwards, M. Burghammer, D. D. Oprian, G. F. X. Schertler, *J. Mol. Biol.* **2007**, *372*, 1179–1188.
- [207] P. C. Edwards, A. D. Antona, M. Fransen, G. Xie, D. D. Oprian, G. F. X. Schertler, **2011**, *45*, 3–8.
- [208] X. Deupi, P. Edwards, a. Singhal, B. Nickle, D. Oprian, G. Schertler, J. Standfuss, *Proc. Natl. Acad. Sci.* **2012**, *109*, 119–124.
- [209] A. Singhal, M. K. Ostermaier, S. a Vishnivetskiy, V. Panneels, K. T. Homan, J. J. G. Tesmer, D. Veprintsev, X. Deupi, V. V Gurevich, G. F. X. Schertler, *EMBO Rep.* **2013**, *14*, 520–6.
- [210] A. Singhal, Y. Guo, M. Matkovic, G. Schertler, X. Deupi, E. C. Y. Yan, **2016**, *17*, 1–10.
- [211] K. Katayama, B. Jastrzebska, S. Filipek, P. D. Kiser, S. Gao, M. Mogi, Á. L. Placeres, K. Gunderson, G. P. Tochtrop, P. Miszta, et al., *Proc. Natl. Acad. Sci.* **2017**, *114*, E2608–E2615.
- [212] G. Xie, A. K. Gross, D. D. Oprian, *Biochemistry* **2003**, *42*, 1995–2001.
- [213] P. R. Robinson, G. B. Cohen, E. a. Zhukovsky, D. D. Oprian, *Neuron* **1992**, *9*, 719–725.
- [214] A. K. Gross, V. R. Rao, D. D. Oprian, *Biochemistry* **2003**, *42*, 2009–2015.
- [215] T. a. Zvyaga, K. Fahmy, F. Siebert, T. P. Sakmar, *Biochemistry* **1996**, *35*, 7536–7545.
- [216] P. J. Reeves, J. Kim, H. G. Khorana, *Proc. Natl. Acad. Sci.* **2002**, *99*, 13413–13418
- [217] A. Singhal, M. K. Ostermaier, S. A. Vishnivetskiy, V. Panneels, T. Kristoff, J. J. G. Tesmer, D. Veprintsev, X. Deupi, V. V Gurevich, F. X. Gebhard, **2013**, 1–8.
- [218] M. Eilers, P. J. Reeves, W. Ying, H. G. Khorana, S. O. Smith, *Proc. Natl. Acad. Sci.* **1999**, *96*, 487–492.
- [219] Werner, I. Lehner, H. Kaur, C. Richter, C. Glaubitz, H. Schwalbe, H. G. Khorana, J. Klein-Seetharaman, *J. Biomol. NMR* **2007**, *37*, 303–312.
- [220] C. E. Eckert, Doctoral Thesis, **2017**.
- [221] D. Chatterjee, C. E. Eckert, C. Slavov, K. Saxena, B. Fürtig, C. R. Sanders, V. V Gurevich, J. Wachtveitl, H. Schwalbe, *Angew. Chem. Int. Ed. Engl.* **2015**, *37232*, 13555–13560.
- [222] K. Fahmy, T. A. Zvyaga, T. P. Sakmar, F. Siebert, *Biochemistry* **1996**, *35*, 15065–15073.
- [223] A. F. L. Creemers, C. H. W. Klaassen, P. H. M. Bovee-Geurts, R. Kelle, U. Kragl, J. Raap, W. J. De Grip, J. Lugtenburg, H. J. M. De Groot, *Biochemistry* **1999**, *38*, 7195–7199.
- [224] M. Concistre, A. Gansmu, N. Mclean, O. G. Johannessen, I. Mari, P. H. M. Bovee-geurts, P. Verdegem, J. Lugtenburg, R. C. D. Brown, W. J. Degrip, *Communication* **2008**, 13–15.
- [225] M. Concistrè, A. Gansmüller, N. McLean, O. G. Johannessen, I. M. Montesinos, P. H. M. Bovee-Geurts, R. C. D. Brown, W. J. DeGrip, M. H. Levitt, *J. Am. Chem. Soc.* **2009**, *131*, 6133–6140.
- [226] T. A. Nakayamas, H. G. Khorana, **1991**, *266*, 4269–4275.
- [227] E. Crocker, M. Eilers, S. Ahuja, V. Hornak, A. Hirshfeld, M. Sheves, S. O. Smith, *J. Mol. Biol.* **2006**, *357*, 163–172.

Appendix

Appendix A1: Plasmids

pE-SUMO

1	AGATCTCGAT	CCC CGCAAAT	TAATACGACT	CACTATAGGG	GAATTGTGAG	CGGATAACAA
61	TTCCCCCTCA	GAAATAATTT	TGTTTAAC TT	TAAGAAGGAG	ATATACCATG	GGTCATCACC
121	ATCATCATCA	CGGGTCCCTG	CAGGACTCAG	AAGTCAATCA	AGAAGCTAAG	CCAGAGGTCA
181	AGCCAGAAGT	CAAGCCTGAG	ACTCACATCA	ATTTAAAGGT	GTCCGATGGA	GTCTCAGAGA
241	TCTTCTTCAA	GATCAAAAAG	ACCCACTCCTT	TAAGAAGGCT	GATGGAAGCG	TTCGCTAAAA
301	GACAGGGTAA	GGAAATGGAC	TCCTTAAGAT	TCTTGTACGA	CGGTATTAGA	ATTCAAGCTG
361	ATCAGGCCCC	TGAAAGATTG	GACATGGAGG	ATAACGATAT	TATTGAGGCT	CACCCGGAAC
421	AGATTGGAGG	TTGAGACCAC	TAGTGGTACC	GGTCTACTA	GAGGATCCGA	ATTTCGAGCTC
481	CGTCGACAAG	CTTGC GGCCG	CACTCGAGCA	CCACCACCAC	CACCACTGAG	ATCCGGCTGC
541	TAACAAAGCC	CGAAAGGAAG	CTGAGTTGGC	TGCTGCCACC	GCTGAGCAAT	AACTAGCATA
601	ACCCCTTGGG	GCCTCTAAC	GGGTCTT GAG	GGGTTTTTTG	CTGAAAGGAG	GAACTATATC
661	CGGATTGGCG	AATGGGACGC	GCCCTGTAGC	GGCGCATTAA	GC CGCGGG	TGTGGTGGTT
721	ACGCGCAGCG	TGACCGCTAC	ACTTGCCAGC	GCCCTAGCGC	CCGCTCCTTT	CGCTTTCTTC
781	CCTTCCCTTC	TCGCCACGTT	CGCCGGCTTT	CCCCGTCAAG	CTCTAAATCG	GGGGCTCCCT
841	TTAGGGTTCC	GATTTAGTGC	TTTACGGCAC	CTCGACCCCA	AAAAACTTGA	TTAGGGTGAT
901	GGTTCACGTA	GTGGGCCATC	GCCCTGATAG	ACGGTTTTTC	GCCCTTTGAC	GTGGAGTCC
961	ACGTTCTTTA	ATAGTGGACT	CTTGTTCCAA	ACTGGAACAA	CACTCAACCC	TATCTCGGTC
1021	TATTCTTTTG	ATTTATAAGG	GATTTTGCCG	ATTTTCGGCT	ATTGGTTAAA	AAATGAGCTG
1081	ATTTAACAAA	AATTTAACGC	GAATTTTAAAC	AAAAATTTAA	CGTTTTACAAT	TTCAGGTGGC
1141	ACTTTTCGGG	GAAATGTGCG	CGGAACCCCT	ATTTGTTTAT	TTTTCTAAAT	ACATTCAAAT
1201	ATGATCCCGC	TCATGAGACA	ATAACCCCTGA	TAAATGCTTC	AATAATATTG	AAAAAGGAAG
1261	AGTATGAGTA	TTCAACATTT	CCGTGTCGCC	CTTATTCCCT	TTTTTGCGGC	ATTTTGCCTT
1321	CCTGTTTTTG	CTCACCCAGA	AACGCTGGTG	AAAGTAAAAG	ATGCTGAAGA	TAGCTTGGGT
1381	GCACGAGTGG	GTTACATCGA	ACTGGATCTC	AACAGCGGTA	AGATCCTTGA	GAGTTTTTCGC
1441	CCCGAAGAAC	GTTTTCCAAT	GATGAGCACT	TTTAAAGTTC	TGCTATGTGG	CGCGGATTA
1501	TCCCGTATTG	ACGCCGGGCA	AGAGCAACTC	GGTCGCGGCA	TACACTATTC	TACAGAATGAC
1561	TTGGTTGAGT	ACTCACCAGT	CACAGAAAAG	CATCTTACGG	ATGGCATTGAC	AGTAAGAGAA
1621	TTATGCAGTG	CTGCCATAAC	CATGAGTGAT	AACACTGCGG	CCAACCTACT	CTGACAACG
1681	ATCGGAGGAC	CGAAGGAGCT	AACCGCTTTT	TTGCACAACA	TGGGGGATCA	TGTAACTCGC
1741	CTTGATCGTT	GGGAACCGGA	GCTGAATGAA	GCCATACCAA	ACGACGAGCG	TGACACCACG
1801	ATGCCTGCAG	CAATGGCAAC	AACGTTGCGC	AAACTATTAA	CTGGCGAACT	ACTTACTCTA
1861	GCTTCCCGGC	AACAATTAAT	AGACTGGATG	GAGCGGATA	AAGTTGCAGG	ACCACTTCTG
1921	CGCTCGGCC	TTCCGGCTGG	CTGGTTTATT	GCTGATAAAT	CTGGAGCCGG	TGAGCGTGGT
1981	TCTCGCGGTA	TCATTGCAGC	ACTGGGGCCA	GATGGTAAAG	CCTCCCGTAT	CGTAGTTATC
2041	TACACGACGG	GGAGTCAGGC	AACTATGGAT	GAACGAAATA	GACAGATCGC	TGAGATAGGT
2101	GCCTCACTGA	TTAAGCATTG	GTAACGTGTA	GACCAAGTTT	ACTCATATAT	ACTTTAGATT
2161	GATTTAAAAC	TTCATTTTTA	ATTTAAAAGG	ATCTAGGTGA	AGATCCTTTT	TGATAATCTC
2221	ATGACCAAAA	TCCCTTAACG	TGAGTTTTCG	TTCCACTGAG	CGTCAGACCC	CGTAGAAAAG
2281	ATCAAAGGAT	CTTCTTGAGA	TCCTTTTTTT	CTGCGCGTAA	TCTGCTGCTT	GCAAAACAAA
2341	AAACCACCGC	TACCAGCGGT	GGTTTGTTTG	CCGGATCAAG	AGCTACCAAC	TCTTTTTCCG
2401	AAGGTAAC TG	GCTTCAGCAG	AGCCGAGATA	CCAATACTG	TCCTTCTAGT	TGAGCCGTAG
2461	TTAGGCCACC	ACTTCAAGAA	CTCTGTAGCA	CCGCCTACAT	ACCTCGCTCT	GCTAATCCTG
2521	TTACCAAGTGG	CTGCTGCCAG	TGGCGATAAG	TCGTGCTTTA	CCGGGTTGGA	CTCAAGACGA
2581	TAGTTACCGG	ATAAGGCGCA	GCGGTGCGGC	TGAACGGGGG	GTTCTGTGAC	ACAGCCCAGC
2641	TTGGAGCGAA	CGACCTACAC	CGAACTGAGA	GTACTACAGC	GTGAGCTATG	AGAAAGCGCC
2701	ACGCTTCCCG	AAGGGAGAAA	GGCGGACAGG	TATCCGGTAA	GCGGCAGGGT	CGGAACAGGA
2761	GAGCGCACGA	GGGAGCTTCC	AGGGGGA AAC	GCCTGGTATC	TTTATAGTCC	TGTCGGGTTT
2821	CGCCACCTCT	GACTTGAGCG	TCGATTTTTG	TGATGCTCGT	CAGGGGGGCG	GAGCCTATGG
2881	AAAAACGCCA	GCAACGCGGC	CTTTTTACGG	TTCTGGCCT	TTTTGTGGCC	TTTTGTCTAC
2941	ATGTTCTTTT	CTGCGTTATC	CCCTGATTCT	GTGGATAACC	GTATTACCGC	CTTTGAGTGA
3001	GCTGATACCG	CTCGCCG CAG	CCGAACGACC	GAGCGCAGCG	AGTCAGTGAG	CGAGGAAGCG
3061	GAAGAGCGCC	TGATGCGGTA	TTTTCTCCTT	ACGCATCTGT	GCCGTATTTT	ACACCGCATA
3121	TATGGTGCAC	TCTCAGTACA	ATCTGCTCTG	ATGCCGCATA	GTTAAGCCAG	TATACACTCC
3181	GCTATCGCTA	CGTACTGCGG	TCATGGCTGC	GCCCGACAC	CGCCCAACAC	CCCTGACCC
3241	GCCCTGACGG	GCTTGTCTGC	TCCCGGCATC	CGTTACAGA	CAAGCTGTGA	CCGTCTCCGG
3301	GAGCTGCATG	TGTCAGAGGT	TTTACCAGTC	ATCACCGAAA	CGCGGAGGC	AGCTGCGGTA
3361	AAGCTCATCA	GCGTGGTCTG	GAAGCGATT C	ACAGATGTCT	GCCTGTTCAT	CCGCGTCCAG
3421	CTCGTTGAGT	TTCTCCAGAA	CGGTTAATGT	CTGGCTTCTG	ATAAAGCGGG	CCATGTTAAG
3481	GGCGGTTTTT	TCCTGTTTTG	TCACTGATGC	CTCCGTGTAA	GGGGGATTTT	TGTTTATGGG
3541	GGTAATGATA	CCGATGAAAC	GAGAGAGGAT	GCTCACGATA	CGGGTACTG	ATGATGAACA
3601	TGCCCGGTTA	CTGGAACGTT	GTGAGGGTAA	ACA ACTGGCG	GTATGGATGC	GGCGGGACCA
3661	GAGAAAAATC	ACTCAGGGTC	AATGCCAGCG	CTTCGTTAAT	ACAGATGTAG	GTGTTCCACA
3721	GGGTAGCCAG	CAGCATCCTG	CGATGCAGAT	CCGGAACATA	ATGGTGCAGG	GCGCTGACTT
3781	CCGCGTTTT C	AGACTTTACG	AAACACGGAA	ACCGAAGACC	ATTCATGTTG	TTGCTCAGGT

3841	CGCAGACGTT	TTGCAGCAGC	AGTCGCTTCA	CGTTCGCTCG	CGTATCGGTG	ATTCATTCTG
3901	CTAACCAAGT	AGGCAACCCC	GCCAGCCTAG	CCGGGTCTCT	AACGACAGGA	GCACGATCAT
3961	CTGCACCCGT	GGGGCCGCCA	TGCCGGCGAT	AATGGCCTGC	TTCTCGCCGA	AACGTTTGGT
4021	GGCGGGACCA	GTGACGAAGG	CTTGAGCGAG	GGCGTGCAAG	ATTCCGAATA	CCGCAAGCGA
4081	CAGGCCGATC	ATCGTCGCGC	TCCAGCGAAA	GCGGTCTCTG	CCGAAAATGA	CCCAGAGCGC
4141	TGCCGGCACC	TGTCCTACGA	GTTGCATGAT	AAAGAAGACA	GTCATAAGTG	CGGCGACGAT
4201	AGTCATGCCC	CGCGCCACC	GGAAGGAGCT	GACTGGGTTG	AAGGCTCTCA	AGGGCATCGG
4261	TCGAGATCCC	GGTGCCTAAT	GAGTGAGCTA	ACTTACATTA	ATTGCGTTGC	GCTCACTGCC
4321	CGCTTTCCAG	TCGGGAAACC	TGTCGTGCCA	GCTGCATTAA	TGAATCGGCC	AACGCGCGGG
4381	GAGAGGCGGT	TTGCGTATTG	GGCGCCAGGG	TGGTTTTTCT	TTTCAACAGT	GAGACGGGCA
4441	ACAGTGATT	GCCCTTACC	GCCTGGCCCT	GAGAGAGTTG	CAGCAAGCGG	TCCACGCTGG
4501	TTTTGCCCCAG	CAGGCGAAAA	TCCTGTTTGA	TGGTGGTTAA	CGGCGGGATA	TAACATGAGC
4561	TGTCTTCGGT	ATCGTCGTAT	CCCACTACCG	AGATATCCGC	ACCAACGCGC	AGCCCGGACT
4621	CGGTAATGGC	GCGCATTGCG	CCCAGCGCCA	TCTGATCGTT	GGCAACCAGC	ATCGCAGTGG
4681	GAACGATGCC	CTCATTACAG	ATTTGCATGG	TTTGTGAAA	ACCGGACATG	GCACTCCAGT
4741	CGCCTTCCCG	TTCCGCTATC	GGTGAAATTT	GATTGCGAGT	GAGATATTTA	TGCCAGCCAG
4801	CCAGACGAG	ACGCGCCGAG	ACAGAACTTA	ATGGGCCCGC	TAACAGCGCG	ATTTGCTGGT
4861	GACCCAATGC	GACCAGATGC	TCCACGCCCA	GTCGCGTACC	GTCCTTCATGG	GAGAAAATAA
4921	TACTGTTGAT	GGGTGTCTGG	TCAGAGACAT	CAAGAAATAA	CGCCGGAACA	TTAGTGCAGG
4981	CAGCTTCCAC	AGCAATGGCA	TCCTGGTCAT	CCAGCGGATA	GTTAATGATC	AGCCCACTGA
5041	CGCGTTGCGC	GAGAAGATTG	TGCACCGCCG	CTTTACAGCG	TTCCGACGCG	CCTTCGTCTA
5101	CCATCGACAC	CACCACGCTG	GCACCCAGTT	GATCGGCGCG	AGATTTAATC	GCCCGGACAA
5161	TTTGGCAGCG	CGCGTGCAGG	GCCAGACTGG	AGGTGGCAAC	GCCAAATCAGC	AACGACTGTT
5221	TGCCCGCCAG	TTGTTGTGCC	ACGCGGTTGG	GAATGTAATT	CAGCTCCGCC	ATCGCCGCTT
5281	CCACTTTTTC	CCGCGTTTTC	GCAGAAACGT	GGTGGCCCTG	GTTCAACCAG	CGGGAACCGG
5341	TCTGATAAGA	GACACCGGCA	TACTCTGCGA	CATCGTATAA	CGTTACTGTT	TTCACTTTCA
5401	CCACCCTGAA	TTGACTCTCT	TCCGGGCGCT	ATCATGCCAT	ACCCGGAAG	GTTTTGCGCC
5461	ATTGATGTTG	GTCCGGGATC	TCGACGCTCT	CCCTTATGCG	ACTCCTGCAT	TAGGAAGCAG
5521	CCAGTAGTA	GGTTGAGGCC	GTTGAGCACC	GCCGCCGCAA	GGAATGGTGC	ATGCAAGGAG
5581	ATGGCGCCCA	ACAGTCCCCC	AGCCACGGGG	GCCCAACGCG	CCTGCCACCA	GAAACAAGCG
5641	CTCATGAGCC	CGAAGTGCGG	AGCCCGATCT	TCCCACATCG	TGATGTCGGC	GATATAGGCG
5701	CCAGCAACCG	CACCTGTGGC	GCCGGTGATG	CCGGCCACGA	TGCGTCCGGC	GTAGAGGATC
5761	G					

pGEX-CS

1	AGCTTATCGA	CTGCACGGTG	CACCAATGCT	TCTGGCGTCA	GGCAGCCATC	GGAAGCTGTG
61	GTATGGCTGT	GCAGGTCGTA	AATCACTGCA	TAATTCGTGT	CGCTCAAGGC	GCACCTCCGT
121	TCTGGATAAT	GTTTTTTGCG	CCGACATCAT	AACGGTCTCG	GCAAATATTC	TGAAATGAGC
181	TGTTGACAAT	TAATCATCGG	CTCGTATAAT	GTGTGGAATT	GTGAGCGGAT	AACAATTTCA
241	CACAGGAAAC	AGTATTCATG	TCCCCTATAC	TAGGTTATTG	GAAAATTAAG	GGCCTTGTGC
301	AACCCACTCG	ACTTCTTTTG	GAATATCTTG	AAGAAAAATA	TGAAGAGCAT	TTGATGAGC
361	GCGATGAAGG	TGATAAATGG	CGAAACAAAA	AGTTTGAATT	GGGTTTGGAG	TTTCCCAATC
421	TTCTTATTA	TATTGATGGT	GATGTTAAAT	TAACACAGTC	TATGGCCATC	ATACGTTATA
481	TAGCTGACAA	GCACAACATG	TTGGGTGGTT	GTCCAAAAGA	GCGTGCAGAG	ATTTCAATGC
541	TTGAAGGAGC	GGTTTTGGAT	ATTAGATACG	GTGTTTCGAG	AATTGCATAT	AGTAAAGACT
601	TTGAAACTCT	CAAAGTTGAT	TTTCTTAGCA	AGTACTCTGA	AATGCTGAAA	ATGTTCGAAG
661	ATCGTTTATG	TCATAAAACA	TATTTAAATG	GTGATCATGT	AACCCATCCT	GACTTCATGT
721	TGTATGACGC	TCTTGATGTT	GTTTATACAA	TGGACCCAAT	GTGCCCTGGAT	GCGTCCCAA
781	AATTAGTTTG	TTTTAAAAAA	CGTATTGAAG	CTATCCCACA	AATTGATAAG	TACTTGAAT
841	CCAGCAAGTA	TATAGCATGG	CCTTTGCAAG	CTATGCCAAG	CACGTTTGGT	GGTGGCGACC
901	ATCTCGAGAA	TCTTTATTTT	CAGGGCGCCA	TGGCACTCAT	CCTTGGCACA	GTCAACGCTA
961	ACATCTGAA	GGAAGTGTTT	GGTGGAGCTC	GGTACCCGGG	GATCCCGGG	AATTATCGT
1021	GACTGACTGA	CGATCTGCCT	CGCGCGTTTC	GGTGATGACG	GTGAAAACCT	CTGACACATG
1081	CAGCTCCCGG	AGACGGTCA	AGCTTGCTGT	TAAGCGGATG	CCGGGAGCAG	ACAAGCCCGT
1141	CAGGGCGCGT	CAGCGGGTGT	TGGCGGGTGT	CGGGGCGCAG	CCATGACCCA	GTCACGTAGC
1201	GATAGCGGAG	TGTATAATTC	TTGAAGACGA	AAGGGCCTCG	TGATACGCCT	ATTTTTATAG
1261	GTAAATGTCA	TGATAATAAT	GGTTTCTTAG	ACGTGAGGTG	GCACTTTTCG	GGGAAATGTG
1321	CGCGGAACCC	CTATTTGTTT	ATTTTTCTAA	ATACATTCAA	ATATGTATCC	GCTCATGAGA
1381	CAATAACCC	GATAAATGCT	TCAATAATAT	TGAAAAAGGA	AGAGTATGAG	TATTTCAACAT
1441	TTCCGTGTGCG	CCCTTATTCC	CCTTTTTCGG	GCATTTTGCC	TTCTGTGTTT	TGCTCACCCA
1501	GAAACGCTGG	TGAAAGTAAA	AGATGCTGAA	GATCAGTTGG	GTGCACGAGT	GGGTTACATC
1561	GAACGTGGATC	TCAACAGCGG	TAAGATCCTT	GAGAGTTTTT	GCCCGAAGA	ACGTTTTCCA
1621	ATGATGAGCA	CTTTTAAAGT	TCTGCTATGT	GCGCGGGTAT	TATCCCGTGT	TGACGCCGGG
1681	CAAGAGCAAC	TCGGTCGCCG	CATACACTAT	TCTCAGAATG	ACTTGGTTGA	GTACTCACCA
1741	GTCACAGAAA	AGCATCTTAC	GGATGGCATG	ACAGTAAGAG	AATTATGCAG	TGCTGCCATA
1801	ACCATGAGTG	ATAACACTGC	GGCCAACTTA	CTTCTGACAA	CGATCGGAGG	ACCGAAGGAG
1861	CTAACCCGCTT	TTTTGCACAA	CATGGGGGAT	CATGTAATCT	GCCTTGATCG	TTGGGAACCG
1921	GAGCTGAATG	AAGCCATACC	AAAGCAGGAG	CGTGACACCA	GATGCCTGCG	AGCAATGCGA
1981	ACAACGTTGC	GCAAACTATT	AACTGGCGAA	CTACTTACTC	TAGTTTCCCG	GCAACAATTA
2041	ATAGACTGGA	TGGAGGCGGA	TAAAGTTGCA	GGACCCTTCT	TGCGTCCGGC	CCTTCCGGCT

Appendix A1: Plasmids

2101	GGCTGGTTTA	TTGCTGATAA	ATCTGGAGCC	GGTGAGCGTG	GGTCTCGCGG	TATCATTGCA
2161	GCACTGGGGC	CAGATGGTAA	GCCCTCCCGT	ATCGTAGTTA	TCTACACGAC	GGGGAGTCAG
2221	GCAACTATGG	ATGAACGAAA	TAGACAGATC	GCTGAGATAG	GTGCCTCACT	GATTAAGCAT
2281	TGGTAACTGT	CAGACCAAGT	TTACTCATAT	ATACTTTAGA	TTGATTTAAA	ACTTCATTTT
2341	TAATTTAAAA	GGATCTAGGT	GAAGATCCTT	TTTGATAATC	TCATGACCAA	AATCCCTTAA
2401	CGTGAGTTTT	CGTTCACACTG	AGCGTCAGAC	CCCGTAGAAA	AGATCAAAGG	ATCTTCTTGA
2461	GATCCTTTTT	TTCTGCGCGT	AATCTGTGTC	TTGCAAACAA	AAAAACCACC	GCTACCAGCG
2521	GTGGTTTGTT	TGCCGGATCA	AGAGCTACCA	ACTCTTTTTTC	CGAAGGTAAC	TGGCTTCAGC
2581	AGAGCGCAGA	TACCAAATAC	TGCTCTTCTA	GTGTAGCCGT	AGTTAGGCCA	CCACTTCAAG
2641	AACTCTGTAG	CACCGCCTAC	ATACCTCGCT	CTGCTAATCC	TGTTACCAGT	GGTGCTGCC
2701	AGTGGCGATA	AGTCGTGTCT	TACCGGGTTG	GACTCAAGAC	GATAGTTACC	GGATAAGGCG
2761	CAGCGGTCGG	GCTGAACGGG	GGGTTCTGTG	ACACAGCCA	GCTTGGAGCG	AACGACTAC
2821	ACCGAACTGA	GATACCTACA	GCGTGAGCTA	TGAGAAAGCG	CCACGCTTCC	CGAAGGGAGA
2881	AAGGCGGACA	GGTATCCGGT	AAGCGGCAGG	GTCGGAACAG	GAGAGCGCAC	GAGGGAGCTT
2941	CCAGGGGAA	ACGCCGTGTA	TCTTTATAGT	CCTGTCGGGT	TTGCGCACT	CTGACTTGAG
3001	CGTCGATTTT	TGTGATGCTC	GTCAGGGGGG	CGGAGCCTAT	GGAAAAACGC	CAGCAACGCG
3061	GCCTTTTTAC	GGTTCTTGGC	CTTTTGCTGG	CCTTTTGCTC	ACATGTTCTT	TCCTGCGTTA
3121	TCCCCTGATT	CTGTGGATAA	CCGTATTACC	GCCTTTGAGT	GAGCTGATAC	CGCTCGCCGC
3181	AGCCGAACGA	CCGAGCGCAG	CGAGTCAGTG	AGCGAGGAAG	CGGAAGAGCG	CCTGATGCGG
3241	TATTTTCTCC	TTACGCATCT	GTGCGGTATT	TCACACCACA	TAAATCCGA	CACCATCGAA
3301	TGGTGCAAAA	CCTTTCGCGG	TATGGCATGA	TAGCGCCCGG	AAGAGAGTCA	ATTAGGGTG
3361	GTGAATGTGA	AACCAAGTAA	GTTATACGAT	GTCGCAGAGT	ATGCCGGTGT	CTCTTATCAG
3421	ACCGTTTCCC	GCGTGGTGAA	CCAGGCCAGC	CACGTTTCTG	CGAAAACGCG	GGAAAAAGTG
3481	GAAGCGGCGA	TGGCGGAGCT	GAATTACATT	CCCACCAGCG	TGGCACAACA	ACTGGCGGGC
3541	AAACAGTCGT	TGCTGATTGG	CGTTGCCACC	TCCAGTCTGG	CCCTGCACGC	GCCGTGCGAA
3601	ATTGTCGCGG	CGATTAATTC	TCGCGCCGAT	CAACTGGGTG	CCAGCGTGGT	GGTGTGATG
3661	GTAGAACGAA	GCGGCGTCGA	AGCCTGTAAA	GCGGCGGTGC	ACAATCTTCT	CGCGCAACGC
3721	GTCAGTGGGC	TGATCATTAA	CTATCCGCTG	GATGACCAGG	ATGCCATTGC	TGTGGAAGCT
3781	GCCTGCACCTA	ATGTTCCGGC	GTTATTTCTT	GATGTCTCTG	ACCAGACACC	CATCAACAGT
3841	ATTATTTTCT	CCCATGAAGA	CGGTACGCGA	CTGGGCGTGG	AGCATCTGGT	CGCATTTGGT
3901	CACCAGCAAA	TCGCGCTGTT	AGCGGGCCCA	TTAAGTTCTG	TCTCGGCGCG	TCTGCGTCTG
3961	GCTGGCTGGC	ATAAATATCT	CACTCGCAAT	CAAATTCAGC	CGATAGCGGA	ACGGGAAGGC
4021	GACTGGAGTG	CCATGTCGCG	TTTTCAACAA	ACCATGCAAA	TGCTGAATGA	GGGCATCGTT
4081	CCCACTGCGA	TGCTGGTTGC	CAACGATCAG	ATGGCGCTGG	GCGCAATGCG	CGCCATTACC
4141	GAGTCCGGGC	TGCGCGTTGG	TGCGGATATC	TCGGTAGTGG	GATACGACGA	TACCGAAGAC
4201	AGTCTCATGTT	ATATCCCGCC	GTCAACCACC	ATCAAACAGG	ATTTTCGCCT	GCTGGGGCAA
4261	ACCAGCGTGG	ACCGCTTGCT	GCAACTCTCT	CAGGGCCAGG	CGGTGAAGGG	CAATCAGCTG
4321	TTGCCCGTCT	CACTGGTGAA	AAGAAAAACC	ACCCTGGCGC	CCAATACGCA	AACCGCCTCT
4381	CCCCGCGCGT	TGGCCGATTCT	ATTAATGCAG	CTGGCACGAC	AGGTTTCCCG	ACTGGAAAAGC
4441	GGGCAGTGAG	CGCAACGCAA	TTAATGTGAG	TTAGTCACT	CATTAGGCAC	CCCAGGCTTT
4501	ACACTTTATG	CTTCCGGCTC	GTATGTTGTG	TGGAATTGTG	AGCGGATAAC	AATTTACAC
4561	AGGAAACAGC	TATGACCATG	ATTACGGATT	CACTGGCCGT	CGTTTTACAA	CGTCGTGACT
4621	GGGAAAACCC	TGGCGTTACC	CAACTTAATC	GCCTTGACGC	ACATCCCCCT	TTCGCCAGCT
4681	GGCGTAATAG	CGAAGAGGCC	CGCACCATC	GCCCTTCCCA	ACAGTTGCGC	AGCCTGAATG
4741	GCGAATGGCG	CTTTGCCTGG	TTTCCGGCAC	CAGAAGCGGT	GCCGGAAAAGC	TGGCTGGAGT
4801	GCGATCTTCC	TGAGGCCGAT	ACTGTCGTCG	TCCCTCAAAA	CTGGCAGATG	CACGGTTACG
4861	ATGCGCCCAT	CTACACCAAC	GTAACCTATC	CCATTACGGT	CAATCCGCGC	TTTGTTCCCA
4921	CGGAGAATCC	GACGGGTTGT	TACTCGCTCA	CATTTAATGT	TGATGAAAGC	TGGCTACAGG
4981	AAGGCCAGAC	GCGAATTATT	TTTGATGGCG	TTGGAATT		

Appendix A2: Buffers and media

SDS- polyacrylamide gel electrophoresis:

4X SDS loading buffer

- 200 mM Tris/HCl (pH 6.8)
- 8% (w/v) SDS
- 40% (v/v) glycerol
- 4% (v/v) 2-mercaptoethanol
- 50 mM EDTA
- 0.08% (w/v) bromphenol blue

SDS-PAGE running buffer

- 0.15 M Tris/HCl (pH 8)
- 250 mM glycine
- 1% (w/v) SDS

Coomassie Brilliant Blue staining solution

- 10% (v/v) ethanol
- 5% (v/v) acetic acid
- 0.1% (v/v) Coomassie solution

Protein expression:

Minimal Medium (pH 7.4)

- 42 mM Na₂HPO₄
- 22 mM KH₂PO₄
- 8.5 mM NaCl
- 2 mM Mg₂SO₄
- mM CaCl₂
- 1X Vitamin mix
- 1X Trace elements
- 100 µg/mL ampicillin
- 1g/L ¹⁵NH₄Cl
- 20 g/L (2%) ¹²C-glucose, 2 g/L ¹³C-glucose

5000X Trace elements

- 50 mM FeCl₃
- 10 mM MnCl₂
- 2 mM CoCl₂
- 2 mM Na₂MoO₄
- 2 mM H₂BO₃
- 2 mM NiSO₄
- 10 mM ZnSO₄
- 2 mM CuCl₂
- 2 mM NaSeO₃

2000X Vitamin mix

- 100 mM Thiamine (vitamin B₁)
- 100 mM Biotin (vitamin B₃)
- 100 mM Niacin (vitamin H)
- 10 mM Cobalamin (vitamin B₁₂)

Lysogeny broth medium after Miller (LB)

- 10 g/L tryptone
- 5g/L yeast extract
- 10g/L NaCl

Protein purification

SUMO-fusion proteins:

Buffer A

- 50 mM potassium phosphate (pH 7 - 8)
- 200 - 300 mM NaCl
- 3 mM DTT or 5 mM 2-mercapto ethanol for small proteins containing cysteines

Buffer B

- 50 mM potassium phosphate (pH 7 - 8)
- 200 - 300 mM NaCl
- 500 mM imidazole
- 3 mM DTT or 5 mM 2-mercaptoethanol for small proteins containing cysteines

Dialysis buffer

- 50 mM potassium phosphate (pH 7 - 8)
- 200 - 300 mM NaCl
- 3 mM DTT or 5 mM 2-mercapto ethanol for small proteins containing cysteines

SEC buffer

- 50 mM potassium phosphate (pH 7 - 8)
- 200 - 300 mM NaCl
- 3 mM DTT or 5 mM 2-mercapto ethanol for small proteins containing cysteines

GST-fusion proteins:

Buffer A

- 25 mM TRIS (pH 8)
- 200 mM NaCl
- 100 μ M ZnCl₂
- 5 mM DTT

Buffer B

- 25 mM TRIS (pH 8)
- 200 mM NaCl
- 100 μ M ZnCl₂
- 5 mM DTT
- 10 mM glutathione

SEC buffer

- 25 mM TRIS (pH 8)
- 200 mM NaCl
- 5 mM DTT

Appendix A3: Purification strategy of peptides and small proteins

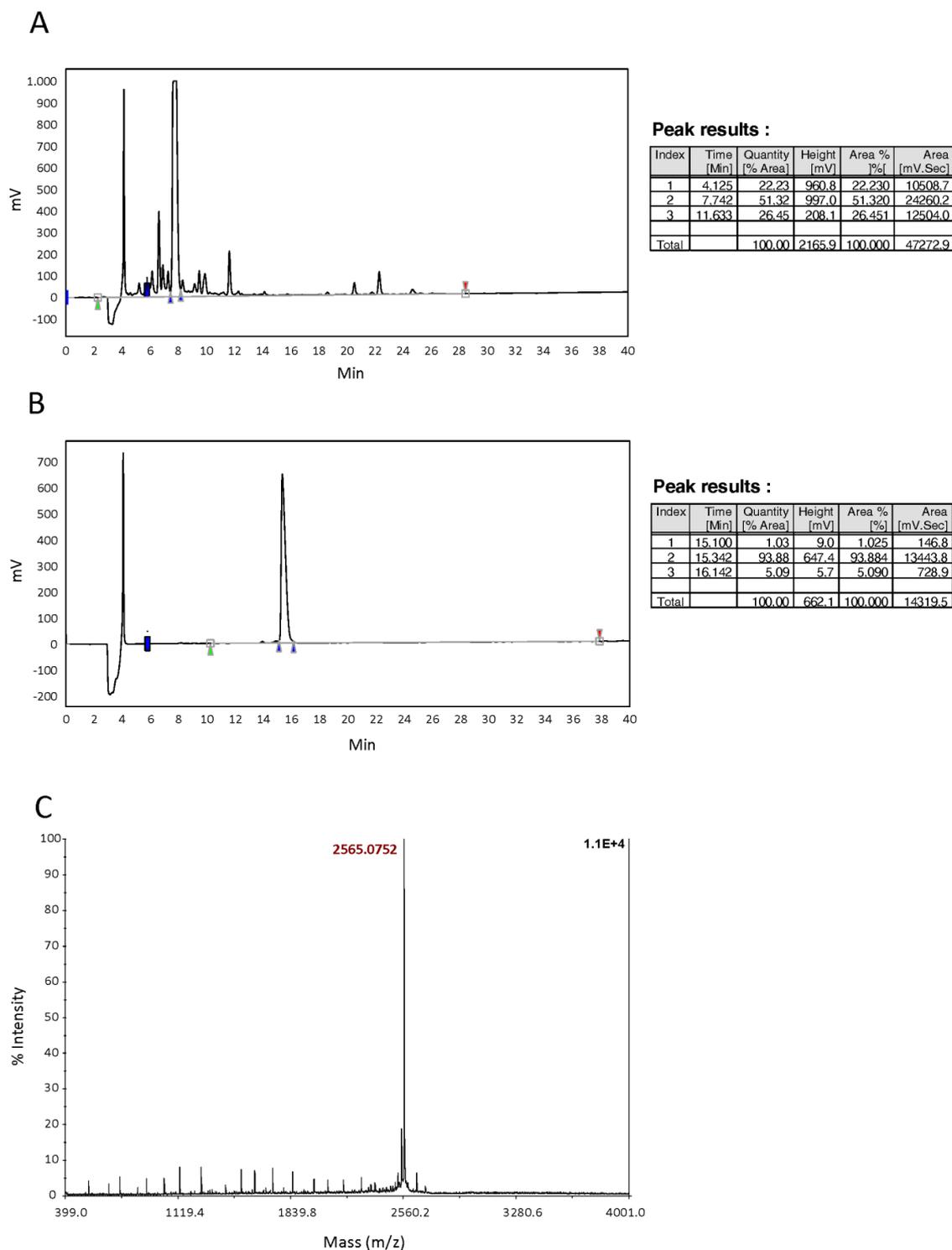


Figure A1: Purification strategy of peptides exemplified for P2. A – Preparative reversed-phase HPLC chromatogram of raw P2 peptide. Purification occurs using a gradient 20-60% in 40 minutes. The pure peptide fraction elutes at 28% acetonitrile. **B** – Analytical reversed-phase HPLC chromatogram of purified P2 peptide. A gradient 20-40% in 40 minutes was used. **C** – ESSI Mass spectrometry spectra P2 peptide. Expected molecular weight is 2565.9 Da.

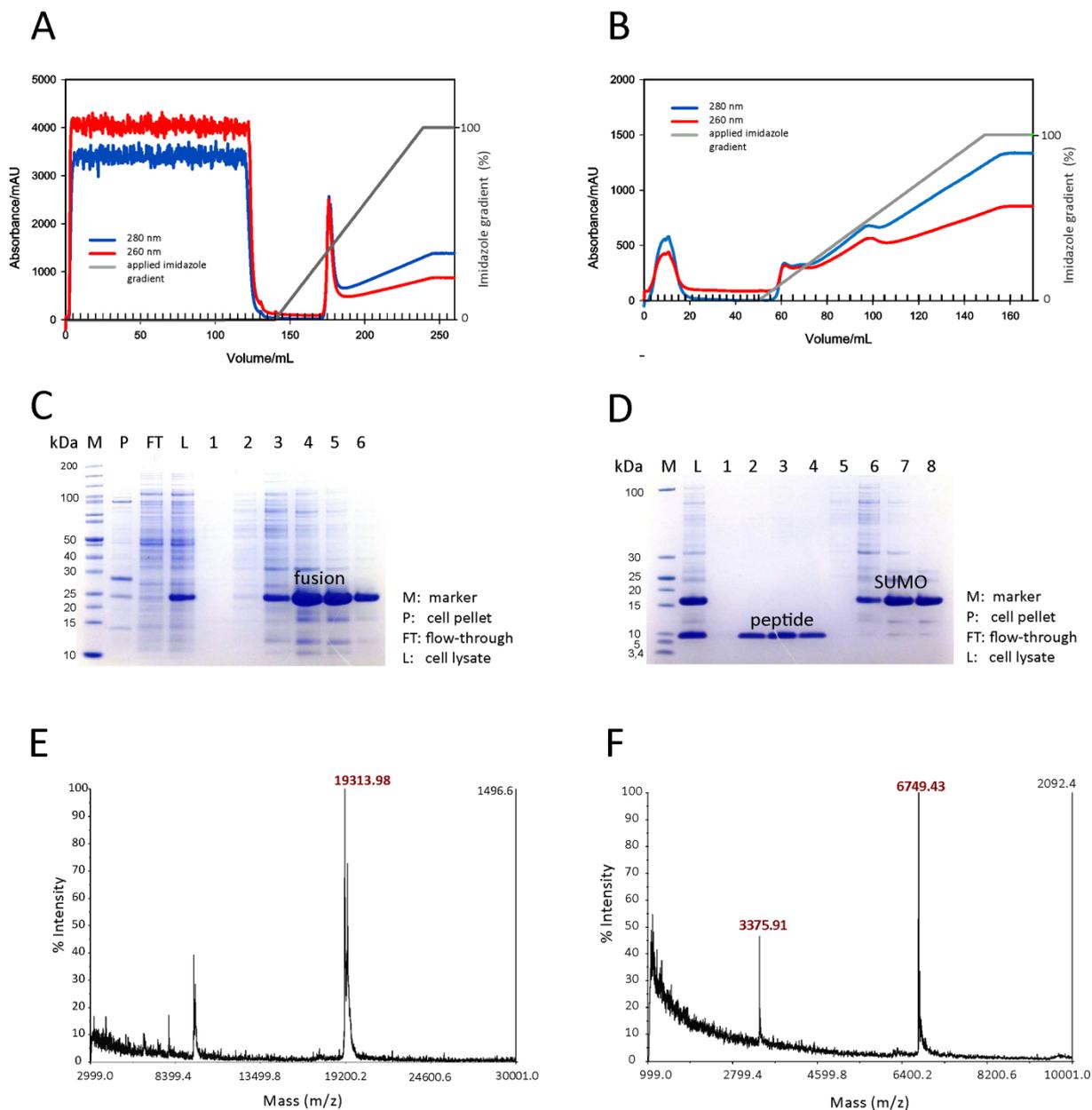


Figure A2: Purification strategy of SUMO-fusion proteins exemplified for P17 small protein. **A** – Ni-NTA affinity chromatogram of SUMO-P17 fusion protein. The absorption at 280 nm and 260 nm are shown in blue and red, respectively (left-handed y-axis) and are plotted as function of the volume. The applied imidazole gradient in percent of buffer B (right-handed y-axis) is shown in gray. The collected fractions are shown on the x-axes. **B** – Reverse Ni-NTA affinity chromatogram of SUMO-P17 fusion protein after cleavage with SUMO-protease. **C, D** - The corresponding SDS-PAGE analyses of relevant fractions from Ni-NTA (**C**) and reverse Ni-NTA (**D**) affinity chromatography. Marker (*PageRuler™ Unstained Protein Ladder, ThermoFisher Scientific*) (M) was used. Visualization was accomplished by coomassie staining. The protein bands are indicated in the picture. **E, F** – MALDI Mass spectrometry spectra of SUMO-P17 fusion protein (**E**) and cleaved isolated P17 small protein (**F**) are shown. Expected molecular weight is 19331 Da for ^{15}N labeled fusion protein and 6761 Da for ^{15}N labeled isolated P17 small protein.

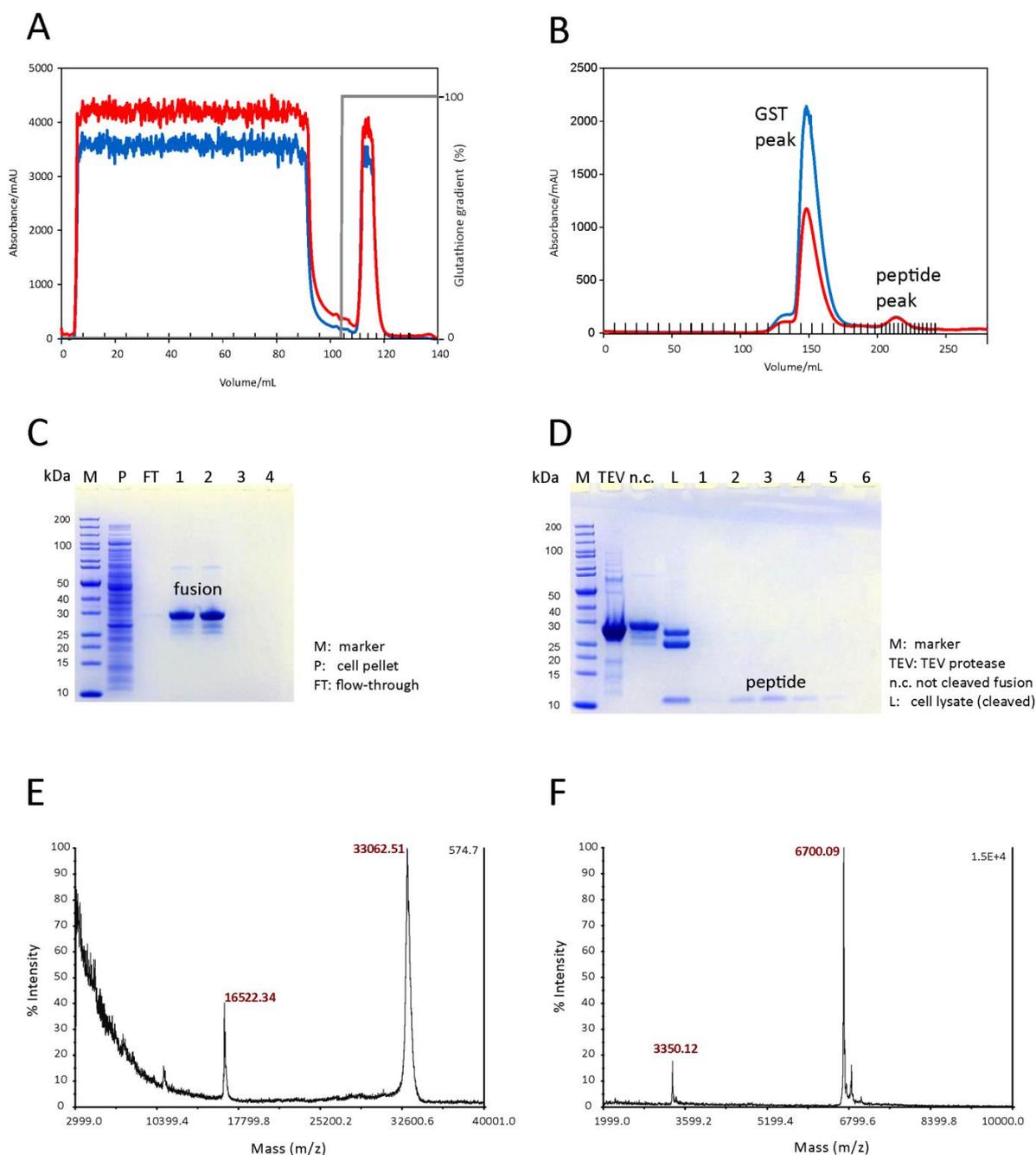


Figure A3: Purification strategy of GST-P19 small protein. **A** – GST affinity chromatogram of GST-P19 fusion protein. The absorption at 280 nm and 260 nm are shown in blue and red, respectively (left-handed y-axis) and are plotted as function of the volume. The applied glutathione gradient in percent of buffer B (right-handed y-axis) is shown in gray. The collected fractions are shown on the x-axes. **B** – Size-exclusion chromatogram of GST-P19 fusion protein after cleavage with TEV-protease. **C, D** –The corresponding SDS-PAGE analyses of relevant fractions from GST affinity chromatogram (**C**) and Size-exclusion chromatogram (**D**). Marker (*PageRuler™ Unstained Protein Ladder, ThermoFisher Scientific*) (M) was used. Visualization was accomplished by coomassie staining. The protein bands are indicated in the picture. **E, F** – MALDI Mass spectrometry spectra of GST-P19 fusion protein (**E**) and cleaved isolated P19 small protein (**F**) are shown. Expected molecular weight is 34001 Da for ^{15}N labeled fusion protein and 6702 Da for ^{15}N labeled isolated P19 small protein.

Appendix A4: Overview of small proteins screened in this study

Table A1. Extinction coefficients, isoelectric point of peptides and small proteins achieved by ProtParam online tool:

Intenal name	Sequence length, number of amino acids	Molecular weight, kDa	ϵ (L·mmol ⁻¹ ·cm ⁻¹)	pI
P1	14	1,6	1490 [b]	10,05
P2	23	2,6	- [a]	10,84
P3	18	1,9	- [a]	9,99
P4	28	3,1	5500	5,43
P5	27	2,9	0 [b]	10,20
P6	29	3,1	- [a]	10,00
P7	53	6,2	12490	8,88
P8	61	7,1	19480	4,39
P9	70	7,7	5500	5,62
P10	71	7,9	4470	10,81
P11	38	4,0	1490	5,42
P12	23	2,8	- [a]	8,56
P13	18	2,2	0 [b]	10,88
P14	31	3,7	0 [b]	10,90
P15	51	5,5	0 [b]	11,26
P16	43	4,8	0 [b]	4,56
P17	60	6,7	8480	6,04
P18	61	6,9	9970	4,65
P19	59	6,5	1490	6,68
P20	14	1,8	16500	9,50

[a] Due to the absence of the tryptophan, tyrosine or cysteine in the sequence, the protein should not be visible by UV spectrophotometry.

[b] Due to the absence of the tryptophan, the computed extinction coefficient error can be more than 10%^[105].

Appendix A4: Overview of small proteins screened in this study

Table A2. Overview of the amino acid sequences of small proteins screened in this study:

Internal name	Sequence
P1	VSYLKRCHLAGIAR
P2	LDSNTSHKNSVRHVLGLAQRVSF
P3	VSKKVLERGVGTTGEARL
P4	MEEVNQIAGGHPTLKDGVCFPPARLFW
P5	VSRGLREGCRFSRHSASHESMPGGSHS
P6	LKIAMGAGLTESRAKEAFKASKKKVAEIV
P7	MGWIEGRDLRIGTQPLPSTSKLNFRNITSFLSFWLNPIPSTCTFIRVYIDFC
P8	VTIWEYDVKEIRFSEWSKAKEDLNNLGVEGWELIKFSNEIDENGMVAAVFKRPVDYVDAAF
P9	MAAFETTRPAPFGAISTFHFVQRMSDLLATVVAWNDARATRAALSKLSDREDDIGLCRGDIDDICALRR
P10	MAYANTTRIGHHGLGDRVSALVASVKLALAQRRIRYQTVRELNSLTTRELSDLGIHRSMITRIAMEAAYGL
P11	MVSMRSCMCCGEPISETRHLGVCIQNGCTSYADACGQ
P12	VPVMKNLADSMMSPPMSSEARKLS
P13	MLVRDLEQLLFKINLLSR
P14	MRPKHRRRASLFRCKNMQ CADGMAVMHIK
P15	MNFSVVVGPRGNQHKSESGSCRILQGSLEVRGRCVASRLPLQTRRPPCV L
P16	MRDTAMSQRKDDHLDIVLDERTAPATVAAGRECIRFELSSDGD
P17	MNKAHFEVFDAAADKYRWRLVHDNGNILADSGEGYASKQKAKQGIESVKRNAPDADVIEA
P18	MSSSPWTANFATEKSKCAADVQRLLKYPQPVVYEVMSSELLRQEMREQFAGAYAASQQSDD
P19	MSESEQRHAHQVSCGINIAGMSAATFKCPDCGQEISRCSKCRKQSNLYECPDCGFMGP
P20	MANTQNISIWWWAR

Appendix A4: Overview of small proteins screened in this study**Table A3. Overview of small protein origin and research groups working on these proteins:**

Internal name	Collaborative name	Organism	Research Group	University
P1	rreB	<i>Bradyrhizobium japonicum</i>	Elena Evguenieva-Hackenberg	Giessen
P2	μP44	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	Kiel
P3	rreR	<i>Dinoroseobacter shibae</i>	Elena Evguenieva-Hackenberg	Giessen
P4	μP37	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	Kiel
P5	na	<i>Sinorhizobium meliloti</i>	Elena Evguenieva-Hackenberg	Giessen
P6	μP31	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	Kiel
P7	μP41	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	Kiel
P8	μP36	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	Kiel
P9	RSP_0557	<i>Rhodobacter sphaeroides</i>	Gabriele Klug	Giessen
P10	RSP_6037	<i>Rhodobacter sphaeroides</i>	Gabriele Klug	Giessen
P11	HVO_2983_A	<i>Haloferax volcanii</i>	Jörg Soppa	Frankfurt am Main
P12	SP26_1_SW	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	Kiel
P13	SP34_2_WW	<i>Methanosarcina mazei</i>	Ruth Schmitz-Streit	Kiel
P14	NGR_c15640	<i>Sinorhizobium fredii</i>	Wolfgang Streit	Hamburg
P15	repX	<i>Sinorhizobium fredii</i>	Wolfgang Streit	Hamburg
P16	NGR_a02780	<i>Sinorhizobium fredii</i>	Wolfgang Streit	Hamburg
P17	2922	<i>Haloferax volcanii</i>	Anita Marchfelder	Ulm
P18	A0101	<i>Haloferax volcanii</i>	Anita Marchfelder	Ulm
P19	HVO_2753	<i>Haloferax volcanii</i>	Jörg Soppa	Frankfurt am Main
P20	TrpL leader	<i>Sinorhizobium meliloti</i>	Elena Evguenieva-Hackenberg	Giessen

Appendix A4: Overview of small proteins screened in this study

Table A4. Classification of small protein origin screened in this study:

Species	Kingdom	Phylum	Class	Order	Family	Genus
<i>Dinoroseobacter shibae</i>	Bacteria	Proteobacteria	Alpha proteobacteria	Rhodobacterales	<i>Rhodobacteraceae</i>	<i>Dinoroseobacter</i>
<i>Rhodobacter sphaeroides</i>	Bacteria	Proteobacteria	Alpha proteobacteria	Rhodobacterales	<i>Rhodobacteraceae</i>	<i>Rhodobacter</i>
<i>Sinorhizobium meliloti</i>	Bacteria	Proteobacteria	Alpha proteobacteria	Rhizobiales	<i>Rhizobiaceae</i>	<i>Sinorhizobium</i>
<i>Sinorhizobium fredii</i>	Bacteria	Proteobacteria	Alpha proteobacteria	Rhizobiales	<i>Rhizobiaceae</i>	<i>Sinorhizobium</i>
<i>Bradyrhizobium japonicum</i>	Bacteria	Proteobacteria	Alpha proteobacteria	Rhizobiales	<i>Bradyrhizobiaceae</i>	<i>Bradyrhizobium</i>
<i>Methanosarcina mazei</i>	Archaea	Euryarchaeota	Methanomicrobia	Methanosarcinales	<i>Methanosarcinaceae</i>	<i>Methanosarcina</i>
<i>Haloferax volcanii</i>	Archaea	Euryarchaeota	Halobacteria	Halobacteriales	<i>Halobacteriaceae</i>	<i>Haloferax</i>

Appendix A5: NMR spectrometer

Table A5. Spectrometer and probe heads used in this study:

Spectrometer	Probe head
AV III 599 MHz	5 mm TCI ^1H , ^{13}C , ^{15}N , Z-GRD
AV III 600 MHz HD	Prodigy 5 mm TCI ^1H , ^{13}C , ^{15}N Z-GRD
AV 600 MHz	5 mm TCI cryo ^1H , ^{13}C , ^{31}P , Z-GRD
AV II 600 MHz	5 mm TCI cryo ^1H , ^{13}C , ^{15}N , Z-GRD
AV III 700 MHz HD	5 mm QCI cryo ^1H , ^{13}C , ^{15}N , ^{31}P , Z-GRD
AV 800 MHz	5 mm TXO cryo ^{13}C , ^1H , ^{15}N , Z-GRD
AV III 800 MHz HD	5 mm TCI cryo ^1H , ^{13}C , ^{15}N , Z-GRD
AV III 950 MHz	5 mm TCI cryo ^1H , ^{13}C , ^{15}N , Z-GRD

Appendix A6: NMR experiments

Table A6. NMR experiments recorded for the structural analysis of P19 protein at 298 K. Sample composition, parameters of the NMR experiments and spectrometer are shown.

	experiment	Avance Bruker	labeling	NS	mixing time [ms]	d1 [s]	NUS
2D	1H-15N BestTrosy	950 MHz	$^{15}\text{N}^{13}\text{C}$	8		0,2	
2D	1H-13C ct HSQC aliphatic	950 MHz	$^{15}\text{N}^{13}\text{C}$	16		1,5	
3D	1H-13C NOESY-HSQC	950 MHz	$^{15}\text{N}^{13}\text{C}$	16	120	1,2	
3D	1H-15N NOESY-HSQC	950 MHz	$^{15}\text{N}^{13}\text{C}$	16	120	1,2	
3D	1H-15N TOCSY-HSQC	600 MHz	$^{15}\text{N}^{13}\text{C}$	16	70	1,2	
3D	HNHA	600 MHz	$^{15}\text{N}^{13}\text{C}$	16		1	
3D	HNCACB	950 MHz	$^{15}\text{N}^{13}\text{C}$	16		0,2	25%
3D	HNCOCACB	950 MHz	$^{15}\text{N}^{13}\text{C}$	16		0,2	25%
3D	HNCO	950 MHz	$^{15}\text{N}^{13}\text{C}$	16		0,2	25%
3D	(H)CC(CO)NH	599 MHz	$^{15}\text{N}^{13}\text{C}$	16		1	25%
3D	H(CC)(CO)NH	599 MHz	$^{15}\text{N}^{13}\text{C}$	16		1	25%
3D	H(C)CH tocsy	600 MHz	$^{15}\text{N}^{13}\text{C}$	16		1	25%
3D	HNN	600 MHz	$^{15}\text{N}^{13}\text{C}$	32		1	25%
3D	HNNH	600 MHz	$^{15}\text{N}^{13}\text{C}$	32		1	25%
3D	(H)N(COCA)NNH	599 MHz	$^{15}\text{N}^{13}\text{C}$	16		1	25%
2D	HBCBCGCDHD	600 MHz	$^{15}\text{N}^{13}\text{C}$	32		1	
2D	HBCBCGCDCEHE	600 MHz	$^{15}\text{N}^{13}\text{C}$	32		1	
2D	zz-exchange	599 MHz	$^{15}\text{N}^{13}\text{C}$	32	400	4	

Appendix A6: NMR experiments

Table A7. NMR experiments recorded for the structural analysis of P17 protein. Sample composition, parameters of the NMR experiments and spectrometer are shown.

	experiment	Avance Bruker	labeling	NS	mixing time [ms]	d1 [s]	NUS
2D	1H-15N HSQC	700 MHz	$^{15}\text{N}^{13}\text{C}$	32		0,22	
2D	1H-13C HSQC aliphatic	800 MHz	^{15}N	128		1,5	
2D	1H-1H NOESY	600 MHz	^{15}N	64	100 300	1,5	
3D	1H-15N NOESY-HSQC	600 MHz	^{15}N	16	120	1,2	
3D	1H-15N TOCSY-HSQC	600 MHz	^{15}N	16	70	1,2	
3D	HNHA	600 MHz	^{15}N	16		1	
3D	HNCACB	700 MHz	$^{15}\text{N}^{13}\text{C}$	32		0,4	25%
3D	HNCOCACB	700 MHz	$^{15}\text{N}^{13}\text{C}$	32		0,4	25%
3D	HNCO	700 MHz	$^{15}\text{N}^{13}\text{C}$	32		0,2	25%
3D	(H)CC(CO)NH	600 MHz	$^{15}\text{N}^{13}\text{C}$	16		1	25%
3D	H(C)CH tocsy	800 MHz	$^{15}\text{N}^{13}\text{C}$	8		1	25%
3D	1H-13C NOESY	800 MHz	$^{15}\text{N}^{13}\text{C}$	16	120	1,2	
3D	1H-13C TOCSY-HSQC aromat.	600 MHz	^{15}N	16		1	25%
2D	HBCBCGCDHD	800 MHz	$^{15}\text{N}^{13}\text{C}$	32		1	
2D	HBCBCGCDCEHE	800 MHz	$^{15}\text{N}^{13}\text{C}$	32		1	
2D	1H-15N HSQC	700 MHz	$^{15}\text{N}^{13}\text{C}$	32		0,22	
2D	1H-13C HSQC aliphatic	800 MHz	^{15}N	128		1,5	
2D	1H-1H NOESY	600 MHz	^{15}N	64	100 300	1,5	
3D	1H-15N NOESY-HSQC	600 MHz	^{15}N	16	120	1,2	

Appendix A7: Prediction of dynamic transition

Table A8: Meta prediction approach of small protein based on a combination of predictors (PrDOS, DisoPred2, VSL2, IUPred)^[134]. Dynamic transition induced by interactions was commutated with FuzPred with a reference to metaPrDos free form. Small proteins are combined in classes with respect to experimental secondary structure screening analysis. The preferred predicted secondary structural elements are highlighted in the table.

NMR analysis	ID	aa	Free form		Bound form	
			structured	disordered	structured	disordered
folded	P19	59	78	22	84,7	15,3
	P17	60	75	25	93,3	6,7
molten globule	P14	31	71	29	96,8	3,2
	P11	38	81,6	18,4	97,4	2,6
	P16	43	62,8	37,2	74,4	25,6
	P18	61	68,9	31,1	75,4	24,6
unstructured	P1	14	64,3	35,7	92,9	7,1
	P20	17	71,4	28,6	92,9	7,1
	P3	18	0	100	5,6	94,4
	P2	23	8,7	91,3	69,6	30,4
	P12	23	0	100	30,4	69,6
	P5	27	0	100	33,3	66,7
	P4	28	64,3	35,7	100	0
degradation	P6	29	9,4	90,6	93,1	6,9
	P15	51	41,2	58,8	98	2
	P7	53	88,5	11,5	88,5	11,5
	P9	70	77,1	22,9	91,4	8,6
not expressed	P10	71	80,3	19,7	91,5	8,5
	P13	18	83,3	16,7	100	0
	P8	61	90,2	9,8	100	0

Appendix A8: NMR secondary structure analysis of small proteins

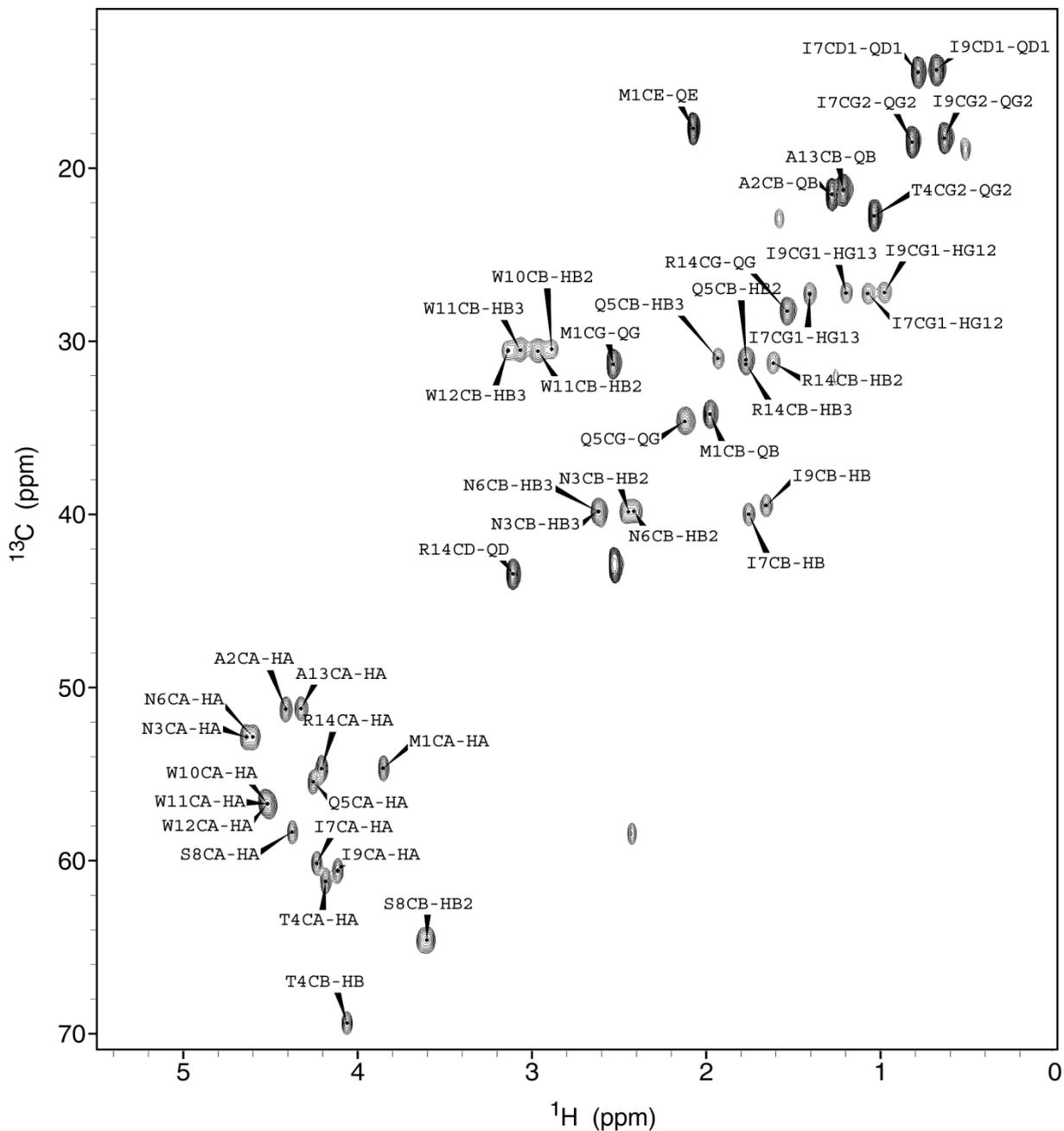


Figure A4. Natural abundance 2D ^1H ^{13}C HSQC spectrum of P20 protein with assignment as indicated in the picture. The experiment was recorded at 600 MHz, 298 K; it contains 5 mM peptide in DMSO and 0.5 mM DSS.

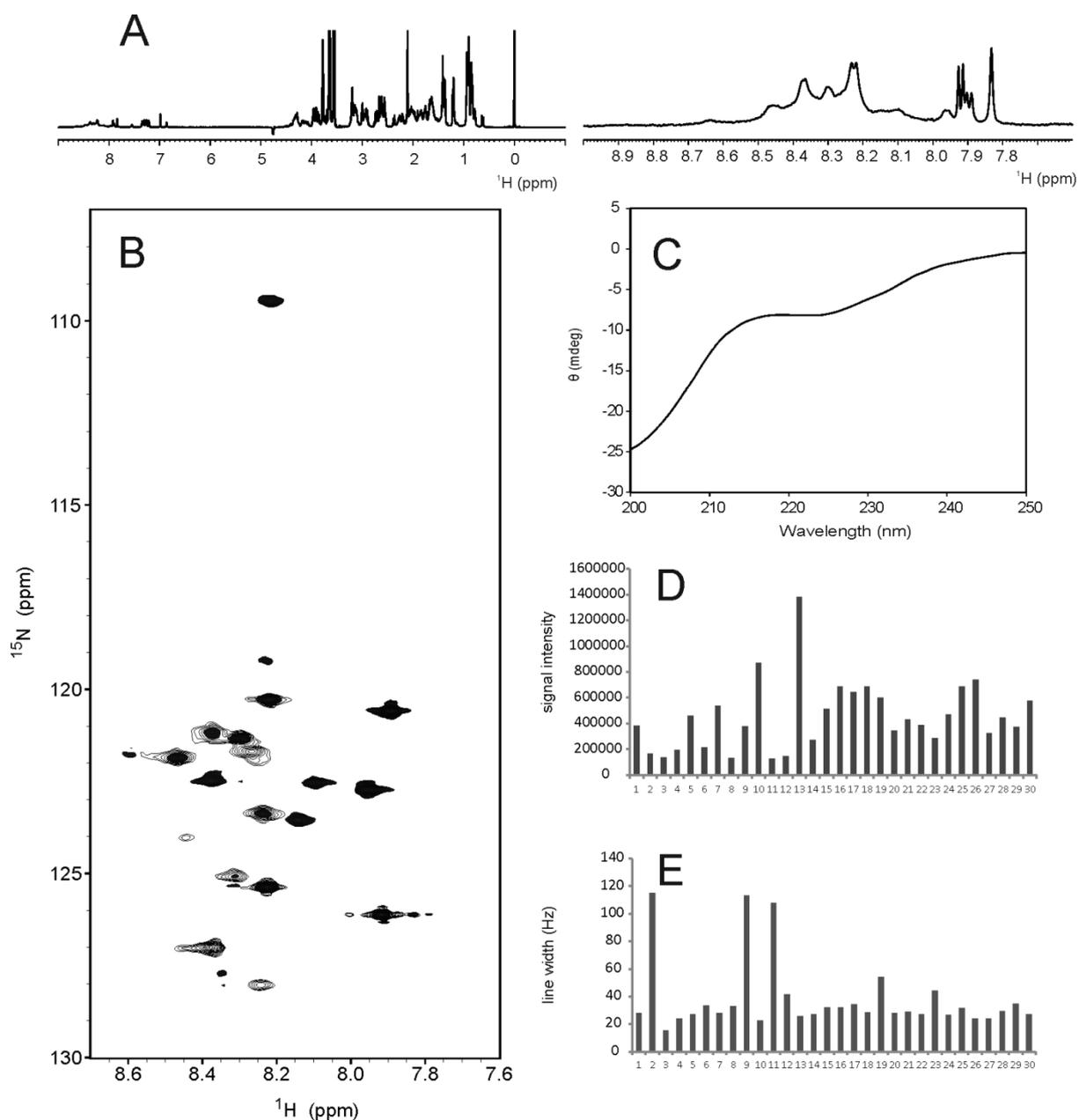


Figure A5. Secondary structure analysis of P16 small protein. **A** - 1D ^1H -NMR spectra with ^{15}N decoupling with amide proton region enlarged recorded at 600 MHz, 298 K. Sample contains 50 mM sodium phosphate buffer pH=7, 300 mM NaCl, 3 mM DTT, 5% D $_2$ O and 0.5 mM DSS. **B** - 2D $^1\text{H}^{15}\text{N}$ HSQC NMR spectrum, recorded at 600 MHz, 298 K and 32 number of scans. **C** - CD-spectrum of 100 μM peptide in 25 mM sodium phosphate buffer pH 7. **D, E** – graphical representation of signal intensity and line width, respectively, from 2D $^1\text{H}^{15}\text{N}$ HSQC.

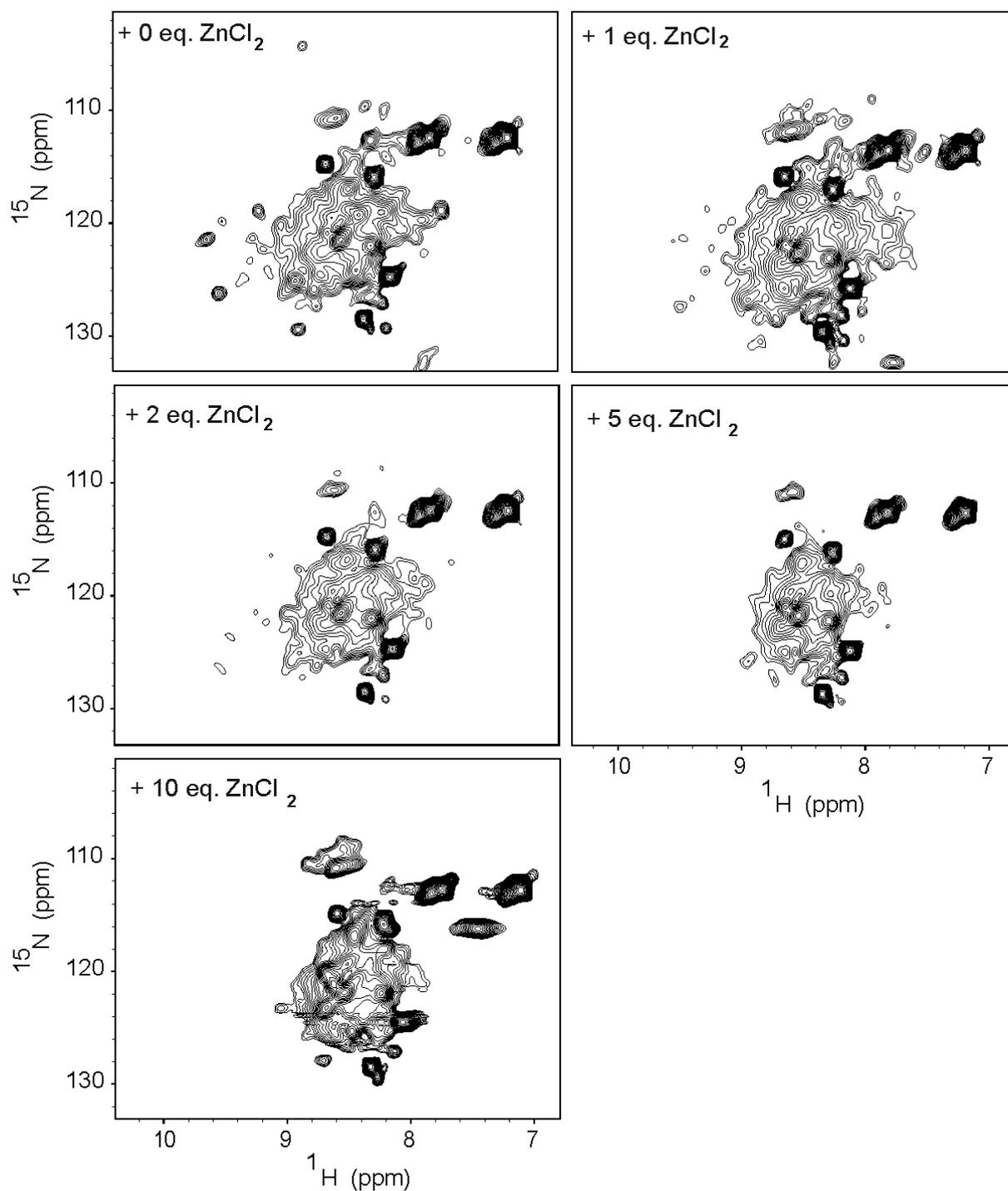


Figure A6. Zinc titration series of the P11 small protein. 2D ^1H - ^{15}N HSQC NMR spectra were recorded at 600 MHz, 298 K. the sample contains 1mM; 50mM NaPi pH 7.5, 3 mM DTT, 5% D₂O, 0.5 mM DSS.

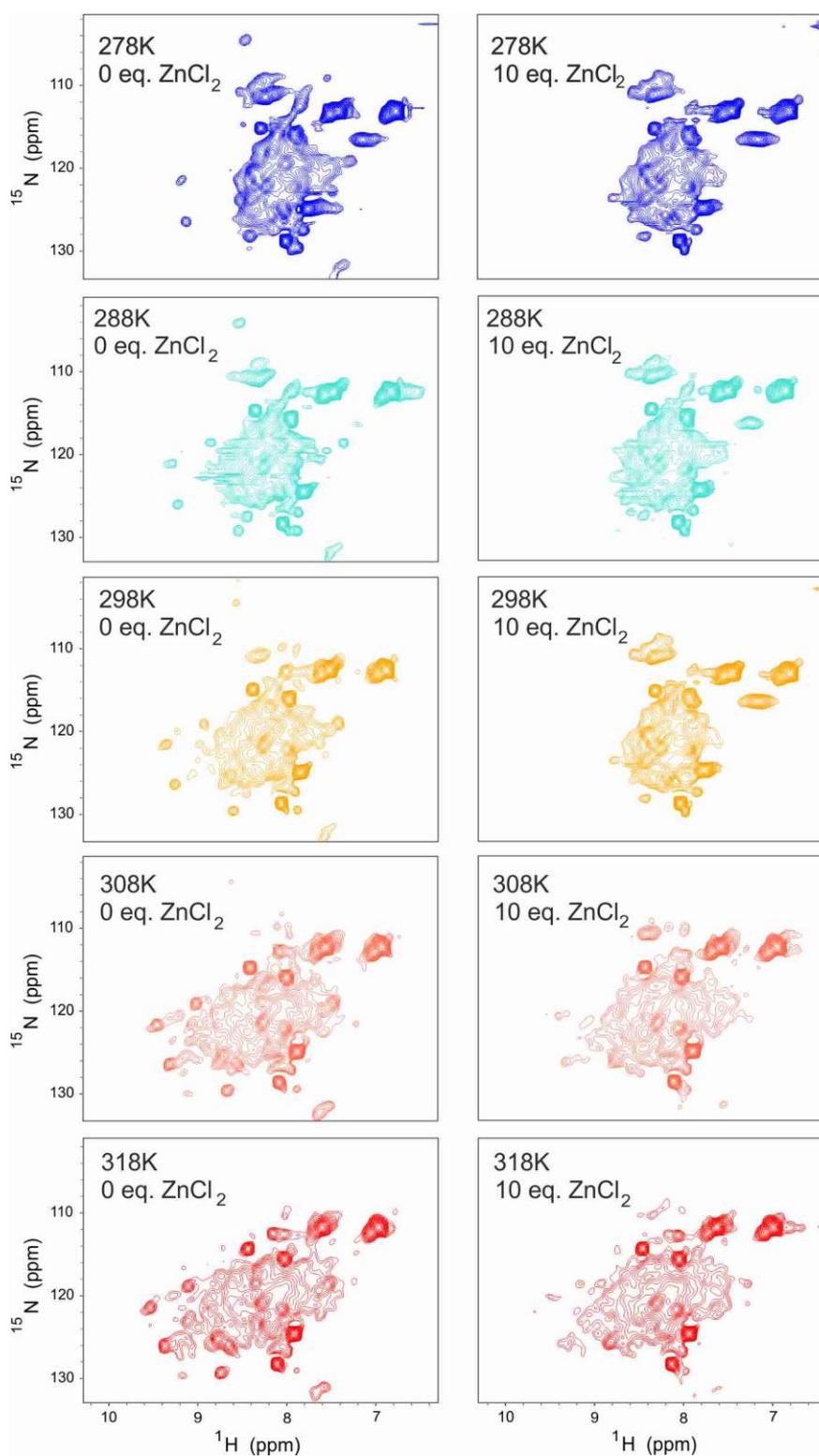


Figure A7. Temperature series of the P11 small protein. 2D $^1\text{H}^{15}\text{N}$ HSQC NMR spectra were recorded at 600 MHz, 298 K without additional zinc ions and with 10 equivalents of zinc chloride. The sample contains 1 mM; 50 mM NaPi pH 7.5, 3 mM DTT, 5% D₂O, 0.5 mM DSS.

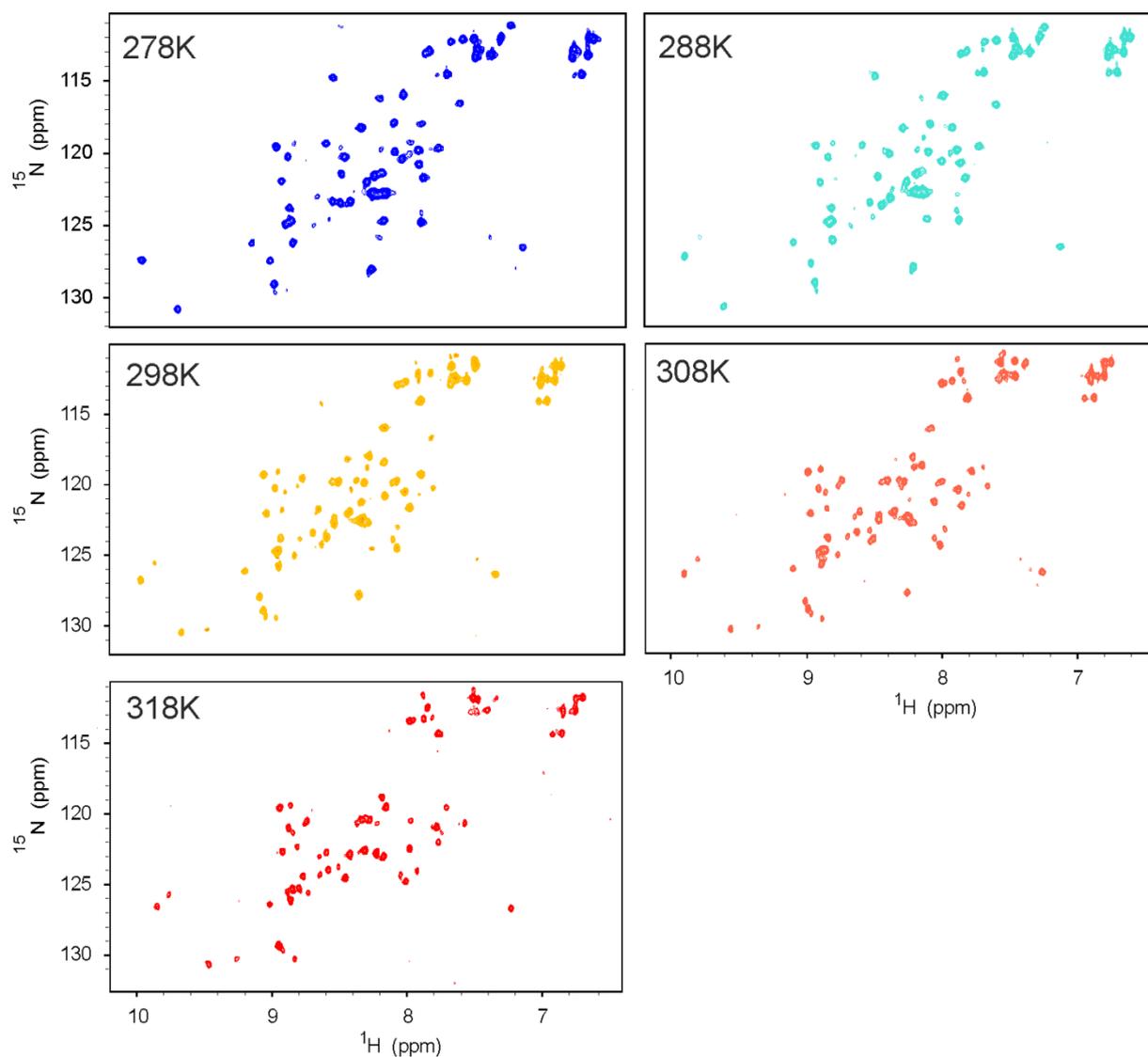


Figure A8. Temperature series of the P19 small protein. 2D ^1H - ^{15}N HSQC spectrum recorded at 800 MHz. The sample contains 25 mM Tris buffer pH 8, 200 mM NaCl, 100 μM ZnCl_2 , 5% D_2O , 0.5 mM DSS.

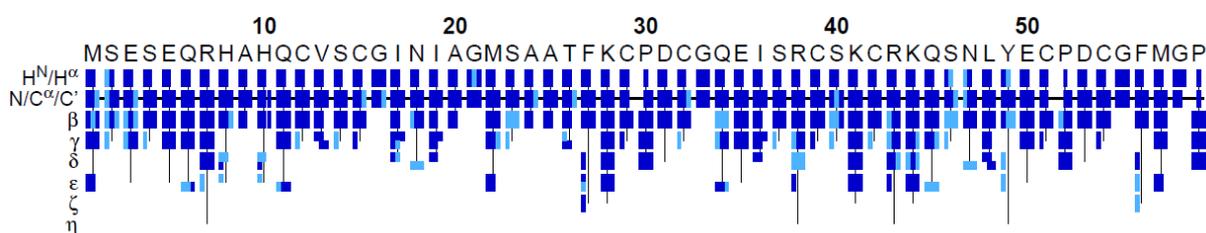
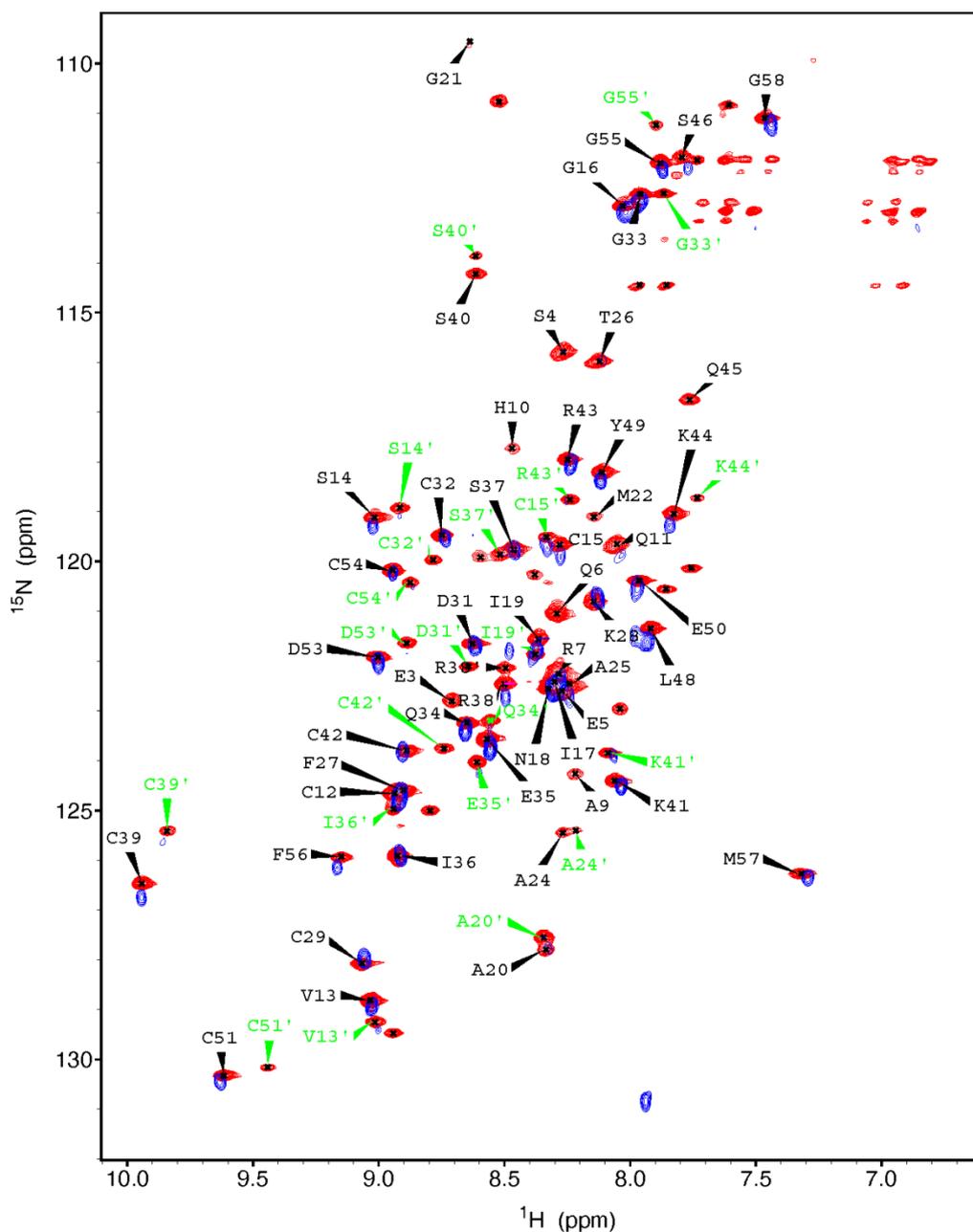


Figure A9: Automated FLYA assignment of P19 small protein. Following NMR experiments were used for the automated assignment: ¹⁵N-HSQC, ¹³C-HSQC, HNC(O), CBCANH, CBCA(CO)NH, HNCA, HN(CA)CO, HNHA, HCCH-TOCSY, ¹⁵N-TOCSY, ¹⁵N-NOESY-HSQC, ¹³C-NOESY-HSQC. The assignment, which is not determined to be strong, is highlighted in light blue.



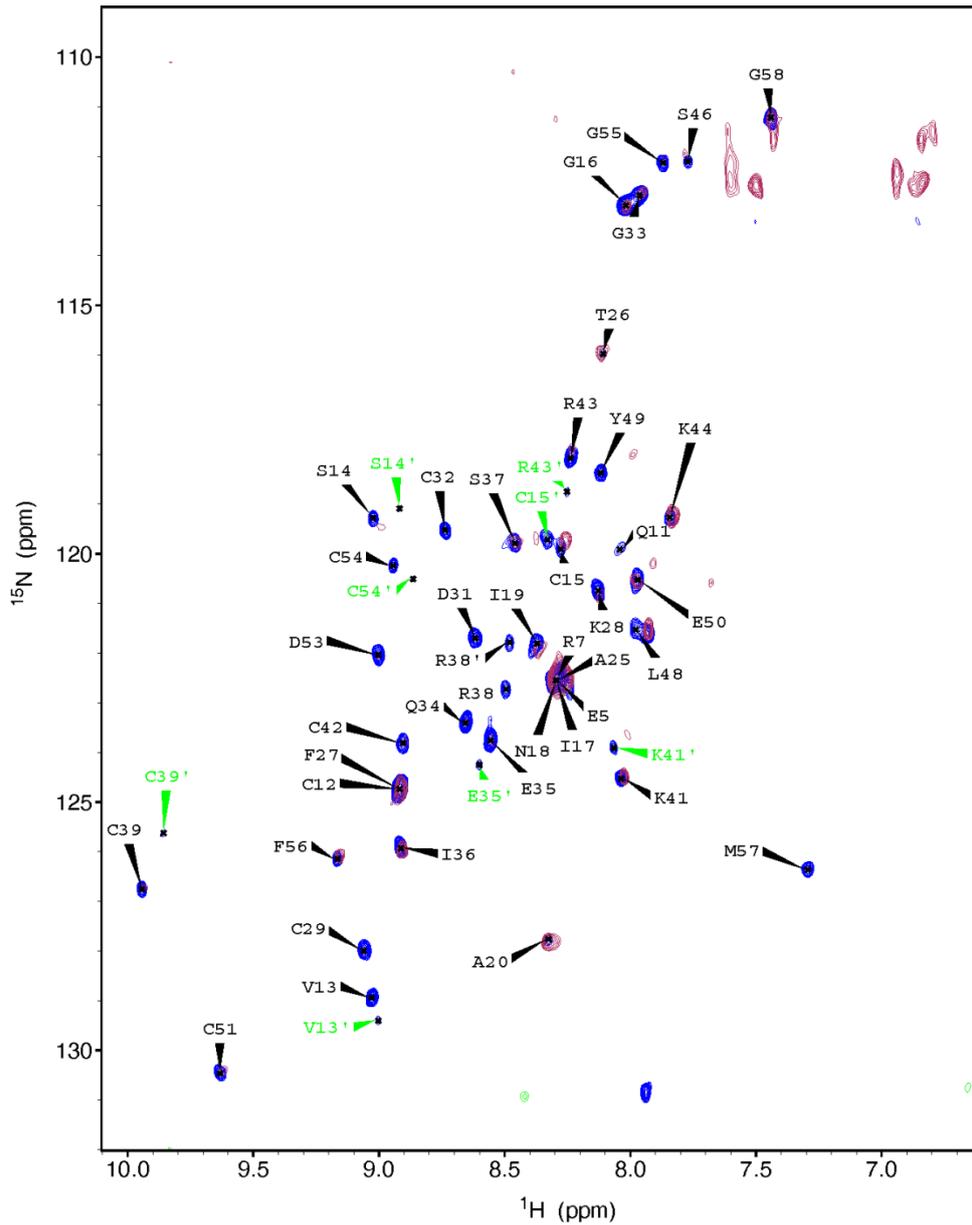
Expressed in M9, 13C15N

950MHz, 298K
 Best troy
 8 scans
 Resolution 10Hz
 2mM, 10mMBisTris pH 7, 200mM NaCl, 100uM
 ZnCl2, 3mM DTT, 5% D2O, 0,5mMDSS

Expressed in LB, natural abundance

700MHz, 298K
 Best troy
 1024 scans
 Resolution 39Hz
 2,5mM, 25mMBisTris pH 7, 200mM NaCl, 100uM
 ZnCl2, 3mM DTT, 5% D2O, 0,5mMDSS

Figure A10. 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum of the P19 small protein expressed in M9 and LB medium.



Expressed in LB, natural abundance
 700MHz, 298K
 Best trosy
 1024 scans
 Resolution 39Hz
 2,5mM, 25mMBisTris pH 7, 200mM NaCl, 100uM
 ZnCl2, 3mM DTT, 5% D2O, 0,5mMDSS

Native protein from HVO, AKSoppa, N-His,
 natural abundance
 600MHz, 298K
 Sofast HMQC
 16k scans
 Resolution 42Hz
 0,33mM, 25mMBisTris pH 7, 200mM NaCl, 100uM
 ZnCl2, 3mM DTT, 5% D2O, 0,5mMDSS

Figure A11. 2D ^1H - ^{15}N HSQC spectrum of the P19 small protein expressed in *E. Coli* and *H.volcanii*.

Appendix A8: NMR secondary structure analysis of small proteins

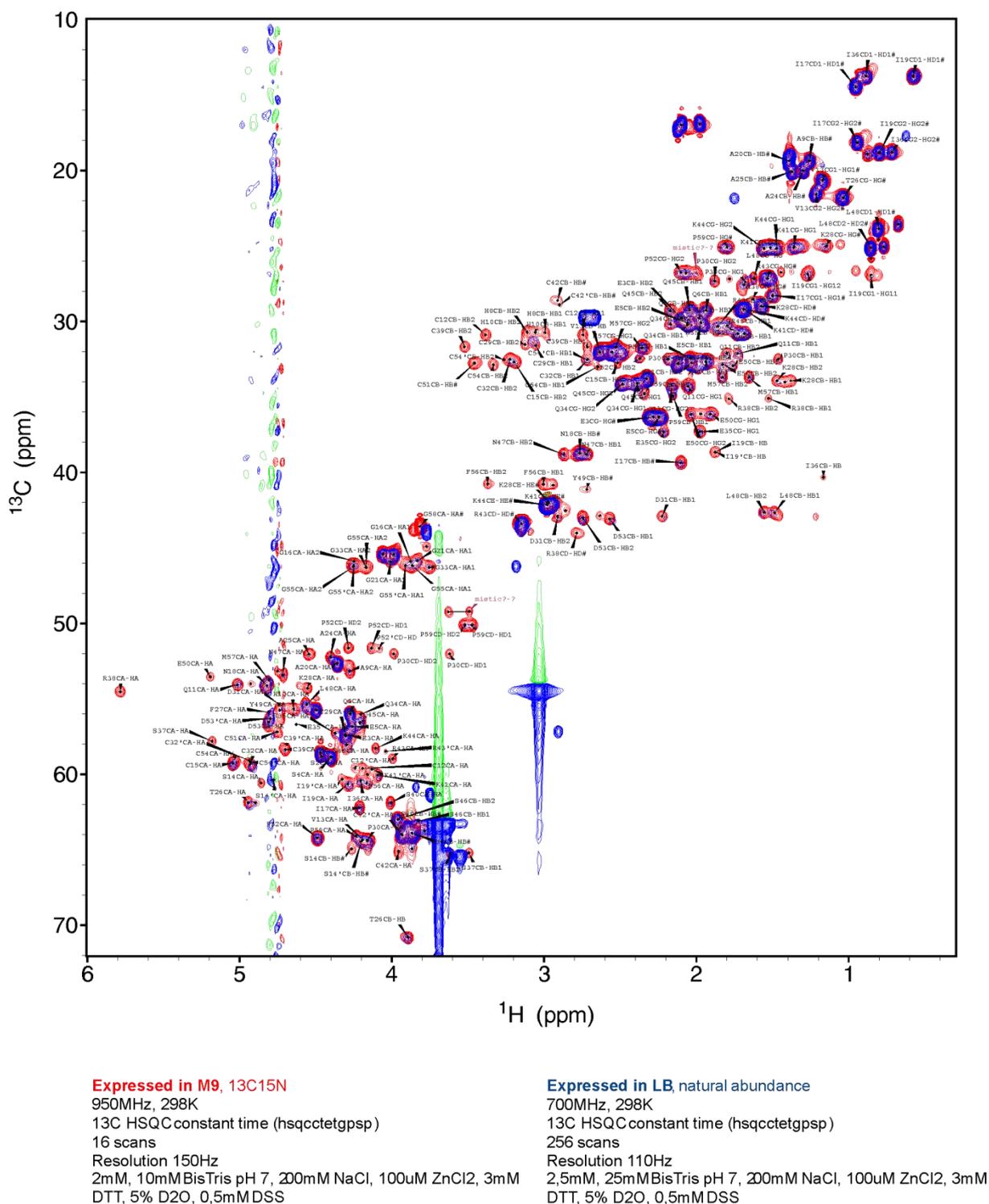
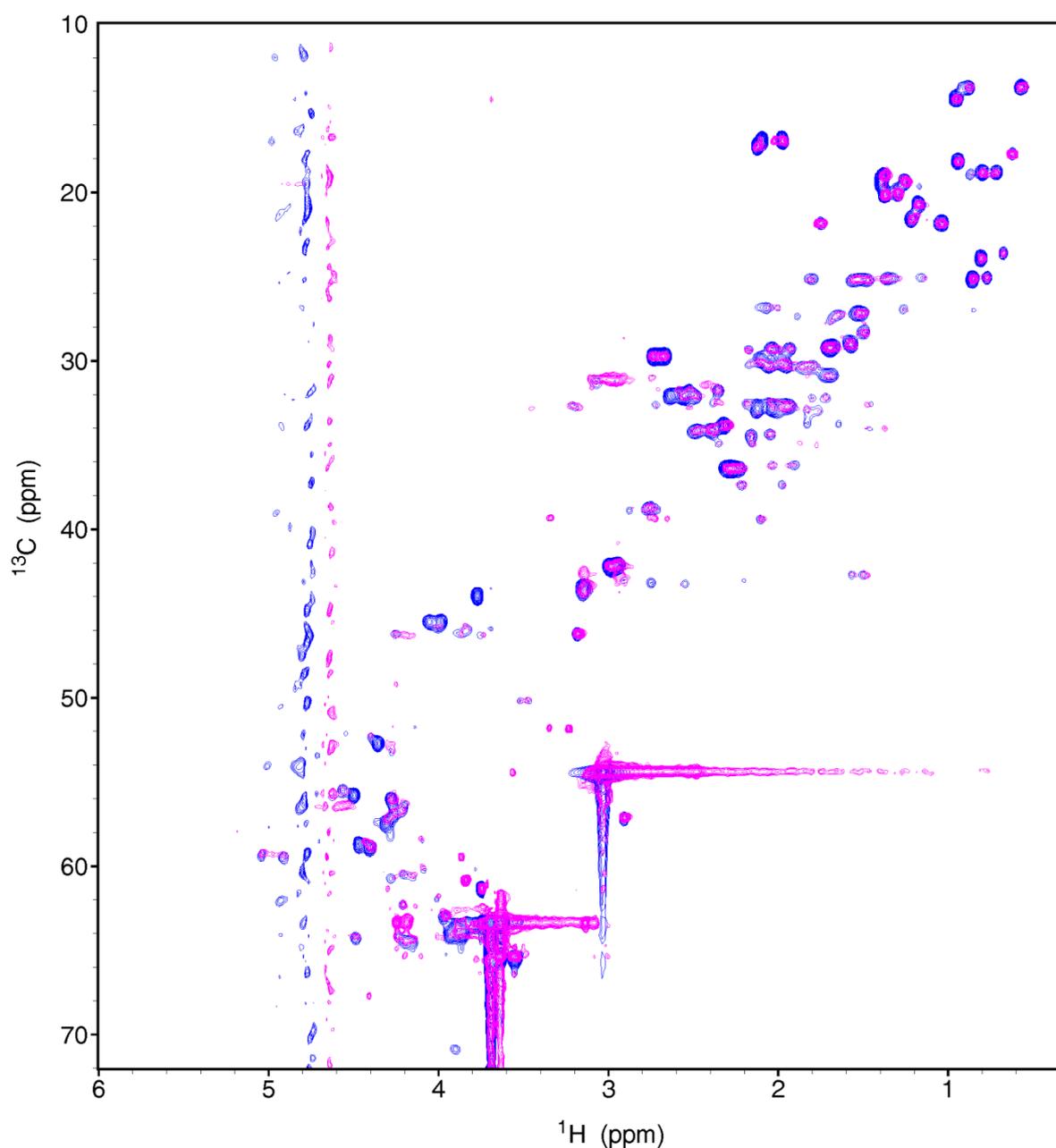


Figure A12. 2D ^1H - ^{13}C HSQC spectrum of the P19 small protein expressed in M9 and LB medium. Signals at 3.7 ppm (^1H) / 64 ppm (^{13}C) and 3.0 ppm (^1H) / 55 ppm (^{13}C) belong to glycerol.



Expressed in LB, natural abundance
 700MHz, 298K
 13C HSQC constant time (hsqcctetgppsp)
 256 scans
 Resolution 110Hz
 2,5mM, 25mMBisTris pH 7, 200mM NaCl, 100uM ZnCl₂, 3mM
 DTT, 5% D₂O, 0,5mMDSS

Native protein from HVO, AKSoppa, N – His, natural
 abundance
 600MHz, 298K
 13C HSQC (hsqcetgppsp)
 1024 scans
 Resolution 94Hz
 0,33mM, 25mMBisTris pH 7, 200mM NaCl, 100uM ZnCl₂,
 3mM DTT, 5% D₂O, 0,5mMDSS

Figure A13: 2D ¹H¹³C HSQC spectrum of the P19 small protein expressed in *E. Coli* and *H.volcanii*. Signals at 3.7 ppm (¹H) / 64 ppm (¹³C) and 3.0 ppm (¹H) / 55 ppm (¹³C) belong to glycerol.

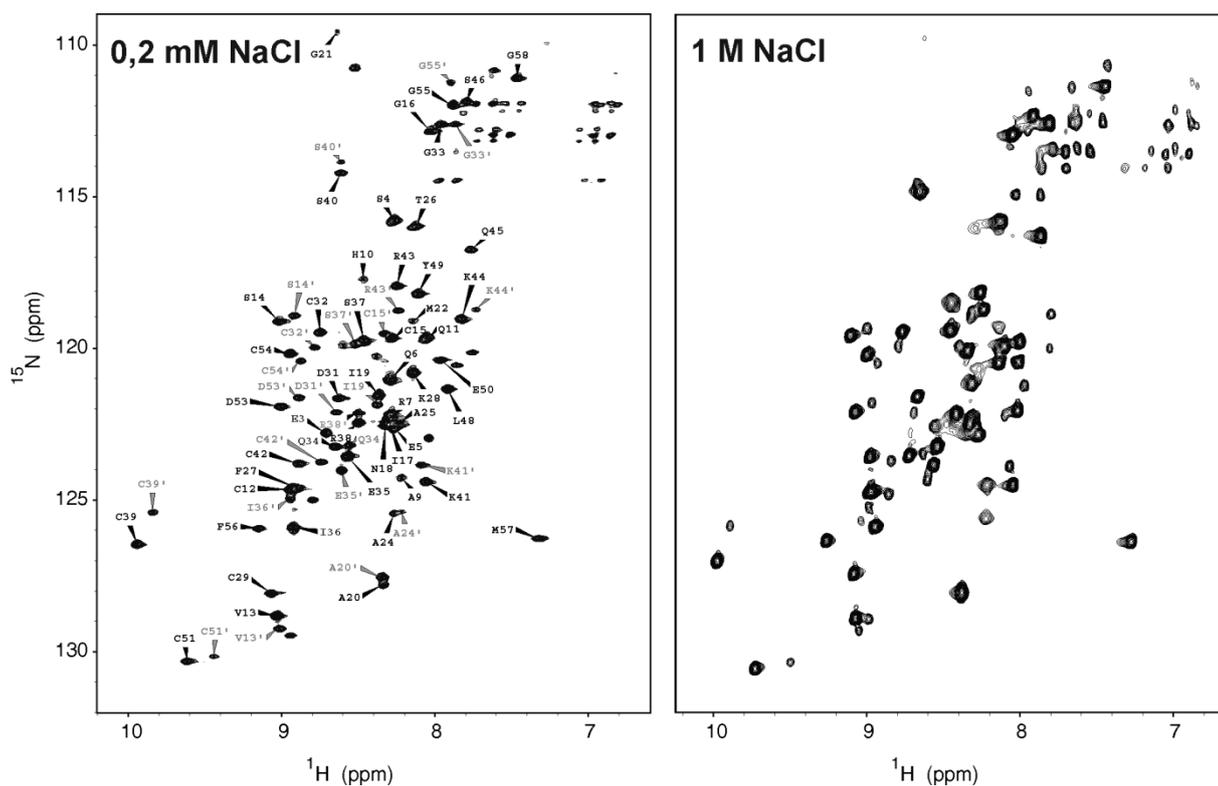


Figure A14. 2D $^1\text{H}^{15}\text{N}$ HSQC spectra of P19 small protein recorded with 0.2 mM and 1 M NaCl. The sample contains 25 mM Tris buffer pH 8, 100 μM ZnCl_2 , 5% D_2O , 0.5 mM DSS; recorded at 800 MHz, 298K.

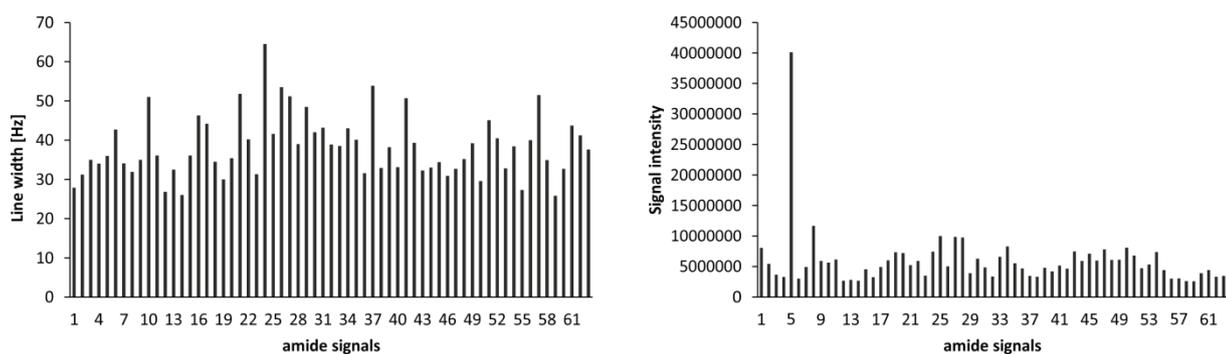


Figure A15. Signal shape analysis of P18 protein. Line width (left) and signal intensity (right) of amide signals extracted from 2D $^1\text{H}^{15}\text{N}$ HSQC spectrum recorded at 600 MHz, 298 K; it contains 1 mM protein, 50 mM sodium phosphate buffer pH 7, 100 mM NaCl, 3 mM DTT, 5% D_2O , 0.5 mM DSS.

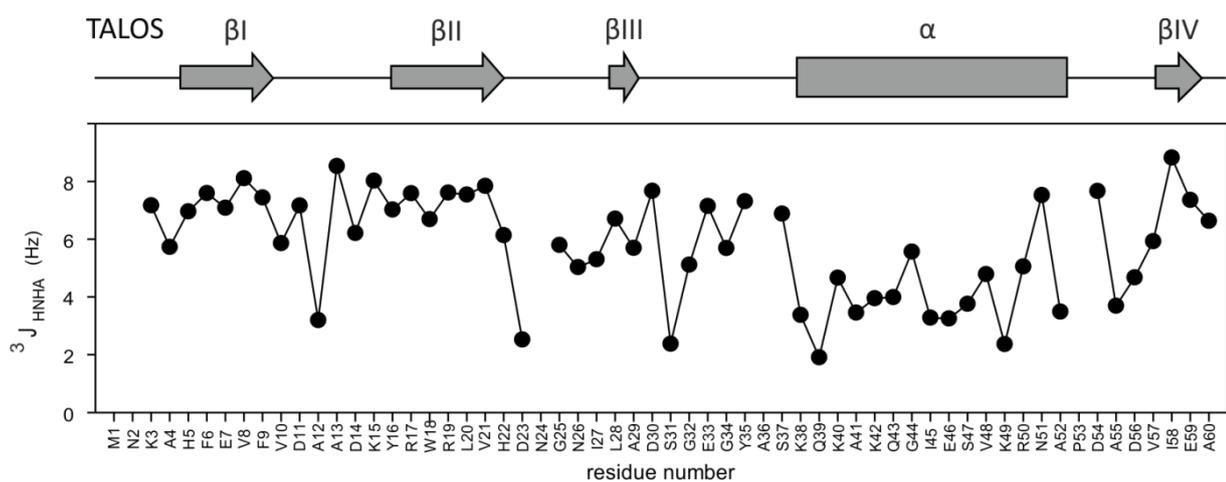


Figure A17. $^3J_{\text{HNHA}}$ coupling constants of P17 small protein. J constants were determined from a 3D HNHA experiment plotted as a function of residue number. On the top the TALOS-N secondary structure prediction is shown.

Appendix A8: NMR secondary structure analysis of small proteins

Table A9: Comparison of P17 to pdb structures with the homology part to P17 protein delivered by sequence-based SWISS model secondary structure prediction ^[125–129] of P17 with respect to identity, similarity and global coverage of their amino acids.

pdb	Description	Organism	Identity	Seq similarity	Coverage	Method	Oligo states	Function	
2k7i	UPF0339 protein Atu0232	<i>Agrobacterium fabrum</i>	43,4	0,42	0,93	NMR	dimer	*	
3bid	UPF0339 protein NMB1088	<i>Neisseria meningitidis</i>	40,7	0,41	0,90	X-ray	dimer	*	
2k8e	UPF0339 protein yegP	<i>Escherichia coli</i>	32,1	0,36	0,93	NMR	monomer	*	
2k49	UPF0339 protein SO_3888	<i>Shewanella oneidensis</i>	28,0	0,34	0,95	NMR	monomer	*	
2dhi	PH domain of Evectin-2	<i>Mus musculus</i>	10,8	0,25	0,62	NMR	monomer	*	
2okq	ybaA	<i>Shigella flexneri</i>	14,3	0,28	0,58	X-ray	dimer	*	
4ei0	DUF4466 family protein	<i>Parabacteroides merdae</i>	15,6	0,29	0,53	X-ray	dimer	*	
5id2	AhpE	<i>Mycobacterium tuberculosis</i>	9,1	0,24	0,55	X-ray	dimer	**	[142]
3drn	Bcp1	<i>Sulfolobus solfataricus</i>	9,3	0,25	0,45	X-ray	monomer	**	[143]
2rp4	Dmp53	<i>Drosophila melanogaster</i>	24,1	0,29	0,44	NMR	tetramer	***	[149]
1j7g	D-tyrosyl-tRNA(Tyr) deacylase	<i>Haemophilus influenzae</i>	25,9	0,34	0,45	X-ray	dimer	****	[144]
2dbo	D-tyrosyl-tRNA(Tyr) deacylase	<i>Aquifex aeolicus</i>	25,9	0,35	0,45	X-ray	dimer	****	
1jke	D-Tyr-tRNA(Tyr) deacylase	<i>Escherichia coli</i>	25,9	0,35	0,45	X-ray	dimer	****	[145]
3knp	D-tyrosyl-tRNA(Tyr) deacylase	<i>Plasmodium falciparum</i>	11,1	0,25	0,45	X-ray	dimer	****	[146]
4nbj	D-aminoacyl-tRNA deacylase (DTD)	<i>Plasmodium falciparum</i>	11,1	0,25	0,45	X-ray	dimer	****	[147]
5j61	D-tyrosyl-tRNA(Tyr) deacylase	<i>Plasmodium falciparum</i>	11,1	0,25	0,45	X-ray	dimer	****	[148]
4nbi	D-aminoacyl-tRNA deacylase (DTD)	<i>Plasmodium falciparum</i>	11,1	0,25	0,45	X-ray	dimer	****	[147]
3ko4	D-Tyr-tRNA(Tyr) deacylase	<i>Plasmodium falciparum</i>	11,1	0,25	0,45	X-ray	dimer	****	[146]
3ko3	D-tyrosyl-tRNA(Tyr) deacylase	<i>Plasmodium falciparum</i>	11,1	0,25	0,45	X-ray	dimer	****	[146]

* - structural genomics unknown function

** - oxidoreductase

*** - transcription

**** - hydrolase

Appendix A9: Rhodopsin sequences

pACMV-tetO : rhodopsin wild type ^[177]

10 20 30 40 50 60
MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLI~~M~~L GFPINFL~~T~~LY
70 80 90 100 110 120
VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH GYFVFGPTGC NLEGGFFATLG
130 140 150 160 170 180
GEIALW~~S~~LVV LAIERYVVVC KPMSNFRFGE NHAIMGVAFT **C**VMALACAAP PLVGWSRYIP
190 200 210 220 230 240
EGMQCSCGID YYTPHEETNN ESFVIYMFVV HFIIPLIVIF FCYQQLVFTV KEAAAQQQES
250 260 270 280 290 300
ATTQKAEKEV TRMVIIMVIA FLICWLPYAG VAFYIFTHQG **S**DFGPIFMTI PAFFAKTSAV
310 320 330 340
YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQLAPA

pcDNA3.1: stabilized (N2C/D282C) wild type (pdb: 5TE3) ^[211]

10 20 30 40 50 60
MCGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLI~~M~~L GFPINFL~~T~~LY
70 80 90 100 110 120
VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH GYFVFGPTGC NLEGGFFATLG
130 140 150 160 170 180
GEIALW~~S~~LVV LAIERYVVVC KPMSNFRFGE NHAIMGVAFT WVMALACAAP PLVGWSRYIP
190 200 210 220 230 240
EGMQCSCGID YYTPHEETNN ESFVIYMFVV HFIIPLIVIF FCYQQLVFTV KEAAAQQQES
250 260 270 280 290 300
ATTQKAEKEV TRMVIIMVIA FLICWLPYAG VAFYIFTHQG **S**CFGPIFMTI PAFFAKTSAV
310 320 330 340
YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA

pcDNA3.1: stabilized (N2C/D282C) G90D mutant (pdb: 4BEZ) ^[209]

10 20 30 40 50 60
MCGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLI~~M~~L GFPINFL~~T~~LY
70 80 90 100 110 120
VTVQHKKLRT PLNYILLNLA VADLFMVFGD FTTTLYTSLH GYFVFGPTGC NLEGGFFATLG
130 140 150 160 170 180
GEIALW~~S~~LVV LAIERYVVVC KPMSNFRFGE NHAIMGVAFT WVMALACAAP PLVGWSRYIP
190 200 210 220 230 240
EGMQCSCGID YYTPHEETNN ESFVIYMFVV HFIIPLIVIF FCYQQLVFTV KEAAAQQQES
250 260 270 280 290 300
ATTQKAEKEV TRMVIIMVIA FLICWLPYAG VAFYIFTHQG **S**CFGPIFMTI PAFFAKTSAV
310 320 330 340
YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA

Appendix A10: Cell culture media and buffers

Media used in cell culture

medium A

- DMEM
- 10% FBS
- 5 ml pen/strep (100 x stock)
- 5 µg/ml Blasticidic
- 200 µg/ml Geneticin
- 200 µg/ml Geneticin

medium B

- labeled individual component medium
- 10% FBS
- 5 ml pen/strep (100 x stock)
- 5 µg/ml Blasticidic
- 200 µg/ml Geneticin
- 2 µg/m tetracycline
- 5 mM sodium buturate

medium C

- DMEM
- 10% FBS
- 5 ml pen/strep (100 x stock)
- 5 µg/ml Blasticidic
- 200 µg/ml Geneticin
- 2 µg/m tetracycline
- 5 mM sodium buturate

Composition of cell culture medium for selective labeling schemes

Compounds	mg/l	Compounds	mg/l
Inorganic salts		Amino Acids	
• CaCl ₂	200	• Arginine x HCl	84
• Fe(NO ₃) ₃ *9H ₂ O	0.1	• Cystine x 2 HCl	64
• KCl	400	• Glycine	30
• MgSO ₄	97.7	• Glutamine	584
• NaH ₂ PO ₄	109	• Histidine x H ₂ O	31.03
• NaCl	6400	• Isoleucine	105
• NaHCO ₃	3700	• Leucine	105
Vitamins		• Lysine	117
• Choline Chloride	4	• Methionine	30
• Folic Acid	4	• Phenylalanine	66
• I-(myo)-Inositol	7.2	• Serine	42
• Niaciamide	4	• Theonine	95
• D-Ca Pantothenate (4°C)	4	• Tyrosine x 2H ₂ O	72.5
• Pyridoxine HCl	4	• Valine	94
• Riboflavin	0.4	• Thryptophan ¹⁵ N	16
• Thiamine HCl	4		
Others compounds			
• Phenol Red	15		
• D-Glucose	4500		

Buffers used for purification of antibody

AB equilibration buffer

- 3 M NaCl (pH 7)
- 1.5 M glycine

AB elution buffer

- 0.1 M citric acid (pH <3)
- 0.15 NaCl

AB pH adjusting buffer

- 0.5 M NaHCO₃ (pH 10-11)

AB dialysis buffer

- 0.1 M NaHCO₃ (pH 8.4)
- 0.5 M NaCl

coupling buffer

- 0.1 M NaHCO₃ (pH 8.4)
- 0.5 M NaCl

bloking buffer

- 1M TRIS (pH 8.4)

AB low pH buffer

- 0.1 M acetic acid (pH <3)

Other buffers

retinal HPLC equilibration buffer.

- 92% hexan and 8% ether

liquid state NMR buffer

- 20 mM phosphate buffer (pH 6.5)
- 10% D2O
- 0.5 mM DSS

DNP juice

- 10% buffer D in H2O
- 60% buffer D in D2O
- 30% d8-glycerol
- 20 mM AMUPOL

buffer D

- 10 mM HEPES
- 50 mM NaCl
- 1 mM EDTA
- 1 mM DTT
- 4% OG
- pH 7.1

buffer E

- 50 mM Tris
- 50 mM sodium acetate
- pH 7

Appendix A11: Rhodopsin purification

Table A10: Retinal Purification details.

Retinal labeling scheme	Starting material	Dissolved in ethanol	Dissolved in hexan/ether	11-cis	Yield
^{12}C	100 mg	50 ml	20 ml	2.1 mg	2.1%
12,13,20- ^{13}C	20 mg	10 ml	6 ml	0.7 mg	3.5%
14,15- ^{13}C	5 mg	5 ml	3 ml	0.6 mg	12%

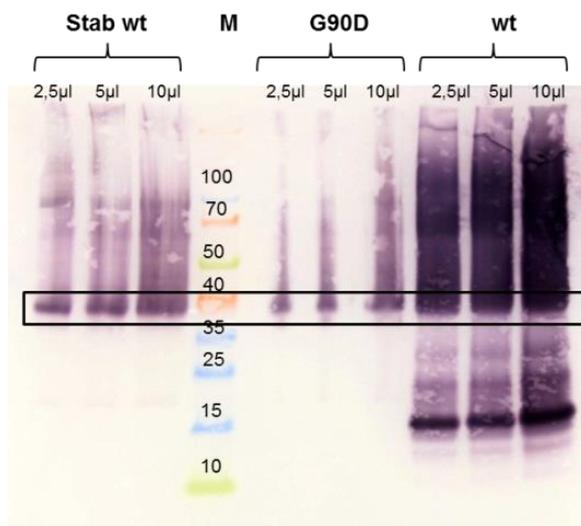


Figure A18. Western blot analysis of test expression of three rhodopsin constructs used in this study. Bands which correspond to the desired rhodopsin molecular weight are highlighted in a black box.

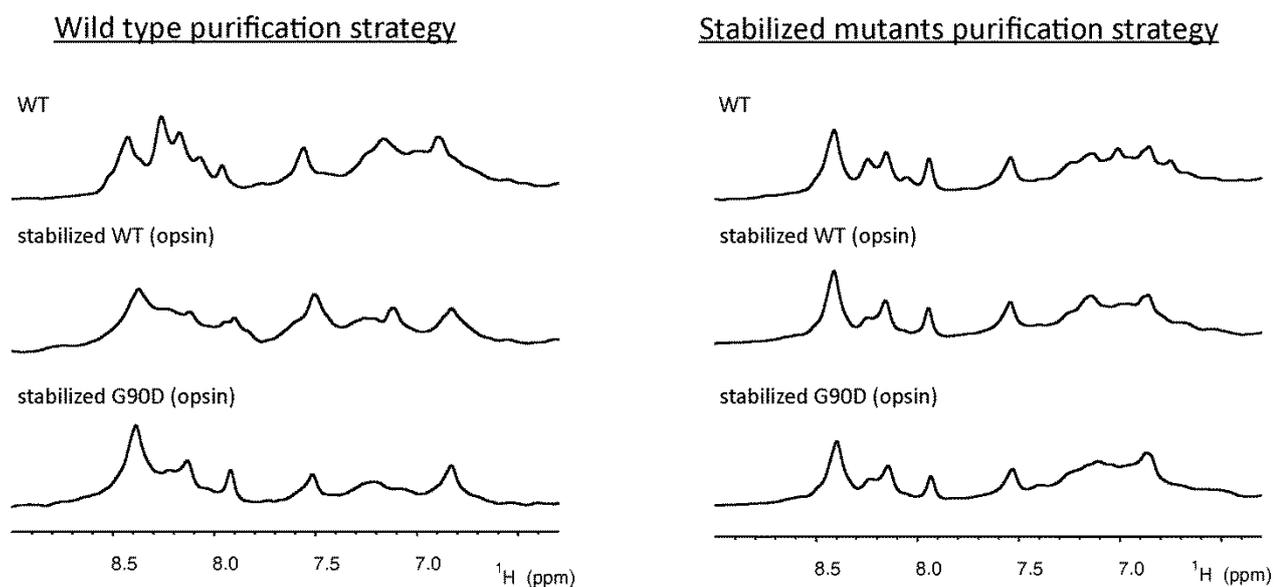


Figure A19. ^1H 1D NMR spectra of amide region of wild type, stabilized wild type and stabilized G90D mutant purified according to the wild type purification protocol (**left panel**) and the optimized protocol (**right panel**).

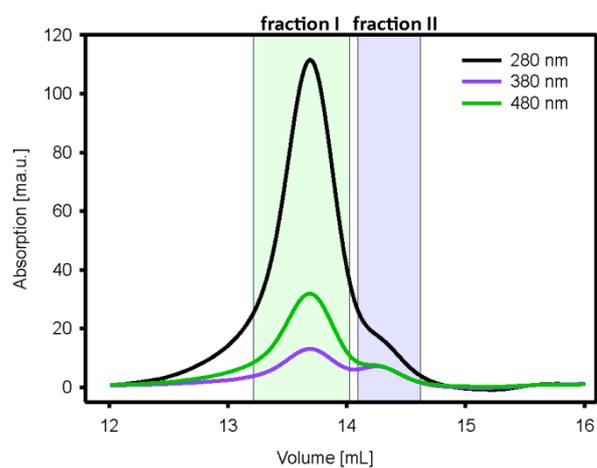


Figure A20. Size-exclusion purification of G90D mutant. Analytical Superdex 200 10/300 size-exclusion chromatogram. G90D mutant was purified in 0.05% DDM at 4°C.

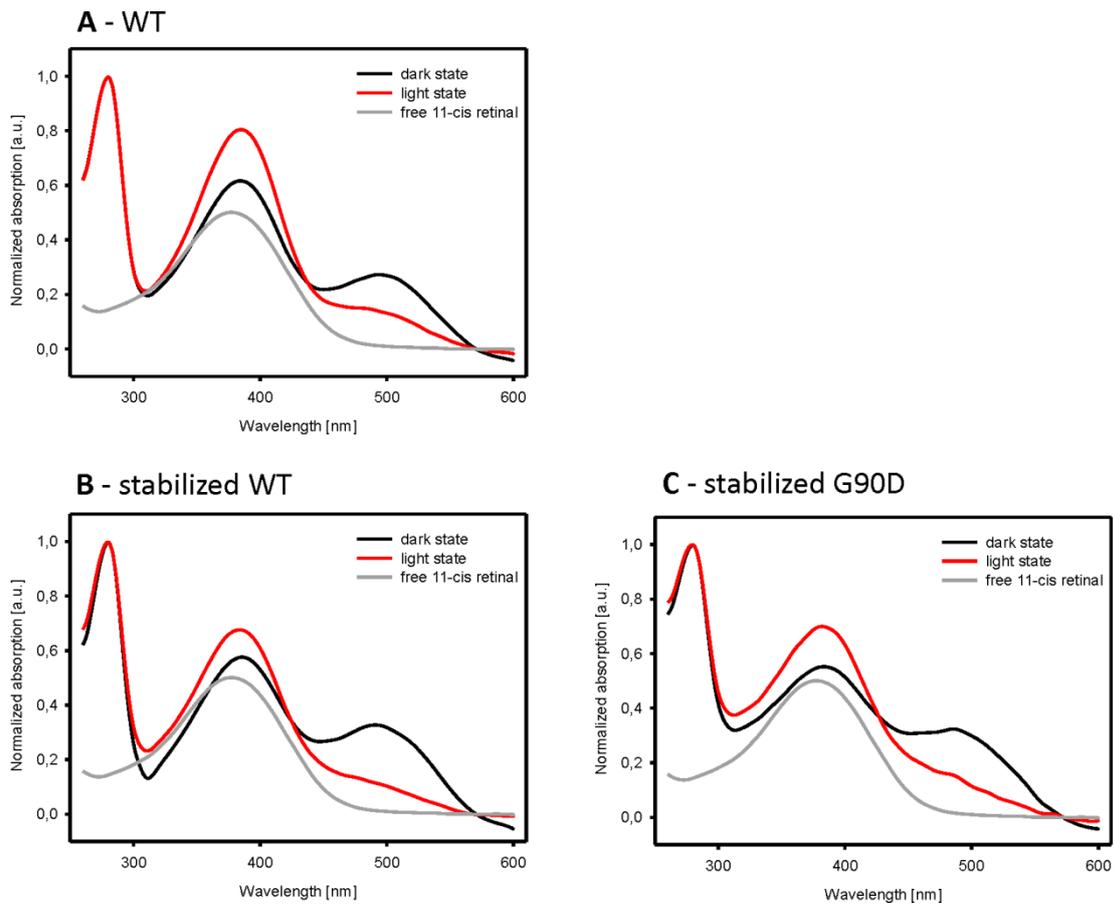


Figure A21. UV/VIS-absorption spectra. Absorption spectra of wild type (A), stabilized wild type (B) and stabilized G90D mutant (C) recorded at the dim light conditions (black) and after illumination (red). The free 11-cis retinal is shown in gray. The absorption spectra were normalized at 280 nm.

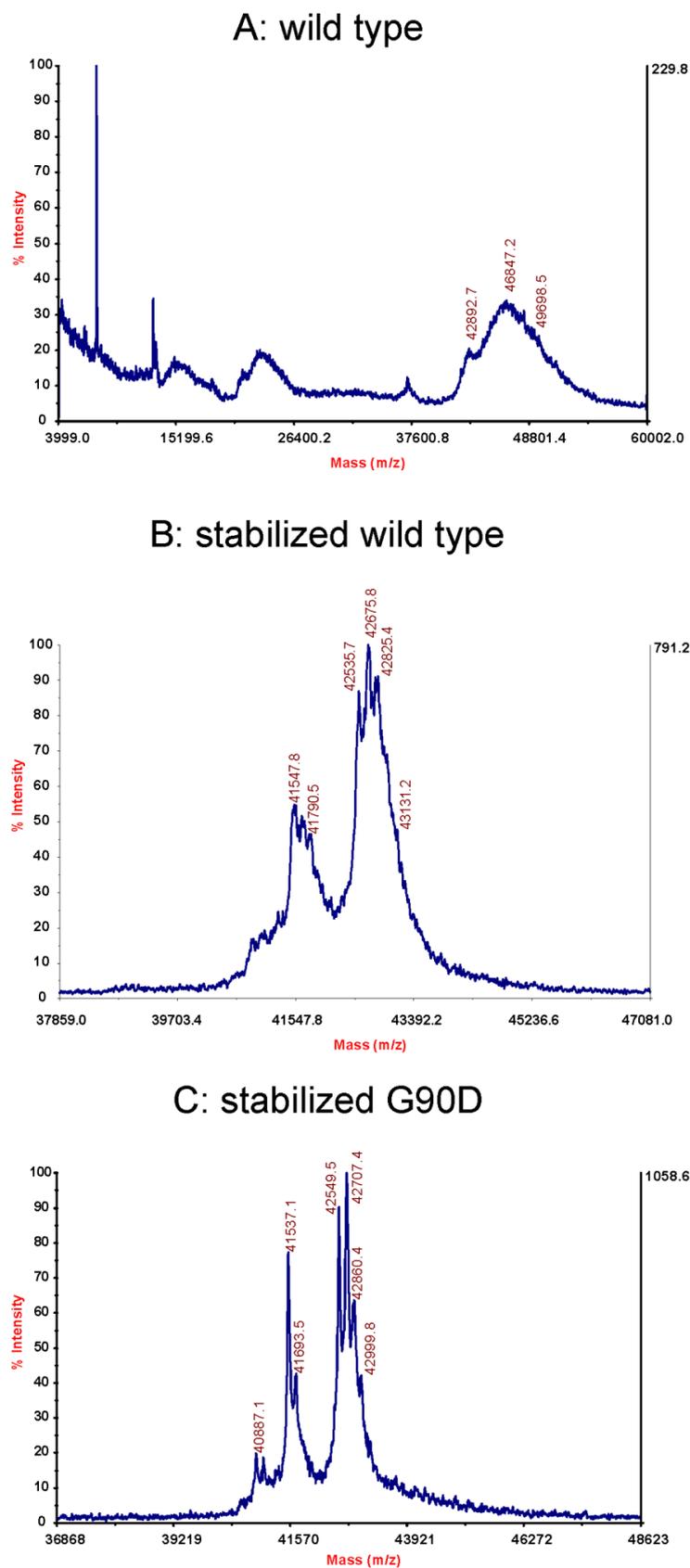


Figure A22. MALDI spectra of three rhodopsin constructs. A - wild type, B - stabilized wild type, C - stabilized G90D mutant. Expected molecular weight is 42 kDa.

Appendix A12: Liquid state NMR experiments

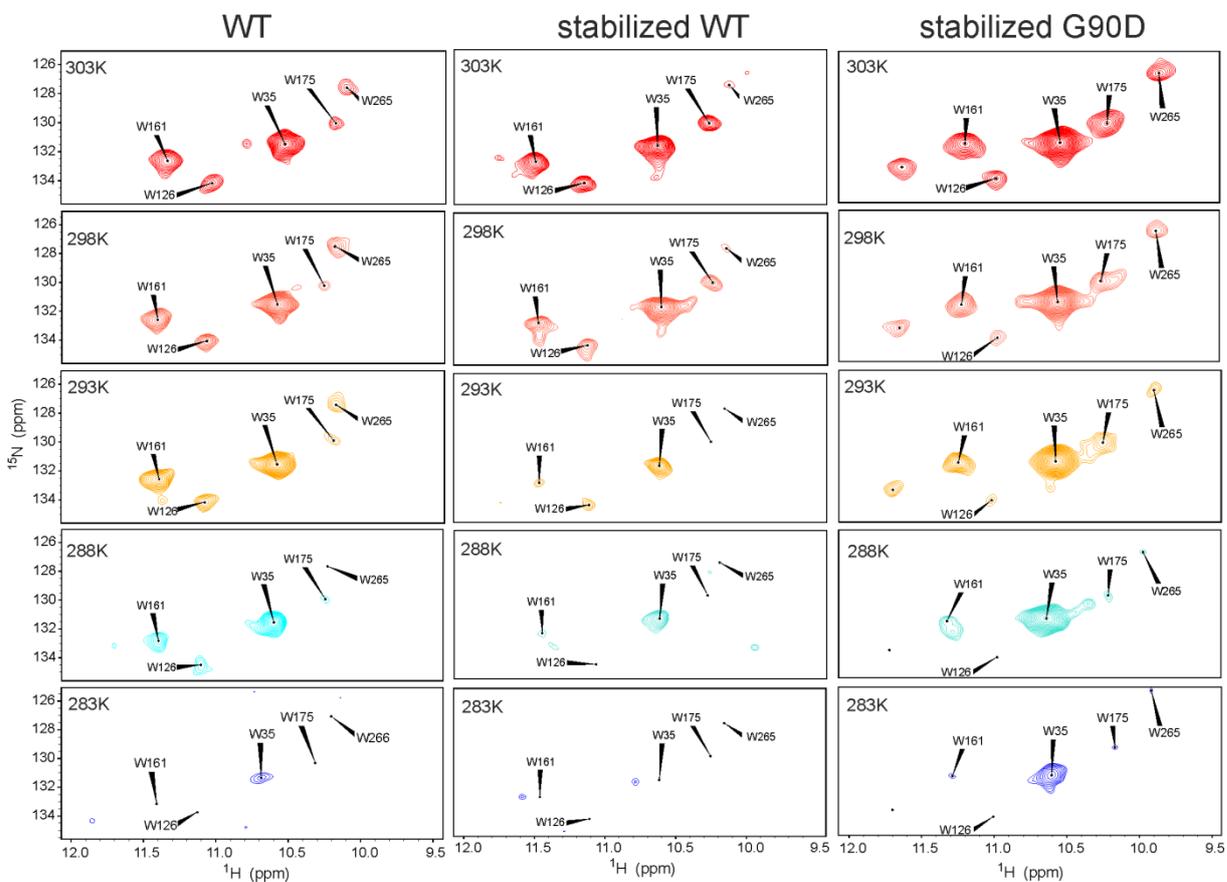


Figure A23. Temperature series of 2D SOFAST-HMQC NMR spectra of rhodopsin constructs in the dark state. All spectra were recorded at 600 MHz, 298 K under light dim conditions.

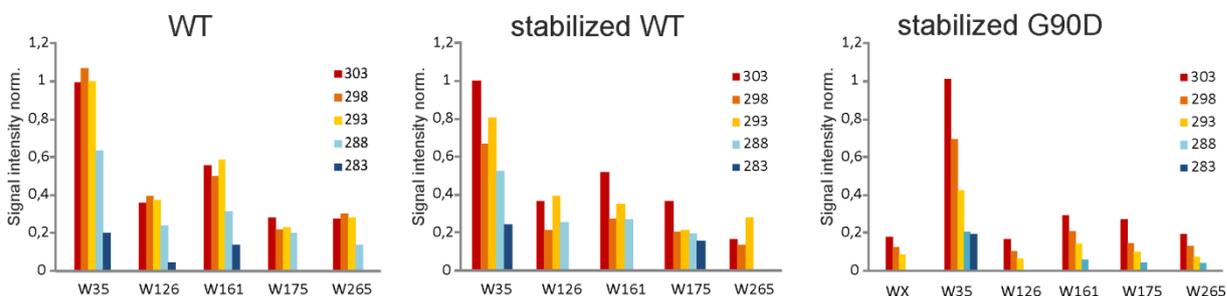


Figure A24. Graphical analysis of temperature series. Signal intensities from 2D SOFAST-HMQC NMR spectra of rhodopsin constructs recorded under light dim conditions were normalized to the highest W35 signal at 303 K.

Acknowledgments

First of all, I would like to thank my Ph.D. supervisor **Prof. Harald Schwalbe** for the opportunity to perform this doctoral thesis in his group, for his continues scientific interest and support in all projects. Also I would like to thank for the inspiring working atmosphere he created in his lab. I sincerely appreciate the ability to participate in different international conferences and workshops, where I could present my work and expand the scientific knowledge.

Further, I would like to thank **Prof. Clemens Glaubitz** for our fruitful collaboration, very helpful discussions about rhodopsin project and for being the second corrector of my thesis. Thank you for your time and for sharing your knowledge.

I want to thank several people for the productive collaborations: **Jiafei Mao** from Prof. Clemens Glaubitz group for the solid state NMR experiments, **Prof. Josef Wachtveitl** and **Elias Eckert** for the rhodopsin kinetic measurements, **Prof. Vladislav Orekhov** and **Maxim Mayzel** (Sweden) for accommodating me in their lab and teaching NUS/TA program, **Prof. Monika Fuxreiter** (Hungary) for very fast and detailed prediction analysis of the small proteins and all cooperation partners from “small proteins” society for providing the protein sequence and fruitful collaboration.

I would especially like to thank **Krishna Saxena** who was involved in all my Ph.D. projects and guided me through the often confusing lab work. Thank you the patience in explaining molecular biology and biochemical stuff to a chemist, for the interest in my work and numerous advices during my whole Ph.D. life.

I highly appreciate all help I got form **Henry Jonker**. Your experience in NMR spectroscopy, protein assignment and structure calculation were crucial for the quality of this work. Thank you for your kindness, patience, support with all computational problems and specially for teaching me the FLYA and CAYNA tricks.

I also want to thank **Christian Richter** for opening the magic of the NMR field for me, **Boris Fürtig** for setting up the laser NMR experiments, our IT-Team **Martin Hähnke** and **Fabian Hiller** for the grateful support with all computer bags. Thank you for your patience in all my NUS troubles and for the introduction to linux.

I address many thanks to **Nusrat Qureshi** and **Anna Niesteruk**. Thank you for brainstorming, for your invaluable support in the lab, office and private stuff throughout all these years.

Acknowledgments

I am sincerely grateful for our secretary team **Kerstin Date** and **Theodora Ruppenthal** for all organization support. I am still impressed with how quickly all problems can be solved by the professionals.

I would like to thank all “**girls room**” members: Nusrat Qureshi, Anna Niesteruk, Deep Chatterjee, Irene Bessi, Florian Lehner, Sven Wahrhaut, Hannah Steinert, Sridhar Sreeramulu, Sam Alshamleh, Nadide Altincekic, Albrecht Völklein, Jasleen Bains und Diana Müller. Thank you for the wonderful atmosphere and patience during my writing phase.

Many thanks to all my coworkers who read and corrected this thesis: Krishna Saxena, Nusrat Qureshi, Anna Niesteruk, Sam Alshamleh, Robbin Schnieders, Dennis Pyper and György Pinter.

Special thanks to **Dennis Pyper** who performed several experiments during his master thesis that contributed to this work.

I would like to thank **Verena Linhard** who helped with the production of peptides and **Deep Chatterjee** who showed me the cell culture lab routine at the very beginning of my Ph.D.

And, of course, many thanks to **all the members of the Schwalbe group**. I will certainly remember our nice lab atmosphere, group trips and balcony discussions.

Finally, I want to thank my family, they made me who I am now, and especially my mom who always believed in me and gave me all possible support in the world.

Publications

N. Kubatova, H.R.A. Jonker, K. Saxena, C. Richter, V. Vogel, S. Schreiber, A. Marchfelder, H. Schwalbe. "Solution structure and dynamics of the small protein HVO_2922 from *Haloferax volcanii*", *ChemBioChem.*, **2019**. doi: 10.1002/cbic.201900085.

N. Kubatova, D. Pyper, H.R.A. Jonker, K. Saxena, L. Rimmel, C. Richter, S. Brantl, E. Evgenieva-Hackenberg, W.R. Hess, G. Klug, A. Marchfelder, J. Soppa, W. Streit, M. Mayzel, V.Y. Orekhov, M. Fuxreiter, R.A. Schmitz, H. Schwalbe. "Rapid biophysical characterization and NMR structure analysis of small proteins from bacteria and archaea", *Angewandte Chemie.*, **2019** submitted.

N. Kubatova, J. Mao, K. Saxena, CE. Eckert, J. Wachtveitl, C. Glaubitz, H. Schwalbe, "Light dynamics of retinal disease relevant G90D bovine rhodopsin mutant", **2019** in preparation.

M. Mehler, CE. Eckert, AJ. Leeder, J. Kaur, T. Fischer, N. Kubatova, LJ. Brown, RCD. Brown, J. Becker-Baldus, J. Wachtveitl, C. Glaubitz. "Chromophore Distortions in Photointermediates of Proteorhodopsin visualized by DNP-enhanced solid-state NMR", *J Am Chem Soc.*, **2017**, 139, 16143-16153.

D. DiGuseppi, B. Milorey, G. Lewis, N. Kubatova, S. Farrell, H. Schwalbe, and R. Schweitzer-Stenner, "Probing the Conformation Dependent Preferential Binding of Ethanol to Cationic Glycylalanyl-glycine in Water/Ethanol by Vibrational and NMR Spectroscopy", *J Phys Chem B.*, **2017**, 121, 5744-5758.

J. Hahn, S. Thalmann, A. Migur, R. Freiherr von Boeselager, N. Kubatova, E. Kubareva, H. Schwalbe, and E. Evgenieva-Hackenberg, "Conserved small mRNA with a unique, extended Shine-Dalgarno sequence," *RNA Biol.*, **2016**, 14, 1353-1363.

J. Grote, D. Krysiak, K. Petersen, S. Güllert, C. Schmeisser, K. U. Förstner, H. B. Krishnan, H. Schwalbe, N. Kubatova, and W. R. Streit, "The Absence of the N-acyl-homoserine-lactone Autoinducer Synthase Genes *tral* and *ngri* Increases the Copy Number of the Symbiotic Plasmid in *Sinorhizobium fredii* NGR234," *Front. Microbiol.*, **2016**. 7: 1858.

S. E. Toal, N. Kubatova, C. Richter, V. Linhard, H. Schwalbe, and R. Schweitzer-Stenner, "Randomizing the unfolded state of peptides (and proteins) by nearest neighbor interactions between unlike residues," *Chemistry*, **2015**, 21, 5173-5192.

Conference contributions

- 24-29.03.2019 SFB 807 – Retreat: Transport and Communication across Biological Membranes, Obergurgl, Austria
Talk: “Light dynamic of retinal disease relevant G90D mutant of bovine rhodopsin”.
- 4-5.03.2019 SPP 2002, Progress Report, Sylt
Talk and Poster: “Rapid NMR and biophysical characterization of the structure and dynamics of small proteins”.
- 29.04-4.05.2018 ENC, 59th Experimental Nuclear Magnetic Resonance Conference, Orlando, Florida
Poster: “Light dynamic of retinal disease relevant rhodopsin G90D mutant”.
- 18.03-22.03.2018 SFB 807 – Retreat: Transport and Communication across Biological Membranes, Hirschegg, Austria
Talk: “Light dynamic of retinal disease relevant rhodopsin G90D mutant”.
- 13.02-15.02.2018 SPP 2002, Kick-Off meeting, Kiel
Poster: “Rapid NMR and biophysical characterization of structure and dynamics of small proteins”.
- 15.01-19.01.2018 Collaboration visit Vladislav Orekhov Group, Swedish NMR Center, University of Gothenburg
- 4.10-6.10.2017 CRC 807 Symposium – New Horizons in Membrane Transport and Communication, Frankfurt am Main
Poster: “Light dynamic of retinal disease relevant rhodopsin G90D mutant”.
- 5.08-12.08.2017 EMBO practical course, Basel
Poster: “NMR and biophysical characterization of small proteins and light dynamics of rhodopsin mutant”.
- 2-6.07.2017 EUROMAR, Warsaw
Poster: “NMR and biophysical characterization of structure and dynamics of small proteins and their complexes”.
- 30-31.03.2017 Review meeting of the Priority Program SPP 2002 “Small Proteins in Prokaryotes, an Unexplored World” Gustav Stresemann Institut e.V. Bonn
- 7-10.09.2015 GDCh FGNMR 37th Annual Meeting, Darmstadt
Poster: “Detection of fast exchangeable unpaired imino protons in RNA with chemical exchange saturation transfer experiments”
- 4-5.11.2014 Bruker user meeting, Ettlingen

Curriculum vitae

Personal details

Name Nina Kubatova
Address Altkoenigstrasse 150, 61440,
Oberursel (Taunus), Germany
Date of Birth 1.09.1988
Place of Birth Odessa, Ukraine
Nationality Ukrainian
E-mail Kubatova@nmr.uni-frankfurt.de



Education

12/2014-today PhD in Chemistry, Goethe University Frankfurt am Main, Germany
Title: *NMR spectroscopy to study dynamic protein structures*
Supervisor: Prof. Dr. Harald Schwalbe
4/2012-12/2014 Master of Science in Chemistry,
Goethe University Frankfurt am Main
4/2008-3/2012 Bachelor of Science in Chemistry,
Goethe University Frankfurt am Main
9/2005-8/2007 Studies in Chemistry, Odessa National Polytechnic University, Ukraine
9/1998-6/2005 Gymnasium N°3, Odessa, Ukraine

Courses and workshops

25.09-29.09.2017 Biomolecular NMR: modern tools for data processing and interpretation dynamics, Swedish NMR Center, University of Gothenburg
5.08-12.08.2017 EMBO practical course, Basel
29.02-4.03.2016 2nd G-NMR School, Goethe University Frankfurt am Main
4/2014-3/2015 Scientific English course, Goethe University Frankfurt am Main

Further experience

4/2015-today	Assistant at the Organic Chemistry practical course for (bio)chemistry students, Goethe University Frankfurt am Main
4/2012-3/2015	Assistant at the Physical Chemistry theoretical course for chemistry students, Goethe University Frankfurt am Main

Languages

Russian	native
German	fluent
English	fluent

PDB entries

6Q2Z	NMR solution structure of the HVO_2922 protein from Haloferax volcanii
------	--

BMRB entries

34334	NMR solution structure of the HVO_2922 protein from Haloferax volcanii
-------	--

List of academic teachers on the Goethe University, Frankfurt

Prof. Dr. Wolf Aßmus, Dr. Markus Braun, Prof. Dr. Bernd Brutschy, Prof. Dr. Andreas Dreuw, Prof. Dr. Ernst Egert, Prof. Dr. Michael Göbel, Prof. Dr. Alexander Heckel, Dr. Rainer Hegger, Prof. Dr. Max C. Holthausen, Prof. Dr. Bernd O. Kolbesen, Prof. Dr. Thomas Prisner, Prof. Dr. Martin U. Schmidt, Prof. Dr. Gerhard Stock, Prof. Dr. Harald Schwalbe, Prof. Dr. Andreas Terfort, Prof. Dr. Josef Wachtveit, Prof. Dr. Matthias Wagner.

